Electronic Journal of Statistics
Vol. 10 (2016) 3179–3216
ISSN: 1935-7524
DOI: 10.1214/16-EJS1204

Functional wavelet regression for linear function-on-function models

Ruiyan Luo

Division of Epidemiology and Biostatistics, Georgia State University School of Public Health, One Park Place, Atlanta, GA 30303
e-mail: rluo@gsu.edu

Xin Qi

Department of Mathematics and Statistics, Georgia State University, 30 Pryor Street, Atlanta, GA 30303
e-mail: xqi3@gsu.edu

and

Yanhong Wang

Department of Mathematics and Statistics, Georgia State University, 30 Pryor Street, Atlanta, GA 30303
e-mail: wangyanhongws@gmail.com

Abstract: We consider linear function-on-function regression models with multiple predictive curves. We first apply the wavelet transformation to the predictive curves and transform the original model to a linear model with functional response and high dimensional multivariate predictors. Based on the best finite dimensional approximation to the signal part in the response curve, we find an expansion of the vector of coefficient functions, which enjoys a good predictive property. To estimate this expansion, we propose a penalized generalized eigenvalue problem followed by a penalized least squares problem. We establish the sparse oracle inequalities for our estimates in the high-dimensional settings. The choices of tuning parameters and the number of components are provided. Simulations studies and application to real datasets demonstrate that our method has good predictive performance and is efficient in dimension reduction.

MSC 2010 subject classifications: 62J05.
Keywords and phrases: Function-on-function regression, wavelet transformation, signal compression, generalized eigenvalue problem, penalized least squares problem.

Received March 2016.

Contents

1. Introduction

With the development of technology and advance of complex data, functional data analysis (FDA) has received much attention in recent years, where the primary unit of observation is a curve or, in general, a function. For a general view of FDA, we refer the reader to Ramsay and Silverman [39], Ferraty and Vieu [11], Bosq [4], Horváth and Kokoszka [16], Hsing and Eubank [17]. Functional regression is a useful tool in FDA. Based on the type of variables, regression in FDA can be classified into three broad groups: scalar-on-function (scalar responses and functional predictors), function-on-scalar (functional responses and scalar predictors), and function-on-function (functional responses and functional predictors). Many methods have been developed for scalar-on-function and function-on-scalar regression models, including but not limited to Ramsay and Dalzell [38], Cardot *et al.* [6], Brown *et al.* [5], Ratcliffe *et al.* [41], Ramsay and Silverman [39], Reiss and Ogden [43], Marx and Eilers [28], James [20], Müller and Stadtmüller [32], Goldsmith *et al.* [13] for linear or generalized linear scalar-on-function regression models, James and Silverman [21], Li and Marx [24], Yao and Müller [53], McLean *et al.* [29] for non-linear scalar-on-function regression models, and Hart and Wehrly [15], Faraway [10], Guo *et al.* [14], Lin *et al.* [25], Morris and Carroll [31], Reiss *et al.* [42] on function-on-scalar regression models.

In this paper, we consider the following function-on-function linear regression model (in population level)

$$y(t) = \mu(t) + \sum_{q=1}^{Q} \int_{\mathcal{S}_q} \beta_q(t, s_q) z_q(s_q) ds_q + \varepsilon(t), \quad t \in \mathcal{T}, \tag{1.1}$$

where $\mathcal{S}_q$ and $\mathcal{T}$ are finite intervals. $z_1(s_1), z_2(s_2), \cdots, z_Q(s_Q)$ are Q predictive curves, $y(t)$ is the response curve, and $\beta_1(t, s_1)$, $\beta_2(t, s_2)$, $\cdots$, $\beta_Q(t, s_Q)$ are Q nonrandom integral kernel functions. No restrictions are imposed on the covariance function of the noise $\varepsilon(t)$ and various within correlation structures in $\varepsilon(t)$ are allowed.

Some work has been done for the linear function-on-function regression model with one predictor curve (in this case, we remove the subscript q). Ramsay and Dalzell [38] first used piecewise Fourier bases for $\beta(t, s)$, and Ramsay and Silverman [39] discussed the method of double expansion of the coefficient surface with basis function: $\sum_{k=1}^{\infty} \sum_{l=1}^{\infty} b_{kl} \eta_k(s) \theta_l(t)$, where $\{\eta_k(s) : k \geq 1\}$ are basis functions for $s \in \mathcal{S}$, $\{\theta_l(t) : l \geq 1\}$ are basis functions for $t \in \mathcal{T}$, and $\{b_{kl} : k, l \geq 1\}$ are the coefficients. Various basis functions can be used in the expansion such as the B-spline basis and the Fourier basis. Yao *et al.* [54] and Wu and Müller [52] used the eigenfunctions of the covariance functions of $x(s)$ and $y(t)$ for basis expansion. Ivanescu *et al.* [18] considered the model (1.1) with $Q > 1$ by representing the function-on-function regression model as a penalized additive model and then fitting the additive model. Scheipl *et al.* [46] extended the method in Ivanescu *et al.* [18] and proposed additive regression models for correlated functional responses, allowing functional random effects. Wang [51] developed a linear mixed function-on-function regression model and estimated parameters by maximizing the log likelihood via the ECME algorithm.

With its ability to extract local features of curves at different levels of resolution and the sparsity of the coefficient vector, wavelet transformation has been used in functional regression models. Brown *et al.* [5], Malloy *et al.* [27], Zhao *et al.* [55], Luo and Qi [26] considered the scalar-on-function models. They conducted wavelet transformations on predictive curves and transformed the scalar-on-function models into regression models with scalar responses and high dimensional multivariate predictors. Brown *et al.* [5] and Malloy *et al.* [27] built Bayesian models and regularized the regression coefficients by placing a spike-and-slab prior, a mixture of a normal distribution and a point mass at zero, on coefficients. Zhao *et al.* [55] applied the LASSO [49] to make feature selection and fit the high-dimensional regression model. Luo and Qi [26] proposed a penalized generalized eigenvalue problem to fit the high-dimensional regression model. Meyer *et al.* [30] considered function-on-function regression models with multilevel functional data. They transformed both the predictive and response curves to wavelet space and conducted variable selection in the Bayesian framework by assuming a spike-and-slab prior on the regression coefficients and vague proper priors on the variance components.

In this paper, we first apply the wavelet transformation to the functional predictors and transform the original function-on-function model (1.1) to a linear model with functional response and high dimensional multivariate predictor variable. With n observations, the transformed model has the form: $\mathbf{Y}(t) = \mu(t)\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}(t) + \boldsymbol{\varepsilon}(t)$, where $\mathbf{1}_n$ is the n-dimensional vector with all elements equal to one, $\mathbf{X}$ is an $n \times p$ matrix of wavelet coefficients of the original predictive curves, $\mathbf{Y}(t) = (y_1(t), \cdots, y_n(t))^{\mathrm{T}}$ is the vector of n observed response functions and $\boldsymbol{\beta}(t) = (\beta_1(t), \cdots, \beta_p(t))^{\mathrm{T}}$ is the vector of p coefficient functions.

For the transformed model, we propose a signal compression approach. Based on the best finite dimensional approximation to the signal function $\mathbf{X}\boldsymbol{\beta}(t)$, we establish an expansion of $\boldsymbol{\beta}(t)$ which has the form $\sum_{j=1}^{\infty} \boldsymbol{\alpha}_j w_j(t)$ and enjoys good prediction properties, where $\boldsymbol{\alpha}_j$ is a p-dimensional vector and $w_j(t)$ is a function. For any k, the truncated expansion $\sum_{j=1}^{k} \boldsymbol{\alpha}_j w_j(t)$ has nearly the smallest prediction errors among all k-dimensional estimates of form $\sum_{j=1}^{k} \mathbf{b}_j v_j(t)$ for arbitrary $\mathbf{b}_j \in \mathbb{R}^p$ and $v_j(t)$. To estimate this expansion, we propose a penalized generalized eigenvalue problem to estimate $\boldsymbol{\alpha}_j$, followed by a penalized least squares problem to estimate $w_j(t)$. We provide the oracle inequalities for our estimates. Simulation studies in various settings for both one and multiple prediction curves demonstrated that our approach has good predictive performance and is efficient in dimension reduction.

The rest of the paper is organized as follows. In Section 2, we discuss the wavelet transformation of model (1.1). Then for the transformed model, we introduce the signal compression approach in Section 3. The theoretical properties will be provided in Section 4. Simulation studies in various settings and application studies are provided in Sections 5 and 6, respectively. We summarize this paper in Section 7 and provide proofs for two theorems in Appendix. All the other proofs and additional simulations and figures are available in the authors' webpage.

2. Wavelet transformation

For notational convenience, without loss of generality, we assume that $\mathcal{T} = [0,1]$ and $\mathcal{S}_q = [0,1]$ for $q = 1,\dots,Q$, in the model (1.1). For the general theory of wavelets and its application in statistics, we refer the reader to the books Daubechies *et al.* [8] and Nason [33]. Let $\psi(x)$ denote a wavelet function (also called the mother wavelet) and $\phi(x)$ denote a scaling function (also called the father wavelet). They are chosen to have compact supports (i.e. they are equal to zero outside a bounded interval) (see Chapters 5 and 6 in Daubechies *et al.* [8] and Sections 2.3 and 2.4 in Nason [33]). Let

$$\phi_k(s) = \phi(s-k), \quad \psi_{j,k}(s) = 2^{j/2}\psi(2^j s - k),$$

where $j = 0, 1, 2, \cdots$ and $k = 0, \pm 1, \pm 2 \cdots$. The two indices j and k represent the dilation and translation, respectively. Then all these functions form a complete orthonormal basis of $L^2(\mathbb{R})$. A larger j indicates that the basis function $\psi_{j,k}(s)$ has a smaller support interval and a finer resolution. We expand the predictive curves in the model (1.1) using this wavelet basis,

$$z_q(s) = \sum_{k=-\infty}^{\infty} \widetilde{x}_k^q \phi_k(s) + \sum_{j=0}^{\infty} \sum_{k=-\infty}^{\infty} x_{jk}^q \psi_{jk}(s),$$

where $\widetilde{x}_k^q = \int_0^1 z_q(s)\phi_k(s)ds$ and $x_{jk}^q = \int_0^1 z_q(s)\psi_{jk}(s)ds$ are the wavelet coefficients. With this wavelet expansion, we have

$$\int_0^1 \beta_q(t,s)z_q(s)ds = \sum_{k=-\infty}^{\infty} \widetilde{x}_k^q \widetilde{\beta}_k^q(t) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} x_{jk}^q \beta_{jk}^q(t),$$

where $\widetilde{\beta}_k^q(t) = \int_0^1 \beta_q(t,s)\phi_k(s)ds$ and $\beta_{jk}^q(t) = \int_0^1 \beta_q(t,s)\psi_{jk}(s)ds$. Since $\beta_q(t,s)$ is smooth, both $\widetilde{\beta}_k^q(t)$ and $\beta_{jk}^q(t)$ are smooth functions. Because $\phi(s)$ has a bounded support and $\widetilde{\beta}_k^q(t) = \int_0^1 \beta_q(t,s)\phi_k(s)ds = \int_0^1 \beta_q(t,s)\phi(s-k)ds$, when the absolute value of k is large enough, we have $\phi(s-k) = 0$ for all $0 \le s \le 1$ and hence $\widetilde{\beta}_k^q(t) = 0$. Similarly, given $j \ge 0$, $\beta_{jk}^q(t)$ is a nonzero function only for a finite number of k's. We define the sets $\mathcal{K} = \{k : \widetilde{\beta}_k^q(t) \text{ is a nonzero function}\}$ and $\mathcal{K}_j = \{k : \widetilde{\beta}_{jk}^q(t) \text{ is a nonzero function}\}$ for each $j \ge 0$. Moreover, in this paper, we assume that $\beta_{jk}^q(t) = 0$ for all j with 2^j large enough for the following two reasons. First, because

$$x_{jk}^q = \int_0^1 z_q(s)\psi_{jk}(s)ds = 2^{-j/2}\int_a^b z_q(2^{-j}(u+k))\psi(u)du,$$

where the interval $[a,b]$ is the support of $\psi(s)$ (i.e. $\psi(s) = 0$ for all $s \notin [a,b]$). Then x_{jk}^q only depends on the values of $z_q(s)$ on an interval shorter than $(b-a)/2^j$. In practice, the curves are all discretely observed. When 2^j is sufficiently large so that $(b-a)/2^j$ is less than the distance between the adjacent observation points, we cannot calculate x_{jk}^q and obtain information about $x_{jk}^q\beta_{jk}^q(t)$. In this case, we can view the term $x_{jk}^q\beta_{jk}^q(t)$ as random noise. Second, since we have assumed that $\beta_q(t,s)$ is smooth, the wavelet coefficients have the l_1 sparsity property in that the sum of the absolute values of these coefficients have a relatively small value. Moreover, $\beta_{jk}^q(t)$ decreases fast as j increases. Thus, $\sum_{k=-\infty}^{\infty} \widetilde{x}_k^q \widetilde{\beta}_k^q(t) + \sum_{j=0}^{M}\sum_{k=-\infty}^{\infty} x_{jk}^q\beta_{jk}^q(t)$ will be a good approximation to $\beta_q(t,s)$ when 2^M is large. Based on these two reasons, we assume

$$\int_0^1 \beta_q(t,s)z_q(s)ds = \sum_{k\in\mathcal{K}} \widetilde{x}_k^q \widetilde{\beta}_k^q(t) + \sum_{j=0}^{M}\sum_{k\in\mathcal{K}_j} x_{jk}^q\beta_{jk}^q(t). \tag{2.1}$$

There are only finite nonzero terms in this expansion. To simplify notation, in the following, we use a single index to replace the triple index (q,j,k) and denote the expansion $\sum_{k\in\mathcal{K}} \widetilde{x}_k^q \widetilde{\beta}_k^q(t) + \sum_{j=0}^{M}\sum_{k\in\mathcal{K}_j} x_{jk}^q\beta_{jk}^q(t)$ by $\sum_{l=1}^{p} x_l\beta_l(t)$, where p is the total number of nonzero terms in this expansion. Then the model (1.1) is transformed to $y(t) = \mu(t) + \sum_{l=1}^{p} x_l\beta_l(t) + \varepsilon(t)$. Since p is typically large, the transformed model is a linear model with functional response and high-dimensional multivariate predictors.

Suppose that we have n independent observations $\{y_i(t), z_{iq}(s), 1 \le q \le Q\}$, $1 \le i \le n$, from the model (1.1). In this paper, we consider the case that for each $1 \le q \le Q$, the sample curves, $z_{iq}(s)$, $1 \le i \le n$, are observed at a common dense set of points in $[0,1]$. The set of observation points can be different for different q. The response curves $y_i(t)$, $1 \le i \le n$, are also densely observed at a common set in $[0,1]$. If for any $1 \le q \le Q$, the observation points for $z_{iq}(s)$ are equally spaced and the number of the observation points is $N_q = 2^{M_q}$ for some positive integer M_q, then we apply the discrete wavelet transformation (DWT) (Nason [33]) which converts the 2^{M_q}-dimensional vector of discrete observations of $z_{iq}(s)$ to a 2^{M_q}-dimensional wavelet coefficient vector. In simulations and applications of this paper, we use the package "wavethresh" (Nason [34]) of the R software (R Core Team [37]) and choose the default Daubechies least-asymmetric wavelet basis functions with filter number ten (Section 2.5.1 in Nason [33]). If the observation points of $z_{iq}(s)$ are not equally spaced or the number of the observation points N_q is not a power of 2, we first approximate $z_{iq}(s)$ by a basis expansion using the method in Chapter 5 in Ramsay and Silverman [39]. The details are provided in Section 6. Then we use the values of the approximation function at 2^{M_q} equally spaced points as new observations and make the DWT transformation. To choose M_q, we note that if M_q is too small, we may lose information in the original discrete observations of $z_{iq}(s)$; if M_q is too large, extra noise will be introduced. To make a balance, we choose M_q satisfying $2^{M_q-1} < N_q < 2^{M_q}$.

For the i-th observation, we concatenate the wavelet coefficient vectors for $z_{iq}(s)$, $1 \leq q \leq Q$, into a new vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathrm{T}}$. Then we have $y_i(t) = \mu(t) + \sum_{l=1}^{p} x_{il}\beta_l(t) + \varepsilon_i(t)$, $1 \leq i \leq n$. Let $\mathbf{Y}(t) = (y_1(t), \cdots, y_n(t))^{\mathrm{T}}$, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^{\mathrm{T}}$, $\boldsymbol{\beta}(t) = (\beta_1(t), \cdots, \beta_p(t))^{\mathrm{T}}$ and $\boldsymbol{\varepsilon}(t) = (\varepsilon_1(t), \cdots, \varepsilon_n(t))^{\mathrm{T}}$. Then we have

$$\mathbf{Y}(t) = \mu(t)\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}(t) + \boldsymbol{\varepsilon}(t). \tag{2.2}$$

In this paper, as in Bickel *et al.* [3], we will assume that $\mathbf{X}$ is a nonrandom matrix and the columns of $\mathbf{X}$ have mean zero. Throughout this paper, we use $\|\cdot\|_{L^2}$ to denote the L^2-norm in $L^2[0,1]$, and $\|\cdot\|_1$ and $\|\cdot\|_2$ to denote the l^1 and l^2-norm for vectors, respectively.

3. Signal compression approach

3.1. Expansion based on signal compression

Prediction based on model (2.2) is closely related to finding an efficient approximation to the signal $\mathbf{X}\boldsymbol{\beta}(t)$. Since $\mathbf{X}\boldsymbol{\beta}(t)$ is an n-dimensional vector of functions of t, we define an L^2 norm for a vector of functions. Let d be any integer and $\mathbf{M}(t) = (m_1(t), \cdots, m_d(t))^{\mathrm{T}}$ be a vector of functions. We define $\|\mathbf{M}\|_{L^2} = \sqrt{\sum_{j=1}^{d} \int_0^1 m_j(t)^2 dt} = \sqrt{\sum_{j=1}^{d} \|m_j\|_{L2}^2}$.

