Journal of Multivariate Analysis 134 (2015) 33–49

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Sparse wavelet regression with multiple predictive curves

Ruiyan Luo [a,*], Xin Qi [b]

[a] Division of Epidemiology and Biostatistics, School of Public Health, Georgia State University, One Park Place, Atlanta, GA 30303, United States

[b] Department of Mathematics and Statistics, Georgia State University, 30 Pryor Street, Atlanta, GA 30303, United States

ARTICLE INFO

Article history:
Received 11 February 2014
Available online 6 November 2014

AMS subject classifications:
primary 62J05
secondary 62G05
62G20

Keywords:
Functional linear model
Wavelet transformation
Sparse regression

ABSTRACT

With the advance of techniques, more and more complicated data are extracted and recorded. In this paper, functional regression models with a scalar response and multiple predictive curves are considered. We transform the functional regression models to multiple linear regression models by using the discrete wavelet transformation. When the number of predictive curves is big, the multiple linear regression model usually has much bigger number of features than the sample size. We apply our correlation-based sparse regression method to the resulted high dimensional regression model. The novel feature of our sparse method is that we impose sparsity penalty on the direction of the estimate of the coefficient vector instead of the estimate itself, and only the direction of the estimate is determined by an optimization problem. The estimation consistency of the coefficient curve for the functional regression model is obtained when both the sample size and the number of curves go to infinity. The effects of the discrete observations are discussed. We compare our method with both functional regression methods and other wavelet based sparse regression methods on both simulated data and four real data sets, including the cases of single and multiple predictive curves. The results indicate that sparse wavelet regression methods are better in extracting local features and our method has good predictive performances in all scenarios.

Published by Elsevier Inc.

1. Introduction

In the traditional functional data, there are one or only a few predictive curves for each subject. However, with the advance of techniques, more and more complicated data are extracted and recorded. The number of curves recorded for each subject can be comparable with or larger than the sample size. For example, as popular techniques in neuroscience, Electroencephalography (EEG), Magnetoencephalography (MEG) and Functional magnetic resonance imaging (fMRI) techniques can record the brain's spontaneous activity over a period of time on different areas of the scalp, or different regions inside the brain. Hundreds or thousands of time series curves can be generated simultaneously for one subject.

In this paper, we will consider the following functional linear regression model with a scalar response variable and multiple predictive curves:

$$y_i = \sum_{j=1}^{K} \int_0^1 Z_{ij}(t) g_j(t) dt + \widetilde{\epsilon}_i, \quad 1 \le i \le n, \tag{1.1}$$

* Corresponding author.
E-mail addresses: rluo@gsu.edu (R. Luo), xqi3@gsu.edu (X. Qi).

http://dx.doi.org/10.1016/j.jmva.2014.10.003
0047-259X/Published by Elsevier Inc.

where y_i is the scalar response in the ith sample, $Z_{ij}(t)$ is the jth predictive function for the ith sample, and $g_j(t)$ is the unknown coefficient function for the jth predictive functional variable, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, K$ with n and K representing the sample size and the number of predictive functions, respectively. We assume that the noises $\widetilde{\epsilon}_i,\ 1 \leq i \leq n$, are i.i.d. and have mean zeros. Without loss of generality, we use $[0,\ 1]$ to denote the interval where the curves are observed, and assume that y and $Z_{ij}(t)$ have been centered.

When K is one or relatively small, many functional regression methods have been proposed, such as the penalized spline method [18,8,24], the functional principal component regression [7,26], the functional partial least squares [22,26,16], and many others. However, when K is relatively big, the number of features, which is K times the number of basis functions, can be much larger than the sample size. Then feature selection or sparsity regularity is necessary as it has been known that the classic linear regression has poor prediction accuracy, principal component analysis and partial least squares are inconsistent [21,19,15,9], in the situation of large p and small n. In [23], we introduce a new framework of regression and multivariate analysis for high dimensional data. The novel feature of our sparse regression method is that we impose sparsity penalty on the direction of the estimate of the coefficient vector instead of the estimate itself, and only the direction of the estimate is determined by an optimization problem. The length of the estimate together with the tuning parameters are determined by a cross validation procedure aiming to minimize the prediction errors. Intensive simulation studies have been performed to demonstrate that the new methods have good predictive abilities compared to the well-known methods such as the LASSO and elastic-net. In this paper, we propose to first conduct discrete wavelet transformation for the predictive curves, and then use our sparse regression method to predict the response variable by the wavelet coefficients. The unknown coefficient curves can be obtained by the inverse discrete wavelet transformation of the estimated coefficients.

The wavelet transformation is orthogonal and has multiscale structures. Hence, it is particularly useful in the extraction of local features of curves at different levels of resolution. The wavelet transformation enjoys the sparsity property in the sense that "the wavelet transformation compresses the energy into a smaller number of coefficients" [11]. For a large variety of functions, the wavelet decomposition allows good representation of the function using only a fraction of the full set of wavelet coefficients. Several wavelet based methods have been proposed for functional linear regression in the last decade. For example, Brown et al. [6] used a Bayesian variable selection method to select a subset of wavelet coefficients for prediction. The intensive computation of this method hinders its application when K is big. Zhao et al. [31] employed the LASSO to wavelet coefficients and selected predictors of the response variable. We will apply our sparse regression method to the multiple linear regression model after wavelet transformation and show that our method has better predictive performance, and can achieve the coefficient curve estimation consistency when both the sample size and the number of curves go to infinity.

The rest of the paper is organized as follows. In Section 2, we review the discrete wavelet transformation and convert the functional linear regression problem to a multiple linear regression problem. In Section 3, we review our sparse regression method. Section 4 provides the consistency of the coefficient curve estimate of our wavelet-based sparse regression method for the functional linear regression when both n and K go to infinity. Simulation study and case studies are presented in Sections 5 and 6, respectively. Section 7 is a short discussion. Algorithmic details, the related theorems, all proofs and additional plots are provided in the online supplementary materials (see Appendix A).

2. Discrete wavelet transformation for discretely observed functional data

There are different types of wavelet basis functions, such as Haar wavelet, Shannon wavelet and so on. In this paper, we choose the Daubechies wavelets with ten vanishing moments which enjoy the nice properties such as compactly supported and smooth. The Daubechies wavelets are a set of complete orthonormal basis of the L^2 space. Hence, any function in L^2 can be written as a linear combination of the Daubechies wavelets and the coefficients are equal to the integrations of the multiplications of the function and the corresponding Daubechies wavelets. However, in practice, all curves are discretely observed and recorded. For notational simplification, we assume that all the curves, $Z_{ij}(t)$, in (1.1) are observed in the same set of equally spaced discrete observation points, $\{t_1, t_2, \ldots, t_N\}$, where $N = 2^J$ and J is a positive number. Our methods can be directly applied to the general case where the numbers of observation points in different curves are different.

The discrete wavelet transformation (DWT) as described by Mallat [17] is a special case of a two-channel subband coder using the conjugate quadrature filters of Smith and Barnwell [27]. The DWT is not a transformation of the curves in $L^2[0,\ 1]$, instead it is the transformation of the vector of discrete observations in the curve [10]. In fact, the DWT is an orthogonal linear transformation from $\mathbb{R}^N$ to $\mathbb{R}^N$. For any $\mathbf{z} \in \mathbb{R}^N$, the DWT converts it to a coefficient vector $\mathbf{d} \in \mathbb{R}^N$ of the N wavelet basis functions: the father wavelet, $\phi(t)$, and the mother wavelets, $\{\psi_{l,k}(t) : 0 \leq l \leq J-1, 0 \leq k \leq 2^l - 1\}$. The index l in $\psi_{l,k}(t)$ denotes the resolution level. A larger l indicates that $\psi_{l,k}(t)$ is a function with finer scale and can capture more detailed local information. The l can be any large positive integer. However, due to the discrete observations, we cannot extract the local information of the curve finer than the resolution level $J-1$. Since the DWT, $\mathbf{z} \to \mathbf{d}$, is orthogonal, there exists an orthogonal matrix $\mathbf{W}$ such that

$$\mathbf{d} = \mathbf{W}\mathbf{z}, \quad \text{and hence,} \quad \mathbf{z} = \mathbf{W}^{\mathrm{T}}\mathbf{d}. \tag{2.1}$$

For any smooth function $f(t)$, let $\mathbf{d} = (c_0, d_{0,0}, \ldots, d_{J-1,2^{J-1}-1})$ be the DWT coefficients of $(f(t_1), \ldots, f(t_N))$, then $\frac{1}{\sqrt{N}}[c_0\phi(t) + \sum_{l=1}^{J-1}\sum_{k=0}^{2^l-1} d_{l,k}\psi_{l,k}(t)]$ is an approximation of $f(t)$ and as $N \to \infty$, the expansion converges to $f(t)$. Moreover,

by the orthogonality of the DWT, we have

$$c_0^2 + \sum_{l=1}^{J-1}\sum_{k=0}^{2^l-1} d_{l,k}^2 = \sum_{k=1}^{N} f(t_k)^2. \tag{2.2}$$

Let $\mathbf{x}_{ij}$ be the N-dimensional vector of the DWT of $(Z_{ij}(t_1), \ldots, Z_{ij}(t_N))$ and $\mathbf{b}_j$ be the N-dimensional vector of the DWT of $(g_j(t_1), \ldots, g_j(t_N))$, where $g_j(t)$ is the coefficient function for $Z_{ij}(t)$ as in model (1.1). Then by (2.1) and due to the orthogonality of $\mathbf{W}$, we have

$$\int_0^1 Z_{ij}(t)g_j(t)dt = \frac{1}{N}\sum_{k=1}^{N} Z_{ij}(t_k)g_j(t_k) + \varepsilon'_{ij} = \frac{1}{N}\mathbf{x}_{ij}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{b}_j + \varepsilon'_{ij} = \frac{1}{N}\mathbf{x}_{ij}^{\mathrm{T}}\mathbf{b}_j + \varepsilon'_{ij}, \tag{2.3}$$

where

$$\varepsilon'_{ij} = \int_0^1 Z_{ij}(t)g_j(t)dt - \frac{1}{N}\sum_{k=1}^{N} Z_{ij}(t_k)g_j(t_k) \tag{2.4}$$

are the errors due to the replacement of integrals by averages. Hence, the functional linear regression model (1.1) is transformed to the following multiple linear regression model,

$$y_i = \frac{1}{N}\sum_{j=1}^{K} \mathbf{x}_{ij}^{\mathrm{T}}\mathbf{b}_j + \epsilon_i, \quad 1 \le i \le n, \tag{2.5}$$

where $\epsilon_i = \widetilde{\epsilon}_i + \sum_{j=1}^{K} \varepsilon'_{ij}$. Let $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ be the vector of responses, $\boldsymbol{\beta} = (\frac{1}{N}\mathbf{b}_1^T, \ldots, \frac{1}{N}\mathbf{b}_K^T)^T$ be a KN-vector and

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_{11}^{\mathrm{T}} & \mathbf{x}_{12}^{\mathrm{T}} & \cdots & \mathbf{x}_{1K}^{\mathrm{T}} \\ \mathbf{x}_{21}^{\mathrm{T}} & \mathbf{x}_{22}^{\mathrm{T}} & \cdots & \mathbf{x}_{2K}^{\mathrm{T}} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{x}_{n1}^{\mathrm{T}} & \mathbf{x}_{n2}^{\mathrm{T}} & \cdots & \mathbf{x}_{nK}^{\mathrm{T}} \end{pmatrix} \tag{2.6}$$

be an $n \times (KN)$ matrix. Then (2.5), or equivalently, the model (1.1) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.7}$$

Suppose that we have obtained an estimation $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, where

$$\widehat{\boldsymbol{\beta}} = \left(\frac{1}{N}\widehat{\mathbf{b}}_1^T, \ldots, \frac{1}{N}\widehat{\mathbf{b}}_K^T\right)^T, \tag{2.8}$$

then the estimate $\widehat{g}_j(t)$ is equal to the linear combination of $\phi(t)$ and $\{\psi_{l,k}(t) : 0 \le l \le J-1, 0 \le k \le 2^l - 1\}$ with the coefficients $\frac{1}{\sqrt{N}}\widehat{\mathbf{b}}_j$, $1 \le j \le K$. The effects of discrete observations on the estimates of $g_j(t)$, $1 \le j \le K$, are two folds. If N is small, that is, the discrete observations are sparse, the errors ε'_{ij} in (2.4) will be large, which lead to large estimation errors. On the other hand, if N is too large, we can see from (2.8) that the magnitudes of nonzero features in $\boldsymbol{\beta}$ will be small so that the feature selection and estimation of $\boldsymbol{\beta}$ will be difficult. This will be discussed in more detail in Section 4.

