Computational Statistics and Data Analysis 136 (2019) 12–29

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Sparse wavelet estimation in quantile regression with multiple functional predictors

Dengdeng Yu[1], Li Zhang[1], Ivan Mizera, Bei Jiang, Linglong Kong *

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada, T6G2G1

ARTICLE INFO

Article history:
Received 30 March 2018
Received in revised form 5 August 2018
Accepted 6 December 2018
Available online 24 January 2019

Keywords:
Functional data analysis
Sparse group lasso
ADMM
Convergence rate
Prediction error bound
ADHD

ABSTRACT

To study quantile regression in partial functional linear model where response is scalar and predictors include both scalars and multiple functions, wavelet basis are usually adopted to better approximate functional slopes while effectively detect local features. The sparse group lasso penalty is imposed to select important functional predictors while capture shared information among them. The estimation problem can be reformulated into a standard second-order cone program and then solved by an interior point method. A novel algorithm is proposed by using alternating direction method of multipliers (ADMM) which was recently employed by many researchers in solving penalized quantile regression problems. The asymptotic properties such as the convergence rate and prediction error bound have been established. Simulations and a real data from ADHD-200 fMRI data are investigated to show the superiority of our proposed method.

1. Introduction

Functional data analysis (FDA) is about the analysis of information on curves, images, functions, or more general objects. It has become a major branch of nonparametric statistics and is a fast evolving area as more data has arisen where the primary object of observation can be viewed as a function (Ramsay, 2006; Wang et al., 2015; Morris, 2015). A standard functional linear model with scalar response and functional covariate is

$$y = \alpha + \int_0^1 x(t)\beta(t)dt + \varepsilon, \tag{1}$$

where the coefficient $\beta(t)$ is a function, and ε is a random error. To estimate the functional coefficient $\beta(t)$, we can use functional basis to approximate it. There are three major choices of functional basis: general basis such as B-spline basis and wavelet basis (Cardot et al., 2003; Zhao et al., 2012), functional principal component basis (Cardot et al., 1999; Cai and Hall, 2006; Müller and Yao, 2008; Kong et al., 2016), and partial least square basis (Delaigle and Hall, 2012). Recently in imaging analysis, Zhao et al. (2012), Wang et al. (2014) and Zhao et al. (2015) successfully adopted wavelet basis with regularizations to estimate the functional slope where the functional covariates are image features located in 1D, 2D and 3D domains respectively.

The functional linear model (1) can be extended to a partial functional linear model with multiple functional covariates

$$y = \alpha + \int_0^1 \boldsymbol{x}^T(t)\boldsymbol{\beta}(t)dt + \boldsymbol{u}^T\boldsymbol{\gamma} + \varepsilon, \tag{2}$$

* Corresponding author.
E-mail address: lkong@ualberta.ca (L. Kong).

[1] These authors contributed equally.

https://doi.org/10.1016/j.csda.2018.12.002
0167-9473/

where covariates $\boldsymbol{u}$ are scalars and $\boldsymbol{\gamma}$ are the coefficients. The functional coefficients $\boldsymbol{\beta}(t)$ can be estimated by using regularization techniques. In particular, penalized principal component basis has been an especially popular choice (Gertheiss et al., 2013; Lian, 2013). Recently, Kong et al. (2016) successfully applied such technique to model (2) in the setting of ultrahigh-dimensional scalar predictors.

In recent years, quantile regression, which was introduced by the seminal work of Koenker and Bassett (1978), has been well developed and recognized in functional linear regression. Many of them are focusing on the functional linear quantile regression model:

$$Q_\tau(y|x(t)) = \alpha_\tau + \int_0^1 x(t)\beta_\tau(t)dt, \tag{3}$$

where $Q_\tau(y|x(t))$ is the τ-th conditional quantile of response y given a functional covariate $x(t)$ for a fixed quantile level $\tau \in (0, 1)$. As an alternative to least squares regression, the quantile regression method is more efficient and robust when the responses are non-normal, errors are heavy tailed or outliers are present. It is also capable of dealing with the heteroscedasticity issues and providing a more complete picture of the response (Koenker, 2005). To estimate the functional coefficient $\beta_\tau(t)$, functional basis can as well be used to approximate it; for instance, general basis like B-spline basis (Cardot et al., 2005; Sun, 2005), functional principle component basis (Kato, 2012; Lu et al., 2014; Tang and Cheng, 2014) and partial quantile basis (Yu et al., 2016).

In this article, we extend model (3) to a partial functional linear quantile regression model with multiple functional covariates

$$Q_\tau(y|\,\boldsymbol{u}, \boldsymbol{x}(t)) = \alpha_\tau + \int_0^1 \boldsymbol{x}^T(t)\boldsymbol{\beta}_\tau(t)dt + \boldsymbol{u}^T\boldsymbol{\gamma}_\tau, \tag{4}$$

where $Q_\tau(y|\,\boldsymbol{u}, \boldsymbol{x}(t))$ is the τ-th conditional quantile of y given scalar covariates $\boldsymbol{u}$ and multiple functions $\boldsymbol{x}(t)$. To our best knowledge, only a few works have studied this model; for example, Yu et al. (2016) used partial quantile basis while (Yao et al., 2017) used penalized principal component basis. Inspired by the success of wavelet basis with regularization in functional linear model (Zhao et al., 2012; Wang et al., 2014; Zhao et al., 2015), we use it to approximate the functional coefficients $\boldsymbol{\beta}_\tau(t)$ in model (4). Wavelet basis can provide a good representation of functional coefficients by using only a small number of basis and are particularly useful for capturing localized functional features. Moreover, the wavelet transform is computationally efficient and hence suitable for dealing with multiple functional predictors.

The penalization we impose is sparse group lasso (Zhao et al., 2014; Simon et al., 2013), which is motivated by the attention deficit hyperactivity disorder (ADHD) study from the ADHD-200 Sample Initiative Project. Our goal is to predict ADHD index at various quantile levels by using both demographic information and functional magnetic resonance imaging (fMRI) data, where the fMRI data consists of 116 functional features, each of which represents a single region of interests (ROI) of human brain. The sparse group lasso technique, by imposing a convex combination of lasso and group lasso penalties, can select important ROIs while capture shared information among them. More specifically, the group lasso penalty makes a sparse selection out of 116 functional features of ROIs, while the lasso penalty induces a sparse representation of each feature. Common wavelet basis is used to represent different features so that the shared information among them can be captured.

There are five major contributions of this paper. First, our conditional quantile framework provides a more suitable modeling of reality especially when the response is heavy tailed (Yao et al., 2017). It is also a compelling choice of dealing with heteroscedasticity issues and can provide a more complete picture of the response (Koenker, 2005). Second, the wavelet basis we adopt provides a good approximation of functional coefficients while effectively detects the local features. The wavelet transform we use is computationally efficient and hence can be easily extended to deal with multiple functional predictors. Third, the proposed sparse group lasso method selects important functional predictors and retains shared information among them as well. It is extremely useful in ADHD-200 fMRI study so that both individual and common information can be captured among the different ROIs. Fourth, the estimation problem is in fact a penalized quantile regression problem, which can be reformulated into a second-order cone program and then easily solved by an interior point method implemented by a powerful R package: Rmosek. We also propose a novel algorithm to solve it by using alternating direction method of multipliers (ADMM). Fifth, we successfully derive the asymptotic properties including the convergence rate and prediction error bound which theoretically warrants good performance of our estimates.

The rest of paper is organized as follows. In Section 2, we review some necessary background on wavelets and provide the penalized quantile objective function with sparse group lasso penalty. The asymptotic properties such as the convergence rate and predictor error bound are established in Section 3. In Section 4, the quantile penalization problem is reformulated into a second-order cone program (SOCP) and solved by an interior point method by using a powerful R package: Rmosek. We also propose a novel algorithm using alternating direction method of multipliers (ADMM). Finite sample simulations and a real data from ADHD-200 fMRI data are investigated in Section 5 to illustrate the superiority of our proposed method.

2. Wavelet-based sparse group lasso

In this section, we first review some necessary background on wavelets. We then provide the penalized quantile objective function with sparse group lasso penalty where the functional coefficients are approximated by wavelet basis. This leads to the sparsities of both the selection and representation of functional features. More specifically, the group lasso selects a sparse set from available functional features, while the lasso induces a sparse representation of the selected functional features.

2.1. Some background on wavelets

Wavelets are basis function that can provide a good approximation of functional coefficients while effectively capture the local features (Zhao et al., 2012). Moreover, the wavelet transform is computationally efficient and hence can be easily extended to deal with multiple functional predictors (Daubechies, 1990). For a given $\tau \in (0, 1)$, let $\beta_{l\tau}(t)$ be one component of $\boldsymbol{\beta}_\tau(t)$ in (4), where $\boldsymbol{\beta}_\tau(t) = (\beta_{1\tau}(t), \ldots, \beta_{m\tau}(t))^T$. Suppose that $\beta_{l\tau}(t)$ is in $L^2[0, 1]$. We can approximate it using wavelet basis. For any wavelet basis in $L^2[0, 1]$, they can be derived by dilating and translating two orthonormal basic functions: a scaling function and a wavelet function, namely $\phi(t)$ and $\psi(t)$ respectively:

$$\varphi_{jk}(t) = \sqrt{2^j}\varphi(2^j t - k), \quad \psi_{jk}(t) = \sqrt{2^j}\psi(2^j t - k),$$

where j and k are integers, $\int_0^1 \varphi(t) = 1$ and $\int_0^1 \psi(t) = 0$. In particular, given a primary resolution level j_0, the wavelet basis are

$$\{\varphi_{j_0,k}\}_{0\le k\le 2^{j_0}-1} \quad \text{and} \quad \{\psi_{j,k}\}_{j_0\le j,\ 0\le k\le 2^j-1}. \tag{5}$$

Therefore, $\beta_{l\tau}(t)$ can be approximated by

$$\beta_{l\tau}(t) = \sum_{k=0}^{2^{j_0}-1} a^l_{j_0k}\varphi_{j_0k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} d^l_{jk}\psi_{jk}(t), \quad \text{for } l = 1, \ldots, m, \tag{6}$$

where $a^l_{j_0k} = \int_0^1 \beta_{l\tau}(t)\varphi_{j_0,k}(t)dt$ is the approximation coefficients at the coarsest resolution j_0, and $d^l_{jk} = \int_0^1 \beta_{l\tau}(t)\psi_{jk}(t)dt$ is the detail coefficients characterizing the fine structures.

In practice, the functional covariates $\boldsymbol{x}(t) = (x_1(t), \ldots, x_m(t))^T$ are discretely observed, for instance without loss of generality, at $N = 2^J$ equally spaced points of $[0, 1]$ with $0 = t_1 < t_2 < \cdots < t_N = 1$. Typically, N can go to infinity as n increases, hence it can be substituted by N_n as a more general case, the discussions of which are deferred to the next section. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ and $\boldsymbol{\beta}_\tau = (\boldsymbol{\beta}_{1\tau}, \ldots, \boldsymbol{\beta}_{m\tau})$, where $\boldsymbol{x}_l = (x_1(t_1), \ldots, x_m(t_N))^T$, $\boldsymbol{\beta}_{l\tau} = (\beta_{l\tau}(t_1), \ldots, \beta_{l\tau}(t_N))^T$ and $l = 1, \ldots, m$. We represent $\boldsymbol{X}$ and $\boldsymbol{\beta}_\tau$ by the wavelet coefficients through discrete wavelet transform (DWT). In particular, let $\boldsymbol{W}$ be an $N \times N$ matrix associated with orthonormal wavelet basis derived from DWT. Suppose $\boldsymbol{C}$ and $\boldsymbol{B}$ are the corresponding wavelet coefficients of $\boldsymbol{X}$ and $\boldsymbol{\beta}_\tau$. Then we have $\boldsymbol{X} = \boldsymbol{W}^T\boldsymbol{C}$, $\boldsymbol{\beta}_\tau = \boldsymbol{W}^T\boldsymbol{B}_\tau$, and the integration in model (4):

$$\int_0^1 \boldsymbol{x}^T(t)\boldsymbol{\beta}_\tau(t)dt \approx \mathrm{vec}(\boldsymbol{X})^T\mathrm{vec}(\boldsymbol{\beta}_\tau)/N = \mathrm{vec}(\boldsymbol{W}^T\boldsymbol{C})^T\mathrm{vec}(\boldsymbol{W}^T\boldsymbol{B})/N = \mathrm{vec}(\boldsymbol{C})^T\mathrm{vec}(\boldsymbol{B})/N.$$

The last equality holds due to the orthonormality of $\mathbf{W}$. From now on, we denote $\boldsymbol{v} = \mathrm{vec}(\boldsymbol{C})^T/N$ and $\boldsymbol{\theta}_\tau = \mathrm{vec}(\boldsymbol{B})$ where $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m)$ and $\boldsymbol{B} = (\boldsymbol{b}_{1\tau}, \ldots, \boldsymbol{b}_{m\tau})$.