We will find a sequence of functions, $w_1(t)$, $w_2(t)$, $\cdots$, in $L^2[0,1]$ and a sequence of n-dimensional vectors $\mathbf{t}_1, \mathbf{t}_2, \cdots$, such that for any $k \geq 1$, $\sum_{j=1}^{k} \mathbf{t}_j w_j(t)$ is the best k-dimensional approximation to $\mathbf{X}\boldsymbol{\beta}(t)$ in the sense that

$$\|\mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{k} \mathbf{t}_j w_j\|_{L^2}^2 = \min_{\substack{\mathbf{r}_j \in \mathbb{R}^n, v_j(t) \in L^2[0,1], \\ 1 \leq j \leq k}} \|\mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{k} \mathbf{r}_j v_j\|_{L^2}^2, \tag{3.1}$$

where the minimum is taken over all possible functions $\{v_1(t), \cdots, v_k(t)\}$ in $L^2[0,1]$ and all possible n-dimensional vectors $\{\mathbf{r}_1, \cdots, \mathbf{r}_k\}$. So $\sum_{j=1}^k \mathbf{t}_j w_j(t)$ has the smallest approximation error among all k-dimensional approximations. To find these two sequences, we consider the generalized singular value decomposition (SVD) of $\mathbf{X}\boldsymbol{\beta}(t)$,

$$\mathbf{X}\boldsymbol{\beta}(t) = \sigma_1 \boldsymbol{\gamma}_1 u_1(t) + \sigma_2 \boldsymbol{\gamma}_2 u_2(t) + \cdots + \sigma_K \boldsymbol{\gamma}_K u_K(t), \tag{3.2}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K > 0$ is the collection of all positive singular values of $\mathbf{X}\boldsymbol{\beta}(t)$ (K is infinity if all $\sigma_k > 0$), $\boldsymbol{\gamma}_k \in \mathbb{R}^n$ and $u_k(t) \in L^2[0,1]$ are the left-singular vector and the right-singular function corresponding to σ_k with $\|\boldsymbol{\gamma}_k\|_2 = 1$ and $\|u_k(t)\|_{L^2} = 1$, respectively. Then $\{\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_K\}$ are orthogonal to each other, and so are $\{u_1(t), \cdots, u_K(t)\}$. We have $\int_0^1 \mathbf{X}\boldsymbol{\beta}(t) u_k(t) dt = \sigma_k \boldsymbol{\gamma}_k$. By the Eckart-Young Theorem, $\sum_{j=1}^k \sigma_k \boldsymbol{\gamma}_k u_k(t)$ is the best k-dimensional approximation to $\mathbf{X}\boldsymbol{\beta}(t)$. For any $1 \leq k \leq K$, we define

$$w_k(t) = \frac{\sigma_k}{\sqrt{n}} u_k(t), \quad \boldsymbol{\alpha}_k = \frac{n}{\sigma_k^2} \int_0^1 \boldsymbol{\beta}(t) w_k(t) dt, \tag{3.3}$$

where $\boldsymbol{\alpha}_k$ is a p-dimensional vector and the integral is coordinate-wise. Let

$$\mathbf{t}_k = \mathbf{X}\boldsymbol{\alpha}_k = \frac{n}{\sigma_k^2} \mathbf{X} \int_0^1 \boldsymbol{\beta}(t) w_k(t) dt = \frac{\sqrt{n}}{\sigma_k} \int_0^1 \mathbf{X}\boldsymbol{\beta}(t) u_k(t) dt = \sqrt{n}\boldsymbol{\gamma}_k. \tag{3.4}$$

Then by (3.3) and (3.4), $\sum_{j=1}^k \mathbf{t}_j w_j(t) = \sum_{j=1}^k \mathbf{X}\boldsymbol{\alpha}_j w_j(t) = \sum_{j=1}^k \sigma_j \boldsymbol{\gamma}_j u_j(t)$ is the best k-dimensional approximation to $\mathbf{X}\boldsymbol{\beta}(t)$ for any $1 \leq k \leq K$. Let

$$\boldsymbol{\beta}_k(t) = \boldsymbol{\alpha}_1 w_1(t) + \boldsymbol{\alpha}_2 w_2(t) + \cdots + \boldsymbol{\alpha}_k w_k(t), \quad 1 \leq k \leq K. \tag{3.5}$$

Then $\boldsymbol{\beta}(t) = \boldsymbol{\beta}_K(t) = \sum_{k=1}^K \boldsymbol{\alpha}_K w_K(t)$ is an expansion of $\boldsymbol{\beta}(t)$ and $\boldsymbol{\beta}_k(t)$, $k < K$, is the truncated expansion. We will show in Theorem 4.1 that $\boldsymbol{\beta}_k(t)$ has nearly the smallest prediction error among all k-dimensional expansions.

As a dimension reduction tool, the principal component analysis (PCA) has been conducted on the wavelet coefficients $\mathbf{X}$ in Johnstone and Lu [22], Røislien and Winje [45], Meyer *et al.* [30]. This dimension reduction procedure only involves the predictor variables and does not depend on the coefficient functions. For our decomposition, by (3.3), $w_k(t)$ is proportional to the k-th right eigenfunction in the SVD of $\mathbf{X}\boldsymbol{\beta}(t)$, hence it is also an eigenfunction of the sample covariance function of $\mathbf{X}\boldsymbol{\beta}(t)$. So the proposed decomposition can be viewed as a PCA procedure of the signal function $\mathbf{X}\boldsymbol{\beta}$, which not only depends on $\mathbf{X}$, but also is adaptive to the coefficient functions. The $w_k(t)$'s capture the major variations in the signal function. As we do not make restrictions on the covariance function of $\varepsilon(t)$, in general, $\{w_k(t) : 1 \leq k \leq K\}$ are not the eigenfunctions of the covariance function of $\mathbf{Y}(t)$. To estimate the decomposition (3.5), below we propose a penalized generalized eigenvalue problem to estimate $\boldsymbol{\alpha}_k$, and then estimate $w_k(t)$ by solving a penalized least squares problem.

3.2. Estimation of $\boldsymbol{\alpha}_k$

Define

$$\mathbf{Z} = \frac{1}{\sqrt{n}}\mathbf{X}, \quad \boldsymbol{\Xi} = \frac{1}{n}\int_0^1 (\mathbf{X}\boldsymbol{\beta}(t))(\mathbf{X}\boldsymbol{\beta}(t))^{\mathrm{T}} dt = \mathbf{Z}\left(\int_0^1 \boldsymbol{\beta}(t)\boldsymbol{\beta}(t)^{\mathrm{T}} dt\right)\mathbf{Z}^{\mathrm{T}}, \tag{3.6}$$

$$\mathbf{S} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}, \quad \mathbf{B} = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{Z} = \mathbf{S}\left(\int_0^1 \boldsymbol{\beta}(t)\boldsymbol{\beta}(t)^{\mathrm{T}} dt\right)\mathbf{S}\ ,$$

where the integrals are coordinate-wise. Then $\mathbf{S}$ is the $p \times p$ sample covariance matrix of $\mathbf{X}$, and the $p \times p$ matrix $\mathbf{B}$ and $n \times n$ matrix $\boldsymbol{\Xi}$ are symmetric and nonnegative definite. The following theorem provides a way to estimate $\boldsymbol{\alpha}_k$.

Theorem 3.1. *(a). We have $K \leq \min\{n, p\}$. For any $1 \leq k \leq K$, $\boldsymbol{\alpha}_k$ defined in (3.3) is the solution to the following generalized eigenvalue problem,*

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \quad \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}, \quad \text{subject to} \quad \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{S}\boldsymbol{\alpha} = 1, \quad \boldsymbol{\alpha}_l^{\mathrm{T}}\mathbf{S}\boldsymbol{\alpha} = 0, \tag{3.7}$$

for all $1 \leq l \leq k-1$.

(b). $\boldsymbol{\Xi}$ has exactly K positive eigenvalues $\mu_1(\boldsymbol{\Xi}) \geq \mu_2(\boldsymbol{\Xi}) \geq \cdots \geq \mu_K(\boldsymbol{\Xi}) > 0$. Moreover, the singular values in (3.2) satisfy $\sigma_k = \sqrt{n\mu_k(\boldsymbol{\Xi})}$, $1 \leq k \leq K$, and the maximum value of (3.7) is equal to $\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_k = \mu_k(\boldsymbol{\Xi}) = \sigma_k^2/n$.

(c). For any $1 \leq k \leq K$, the approximation error of the best k dimensional approximation to $\mathbf{X}\boldsymbol{\beta}(t)$ is

$$\begin{aligned}\frac{1}{n}\|\mathbf{X}\boldsymbol{\beta} - \sum_{i=1}^{k}\sigma_i\boldsymbol{\gamma}_i u_i\|_{L^2}^2 &= \frac{1}{n}\left\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\sum_{i=1}^{k}\boldsymbol{\alpha}_i w_i\right\|_{L^2}^2 \\ =&\frac{1}{n}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}_k\|_{L^2}^2 = \frac{1}{n}\sum_{i=k+1}^{K}\sigma_i^2 = \sum_{i=k+1}^{K}\mu_i(\boldsymbol{\Xi}).\end{aligned} \tag{3.8}$$

By Theorem 3.1(b), $\mu_k(\boldsymbol{\Xi}) = \sigma_k^2/n$ can be viewed as a measure of the magnitude of the signal in the k-th component of the SVD (3.2) of $\mathbf{X}\boldsymbol{\beta}(t)$. Through the SVD decomposition, the signal is compressed into the first few terms in the decomposition as much as possible. So we call our method the *wavelet based signal compression approach* (*wSigComp*). By (3.8), even if K is not small, as long as $\mu_k(\boldsymbol{\Xi})$ decreases fast enough, $\mathbf{X}\boldsymbol{\beta}(t)$ can be well approximated by the first few components.

To estimate $\boldsymbol{\alpha}_k$ from samples, we define

$$\widehat{\mathbf{B}} = \frac{1}{n^2}\mathbf{X}^{\mathrm{T}}\left\{\int_0^1 [\mathbf{Y}(t) - \bar{y}(t)\mathbf{1}_n][\mathbf{Y}(t) - \bar{y}(t)\mathbf{1}_n]^{\mathrm{T}} dt\right\}\mathbf{X} \tag{3.9}$$

as an estimate of $\mathbf{B}$, where $\bar{y}(t)$ is the sample mean of $y_1(t), \cdots, y_n(t)$. In high-dimensional settings, the solution to (3.7) with $\mathbf{B}$ replaced by $\widehat{\mathbf{B}}$ may not be a consistent estimate of $\boldsymbol{\alpha}_k$ as both the sample size n and the dimension p of $\mathbf{x}_i$

go to infinity. We recall that $\boldsymbol{\beta}(t)$ is the collection of all wavelet coefficients of the smooth coefficient functions $\beta_q(t, s_q)$, $1 \leq q \leq Q$, in the original function-on-function regression model (1.1). Therefore, $\boldsymbol{\beta}(t)$ has the sparsity property in the l_1 sense, which, together with (3.3), implies that $\boldsymbol{\alpha}_k$ is a sparse vector for any $1 \leq k \leq K$. So we propose a penalized generalized eigenvalue problem with sparsity penalty to estimate $\boldsymbol{\alpha}_k$. We propose to get the estimate $\widehat{\boldsymbol{\alpha}}_k$ of $\boldsymbol{\alpha}_k$ by solving

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{\boldsymbol{\alpha}^{\mathrm{T}} \widehat{\mathbf{B}} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{S} \boldsymbol{\alpha} + \tau \|\boldsymbol{\alpha}\|_\lambda^2}, \quad \text{subject to} \quad \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{S} \boldsymbol{\alpha} = 1, \quad \widehat{\boldsymbol{\alpha}}_l^{\mathrm{T}} \mathbf{S} \boldsymbol{\alpha} = 0, \tag{3.10}$$

for all $1 \leq l \leq k-1$, where $\|\boldsymbol{\alpha}\|_\lambda^2 = (1-\lambda)\|\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1^2$, and both $\tau \geq 0$ and $0 \leq \lambda < 1$ are tuning parameters. In the penalty $\tau \|\boldsymbol{\alpha}\|_\lambda^2$, the l_2 term is used to overcome the singularity problem of $\mathbf{S}$ and the l_1 term encourages the sparsity of $\widehat{\boldsymbol{\alpha}}_k$. This penalty $\tau \|\boldsymbol{\alpha}\|_\lambda^2$ was introduced in Qi *et al.* [36] for sparse principal component analysis and used in Qi *et al.* [35] for sparse regression and sparse discriminant analysis.

Suppose that $\{y_i(t), 1 \leq i \leq n\}$ are observed at L common observation points $0 = t_1 < t_2 < \cdots < t_L = 1$. For any continuous functions $g(t)$, $0 \leq t \leq 1$, we approximate the integral by $\int_c^d g(v)dv \approx \sum_{\ell=1}^L \delta_\ell g(t_\ell)$, where $\{\delta_\ell : 1 \leq \ell \leq L\}$ are weights. There are several choices for weights by different interpolation formulas. For example, for equally spaced observation points, we can choose $\delta_\ell = 1/L$; for unequally spaced observation points, we can choose $\delta_1 = (t_2 - t_1)/2$, $\delta_\ell = (t_{(\ell+1)} - t_{(\ell-1)})/2$ for $1 < \ell < L$, $\delta_L = (t_L - t_{(L-1)})/2$ based on the trapezoidal formula. We use these approximations to calculate the integrals in the expression (3.9) of $\widehat{\mathbf{B}}$.

To solve the penalized optimization problem (3.10), we first note that due to the scale-invariant property, (3.10) is equivalent to

$$\max \quad \boldsymbol{\alpha}^{\mathrm{T}} \widehat{\mathbf{B}} \boldsymbol{\alpha}, \quad \text{subject to} \quad \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{S} \boldsymbol{\alpha} + \tau \|\boldsymbol{\alpha}\|_\lambda^2 \leq 1 \text{ and } \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{S} \widehat{\boldsymbol{\alpha}}_l = 0, \tag{3.11}$$

for all $1 \leq l \leq k-1$. The solution to (3.11) differs from that of (3.10) only by a scale factor. An algorithm to solve a more general optimization problem than (3.11) has been proposed in Qi *et al.* [35]. We apply the algorithm to solve (3.11) and then scale the solution to obtain the estimate $\widehat{\boldsymbol{\alpha}}_k$.

3.3. Estimation of $\boldsymbol{\mu(t)}$ and $\boldsymbol{w_k(t)}$

With the estimates $\widehat{\boldsymbol{\alpha}}_k$, $1 \leq k \leq K$, we next estimate $\mu(t)$ and $w_k(t)$, $1 \leq k \leq K$. Let $\mathbf{T} = [\mathbf{t}_1, \cdots, \mathbf{t}_K]$ and $\mathbf{W}(t) = (w_1(t), \cdots, w_K(t))^{\mathrm{T}}$, where $\mathbf{t}_k$ and w_k are the left-singular vector and right-singular function of the SVD of $\mathbf{X}\boldsymbol{\beta}(t)$, respectively, as seen from (3.2), (3.3) and (3.4). Since $\mathbf{X}\boldsymbol{\beta}(t) = \sum_{k=1}^K \mathbf{t}_k w_k(t) = \mathbf{T}\mathbf{W}(t)$, the model (2.2) can be transformed to

$$\mathbf{Y}(t) = \mu(t)\mathbf{1}_n + \sum_{k=1}^K \mathbf{t}_k w_k(t) + \boldsymbol{\varepsilon}(t) = \mu(t)\mathbf{1}_n + \mathbf{T}\mathbf{W}(t) + \boldsymbol{\varepsilon}(t), \tag{3.12}$$

where $\mathbf{t}_1, \cdots, \mathbf{t}_K$ can be viewed as new predictors and $w_1(t), \cdots, w_K(t)$ are the coefficient functions. The true value of $\mathbf{t}_k$ is not observed but can be estimated by $\widehat{\mathbf{t}}_k = \mathbf{X}\widehat{\boldsymbol{\alpha}}_k$, $1 \le k \le K$. We propose to estimate $\mu(t)$ and $\mathbf{W}(t)$ by regressing $\mathbf{Y}(t)$ on $\widehat{\mathbf{T}} = [\widehat{\mathbf{t}}_1, \cdots, \widehat{\mathbf{t}}_K]$. It follows from the definition of $w_k(t)$ in (3.3) and the smoothness of $\boldsymbol{\beta}(t)$ that $\{w_1(t), \cdots, w_K(t)\}$ are smooth functions. To take care of this property, we use the penalized least squares method for the function-on-scalar regression in Chapter 13 of Ramsay and Silverman [39]. Specifically, the estimates $\widehat{\mu}(t)$ and $\widehat{\mathbf{W}}(t) = (\widehat{w}_1(t), \cdots, \widehat{w}_K(t))^{\mathrm{T}}$ are the solution to

$$\min_{\substack{\nu(t), v_1(t),\\ \cdots, v_K(t)}} \left\{ \frac{1}{n} \left\| \mathbf{Y} - \nu \mathbf{1}_n - \sum_{k=1}^{K} \widehat{\mathbf{t}}_k v_k \right\|_{L^2}^2 + \eta \left(\|\nu''\|_{L^2}^2 + \sum_{k=1}^{K} \|v_k''\|_{L^2}^2 \right) \right\}, \tag{3.13}$$

where the first term is the mean squared residuals, the second term is the smoothness penalty, and η is the smooth tuning parameter.