In the following, we will focus on the estimation of $\boldsymbol{\beta}$ in (2.7). Let $p = KN$ be the total number of features, then p can be much larger than the sample size in the case of multiple curves.

When $p \gg n$, the ordinary least squares (OLS) has poor prediction accuracy and meets with problems of interpretation. Various penalization techniques have been proposed to improve OLS, such as ridge regression [14], LASSO [28], elastic net [33], supervised principal components [4], sparse partial least squares regression [9], smoothly clipped absolute deviation (SCAD) [12], minimax concave penalty (MCP) [29], and others. There are two strategies to determine the final estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. In the first strategy, $\widehat{\boldsymbol{\beta}}$ is determined by a regularized optimization problem, such as LASSO and elastic net. But the regularization process leads to that $\mathbf{X}\widehat{\boldsymbol{\beta}}$ is not the projection of $\mathbf{y}$ along the direction of $\mathbf{X}\widehat{\boldsymbol{\beta}}$. When the sparsity tuning parameter is large, $\mathbf{X}\widehat{\boldsymbol{\beta}}$ is far away from the projection of $\mathbf{y}$ along the direction of $\mathbf{X}\widehat{\boldsymbol{\beta}}$, which may lead to large bias in the estimate. In the second strategy, variable selection methods are used to identify the set of relevant variables, and the final model is constructed using only the selected variables. Hence, all the coefficients of the selected variables are estimated in a separate step other than the variable selection step. The performance of the final model will heavily rely on the variable selection.

Our sparse regression method [23] for (2.7) will be introduced in Section 3. It is different from the above two strategies. In our approach, the sparsity penalty is imposed on the direction of the estimate $\widehat{\boldsymbol{\beta}}$. Only the direction of $\widehat{\boldsymbol{\beta}}$ is determined by our penalized optimization problem, and the length of $\widehat{\boldsymbol{\beta}}$ and the tuning parameters are determined by a cross validation procedure to achieve the largest prediction accuracy.

3. Correlation-based sparse regression (CSR)

To introduce our method, we first consider the OLS method. In OLS, the estimate $\widehat{\boldsymbol{\beta}}$ is usually obtained by minimizing $\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$, where $\|\cdot\|_2$ is the L_2 norm. Alternatively, there is an equivalent two-step procedure to obtain the OLS estimator. First, we obtain the direction of $\widehat{\boldsymbol{\beta}}$ by solving

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha}, \quad \text{subject to } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha} \le 1, \tag{3.1}$$

where we assume that $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is full rank for the moment. Let $\widetilde{\boldsymbol{\alpha}}$ be a solution to (3.1). Second, by rescaling $\widetilde{\boldsymbol{\alpha}}$, we get $\widehat{\boldsymbol{\beta}} = (\mathbf{y}^{\mathrm{T}}\mathbf{X}\widetilde{\boldsymbol{\alpha}}/\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widetilde{\boldsymbol{\alpha}})\widetilde{\boldsymbol{\alpha}}$, i.e., $\mathbf{X}\widehat{\boldsymbol{\beta}}$ is the orthogonal projection of $\mathbf{y}$ along the direction of $\mathbf{X}\widetilde{\boldsymbol{\alpha}}$. It is easy to see that $\widetilde{\boldsymbol{\alpha}}$ is also the solution to

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p}(\mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha})^2 = \max_{\boldsymbol{\alpha}\in\mathbb{R}^p}(n-1)^2\mathrm{Cov}^2(\mathbf{y},\mathbf{X}\boldsymbol{\alpha}), \quad \text{subject to } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha} \le 1, \tag{3.2}$$

which is equivalent to

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{(n-1)^2\mathrm{Cov}^2(\mathbf{y},\mathbf{X}\boldsymbol{\alpha})}{\left(\boldsymbol{\alpha}^{T}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha}\right)\cdot\|\mathbf{y}\|_2^2} = \max_{\boldsymbol{\alpha}\in\mathbb{R}^p}\mathrm{Corr}^2(\mathbf{y},\mathbf{X}\boldsymbol{\alpha}), \tag{3.3}$$

where we have assumed that $\mathbf{y}$ and $\mathbf{X}$ are both centered. Hence, solving (3.1) is equivalent to finding the direction $\widetilde{\boldsymbol{\alpha}}$ such that $\mathbf{y}$ and $\mathbf{X}\widetilde{\boldsymbol{\alpha}}$ have the largest correlation among all directions. As a comparison, the partial least squares regression (PLS), another popular regression method, maximizes the covariance between the response and the linear combinations of predictor variables. Note that for any $\boldsymbol{\alpha}$, the correlation between $\mathbf{y}$ and $\mathbf{X}\boldsymbol{\alpha}$ only depends on the direction, rather than the length, of $\boldsymbol{\alpha}$. Our regularized regression method consists of two steps: (1) get the sparse direction, (2) determine the length of the estimate.

3.1. Determination of sparse direction

The direction $\widetilde{\boldsymbol{\alpha}}$ of the sparse estimate of $\boldsymbol{\beta}$ is obtained by solving

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha}, \quad \text{subject to } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha} + \tau\|\boldsymbol{\alpha}\|_\lambda^2 \le 1, \tag{3.4}$$

where $\|\boldsymbol{\alpha}\|_\lambda^2 = (1-\lambda)\|\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1^2$, and both $\tau \ge 0$ and $0 \le \lambda \le 1$ are tuning parameters. The introduction of $\|\boldsymbol{\alpha}\|_2^2$ in the constraint aims to overcome the potential multicollinearity problems or the singularity problem when $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not full rank. The l_1 term in the constraint of (3.4) leads to sparse solutions. We use $\|\boldsymbol{\alpha}\|_1^2$ instead of $\|\boldsymbol{\alpha}\|_1$ as in the elastic net so that the solution to

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha}, \quad \text{subject to } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha} + \tau\|\boldsymbol{\alpha}\|_\lambda^2 \le t,$$

where t is any positive number, differs from the solution to (3.4) only by a multiplicative constant and thus the two solution vectors have the same directions. Hence, the sparsity penalty is actually imposed on the direction of the coefficient vector. Both λ and τ can control the sparsity of $\widetilde{\boldsymbol{\alpha}}$. Larger values of τ and λ lead to sparser $\widetilde{\boldsymbol{\alpha}}$. When $\tau > 0$ and $0 \le \lambda < 1$, the feasible region is strictly convex, and hence the solution to (3.4) is unique.

In [23], we point out that the optimization problem (3.4) has a penalized version (3.5):

$$\min_{\boldsymbol{\beta}} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \tau\left[(1-\lambda)\|\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1^2\right], \tag{3.5}$$

where τ and λ are the same parameters as in (3.4). The solutions to (3.4) and (3.5) differ only by a scaling factor. That is, let $\boldsymbol{\alpha}^*$ be the solution to (3.4), then the solution to (3.5) is

$$\boldsymbol{\beta}^* = \frac{\mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha}^*}{\|\mathbf{X}\boldsymbol{\alpha}^*\|_2^2 + \tau\|\boldsymbol{\alpha}^*\|_\lambda^2}\boldsymbol{\alpha}^*. \tag{3.6}$$

The essential difference is that in (3.4), we do not determine the length, whereas both the direction and length are determined in (3.5).

The major difference between (3.5) and the elastic-net problem is that the squared l_1 norm is used in (3.5) instead of the l_1 norm itself. This difference makes (3.5) (and (3.4)) enjoy scale invariant properties which are not possessed by the elastic-net. Specifically,

(a) If $\boldsymbol{\beta}^*$ is the solution to (3.5), then $c\boldsymbol{\beta}^*$ is the solution to (3.5) with $\mathbf{y}$ replaced by $c\mathbf{y}$, where c is any positive scaling constant.
(b) If $\boldsymbol{\beta}^*$ is the solution to (3.5), then $\boldsymbol{\beta}^*/c$ is the solution to

$$\max_{\boldsymbol{\beta}} \|\mathbf{y}-c\mathbf{X}\boldsymbol{\beta}\|_2^2 + c\tau\left[(1-\lambda)\|\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1^2\right], \tag{3.7}$$

where c is any positive scaling constant.

Hence, scaling $\mathbf{y}$ does not affect the direction of the estimate of the coefficients. When we scale $\mathbf{X}$, we just need to rescale τ by the same amount, then the direction of the coefficient vector estimator is unchanged. However, the elastic-net does not have this property. When $\mathbf{y}$ is scaled, the direction of the estimate of the coefficients is changed.

We use (3.4) instead of the penalized version (3.5) for two reasons. First, we actually develop efficient algorithms for the following more general optimization problem

$$\max_{\mathbf{u}} \mathbf{c}^{\mathrm{T}}\mathbf{u}, \quad \text{subject to } \mathbf{u}^{\mathrm{T}}\mathbf{C}\mathbf{u} + \tau\|\mathbf{u}\|_{\lambda}^2 \le 1, \qquad \mathbf{D}\mathbf{u} = 0, \tag{3.8}$$

where $\mathbf{c}$ is a nonzero vector, $\mathbf{C}$ is a nonnegative definite symmetric matrix and $\mathbf{D}$ is a matrix. (3.4) is only a special case of (3.8) with $\mathbf{u} = \boldsymbol{\alpha}$, $\mathbf{c} = \mathbf{y}^{\mathrm{T}}\mathbf{X}$, $\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{D} = \mathbf{0}$. Second, it can be seen from (3.6) that, when the tuning parameter τ is large, the length of the solution $\boldsymbol{\beta}^*$ to (3.5) is quite small and can be far away from the optimal one, which can lead to large prediction errors. Instead, in (3.5), we only determine the direction of the estimate and then the length is chosen to minimize the prediction errors. In addition, the tuning parameters τ and λ need to be chosen to minimize the prediction errors. Hence, we design a cross-validation procedure to choose the tuning parameters and the length of the estimate simultaneously. The details are described in Section 3.2.