2.2. Model estimation

Using wavelet basis by DWT, model (4) becomes

$$Q_\tau(y|\boldsymbol{u}, \boldsymbol{x}(t)) \approx \alpha_\tau + \boldsymbol{v}^T\boldsymbol{\theta}_\tau + \boldsymbol{u}^T\boldsymbol{\gamma}_\tau. \tag{7}$$

Given n identical copies of data triplets $(\boldsymbol{X}_i, \boldsymbol{u}_i, y_i)$, where $\boldsymbol{X}_i$ and $\boldsymbol{u}_i$ are the observed functional and scalar covariates respectively, and y_i is the corresponding response, the parameters in (7) can be estimated by minimizing a regular quantile loss function. However, to find the important functional covariates in predicting responses while preserve a desired sparse representation of the coefficients, an appropriate penalty has to be imposed. In this paper, we propose to use the sparse group lasso penalty

$$P_{\lambda_1,\lambda_2}(\boldsymbol{\theta}) = \lambda_1\sum_{l=1}^m \|\boldsymbol{b}_l\|_1 + \lambda_2\sum_{l=1}^m \|\boldsymbol{b}_l\|_2, \tag{8}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ represent the L_1 and L_2 norms respectively, and λ_1 and λ_2 are two nonnegative tuning parameters. The sparse group lasso penalty includes two components, namely a lasso and a group lasso penalties, where the lasso penalty $\|\cdot\|_1$ induces sparsity in each functional coefficient and the group lasso penalty $\|\cdot\|_2$ selects functional coefficients. Common information among functional covariates can be retained by using the same wavelet basis to approximate the functional coefficients. Moreover, the sparse group lasso warrants the selection of important functional coefficients while captures distinct traits carried by individual functional covariates. Specifically, the parameters α_τ, $\boldsymbol{\gamma}_\tau$, and $\boldsymbol{\theta}_\tau$ can be estimated by

$$(\hat{\alpha}_\tau, \hat{\boldsymbol{\gamma}}_\tau, \hat{\boldsymbol{\theta}}_\tau) = \arg\min_{\alpha,\boldsymbol{\gamma},\boldsymbol{\theta}} \sum_{i=1}^n \rho_\tau\left(y_i - \alpha - \boldsymbol{u}_i^T\boldsymbol{\gamma} - \boldsymbol{v}_i^T\boldsymbol{\theta}\right) + P_{\lambda_1,\lambda_2}(\boldsymbol{\theta}), \tag{9}$$

where $\rho_\tau(x) = x(\tau - \mathbf{1}(x < 0))$ is the quantile check function (Koenker, 2005).

To combine information from different quantiles, Zou and Yuan (2008) proposed composite quantile regression, which simultaneously considers multiple regression quantiles at different levels. With homoscedasticity assumption, where all

conditional regression quantiles have the same slope, the composite quantile estimate is more efficient than the one from a single level and has in recent years begun to gain its popularity in many fields (Kai et al., 2010; Fan and Lv, 2010; Bradic et al., 2011; Kai et al., 2011; Yu et al., 2016). In this paper, we propose to use composite quantile regression with sparse group lasso penalty in our functional data analysis framework. Let $0 < \tau_1 < \cdots < \tau_k < 1$ denote the selected quantile levels and then the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ can be estimated by

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\theta}}{\arg\min} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - \boldsymbol{u}_i^T \boldsymbol{\gamma} - \boldsymbol{v}_i^T \boldsymbol{\theta}) + P_{\lambda_1,\lambda_2}(\boldsymbol{\theta}), \tag{10}$$

where $\boldsymbol{\alpha} = (\alpha_1 \ldots, \alpha_K)$ is a vector of intercepts. Typically, we can choose $K = 9$ and use equally spaced quantiles (Kai et al., 2010; Zou and Yuan, 2008). Note that quantile estimate (9) at a single level is just a special case of composite quantile estimate (10) with $K = 1$. In the following, we will focus on the composite quantile regression case of (10).

3. Asymptotics

In this section, we investigate the asymptotic properties of our proposed estimates when both the sample size n and the number of discrete points N_n, where the functional covariates are observed, tend to infinity. Let $\lambda_{1,n}$ and $\lambda_{2,n}$ denote the tuning parameters when the sample size is n. To derive the asymptotic properties, we impose the following conditions:

A1. The model errors $\varepsilon_1, \ldots, \varepsilon_n$ are independently following a distribution F, with density f to be bounded away from zero and infinity, and its derivative f' to be continuous and uniformly bounded at their τ-th quantiles.

A2. There exist a constants c_1 such that

$$0 < \rho_n \leq \rho_{\max}(\frac{1}{n}\boldsymbol{A}_n^T \boldsymbol{A}_n) < c_1 < \infty,$$

where $\boldsymbol{A}_n = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)^T$ is the design matrix with $\boldsymbol{a}_i = (1, \boldsymbol{v}_i^T, \boldsymbol{u}_i^T)^T$, and we denote by ρ_n the smallest eigenvalue of $\frac{1}{n}\boldsymbol{A}_n^T \boldsymbol{A}_n$ and $\rho_{\max}(\cdot)$ is the largest eigenvalue of $\frac{1}{n}\boldsymbol{A}_n^T \boldsymbol{A}_n$ respectively.

A3. There exists a constant M such that $\|\boldsymbol{a}_i\|_2 < M$ for all i.

A4. The functional slope $\beta_l(t)$s are d times differentiable in the Sobolev sense, and the wavelet basis has w vanishing moments, where $w > d$.

A5. $\lambda_{1,n} = O(\sqrt{n\rho_n})$ and $\lambda_{2,n} = O(\sqrt{n\rho_n})$.

A6. $\frac{N_n}{n\rho_n} \to 0$.

These regularity conditions might not be the weakest ones but are commonly assumed among literatures of quantile regression and functional linear model. Condition (A1) is standard for quantile regression (Koenker, 2005; Zhao et al., 2014), which regulates the behavior of the conditional density of the response in a neighborhood of the conditional quantile and is crucial to the asymptotic properties of quantile estimators (Koenker and Bassett, 1978). Condition (A2) is a classical condition in functional linear regression literature (Delaigle and Hall, 2012). It ensures the eigenvalues of the covariance matrix go to neither zero nor infinity too quickly. Similar conditions as (A3) - (A6) can be found in Zhao et al. (2012, 2015), among others. Condition (A4) guarantees that the space spanned by the wavelet basis can well approximate the functional slopes with small approximation errors. Condition (A6) implies that to allow for estimation of β with appropriate asymptotic properties, n should grow faster than N_n. Note the wavelet basis has w vanishing moments if and only if its scaling function φ can generate polynomials of degree at most w.

Theorem 3.1. *Let $\hat{\beta}_{l,n}$ be the estimator resulting from (10) and β_l is the true coefficient function. If Conditions (A1)–(A6) hold, then*

$$\|\hat{\beta}_{l,n} - \beta_l\|_2^2 = O_p\left(\frac{N_n}{n\rho_n}\right) + o_p\left(\frac{1}{N_n^{2d}}\right).$$

A detailed proof of this theorem is provided in the Appendix. The accuracy of $\hat{\beta}$ relies on both n , the smallest eigenvalue and N_n. The approximation error rate of $\hat{\beta}$ towards β are controlled by two terms. The first term is of the same order of $N/(n\rho_n)$ which is a typical result of estimating, while the second term is of the lower order of $1/N_n^{2d}$ which is mainly due to approximation by wavelets. In particular, the approximation error rate is dominated by the second term if N_n^{2d+1} is of the lower order of n. Otherwise, it is dominated by the first term. Under some further conditions, we can have the following theorem for the prediction error bound:

Theorem 3.2. *Suppose $x_l(t)$ is square integrable on $[0, 1]$ and $\boldsymbol{F}^{-1}(\tau) = 0$. If Conditions (A1)–(A6) hold, then*

$$\|\hat{y} - Q_\tau(y|\, \boldsymbol{u}, \boldsymbol{x}(t))\|_2^2 = O_p(\frac{N_n}{n\rho_n}) + o_p(\frac{1}{N_n^{2d}}),$$

where $Q_\tau(y|\, \boldsymbol{u}, \boldsymbol{x}(t))$ and $\hat{y}$ are the true and estimated τ-th conditional quantile, respectively.

The proof follows that from Theorem 3.1 and the Cauchy–Schwarz inequality, the details of which are omitted in this paper. Similarly as in Theorem 3.1, L_2 prediction error rate depends on the same two terms from estimating and approximation by wavelets respectively, while the estimation errors caused by $\hat{\alpha}$ and $\hat{\gamma}$ is absorbed by the first term. The following lemma establishes the sparsity property of our estimator. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_{20})$ where $\boldsymbol{\theta}_{10}$ be the nonzero coefficients and $\boldsymbol{\theta}_{20} = 0$, then we have the following results.

Theorem 3.3. *If conditions (A1)–(A6) hold, then with probability tending to one, for any given $\boldsymbol{\theta}_1$ satisfying $\|(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\theta}_1) - (\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \boldsymbol{\theta}_{10})\| = O_P(\frac{N_n}{n\rho_n})$ and any constant C,*

$$Q((\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\theta}_1, \mathbf{0})) = \min_{\|\boldsymbol{\theta}_2\| \le C\sqrt{N_n/(n\rho_n)}} Q(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

The proof can easily follow the proof of Lemma 1 in Wu and Liu (2009). Therefore, from Theorem 3.3, the model selection consistency has been established.

4. Implementations

Due to the non-smoothness of loss function, quantile estimator does not enjoy the nice asymptotic properties, as well as computational easiness, as what ordinary least square estimator does. After illustrating asymptotic theory of the proposed quantile estimator, it becomes of great importance to have an efficient algorithm to obtain it. In this section, we reformulate the optimization problem (10) into a second-order cone program (SOCP) and implement it by interior point method using a powerful R package: **Rmosek** (Aps, 2015). Alternatively we propose a novel algorithm to solve problem (10) by using alternating direction method of multipliers (ADMM) which was a technique recently employed by many researchers in solving penalized quantile regression problems. In the end, we discuss some practical rules to choose tuning parameters.

4.1. A second-order cone program

Let the superscripts $^+$ and $^-$ denote the positive and negative parts of a vector. For unknown parameter $\boldsymbol{\theta}$ in (10), we write: $\boldsymbol{\theta} = \boldsymbol{\theta}^+ - \boldsymbol{\theta}^-$ and $\|\boldsymbol{\theta}\|_1 = \|\boldsymbol{\theta}^+\|_1 + \|\boldsymbol{\theta}^-\|_1$. Similarly, we have $\boldsymbol{b} = \boldsymbol{b}^+ - \boldsymbol{b}^-$ and $\|\boldsymbol{b}\|_1 = \|\boldsymbol{b}^+\|_1 + \|\boldsymbol{b}^-\|_1$. Then problem (10) can be reformulated as the following standard second-order cone program:

$$\begin{aligned} \min \ & \sum_{k=1}^{K}\sum_{i=1}^{n}(\tau_k r_{ki}^+ + (1-\tau_k) r_{ki}^-) + \lambda_1 \sum_{l=1}^{m}(\|\boldsymbol{b}_l^+\|_1 + \|\boldsymbol{b}_l^-\|_1) + \lambda_2 \sum_{l=1}^{m} z_l \\ \text{subject to } & -r_{ki}^- \le y_i - \alpha_k - \boldsymbol{u}_i^T\boldsymbol{\gamma} - \boldsymbol{v}_i^T(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) \le r_{ki}^+ \\ & \sqrt{\|\boldsymbol{b}_l^+\|_2^2 + \|\boldsymbol{b}_l^-\|_2^2} \le z_l \\ & \boldsymbol{\theta}^+ \succeq 0,\ \boldsymbol{\theta}^- \succeq 0,\ z_l \ge 0,\ r_{ki}^+ \ge 0,\ r_{ki}^- \ge 0. \end{aligned} \tag{11}$$

where r_{ki}^+, r_{ki}^- and z_l are three nonnegative slack variables, and the constraint $\sqrt{\|\boldsymbol{b}_l^+\|_2^2 + \|\boldsymbol{b}_l^-\|_2^2} \le z_l$ implies a second order cone of dimension $2N + 1$ (Lobo et al., 1998) denoted as

$$\mathbb{Q}_l^{2N+1} = \left\{ (z_l, \boldsymbol{b}_l^+, \boldsymbol{b}_l^-) \in \mathbb{R}^{2N+1} \,\middle|\, z_l \ge \sqrt{\|\boldsymbol{b}_l^+\|_2^2 + \|\boldsymbol{b}_l^-\|_2^2} \right\}.$$

The reformulation is guaranteed by the fact that for each component of optimal $\boldsymbol{b}_l$, either $b_{l,j}^+ = 0$ or $b_{l,j}^- = 0$ would be held. Otherwise, for optimal $\boldsymbol{b}_l$, if there exist l and j_0 such that $b_{l,j_0}^+ > 0$ and $b_{l,j_0}^- > 0$, we can replace b_{l,j_0}^+ and b_{l,j_0}^- by $b_{l,j_0}^{(\text{new})+}$ and $b_{l,j_0}^{(\text{new})-}$ respectively with

$$b_{l,j_0}^{(\text{new})+} = \begin{cases} 0 & \text{if } b_{l,j_0}^+ < b_{l,j_0}^-, \\ b_{l,j_0}^+ - b_{l,j_0}^- & \text{otherwise,} \end{cases} \qquad b_{l,j_0}^{(\text{new})-} = \begin{cases} 0 & \text{if } b_{l,j_0}^+ > b_{l,j_0}^-, \\ b_{l,j_0}^- - b_{l,j_0}^+ & \text{otherwise.} \end{cases}$$

As a result, the objective function in (11) decreases, which contradicts with the fact that $\boldsymbol{b}_l$ being optimal.