In a general function-on-scalar regression model, all the coefficient functions are estimated simultaneously by solving (3.13), which leads to a heavy computational load. In our situation, we can estimate $\mu(t)$ and $w_1(t), \cdots, w_K(t)$ separately, which improves the computational efficiency. As $\widehat{\mathbf{t}}_k = \mathbf{X}\widehat{\boldsymbol{\alpha}}_k$, due to the constraints in (3.10), $\widehat{\mathbf{t}}_k^{\mathrm{T}} \widehat{\mathbf{t}}_j / n = \widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}} \mathbf{S} \widehat{\boldsymbol{\alpha}}_j = 1$ if $k = j$ and 0 if $k \neq j$. So $\widehat{\mathbf{t}}_1, \cdots, \widehat{\mathbf{t}}_K$ are orthogonal to each other and $\|\widehat{\mathbf{t}}_k\|_2^2 / n = 1$. Moreover, since the column means of $\mathbf{X}$ are zero, $\widehat{\mathbf{t}}_k$ also has mean zero for any $1 \le k \le K$. This means that $\widehat{\mathbf{t}}_1, \cdots, \widehat{\mathbf{t}}_K$ are also orthogonal to $\mathbf{1}_n$. Hence, the penalized least squares problem (3.13) can be decomposed as $K + 1$ sub-problems. Specifically, $\widehat{\mu}(t)$ is the solution to

$$\min_{\nu(t)} \left[\|\nu - \bar{y}\|_{L^2}^2 + \eta \|\nu''\|_{L^2}^2 \right], \tag{3.14}$$

and for any $1 \le k \le K$, $\widehat{w}_k(t)$ is the solution to

$$\min_{v_k(t)} \left[\|v_k - \widehat{w}_k^0\|_{L^2}^2 + \eta \|v_k''\|_{L^2}^2 \right], \tag{3.15}$$

where $\widehat{w}_k^0(t) = \frac{1}{n} \widehat{\mathbf{t}}_k^{\mathrm{T}} \mathbf{Y}(t)$. For any k, $\widehat{w}_k(t)$ only depends on $\widehat{\mathbf{t}}_k = \mathbf{X}\widehat{\boldsymbol{\alpha}}_k$. Because $\widehat{\boldsymbol{\alpha}}_k$ only depends on $\widehat{\boldsymbol{\alpha}}_1, \cdots, \widehat{\boldsymbol{\alpha}}_{k-1}$ and does not depend on $\widehat{\boldsymbol{\alpha}}_{k+1}, \widehat{\boldsymbol{\alpha}}_{k+2}, \cdots$, the estimates $\widehat{\boldsymbol{\alpha}}_k$, $\widehat{w}_k(t)$ and $\widehat{\boldsymbol{\beta}}_k(t)$ do not depend on the choice of the number of components. This property helps to improve the computational efficiency in practice.

To solve (3.14) and (3.15), we approximate the solutions by basis expansions. Let $\boldsymbol{\Phi}(t) = (\Phi_1(t), \cdots, \Phi_P(t))^{\mathrm{T}}$, where Φ_j, $1 \le j \le P$, are P basis functions in $L^2[0,1]$. Since we only observe $\mathbf{Y}(t)$ at $t_1, \cdots, t_L$, we approximate (3.14) by the following problem,

$$\min_{\mathbf{h} \in \mathbb{R}^p} \left[\sum_{\ell=1}^{L} \{\boldsymbol{\Phi}(t_\ell)^{\mathrm{T}} \mathbf{h} - \bar{y}(t_\ell)]^2 \delta_\ell + \eta \mathbf{h}^{\mathrm{T}} \left(\int_c^d \boldsymbol{\Phi}''(t) \boldsymbol{\Phi}''(t) dt \right) \mathbf{h} \right], \tag{3.16}$$

where $\mathbf{h}$ is the expansion coefficient vector of $\mu(t)$. The solution of (3.16) is

$$\widehat{\mathbf{h}} = \left[\sum_{\ell=1}^{L} \mathbf{\Phi}(t_\ell)\mathbf{\Phi}(t_\ell)^{\mathrm{T}}\delta_\ell + \eta\int_c^d \mathbf{\Phi}''(t)\mathbf{\Phi}''(t)dt\right]^{-1}\left[\sum_{\ell=1}^{L} \mathbf{\Phi}(t_\ell)\bar{y}(t_\ell)\delta_\ell\right],$$

and the estimate $\widehat{\mu}(t) = \mathbf{\Phi}(t)^{\mathrm{T}}\widehat{\mathbf{h}}$. We approximate (3.15) by

$$\min_{\mathbf{d}\in\mathbb{R}^p}\left[\sum_{\ell=1}^{L}[\mathbf{\Phi}(t_\ell)^{\mathrm{T}}\mathbf{d} - \widehat{w}_\ell^0(t_\ell)]^2\delta_\ell + \eta\mathbf{d}^{\mathrm{T}}\left(\int_c^d \mathbf{\Phi}''(t)\mathbf{\Phi}''(t)dt\right)\mathbf{d}\right], \tag{3.17}$$

which has solution

$$\widehat{\mathbf{d}}_k = \left[\sum_{\ell=1}^{L} \mathbf{\Phi}(t_\ell)\mathbf{\Phi}(t_\ell)^{\mathrm{T}}\delta_\ell + \eta\int_c^d \mathbf{\Phi}''(t)\mathbf{\Phi}''(t)dt\right]^{-1}\left[\sum_{\ell=1}^{L} \mathbf{\Phi}(t_\ell)\widehat{w}_\ell^0(t_\ell)\delta_\ell\right].$$

Then the estimate $\widehat{w}_k(t) = \mathbf{\Phi}(t)^{\mathrm{T}}\widehat{\mathbf{d}}_k$.

Finally, $\boldsymbol{\beta}_k(t)$ is estimated as $\widehat{\boldsymbol{\beta}}_k(t) = \widehat{\boldsymbol{\alpha}}_1\widehat{w}_1(t) + \widehat{\boldsymbol{\alpha}}_2\widehat{w}_2(t) + \cdots + \widehat{\boldsymbol{\alpha}}_k\widehat{w}_k(t)$ for $1 \le k \le K$, and $\boldsymbol{\beta}(t)$ is estimated by $\widehat{\boldsymbol{\beta}}(t) = \widehat{\boldsymbol{\beta}}_K(t)$.

3.4. Choice of the number of components and tuning parameters

We first consider the choice of (τ, λ). Usually one jointly chooses two tuning parameters in a two dimensional grid. But in our situations, the two tuning parameters are not equally important. Theorem 4.2 in section 4 implies that λ is not essential for the convergence rates in our theoretical results. We can choose λ to be any number as long as it is bounded away from zero and does not affect the convergence rates. On the other hand, in the penalty $\tau\|\boldsymbol{\alpha}\|_\lambda^2 = \tau(1-\lambda)\|\boldsymbol{\alpha}\|_2^2 + \tau\lambda\|\boldsymbol{\alpha}\|_1^2$, the sparsity is mainly determined by the l_1-norm part $\tau\lambda\|\boldsymbol{\alpha}\|_1^2$. Roughly speaking, the effect of (τ, λ) on the sparsity of our estimates is mainly through their product $\tau\lambda$ and thus a small τ with a large λ has a similar effect as a large τ with a small λ. Hence, to improve the computational efficiency, we do not consider all possible pairs of (τ, λ) in a two dimensional grid. Instead, we select the parameters from a set of paired values where with the increase of τ, the value of λ also increases. Specifically, in the following simulation studies and applications, we choose the paired-value for (τ, λ) from 7 pairs, $(0.01, 0.01)$, $(0.05, 0.01)$, $(0.05, 0.05)$, $(0.1, 0.05)$, $(0.1, 0.1)$, $(0.5, 0.1)$, $(0.5, 0.2)$. The smoothness tuning parameter η is chosen from $\{10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$.

For the i-th pair of (τ, λ), we first determine the maximum number of components $\widehat{K}_i$ we need to calculate, $1 \le i \le 7$. The optimal number of components will be chosen between 1 and these maximum numbers. As mentioned after Theorem 3.1, $\mu_k(\mathbf{\Xi}) = \boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_k$ measures the signal magnitude of the k-th component. As $\widehat{\mu}_k(\mathbf{\Xi}) = \widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_k$ is an estimate of $\mu_k(\mathbf{\Xi})$, we only compute the first few components with large values of $\widehat{\mu}_k(\mathbf{\Xi})$ and stop when the value becomes small enough. On the other hand, by Theorem 3.1 (a), the number of components cannot exceed $\min(n, p)$. Based on these considerations, we define

$$\widehat{K}_i = \min\{n, p, \widehat{K}_i^{(1)}\}, \tag{3.18}$$

$$\text{where}\quad \widehat{K}_i^{(1)} = \min\left\{k > 1 : \frac{\widehat{\mu}_k(\boldsymbol{\Xi})}{\widehat{\mu}_1(\boldsymbol{\Xi}) + \cdots + \widehat{\mu}_k(\boldsymbol{\Xi})} \le 0.02\right\}.$$

So we stop solving the sequential problems (3.10) when the number of components reaches any of the two numbers n, p, or when the ratio between $\widehat{\mu}_k(\boldsymbol{\Xi})$ and the cumulative sum $\widehat{\mu}_1(\boldsymbol{\Xi}) + \cdots + \widehat{\mu}_k(\boldsymbol{\Xi})$ is less than 2%. Once we have determined the $\widehat{K}_i$ for all $1 \le i \le 7$, we use the cross-validation method to determine the tuning parameters and the optimal number of components simultaneously. We summarize the details of the procedure in the following algorithm.

Algorithm 3.1. *1. For the i-th paired value of (τ, λ), $1 \le i \le 7$, we use the whole data set to determine $\widehat{K}_i$ by (3.18).*

2. We use the five-fold cross-validation to determine the number of components and the tuning parameters. Specifically, we split the whole data set into five subsets and repeat the following procedure. For $1 \le l \le 5$, we use the l-th subset as the l-th validation set and all other observations as the l-th training set. Then for the i-th pair of values for (τ, λ) and the l-th training set,

(a) we estimate the first $\widehat{K}_i$ components, $\widehat{\boldsymbol{\alpha}}_1^{(li)}, \ldots, \widehat{\boldsymbol{\alpha}}_{\widehat{K}_i}^{(li)}$.

(b) For each $1 \le j \le 6$, we use the j-th value for η and $\{\widehat{\boldsymbol{\alpha}}_k^{(li)}, 1 \le k \le \widehat{K}_i\}$, to obtain the estimates $\widehat{\mu}^{(lij)}(t)$ and $\widehat{w}_k^{(lij)}(t)$, $1 \le k \le \widehat{K}_i$. Then for each $1 \le k \le \widehat{K}_i$, we define the estimate $\widehat{\boldsymbol{\beta}}^{(lijk)}(t) = \widehat{\boldsymbol{\alpha}}_1^{(li)} \widehat{w}_1^{(lij)}(t) + \cdots + \widehat{\boldsymbol{\alpha}}_k^{(li)} \widehat{w}_k^{(lij)}(t)$ of $\boldsymbol{\beta}(t)$.

(c) We apply $\widehat{\mu}^{(lij)}(t)$ and $\widehat{\boldsymbol{\beta}}^{(lijk)}(t)$ to the l-th validation data set to obtain the predicted curves $\{\widehat{y}_m^{(lijk)}, 1 \le m \le n_l\}$ and calculate the validation error $e_{ijk}^{(l)} = \sum_{m=1}^{n_l} \sum_{\ell=1}^{L} (\widehat{y}_m^{(lijk)}(t_\ell) - y_m^{(l)}(t_\ell))^2 \delta_\ell / n_l$, where $\{y_m^{(l)}, 1 \le m \le n_l\}$ are the observed response curves and n_l is the number of observations in the l-th validation set.

(d) Finally, we calculate the average validation error, $\bar{e}_{ijk} = (e_{ijk}^{(1)} + \cdots + e_{ijk}^{(5)})/5$, for the i-th pair of values for (τ, λ), the j-th value for η and the first k components.

Let $\bar{e}_{i_0 j_0 K_0} = \min_{i,j,k} \bar{e}_{ijk}$. Then the i_0-th pair of values for (τ, λ) and the j_0-th value for η are chosen, and the optimal number of components is $\widehat{K}_{opt} = K_0$.

4. Oracle inequalities in high-dimensional settings

Now we provide oracle inequalities for the estimates of $\boldsymbol{\alpha}_k$, $w_k(t)$ and $\boldsymbol{\beta}_k(t)$, $1 \le k \le K$, in high-dimensional settings. These oracle inequalities hold for any n, p and $\boldsymbol{\beta}(t)$ which satisfies the conditions in this section.

We introduce some notation. For any $d \times d$ symmetric and nonnegative definite matrix $\mathbf{M}$, where d is any positive integer, we define two norms, *the operator norm* $\|\mathbf{M}\| = \sup_{\mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\|_2 = 1} \|\mathbf{M}\mathbf{v}\|_2 = \lambda_{max}(\mathbf{M})$ and *the max norm* $\|\mathbf{M}\|_\infty =$

$\max_{1\le k,l\le d}|M_{kl}|$, where M_{kl} is the (k,l)-th entry of $\mathbf{M}$, and $\lambda_{max}(\mathbf{M})$ denotes the largest eigenvalue of $\mathbf{M}$. We adopt the notation in Bickel *et al.* [3]. For any p-dimensional vector $\mathbf{a} = (a_1, \cdots, a_p)^{\mathrm{T}}$, let $J(\mathbf{a}) = \{j \in \{1, \cdots, p\} : a_j \neq 0\}$ denote the collection of indices of nonzero coordinates of $\mathbf{a}$ and $\mathcal{M}(\mathbf{a}) = |J(\mathbf{a})|$ denote the number of nonzero coordinates of $\mathbf{a}$, where $|J(\mathbf{a})|$ is the cardinality of $J(\mathbf{a})$. $\mathcal{M}(\mathbf{a})$ measures the sparsity of $\mathbf{a}$. Similarly, we define $J(\boldsymbol{\beta}(t)) = \{j \in \{1, \cdots, p\} : \beta_j(t) \neq \mathbf{0}\}$ and $\mathcal{M}(\boldsymbol{\beta}(t)) = |J(\boldsymbol{\beta}(t))|$. Then it follows from the definitions of $\boldsymbol{\alpha}_k$ in (3.3) and $\boldsymbol{\beta}_k$ in (3.5) that for any $1 \le k \le K$,

$$J(\boldsymbol{\alpha}_k) \subset J(\boldsymbol{\beta}(t)), \quad \mathcal{M}(\boldsymbol{\alpha}_k) \le \mathcal{M}(\boldsymbol{\beta}(t)), \quad \mathcal{M}(\boldsymbol{\beta}_k(t)) \le \mathcal{M}(\boldsymbol{\beta}(t)). \tag{4.1}$$

Before we provide the main results, we first show that $\boldsymbol{\beta}_k(t)$ has nearly the smallest prediction error among all k-dimensional estimates when n is large. Let $\mathbf{x}^{\text{new}}$ be the vector of wavelet coefficients of a new observation of predictive curves and has the covariance matrix $\boldsymbol{\Sigma}$. The corresponding new response is $y^{\text{new}}(t) = \mu(t) + (\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t) + \varepsilon^{\text{new}}(t)$, where $\varepsilon^{\text{new}}(t)$ is independent of $\mathbf{x}^{\text{new}}$.

Theorem 4.1. *Suppose that $\|\mathbf{S} - \boldsymbol{\Sigma}\|_\infty \le C\sqrt{\frac{\ln p}{n}}$, where C is a constant which does not depend on n and p. Let $s = \mathcal{M}(\boldsymbol{\beta}(t))$, then we have*

$$\begin{aligned}&E\left[\|\mu(t) + (\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}_k(t) - y^{\text{new}}(t)\|_{L^2}^2\right]\\ \le& \min_{\widetilde{\boldsymbol{\beta}}_k} E\left[\|\mu(t) + (\mathbf{x}^{\text{new}})^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_k(t) - y^{\text{new}}(t)\|_{L^2}^2\right] + 2s\sqrt{\frac{\ln p}{n}}C\|\boldsymbol{\beta}(t)\|_{L^2}^2,\end{aligned}$$

where the minimum is taken over all possible $\widetilde{\boldsymbol{\beta}}_k$ of the forms $\sum_{j=1}^k \mathbf{b}_j v_j(t)$ with arbitrary $\mathbf{b}_j \in \mathbb{R}^p$ and $v_j(t) \in L^2[0,1]$.

Since $\boldsymbol{\beta}(t)$ is the collection of the wavelet coefficient functions for $\beta_q(t,s)$, $1 \le q \le Q$, $\|\boldsymbol{\beta}(t)\|_{L^2}^2 = \sum_{q=1}^Q \int_0^1\int_0^1 \beta_q(t,s)^2 dtds$. The term $s\sqrt{\ln(p)/n}$ is the convergence rate of the LASSO and the Dantzig selector provided in Bickel and Levina [2]. Since $\boldsymbol{\beta}(t)$ is sparse, $s\sqrt{\ln(p)/n}\|\boldsymbol{\beta}(t)\|_{L^2}^2$ is small when n and p are large. Thus, the prediction error of $\boldsymbol{\beta}_k(t)$ is close to the smallest one among all k-dimensional estimates. We assume that $\|\mathbf{S} - \boldsymbol{\Sigma}\|_\infty \le C\sqrt{\ln(p)/n}$ because it has been shown (Equation (A14) in Bickel and Levina [2]) that $\sqrt{\ln(p)/n}$ is the order of the max norm of the difference between the sample covariance matrix and population covariance matrix of p-dimensional multivariate normal distribution.

We state three regularity conditions for the main theorem. In the setting of large p and small n, the identification problem exists for the model (2.2). That is, there exists $\widetilde{\boldsymbol{\beta}}(t) \neq \boldsymbol{\beta}(t)$ such that $\mathbf{X}\widetilde{\boldsymbol{\beta}}(t) = \mathbf{X}\boldsymbol{\beta}(t)$. Bickel *et al.* [3] imposed the restricted eigenvalue assumptions on $\mathbf{X}$. We will make the same assumption below in Condition 1.

Condition 1. *Let $s = \mathcal{M}(\boldsymbol{\beta}(t))$ and*

$$\kappa = \min_{\substack{J_0 \subset \{1,\cdots,p\},\\ |J_0| \le s}} \; \min_{\substack{\mathbf{0}\neq\boldsymbol{\delta}\in\mathbf{R}^p,\\ \left\|\boldsymbol{\delta}_{J_0^c}\right\|_1 \le c\left\|\boldsymbol{\delta}_{J_0}\right\|_1}} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\sqrt{n}\|\boldsymbol{\delta}_{J_0}\|_2},$$

where $c > 1$ and $\kappa > 0$ are two constants, $\boldsymbol{\delta}_{J_0}$ and $\boldsymbol{\delta}_{J_0^c}$ are the subvectors of $\boldsymbol{\delta}$ consisting of the coordinates of $\boldsymbol{\delta}$ with indices belonging to J_0 and J_0^c, respectively.