3.2. Choices of tuning parameters and determination of the length of the estimate

We use cross-validation (CV) to choose the tuning parameters. In our method, the length of the estimate is not determined by the optimization problem itself. Instead, it is viewed as a special tuning parameter and will be chosen to maximize the predictive ability. To measure the prediction accuracy of the models corresponding to different values of λ and τ, we must consider the effect of the length of the estimate. Given a pair of λ and τ, we choose the length to minimize the prediction errors which are used as a criterion to choose λ and τ. Hence, our cross-validation procedure is different from those in the Lasso and the elastic-net.

Specifically, we repeat the following procedure 10 times. In the ith repeat, $1 \le i \le 10$, we randomly split the whole data set into a training set and a validation set, where the validation set has one third of all the observations. Let $\widehat{\boldsymbol{\mu}}_X^{(\text{train})}$ denote the matrix with the same size as $\mathbf{X}_{\text{train}}$ and the values in each column equal to the mean of the corresponding column of $\mathbf{X}_{\text{train}}$. Let $\widehat{\boldsymbol{\mu}}_y^{(\text{train})}$ be the vector with each value equal to the mean of $\mathbf{y}_{\text{train}}$. For any pair (τ, λ), the direction, $\widetilde{\boldsymbol{\alpha}}(\tau, \lambda)$, is the solution to

$$\begin{aligned} &\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \quad (\mathbf{y}_{\text{train}} - \widehat{\boldsymbol{\mu}}_y^{(\text{train})})^{\mathrm{T}}(\mathbf{X}_{\text{train}} - \widehat{\boldsymbol{\mu}}_X^{(\text{train})})\boldsymbol{\alpha}, \\ &\text{subject to } \boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{X}_{\text{train}} - \widehat{\boldsymbol{\mu}}_X^{(\text{train})})^{\mathrm{T}}(\mathbf{X}_{\text{train}} - \widehat{\boldsymbol{\mu}}_X^{(\text{train})})\boldsymbol{\alpha} + \tau\|\boldsymbol{\alpha}\|_{\lambda}^2 \le 1. \end{aligned} \tag{3.9}$$

Once the direction is determined, we calculate the mean squared error,

$$\text{MSE}(\tau, \lambda, i) = \min_{\delta\in\mathbb{R}} \|\mathbf{y}_{\text{valid}} - \widehat{\boldsymbol{\mu}}_y^{(\text{train})} - \delta(\mathbf{X}_{\text{valid}} - \widehat{\boldsymbol{\mu}}_X^{(\text{train})})\widetilde{\boldsymbol{\alpha}}(\tau, \lambda)\|_2^2. \tag{3.10}$$

In other words, we choose the length to minimize the prediction errors on the validation set. We choose the pair (τ_0, λ_0) which minimizes $10^{-1}\sum_{i=1}^{10}\text{MSE}(\tau, \lambda, i)$. Then the direction $\widetilde{\boldsymbol{\alpha}}$ of the final estimate solves

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \quad (\mathbf{y} - \widehat{\boldsymbol{\mu}}_y)^{\mathrm{T}}(\mathbf{X} - \widehat{\boldsymbol{\mu}}_X)\boldsymbol{\alpha}, \quad \text{subject to } \boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{X} - \widehat{\boldsymbol{\mu}}_X)^{\mathrm{T}}(\mathbf{X} - \widehat{\boldsymbol{\mu}}_X)\boldsymbol{\alpha} + \tau_0\|\boldsymbol{\alpha}\|_{\lambda_0}^2 \le 1,$$

where $\widehat{\boldsymbol{\mu}}_X$ and $\widehat{\boldsymbol{\mu}}_y$ are the mean matrix and the mean vector of the whole data set $\mathbf{X}$ and $\mathbf{y}$, respectively. The final estimate $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{\boldsymbol{\beta}} = \frac{(\mathbf{y} - \widehat{\boldsymbol{\mu}}_y)^{\mathrm{T}}(\mathbf{X} - \widehat{\boldsymbol{\mu}}_X)\widetilde{\boldsymbol{\alpha}}}{\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}(\mathbf{X} - \widehat{\boldsymbol{\mu}}_X)^{\mathrm{T}}(\mathbf{X} - \widehat{\boldsymbol{\mu}}_X)\widetilde{\boldsymbol{\alpha}}}\widetilde{\boldsymbol{\alpha}}.$$

It is easy to see that $(\mathbf{X} - \widehat{\boldsymbol{\mu}}_X)\widehat{\boldsymbol{\beta}}$ is the projection of $\mathbf{y} - \widehat{\boldsymbol{\mu}}_y$ along the direction of $(\mathbf{X} - \widehat{\boldsymbol{\mu}}_X)\widetilde{\boldsymbol{\alpha}}$.

Another important difference between our CV procedure and the usual CV procedure is that we do not split the whole data set into 10 subsets as the usual 10-fold CV procedure where in each repeat, one subset is selected as the validation set and all the other observations as the training set. The main reason is that in our approach, the length is not determined by the training set. Hence, we increase the size of the validation set to more accurately estimate the prediction errors on the validation set.

4. Asymptotic results

For the multivariate regression model (2.7), Zhao and Yu [32] proved that the Lasso is model selection consistent when both n and p go to infinity. In [23], we proved that our correlation-based sparse regression method described in Section 3 is simultaneously model selection consistent and parameter estimation consistent under the similar setting. The main purpose of this section is to prove that our estimate of the coefficient function through the wavelet transformation is consistent for

the functional regression model (1.1) as both n and the number K of curves go to infinity. Since we need the theoretical result in [23], we briefly describe it below.

Assume that we have a sequence of linear regression models

$$\mathbf{y}^n = \mathbf{X}^n \boldsymbol{\beta}^n + \boldsymbol{\varepsilon}^n, \tag{4.1}$$

where $\boldsymbol{\varepsilon}^n = (\varepsilon_1^n, \ldots, \varepsilon_n^n)^{\mathrm{T}}$ is a vector of i.i.d. random variables with mean zero, $\mathbf{y}^n$ is the n-dimensional response vector and $\mathbf{X}^n$ is the $n \times p_n$ data matrix. We consider the situation where both n and p_n go to infinity. Without loss of generality, suppose that the first q_n coordinates of $\boldsymbol{\beta}^n$ are nonzero and the others are zero. Let $\boldsymbol{\beta}^n = (\boldsymbol{\beta}_1^{n\mathrm{T}}, \boldsymbol{\beta}_2^{n\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_1^n$ and $\boldsymbol{\beta}_2^n = \mathbf{0}$ are the first q_n and the last $p_n - q_n$ coordinates of $\boldsymbol{\beta}^n$, respectively.

Suppose that $\widehat{\boldsymbol{\beta}}^n = (\widehat{\boldsymbol{\beta}}_1^{n\mathrm{T}}, \widehat{\boldsymbol{\beta}}_2^{n\mathrm{T}})^{\mathrm{T}}$ is the solution to (3.4) with tuning parameters λ_n and τ_n, where $\widehat{\boldsymbol{\beta}}_1^n$ and $\widehat{\boldsymbol{\beta}}_2^n$ are the first q_n and the last $p_n - q_n$ coordinates of $\widehat{\boldsymbol{\beta}}^n$, respectively. Then our estimate is

$$\widehat{\boldsymbol{\gamma}}^n = \frac{(\mathbf{y}^n)^{\mathrm{T}} \mathbf{X}^n \widehat{\boldsymbol{\beta}}^n}{\widehat{\boldsymbol{\beta}}^{n\mathrm{T}} \mathbf{X}^{n\mathrm{T}} \mathbf{X}^n \widehat{\boldsymbol{\beta}}^n} \widehat{\boldsymbol{\beta}}^n, \tag{4.2}$$

that is, $\mathbf{X}^n \widehat{\boldsymbol{\gamma}}^n$ is the projection of $\mathbf{y}^n$ along the direction of $\mathbf{X}^n \widehat{\boldsymbol{\beta}}^n$. Estimator $\widehat{\boldsymbol{\beta}}^n$ is said to have the same sign as $\boldsymbol{\beta}^n$ if each coordinate of $\widehat{\boldsymbol{\beta}}_1^n$ has the same sign as the corresponding coordinate of $\boldsymbol{\beta}_1^n$ and $\widehat{\boldsymbol{\beta}}_2^n = \mathbf{0}$. If $P(\widehat{\boldsymbol{\beta}}^n$ has the same sign as $\boldsymbol{\beta}^n) \to 1$, $\widehat{\boldsymbol{\beta}}^n$ is model selection consistent. If $\|\widehat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}^n\|_2 \to 0$ in probability, $\widehat{\boldsymbol{\beta}}^n$ is parameter estimation consistent.

We consider a setting which is essentially the same as in [32]. Let $\mathbf{X}^n = (\mathbf{X}_1^n, \mathbf{X}_2^n)$, where $\mathbf{X}_1^n$ and $\mathbf{X}_2^n$ are the submatrices corresponding to $\boldsymbol{\beta}_1^n$ and $\boldsymbol{\beta}_2^n$, respectively. Let $\mathbf{C}^n = \mathbf{X}^{n\mathrm{T}} \mathbf{X}^n / n$, $\mathbf{C}_{11}^n = \mathbf{X}_1^{n\mathrm{T}} \mathbf{X}_1^n / n$ and $\mathbf{C}_{21}^n = \mathbf{X}_2^{n\mathrm{T}} \mathbf{X}_1^n / n$. Assume that there exist constants $0 \le c_1 < c_2 \le 1$, $0 < c_4 < \frac{c_1}{2} < \frac{c_0}{2}$, $-\infty < c_3 < \infty$, positive M_1, M_2, M_3 and a positive integer k such that the following conditions hold.

Condition 1. 1. *The largest singular values of* $\mathbf{C}_{21}$ *are less than* $O(n^{-c_0})$.
2. *All the eigenvalues of* $\mathbf{C}^n$ *are less than* M_1, *and all the eigenvalues of* $\mathbf{C}_{11}^n$ *are greater than* M_2.
3. $n^{\frac{1-c_2}{2}} \min_{1 \le j \le q} |\boldsymbol{\beta}_j^n| \ge M_3$, $\|\boldsymbol{\beta}_1^n\|_2 \sim n^{c_3}$, $\sup_n E[(\varepsilon_i^n)^{2k}] < \infty$, $q_n = O(n^{c_1})$, $p_n \le O(n^{c_4 k})$.

In [23], we proved the following theorem, which implies that our correlation-based sparse regression method can simultaneously achieve the model selection consistency and parameter estimation consistency.