Various optimization strategies can be applied to solve SOCP (11) such as interior point method (Koenker and Park, 1996) and the simplex method (Koenker, 2005). In this paper, we choose to use interior point method. The R package we use is **Rmosek** (Aps, 2015). The technique proposed to reformulate our problem into a SOCP can be easily adapted to other penalized quantile regression problems; for example, quantile ridge regression (Wu and Liu, 2009).

4.2. ADMM algorithm

Although problem (10) is convex, solving it can be very slow partially due to large scale data in the application and the non-smooth terms in the objective that prevent fast gradient method being applied. In this section, we explore the additive structure of the objective function, namely, decompose it into two sub convex problems, and then propose a novel and

efficient algorithm by using alternating direction method of multipliers (ADMM) (Gabay and Mercier, 1976). This powerful tool was originated in 1950s and developed during 1970s (Hestenes, 1969; Gabay and Mercier, 1976). It has been popularized in recent years among quantile regression literature (Boyd et al., 2011; Gao and Kong, 2015; Kong et al., 2015).

Denote $L_n(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{k=1}^{K}\sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - \boldsymbol{u}_i^T\boldsymbol{\gamma} - \boldsymbol{v}_i^T\boldsymbol{\theta})$. The minimization problem (10) can be rewritten as

$$\begin{aligned} \min \quad & L_n(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\gamma}) + P_{\lambda_1,\lambda_2}(\boldsymbol{\theta}^*) \\ \text{subject to} \quad & \boldsymbol{\theta} = \boldsymbol{\theta}^*, \end{aligned}$$

where $L_n(\cdot)$ and $P_{\lambda_1,\lambda_2}(\cdot)$ are two convex functions. Applying augmented lagrangian (Hestenes, 1969), we have

$$L_{n,\eta}(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\theta}^*, \boldsymbol{\mu}) = L_n(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\gamma}) + P_{\lambda_1,\lambda_2}(\boldsymbol{\theta}^*) + \boldsymbol{\mu}^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{\eta}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2. \tag{12}$$

Let $\boldsymbol{w} = \boldsymbol{\mu}/\eta$. The ADMM algorithm to obtain the minimizer of (12) follows a three-step iterative scheme:

$$\begin{aligned} (\boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\theta}^{(l+1)}, \boldsymbol{\gamma}^{(l+1)}) &= \underset{\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\gamma}}{\operatorname{argmin}}\, L_n(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\gamma}) + \frac{\eta}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{*(l)} + \boldsymbol{w}^{(l)}\|_2^2 \\ \boldsymbol{\theta}^{*(l+1)} &= \underset{\theta^*}{\operatorname{argmin}}\, P_{\lambda_1,\lambda_2}(\boldsymbol{\theta}^*) + \frac{\eta}{2}\|\boldsymbol{\theta}^{(l+1)} - \boldsymbol{\theta}^* + \boldsymbol{w}^{(l)}\|_2^2 \\ \boldsymbol{w}^{(l+1)} &= \boldsymbol{w}^{(l)} + \eta(\boldsymbol{\theta}^{(l+1)} - \boldsymbol{\theta}^{*(l+1)}). \end{aligned} \tag{13}$$

For the first step of (13), it can be reformulated as a SOCP:

$$\begin{aligned} \min \quad & \sum_{k=1}^{K}\sum_{i=1}^{n} \rho_{\tau_k}(r_{ik}) + \frac{\eta}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{*(l)} + \boldsymbol{w}^{(l)}\|_2^2 \\ \text{subject to} \quad & y_i - \alpha_k - \boldsymbol{u}_i^T\boldsymbol{\gamma} - \boldsymbol{v}_i^T\boldsymbol{\theta} = r_{ik}, \quad \text{for } i = 1, \ldots, n; \quad k = 1, \ldots, K, \end{aligned}$$

which can be easily solved by following an ADMM scheme:

$$\begin{aligned} r_{ik}^{(j+1)} &= \underset{r_{ik}}{\operatorname{argmin}}\, \rho_{\tau_k}(r_{ik}) + \frac{\eta_1}{2}(y_i - \alpha_k^{(j)} - \boldsymbol{u}_i^T\boldsymbol{\gamma}^{(j)} - \boldsymbol{v}_i^T\boldsymbol{\theta}^{(j)} - r_{ik} + z_{ik}^{(j)})^2 \\ (\boldsymbol{\alpha}^{(j+1)}, \boldsymbol{\theta}^{(j+1)}, \boldsymbol{\gamma}^{(j+1)}) &= \underset{\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\gamma}}{\operatorname{argmin}}\, \frac{\eta}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{*(l)} + \boldsymbol{w}^{(l)}\|_2^2 + \frac{\eta_1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}(y_i - \alpha_k - \boldsymbol{u}_i^T\boldsymbol{\gamma} - \boldsymbol{v}_i^T\boldsymbol{\theta} - r_{ik}^{(j+1)} + z_{ik}^{(j)})^2 \\ z_{ik}^{(j+1)} &= z_{ik}^{(j)} + \eta_1(y_i - \alpha_k^{(j+1)} - \boldsymbol{u}_i^T\boldsymbol{\gamma}^{(j+1)} - \boldsymbol{v}_i^T\boldsymbol{\theta}^{(j+1)} - z_{ik}^{(j+1)}). \end{aligned} \tag{14}$$

The first step of (14) can be explicitly solved by the soft thresholding operator. The second step can be easily approximated by a standard ridge regression therefore has a closed form.

The second step of (13) can be simplified by the soft thresholding operator. That is,

$$\begin{aligned} \boldsymbol{v}^* &= sgn(\boldsymbol{\theta}^{(l+1)} + \boldsymbol{w}^{(l)}) \cdot \max(|\boldsymbol{\theta}^{(l+1)} + \boldsymbol{w}^{(l)}| - \frac{\lambda_1}{\eta},\ 0) \\ \boldsymbol{\theta}^{*(l+1)} &= \frac{\boldsymbol{v}^*}{\|\boldsymbol{v}^*\|_2}\max(\|\boldsymbol{v}^*\|_2 - \frac{\lambda_2}{\eta}, 0), \end{aligned}$$

where $sgn(\cdot)$ is the sign function.

A typical stopping criterion with primal and dual residuals denoted respectively by r_{primal} and r_{dual} (Boyd et al., 2011) can be chosen as:

$$\|\boldsymbol{\theta}^{(l)} - \boldsymbol{\theta}^{*(l)}\|_2 \le r_{\text{primal}} \quad \text{and} \quad \|\eta(\boldsymbol{\theta}^{*(l)} - \boldsymbol{\theta}^{*(l-1)})\|_2 \le r_{\text{dual}},$$

with

$$\begin{aligned} r_{\text{primal}} &= \sqrt{mN}\epsilon_{\text{abs}} + \epsilon_{\text{rel}} \cdot \max\{\|\boldsymbol{\theta}^{(l)}\|_2,\ \|\boldsymbol{\theta}^{*(l)}\|_2\}, \\ r_{\text{dual}} &= \sqrt{mN + q + K}\epsilon_{\text{abs}} + \epsilon_{\text{rel}} \cdot \|\boldsymbol{w}^{(l)}\|_2, \end{aligned}$$

where q is the dimension of $\boldsymbol{\gamma}$, and parameters ϵ_{abs} and ϵ_{rel} are two predefined absolute and relative tolerances which can be set as 10^{-4} and 10^{-2} respectively.

Instead of tackling the original problem directly, ADMM decompose it into several sub convex problems then deal with them separately by iteration. In each iteration, the sub problem can be easily and efficiently solved by the soft thresholding operator or approximated to have a closed form. Therefore, the ADMM algorithm derived is much faster and more efficient than other general techniques.

4.3. Selection of tuning parameters

The proposed method involves selection of two nonnegative tuning parameters, namely λ_1 and λ_2, which control the severity of penalization towards model complexity. Specifically, λ_1 controls sparsity in each functional coefficient while λ_2 controls the number of selected functional coefficients. Although many options exist for selecting tuning parameters, such as AIC, BIC and cross validation, there is no agreed-upon selection criterion in general. After showing that AIC and cross validation may fail to consistently identify the true model, Zhang et al. (2010) proposed to use the generalized information criterion (GIC), encompassing the commonly used AIC and BIC, and illustrated the corresponding asymptotic consistency. More recently, Zheng et al. (2015) used the GIC to make consistent model selection for quantile regression in ultra-high dimensional settings. In this paper, we propose to use the GIC:

$$(\hat{\lambda}_1, \hat{\lambda}_2) = \underset{\lambda_1,\lambda_2}{\arg\min} \frac{1}{K} \sum_{k=1}^{K} \ln\left(\frac{1}{n} \sum_{i=1}^{n} \rho_{\tau_k}\left(y_i - \hat{y}_{ki}\right) \right) + \phi_n \|\hat{\boldsymbol{\theta}}_{\lambda_1,\lambda_2}\|_0, \tag{15}$$

where $\hat{\boldsymbol{\theta}}_{\lambda_1,\lambda_2}$ is a solution of problem (10), $\|\cdot\|_0$ denotes L_0 norm (total number of non-zero elements in a vector), ϕ_n is a sequence converging to zero with n goes to infinity, and $\hat{y}_{ki}$ is calculated from (7) with $\tau = \tau_k$.

In addition, we can also use the validation set (Li et al., 2007; Wu and Liu, 2009) to select gold standard λ_1 and λ_2 that minimize the prediction error. Simulations in Section 5 demonstrate a satisfactory behavior of the proposed criterion compared with the validation set method.

5. Numerical studies

In this section, we compare performances of the proposed sparse group lasso method with group lasso and lasso methods using simulations and a real data from ADHD-200 fMRI sample (Mennes et al., 2013). In addition, we also compared our methods with wavelet versions of elastic net (Zou and Hastie, 2005) and sparse partial least squares (Chun and Keleş, 2010). As pointed out by Luo and Qi (2015); for some other methods such as functional principal components and functional partial least squares, the corresponding R-codes may not be applicable to the case of multiple predictive curves, we omit the comparison with them. We also compare the tuning parameters selected by the GIC approach we proposed and the validation set approach. In our numerical studies, we employ least-asymmetric wavelets of Daubechies with 6 vanishing moments and fix the tuning parameter ratio $\lambda_1/\lambda_2 = 0.5$ (Simon et al., 2013). To simplify notations, we use qSGL, qL, qGL, sPLS and EN to represent the quantile sparse group lasso, lasso, group lasso, sparse partial least squares and elastic net methods respectively.

5.1. Simulations

Our data are randomly generated using 12 functional covariates and 2 scalar covariates in a setting similar to Collazos et al. (2016). In particular, the model is of the form:

$$y_i = \alpha + \boldsymbol{u}_i^T \boldsymbol{\gamma} + \int_0^1 \boldsymbol{x}_i(t)^T \boldsymbol{\beta}(t) dt + \sigma \varepsilon_i \quad \text{for } i = 1 \ldots, n,$$

where $\boldsymbol{u}_i = (u_{i1}, u_{i2})^T$ with $u_{i1} \sim N(0, 1)$ and $u_{i2} \sim$ Bernoulli (0.5), and the coefficients $\boldsymbol{\gamma} = (0.32/256,\ 0.32/256)^T$ and $\alpha = -0.5$. The functional covariates $\boldsymbol{x}_i(t) = (x_{i1}(t), \ldots, x_{i12}(t))^T$ are observed on an equally spaced grid of $N = 256$ points on $[0, 1]$ with

$$\begin{aligned}
x_{i1}(t) &= \sqrt{.84}\omega_{i1}(t) + .4\omega_{i6}(t), \quad x_{i2}(t) = \sqrt{.98}\omega_{i2}(t) + .1\omega_{i1}(t) + .1\omega_{i5}(t),\\
x_{i3}(t) &= \sqrt{.84}\omega_{i3}(t) + .4\omega_{i4}(t), \quad x_{i5}(t) = \sqrt{.99}\omega_{i5}(t) + .1\omega_{i2}(t),\\
x_{il}(t) &= \omega_{il}(t) \quad \text{for } l = 4, 6, 7, \ldots, 12;
\end{aligned}$$

where

$$\omega_{il}(t) = z_{il}(t) + \epsilon_{il},\ \epsilon_{il} \sim N\left(0, \left(.05 r_{x_{il}}\right)^2\right), \quad \text{for } l = 1 \ldots, 12,$$

with $r_{x_{il}} = \max_i (z_{il}(t)) - \min_i (z_{il}(t))$ and

$$\begin{aligned}
z_{i1}(t) &= \cos(2\pi(t - a_1)) + a_2, \mathbb{T}_1 = [0,\ 1],\ a_1 \sim N(-4,\ 3^2),\ a_2 \sim N(7,\ 1.5^2),\\
z_{i2}(t) &= b_1 t^3 + b_2 t^2 + b_3 t, \mathbb{T}_2 = [-1,\ 1],\ b_1 \sim N(-3,\ 1.2^2),\ b_2 \sim N(2, .5^2),\ b_3 \sim N(-2\ , 1),\\
z_{i3}(t) &= \sin(2(t - c_1)) + c_2 t, \mathbb{T}_3 = [0,\ \pi/3],\ c_1 \sim N(-2,\ 1),\ c_2 \sim N(3, 1.5^2),\\
z_{i4}(t) &= d_1 \cos(2t) + d_2 t, \mathbb{T}_4 = [-2,\ 1],\ d_1 \sim U(2,\ 7),\ d_2 \sim N(2, .4^2),\\
z_{i5}(t) &= e_1 \sin(\pi t) + e_2, \mathbb{T}_5 = [0,\ \pi/3],\ e_1 \sim U(3,\ 7),\ e_2 \sim N(0, 1),
\end{aligned}$$