Although this assumption does not lead to the identification of the model among all possible coefficients, it makes the model identifiable among all sparse coefficients. In fact, under this assumption, for any two sparse p-dimensional vectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ with sparsity $\mathcal{M}(\boldsymbol{\alpha}) \leq s$ and $\mathcal{M}(\boldsymbol{\alpha}') \leq s$, if $\mathbf{X}\boldsymbol{\alpha} = \mathbf{X}\boldsymbol{\alpha}'$, then we have $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$ (see the second remark after Theorem 7.3 in Bickel *et al.* [3]). Therefore, the model (2.2) is identifiable among all vectors sparser than or as sparse as $\boldsymbol{\beta}(t)$. That is, if $\mathbf{X}\widetilde{\boldsymbol{\beta}}(t) = \mathbf{X}\boldsymbol{\beta}(t)$ and $\mathcal{M}(\widetilde{\boldsymbol{\beta}}(t)) \leq \mathcal{M}(\boldsymbol{\beta}(t))$, then $\widetilde{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}(t)$. Because $\mathbf{X}\boldsymbol{\beta}(t) = \mathbf{X}\boldsymbol{\beta}_K(t)$ and $\mathcal{M}(\boldsymbol{\beta}_K(t)) \leq \mathcal{M}(\boldsymbol{\beta}(t))$ by (4.1), we have $\boldsymbol{\beta}(t) = \boldsymbol{\beta}_K(t)$.

The next regularity condition is on the distribution of the random noise function $\varepsilon_i(t)$.

<u>Condition</u> 2. *The noise functions $\varepsilon_i(t)$, $1 \leq i \leq n$, are Gaussian processes taking values in $L^2[0,1]$. $\{\varepsilon_1(t), \cdots, \varepsilon_n(t)\}$ have the same distribution and are between-function independent.*

Note that we do not make restrictions on the within-function covariance of $\varepsilon_i(t)$. We allow $\varepsilon_i(t')$ and $\varepsilon_i(t'')$ to be correlated for any $0 \leq t', t'' \leq 1$. We define the median M_ε and variance σ^2 for $\varepsilon_i(t)$. M_ε is defined to be the median of the real-valued random variable $\|\varepsilon_i(t)\|_{L^2}$ and the variance is defined as (Section 3.1 in Ledoux and Talagrand [23])

$$\sigma^2 = \sup_{u(t)\in L^2[0,1], \|u(t)\|_{L^2}=1} E\left[\left\{\int_0^1 u(t)\varepsilon_i(t)dt\right\}^2\right]. \tag{4.2}$$

$\int_0^1 u(t)\varepsilon_i(t)dt$ is the length of the projection of $\varepsilon_i(t)$ onto the direction of $u(t)$ and has a normal distribution. Therefore, (4.2) means that σ^2 is the maximum of the variances of the projections of $\varepsilon_i(t)$ along all the possible directions in $L^2[0,1]$. In our theoretical development, we need to estimate the tail probabilities of the norms of $L^2[0,1]$-valued Gaussian variables which can be controlled by M_ε and σ^2 (Section 3.1 in Ledoux and Talagrand [23]).

<u>Condition</u> 3. *All the diagonal elements of $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/n$ are equal to 1. All the positive eigenvalues, $\mu_1(\boldsymbol{\Xi})$, $\cdots$, $\mu_K(\boldsymbol{\Xi})$ of $\boldsymbol{\Xi}$ are different. Let*

$$c_2 = \min\left\{\frac{\mu_1(\boldsymbol{\Xi}) - \mu_2(\boldsymbol{\Xi})}{\mu_1(\boldsymbol{\Xi})}, \quad \frac{\mu_2(\boldsymbol{\Xi}) - \mu_3(\boldsymbol{\Xi})}{\mu_2(\boldsymbol{\Xi})}, \cdots, \frac{\mu_{K-1}(\boldsymbol{\Xi}) - \mu_K(\boldsymbol{\Xi})}{\mu_{K-1}(\boldsymbol{\Xi})}\right\},$$
$$c_3 = \mu_1(\boldsymbol{\Xi})/\mu_K(\boldsymbol{\Xi}).$$

Bickel *et al.* [3] assumed that the diagonal elements of $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/n$ are equal to 1 as they derived the oracle inequalities for the Lasso and the Dantzig selector. This assumption can be satisfied by scaling $\mathbf{X}$. The c_2 measures how well the eigenvalues of $\boldsymbol{\Xi}$ can be separated.

We first provide upper bounds on the l_1 norms (that is, the l_1 sparsity) of $\widehat{\boldsymbol{\alpha}}_k$, $1 \leq k \leq K$, and the oracle inequalities for them in the following theorem. Although we use the same tuning parameters (τ, λ) in (3.10) for all components for computational efficiency in practice, in our theoretical results, we allow different tuning parameters for different components. We use $(\tau^{(k)}, \lambda^{(k)})$ to denote the tuning parameters for the k-th component, $1 \leq k \leq K$. Let $\widehat{\boldsymbol{\gamma}}_k = \mathbf{Z}\widehat{\boldsymbol{\alpha}}_k = \widehat{\mathbf{t}}_k/\sqrt{n}$, which are estimates of $\boldsymbol{\gamma}_k = \mathbf{Z}\boldsymbol{\alpha}_k = \mathbf{t}_k/\sqrt{n}$ for $1 \leq k \leq K$. Define $\varpi = \sqrt{\ln(p)/n}$ and recall that $s = \mathcal{M}(\boldsymbol{\beta}(t))$.

Theorem 4.2. *Assume that Conditions 1-3 hold. Suppose that*

$$\mu_1(\boldsymbol{\Xi}) \geq \hbar^2 C_0^2 \varpi^2 s/\kappa^2, \quad \textit{where } C_0 = 2\max\{M_\varepsilon/\sqrt{\ln 2}, 2\sigma\}, \tag{4.3}$$

κ is the constant in Condition 1 and $\hbar$ is a constant. Let the tuning parameters $(\tau^{(k)}, \lambda^{(k)})$, $1 \leq k \leq K$, satisfy conditions

$$\tau^{(k)} = \frac{A^{(k)} C_0 \varpi}{\|\boldsymbol{\alpha}_1\|_1 \sqrt{\mu_1(\boldsymbol{\Xi})}}, \qquad c^{-1} + \delta_0 < \lambda^{(k)} \leq 1, \tag{4.4}$$

where $A^{(k)}$ and δ_0 are positive constants such that $c^{-1} + \delta_0 < 1$, and c is the constant in Condition 1.

(a). For the first component ($k = 1$), there exist constants A_1^L and $\hbar_0$ which only depend on c, c_2 and δ_0, where c_2 is the constant in Condition 3(a), such that with probability at least $1 - 2e^{M_\varepsilon^2/2\sigma^2} p^{1 - C_0^2/4\sigma^2}$, if $A^{(1)} \geq A_1^L$ and $\hbar \geq \hbar_0$, we have

$$\begin{aligned} &\|\widehat{\boldsymbol{\alpha}}_1\|_1 \leq \sqrt{6c}\|\boldsymbol{\alpha}_1\|_1 \leq \sqrt{6cs}/\kappa, \\ &\|\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1\|_1 \leq 4(1+c)(1+\sqrt{6c})c_2^{-1} A^{(1)} C_0 \kappa^{-2} \mu_1(\boldsymbol{\Xi})^{-1/2} \varpi s, \\ &\frac{1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1)\|_2^2 \leq 16 c_2^{-2}(1+\sqrt{6c})^2 (A^{(1)} C_0/\kappa)^2 \mu_1(\boldsymbol{\Xi})^{-1} \varpi^2 s. \end{aligned} \tag{4.5}$$

(b). For the higher order components ($1 < k \leq K$), we further assume that

$$\begin{aligned} \max_{1 \leq k \leq K} \|\boldsymbol{\alpha}_k\|_1 &\leq c_4 \min_{1 \leq k \leq K} \|\boldsymbol{\alpha}_k\|_1, \\ \kappa^{-2} \mu_1(\boldsymbol{\Xi})^{-1/2} C_0 \varpi s &\leq c_5 \|\boldsymbol{\alpha}_1\|_1, \end{aligned} \tag{4.6}$$

where c_4 and c_5 are two constants. Then there exist constants $\hbar_0$, $A_j^L < A_j^U$, $1 \leq j \leq K$, which only depend on δ_0, c, $c_2 \sim c_5$, such that with probability at least $1 - 2e^{M_\varepsilon^2/2\sigma^2} p^{1 - C_0^2/4\sigma^2}$, for any $1 \leq k \leq K$, if $A_j^L \leq A^{(j)} \leq A_j^U$, $1 \leq j < k$, $A^{(k)} \geq A_k^L$ and $\hbar \geq \hbar_0$, we have

$$\begin{aligned} &\|\widehat{\boldsymbol{\alpha}}_k\|_1 \leq D_{k,1}\sqrt{s}/\kappa, \\ &\|\widehat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k\|_1 \leq D_{k,4} A^{(k)} C_0 \kappa^{-2} \mu_1(\boldsymbol{\Xi})^{-1/2} \varpi s, \\ &\frac{1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k)\|_2^2 \leq D_{k,2} (A^{(k)} C_0/\kappa)^2 \mu_1(\boldsymbol{\Xi})^{-1} \varpi^2 s\ , \end{aligned} \tag{4.7}$$

where $D_{k,1}$, $D_{k,2}$ and $D_{k,4}$ are constants only depending on δ_0, c, $c_2 \sim c_5$.

Although we have two tuning parameters, $\tau^{(k)}$ and $\lambda^{(k)}$, for each k, by Theorem 4.2, $\lambda^{(k)}$ is not essential for the convergence rates. Actually, it can be any number in a subinterval of $(c^{-1}, 1]$ and does not affect the convergence rates.

Next, based on Theorem 4.2, we provide the oracle inequalities for $\widehat{\mathbf{W}}(t)$, $\widehat{\boldsymbol{\beta}}_k(t)$, and $\mathbf{X}\widehat{\boldsymbol{\beta}}_k(t)$, $1 \le k \le K$. Bickel *et al.* [3] provided the oracle inequality for the coefficient vector under the l_1-norm. We extend the l_1 norm of a usual vector to a vector of functions. For any $\mathbf{M}(t) = (m_1(t), \cdots, m_p(t))^{\mathrm{T}}$, we define the $L_{1,2}$-norm: $\|\mathbf{M}\|_{1,2} = \left\{\int_0^1 \left(\sum_{i=1}^p |m_i(t)|\right)^2 dt\right\}^{1/2}$. The $L_{1,2}$-norm is stronger than the L_2-norm, that is, $\|\mathbf{M}\|_{1,2} \ge \|\mathbf{M}\|_{L^2}$. We will provide the oracle inequalities for $\widehat{\boldsymbol{\beta}}_k(t)$ and $\widehat{\boldsymbol{\beta}}(t)$ based on the $L_{1,2}$-norm.

Theorem 4.3. *Suppose that all the conditions in Theorem 4.2 hold and* $0 \le \eta \le C_0 C_\beta^{-1} s^{-1}$, *where* $C_\beta = \max_{1\le q\le Q} \int_0^1 \int_0^1 \left(\frac{\partial^2 \beta_q(t,s)}{\partial t^2}\right)^2 dsdt$ *and* C_0 *is defined in (4.3) in Theorem 4.2. Then with probability at least* $1 - 2e^{M_\varepsilon^2/2\sigma^2} p^{1-C_0^2/4\sigma^2}$, *for any* $1 \le K_0 \le K$, *if* $A_k^L \le A^{(k)} \le A_k^U$, $1 \le k < K_0$, *and* $A^{(K_0)} \ge A_{K_0}^L$, *we have*

$$\begin{aligned}
&\|\widehat{w}_k(t) - w_k(t)\|_{L^2} \le L_{k,1} C_0 \varpi \sqrt{s} \kappa^{-1}, 1 \le k \le K_0,\\
&\|\widehat{\boldsymbol{\beta}}_{K_0}(t) - \boldsymbol{\beta}_{K_0}(t)\|_{1,2} \le L_{K_0,3} C_0 \kappa^{-2} \varpi s,\\
&\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{K_0}(t) - \mathbf{X}\boldsymbol{\beta}_{K_0}(t)\|_{L^2} \le \sqrt{n} L_4 C_0 \kappa^{-1} \varpi \sqrt{s},
\end{aligned}$$

where $L_{k,1}$, $L_{k,3}$ *and* $L_{k,4}$ *are constants only depending on* $A^{(j)}$, $1 \le j \le k$, c, $\hbar$ *and* $c_2 \sim c_5$. *In particular, when* $K_0 = K$, *we have*

$$\|\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)\|_{1,2} \le L_{K,3} C_0 \kappa^{-2} \varpi s, \quad \|\mathbf{X}\widehat{\boldsymbol{\beta}}(t) - \mathbf{X}\boldsymbol{\beta}(t)\|_{L^2} \le \sqrt{n} L_{K,4} C_0 \kappa^{-1} \varpi \sqrt{s}.$$

Therefore, for any $K_0 \le K$, $\sum_{k=1}^{K_0} \mathbf{X}\widehat{\boldsymbol{\alpha}}_k \widehat{w}_k(t)$ is an estimate of the best K_0 dimensional approximation to $\mathbf{X}\boldsymbol{\beta}(t)$. The upper bounds of $\|\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)\|_{1,2}$ and $\|\mathbf{X}\widehat{\boldsymbol{\beta}}(t) - \mathbf{X}\boldsymbol{\beta}(t)\|_{L^2}$ in Theorem 4.3 are the same as those for the Lasso and the Dantzig selector [3] except the constants.

5. Simulation studies

We study the performance of our method for the case that there is only one predictive curve in Section 5.1 and the case with multiple predictive curves in Section 5.2. In all simulation studies, the predictive functions are defined in $0 \le s \le 2$ with 128 equally spaced discrete observation points, and the response functions are defined in $0 \le t \le 1$ and there are 60 equally spaced observation points.

5.1. Simulation 1: One predictive curve

In this section, we consider the model with one predictive curve and compare our method (*wSigComp*) with the functional linear regression via the Principal

Analysis by Conditional Estimation (PACE) algorithm (*PACE-reg*) by Yao *et al.* [54], and the penalized function-on-function regression (*pffr*) by Ivanescu *et al.* [18] and (*pffr.pc*) by Scheipl *et al.* [46]. Our method is implemented in R and use 40 cubic B-spline basis as the basis functions $\mathbf{\Phi}(t)$ ((3.16) and (3.17) in section 3.3) for both $\mu(t)$ and $\widehat{w}_k(t)$. Both *pffr* and *pffr.pc* are implemented in the R package "refund" (Crainiceanu *et al.* [7]). We use the default settings except that we use 40 basis functions for both of them. The *PACE-reg* is downloaded from http://www.stat.ucdavis.edu/PACE/. It is implemented in matlab (The MathWorks [48]) and we use the default setting. We also consider the method *linmod* for functional linear regression by Ramsay and Silverman [39], which is implemented in the R package "fda" (J. O. Ramsay and Hooker [19]). We use 40 cubic B-spline basis for both s and t, and choose both smoothness parameters from $\{10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$. As numerical problems frequently occurred due to the singularity of some matrices, we do not provide results from *linmod* in this simulation, but will consider *linmod* in Section 6.1 for a real data set.

We consider two settings similar to those in Ivanescu *et al.* [18]. In supplementary materials on our webpage, we provide additional simulations for the model with one predictive curve which is generated from Gaussian processes. In Setting 1,

$$\beta(t,s) = \cos(2\pi t)\sin(\pi s), \quad z(s) = \sum_{m=1}^{10} \frac{1}{m^2}\left\{\varsigma_{1,m}\sin(m\pi s) + \varsigma_{2,m}\cos(m\pi s)\right\},$$

for $0 \le s \le 2$ and $0 \le t \le 1$, where $\varsigma_{j,m}$, $1 \le j \le 2$ and $1 \le m \le 10$, are independent standard normal random variables. In Setting 2,

$$\beta(t,s) = \sqrt{ts}/1.2, \quad z(s) = \sum_{m=1}^{40} \frac{2\sqrt{2}}{\pi m}\varrho_m \sin(m\pi s),$$

for $0 \le s \le 2$ and $0 \le t \le 1$, where ϱ_m, $1 \le m \le 40$, are independent standard normal random variables. In both settings, the intercept function is $\beta_0(t) = 2e^{-(t-1)^2}$, and the noise $\varepsilon(t)$ is generated from the Gaussian process with covariance function $\mathbf{\Sigma}_\varepsilon(t,t') = \sigma^2\rho^{\{10|t-t'|\}^2}$. We fix $\sigma^2 = 0.1$ and choose the within-function correlation $\rho = 0$ or 0.7. When $\rho = 0$, $\varepsilon(t)$ is Gaussian white noise. When ρ is bigger, the within-function correlation in $\varepsilon(t)$ is stronger and the sample noise curve is smoother. We consider all the 4 combinations of two types of $(z(s), \beta(t,s))$ and two values of ρ. For each combination, we repeat the following procedure 50 times. In each repeat, we generate 100 discretely observed random samples $\{z_i(s_k), \varepsilon_i(t_\ell) | 1 \le i \le 100, 1 \le k \le 128, 1 \le \ell \le 60\}$, where $\{s_k, 1 \le k \le 128\}$ is the set of equally spaced observation points in $[0,2]$ and $\{t_\ell, 1 \le \ell \le 60\}$ is the set of equally spaced observation points in $[0,1]$. Then we calculate $y_i(t_\ell)$ based on the model (1.1) and use $\{z_i(s_k), y_i(t_\ell) | 1 \le i \le 100, 1 \le k \le 128, 1 \le \ell \le 60\}$ as the training data. Similarly the test data set $\{z_i^{\text{test}}(s_k), y_i^{\text{test}}(t_\ell) | 1 \le i \le 500, 1 \le k \le 128, 1 \le \ell \le 60\}$ is generated with size of 500. We use the training data to choose the tuning parameters and fit the model for each method. Then the final model is applied to the test data to calculate the predicted curves $\widehat{y}_i^{\text{predict}}(t)$, $1 \leq i \leq 500$. We calculate the mean squared prediction error for this repeat by

$$MSPE = \frac{1}{500}\sum_{i=1}^{500}\left\{\frac{1}{60}\sum_{\ell=1}^{60}\left(\widehat{y}_i^{\text{predict}}(t_\ell) - y_i^{\text{test}}(t_\ell)\right)^2\right\}. \tag{5.1}$$

We report the averages and standard deviations of the MSPEs of 50 replicates in Table 1. In general, the *wSigComp* method has the smallest prediction error. A stronger within-function correlation in $\varepsilon(t)$ tends to increase the prediction error of all methods. The *wSigComp* chooses 2 components in all cases, and the *PACE-reg* chooses about 7 components for $z(s)$ and 1 component for $y(t)$. The *pffr.pc* chooses about 8 and 15 components on average for the two settings, respectively. The averages and standard deviations of the running time for one repeat over 50 replicates are also provided in Table 1, which shows that the *wSigComp* method is very computational efficient.