Theorem 4.1. *Under Condition 1, if we choose* $\tau_n = n^{d_1}$ *and* $\lambda_n = n^{d_2}$, *where* $-\infty < d_1 < \infty$ *and* $d_2 \le 0$ *are two constants satisfying*

$$-c_0 < d_2 < -\frac{c_1}{2}, \qquad \frac{1}{2} + c_4 < d_1 + \max(0, c_1 + d_2) + c_3 < \frac{1 + c_2}{2}, \tag{4.3}$$

then we have

$$P\left(\widehat{\boldsymbol{\beta}}^n \text{ has the same sign as } \boldsymbol{\beta}^n\right) \ge 1 - O\left(n^{-\delta k}\right), \tag{4.4}$$
$$P\left(\widehat{\boldsymbol{\gamma}}^n \text{ has the same sign as } \boldsymbol{\beta}^n\right) \ge 1 - O\left(n^{-\delta k}\right),$$

where δ *is a positive constant only depending on* $c_0 \sim c_4$ *and* $d_1 \sim d_2$. *Moreover, both* $\widehat{\boldsymbol{\beta}}^n$ *and* $\widehat{\boldsymbol{\gamma}}^n$ *are consistent estimates of* $\boldsymbol{\beta}^n$. *That is,*

$$\|\widehat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}^n\|_2 \to 0, \qquad \|\widehat{\boldsymbol{\gamma}}^n - \boldsymbol{\beta}^n\|_2 \to 0, \tag{4.5}$$

in probability as $n \to \infty$.

In the following, we will prove the asymptotic consistency of coefficient curves for functional regression model (1.1) using our wavelet based CSR. Consider a sequence of functional linear regression models

$$y_i^n = \sum_{j=1}^{K_n} \int_0^1 Z_{ij}^n(t) g_j^n(t) dt + \widetilde{\epsilon}_i^n, \quad 1 \le i \le n, \tag{4.6}$$

where $\mathbf{Z}_i^n(t) = (Z_{i1}^n(t), \ldots, Z_{iK_n}^n(t))$, $1 \le i \le n$, are i.i.d. random vectors of curves, and $\widetilde{\epsilon}_i^n$, $1 \le i \le n$, are i.i.d. random variables. Then by wavelet transformation, (4.6) is converted to the following multiple linear regression model as in Section 2,

$$\mathbf{y}^n = \mathbf{X}^n \boldsymbol{\beta}^n + \boldsymbol{\varepsilon}^n, \tag{4.7}$$

where

$$\boldsymbol{\beta}^n = \left(\frac{1}{N_n}\mathbf{b}_1^{n\mathrm{T}}, \ldots, \frac{1}{N_n}\mathbf{b}_{K_n}^{n\ T}\right)^{\mathrm{T}}, \tag{4.8}$$

K_n and N_n are the numbers of curves and the number of points observed in each curve, respectively, $\mathbf{b}_j^n = (b_{j1}^{n\ T}, \ldots, b_{jN_n}^{n\ T})^T$ is the DWT of $(g_j^n(t_1^n), \ldots, g_j^n(t_{N_n}^n))$, and $(t_1^n, \ldots, t_{N_n}^n)$ are the equally spaced observation points.

The requirements that $n^{\frac{1-c_2}{2}} \min_{1\le j\le q} |\boldsymbol{\beta}_j^n| \ge M_3$ and $\sup_n E[(\varepsilon_i^n)^{2k}] < \infty$ in Condition 1 prevent the application of Theorem 4.1 to (4.7). The term $\boldsymbol{\varepsilon}^n$ in (4.7) is the sum of two components: $\widetilde{\epsilon}_i^n$, the noise term in the original functional regression model, and $\sum_{j=1}^{K_n} \varepsilon'_{ij}$ as defined in (2.4). The term $\sum_{j=1}^{K_n} \varepsilon'_{ij}$ captures the effect of discretization and is the error produced by replacing the integral in the functional regression model with the average of discretely observed values. To make the $2k$th moment of $\boldsymbol{\varepsilon}^n$ uniformly bounded, we have to control the magnitude of the term $\sum_{j=1}^{K_n} \varepsilon'_{ij}$ by choosing N_n, the number of the observation points, large enough. But, if N_n is too large, by the definition of $\boldsymbol{\beta}^n$ in (4.8), the magnitude of the nonzero coordinates will be so small that $n^{\frac{1-c_2}{2}} \min_{1\le j\le q} |\boldsymbol{\beta}_j^n| \ge M_3$ will not hold and it is difficult to select correct features and obtain consistent estimation. We impose the following regularity conditions.

Condition 2. 1. $\sup_n E[(\widetilde{\epsilon}_i^n)^{2k}] < \infty$, $\sup_n \max_{1\le j\le K_n} \sup_{0\le t\le 1} |g_j'(t)| < \infty$, *where* $g_j'(t)$ *is the derivative of* $g_j(t)$, *and* $\sup_n \max_{1\le j\le K_n} \sup_{0\le t\le 1} |g_j(t)| < \infty$.

2. $Z_{ij}(t),\ 1 \le j \le K_n$, *are Gaussian processes and satisfy*

$$\sup_n \sup_{j\in I^n} \sup_{|t-s|<\eta} E\left[\left(Z_{ij}^n(t) - Z_{ij}^n(s)\right)^2\right] = O(\eta^{r_1}), \tag{4.9}$$

where I^n *is the collection of* j*'s which satisfy* $g_j(t) \not\equiv 0$, *and* $1 < r_1 \le 2$.

3. *The number of nonzero* b_{jk}^n *is* $O(n^{c_1})$, $N_n \sim O(n^{r_2})$, $K_n \le O(n^{c_4 k - r_2})$ *and*

$$\min_{b_{jk}^n \ne 0} |b_{jk}^n| \ge M_3 n^{-\frac{1-c_2}{2}+r_2}, \qquad \sum_{j=1}^{K_n} \|\mathbf{b}_j^n\|_2^2 \sim n^{2c_3+2r_2},$$

where $r_2 > 0$ *and satisfies*

$$\frac{2c_3 + c_1}{r_1 - 1} < r_2 < 1 - c_2. \tag{4.10}$$

Here we consider the case that the predictive curves are Gaussian processes. The choice of N_n is related to the smoothness of the predictive curves. For smoother predictive curves, we can make sparser observations, otherwise denser observations are needed. The smoothness of the predictive curves is controlled by the asymptotic behavior of the covariance between the $Z_{ij}(t)$ and $Z_{ij}(s)$ as s approaches t. In (4.9), the larger is r_1, the smoother is Z_{ij}. In order that (4.10) is satisfied, when r_1 is small, we have to choose a large r_2, that is, choose a large N_n. Many commonly-used covariance functions for Gaussian processes satisfy (4.9) for different r_1 [25]. By (4.10), we need to choose a larger r_2 and hence a larger N_n for larger c_1 and c_3 which implies that the number and magnitude of true features are large. The right inequality in (4.10) implies that we cannot make N_n too large.

Theorem 4.2. *Under Condition 2, if we choose* $\tau_n = n^{d_1}$ *and* $\lambda_n = n^{d_2}$, *where* $-\infty < d_1 < \infty$ *and* $d_2 \le 0$ *are two constants satisfying*

$$-c_0 < d_2 < -\frac{c_1}{2}, \qquad \frac{1}{2} + c_4 < d_1 + \max(0, c_1 + d_2) + c_3 < \frac{1 + c_2}{2},$$

then we have

$$\sum_{j=1}^{K_n} \|\widehat{g}_j^n(t) - g_j^n(t)\|_2^2 \to 0 \tag{4.11}$$

in probability as $n \to \infty$.

5. Simulation studies

We study the performance of our method in two cases. In the first case, there is only one predictive curve and the combinations of different levels of smoothness of the predictive curve and coefficient curve will be considered. In the second case, there are multiple predictive curves for each sample and in addition to different levels of smoothness of the predictive curves and coefficient curves, we also consider different correlation patterns between the multiple predictive curves. Our method will be compared to various sparse regression methods and/or functional regression methods in different scenarios in both cases.

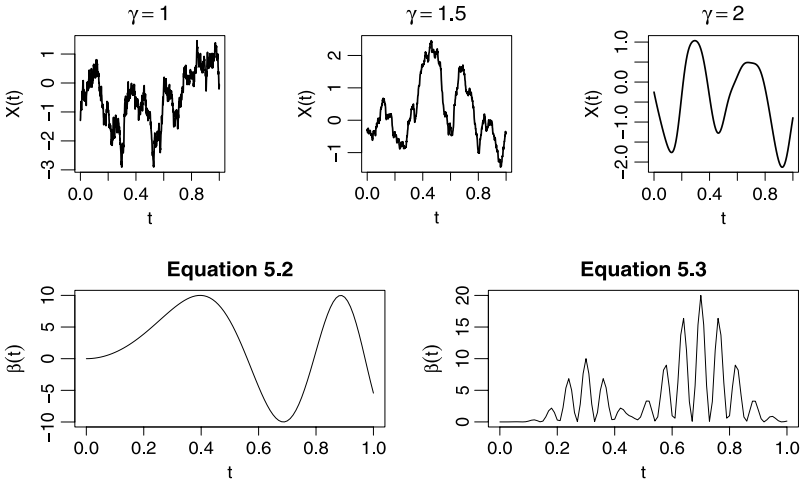

Fig. 1. The top panel is one sample of the predictive curve for different γ values. From left to right: $\gamma = 1, 1.5, 2$. The bottom panel shows the coefficient curves. Left: function (5.2); Right: function (5.3).

5.1. One predictive curve

The predictive curve $X(t)$ is a Gaussian process with a gamma-exponential covariance function [25]:

$$\mathrm{Cov}(X(t), X(s)) = e^{-(10|t-s|)^{\gamma}}, \quad 0 < \gamma \leq 2,\ 0 \leq t,\ s \leq 1, \tag{5.1}$$

where γ controls the smoothness of curves. We take $\gamma = 1, 1.5, 2$, respectively, in this study and one sample curve for each of the three values is drawn in the top panel of Fig. 1. Among the three, the curve is smoothest for $\gamma = 2$ and most wiggly for $\gamma = 1$. We investigate the following two coefficient functions:

$$\beta_{(1)}(t) = 10\sin(10t^2), \tag{5.2}$$

$$\beta_{(2)}(t) = 10[\cos\{50(t-0.3)\}]^2 e^{-100(t-0.3)^2} + 20[\cos\{50(t-0.7)\}]^2 e^{-50(t-0.7)^2}. \tag{5.3}$$

The two curves are plotted in the bottom panel of Fig. 1. It can be seen that $\beta_{(1)}(t)$ is smoother than $\beta_{(2)}(t)$ and $\beta_{(2)}(t)$ has more local variations than $\beta_{(1)}(t)$. Since in practice, we cannot simulate the values of $X(t)$ at all $0 \leq t \leq 1$, we simulate the predictive curves at 1024 equally spaced points, $0 = s_1 < s_2 < \cdots < s_{1024} = 1$, and use $\sum_{i=1}^{1024} X(s_i)\beta(s_i)/1024$ to approximate $\int_0^1 X(t)\beta(t)dt$. Finally, for each $X(t)$, we take its values at 128 equally spaced points, $0 = t_1 = s_1 < t_2 = s_9 < \cdots < t_{128} = s_{1024} = 1$, as the discrete observations and let $\widetilde{\epsilon}_i \sim N(0, 0.01^2)$. The training data size is 100 and test data size is 700. We carried out 100 simulations for each setting.

We compare our method, denoted by W-CSR (correlation-based sparse regression on wavelet coefficients), with several sparse regression methods: LASSO (W-Lasso) [28], elastic net (W-EN) [33], and sparse partial least squares regression (W-SPLS) [9] on 128 wavelet coefficients, and several functional regression methods: functional linear model with basis representation (FReg.basis) [24], functional principal component regression (FPCR) [7] and the functional partial least squares (FPLS) [22]. The LASSO and EN are implemented in the R package "glmnet", the SPLS in "spls", and the FReg.basis, FPC and FPLS in "fda.usc".