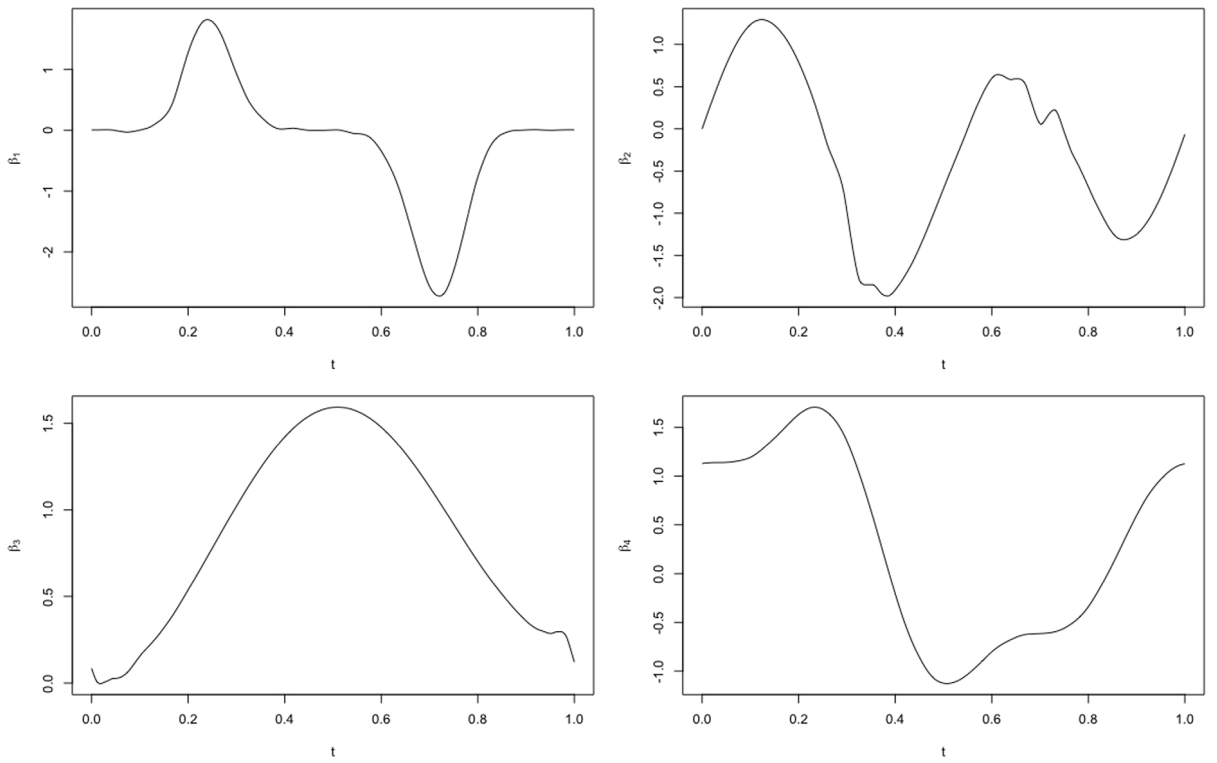

Fig. 1. Slope functions of $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_4$.

$$z_{i6}(t) = f_1 e^{-t/3} + f_2 t + f_3,\ \mathbb{T}_6 = [-1,\ 1],\ f_1 \sim N(4,\ 2^2),\ f_2 \sim N(-3,\ .5^2),\ f_3 \sim N(1,\ 1),$$

$$z_{il}(t) = 5\sqrt{2}\sum_{j=1}^{49}\cos(j\pi t)g_j + 5h,\ \mathbb{T}_l = [0,\ 1],\ g_j \sim N\left(0,\ \ (j+1)^{-2}\right),\ h \sim N(0,\ 1),\ \text{for } l = 7, \ldots, 12.$$

The functional coefficients $\boldsymbol{\beta}(t)$ are generated based on the following 4 functions:

$$\begin{aligned} f_1(t) &= .03f(t, 20, 60) - .05f(t, 50, 20),\\ f_2(t) &= 4\sin(4\pi x) - \text{sign}(x - .3) - \text{sign}(.72 - x),\\ f_3(t) &= -3\cos(2\pi t) + 3e^{t^2}/(t^3 + 1),\\ f_4(t) &= .1\sin(2\pi t) + .2\cos(2\pi t) + .3\sin^2(2\pi t) + .4\cos^3(2\pi t) + .5\sin^3(2\pi t), \end{aligned}$$

where $f(t, \alpha, \beta)$ is the density function for beta distribution: Beta(α, β). Note $f_1(t)$ has also been considered by Zhao et al. (2012); the second function f_2, the so-called "Heavi-Sine" function, is one of test functions from Donoho and Johnstone (1994) which is very popular among wavelet literature (Antoniadis et al., 2001); and f_4 was proposed by Lin et al. (2013).

To generate the functional slopes $\beta_1(t), \ldots \beta_4(t)$, we first apply DWT for $f_1, \ldots, f_4$ and select the wavelet coefficients with absolute values greater than .1; and based on the inverse DWT of the selected coefficients, we generate normalized $\beta_1(t), \ldots \beta_4(t)$, each of which possesses sparsity and is shown in Fig. 1. The rest of slopes are set to be zero, i.e., $\beta_l(t) = 0$ for $l = 5,\ \ldots,\ 12$. The error term ε_i is drawn from the following distributions: (1) Standard normal : $N(0,\ 1)$; (2) Mixed-variance: $.95N(0,\ 1) + .05N(0,\ 10)$; (3) t distribution with 3 degrees of freedom: t_3; (4) Standard Cauchy: $C(0,\ 1)$. The signal-to-noise (SNR) ratio, defined as μ/σ in this paper, is chosen from three different levels: SNR $= 1,\ 5,\ 10$, where μ is the mean of signal and σ is the standard deviation of the noise.

The sizes of the training, tuning and testing data sets are n, n and $10n$ respectively. We select the tuning parameters via a grid search using the GIC and validation set methods through the tuning data set. In GIC, ϕ_ns are $5p_n$, $5p_n$ and p_n for the quantile sparse group lasso, lasso and group lasso methods respectively, while $p_n = \log(\log(n))\log(\log(p))/(10n)$. The validation set method is used to select the gold standard (GS) tuning parameters that minimize the prediction error of tuning data sets (Li et al., 2007; Zou and Yuan, 2008; Wu and Liu, 2009).

In our simulations, we choose $n = 200, 400$, set $\tau = 0.5$, and use 100 Monte Carlo repetitions. We use the following five criteria of the performance, namely, the group accuracy (GA), variable accuracy (VA), mean absolute prediction error (MAPE), mean integrated square errors (MISE) and individual integrated square errors (ISE). The group accuracy (GA) is the

Table 1
Simulation summary of SNR = 5. The first column n is the size of training data. The second column is the type of noise. The third column is the method we used, qSGL for the quantile sparse group lasso, qL for the quantile Lasso, and qGL for the quantile group lasso. GS means λ was selected by the validation method (gold standard). GIC means λ selected via the GIC criterion. MISE stands for mean integrated errors. MAPE, GA and VA indicate mean absolute prediction error, group accuracy and variable accuracy, respectively.

			GS				GIC			
n	Noise	Method	MISE	GA	VA	MAPE	MISE	GA	VA	MAPE
200	1	qSGL	0.118	0.983	0.975	0.009	0.140	0.967	0.935	0.010
		qL	0.232	0.975	0.983	0.011	0.259	0.942	0.966	0.013
		qGL	0.143	0.917	0.589	0.011	0.144	0.917	0.554	0.011
	2	qSGL	0.122	0.983	0.972	0.009	0.139	0.967	0.934	0.011
		qL	0.237	0.975	0.983	0.011	0.271	0.942	0.965	0.013
		qGL	0.144	0.917	0.586	0.011	0.144	0.917	0.554	0.011
	3	qSGL	0.114	1.000	0.966	0.008	0.138	0.950	0.933	0.010
		qL	0.215	0.983	0.984	0.009	0.264	0.958	0.965	0.012
		qGL	0.139	0.917	0.599	0.010	0.144	0.917	0.549	0.011
	4	qSGL	0.096	1.000	0.966	0.005	0.142	0.950	0.933	0.010
		qL	0.149	1.000	0.982	0.007	0.289	0.958	0.965	0.013
		qGL	0.119	0.942	0.661	0.008	0.144	0.917	0.550	0.011
400	1	qSGL	0.110	1.000	0.976	0.006	0.135	0.933	0.896	0.008
		qL	0.188	1.000	0.986	0.007	0.200	0.942	0.934	0.010
		qGL	0.124	0.925	0.581	0.008	0.133	0.917	0.520	0.008
	2	qSGL	0.109	1.000	0.974	0.006	0.133	0.925	0.897	0.009
		qL	0.191	1.000	0.985	0.008	0.208	0.942	0.933	0.010
		qGL	0.125	0.917	0.577	0.008	0.133	0.917	0.521	0.008
	3	qSGL	0.099	1.000	0.969	0.005	0.127	0.942	0.893	0.008
		qL	0.165	1.000	0.984	0.007	0.201	0.950	0.930	0.010
		qGL	0.116	0.933	0.592	0.007	0.132	0.925	0.524	0.008
	4	qSGL	0.046	1.000	0.935	0.003	0.117	0.958	0.882	0.008
		qL	0.051	1.000	0.962	0.003	0.175	0.950	0.926	0.009
		qGL	0.092	0.992	0.623	0.005	0.128	0.950	0.526	0.008

proportion of correctly picked up and dropped off functional components, that is GA $= E\left(\left(|\widehat{M} \cap M_0| + |\widehat{M}^c \cap M_0^c|\right)/12\right)$ with $M_0 = \{l : \beta_l(t) \neq 0\}$ and $\widehat{M} = \{l : \hat{\beta}_l(t) \neq 0\}$. The variable accuracy (VA) is defined similarly as GA by simply replacing the M_0 and $\widehat{M}$ as the true and estimated index sets of non-zero wavelet coefficients. The mean absolute prediction error (MAPE) is MAPE$= E(|\hat{y} - y|)$. The mean integrated square errors (MISE) of the 12 estimated functional coefficients:

$$\text{MISE} = \frac{1}{12}\sum_{l=1}^{12}\int_0^1 (\hat{\beta}_l(t) - \beta_l(t))^2 dt,$$

as well as the individual integrated square error (ISE):

$$\text{ISE}_l = \int_0^1 (\hat{\beta}_l(t) - \beta_l(t))^2 dt,$$

is used to measure the estimation accuracy of functional coefficients.

Due to space limit, we only discuss the results of SNR $= 5$. The results for the other two SNRs are both in favor of our method and deferred to the Appendix. As shown in Tables 1 and 3, in general, the performance of qSGL method is better than the other methods in terms of mean integrated square errors (MISEs) and mean absolute prediction errors (MAPEs). For different error types, our proposed GIC approach is only slightly outperformed by the gold standards. As the sample size increases, the MISEs and MAPEs decrease, which is consistent with our theoretical results. For group accuracy (GA), both qSGL and qL performs slightly better than the other methods in most cases, while qL performs quite well in terms of variable accuracy (VA). However, in the case of GIC, the sparse group lasso method outperforms the two competitors regarding both GA and VA, especially for larger sample sizes. In Table 2, it shows that the ISEs of sparse group lasso are mostly smaller than the other two methods. It also shows that in the GS, the ISE of $\hat{\beta}_1(t)$ and $\hat{\beta}_3(t)$ are always less than the other slope functions in most cases regardless the methods used. It might be due to the fact that $\beta_1(t)$ is smoother than the other slopes; see Fig. 1. In Table 3, the qSGL method selected by the proposed GIC out-performs both the EN and sPLS selected by cross validations. It might be partially due to the fact that our proposed method is resistant against error distribution and can capture both individual and shared information among the slopes.