TABLE 1
The averages (and standard deviations) of MSPEs and running time over 50 replicates for the simulation 1 in Section 5.1

$z(s), \beta(t,s)$	ρ	*wSigComp*	*pffr.pc*	*pffr*	*PACE-reg*
Prediction Error (MSPE)					
Setting 1	0	0.101(0.001)	0.110(0.006)	0.107(0.001)	0.502(0.201)
	0.7	0.106(0.003)	0.117(0.008)	0.124(0.005)	0.513(0.202)
Setting 2	0	0.104(0.001)	0.129(0.011)	0.103(0.001)	0.313(0.070)
	0.7	0.111(0.005)	0.137(0.010)	0.121(0.006)	0.307(0.078)
Running time in seconds					
Setting 1	0	3.9(0.5)	87.9(27.2)	6567.0(5136.9)	400.1(111.6)
	0.7	4.1(0.4)	106.3(37.7)	6130.3(1390.0)	454.6(222.3)
Setting 2	0	3.7(0.3)	734.0(173.3)	3582.3(1446.9)	646.5(151.2)
	0.7	4.0(0.4)	684.0(124.0)	3160.7(1722.8)	508.2(199.3)

5.2. *Multiple predictive curves*

We consider two sets of simulation studies with multiple predictive curves. In both simulations, we generate the predictive curves from Gaussian processes and study the effects of the correlation between multiple predictive curves and their smoothness on the predictive performance. We consider three types of Gaussian processes with different smoothness levels. Their covariance functions are given by

$$\begin{aligned}&\boldsymbol{\Sigma}_1(s,s') = e^{-\{10|s-s'|\}^2}, \quad \boldsymbol{\Sigma}_2(s,s') = \left\{1 + 20|s-s'| + \frac{20}{3}(s-s')^2\right\}e^{-20|s-s'|},\\ &\boldsymbol{\Sigma}_3(s,s') = e^{-\{10|s-s'|\}^{1.5}}.\end{aligned} \tag{5.2}$$

The first one in (5.2) is the squared exponential covariance function and the corresponding Gaussian process has mean square derivatives of all orders (Chapter 4 in Rasmussen and Williams [40]). The second one belongs to the Matérn class and the corresponding Gaussian process has the second order mean square derivative. The last one is the γ-exponential covariance function with $\gamma = 1.5$ and the Gaussian process is mean square continuous but not mean square differentiable. We plot sample curves for each of the three Gaussian processes in Figure 1.

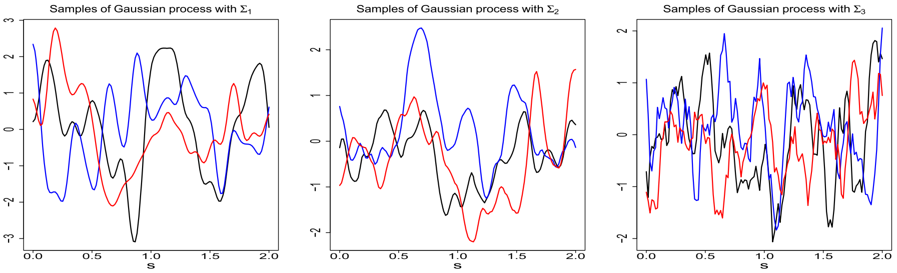

Fig 1. *Three sample curves from the Gaussian processes generated with the covariance function* $\mathbf{\Sigma}_1$*,* $\mathbf{\Sigma}_2$ *and* $\mathbf{\Sigma}_3$ *in* (5.2)*, respectively.*

In both simulations, the predictive curves $z_1(s), \cdots, z_Q(s)$ are generated from the Gaussian process with one of the covariance functions in (5.2). We model the correlation between curves $z_1(s), \cdots, z_Q(s)$ in the following way. Let $\mathbf{S}$ be the $Q \times Q$ matrix with the (i, j)-th entry equal to ρ_{curve} if $i \neq j$ and 1 if $i = j$, where $0 \leq \rho_{\text{curve}} \leq 1$ controls the correlation between the predictive curves. We decompose $\mathbf{S} = \mathbf{\Delta}\mathbf{\Delta}^{\mathrm{T}}$, where $\mathbf{\Delta}$ is a $Q \times Q$ matrix. Given one of the covariance functions in (5.2), we generate Q independent curves $u_1(s), \cdots, u_Q(s)$ from the corresponding Gaussian process. Let

$$(z_1(s), z_2(s), \cdots, z_Q(s)) = (u_1(s), u_2(s), \cdots, u_Q(s))\mathbf{\Delta}^{\mathrm{T}}. \tag{5.3}$$

Then each of $z_1(s), z_2(s), \cdots, z_Q(s)$ is a Gaussian process with the same covariance function as $u_i(s)$ and given any $0 \leq s \leq 2$, $(z_1(s), z_2(s), \cdots, z_Q(s))$ is a Q-dimensional normal random vector with covariance matrix $\mathbf{S}$. Therefore, when $\rho_{\text{curve}} = 0$, $z_1(s), z_2(s), \cdots, z_Q(s)$ are independent and when ρ_{curve} is large, strong correlations exist among $z_1(s), z_2(s), \cdots, z_Q(s)$. In Figure 2, we plot one sample from $(z_1(s), z_2(s), \cdots, z_Q(s))$ for $\rho_{\text{curve}} = 0$ and 0.7, respectively, with $Q = 4$ and the covariance function $\mathbf{\Sigma}_1$. When $\rho_{\text{curve}} = 0.7$, strong positive correlations exist among $z_1(s), z_2(s), \cdots, z_Q(s)$ and the sample curves show similar trends.

For the coefficient surface functions, in one case, we specify the explicit expression for $\beta_q(t, s)$, $1 \leq q \leq Q$, with Q fixed. In another case, we evaluate our methods for different number of predictive curves Q and various $\beta_q(t, s)$ with different roughness levels.

5.2.1. Simulation 2:

In this simulation, we take $Q = 4$, $\beta_0(t) = 0$ and

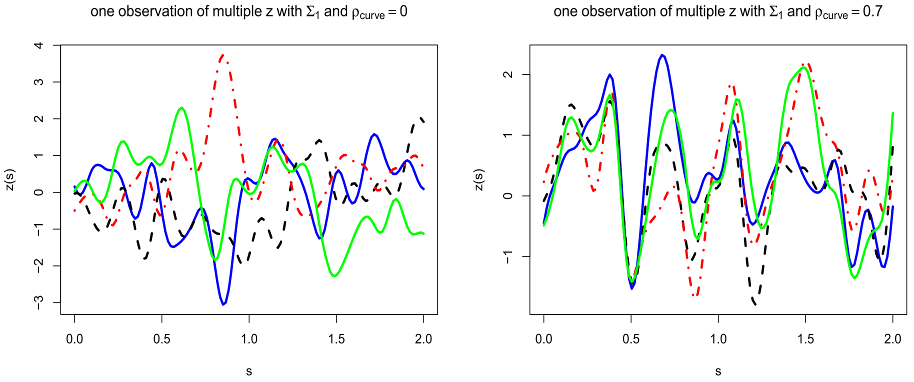

FIG 2. *One sample of* $(z_1(s), z_2(s), \cdots, z_4(s))$ *defined in* (5.3) *for* $\rho_{\text{curve}} = 0$ *or* 0.7*, respectively, with covariance function* $\mathbf{\Sigma}_1$*.*

$$\beta_1(t,s) = \{10(t-0.5)(1-s)\}^2, \quad \beta_2(t,s) = 20e^{-\{5(t-0.5)^2+3(s-1)^2\}},$$
$$\beta_3(t,s) = 15e^{-\{5(t-0.5)^2+5(s-0.5)^2\}} + 20e^{-\{5(t-0.5)^2+5(s-1.5)^2\}},$$
$$\beta_4(t,s) = 10\sin \pi t \sin(3\pi s/2),$$

which are plotted in Figure 3. $(z_1(s), z_2(s), \cdots, z_4(s))$ is generated by (5.3) with $\rho_{\text{curve}} = 0$ and 0.7, respectively. The noise $\varepsilon(t)$ is generated in the same way as in Section 5.1. We consider two noise levels $\sigma^2 = 0.1, 0.25$, and fix $\rho = 0$. The averages and standard deviations of MSPEs of 50 repeats for our method are listed in Table 2. The averages and standard deviations of the number of components chosen by our method are given in Table 3.

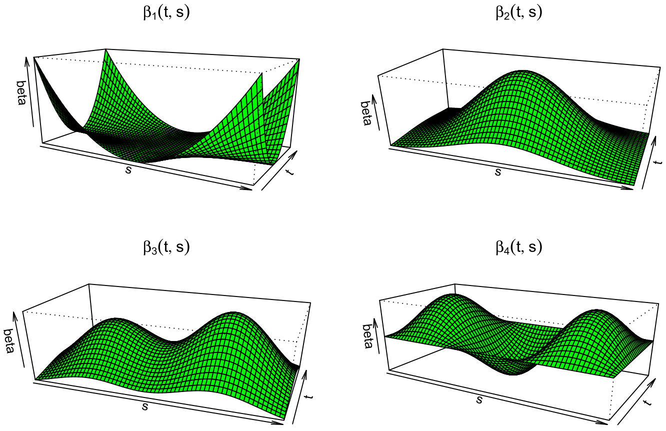

FIG 3. *The plots of* $\beta_1(t,s)$*,* $\beta_2(t,s)$*,* $\beta_3(t,s)$ *and* $\beta_4(t,s)$ *in Simulation 2.*

Although there are four $\beta_q(t,s)$, our method chooses 3 components in most of the repeats and chooses 2 components in the others. As an example, we choose one repeat with $\rho_{\text{curve}} = 0.7$, $\sigma^2 = 0.25$ and $z_q(s)$ generated using $\mathbf{\Sigma}_1$ to plot

TABLE 2
The averages and standard deviations (in parenthesis) of the MSPEs of 50 repeats for our method in Simulation 2.

Covariance of $x_i(s)$	$\sigma^2 = 0.1$		$\sigma^2 = 0.25$	
	$\rho_{\text{curve}} = 0$	$\rho_{\text{curve}} = 0.7$	$\rho_{\text{curve}} = 0$	$\rho_{\text{curve}} = 0.7$
$\mathbf{\Sigma}_1$	0.120(0.007)	0.129(0.013)	0.283(0.009)	0.290(0.020)
$\mathbf{\Sigma}_2$	0.121(0.006)	0.129(0.010)	0.277(0.006)	0.294(0.029)
$\mathbf{\Sigma}_3$	0.121(0.003)	0.137(0.025)	0.281(0.008)	0.292(0.020)

TABLE 3
The averages and standard deviations (in parenthesis) of the numbers of components of 50 repeats for our method in Simulation 2.

Covariance of $z_q(s)$	$\sigma^2 = 0.1$		$\sigma^2 = 0.25$	
	$\rho_{\text{curve}} = 0$	$\rho_{\text{curve}} = 0.7$	$\rho_{\text{curve}} = 0$	$\rho_{\text{curve}} = 0.7$
$\mathbf{\Sigma}_1$	3(0)	2.98(0.14)	3(0)	2.96(0.20)
$\mathbf{\Sigma}_2$	3(0)	3(0)	3(0)	2.88(0.33)
$\mathbf{\Sigma}_3$	3(0)	2.90(0.30)	3(0)	2.96(0.20)

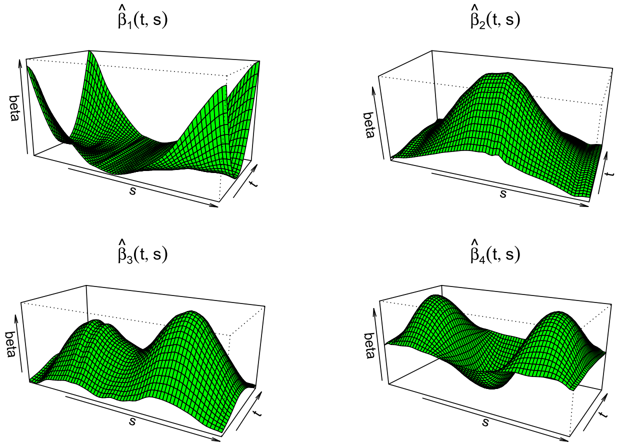

FIG 4. *The estimates $\widehat{\beta}_q(t,s)$, $1 \le q \le 4$, in one repeat for $\rho_{\text{curve}} = 0.7$, $\sigma^2 = 0.25$ and $z_q(s)$ with covariance function $\mathbf{\Sigma}_1$ in Simulation 2.*

the estimates $\widehat{\beta}_j(t,s)$, $1 \le j \le 4$, in Figure 4, and $\widehat{w}_j(t)$, $1 \le j \le 3$ in Figure 5, where three components were chosen. The shape of $\widehat{w}_1(t)$ implies that the most important variation in the signal function is in the middle of the interval $[0,1]$. $\widehat{w}_2(t)$ reflects the variations of the contrast between the values of the signal functions in the middle and those in the two ends. The third one only account for very small proportion of the variations in the signal function.

5.2.2. Simulation 3:

In previous simulations, Q is fixed and the coefficient surfaces $\beta_q(t,s)$, $1 \le q \le Q$, have explicit expressions. In this section, we will consider different

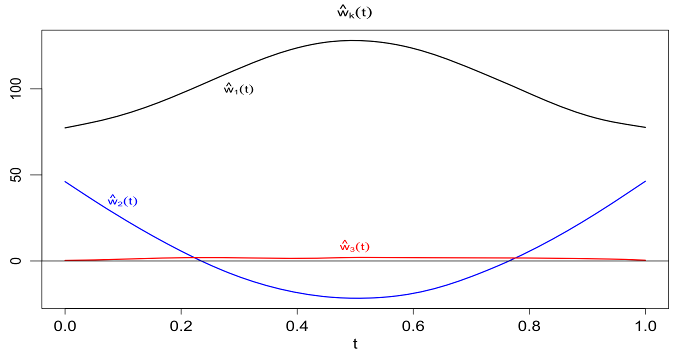

FIG 5. *The estimates $\widehat{w}_k(t)$, $1 \leq k \leq 3$, in one repeat for $\rho_{\text{curve}} = 0.7$, $\sigma^2 = 0.25$ and $z_q(s)$ with covariance function $\mathbf{\Sigma}_1$ in Simulation 2.*

number of predictive curves Q and test our method for various $\beta_q(t, s)$ with different roughness. We note that any coefficient surface function $\beta(s, t)$ can be expanded of form $\beta(t, s) = \sum_{i=1}^{\infty} \zeta_i(t)\xi_i(s)$, where the terms in the expansion can be finite or infinite. We will first randomly generate a finite number of pairs $\{(\zeta_i(t), \xi_i(s)), 1 \leq i \leq k\}$ and obtain a coefficient surface which is equal to $\sum \zeta_i(t)\xi_i(s)$. In this way, we can obtain various coefficient functions and control their roughness by controlling the roughness of $\{(\zeta_i(t), \xi_i(s))\}$.

Specifically, let

$$\beta_q(t, s) = \{\zeta_{1q}(t)\xi_{1q}(s) + \zeta_{2q}(t)\xi_{2q}(s) + \zeta_{3q}(t)\xi_{3q}(s)\}/q^2, \quad 1 \leq q \leq Q, \quad (5.4)$$

where $\zeta_{jq}(t)$ $(1 \leq j \leq 3,\ 1 \leq q \leq Q)$ and $\xi_{jq}(t)$ $(1 \leq j \leq 3,\ 1 \leq q \leq Q)$ are independently generated from the Gaussian process with the same covariance function. We consider two covariance functions: $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_3$ in (5.2). We plot examples of the coefficient surface functions generated in (5.4) using $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_3$, respectively, in Figure 6. We consider $Q = 1, 5, 10$, and 30. $(z_1(s), z_2(s), \cdots, z_Q(s))$ is generated from (5.3) with $\rho_{\text{curve}} = 0$ or 0.7. The noise $\varepsilon(t)$ is generated in the same way as in Simulation 1 and we fix $\sigma^2 = 0.1$ and $\rho = 0$. We list the averages and standard deviations of MSPEs and the number of selected components in Tables 4 and 5, respectively. The MSPE increases as Q increases because the model becomes more complicated. In this simulation, strong correlation between multiple functional predictors helps the predictive accuracy, especially for a large Q. Generally, the prediction errors for a noisier $\beta_q(t, s)$ are larger. But when $Q = 30$, the difference in the prediction errors for the two types of coefficient surface functions is not significant. We also observe that the number of the selected components does not always increase as Q increases. Instead, the average number reaches the maximum at $Q = 5$ or $Q = 10$ and then is almost unchanged or slightly decreases with further increase of Q. A possible explanation comes from the trade-off between bias and variance. Selecting more components reduces bias but increases variance in prediction. Given the number Q of predictive curves, there is no great difference in the running time from different settings. When $\rho_{\text{curve}} = 0$, and $z_q(s)$ and $\beta_q(t, s)$ are both generated by $\mathbf{\Sigma}_3$, the

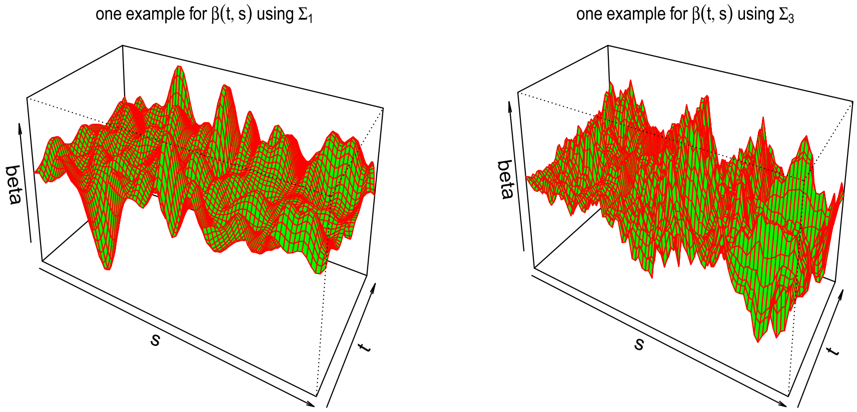

FIG 6. *Examples of $\beta(t, s)$ generated in (5.4) using $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_3$, respectively, in Simulation 3.*

TABLE 4
The averages and standard deviations (in parenthesis) of the MSPEs of 50 repeats for our method in Simulation 3.