In our method, W-CSR, the tuning parameters are selected from the grid of $\tau = 0.01, 0.1, 1, 10, 100, 500, 1000$ and $\lambda = 0.001, 0.01, 0.1, 0.2, \ldots, 0.8, 0.9$. The EN regression method solves the penalized least squares problem

$$\max_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\left[\{(1-\alpha)/2\}\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1\right], \tag{5.4}$$

where $\lambda \geq 0$ and $0 \leq \alpha \leq 1$ are tuning parameters. Here we use the parameterization adopted in"glmnet". Note that the λ and α in the EN play the similar roles as τ and λ in our method. Given α, the EN solves (5.4) for all λ and a cross-validation procedure is available in "glmnet" to choose the best λ. For the purpose of comparison, we will consider the grid $\alpha = 0.001, 0.01, 0.1, 0.2, \ldots, 0.8, 0.9$. For each value of α in this grid, we use the CV procedure in "glmnet" to choose the optimal λ. Then the final choice of the tuning parameter is the α and the corresponding λ with the smallest prediction error on the validation set. For all the other methods, the default cross-validation methods are used to select the corresponding tuning parameters.

We summarize the averages and standard deviations of the mean squared prediction error (MSPE) in Table 1. Table 1 reveals that our method has lowest MSPE in all scenarios.

Table 1
The averages and standard deviations of the MSPE for simulation with $K = 1$. For each setting, the first row is the average of MSPE, and the second row is the standard deviation of MSPE. The second column of each competing method is the ratio of the average of MSPE of our method over that of the competing method.

$\beta(t)$	γ	W-CSR	W-SPLS		W-LASSO		W-EN		FReg.basis		FPC		FPLS	
$\beta_{(1)}(t)$	1	1.3e−2	2.0e−2	0.62	1.7e−2	0.74	1.8e−2	0.71	3.4e−2	0.37	3.3e−2	0.38	2.0e−2	0.62
		3.9e−3	5.2e−3		3.2e−3		3.4e−3		2.0e−3		6.1e−3		5.7e−3	
	1.5	3.9e−3	7.7e−3	0.51	7.7e−3	0.51	8.2e−3	0.48	1.8e−2	0.22	1.2e−2	0.32	8.9e−3	0.44
		9.4e−4	2.5e−3		1.4e−3		1.5e−3		1.2e−3		3.4e−3		1.7e−3	
	2	1.6e−4	1.8e−3	0.09	3.8e−3	0.04	3.8e−3	0.04	1.1e−2	0.01	2.6e−3	0.06	1.3e−3	0.13
		3.5e−5	5.7e−3		7.2e−4		6.7e−4		7.8e−4		2.4e−3		1.2e−3	
$\beta_{(2)}(t)$	1	9.0e−3	1.0e−2	0.87	1.1e−2	0.79	1.2e−2	0.73	1.9e−2	0.48	5.4e−2	0.16	9.2e−3	0.98
		1.6e−3	3.7e−3		2.1e−3		2.5e−3		1.3e−3		3.3e−2		2.2e−3	
	1.5	2.0e−3	5.5e−3	0.37	7.0e−3	0.29	7.8e−3	0.26	1.6e−2	0.13	3.3e−2	0.06	3.2e−3	0.63
		7.3e−4	2.6e−3		1.6e−3		2.0e−3		1.1e−3		2.7e−2		1.7e−3	
	2	3.4e−4	6.2e−3	0.05	6.7e−3	0.05	7.1e−3	0.05	1.6e−2	0.02	2.5e−2	0.01	3.2e−3	0.11
		1.2e−4	8.5e−3		1.5e−3		1.7e−3		1.1e−3		1.8e−2		2.9e−3	

5.2. Multiple predictive curves

In the second simulation study, there are ten ($K = 10$) predictive curves $X_k(t)$, $1 \le k \le 10$. The ten curves are all Gaussian processes with the same gamma-exponential covariance (that is, they have the same marginal distributions). We will consider $\gamma = 1, 1.5, 2$, respectively, and different correlation patterns between predictive curves. Similar to Section 5.1, for each predictive curve, we simulate its values at 1024 equally spaced points, $0 = s_1 < s_2 < \cdots < s_{1024} = 1$. Let $\mathbf{\Sigma}$ be the 1024×1024 covariance matrix of $(X_k(s_1), \ldots, X_k(s_{1024}))$ ($\mathbf{\Sigma}$ is the same for all $1 \le k \le 10$) and the (i, j)th element of $\mathbf{\Sigma}$ is equal to $\exp\{-(10|s_i - s_j|)^\gamma\}$, $1 \le i, j \le 1024$. To model the correlations between the 10 curves, we use the following procedure. Let $\mathbf{R}$ be a 10×10 matrix with the (k, l)th entry $\mathbf{R}_{k,l} = \rho^{|k-l|}$, which is the covariance between the $X_k(t)$ and $X_l(t)$ for any t, where $1 \le k, l \le 10$. Let Z be a 1024×10 matrix with each element independently having the standard normal distribution. Then the kth column of $\mathbf{\Sigma}^{1/2} Z \mathbf{R}^{1/2}$ is one realization of the vector $(X_k(s_1), \ldots, X_k(s_{1024}))$, where $\mathbf{\Sigma}^{1/2}$ and $\mathbf{R}^{1/2}$ are matrices satisfying $\mathbf{\Sigma}^{1/2}\left[\mathbf{\Sigma}^{1/2}\right]^T = \mathbf{\Sigma}$ and $\mathbf{R}^{1/2}\left[\mathbf{R}^{1/2}\right]^T = \mathbf{R}$, respectively, and $1 \le k \le 10$. It can be seen that if $\rho = 0$, then the 10 predictive curves are independent. Here we choose $\rho = 0.1, 0.5$ and 0.9 to represent the different correlation levels. Finally, for each $X(t)$, we take its values at 128 equally spaced points, $0 = t_1 = s_1 < t_2 = s_9 < \cdots < t_{128} = s_{1024} = 1$, as the discrete observations and let $\widetilde{\epsilon_i} \sim N(0, 0.01^2)$.

We take the following coefficient functions:

$$\beta_1(t) = 2\sin(10t^2), \tag{5.5}$$

$$\beta_2(t) = 2[\cos\{50(t - 0.3)\}]^2 e^{-100(t-0.3)^2} + 2[\cos\{50(t - 0.7)\}]^2 e^{-50(t-0.7)^2}, \tag{5.6}$$

$$\beta_3(t) = \begin{cases} 0 & \text{if } \dfrac{2i}{10} \le t < \dfrac{2i+1}{10}, \quad i = 0, 1, 2, 3, 4, \\ 1 & \text{if } \dfrac{2i+1}{10} \le t < \dfrac{2(i+1)}{10}, \quad i = 0, 1, 2, 3, 4 \end{cases} \tag{5.7}$$

$$\beta_j(t) = 0 \quad \text{for } j = 4, 5, \ldots, 10. \tag{5.8}$$

The functions (5.5) and (5.6) are similar to (5.2) and (5.3), respectively. The third function (5.7) is nonsmooth and piecewise constant.

The training data size is 60 and test data size is 740. We carried out 100 simulations for each setting. Since the R functions for FPC and FPLS cannot be applied to the case of multiple predictive curves and in such case, and there is no proper cross validation method available for FReg.basis, we compare our method with other wavelet-based sparse methods. From Table 2, we also observe the good performance of our method.

6. Case studies

In this section, we consider four different data sets. There is one predictive curve in the first data set, and multiple in others.

6.1. Near-infrared spectroscopy data

This data set was collected from an experiment to test the feasibility of near-infrared (NIR) spectroscopy in measuring the composition of biscuit dough pieces [6,20]. The purpose of the study is to predict the percentage of each of the four

Table 2
The averages and standard deviations of the MSPE for simulation with $K = 10$. For each setting, the first row is the average of MSPE, and the second row is the standard deviation of MSPE. The second column of each competing method is the ratio of the average of MSPE of our method over that of the competing method.

ρ	γ	W-CSR	W-SPLS		W-LASSO		W-Ridge		W-EN	
0.1	1	0.047	0.095	0.49	0.120	0.39	0.521	0.09	0.134	0.35
		0.016	0.066		0.047		0.031		0.057	
	1.5	0.044	0.118	0.38	0.133	0.33	0.534	0.08	0.154	0.29
		0.016	0.087		0.062		0.030		0.076	
	2	0.042	0.090	0.47	0.072	0.58	0.581	0.07	0.083	0.51
		0.013	0.059		0.042		0.043		0.049	
0.5	1	0.039	0.120	0.33	0.118	0.33	0.645	0.06	0.125	0.31
		0.018	0.045		0.037		0.041		0.038	
	1.5	0.035	0.127	0.27	0.126	0.28	0.645	0.05	0.133	0.26
		0.014	0.046		0.035		0.045		0.038	
	2	0.035	0.103	0.34	0.069	0.50	0.702	0.05	0.076	0.46
		0.014	0.045		0.039		0.045		0.041	
0.9	1	0.030	0.068	0.44	0.049	0.60	0.709	0.04	0.051	0.58
		0.009	0.027		0.019		0.058		0.020	
	1.5	0.025	0.067	0.37	0.052	0.48	0.695	0.04	0.055	0.45
		0.010	0.026		0.018		0.048		0.019	
	2	0.024	0.056	0.42	0.026	0.90	0.668	0.04	0.028	0.85
		0.014	0.026		0.009		0.078		0.010	

Table 3
The averages and standard deviations of the MSPE for the NIR spectroscopy data. For each model shown in column 1, the first row is the average of MSPE, and the second row is the standard deviation of MSPE. The second column of each competing method is the ratio of the average of MSPE of our method over that of the competing method.