5.2. Real data

The real data we use is a subset of the ADHD-200 Sample Initiative Project (Mennes et al., 2013), which studies attention deficit hyperactivity disorder (ADHD), the most commonly diagnosed mental disorder of childhood which may persist into

Table 2
Individual functional L_2 error of SNR = 5. The first column n is the size of training data. The second column is the noise type. The third column is the method we used. ISE1: $||\hat{\beta}_1 - \beta_1||_2^2$; ISE2: $||\hat{\beta}_2 - \beta_2||_2^2$; ISE3: $||\hat{\beta}_3 - \beta_3||_2^2$; ISE4: $||\hat{\beta}_4 - \beta_4||_2^2$.

			GS				GIC			
n	Noise	Method	ISE1	ISE2	ISE3	ISE4	ISE1	ISE2	ISE3	ISE4
200	1	qSGL	0.278	0.805	0.362	0.391	0.782	0.863	0.349	0.457
		qL	0.620	1.079	3.081	1.576	1.284	1.101	3.225	1.563
		qGL	0.913	0.807	0.330	0.419	0.948	0.807	0.342	0.420
	2	qSGL	0.346	0.811	0.370	0.391	0.774	0.848	0.361	0.455
		qL	0.815	1.051	3.335	1.598	1.319	1.159	2.958	2.042
		qGL	0.954	0.814	0.349	0.410	0.989	0.813	0.356	0.415
	3	qSGL	0.187	0.789	0.327	0.393	0.758	0.814	0.348	0.563
		qL	0.431	1.037	2.520	1.396	1.159	1.158	3.206	2.014
		qGL	0.912	0.800	0.335	0.397	1.084	0.816	0.364	0.413
	4	qSGL	0.069	0.873	0.097	0.325	1.174	0.842	0.358	0.452
		qL	0.098	0.988	0.713	0.685	1.977	1.055	2.826	1.717
		qGL	0.460	0.732	0.169	0.319	1.110	0.794	0.311	0.404
400	1	qSGL	0.189	0.803	0.239	0.411	1.471	0.841	0.310	0.483
		qL	0.281	1.021	2.485	1.287	1.879	0.992	1.938	1.231
		qGL	0.946	0.770	0.262	0.406	1.413	0.778	0.285	0.475
	2	qSGL	0.208	0.818	0.224	0.414	1.405	0.850	0.320	0.487
		qL	0.279	1.023	2.178	1.293	1.838	0.994	2.108	1.357
		qGL	0.928	0.776	0.245	0.405	1.437	0.781	0.270	0.464
	3	qSGL	0.143	0.779	0.248	0.389	1.198	0.800	0.334	0.518
		qL	0.173	0.965	1.676	1.024	1.484	0.968	2.226	1.573
		qGL	0.762	0.757	0.262	0.392	1.313	0.780	0.328	0.464
	4	qSGL	0.020	0.703	0.052	0.168	1.039	0.843	0.227	0.453
		qL	0.024	0.904	0.206	0.299	1.510	0.986	1.422	1.419
		qGL	0.257	0.661	0.112	0.261	1.158	0.759	0.264	0.420

Table 3
Simulation summary of SNR=5. The first column *n* is the size of training data. The second column is the type of noise. The third column is the method we used, sPLS for the sparse partial least squares, EN for elastic net. sPLS and EN are selected based on cross validation while sSGL is based on GIC. MISE stands for mean integrated errors. MAPE, GA and VA indicate mean absolute prediction error, group accuracy and variable accuracy, respectively. ISE1: $||\hat{\beta}_1 - \beta_1||_2^2$; ISE2: $||\hat{\beta}_2 - \beta_2||_2^2$; ISE3: $||\hat{\beta}_3 - \beta_3||_2^2$; ISE4: $||\hat{\beta}_4 - \beta_4||_2^2$.

n	Noise	Method	MISE	GA	VA	MAPE	ISE1	ISE2	ISE3	ISE4
200	1	qSGL	0.140	0.967	0.935	0.010	0.143	0.882	0.136	0.328
		sPLS	0.447	0.917	0.990	0.426	1.567	1.124	1.757	0.878
		EN	4.488	0.833	0.508	0.166	2.614	11.252	16.593	5.889
	2	qSGL	0.139	0.967	0.934	0.011	0.150	0.869	0.132	0.318
		sPLS	0.672	0.917	0.990	0.430	1.707	1.181	4.142	0.984
		EN	4.514	0.833	0.508	0.166	2.626	11.314	16.713	5.922
	3	qSGL	0.138	0.950	0.933	0.010	0.139	0.853	0.114	0.370
		sPLS	0.561	0.917	0.990	0.405	1.687	1.217	2.880	0.909
		EN	4.498	0.833	0.510	0.166	2.618	11.265	16.644	5.898
	4	qSGL	0.133	0.925	0.897	0.009	0.119	0.814	0.143	0.225
		sPLS	0.708	0.917	0.989	0.364	1.888	1.309	4.253	0.987
		EN	4.471	0.833	0.508	0.166	2.608	11.206	16.554	5.865
400	1	qSGL	0.135	0.933	0.896	0.008	0.116	0.825	0.140	0.237
		sPLS	0.394	0.917	0.992	0.450	1.635	0.904	1.519	0.648
		EN	4.275	0.833	0.504	0.113	2.370	10.557	16.107	5.882
	2	qSGL	0.133	0.925	0.897	0.009	0.119	0.814	0.143	0.225
		sPLS	0.394	0.917	0.992	0.450	1.635	0.904	1.519	0.648
		EN	4.285	0.833	0.504	0.113	2.374	10.583	16.146	5.896
	3	qSGL	0.127	0.942	0.893	0.008	0.115	0.785	0.120	0.220
		sPLS	0.394	0.917	0.992	0.450	1.635	0.904	1.519	0.648
		EN	4.264	0.833	0.504	0.113	2.366	10.530	16.065	5.867
	4	qSGL	0.117	0.958	0.882	0.008	0.098	0.688	0.093	0.197
		sPLS	0.393	0.917	0.992	0.450	1.641	0.900	1.517	0.644
		EN	4.277	0.833	0.504	0.113	2.371	10.563	16.115	5.884

adulthood. ADHD is characterized by problems related to paying attention, hyperactivity, or impulsive behavior. The data set is a filtered preprocessed resting state fMRI data from New York University Child Study Centre using the Anatomical

Table 4
Selected ROIs for the ADHD-200 fMRI Data set..

Method	Significant ROIs
qSGL	"Temporal R" "Cerebellum R" "Frontal R" "Occipital R" "Olfactory R" "SupraMarginal R" "Caudate R" "Vermis" "Cuneus L" "Parietal R" "Frontal L" "Precuneus R" "Temporal L" "Cerebellum L" "Precentral R"
qL	"Frontal R" "Caudate R" "Temporal R" "Cuneus L" "SupraMarginal R" "Parietal R" "Lingual L" "Frontal L" "Precuneus R" "Vermis" "Fusiform R" "Pallidum L" "Olfactory R" "Precentral R" "Cingulum L" "Cuneus R" "Parietal L" "Temporal L" "Angular L" "Cerebellum R"
qGL	"Caudate R" "Frontal R" "Cerebellum R" "Vermis" "Olfactory R" "Temporal R" "Precentral R" "SupraMarginal R" "Frontal L"

Table 5
Selected ROIs for the suggested 7 regions, 'R' and 'L' indicate the region is selected from the right brain and left brain, respectively. Blank means the brain region is not chosen.

Significant regions	qSGL	qL	qGL
Cerebellum	R L	R	R
Temporal	R L	R L	R
Vermis	R L	R L	R L
Parietal	R	R L	
Occipital	R		
Cingulum		L	
Frontal	R L	L	R L

Automatic Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002). In the data set, there are 172 equally spaced time courses in the filtering and AAL contains 116 Regions of Interests (ROIs) fractionated into functional space using nearest-neighbor interpolation. Each of 172 time courses is then smoothed to 64 equally to apply DWT. After cleaning the raw data that fails in quality control or has missing data, we have 120 individuals in final analysis. Grouping ROIs in terms of their anatomical functions and averaging within each group the corresponding time courses, we have 59 averaged time courses of grouped ROIs serving as functional predictors, each of which has 64 equally spaced time points. In addition, 8 scalar covariates are considered, including gender, age, handedness, diagnosis status, medication status, Verbal IQ, Performance IQ and Full4 IQ. The response of interest is the ADHD index, a measurement of severity of mental disorder.

We apply partial functional linear quantile regression model (4) with 59 functional covariates and 8 scalar covariates. In order to select the significant functional covariates from 59 ROIs, we use the procedure proposed by Meinshausen and Bühlmann (2010) to obtain stable selections from 100 bootstrap samples. The tuning parameters are chosen by GIC. The boxplots of L_2 norms of the estimated slope functions from bootstrap samples are shown in Figs. 2–4 in the Appendix. The selection criterion is that the median of corresponding L_2 norm should be greater than 10^{-5}.

In neurological science literature on ADHD, it has been shown that the 7 regions of cerebellum, temporal, vermis, parietal, occipital, cingulum and frontal are commonly discovered to be significantly related to ADHD symptoms from various studies (Max et al., 2005; Konrad and Eickhoff, 2010; Tomasi and Volkow, 2012). We first evaluate the performances of qSGL, qL and qGL methods in terms of the selection of these 7 regions, which are essentially 14 ROIs including the left and right parts. In Tables 4 and 5, we list the selected ROIs from three different methods. In particular, qSGL, qL and qGL select 15, 20 and 9 ROIs respectively. In terms of those 7/14 commonly discovered regions/ROIs, both our proposed qSGL and qGL methods have lower false discovery rates (33%) than the qL method (55%), while our method is superior to the qGL as it identifies more true positives (10 vs 6). Moreover, "Occipital R", the right occipital region, can only be identified by our method. While both Tables 4 and 5 confirm that most of the selected ROIs are coming from the 7/14 mostly discovered regions/ROIs, the three methods also suggest three other common ROIs: "Olfactory R", "Supramarginal R", and "Caudate R", namely right olfactory, right supramarginal, and right caudate regions respectively, which have been evidently important as suggested by some ADHD studies. For instance, Schrimsher et al. (2002) revealed a relationship between caudate asymmetry and some symptoms related to ADHD. The findings of Sidlauskaite et al. (2015) imply the supramarginal gyrus is associated with the ADHD symptom scores.

6. Discussion

This article studies quantile regression in partial functional linear model where response is scalar and predictors include both scalars and multiple functions. We adopt wavelet basis to well approximate functional slopes while effectively detect local features. A sparse group lasso method is proposed to select important functional predictors while capture shared information among them. We reformulate the proposed problem into a standard second-order cone program and then solve it by an interior point method. A novel and efficient algorithm by using alternating direction method of multipliers

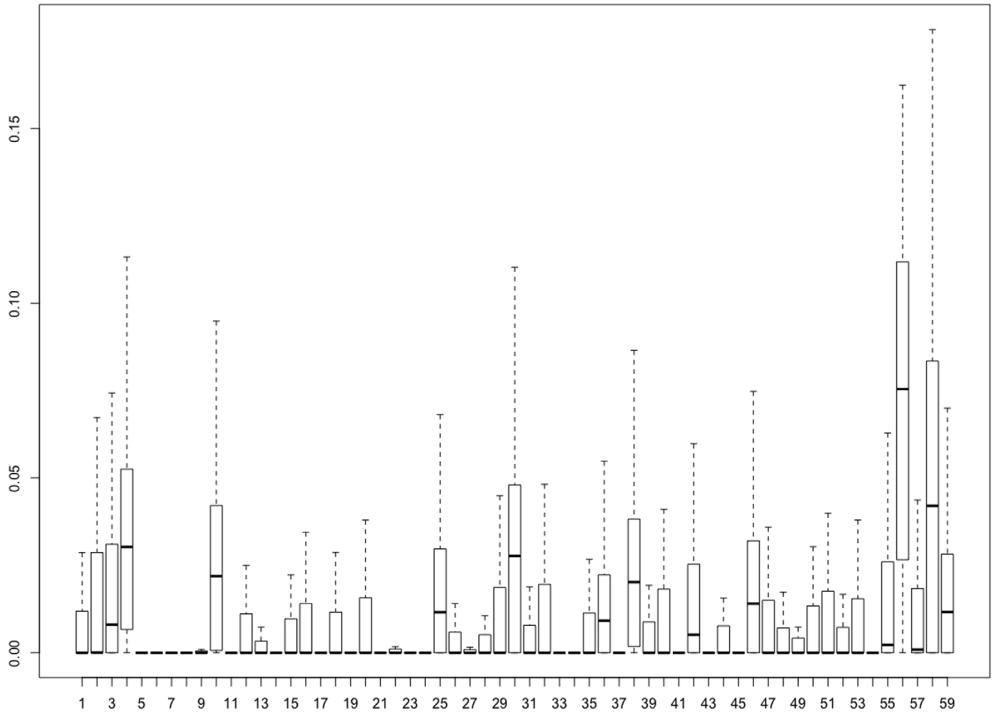

Fig. 2. Boxplot of L_2 norm for each slope function, by using the quantile spare group lasso method.