$\beta_q(t, s)$'s are generated by $\mathbf{\Sigma}_1$					
$z_q(s)$ by	ρ_{curve}	$Q = 1$	$Q = 5$	$Q = 10$	$Q = 30$
$\mathbf{\Sigma}_1$	0	0.103 (0.001)	0.137 (0.007)	0.167 (0.018)	0.239 (0.053)
	0.7	0.103 (0.001)	0.129 (0.008)	0.145 (0.009)	0.178 (0.023)
$\mathbf{\Sigma}_2$	0	0.103 (0.001)	0.133 (0.006)	0.160 (0.015)	0.234 (0.041)
	0.7	0.104 (0.001)	0.126 (0.005)	0.144 (0.010)	0.178 (0.020)
$\mathbf{\Sigma}_3$	0	0.104 (0.001)	0.137 (0.007)	0.165 (0.014)	0.244 (0.045)
	0.7	0.104 (0.001)	0.129 (0.005)	0.144 (0.009)	0.177 (0.023)

$\beta_q(t, s)$'s are generated by $\mathbf{\Sigma}_3$					
$z_q(s)$ by	ρ_{curve}	$Q = 1$	$Q = 5$	$Q = 10$	$Q = 30$
$\mathbf{\Sigma}_1$	0	0.109 (0.002)	0.143 (0.005)	0.171 (0.015)	0.231 (0.031)
	0.7	0.110 (0.002)	0.137 (0.006)	0.151 (0.009)	0.182 (0.020)
$\mathbf{\Sigma}_2$	0	0.110 (0.002)	0.142 (0.007)	0.165 (0.015)	0.230 (0.037)
	0.7	0.111 (0.003)	0.136 (0.006)	0.151 (0.009)	0.178 (0.017)
$\mathbf{\Sigma}_3$	0	0.110 (0.002)	0.144 (0.006)	0.169 (0.014)	0.227 (0.030)
	0.7	0.111 (0.002)	0.137 (0.005)	0.153 (0.010)	0.184 (0.020)

averages and standard deviations (in parenthesis) of the running time (including the cross-validation procedure) in seconds over 50 repeats for $Q = 1, 5, 10, 30$ are 121.294 (25.563), 235.129 (85.829), 297.371 (84.89) and 498.801 (113.91), respectively.

6. Application to real datasets

6.1. Diffusion tensor imaging data

In the human brain, white matter tracts consist of axons that connect nerve cells and transmit information via electrical nerve impulses. Axons are surrounded by

TABLE 5

The averages and standard deviations (in parenthesis) of the number of components selected by our method in 50 repeats in Simulation 3.

$\beta_q(t,s)$'s are generated by $\boldsymbol{\Sigma}_1$					
$z_q(s)$ by	ρ_{curve}	$Q=1$	$Q=5$	$Q=10$	$Q=30$
$\boldsymbol{\Sigma}_1$	0	3 (0)	4.756 (0.799)	4.891 (0.654)	4.491 (0.848)
	0.7	3 (0)	4.757 (0.76)	4.689 (0.612)	4.649 (0.668)
$\boldsymbol{\Sigma}_2$	0	3 (0)	4.889 (0.622)	4.793 (0.734)	4.722 (0.763)
	0.7	3 (0)	4.794 (0.729)	4.747 (0.614)	4.712 (0.536)
$\boldsymbol{\Sigma}_3$	0	3 (0)	4.812 (0.693)	4.791 (0.721)	4.564 (0.958)
	0.7	3 (0)	4.767 (0.568)	4.674 (0.694)	4.865 (0.715)

$\beta_q(t,s)$'s are generated by $\boldsymbol{\Sigma}_3$					
$z_q(s)$ by	ρ_{curve}	$Q=1$	$Q=5$	$Q=10$	$Q=30$
$\boldsymbol{\Sigma}_1$	0	3 (0)	4.769 (0.731)	4.985 (0.668)	4.921 (0.867)
	0.7	3 (0)	4.725 (0.723)	4.661 (0.71)	4.745 (0.645)
$\boldsymbol{\Sigma}_2$	0	3 (0)	4.8 (0.67)	4.83 (0.778)	4.627 (0.72)
	0.7	3 (0)	4.653 (0.561)	4.569 (0.64)	4.745 (0.688)
$\boldsymbol{\Sigma}_3$	0	3 (0)	4.926 (0.723)	4.939 (0.747)	4.562 (0.796)
	0.7	3 (0)	4.824 (0.623)	4.604 (0.61)	4.771 (0.722)

a white fatty insulation called myelin, which increases the speed of transmission of nerve signals. Changes in water diffusion in the brain could potentially be associated with demyelination, a disease of the nervous system in which the myelin sheath of neurons is damaged. Diffusion tensor imaging (DTI) tractography [44] is a magnetic resonance imaging technique that studies white-matter tracts by measuring the diffusivity of water in the brain: in white-matter tracts, water diffuses anisotropically (perfectly organized and synchronized movement of all water molecules in one direction) in the direction of the tract, while elsewhere water diffuses isotropically (Brownian motion). One of the diffusion measures is fractional anisotropy (FA) which takes values between zero and one. A value of zero means that diffusion is isotropic. A value of one means that diffusion occurs only along the direction of the white-matter tracts and is fully restricted along all other directions. Tievsky *et al.* [50], Song *et al.* [47], Ivanescu *et al.* [18] have used FA as a proxy variable for demyelination of the white matter tracts, and assume larger FA values are closely associated with less demyelination and fewer lesions.

We use the DTI data in the R package "refund", which consists of FA tract profiles for the corpus callosum (CCA) and the right corticospinal tract (RCTS) for 142 individuals at one or multiple visits. The individuals are either multiple sclerosis (MS) cases or controls (MS is a demyelinating autoimmune-mediated disease). Each profile contains two curves: the FA values along the CCA tract and the RCTS tract from one individual in a visit. We will call them CCA curve and RCTS curve, respectively. Using a similar data set, Goldsmith *et al.* [12] predicted multiple sclerosis cases and controls based on functional predictors, and Ivanescu *et al.* [18] built function-on-function regression models to study the spatial associations between functional predictors and responses. We consider

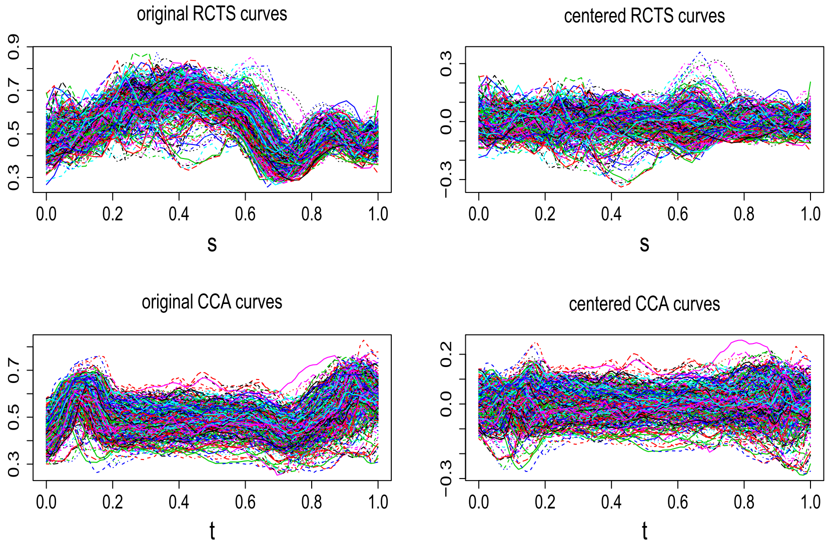

Fig 7. *Original and centered sample curves for RCTS ($z(s)$) and CCA ($y(t)$).*

a function-on-function regression model with CCA curve as the response $y(t)$ and RCTS curve as the predictor $z(s)$. After removing the first 12 observation points with a large number of missing values for RCST and then removing 6 pairs of curves with missing values, we obtain 376 sample curves for each of the two functional variables. There are 93 observation points for $y(t)$ and 43 points for $z(s)$. For simplicity, we use the relative distances, the position along the tract divided by the length of the tract, so that $y(t)$ and $z(s)$ are both defined in $[0, 1]$. Figure 7 displays the original and centered sample curves for RCTS and CCA.

To evaluate the predictive performance of our method, *PACE-reg*, *linmod*, *pffr* and *pffr.pc* on this data set, we randomly choose 200 observations as the training data and the remaining as the test data. We repeat the following procedure 50 times. In each repeat, for *PACE-reg*, *linmod*, *pffr* and *pffr.pc*, we use exactly the same way as in the simulation study to choose tuning parameters and fit the model using the training set and calculate the MSPE based on the test set. For our method, since the number of observation points of the predictive curve $z(s)$ is not a power of two, we first make basis expansions for the sample curves $z_i(s)$, $1 \le i \le n$, using 40 cubic B-spline basis with equally spaced knots and the smoothing method with a roughness penalty in Chapter 5 of Ramsay and Silverman [39]. We choose the tuning parameter for this roughness penalty from the set $\{10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$ by minimizing the generalized cross-validation statistic (GCV) which is calculated using the smooth.basis function in the R package "fda" on the training data set. Then we use the selected tuning parameter and the function smooth.basis to calculate the basis expansions for all the sample curves $z_i(s)$, $1 \le i \le 376$. By the rule introduced in Section 2, we evaluate these basis expansions at $2^6 = 64$ equally spaced points and perform the

DWT to obtain 64-dimensional wavelet coefficient vectors $\mathbf{x}_i$, $1 \le i \le 376$, as our predictors, using the R package "wavethresh" and the default Daubechies least-asymmetric wavelet basis functions with filter number ten. Finally we choose the tuning parameters and fit the model in the same way as in the simulation studies for our method. The averages and the standard deviations of MSPEs over 50 repeats are: *wSigComp*: 0.00323 (0.00016); *PACE-reg*: 0.00329 (0.0002); *linmod*: 0.00370 (0.0002); *pffr*: 0.00323 (0.0002); *pffr.pc*: 0.00329 (0.0002). The averages and the standard deviations of the numbers of components chosen are: *wSigComp*: 4.15 (1.35); *PACE-reg*: 15.6 (1.62) (the PC for RCTS) and 4 (0) (the PC for CCA); *pffr.pc*: 15 (0). In summary, our method uses less components and achieves competitive prediction accuracy compared to other methods.

The most frequently chosen tuning parameters in the 50 repeats of our method are $(\tau, \lambda) = (0.01, 0.01)$, $\eta = 10^{-6}$ and $K = 5$. We fit a model using these tuning parameters and all the 376 observations in this data set to obtain the estimates $\widehat{\boldsymbol{\alpha}}_k$ and $\widehat{w}_k(t)$, $1 \le k \le 5$, where $\widehat{\boldsymbol{\alpha}}_k$ is a 64-dimensional vector. To obtain the estimate $\widehat{\beta}(t,s)$ of the coefficient surface $\beta(t,s)$, we use inverse DWT to transfer $\widehat{\boldsymbol{\alpha}}_k$ back to the function space and obtain a function $\widehat{\psi}_k(s)$ for any $1 \le k \le 5$. Then we have $\widehat{\beta}(t,s) = \widehat{\psi}_1(s)\widehat{w}_1(t) + \cdots + \widehat{\psi}_5(s)\widehat{w}_5(t)$. We plot $\widehat{\psi}_k(s)$, $\widehat{w}_k(t)$, and $\widehat{\beta}(t,s)$ in Figure 8. For $1 \le k \le 5$, $\widehat{w}_k(t)$ is the estimate of $w_k(t)$, the k-th eigenfunction of the covariance function of the signal function. These five components account for about 83%, 7%, 3%, 2%, 2% of the variations in the signal function, respectively. $\widehat{w}_1(t)$ is positive throughout the tract, and has two peaks at 0.17 and 0.91 (the relative distance along the CCA tract), implying that large variations in the signal part of CCA curves exist around these two locations. For any $z_i(s)$, the predicted curve $y_{\text{pred},i}(t)$ for $z_i(s)$ is equal to the mean response curve plus a linear combination of the five functions $\widehat{w}_1(t), \cdots, \widehat{w}_5(t)$ with the coefficient of $\widehat{w}_k(t)$ given by $\int_0^1 [z_i(s) - \bar{z}(s)]\widehat{\psi}_k(s)ds$. Since the first component is the most important and $\widehat{\psi}_1(s)$ has wider peaks around $s = 0.4$ and $s = 0.85$ (Figure 8), the predicted CCA curve will have relatively large values if the predictor RCTS curve $z_i(s)$ has relative large values around $s = 0.4$ and $s = 0.85$. This relationship is illustrated in Figure 9, where we plot 30 centered predictive RCTS curves with the corresponding centered predicted CCA curves. We draw the predicted CCA curves above the mean CCA curve in blue and those below the mean in red. One can see that the values of $z(s)$ in $0.25 \le s \le 0.45$ have great effects on the predicted CCA curves.

6.2. Daily air quality data

The *Air Quality* data, available in UCI Machine Learning Repository (Bache and Lichman [1]), were recorded by an array of five metal oxide chemical sensors embedded in an air quality chemical multisensor device located in a significantly polluted area, at road level, within an Italian city. This dataset contains the hourly averages of the concentration values of five different atmospheric pollutants in each day. The five pollutants are Nitrogen dioxide (NO_2), Carbon monoxide (CO), Non-methane hydrocarbons (NMHC), total Nitrogen Oxides

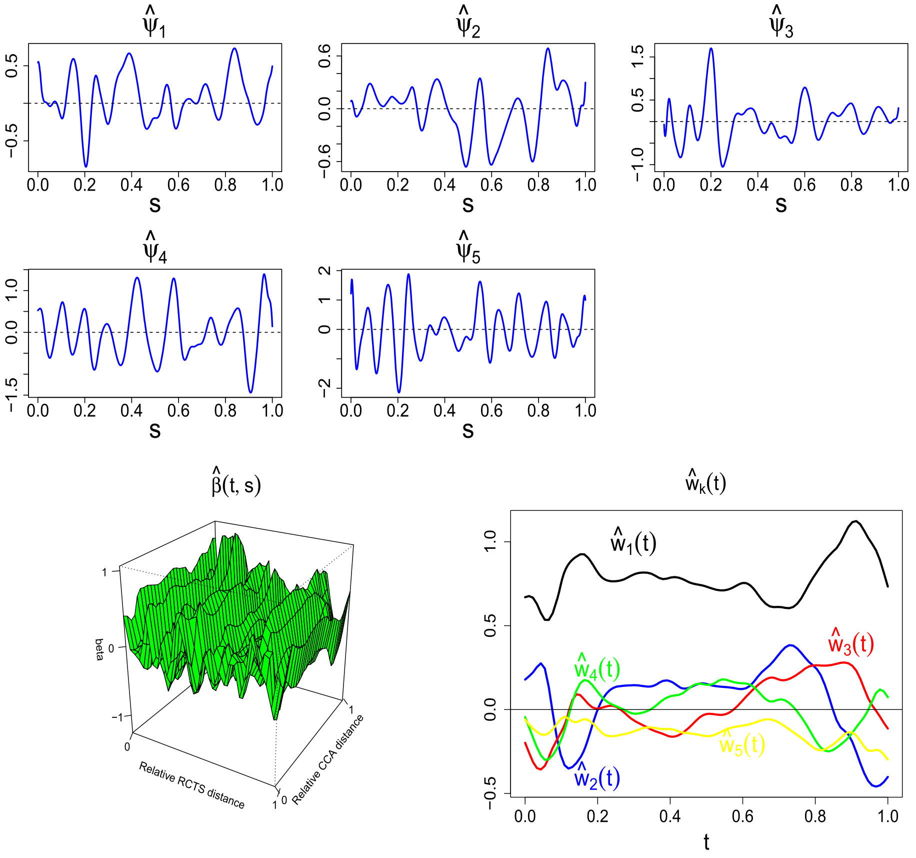

Fig 8. *The estimates: $\widehat{\psi}_k(s)$, $\widehat{\beta}(t, s)$ and $\widehat{w}_k(t)$, $1 \leq k \leq 5$, for the DTI data set.*

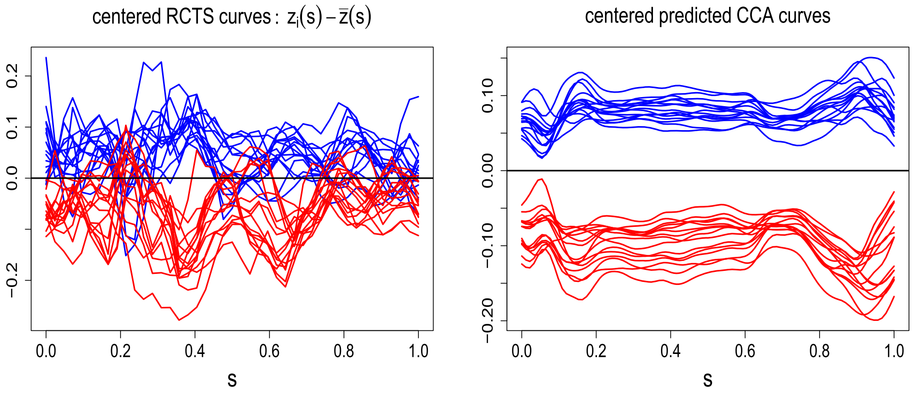

Fig 9. *Relationship between centered predictive RCTS curves and centered response CCA curves. The centered response CCA curves are shown in blue if they are above zero, and in red if they are below zero.*

(NO_x), and Benzene(C_6H_6). In addition, the temperature (in Celsius) and relative humidity (in Percentages) were also recorded hourly in each day. For the details of the experiment, we refer the reader to De Vito *et al.* [9]. We view the 24 hourly averaged concentration values of each pollutant in each day as a discretely observed curve. Together with the temperature and relative humidity, we have seven functional variables. With the removal of missing values, we obtained 355 sample curves for each of the seven functional variables. We plot all the sample curves in Figure 10. For convenience, we scale the 24 observation time points to the interval $[0, 1]$.