Model	W-CSR	W-SPLS		W-LASSO		W-Ridge		W-EN		FReg.basis		FPC		FPLS	
Fat $\sim X$	0.045	0.058	0.78	0.050	0.90	0.496	0.09	0.050	0.90	0.064	0.71	0.042	1.08	0.034	1.32
	0.013	0.022		0.013		0.090		0.014		0.054		0.010		0.009	
Fat $\sim X'$	0.056	0.066	0.84	0.055	1.01	0.206	0.27	0.055	1.01	0.048	1.16	0.050	1.12	0.035	1.58
	0.019	0.021		0.016		0.048		0.019		0.020		0.013		0.009	
Fat $\sim X''$	0.082	0.091	0.91	0.082	1.01	0.240	0.34	0.079	1.04	0.065	1.27	0.296	0.28	0.077	1.07
	0.024	0.030		0.024		0.060		0.025		0.042		0.127		0.025	
Sucr $\sim X$	0.173	0.199	0.87	0.207	0.83	0.272	0.64	0.206	0.84	0.222	0.78	0.196	0.88	0.206	0.84
	0.053	0.064		0.053		0.055		0.046		0.112		0.093		0.088	
Sucr $\sim X'$	0.182	0.194	0.94	0.203	0.90	0.203	0.90	0.215	0.85	0.237	0.77	0.212	0.86	0.237	0.77
	0.071	0.079		0.071		0.037		0.061		0.106		0.097		0.110	
Sucr $\sim X''$	0.162	0.209	0.77	0.179	0.90	0.205	0.79	0.198	0.82	0.323	0.50	0.188	0.86	0.201	0.80
	0.060	0.071		0.052		0.038		0.048		0.128		0.048		0.063	
Flour $\sim X$	0.162	0.174	0.93	0.228	0.71	0.227	0.71	0.225	0.72	0.200	0.81	0.225	0.72	0.237	0.68
	0.083	0.101		0.078		0.060		0.073		0.117		0.102		0.096	
Flour $\sim X'$	0.163	0.172	0.95	0.219	0.74	0.187	0.87	0.208	0.78	0.226	0.72	0.209	0.78	0.240	0.68
	0.102	0.110		0.089		0.053		0.080		0.147		0.126		0.132	
Flour $\sim X''$	0.164	0.186	0.88	0.216	0.76	0.193	0.85	0.207	0.79	0.294	0.56	0.176	0.93	0.226	0.73
	0.082	0.108		0.075		0.052		0.075		0.124		0.066		0.099	
Water $\sim X$	0.138	0.156	0.88	0.154	0.90	0.214	0.64	0.152	0.90	0.200	0.69	0.173	0.79	0.175	0.79
	0.066	0.072		0.046		0.059		0.048		0.086		0.071		0.068	
Water $\sim X'$	0.137	0.141	0.97	0.149	0.92	0.149	0.92	0.151	0.90	0.218	0.63	0.161	0.85	0.174	0.79
	0.068	0.066		0.068		0.057		0.065		0.103		0.088		0.082	
Water $\sim X''$	0.140	0.149	0.94	0.156	0.90	0.162	0.86	0.153	0.91	0.244	0.57	0.193	0.72	0.168	0.83
	0.075	0.076		0.064		0.063		0.058		0.190		0.088		0.087	

constituents: *fat, sucrose, dry flour, and water*, based on the NIR reflectance spectrum. There were 72 dough pieces measured. An NIR reflectance spectrum, denoted by $X(t)$, is available for each dough piece. The original spectral data consists of 700 points measured from 1100 to 2498 nanometers (nm) in steps of 2 nm. The first 140 and last 49 wavelengths, which were thought to contain little useful information, were removed, and we use the remained 512 points for wavelet transformation. To be comparable, the Freg.basis, FPCR and FPLS also use these 512 observed points for prediction. The data set is available in the R package "ppls". In this example, we will consider 12 different functional regression models, which take one of the four variables, *fat, sucrose, dry flour, water*, as the response and one of the three curves, $X(t), X'(t), X''(t)$, as the predictive

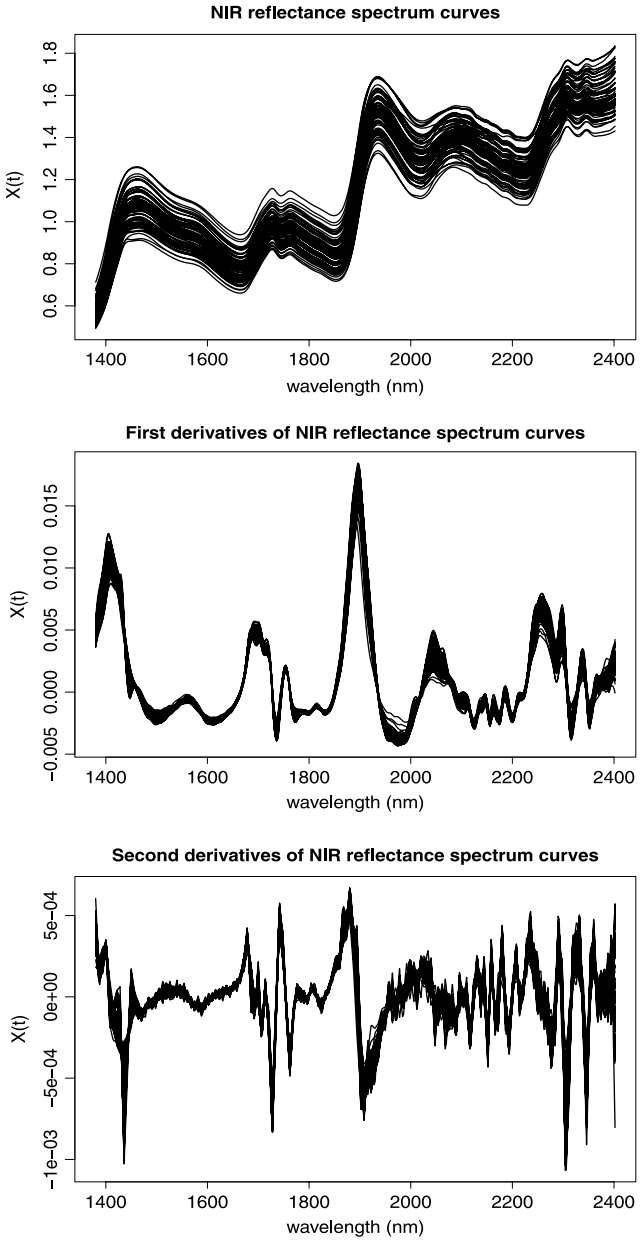

Fig. 2. $X(t)$, $X'(t)$, $X''(t)$ for all 72 observed curves in near-infrared spectroscopy data.

curve. Here $X'(t)$ and $X''(t)$ are the first and second derivatives of $X(t)$. We plot $X(t)$, $X'(t)$ and $X''(t)$ for all 72 observed curves in Fig. 2. As can be seen, $X(t)$ is relatively smooth, and $X'(t)$ and $X''(t)$ have more local variations.

We randomly split the 72 observations into a training set with 39 observations and a test set with 33 observations. We apply all the methods and the corresponding cross-validation procedures to the training set to select tuning parameters and build the predictive models. For the wavelet based sparse regression methods, we use 512 wavelet basis functions and transform the functional data set to a 72×512 multivariate data set. Then all the models are applied to the test data to compute the MSPE. We repeat the procedure 100 times and calculate the averages and standard deviations of the 100 MSPEs for each model. The results are summarized in Table 3. Our method has lowest predictive errors in all the cases except when the response is fat. To compare the estimates of the coefficient curve $g(t)$ in (1.1), we plot the estimated curves of all the methods. The estimated coefficient curves for the model, Fat $\sim X(t)$ and Sucrose $\sim X(t)$, are presented in Figs. 3 and 4, respectively. The curves for other response variables are available in the supplementary materials (see Appendix A). From Fig. 3, it can be seen that our method, W-LASSO, W-EN and W-SPLS all capture a common local feature lying around the wavelength 1700 nm. Since the ridge regression does not make feature selection, no clear local feature can be found. The curve in FReg.basis is the most smooth, and those for FPCR and FPLS have similar patterns and their largest peaks also

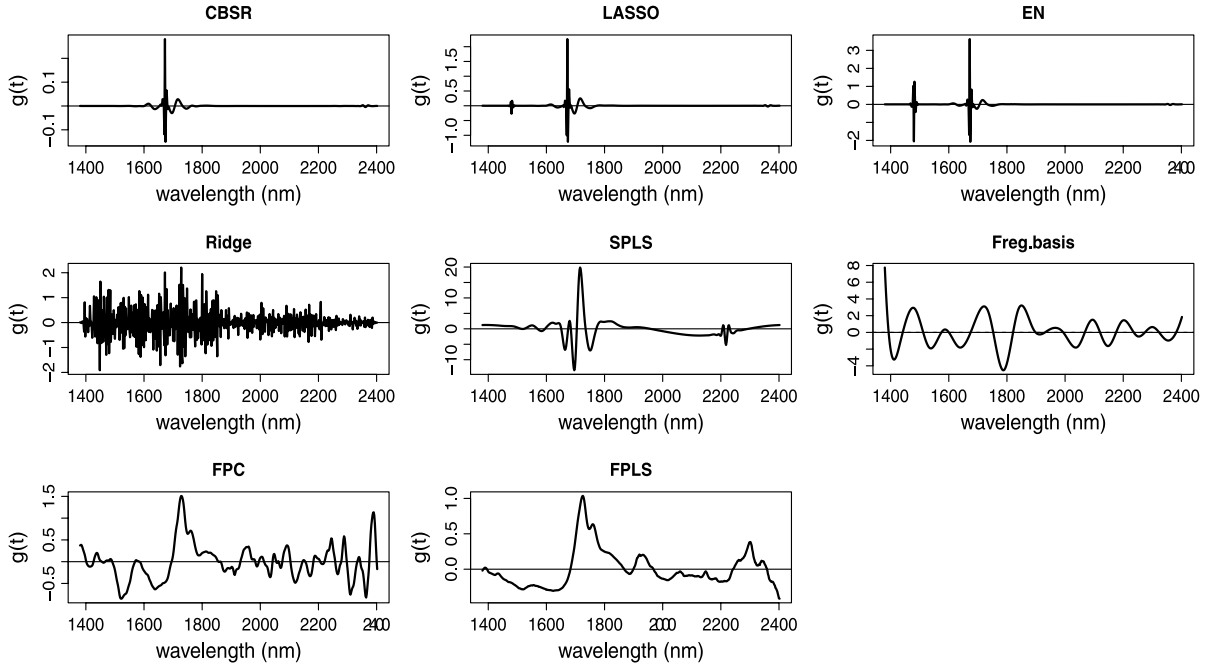

Fig. 3. The estimated coefficient function $\hat{g}(t)$ from all methods for the functional regression model: Fat $\sim X(t)$, in near-infrared spectroscopy data.

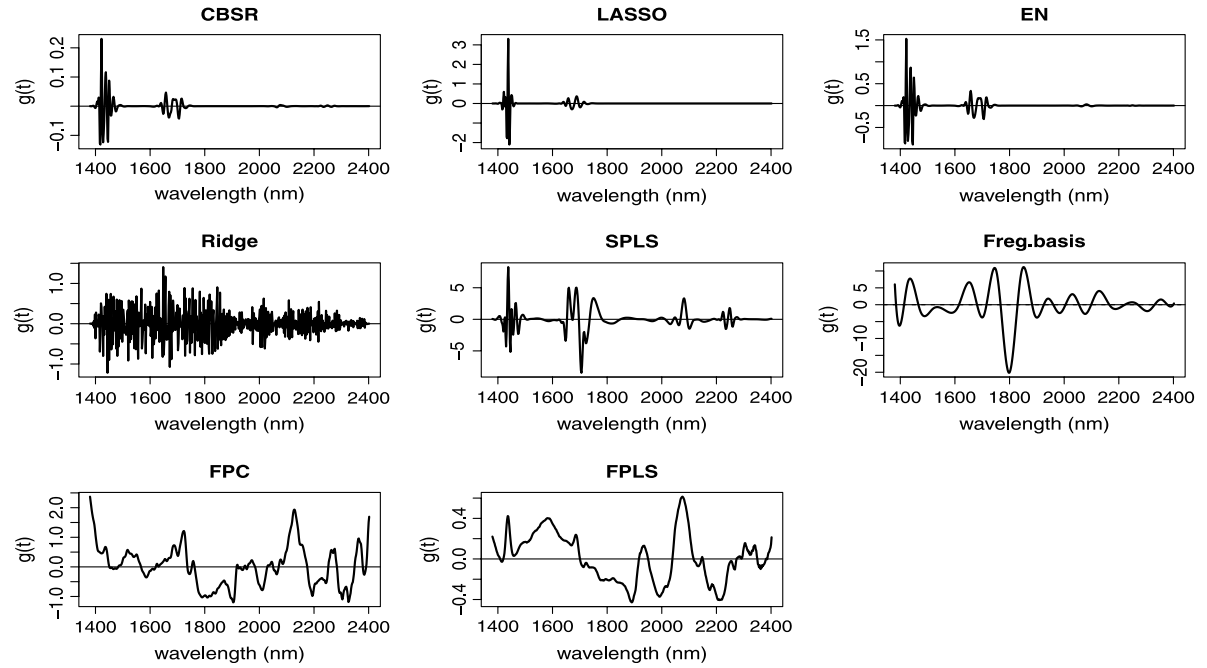

Fig. 4. The estimated coefficient function $\hat{g}(t)$ from all methods for the functional regression model: Sucrose $\sim X(t)$, in near-infrared spectroscopy data.

appear around the wavelength 1700 nm. From Fig. 4, we see that the percentage of sucrose is more related to the local features around the wavelength 1400 nm.