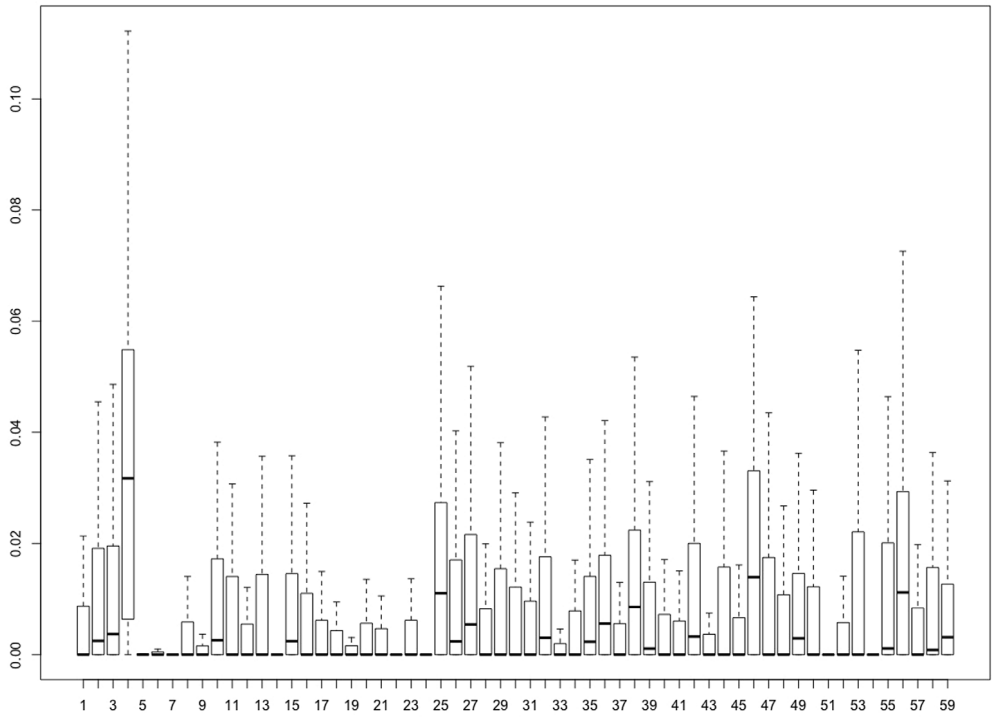

Fig. 3. Boxplot of L_2 norm for each slope function, by using the quantile lasso method.

(ADMM) is utilized to solve the optimization problem. In addition, we successfully derive the asymptotic properties including the convergence rate and prediction error bound which guarantee a good theoretical performance of the proposed

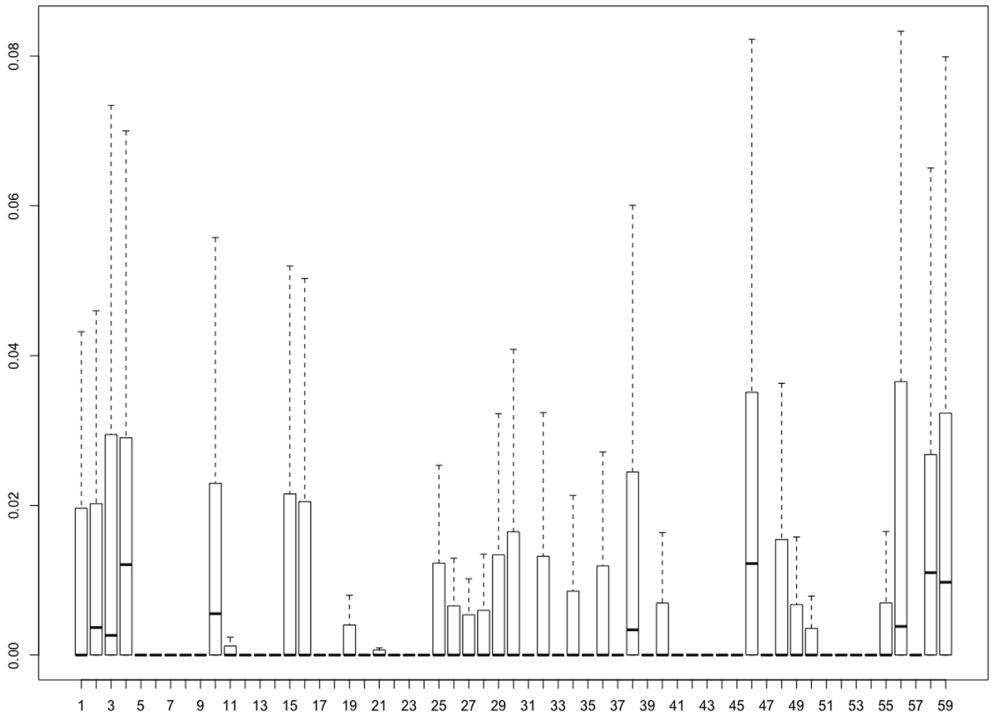

Fig. 4. Boxplot of L_2 norm for each slope function, by using the quantile group lasso method.

method. Simulation studies demonstrate that our proposed method is more effective in estimating coefficients and making predictions while capable of identifying non-zero functional components and wavelet coefficients. We analyze a real data from ADHD-200 fMRI data set and show the superiority of our method. Moreover, our analysis makes some new discovery about other brain regions that are evidently important in making diagnosis.

There are several topics that merit further research. Other asymptotic properties, such as the model selection consistency and asymptotic normality, of our proposed method could be developed. The technique proposed to reformulate our problem into a second order cone program (SOCP) could be further adapted to other penalized quantile regression problems; for example, quantile ridge regression (Wu and Liu, 2009). Moreover, to estimate the functional slopes, the wavelet-based technique can also be used together with principal component analysis or partial least squares methods (Reiss et al., 2015).

Acknowledgments

We wish to thank the Editor, Associate Editor and two referees for their valuable comments that have helped greatly improve the quality of this work. This research was partially supported by fundings from the Natural Sciences and Engineering Research Council of Canada (NSERC) to Dr. Ivan Mizera, from the University of Alberta and the Natural Sciences and Engineering Research Council of Canada (NSERC) to Dr. Bei Jiang, from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Statistical Sciences Institute (CANSSI) to Dr. Linglong Kong.

Appendix

See Tables 6–11.

A.1. Proof of Theorem 3.1

Proof. First, we introduce some notation. The orthonormal wavelet basis set of $L^2[0,1]$ is defined as $\{\varphi_{j_0k}, k = 1, \ldots, 2^{j_0}\} \cup \{\psi_{jk},\ j \geq j_0,\ k = 1, \ldots, 2^j\}$. Without loss of generality, the wavelet basis is ordered according to the scales from the coarsest level J_0 to the finest one. Let $\mathbb{V}_{N_n} := \text{Span}\{\boldsymbol{\varphi}_1, \ \ldots, \ \boldsymbol{\varphi}_{N_n}\}$ be the space spanned by the first N_n basis function, for example, if $N_n = 2^{j_0+t}$, then the collection of $\{\varphi_{j_0k}, k = 1 \ldots, 2^{j_0}\} \cup \{\psi_{jk}, j_0 \leq j \leq j_0 + t - 1,\ k = 1, \ldots,\ 2^j\}$ is the basis of $\mathbb{V}_{N_n}$. Let $\boldsymbol{b}^j_{N_n}$ be an $N_n \times 1$ parameter vector with elements $b^j_k = \langle \beta_j(t), \varphi_k \rangle$. In addition, let $\beta^j_{N_n}$ be the functions reconstructed from the vector $\boldsymbol{b}^j_{N_n}$. Here $\beta^j_{N_n}$ is a linear approximation to β_j by the first N_n wavelet coefficients, while $\hat{\beta}_j$ denotes the function reconstructed from the wavelet coefficients $\hat{\boldsymbol{b}}_j$ from (10).

Table 6
Simulation summary of SNR = 1, as for Table 1.

				GS				GIC		
n	Noise	Method	MISE	GA	VA	MAPE	MISE	GA	VA	MAPE
200	1	qSGL	0.160	0.967	0.981	0.028	0.622	0.633	0.932	0.043
		qL	0.544	0.933	0.984	0.033	1.375	0.633	0.961	0.048
		qGL	0.339	0.825	0.584	0.036	0.368	0.783	0.566	0.036
	2	qSGL	0.174	0.942	0.981	0.030	0.633	0.617	0.932	0.043
		qL	0.598	0.917	0.983	0.035	1.475	0.633	0.962	0.048
		qGL	0.350	0.800	0.574	0.036	0.378	0.783	0.551	0.036
	3	qSGL	0.146	0.967	0.982	0.021	0.589	0.617	0.926	0.042
		qL	0.455	0.925	0.985	0.024	1.383	0.700	0.959	0.050
		qGL	0.276	0.892	0.574	0.028	0.371	0.825	0.542	0.034
	4	qSGL	0.114	1.000	0.975	0.007	0.660	0.658	0.924	0.043
		qL	0.208	0.992	0.986	0.009	1.473	0.700	0.958	0.050
		qGL	0.143	0.925	0.602	0.011	0.399	0.783	0.539	0.035
400	1	qSGL	0.138	0.975	0.986	0.016	0.867	0.492	0.882	0.035
		qL	0.433	0.933	0.985	0.020	1.436	0.600	0.927	0.040
		qGL	0.278	0.900	0.564	0.023	0.539	0.700	0.520	0.029
	2	qSGL	0.139	1.000	0.987	0.016	0.871	0.450	0.884	0.035
		qL	0.410	0.925	0.985	0.019	1.476	0.592	0.929	0.041
		qGL	0.270	0.917	0.562	0.023	0.544	0.667	0.522	0.029
	3	qSGL	0.130	0.992	0.986	0.012	0.837	0.458	0.871	0.035
		qL	0.324	0.933	0.986	0.015	1.434	0.583	0.923	0.039
		qGL	0.215	0.917	0.566	0.018	0.523	0.725	0.542	0.030
	4	qSGL	0.079	1.000	0.958	0.004	0.885	0.575	0.778	0.032
		qL	0.121	0.992	0.978	0.005	1.464	0.642	0.916	0.035
		qGL	0.108	0.983	0.613	0.006	0.562	0.683	0.533	0.027

Table 7
Individual functional L_2 error when SNR = 1, as for Table 2.

				GS				GIC		
n	Noise	Method	ISE1	ISE2	ISE3	ISE4	ISE1	ISE2	ISE3	ISE4
200	1	qSGL	0.278	0.805	0.362	0.391	0.782	0.863	0.349	0.457
		qL	0.620	1.079	3.081	1.576	1.284	1.101	3.225	1.563
		qGL	0.913	0.807	0.330	0.419	0.948	0.807	0.342	0.420
	2	qSGL	0.346	0.811	0.370	0.391	0.774	0.848	0.361	0.455
		qL	0.815	1.051	3.335	1.598	1.319	1.159	2.958	2.042
		qGL	0.954	0.814	0.349	0.410	0.989	0.813	0.356	0.415
	3	qSGL	0.187	0.789	0.327	0.393	0.758	0.814	0.348	0.563
		qL	0.431	1.037	2.520	1.396	1.159	1.158	3.206	2.014
		qGL	0.912	0.800	0.335	0.397	1.084	0.816	0.364	0.413
	4	qSGL	0.069	0.873	0.097	0.325	1.174	0.842	0.358	0.452
		qL	0.098	0.988	0.713	0.685	1.977	1.055	2.826	1.717
		qGL	0.460	0.732	0.169	0.319	1.110	0.794	0.311	0.404
400	1	qSGL	0.189	0.803	0.239	0.411	1.471	0.841	0.310	0.483
		qL	0.281	1.021	2.485	1.287	1.879	0.992	1.938	1.231
		qGL	0.946	0.770	0.262	0.406	1.413	0.778	0.285	0.475
	2	qSGL	0.208	0.818	0.224	0.414	1.405	0.850	0.320	0.487
		qL	0.279	1.023	2.178	1.293	1.838	0.994	2.108	1.357
		qGL	0.928	0.776	0.245	0.405	1.437	0.781	0.270	0.464
	3	qSGL	0.143	0.779	0.248	0.389	1.198	0.800	0.334	0.518
		qL	0.173	0.965	1.676	1.024	1.484	0.968	2.226	1.573
		qGL	0.762	0.757	0.262	0.392	1.313	0.780	0.328	0.464
	4	qSGL	0.020	0.703	0.052	0.168	1.039	0.843	0.227	0.453
		qL	0.024	0.904	0.206	0.299	1.510	0.986	1.422	1.419
		qGL	0.257	0.661	0.112	0.261	1.158	0.759	0.264	0.420

By the Parseval theorem, we have $\|\hat{\beta}_j - \beta_j\|_{L_2}^2 = \|\hat{b}^j_{N_n} - b^j_{N_n}\|_2^2 + \sum_{k=N_n+1}^{\infty} {\theta_k^j}^2$. To derive the convergence rate of $\hat{\beta}_j$ to β_j, we bound the error in estimating $\beta^j_{N_n}$ by $\hat{\beta}_j$ and the error in approximating β_j by β_{N_n}. By the Theorem 9.5 of Mallat (2008),