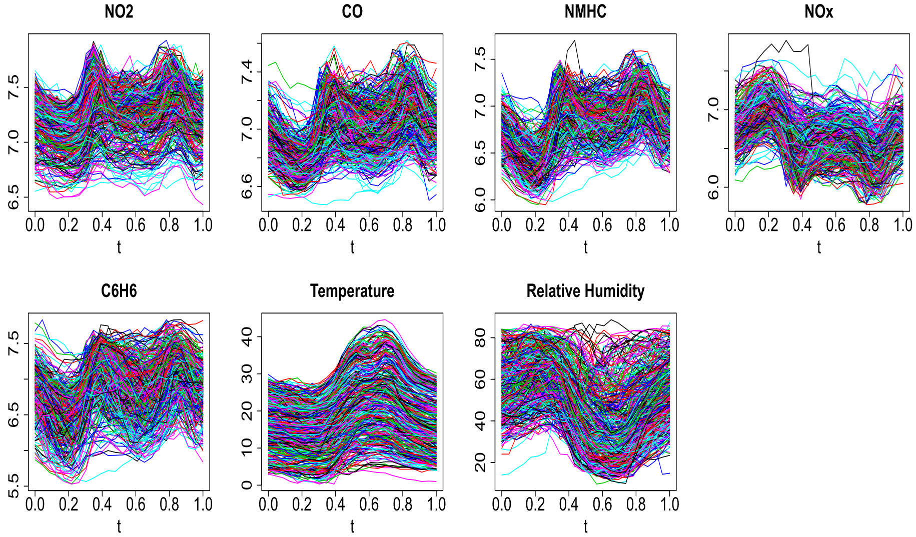

Fig 10. *The 355 sample curves for each of the seven variables in the air quality data.*

To study the relationship between the five pollutants, we investigate to what extent we can predict the daily curve of one pollutant by the other pollutants together with the temperature and relative humidity. We take the curve of NO_2 as the response and the other six curves as predictors. Since the number of observation points for all predictive curves is 24 which is not a power of two, we make basis expansions for the sample curves in the same way as in the analysis of the DTI data in Section 6.1. After obtaining the basis expansions, by the rule in Section 2, we evaluate each basis expansion at $2^5 = 32$ equally spaced points and perform the DWT to obtain a 32-dimensional wavelet coefficient vector for each predictor curve. Then we combine the six wavelet coefficient vectors for the six predictor curves into a 192-dimensional vector as our predictors. We randomly choose 200 observations as the training data, and take the other 155 observations as the test data. After obtaining the fitted model from the training data, we apply it to the test data and get the MSPE. In addition, for each repeat, we also calculate the average functional R^2 which has been used in Meyer *et al.* [30] and can be approximately calculated by

$$R^2_{\text{ave}} = \int_0^1 R^2(t)dt \approx \sum_{k=1}^{24} \left\{ 1 - \frac{\sum_{i=1}^{200} (Y_{\text{train},i}(t_k) - \widehat{Y}_i(t_k))^2}{\sum_{i=1}^{200} (Y_{\text{train},i}(t_k) - \bar{Y}_{\text{train}}(t_k))^2} \right\} \Big/ 24 \, ,$$

where $Y_{\text{train},i}(t)$ is the i-th response curve in the training set, $\widehat{Y}_i(t)$ is the corresponding predicted or fitted curve, $\bar{Y}_{\text{train}}(t)$ is the mean response curve in the training set and $0 = t_1 < t_2 < \cdots < t_{24} = 1$ are 24 equally spaced observation time points. We repeat this procedure 50 times. The average of the MSPEs in 50 iterations is 0.0089, with standard deviation of 0.0008. The average of the R^2_{ave} in 50 iterations is 88.0%, with standard deviation of 0.7%. The *wSigComp* selects 6 components in most repeats and select 5 or 7 components in other repeats.

Finally, we fit the model using all the 355 observations to obtain the estimates $\widehat{\boldsymbol{\alpha}}_{kq}$ and $\widehat{w}_k(t)$, where $\widehat{\boldsymbol{\alpha}}_{kq}$ is a 32-dimensional vector corresponding to the q-th functional predictor and the k-th component, $1 \leq k \leq 6$ and $1 \leq q \leq 6$. We apply the inverse DWT to $\widehat{\boldsymbol{\alpha}}_k$ and obtain $\widehat{\psi}_{kj}(s)$ and then $\widehat{\widehat{\beta}}_q(t,s) = \widehat{\widehat{\psi}}_{1q}(s)\widehat{w}_1(t) + \cdots + \widehat{\psi}_{6q}(s)\widehat{w}_6(t)$. We plot $\widehat{\psi}_{kq}(s)$'s and $\widehat{w}_k(t)$'s in Figure 11. These six components account for about 83%, 8.5%, 3%, 2%, 1.5%, 1% of the variations in the signal function, respectively. The first component accounts for most of the variations and is the most important. $\widehat{w}_1(t)$ has a peak around 0.33 corresponding to eight o'clock in the morning, which implies that the predicted daily pollution level of NO_2 has a large variation around eight o'clock in the morning. The R^2_{ave} for this fitted model is 94.7%.

7. Discussion

We consider the linear function-on-function regression models with multiple predictive curves. We first apply the wavelet transformation to the predictive curves and transform the original model to a linear model with functional response and high dimensional multivariate predictors. Based on the best finite dimensional approximation to the signal part in the response curve, we find an expansion of the vector of coefficient functions, which enjoys a good predictive property. For any k, the truncated expansion has nearly the smallest prediction error among all k-dimensional estimates. To estimate this expansion, we propose a penalized generalized eigenvalue problem followed by a penalized least squares problem. We provide the sparse oracle inequalities for our estimates in the high-dimensional settings. The choices of tuning parameters and the number of components are discussed. Simulation studies and applications to two real data sets demonstrate that our method has good predictive performance and is efficient in dimension reduction.

Appendix A: R code and supplementary material

The R code for *wSigComp* is available on http://sites.gsu.edu/rluo/software/. Supplementary material for additional proofs and simulations is available on http://sites.gsu.edu/rluo/publications/.

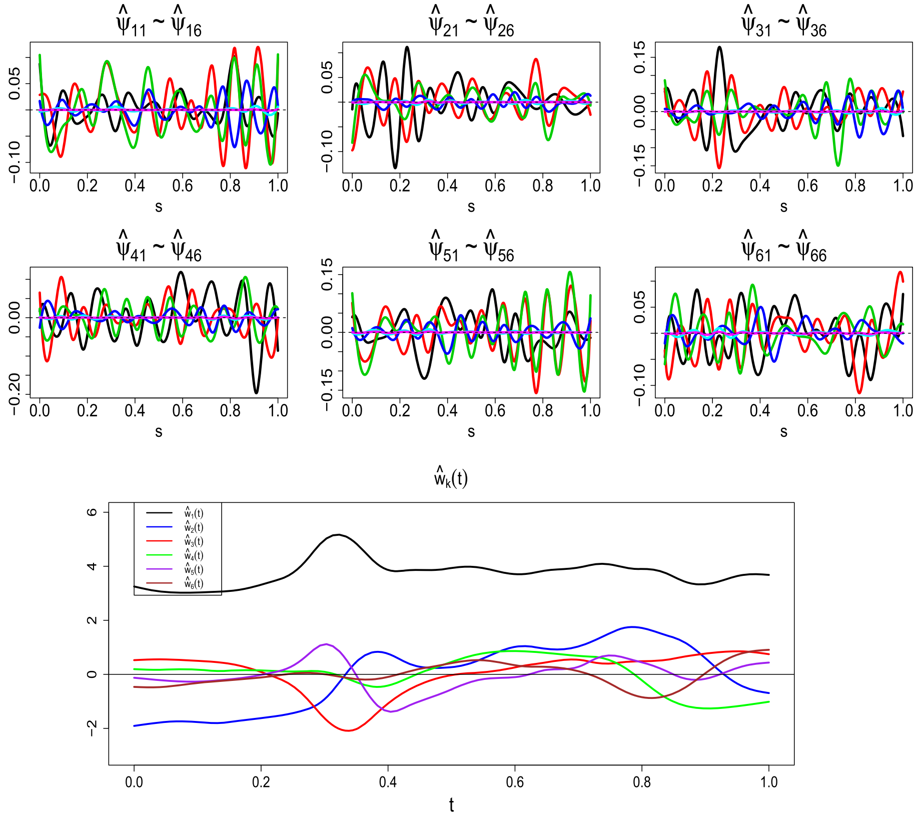

FIG 11. *The estimates $\widehat{\psi}_{kq}(s)$ and $\widehat{w}_k(t)$, $1 \leq k \leq 6$ and $1 \leq q \leq 6$, for the air quality data set.*

Appendix B: Proofs of theorems

Let $\boldsymbol{\varepsilon}(t) = (\varepsilon_1(t), \cdots, \varepsilon_n(t))^{\mathrm{T}}$ be the vector of the noise functions. Define

$$\bar{\varepsilon}(t) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i(t), \quad \boldsymbol{\varrho}(t) = \frac{1}{\sqrt{n}}\left[\boldsymbol{\varepsilon}(t) - \bar{\varepsilon}(t)\mathbf{1}_n\right], \tag{B.1}$$

where $\bar{\varepsilon}(t)$ is the mean noise function and $\boldsymbol{\varrho}(t)$ is an n-dimensional vector of functions.

B.1. Proof of Theorem 3.1

First, by the definition of $\boldsymbol{\Xi}$ in (3.6), its rank is less than the rank of $\mathbf{Z}$ which is an $n \times p$ matrix. Therefore, we have $K \leq \min\{n, p\}$.

Since $\boldsymbol{\gamma}_k$, $1 \leq k \leq K$, are the left-singular vectors of $\mathbf{X}\boldsymbol{\beta}(t)$, they are the first K eigenvectors of the matrix $\int_0^1 (\mathbf{X}\boldsymbol{\beta}(t))(\mathbf{X}\boldsymbol{\beta}(t))^{\mathrm{T}} dt$ with corresponding eigenvalue σ_k^2 (the squared singular value). By (3.6), $\int_0^1 (\mathbf{X}\boldsymbol{\beta}(t))(\mathbf{X}\boldsymbol{\beta}(t))^{\mathrm{T}} dt = n\boldsymbol{\Xi}$. Therefore, $\boldsymbol{\gamma}_k$, $1 \leq k \leq K$, are the first K eigenvectors of $\boldsymbol{\Xi}$ and we have $\sigma_k = \sqrt{n\mu_k(\boldsymbol{\Xi})}$. Similarly, the right-singular functions $u_k(t)$, $1 \leq k \leq K$, are the first K eigenfunctions of the covariance function $(\mathbf{X}\boldsymbol{\beta}(t))^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta}(s)) = \boldsymbol{\beta}(t)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}(s)$ with eigenvalues $\sigma_k^2 = n\mu_k(\boldsymbol{\Xi})$.

We will prove part (a) by induction. When $k = 1$, we first show that the maximum value of (3.7) is less than or equal to $\mu_1(\boldsymbol{\Xi})$. Let $\boldsymbol{\alpha} \in \mathbb{R}^p$ be a vector satisfying the constraint in (3.7) with $k = 1$, that is, $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{S}\boldsymbol{\alpha} = 1$. Define $\mathbf{v} = \mathbf{Z}\boldsymbol{\alpha}$. Then we have $\|\mathbf{v}\|_2^2 = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{S}\boldsymbol{\alpha} = 1$. By the definition of $\mathbf{B}$ in (3.6),

$$\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{Z}\boldsymbol{\alpha} = \mathbf{v}^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{v} \leq \mu_1(\boldsymbol{\Xi})\|\mathbf{v}\|_2^2 = \mu_1(\boldsymbol{\Xi}).$$

Therefore, the maximum value of (3.7) is not greater than $\mu_1(\boldsymbol{\Xi})$. On the other hand, by (3.4),

$$\mathbf{Z}\boldsymbol{\alpha}_1 = \frac{1}{\sqrt{n}}\mathbf{X}\boldsymbol{\alpha}_1 = \boldsymbol{\gamma}_1, \tag{B.2}$$

which leads to

$$\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{Z}\boldsymbol{\alpha}_1 = \boldsymbol{\gamma}_1^{\mathrm{T}}\boldsymbol{\Xi}\boldsymbol{\gamma}_1 = \mu_1(\boldsymbol{\Xi})\boldsymbol{\gamma}_1^{\mathrm{T}}\boldsymbol{\gamma}_1 = \mu_1(\boldsymbol{\Xi}).$$

Hence, when $k = 1$, $\boldsymbol{\alpha}_1$ is the solution to (3.7) and the maximum value is $\mu_1(\boldsymbol{\Xi}) = \sigma_1^2/n$.

Now we assume that for any $1 \leq j < k$, $\boldsymbol{\alpha}_j$ is the solution to (3.7) with $k = j$ and the maximum value is $\boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_j = \mu_j(\boldsymbol{\Xi})$. Based on this induction hypothesis, we will prove that $\boldsymbol{\alpha}_k$ is the solution to (3.7) and the maximum value is $\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_k = \mu_k(\boldsymbol{\Xi})$. For any $\boldsymbol{\alpha} \in \mathbb{R}^p$ satisfying the constraints in (3.7) , that is, $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{S}\boldsymbol{\alpha} = 1$ and $\boldsymbol{\alpha}_l^{\mathrm{T}}\mathbf{S}\boldsymbol{\alpha} = 0$ for all $1 \leq l \leq k-1$, let $\mathbf{v} = \mathbf{Z}\boldsymbol{\alpha}$. Then we have $\|\mathbf{v}\|_2^2 = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{S}\boldsymbol{\alpha} = 1$ and $\boldsymbol{\alpha}_l^{\mathrm{T}}\mathbf{S}\boldsymbol{\alpha} = \boldsymbol{\alpha}_l^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\alpha} = \boldsymbol{\gamma}_l^{\mathrm{T}}\mathbf{v} = 0$ for all $1 \leq l \leq k-1$. Therefore, $\mathbf{v}$ is orthogonal to the first $k-1$ eigenvectors of $\boldsymbol{\Xi}$, and then we have

$$\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{Z}\boldsymbol{\alpha} = \mathbf{v}^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{v} \leq \mu_k(\boldsymbol{\Xi})\|\mathbf{v}\|_2^2 = \mu_k(\boldsymbol{\Xi}).$$

So the maximum value of (3.7) is not greater than $\mu_k(\boldsymbol{\Xi})$. On the other hand,

$$\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{Z}\boldsymbol{\alpha}_k = \boldsymbol{\gamma}_k^{\mathrm{T}}\boldsymbol{\Xi}\boldsymbol{\gamma}_k = \mu_k(\boldsymbol{\Xi})\boldsymbol{\gamma}_k^{\mathrm{T}}\boldsymbol{\gamma}_k = \mu_k(\boldsymbol{\Xi}).$$

Hence, $\boldsymbol{\alpha}_k$ is the solution to (3.7) and the maximum value is $\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_k = \mu_k(\boldsymbol{\Xi})$. By induction, Part (a) holds for any $1 \leq k \leq K$.

For part (b), we just need to prove that $\boldsymbol{\Xi}$ has exactly K positive eigenvalues $\mu_1(\boldsymbol{\Xi}) \geq \mu_2(\boldsymbol{\Xi}) \geq \cdots \geq \mu_K(\boldsymbol{\Xi}) > 0$. Let $\boldsymbol{\gamma}_{K+1}$ be the $(K+1)$-th eigenvector of $\boldsymbol{\Xi}$ with the eigenvalue $\mu_{K+1}(\boldsymbol{\Xi})$. We will show that $\mu_{K+1}(\boldsymbol{\Xi}) = 0$. Since $\boldsymbol{\gamma}_{K+1}$ is orthogonal to the first K eigenvectors $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_K$, by the SVD decomposition (3.2) of $\mathbf{X}\boldsymbol{\beta}(t)$, we have $(\mathbf{X}\boldsymbol{\beta}(t))^{\mathrm{T}}\boldsymbol{\gamma}_{K+1} = 0$. By the definition of $\boldsymbol{\Xi}$ in (3.6), we have $\mu_{K+1}(\boldsymbol{\Xi}) = \boldsymbol{\gamma}_{K+1}\boldsymbol{\Xi}\boldsymbol{\gamma}_{K+1} = 0$.

For part (c), due to the orthogonality of $\{\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_K\}$ and $\{u_1(t), \cdots, u_K(t)\}$,

$$\|\mathbf{X}\boldsymbol{\beta}(t) - \sum_{i=1}^{k} \sigma_k \boldsymbol{\gamma}_k u_k(t)\|_{L^2}^2 = \| \sum_{i=k+1}^{K} \sigma_k \boldsymbol{\gamma}_k u_k(t)\|_{L^2}^2 = \sum_{i=k+1}^{K} \sigma^2 = n \sum_{i=k+1}^{K} \mu_i(\boldsymbol{\Xi}).$$

The proof is completed.