6.2. CT slice data

The data was retrieved from Computed Tomography (CT) slice images from different patients [13], and available in UCI Machine Learning Repository [3]. The purpose of the study is to construct a predictive model to determine relative locations of CT slices on axial axis (for example, the slice is on the top or bottom of head) based on the distributions of bone structures

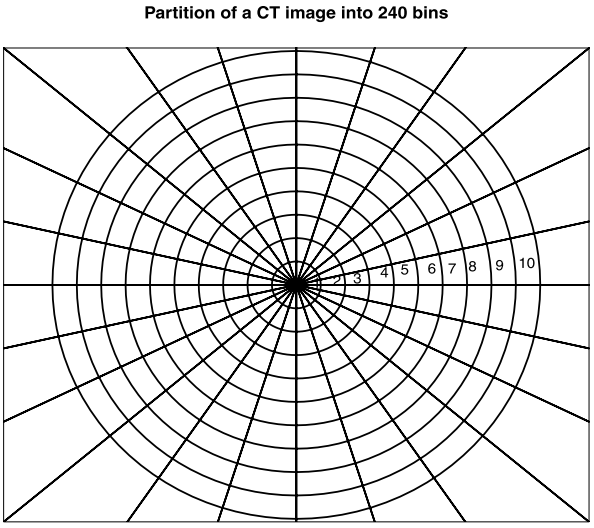

Fig. 5. Partition of a CT image into 240 bins by using 24 circle sectors and 10 shells, where the 10 bins for the first sector are labeled.

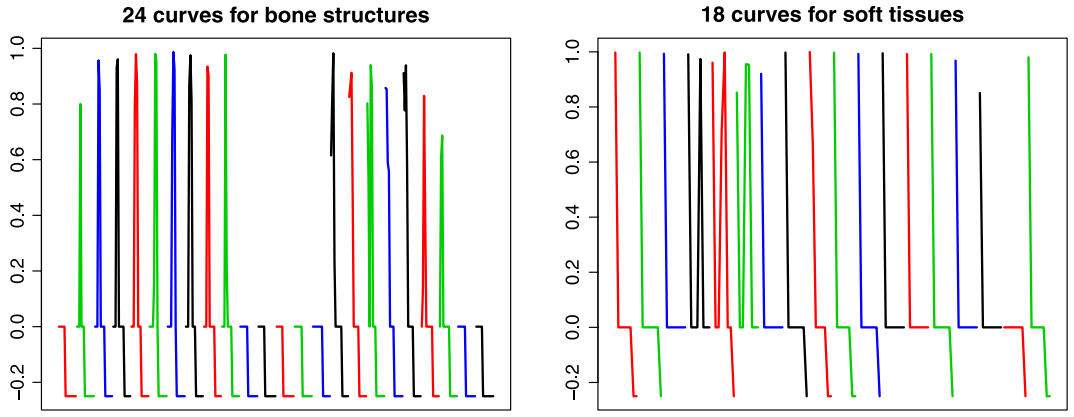

Fig. 6. The distribution curves of bone structures and soft tissues for one slice image in CT slice data. (Colorful curves are drawn in the web version of this article.)

and soft tissues in the CT slices image. The response variable is the relative position ($\in [0, 1]$) of the slice in the body region from the top of the head to the end of the coccyx. In [13], the distributions of bone structures and soft tissues in an image were extracted in the following way. In each image, the polar coordinate system is used to partition the image into 24 circle sectors based on the angular coordinate with central angle 15° in each sector as in Fig. 5. Each sector is further partitioned into 10 bins by 10 shells (the 10 bins for the first sector are labeled in Fig. 5). One measurement of bone structures is obtained from each bin. Hence, the 10 observations in each circle sector can be regarded as the discrete observations of the distribution curve of bone structures along the direction of this sector. Totally there are 24 such discretely observed curves for each CT slice image.

We plot the 24 curves for one CT slice image in Fig. 6, where in order to make a more clear presentation, we concatenate the curves together and used different colors for different curves. The value -0.25 means the corresponding bins are outside the region of the patient's body in the image. To facilitate the wavelet transformation, we add 6 additional values all equal to -0.25 after the ten values for each curve. The similar procedure was used to extract the distribution of soft tissues as the second predictor, but 18 sectors and 8 shells were used. The 18 curves for soft tissues are plotted in Fig. 6. The original data contains 53 500 CT images from 74 different patients. We consider the data for the first subject which contain 583 slice images in various positions. We consider three regression models. In the first model, the 24 curves (16 basis functions for each curve) for bone structures are used as the predictors, in the second model, the 18 curves (8 basis functions for each curve) for soft tissues are used as the predictors, and in the third model, all the 42 curves are used.

We randomly split the 538 observations into a training set with 100 observations and a test set with 438 observations. The procedure is repeated 100 times and the averages and standard deviations of the MSPEs are reported in Table 4. In all the three models, our method have the smallest prediction errors. To compare the estimates of the coefficient curves $g_i(t)$ in (1.1), we plot the estimated curves in one repeat for the first and second models in Figs. 7 and 8, respectively, where different colors represent the coefficients for different predictive curves.

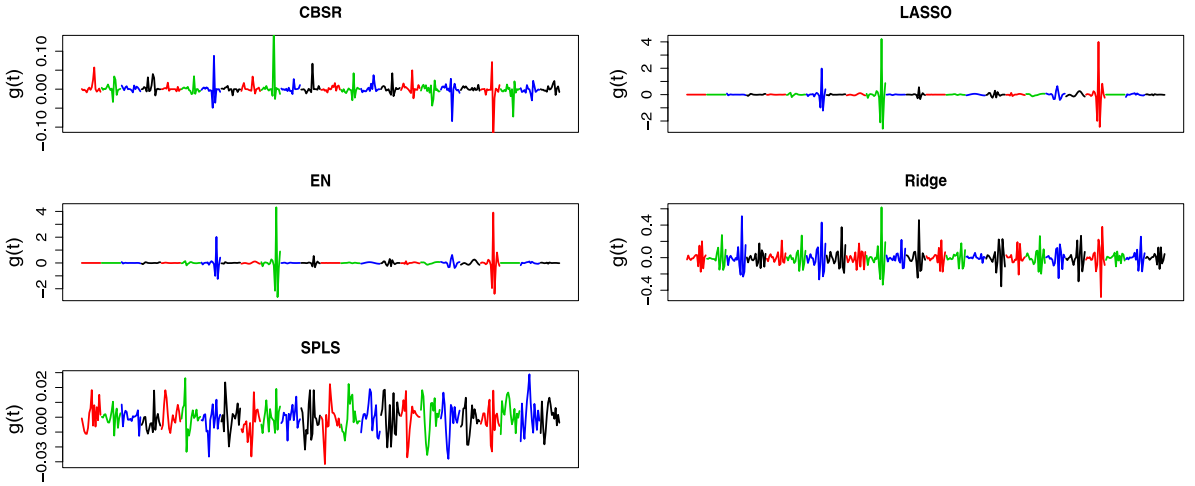

Fig. 7. The coefficient curves for the model with the 24 distribution curves of bone structures as predictive curves in CT slice data. (Colorful curves are drawn in the web version of this article.)

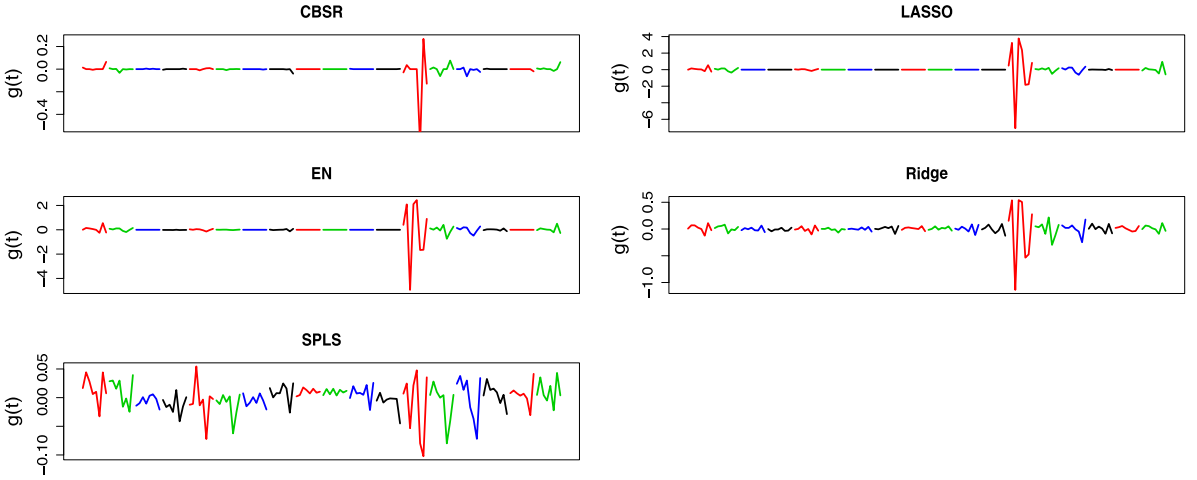

Fig. 8. The coefficient curves for the model with the 18 distribution curves soft tissues as predictive curves in CT slice data. (Colorful curves are drawn in the web version of this article.)

Table 4
The averages and standard deviations of the MSPE for the CT image data. For each model shown in column 1, the first row is the average of MSPE, and the second row is the standard deviation of MSPE. The second column of each competing method is the ratio of the average of MSPE of our method over that of the competing method.