Table 8
Simulation summary of SNR = 1. The first column n is the size of training data. The second column is the type of noise. The third column is the method we used, sPLS for the sparse partial least squares, EN for elastic net. sPLS and EN are selected based on cross validation while sSGL is based on GIC. MISE stands for mean integrated errors. MAPE, GA and VA indicate mean absolute prediction error, group accuracy and variable accuracy, respectively. ISE1: $||\hat{\beta}_1 - \beta_1||_2^2$; ISE2: $||\hat{\beta}_2 - \beta_2||_2^2$; ISE3: $||\hat{\beta}_3 - \beta_3||_2^2$; ISE4: $||\hat{\beta}_4 - \beta_4||_2^2$.

n	Noise	Method	MISE	GA	VA	MAPE	ISE1	ISE2	ISE3	ISE4
200	1	qSGL	0.622	0.633	0.932	0.043	0.782	0.863	0.349	0.457
		sPLS	0.586	0.917	0.991	0.378	1.734	1.287	3.065	0.932
		EN	4.529	0.833	0.509	0.165	2.632	11.384	16.710	5.947
	2	qSGL	0.633	0.617	0.932	0.043	0.774	0.848	0.361	0.455
		sPLS	0.843	0.900	0.991	0.384	2.732	1.357	5.017	0.968
		EN	4.501	0.833	0.508	0.166	2.619	11.345	16.579	5.913
	3	qSGL	0.589	0.617	0.926	0.042	0.758	0.814	0.348	0.563
		sPLS	0.529	0.917	0.991	0.394	1.620	1.237	2.522	0.928
		EN	4.504	0.833	0.508	0.166	2.621	11.272	16.646	5.894
	4	qSGL	0.871	0.450	0.884	0.035	1.405	0.850	0.320	0.487
		sPLS	0.648	0.917	0.991	0.401	1.860	1.292	3.649	0.913
		EN	4.529	0.833	0.510	0.166	2.630	11.343	16.799	5.925
400	1	qSGL	0.867	0.492	0.882	0.035	1.471	0.841	0.310	0.483
		sPLS	0.393	0.917	0.992	0.451	1.641	0.900	1.516	0.644
		EN	4.278	0.833	0.504	0.113	2.372	10.565	16.118	5.886
	2	qSGL	0.871	0.450	0.884	0.035	1.405	0.850	0.320	0.487
		sPLS	0.393	0.917	0.992	0.450	1.634	0.904	1.519	0.648
		EN	4.286	0.833	0.504	0.113	2.374	10.585	16.151	5.898
	3	qSGL	0.837	0.458	0.871	0.035	1.198	0.800	0.334	0.518
		sPLS	0.394	0.917	0.992	0.450	1.635	0.904	1.520	0.648
		EN	4.259	0.833	0.504	0.113	2.364	10.515	16.044	5.862
	4	qSGL	0.885	0.575	0.778	0.032	1.039	0.843	0.227	0.453
		sPLS	0.393	0.917	0.992	0.451	1.642	0.901	1.517	0.644
		EN	4.280	0.833	0.504	0.113	2.373	10.571	16.121	5.887

Table 9
Simulation summary of SNR=10, as for Table 1.

			GS				GIC			
n	Noise	Method	MISE	GA	VA	MAPE	MISE	GA	VA	MAPE
200	1	qSGL	0.106	0.983	0.964	0.006	0.107	0.983	0.936	0.007
		qL	0.176	0.992	0.981	0.008	0.175	0.983	0.967	0.009
		qGL	0.124	0.933	0.601	0.008	0.124	0.933	0.551	0.008
	2	qSGL	0.108	0.992	0.971	0.007	0.107	0.992	0.935	0.007
		qL	0.183	0.983	0.981	0.008	0.180	0.992	0.967	0.009
		qGL	0.124	0.925	0.597	0.009	0.124	0.925	0.555	0.009
	3	qSGL	0.103	1.000	0.965	0.006	0.108	0.983	0.934	0.007
		qL	0.172	1.000	0.982	0.008	0.183	0.983	0.965	0.009
		qGL	0.122	0.917	0.634	0.008	0.124	0.917	0.554	0.008
	4	qSGL	0.088	1.000	0.959	0.005	0.107	0.975	0.933	0.007
		qL	0.146	1.000	0.981	0.006	0.198	0.975	0.965	0.009
		qGL	0.116	0.950	0.679	0.008	0.124	0.925	0.552	0.008
400	1	qSGL	0.095	1.000	0.969	0.005	0.082	1.000	0.898	0.005
		qL	0.137	1.000	0.982	0.006	0.116	1.000	0.935	0.007
		qGL	0.104	0.975	0.581	0.006	0.105	0.967	0.520	0.006
	2	qSGL	0.095	1.000	0.970	0.005	0.082	0.992	0.902	0.005
		qL	0.142	1.000	0.982	0.006	0.117	1.000	0.936	0.007
		qGL	0.104	0.967	0.584	0.006	0.106	0.967	0.514	0.006
	3	qSGL	0.079	1.000	0.955	0.004	0.078	1.000	0.900	0.005
		qL	0.120	1.000	0.979	0.005	0.114	0.992	0.935	0.007
		qGL	0.101	0.992	0.590	0.006	0.105	0.983	0.515	0.006
	4	qSGL	0.034	1.000	0.927	0.003	0.071	1.000	0.887	0.005
		qL	0.036	1.000	0.957	0.003	0.097	1.000	0.932	0.006
		qGL	0.090	1.000	0.618	0.005	0.104	0.975	0.513	0.006

Table 10
Individual functional L_2 error when SNR=10, as for Table 2..

n	Noise	Method	GS ISE1	GS ISE2	GS ISE3	GS ISE4	GIC ISE1	GIC ISE2	GIC ISE3	GIC ISE4
200	1	qSGL	0.052	0.862	0.077	0.269	0.065	0.825	0.089	0.245
		qL	0.066	0.980	0.460	0.568	0.060	0.976	0.444	0.477
		qGL	0.346	0.709	0.141	0.284	0.346	0.709	0.141	0.285
	2	qSGL	0.059	0.858	0.083	0.286	0.067	0.819	0.088	0.246
		qL	0.065	0.974	0.510	0.603	0.061	0.992	0.430	0.512
		qGL	0.347	0.710	0.142	0.283	0.347	0.710	0.142	0.283
	3	qSGL	0.039	0.841	0.082	0.270	0.063	0.819	0.080	0.277
		qL	0.058	0.990	0.435	0.553	0.069	0.984	0.452	0.546
		qGL	0.338	0.706	0.138	0.283	0.347	0.708	0.141	0.285
	4	qSGL	0.024	0.766	0.055	0.206	0.072	0.788	0.106	0.255
		qL	0.031	0.971	0.303	0.428	0.072	0.973	0.657	0.517
		qGL	0.298	0.698	0.122	0.267	0.352	0.708	0.136	0.284
400	1	qSGL	0.027	0.831	0.065	0.208	0.044	0.627	0.084	0.142
		qL	0.024	0.940	0.288	0.374	0.033	0.749	0.244	0.210
		qGL	0.233	0.650	0.112	0.246	0.234	0.645	0.120	0.250
	2	qSGL	0.030	0.838	0.062	0.204	0.043	0.625	0.084	0.139
		qL	0.024	0.962	0.304	0.395	0.033	0.732	0.255	0.220
		qGL	0.237	0.651	0.112	0.244	0.238	0.647	0.120	0.248
	3	qSGL	0.020	0.702	0.058	0.157	0.045	0.600	0.072	0.130
		qL	0.021	0.888	0.224	0.285	0.034	0.736	0.242	0.201
		qGL	0.213	0.641	0.111	0.239	0.229	0.642	0.124	0.250
	4	qSGL	0.008	0.320	0.024	0.055	0.039	0.546	0.058	0.117
		qL	0.007	0.299	0.050	0.066	0.034	0.586	0.184	0.194
		qGL	0.156	0.617	0.091	0.210	0.225	0.640	0.117	0.246

Table 11
Simulation summary of SNR = 10. The first column n is the size of training data. The second column is the type of noise. The third column is the method we used, sPLS for the sparse partial least squares, EN for elastic net. sPLS and EN are selected based on cross validation while sSGL is based on GIC. MISE stands for mean integrated errors. MAPE, GA and VA indicate mean absolute prediction error, group accuracy and variable accuracy, respectively. ISE1: $||\hat{\beta}_1 - \beta_1||_2^2$; ISE2: $||\hat{\beta}_2 - \beta_2||_2^2$; ISE3: $||\hat{\beta}_3 - \beta_3||_2^2$; ISE4: $||\hat{\beta}_4 - \beta_4||_2^2$.

n	Noise	Method	MISE	GA	VA	MAPE	ISE1	ISE2	ISE3	ISE4
200	1	qSGL	0.107	0.983	0.936	0.007	0.065	0.825	0.089	0.245
		sPLS	0.448	0.917	0.991	0.434	1.519	1.165	1.767	0.882
		EN	4.491	0.833	0.508	0.166	2.615	11.256	16.611	5.892
	2	qSGL	0.107	0.992	0.935	0.007	0.067	0.819	0.088	0.246
		sPLS	0.585	0.917	0.989	0.405	1.702	1.192	3.139	0.934
		EN	4.517	0.833	0.508	0.166	2.627	11.318	16.725	5.925
	3	qSGL	0.108	0.983	0.934	0.007	0.063	0.819	0.080	0.277
		sPLS	0.541	0.917	0.991	0.395	1.647	1.223	2.692	0.895
		EN	4.514	0.833	0.509	0.166	2.625	11.313	16.701	5.920
	4	qSGL	0.082	0.992	0.902	0.005	0.043	0.625	0.084	0.139
		sPLS	0.672	0.917	0.989	0.398	1.850	1.236	3.949	0.974
		EN	4.486	0.833	0.508	0.166	2.613	11.241	16.610	5.884
400	1	qSGL	0.082	1.000	0.898	0.005	0.044	0.627	0.084	0.142
		sPLS	0.394	0.917	0.992	0.450	1.635	0.904	1.519	0.648
		EN	4.275	0.833	0.504	0.113	2.370	10.557	16.107	5.882
	2	qSGL	0.082	0.992	0.902	0.005	0.043	0.625	0.084	0.139
		sPLS	0.394	0.917	0.992	0.450	1.635	0.904	1.519	0.648
		EN	4.277	0.833	0.504	0.113	2.371	10.564	16.117	5.885
	3	qSGL	0.078	1.000	0.900	0.005	0.045	0.600	0.072	0.130
		sPLS	0.394	0.917	0.992	0.450	1.635	0.904	1.519	0.648
		EN	4.264	0.833	0.504	0.113	2.366	10.530	16.065	5.867
	4	qSGL	0.071	1.000	0.887	0.005	0.039	0.546	0.058	0.117
		sPLS	0.393	0.917	0.992	0.450	1.641	0.900	1.517	0.644
		EN	4.277	0.833	0.504	0.113	2.371	10.563	16.115	5.884

the linear approximation error goes to zero as

$$\sum_{k=N_n+1}^{\infty} {b_k^j}^2 = o(N_n^{-2d}). \tag{16}$$

Let $\Upsilon^0 = (\boldsymbol{\alpha}^0, \boldsymbol{\gamma}^0, \boldsymbol{\theta}^0)$ be the true coefficients with $\boldsymbol{\theta}^0 = \text{vec}^T\left(\boldsymbol{b}^1_{N_n}, \ldots, \boldsymbol{b}^m_{N_n}\right)$. To obtain the result, we show that for any given $\varepsilon > 0$, there exists a constant C such that

$$\Pr\left\{\inf_{\|\boldsymbol{z}\|=C} L_n\left(\Upsilon^0 + r_n\boldsymbol{z}\right) + P_{\lambda_1,\lambda_2}\left(\boldsymbol{\theta}^0 + r_n\boldsymbol{z}_\theta\right) > L_n\left(\Upsilon^0\right) + P_{\lambda_1,\lambda_2}\left(\boldsymbol{\theta}^0\right)\right\} \geq 1-\varepsilon, \tag{17}$$

where $r_n = \sqrt{N_n/n\rho_n}$ and $\boldsymbol{z} = (z_1, \ldots, z_k, \boldsymbol{z}_\gamma, \boldsymbol{z}_\theta)$ is a vector with the same length of vector Υ^0. This implies that there exists a local minimizer in the ball $\{\Upsilon^0 + r_n z : \|z\| \leq C\}$ with probability at least $1-\varepsilon$. Hence, there is a local minimizer $\widehat{\Upsilon}$ such that $\|\widehat{\Upsilon} - \Upsilon^0\| = O_p(r_n)$.