B.2. Proof of Theorem 4.1

First, we have

$$\begin{aligned} &\min_{\widetilde{\boldsymbol{\beta}}_k} E\left[\|\mu(t) + (\mathbf{x}^{\text{new}})^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}_k(t) - y^{\text{new}}(t)\|_{L^2}^2\right] \qquad \text{(B.3)} \\ =& \min_{\substack{\mathbf{b}_j \in \mathbb{R}^p, v_j(t) \in L^2[0,1], \\ 1 \le j \le k}} E\left[\| \sum_{j=1}^{k} (\mathbf{x}^{\text{new}})^{\mathrm{T}} \mathbf{b}_j v_j(t) - (\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}(t) - \varepsilon^{\text{new}}(t)\|_{L^2}^2\right] \\ =& \min_{\substack{\mathbf{b}_j \in \mathbb{R}^p, v_j(t) \in L^2[0,1], \\ 1 \le j \le k}} E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}(t) - \sum_{j=1}^{k} (\mathbf{x}^{\text{new}})^{\mathrm{T}} \mathbf{b}_j v_j(t)\|_{L^2}^2\right] + E[\|\varepsilon^{\text{new}}(t)\|_{L^2}^2], \end{aligned}$$

and similarly,

$$\begin{aligned} &E\left[\|\mu(t) + (\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}_k(t) - y^{\text{new}}(t)\|_{L^2}^2\right] \qquad \text{(B.4)} \\ =&E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}(t) - (\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}_k(t)\|_{L^2}^2\right] + E[\|\varepsilon^{\text{new}}(t)\|_{L^2}^2]. \end{aligned}$$

Next, $(\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}(t)$ is a stochastic process in $[0, 1]$ and its Karhunen-Loève expansion is given by $(\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}(t) = \sum_{k=1}^{\infty} Z_k \phi_k(t)$, where $\phi_k(t)$ is the k-th eigenfunction of the covariance function of $(\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}(t)$ and

$$Z_k = \int_0^1 (\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}(t) \phi_k(t) dt = (\mathbf{x}^{\text{new}})^{\mathrm{T}} \mathbf{b}_k^0, \quad \mathbf{b}_k^0 = \int_0^1 \boldsymbol{\beta}(t) \phi_k(t) dt. \qquad \text{(B.5)}$$

It is well known that the truncated Karhunen-Loève expansion has the minimum mean integrated squared error. That is, for any $k \ge 1$, we have

$$\begin{aligned} &E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}(t) - \sum_{j=1}^{k} (\mathbf{x}^{\text{new}})^{\mathrm{T}} \mathbf{b}_j^0 \phi_j(t)\|_{L^2}^2\right] \\ =& \min_{\substack{\widetilde{Z}_j, v_j(t), \\ 1 \le j \le k}} E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}(t) - \sum_{j=1}^{k} \widetilde{Z}_j v_j(t)\|_{L^2}^2\right], \end{aligned}$$

where the minimum is taken over all possible random variables $\widetilde{Z}_j$ and all possible nonrandom function $v_j(t)$, $1 \le j \le k$. Therefore, we have

$$\min_{\substack{\mathbf{b}_j \in \mathbb{R}^p, v_j(t) \in L^2[0,1], \\ 1 \le j \le k}} E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}} \boldsymbol{\beta}(t) - \sum_{j=1}^{k} (\mathbf{x}^{\text{new}})^{\mathrm{T}} \mathbf{b}_j v_j(t)\|_{L^2}^2\right]$$

$$
\begin{aligned}
&\leq E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-\sum_{j=1}^{k}(\mathbf{x}^{\text{new}})^{\mathrm{T}}\mathbf{b}_j^0\phi_j(t)\|_{L^2}^2\right]\\
&=\min_{\substack{\widetilde{Z}_j,v_j(t),\\ 1\leq j\leq k}} E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-\sum_{j=1}^{k}\widetilde{Z}_j v_j(t)\|_{L^2}^2\right]\\
&\leq \min_{\substack{\mathbf{b}_j\in\mathbb{R}^p,v_j(t)\in L^2[0,1],\\ 1\leq j\leq k}} E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-\sum_{j=1}^{k}(\mathbf{x}^{\text{new}})^{\mathrm{T}}\mathbf{b}_j v_j(t)\|_{L^2}^2\right],
\end{aligned}
$$

which implies

$$
\begin{aligned}
&E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-\sum_{j=1}^{k}(\mathbf{x}^{\text{new}})^{\mathrm{T}}\mathbf{b}_j^0\phi_j(t)\|_{L^2}^2\right]\\
=&\min_{\substack{\mathbf{b}_j\in\mathbb{R}^p,v_j(t)\in L^2[0,1],\\ 1\leq j\leq k}} E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-\sum_{j=1}^{k}(\mathbf{x}^{\text{new}})^{\mathrm{T}}\mathbf{b}_j v_j(t)\|_{L^2}^2\right]. \qquad \text{(B.6)}
\end{aligned}
$$

On the other hand, by (B.5), we have $\mathcal{M}(\boldsymbol{\beta}(t)) = \mathcal{M}\Big(\boldsymbol{\beta}(t) - \sum_{j=1}^{k}(\mathbf{x}^{\text{new}})^{\mathrm{T}}\mathbf{b}_j v_j(t)\Big)$. Then we have

$$
\left|E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-\sum_{j=1}^{k}(\mathbf{x}^{\text{new}})^{\mathrm{T}}\mathbf{b}_j^0\phi_j(t)\|_{L^2}^2\right]-\frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}(t)-\sum_{j=1}^{k}\mathbf{b}_j^0\phi_j(t)\|_{L^2}^2\right| \qquad \text{(B.7)}
$$

$$
\begin{aligned}
=&\left|\int_0^1\left(\boldsymbol{\beta}(t)-\sum_{j=1}^{k}\mathbf{b}_j^0\phi_j(t)\right)\boldsymbol{\Sigma}\left(\boldsymbol{\beta}(t)-\sum_{j=1}^{k}\mathbf{b}_j^0\phi_j(t)\right)dt\right.\\
&\left.-\int_0^1\left(\boldsymbol{\beta}(t)-\sum_{j=1}^{k}\mathbf{b}_j^0\phi_j(t)\right)\mathbf{S}\left(\boldsymbol{\beta}(t)-\sum_{j=1}^{k}\mathbf{b}_j^0\phi_j(t)\right)dt\right|\\
=&\left|\int_0^1\left(\boldsymbol{\beta}(t)-\sum_{j=1}^{k}\mathbf{b}_j^0\phi_j(t)\right)(\mathbf{S}-\boldsymbol{\Sigma})\left(\boldsymbol{\beta}(t)-\sum_{j=1}^{k}\mathbf{b}_j^0\phi_j(t)\right)dt\right|\\
\leq&\|\mathbf{S}-\boldsymbol{\Sigma}\|_\infty\int_0^1\|\boldsymbol{\beta}(t)-\sum_{j=1}^{k}\mathbf{b}_j^0\phi_j(t)\|_1^2dt\\
\leq&\|\mathbf{S}-\boldsymbol{\Sigma}\|_\infty\mathcal{M}\left(\boldsymbol{\beta}(t)-\sum_{j=1}^{k}(\mathbf{x}^{\text{new}})^{\mathrm{T}}\mathbf{b}_j v_j(t)\right)\int_0^1\|\boldsymbol{\beta}(t)-\sum_{j=1}^{k}\mathbf{b}_j^0\phi_j(t)\|_2^2dt\\
=&\|\mathbf{S}-\boldsymbol{\Sigma}\|_\infty\mathcal{M}(\boldsymbol{\beta}(t))\|\boldsymbol{\beta}(t)-\sum_{j=1}^{k}\mathbf{b}_j^0\phi_j(t)\|_{L^2}^2
\end{aligned}
$$

$$\leq \|\mathbf{S}-\boldsymbol{\Sigma}\|_\infty \mathcal{M}(\boldsymbol{\beta}(t))\|\boldsymbol{\beta}(t)\|_{L^2}^2 \leq C\|\boldsymbol{\beta}(t)\|_{L^2}^2\mathcal{M}(\boldsymbol{\beta}(t))\sqrt{\frac{\ln p}{n}}$$

where the inequality in the third line from the last follows from the Cauchy-Schwarz inequality and and the first inequality in the last line is because each coordinate function of $\sum_{j=1}^k \mathbf{b}_j^0\phi_j(t)$ is the projection of the corresponding coordinate function of $\boldsymbol{\beta}(t)$ onto the space spanned by $\{\phi_i, 1 \leq i \leq k\}$. Similarly, we have

$$\left| E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}_k(t)\|_{L^2}^2\right] - \frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}(t)-\mathbf{X}\boldsymbol{\beta}_k(t)\|_{L^2}^2 \right| \tag{B.8}$$
$$\leq C\|\boldsymbol{\beta}(t)\|_{L^2}^2\mathcal{M}(\boldsymbol{\beta}(t))\sqrt{\frac{\ln p}{n}}.$$

By (B.6), (B.7) and (B.8), we have

$$E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}_k(t)\|_{L^2}^2\right] \leq \frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}(t)-\mathbf{X}\boldsymbol{\beta}_k(t)\|_{L^2}^2 \tag{B.9}$$
$$+\left| E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}_k(t)\|_{L^2}^2\right] - \frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}(t)-\mathbf{X}\boldsymbol{\beta}_k(t)\|_{L^2}^2 \right| \tag{B.10}$$
$$\leq \frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}(t)-\mathbf{X}\boldsymbol{\beta}_k(t)\|_{L^2}^2 + C\|\boldsymbol{\beta}(t)\|_{L^2}^2\mathcal{M}(\boldsymbol{\beta}(t))\sqrt{\frac{\ln p}{n}}$$
$$\leq \frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}(t)-\sum_{j=1}^k \mathbf{b}_j^0\phi_j(t)\|_{L^2}^2 + C\|\boldsymbol{\beta}(t)\|_{L^2}^2\mathcal{M}(\boldsymbol{\beta}(t))\sqrt{\frac{\ln p}{n}}$$
$$\leq \left| E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-\sum_{j=1}^k(\mathbf{x}^{\text{new}})^{\mathrm{T}}\mathbf{b}_j^0\phi_j(t)\|_{L^2}^2\right] - \frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}(t)-\sum_{j=1}^k \mathbf{b}_j^0\phi_j(t)\|_{L^2}^2 \right|$$
$$+E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-\sum_{j=1}^k(\mathbf{x}^{\text{new}})^{\mathrm{T}}\mathbf{b}_j^0\phi_j(t)\|_{L^2}^2\right] + C\|\boldsymbol{\beta}(t)\|_{L^2}^2\mathcal{M}(\boldsymbol{\beta}(t))\sqrt{\frac{\ln p}{n}}$$
$$\leq E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-\sum_{j=1}^k(\mathbf{x}^{\text{new}})^{\mathrm{T}}\mathbf{b}_j^0\phi_j(t)\|_{L^2}^2\right] + 2C\|\boldsymbol{\beta}(t)\|_{L^2}^2\mathcal{M}(\boldsymbol{\beta}(t))\sqrt{\frac{\ln p}{n}}$$
$$=\min_{\substack{\mathbf{b}_j\in\mathbb{R}^p, v_j(t)\in L^2[0,1],\\ 1\leq j\leq k}} E\left[\|(\mathbf{x}^{\text{new}})^{\mathrm{T}}\boldsymbol{\beta}(t)-\sum_{j=1}^k(\mathbf{x}^{\text{new}})^{\mathrm{T}}\mathbf{b}_j v_j(t)\|_{L^2}^2\right]$$
$$+2C\|\boldsymbol{\beta}(t)\|_{L^2}^2\mathcal{M}(\boldsymbol{\beta}(t))\sqrt{\frac{\ln p}{n}}, \tag{B.11}$$

where the third inequality follows from the facts that $\mathbf{X}\boldsymbol{\beta}_k(t)$ is the best k-dimensional approximation to $\mathbf{X}\boldsymbol{\beta}(t)$. (B.3), (B.4), and (B.11) give the theorem.

References

[1] Bache, K. and Lichman, M. (2013) UCI machine learning repository. http://archive.ics.uci.edu/ml.

[2] Bickel, P. J. and Levina, E. (2008) Regularized estimation of large covariance matrices. *The Annals of Statistics*, 199–227. MR2387969

[3] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 1705–1732. MR2533469

[4] Bosq, D. (2012) *Linear processes in function spaces: theory and applications*, vol. 149. Springer Science & Business Media. MR1783138

[5] Brown, P. J., Fearn, T. and Vannucci, M. (2001) Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, **96**, 398–408. MR1939343

[6] Cardot, H., Ferraty, F. and Sarda, P. (1999) Functional linear model. *Statistics and Probability Letters*, **45**, 11–22. MR1718346

[7] Crainiceanu, C., Reiss, P., Goldsmith, J., Huang, L., Huo, L. and Scheipl, F. (2014) refund: Regression with functional data. R package version 3.0.1. http://CRAN.R-project.org/package=refund.

[8] Daubechies, I. *et al.* (1992) *Ten lectures on wavelets*, vol. 61. SIAM. MR1162107

[9] De Vito, S., Massera, E., Piga, M., Martinotto, L. and Di Francia, G. (2008) On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, **129**, 750–757.

[10] Faraway, J. J. (1997) Regression analysis for a functional response. *Technometrics*, **39**, 254–261. MR1462586

[11] Ferraty, F. and Vieu, P. (2006) *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media. MR2229687

[12] Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2011) Penalized functional regression. *Journal of Computational and Graphical Statistics*, **20**. MR2878950

[13] Goldsmith, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2012) Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**, 453–469. MR2914521

[14] Guo, W., Dai, M., Ombao, H. C. and von Sachs, R. (2003) Smoothing spline anova for time-dependent spectral analysis. *Journal of the American Statistical Association*, **98**, 643–652. MR2011677

[15] Hart, J. D. and Wehrly, T. E. (1986) Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association*, **81**, 1080–1088. MR0867635

[16] Horváth, L. and Kokoszka, P. (2012) *Inference for functional data with applications*, vol. 200. Springer Science & Business Media. MR2920735

[17] Hsing, T. and Eubank, R. (2015) *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons. MR3379106

[18] Ivanescu, A. E., Staicu, A.-M., Scheipl, F. and Greven, S. (2014) Penalized function-on-function regression. *Computational Statistics*, 1–30. MR3357075

[19] J. O. Ramsay, Hadley Wickham, S. G. and Hooker, G. (2014) fda: Functional data analysis. R package version 3.0.1. http://CRAN.R-project.org/package=fda.

[20] James, G. M. (2002) Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 411–432. MR1924298

[21] James, G. M. and Silverman, B. W. (2005) Functional adaptive model estimation. *Journal of the American Statistical Association*, **100**, 565–576. MR2160560

[22] Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, **104**, 682–693. MR2751448

[23] Ledoux, M. and Talagrand, M. (2011) *Probability in Banach Spaces: Isoperimetry and Processes*. Classics in Mathematics. Springer. MR2814399

[24] Li, B. and Marx, B. D. (2008) Sharpening p-spline signal regression. *Statistical Modelling*, **8**, 367–383. MR2749829

[25] Lin, X., Wang, N., Welsh, A. H. and Carroll, R. J. (2004) Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. *Biometrika*, **91**, 177–193. MR2050468

[26] Luo, R. and Qi, X. (2015) Sparse wavelet regression with multiple predictive curves. *Journal of Multivariate Analysis*, **134**, 33–49. MR3296032

[27] Malloy, E. J., Morris, J. S., Adar, S. D., Suh, H., Gold, D. R. and Coull, B. A. (2010) Wavelet-based functional linear mixed models: an application to measurement error–corrected distributed lag models. *Biostatistics*, **11**, 432–452.

[28] Marx, B. D. and Eilers, P. H. (1999) Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics*, **41**, 1–13.

[29] McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F. and Ruppert, D. (2014) Functional generalized additive models. *Journal of Computational and Graphical Statistics*, **23**, 249–269. MR3173770

[30] Meyer, M. J., Coull, B. A., Versace, F., Cinciripini, P. and Morris, J. S. (2015) Bayesian function-on-function regression for multilevel functional data. *Biometrics*, **71**, 563–574. MR3402592

[31] Morris, J. S. and Carroll, R. J. (2006) Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 179–199. MR2188981

[32] Müller, H.-G. and Stadtmüller, U. (2005) Generalized functional linear models. *Annals of Statistics*, 774–805. MR2163159

[33] Nason, G. (2010) *Wavelet methods in statistics with R*. Springer. MR2445580

[34] Nason, G. (2013) wavethresh: Wavelets statistics and transforms. R package version 4.6.6. https://cran.r-project.org/web/packages/wavethresh/index.html.

[35] Qi, X., Luo, R., Carroll, R. J. and Zhao, H. (2015) Sparse regression by projection and sparse discriminant analysis. *Journal of Computational and Graphical Statistics*, **24**, 416–438. MR3357388

[36] Qi, X., Luo, R. and Zhao, H. (2013) Sparse principal component analysis by choice of norm. *Journal of Multivariate Analysis*, **114**, 127–160. MR2993878

[37] R Core Team (2015) R: A language and environment for statistical computing,. R Foundation for Statistical Computing, Vienna, Austria. `http://CRAN.R-project.org/package=refund`.

[38] Ramsay, J. O. and Dalzell, C. (1991) Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 539–572. MR1125714

[39] Ramsay, J. O. and Silverman, B. W. (2005) *Functional data analysis. 2nd Edition.* New York: Springer. MR2168993

[40] Rasmussen, C. E. and Williams, C. K. I. (2005) *Gaussian processes for machine learning (adaptive computation and machine learning series)*, chap. 4. The MIT Pres. MR2514435

[41] Ratcliffe, S. J., Heller, G. Z. and Leader, L. R. (2002) Functional data analysis with application to periodically stimulated foetal heart rate data. ii: Functional logistic regression. *Statistics in Medicine*, **21**, 1115–1127.

[42] Reiss, P. T., Huang, L. and Mennes, M. (2010) Fast function-on-scalar regression with penalized basis expansions. *The International Journal of Biostatistics*, **6**. MR2683940

[43] Reiss, P. T. and Ogden, R. T. (2007) Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, **102**, 984–996. MR2411660

[44] Roceanu, A., Onub, M., Antochi, F. and Bajenaru, O. (2012) Diffusion tensor imaging (dti)-a new imaging technique applyed in multiple sclerosis. *Maedica a Journal of Clinical Medicine*, **7**.

[45] Røislien, J. and Winje, B. (2013) Feature extraction across individual time series observations with spikes using wavelet principal component analysis. *Statistics in Medicine*, **32**, 3660–3669. MR3095504

[46] Scheipl, F., Staicu, A.-M. and Greven, S. (2015) Functional additive mixed models. *Journal of Computational and Graphical Statistics*, **24**, 477–501. MR3357391

[47] Song, S.-K., Sun, S.-W., Ramsbottom, M. J., Chang, C., Russell, J. and Cross, A. H. (2002) Dysmyelination revealed through mri as increased radial (but unchanged axial) diffusion of water. *Neuroimage*, **17**, 1429–1436.

[48] The MathWorks (2011) Matlab and statistics toolbox release 2011. The MathWorks, Inc., Natick, Massachusetts, United States.

[49] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**, 267–288. MR1379242

[50] Tievsky, A. L., Ptak, T. and Farkas, J. (1999) Investigation of apparent diffusion coefficient and diffusion tensor anisotropy in acute and chronic multiple sclerosis lesions. *American Journal of Neuroradiology*, **20**, 1491–1499.

[51] Wang, W. (2014) Linear mixed function-on-function regression models. *Biometrics*, **70**, 794–801. MR3295740

[52] Wu, S. and Müller, H.-G. (2011) Response-adaptive regression for longitudinal data. *Biometrics*, **67**, 852–860. MR2829259

[53] Yao, F. and Müller, H.-G. (2010) Functional quadratic regression. *Biometrika*, **97**, 49–64. MR2594416

[54] Yao, F., Müller, H.-G., Wang, J.-L. *et al.* (2005) Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, **33**, 2873–2903. MR2253106

[55] Zhao, Y., Ogden, R. T. and Reiss, P. T. (2012) Wavelet-based lasso in functional linear regression. *Journal of Computational and Graphical Statistics*, **21**, 600–617. MR2970910