Model	W-CSR	W-SPLS		W-LASSO		W-Ridge		W-EN	
$Y \sim$ bone structures	0.010 0.005	0.017 0.006	0.59	0.016 0.007	0.60	0.061 0.009	0.16	0.016 0.007	0.60
$Y \sim$ soft tissues	0.035 0.009	0.044 0.007	0.79	0.040 0.011	0.88	0.19 0.021	0.19	0.037 0.010	0.95
$Y \sim$ both	0.0075 0.0027	0.0107 0.0024	0.70	0.0130 0.0050	0.57	0.0325 0.0052	0.23	0.0107 0.0034	0.70

6.3. EEG data

This data set is available in UCI Machine Learning Repository [3]. It was collected from a large study to examine EEG correlates of genetic predisposition to alcoholism [30]. It contains measurements from 64 electrodes placed on subjects' scalps which were sampled at 256 Hz (3.9-ms epoch) for 1 s. There were two groups of subjects: alcoholic and control. In each group, there were 10 subjects and for each subject, 30 trials were performed. Hence, the sample size is 600 and in each sample, 64 curves were observed simultaneously. The response variable is categorical and has two levels: "alcoholic"

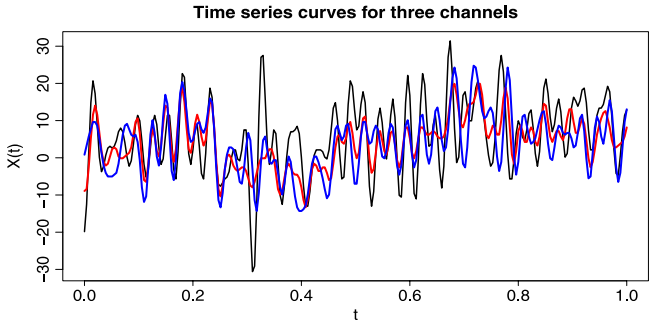

Fig. 9. Three of the 64 curves for one trial of one subject in EEG data.

Table 5
The averages and standard deviations of the misclassification rates for the EEG data. For each model shown in column 1, the first row is the average of misclassification rates, and the second row is the standard deviations. The second column of each competing method is the ratio of the average of misclassification rates of our method over that of the competing method.

Model	W-CSR	W-SPLS		W-LASSO		W-Ridge		W-EN	
A single stimulus	0.16 0.05	0.19 0.055	0.84	0.20 0.048	0.82	0.21 0.055	0.77	0.20 0.05	0.81
Two stimuli	0.22 0.03	0.28 0.03	0.79	0.31 0.04	0.71	0.29 0.03	0.76	0.29 0.04	0.76
All trials	0.178 0.025	0.233 0.029	0.76	0.256 0.032	0.70	0.243 0.026	0.73	0.237 0.031	0.75

and "control". Hence, it is a binary classification problem. We use the wavelet based sparse regression methods to perform classification as in [33]. Specifically, we define the response variable $y = 1$ if the corresponding subject is alcoholic and $y = -1$ if the subject is in the control group. For each subject, 64 curves were measured simultaneously. In Fig. 9, we plot three of the 64 curves for one trial of one subject. All the three curves show many local variations and are positively correlated.

For each curve, we use 256 wavelet basis functions to transform it and hence, we obtain a vector of $256 \times 64 = 16\,384$ coefficients as new predictors. The sparse regression methods are used to construct predictive regression models. If the predicted value of an observation is negative, it will be assigned to the alcoholic group, otherwise, it will be assigned to the control group. The cross validation procedure for MSPE is not appropriate in this example, instead we randomly split the data set into three subsets: the training, validation and test data. For each method, the model is built based on the training data, the tuning parameters are selected by using the validation data and minimizing the misclassification rate and the final model is applied to the test data to obtain the misclassification rate. In each trial, the subject was exposed to either a single stimulus or to two stimuli, hence we will consider three situations. In the first situation, only the trials with a single stimulus are considered. The sample size is 200, and we use 100, 50 and 50 observations in the training, validation and test subsets, respectively. In the second situation, only the trials with two stimuli are considered. The sample size is 400, and there are 100, 100 and 200 observations in the training, validation and test subsets, respectively. In the last situation, all the trials are considered. The sample size is 600, and there are 200, 200 and 200 observations in the training, validation and test subsets, respectively. For each situation, the procedure is repeated 100 times and the averages and standard deviations of misclassification rates are reported in Table 5. Our methods has the smallest misclassification rates in all three situations.

6.4. Daily and sports activities data

This data, available in UCI Machine Learning Repository, records several daily and sports activities each performed by 8 subjects in their own style for 5 min [2,5,1]. The 5 min signals were divided into 5 s segments, so that $60 \times 8 = 480$ signal segments were obtained for each activity and will be treated as replicates for that activity. For each activity, nine sensors (x, y, z accelerometers, x, y, z gyroscopes, x, y, z magnetometers) were placed on each of five body parts: torso, right arm, left arm, right leg, left leg, and were calibrated to acquire data at 25 Hz sampling frequency. So for each signal segment, there were 45 curves with $25 \times 5 = 125$ sample points in each curve. The purpose of the study is to classify the activities based on the curves recorded by the 45 sensors. Since in this paper, we focus on the regression methods, we will consider several pairs of activities and apply the regression methods to each pair for classifications. We extract the frequency curve of each of the 45 curves by the Fast Fourier Transformation and apply the wavelet based sparse regression methods to these frequency curves. Considering the periodicity of the frequency curves, we use the first 64 frequency points and hence 64 wavelet basis functions to do transformation. Hence, for each observation, a vector with 2880 wavelet coefficients is obtained as the predictors. We consider the following three pairs of activities: sitting vs. standing, lying on back vs. on

Table 6
The averages and standard deviations of the misclassification rates for the Daily and Sports Activities Data. For each model shown in column 1, the first row is the average of misclassification rates, and the second row is the standard deviations.

Model	W-CSR	W-SPLS	W-LASSO	W-Ridge	W-EN
Sitting vs. standing	0	0.0085	0.0004	0.055	0.0067
	0	0.011	0.0008	0.0287	0.0067
Lying on back vs. on right side	0	0.0017	0.0019	0.0031	0.0013
	0	0.003	0.002	0.009	0.0020
Walking on a treadmill in flat vs. 15°	0.015	0.027	0.027	0.046	0.023
	0.017	0.024	0.023	0.031	0.022

right side, and walking on a treadmill with a speed of 4 km/h in flat vs. 15° inclined positions. For each pair, we randomly choose the training, validation and test subsets with 50, 100 and 810 observations, respectively. The procedure is repeated 100 times and the averages and standard deviations of misclassification rates are reported in Table 6. Our method has the smallest misclassification rates in all three situations.

7. Discussion

In this paper, functional regression models with a scalar response and multiple predictive curves are considered. We transform the functional regression models to multiple linear regression models by using the discrete wavelet transformation. When the number of predictive curves is big, the multiple linear regression model usually has much bigger number of features than the sample size. Our correlation-based sparse regression method is applied to the resulted high dimensional regression model. The novel feature of our sparse method is that we impose sparsity penalty on the direction of the estimate of the coefficient vector instead of the estimate itself, and only the direction of the estimate is determined by optimization problem. The estimation consistency of the coefficient curve for the functional regression model is obtained when both the sample size and the number of curves go to infinity. The effects of the discrete observations are discussed. We compare our method with both functional regression methods and other wavelet based sparse regression methods in simulation studies and on four real data sets, including the cases of single and multiple predictive curves. The results indicate that sparse wavelet regression methods are better in extracting local features and our method has good predictive performances in all the scenarios.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2014.10.003.

References

[1] K. Altun, B. Barshan, Human activity recognition using inertial/magnetic sensor units, in: Human Behavior Understanding, Springer, 2010, pp. 38–51.
[2] K. Altun, B. Barshan, O. Tunçel, Comparative study on classifying human activities with miniature inertial and magnetic sensors, Pattern Recognit. 43 (2010) 3605–3620.
[3] K. Bache, M. Lichman, UCI Machine Learning Repository, 2013. URL: http://archive.ics.uci.edu/ml.
[4] E. Bair, T. Hastie, D. Paul, R. Tibshirani, Prediction by supervised principal components, J. Amer. Statist. Assoc. 101 (2006) 119–137.
[5] B. Barshan, M.C. Yüksek, Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units, Comput. J. (2013) bxt075.
[6] P. Brown, T. Fearn, M. Vannucci, Bayesian wavelet regression on curves with application to a spectroscopic calibration problem, J. Amer. Statist. Assoc. 96 (2001) 398–408.
[7] H. Cardot, F. Ferraty, P. Sarda, Functional linear model, Statist. Probab. Lett. 45 (1999) 11–22.
[8] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statist. Sinica 13 (2003) 571–592.
[9] H. Chun, S. Keles, Sparse partial least squares regresssion for simultaneous dimension reduction and variable selection, J. R. Stat. Soc. 72 (2010) 3–25.
[10] I. Daubechies, et al., Ten Lectures on Wavelets, vol. 61, SIAM, 1992.
[11] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, D. Picard, Wavelet shrinkage: asymptopia? J. R. Stat. Soc. Ser. B (1995) 301–369.
[12] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[13] F. Graf, H.P. Kriegel, S. Pölsterl, M. Schubert, A. Cavallaro, 2D image registration in ct images using radial image, in: Proceedings of the 14th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI, Toront, CA, 2011.
[14] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.
[15] I.M. Johnstone, A.Y. Lu, On consistency and sparsity for principal components analysis in high dimensions, J. Amer. Statist. Assoc. 104 (486) (2009) 682–693.
[16] N. Krämer, A.-L. Boulesteix, G. Tutz, Penalized partial least squares with applications to *b*-spline transformations and functional data, Chemometr. Intell. Lab. Syst. 94 (2008) 60–69.
[17] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 674–693.
[18] B.D. Marx, P.H. Eilers, Generalized linear regression on sampled signals and curves: a *p*-spline approach, Technometrics 41 (1999) 1–13.
[19] B. Nadler, Finite sample approximation results for principal component analysis: a matrix perturbation approach, Ann. Statist. 36 (2008) 2791–2817.
[20] B.G. Osborne, T. Fearn, A.R. Miller, S. Douglas, Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs, J. Sci. Food Agric. 35 (1984) 99–105.
[21] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, Statist. Sinica 17 (2007) 1617–1642.
[22] C. Preda, G. Saporta, PLS regression on a stochastic process, Comput. Statist. Data Anal. 48 (2005) 149–158.
[23] X. Qi, R. Luo, R.J. Carroll, H. Zhao, Sparse regression by projection and sparse discriminant analysis, J. Comput. Graph. Statist. (2014) in press, URL: http://dx.doi.org/10.1080/10618600.2014.907094.

[24] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, second ed., Springer, New York, 2005.
[25] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, in: Adaptive Computation and Machine Learning Series, The MIT Press, 2005, (Chapter 4).
[26] P.T. Reiss, R.T. Ogden, Functional principal component regression and functional partial least squares, J. Amer. Statist. Assoc. 102 (2007) 984–996.
[27] M. Smith, T.P. Barnwell, Exact reconstruction techniques for tree-structured subband coders, IEEE Trans. Acoust. Speech Signal Process. 34 (1986) 434–441.
[28] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[29] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2010) 894–942.
[30] X.L. Zhang, H. Begleiter, B. Porjesz, W. Wang, A. Litke, Event related potentials during object recognition tasks, Brain Res. Bull. 38 (1995) 531–538.
[31] Y. Zhao, R.T. Ogden, P.T. Reiss, Wavelet-based Lasso in functional linear regression, J. Comput. Graph. Statist. 21 (2012) 600–617.
[32] P. Zhao, B. Yu, On model selection consistency of Lasso, J. Mach. Learn. Res. 7 (2006) 2541–2563.
[33] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B 67 (2005) 301–320.