To show (17), we compare $L_n(\boldsymbol{\Upsilon}^0) + P_n(\boldsymbol{\theta}^0)$ with $L_n(\boldsymbol{\Upsilon}^0 + r_n\boldsymbol{z}) + P_n(\theta^0 + r_n\boldsymbol{z}_\theta)$. By using the Knight identity,

$$\rho_\tau(u-v) - \rho_\tau(u) = -v\varrho_\tau(u) + \int_0^v (I(u \leq t) - I(u \leq 0))dt,$$

where $\varrho_\tau(u) = \tau - I(u < 0)$, we have

$$\begin{aligned}
I &:= L_n(\boldsymbol{\Upsilon}^0 + r_n\boldsymbol{v}) - L_n(\boldsymbol{\Upsilon}^0) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n}[\rho_{\tau_k}(e_{ki} - d_{ki}) - \rho_{\tau_k}(e_{ki})] \\
&= -\sum_{k=1}^{K}\sum_{i=1}^{n}[-d_{ki}\varrho_{\tau_k}(e_{ki})] + \sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^{d_{ki}}(I(e_{ki} \leq t) - I(e_{ki} \leq 0))dt \\
&> \frac{\rho_n n r_n^2}{2}\|\boldsymbol{z}\|_2^2 \min_k\{f(F^{-1}(\tau_k) + o(N_n^{-2d})) + o_p(1)\} - K\sqrt{N_n}c_2\|\boldsymbol{z}\|_2 \\
&> \frac{N_n}{2}\|\boldsymbol{z}\|_2^2 \min_k\{f(F^{-1}(\tau_k) + o(N_n^{-2d})) + o_p(1)\} - K\sqrt{N_n}c_2\|\boldsymbol{z}\|_2,
\end{aligned}$$

where the last two steps were from the following lemma.

Finally, since $r_n \to 0$ and $\|\boldsymbol{z}\|_2 \leq C$, we have

$$\begin{aligned}
II := P_n(\theta^0 + r_n\boldsymbol{z}_\theta) - P_n(\theta^0) &\leq \lambda_1 r_n\|\boldsymbol{z}_\theta\|_1 + \lambda_2 r_n\sum_{j=1}^{m}\|\boldsymbol{z}_{\theta_j}\|_2 \\
&\leq \lambda_1 r_n\sqrt{mN_n}\|\boldsymbol{z}_\theta\|_2 + \lambda_2 r_n m\|\boldsymbol{z}_\theta\|_2 \\
&= O_p(N_n\|\boldsymbol{z}_\theta\|_2).
\end{aligned}$$

Since II is bounded by $N_n\|\boldsymbol{z}_\theta\|_2$,we can choose a C such that the II is dominated by the term I_2 on $\|u\| = C$ uniformly. So $Q_n(\Sigma^0 + r_n u) - Q_n(\Sigma^0) > 0$ holds uniformly on $\|u\| = C$.

To show the upper bound of I, we need to first prove a lemma that is similar the Lemma 3 in Wu and Liu (2009). Denote a linear approximation to $\rho_{\tau_k}(e_{k_i} - t)$ by $D_{k_i} = (1-\tau_k)\{e_{k_i} < 0\} - \tau_k\{e_{k_i} \geq 0\}$. Based on the (Wu and Liu, 2009), D_{k_i} can be thought of as the first derivative of$\rho_{\tau_k}(e_{k_i} - t)$ at $t = 0$ and $E(D_{k_i}) = o(N_n^{-2d})$. Define $R_{i,k}(\boldsymbol{v}) = \rho_{\tau_k}(e_{k_i} - b_{k_i}) - \rho_{\tau_k}(e_{k_i}) - D_{k_i}b_{k_i}$ and $W_k = \sum_{i=1}^{n} r_n D_{k_i}A_i$.

Lemma A.1. *Under condition (A1) and condition (A2), we have*

$$I > \frac{\rho_n n r_n^2}{2}\|\boldsymbol{z}\|_2^2 \min_k\{f(F^{-1}(\tau_k) + o(N_n^{-2d})) + o_p(1)\} - K\sqrt{N_n}c_2\|\boldsymbol{z}\|_2. \tag{18}$$

Proof. By using the Lemma 3 in Wu and Liu (2009), we had

$$I = \sum_{k=1}^{K}\{f(F^{-1}(\tau_k) + o(N_n^{-2d})) + o_p(1)\}\boldsymbol{z}^T\left(\frac{r_n^2}{2}\boldsymbol{A}_n^T\boldsymbol{A}_n\right)\boldsymbol{z} + \sum_{k=1}^{K}\boldsymbol{W}_k^T\boldsymbol{z} + o_p(1). \tag{19}$$

Clearly, the first term $\sum_{k=1}^{K}\{f(F^{-1}(\tau_k) + o(N_n^{-2d})) + o_p(1)\}\boldsymbol{z}^T\left(\frac{r_n^2}{2}\boldsymbol{A}_n^T\boldsymbol{A}_n\right)\boldsymbol{z} \geq \frac{\rho_n n r_n^2}{2}\|\boldsymbol{z}\|_2^2 \min_k\{f(F^{-1}(\tau_k) + o(N_n^{-2d}))$, from the condition 1 and condition 2. By the $|D_{k_i}| < 1$, we had $|W_k^T z|^2 < r_n^2 z(\boldsymbol{A}_n^T\boldsymbol{A})z < c_2^2 N_n\|z\|_2^2$. □

Based on above lemma, we can find that I is dominated by the quadratic term $\frac{\rho_n n r_n^2}{2}\|\boldsymbol{z}\|_2^2 \min_k\{f(F^{-1}(\tau_k) + o(N_n^{-2d})) + o_p(1)\}$. This completes the proof. □

References

Antoniadis, A., Bigot, J., Sapatinas, T., 2001. Wavelet estimators in nonparametric regression: a comparative simulation study. J. Stat. Softw. 6, 1–83.

Aps, M., 2015. Rmosek: The R to MOSEK optimization interface. URL http://rmosek.r-forge.r-project.org/, http://www.mosek.com/. R package version 7(2).

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. 3 (1), 1–122.

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

Cai, T.T., Hall, P., 2006. Prediction in functional linear regression. Ann. Statist. 34 (5), 2159–2179.

Cardot, H., Crambes, C., Sarda, P., 2005. Quantile regression when the covariates are functions. Nonparametr. Stat. 17 (7), 841–856.

Cardot, H., Ferraty, F., Sarda, P., 1999. Functional linear model. Statist. Probab. Lett. 45 (1), 11–22.

Cardot, H., Ferraty, F., Sarda, P., 2003. Spline estimators for the functional linear model. Statist. Sinica 13 (3), 571–592.

Chun, H., Keleş, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (1), 3–25.

Collazos, J.A., Dias, R., Zambom, A.Z., 2016. Consistent variable selection for functional regression models. J. Multivariate Anal. 146, 63–71.

Daubechies, I., 1990. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inform. Theory 36 (5), 961–1005.

Delaigle, A., Hall, P., 2012. Methodology and theory for partial least squares applied to functional data. Ann. Statist. 40 (1), 322–352.

Donoho, D.L., Johnstone, J.M., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81 (3), 425–455.

Fan, J., Lv, J., 2010. A selective overview of variable selection in high dimensional feature space. Statist. Sinica 20 (1), 101.

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

Gao, J., Kong, L., 2015. Quantile, composite quantile regression and regularized versions [R package cqrReg version 1.2]. Comprehensive R Archive Network (CRAN).

Gertheiss, J., Maity, A., Staicu, A.-M., 2013. Variable selection in generalized functional linear models. Stat 2 (1), 86–101.

Hestenes, M.R., 1969. Multiplier and gradient methods. J. Optim. Theory Appl. 4 (5), 303–320.

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

Kai, B., Li, R., Zou, H., 2011. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. Ann. Stat. 39 (1), 305.

Kato, K., 2012. Estimation in functional linear quantile regression. Ann. Statist. 40 (6), 3108–3136.

Koenker, R., 2005. Quantile Regression. Cambridge university press.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46 (1), 33–50.

Koenker, R., Park, B.J., 1996. An interior point algorithm for nonlinear quantile regression. J. Econometrics 71 (1), 265–283.

Kong, L., Shu, H., Heo, G., He, Q.C., 2015. Estimation for bivariate quantile varying coefficient model. arXiv preprint arXiv:1511.02552.

Kong, D., Xue, K., Yao, F., Zhang, H.H., 2016. Partially functional linear regression in high dimensions. Biometrika 103 (1), 147–159.

Konrad, K., Eickhoff, S.B., 2010. Is the adhd brain wired differently? a review on structural and functional connectivity in attention deficit hyperactivity disorder. Hum. Brain Mapp. 31 (6), 904–916.

Li, Y., Liu, Y., Zhu, J., 2007. Quantile regression in reproducing kernel hilbert spaces. J. Amer. Statist. Assoc. 102 (477), 255–268.

Lian, H., 2013. Shrinkage estimation and selection for multiple functional regression. Statist. Sinica 51–74.

Lin, C.-Y., Bondell, H., Zhang, H.H., Zou, H., 2013. Variable selection for non-parametric quantile regression via smoothing spline analysis of variance. Stat 2 (1), 255–268.

Lobo, M.S., Vandenberghe, L., Boyd, S., Lebret, H., 1998. Applications of second-order cone programming. Linear Algebra Appl. 284 (1–3), 193–228.

Lu, Y., Du, J., Sun, Z., 2014. Functional partially linear quantile regression model. Metrika 77 (2), 317–332.

Luo, R., Qi, X., 2015. Sparse wavelet regression with multiple predictive curves. J. Multivariate Anal. 134, 33–49.

Mallat, S., 2008. A wavelet tour of signal processing, third edition: the sparse way, 3rd ed. Academic Press.

Max, J.E., Manes, F.F., Robertson, B.A., Mathews, K., Fox, P.T., Lancaster, J., 2005. Prefrontal and executive attention network lesions and the development of attention-deficit/hyperactivity symptomatology. J. Am. Acad. Child Adolesc. Psychiatry 44 (5), 443–450.

Meinshausen, N., Bühlmann, P., 2010. Stability selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (4).

Mennes, M., Biswal, B.B., Castellanos, F.X., Milham, M.P., 2013. Making data sharing work: the fcp/indi experience. NeuroImage 82, 683–691.

Morris, J.S., 2015. Functional regression. Annu. Rev. Stat. Appl. 2.

Müller, H.-G., Yao, F., 2008. Functional additive models. J. Amer. Statist. Assoc. 103 (484), 1534–1544.

Ramsay, J.O., 2006. Functional Data Analysis. Wiley Online Library.

Reiss, P.T., Huo, L., Zhao, Y., Kelly, C., Ogden, R.T., 2015. Wavelet-domain regression and predictive inference in psychiatric neuroimaging. Ann. Appl. Stat. 9 (2), 1076.

Schrimsher, G.W., Billingsley, R.L., Jackson, E.F., Moore, B.D., 2002. Caudate nucleus volume asymmetry predicts attention-deficit hyperactivity disorder (adhd) symptomatology in children. J. Child Neurol. 17 (12), 877–884.

Sidlauskaite, J., Caeyenberghs, K., Sonuga-Barke, E., Roeyers, H., Wiersema, J.R., 2015. Whole-brain structural topology in adult attention-deficit/hyperactivity disorder: preserved global - disturbed local network organization. Neuroimage Clin. 9, 506–512.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A sparse-group lasso. J. Comput. Graph. Statist. 22 (2), 231–245.

Sun, Y., 2005. Semiparametric efficient estimation of partially linear quantile regression models. Ann. Econ. Financ. 6 (1), 105.

Tang, Q., Cheng, L., 2014. Partial functional linear quantile regression. Sci. China Math. 57 (12), 2589–2608.

Tomasi, D., Volkow, N.D., 2012. Abnormal functional connectivity in children with attention-deficit/hyperactivity disorder. Biol. Psychiatry 71 (5), 443–450.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15 (1), 273–289.

Wang, J.-L., Chiou, J.-M., Müller, H.-G., 2015. Review of functional data analysis. Annu. Rev. Stat. Appl. 1, 41.

Wang, X., Nan, B., Zhu, J., Koeppe, R., 2014. Regularized 3d functional regression for brain image data via Haar wavelets. Ann. Appl. Stat. 8 (2), 1045.

Wu, Y., Liu, Y., 2009. Variable selection in quantile regression. Statist. Sinica 19 (2), 801.

Yao, F., Sue-Chee, S., Wang, F., 2017. Regularized partially functional quantile regression. J. Multivariate Anal. 156, 39–56.

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

Zhang, Y., Li, R., Tsai, C.-L., 2010. Regularization parameter selections via generalized information criterion. J. Amer. Statist. Assoc. 105 (489), 312–323.

Zhao, Y., Chen, H., Ogden, R., 2015. Wavelet-based weighted lasso and screening approaches in functional linear regression. J. Comput. Graph. Statist. 24 (3), 655–675.

Zhao, Y., Ogden, R., Reiss, P.T., 2012. Wavelet-based lasso in functional linear regression. J. Comput. Graph. Statist. 21 (3), 600–617.

Zhao, W., Zhang, R., Liu, J., 2014. Sparse group variable selection based on quantile hierarchical lasso. J. Appl. Stat. 41 (8), 1658–1677.

Zheng, Q., Peng, L., He, X., 2015. Globally adaptive quantile regression with ultra-high dimensional data. Ann. Statist. 43 (5), 2225.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2), 301–320.

Zou, H., Yuan, M., 2008. Composite quantile regression and the oracle model selection theory. Ann. Statist. 36 (3), 1108–1126.