Journal of Multivariate Analysis 168 (2018) 105–118

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Sparse estimation for functional semiparametric additive models

Peijun Sang, Richard A. Lockhart, Jiguo Cao *

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A1S6

ARTICLE INFO

Article history:
Received 5 January 2018
Available online 11 July 2018

MSC:
62G08

Keywords:
Functional data analysis
Functional linear model
Functional principal component analysis

ABSTRACT

We propose a functional semiparametric additive model for the effects of a functional covariate and several scalar covariates and a scalar response. The effect of the functional covariate is modeled nonparametrically, while a linear form is adopted to model the effects of the scalar covariates. This strategy can enhance flexibility in modeling the effect of the functional covariate and maintain interpretability for the effects of scalar covariates simultaneously. We develop the method for estimating the functional semiparametric additive model by smoothing and selecting non-vanishing components for the functional covariate. Asymptotic properties of our method are also established. Two simulation studies are implemented to compare our method with various conventional methods. We demonstrate our method with two real applications.

1. Introduction

High-dimensional data sets of large volume and complex structure are rapidly emerging in various fields. Functional data analysis, due to its great flexibility and wide applications in dealing with high-dimensional data, has received considerable attention. One important problem in functional data analysis is functional linear regression (FLR). One type of FLR models the relationship between a functional covariate and a univariate scalar response of interest. Due to potential lack of fit with FLR models, [24] proposed functional additive models (FAM), in which a scalar response depends on an additive form of the functional principal component (FPC) scores of a functional covariate. A local linear smoother was employed to estimate each component in the additive form and consistency was established for this estimator.

However, in many cases, not only functional covariates but also some scalar covariates may play a role in explaining variation of response. For instance, the Tecator dataset (see Section 4.1 for a more detailed description), which consists of three contents (fat, water, and protein) and 100-channel spectral trajectories of absorbance, has been analyzed with various models, where the response of interest is one of the three contents. Previous studies have focused on regressing the response on the spectral trajectories, which can be viewed as a functional covariate. Zhu et al. [36], for example, employed a regularized functional additive model, where scaled FPC scores are treated as covariates to predict the protein content. However, pairwise scatter plots of the three contents suggest that the other two contents are highly correlated with the protein content as well; thus it may be beneficial to add them into the regression model. In light of this fact, we aim to build a model which can incorporate the effects of both the spectral trajectories and the fat and water contents on the prediction of the protein content.

Motivated by the above example, we propose a functional semiparametric additive model (FSAM) to describe the relationship between a functional covariate, a finite number of scalar covariates and a response variable of interest. In this

* Corresponding author.
E-mail address: jiguo_cao@sfu.ca (J. Cao).

https://doi.org/10.1016/j.jmva.2018.06.010
0047-259X/

model, the effect of a functional covariate is represented by its scaled leading FPC scores while scalar covariates are modeled linearly. As a result, this model enables us to acquire flexibility in calibrating the effect of the functional covariate while retaining easy interpretation of the effects of the scalar covariates. There are two main difficulties associated with this new model: the first one is the model estimation and the second concerns theoretical properties. Obviously, the estimation of the effect of a functional covariate may affect that of scalar covariates and vice versa. To address this issue, we propose an iterative updating algorithm, which is similar in spirit to the EM algorithm, to account for the interdependence between these two estimated effects. In addition, only the nonparametric effect of the functional covariate needs to be regularized; this adds additional difficulties in estimation. In the theoretical aspect, we aim to establish consistency for the parametric part and the nonparametric part, respectively. Separating these two effects is more difficult than developing theoretical properties with only a nonparametric part as in a FAM.

A semiparametric additive model (sometimes described under alternative names like partially linear model) can be viewed as a special version of a generalized additive model in which the mean response is assumed to have a linear relationship with one or more of the covariates, but the relation with other covariates cannot be easily modeled in a parametric form [26,29]. Numerous methods have been proposed to fit such models. The method of penalized least squares [7,17,32] has played a major role in this regard. Chen et al. [4] employed a piecewise polynomial to approximate the nonparametric part and developed asymptotic properties of the least squares estimator of the coefficients in the parametric part. Fan and Gijbels [8] estimated the nonparametric part using a local polynomial and derived asymptotic properties of their estimators as well. A comprehensive review of different approaches to fitting a semiparametric additive model can be seen in [15].

For the case when both a functional covariate and scalar covariates are involved to predict the mean response, Shin [28] considered a functional partially linear model in which the effect of a functional covariate is modeled via a finite-dimensional linear combination of principal component scores. A similar model was proposed by Lu et al. [22] to model the quantile function of the response variable. Even though both papers derived asymptotic properties of their estimators, they did not consider selection of functional principal components for the functional covariate. Kong et al. [19] extended the above work to the situation when multiple functional covariates and high-dimensional scalar covariates are encountered. The effect of each functional covariate is represented via a truncated linear combination of FPC scores, the truncation level of which is allowed to increase as sample size increases. To identify important features, reduce variability and enhance interpretability, they proposed to combine regularization of each functional covariate with a penalty on high-dimensional scalar covariates. Ivanescu et al. [18] proposed a general regression framework which considered a functional response and two functional covariates.

This article has three main contributions. First, in comparison with previous work on functional partial linear regression, our model allows for a more general representation in terms of the effect of a functional covariate. In addition, using a special regularization scheme, our method can select the non-vanishing functional principal components for a functional covariate. Last but not least, we derive asymptotic properties of the estimator.

The remainder of this paper is organized as follows. Section 2 introduces FSAM and our method to estimate FSAM using a special regularization scheme and implementing an iterative updating algorithm. Section 3 evaluates the finite-sample performance of our proposed estimation method in comparison with three alternative methods using some simulation studies. In Section 4, our method is demonstrated in two real examples. Some asymptotic results for the proposed estimation method are provided in Section 5. Section 6 concludes this article. Additional results of numerical studies and theoretical proofs are given in the Online Supplement.

2. Model and estimation method

2.1. Functional semiparametric additive model

Let $X(t)$ denote a square integrable stochastic process on a domain $\mathcal{I} = [0, T]$ and Y denote a scalar random variable. A functional regression model characterizes the relationship between the scalar response Y and the random function $X(t)$. A typical example is the functional linear model: $Y = \int_{\mathcal{I}} X(t)\beta(t)\mathrm{d}t$, where $\beta(t)$ is a square integrable function on $[0, T]$ as well.

To account for the effect of some scalar predictors in a functional regression model, several functional partial linear models have been proposed; see [19,22,28]. In these papers, the effect of the functional predictor is modeled nonparametrically while a linear form is adopted to model the effect of scalar predictors. For instance, Shin [28] considered the following model:

$$\mathrm{E}(Y|X, \boldsymbol{z}) = \int_{\mathcal{I}} X(t)\beta(t)\mathrm{d}t + \boldsymbol{z}^{\top}\boldsymbol{\alpha}, \tag{1}$$

where $\boldsymbol{z} = (z_1, \ldots, z_p)^{\top}$ is a p-dimensional scalar covariate and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^{\top} \in \mathbb{R}^p$ is the corresponding coefficient vector.

Let $m(t)$ and $G(s, t)$ denote the mean function and covariance function of $X(t)$, respectively. The covariance function $G(s, t)$ can be expressed as

$$G(s, t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t),$$

where $\lambda_1, \lambda_2, \ldots$ are the eigenvalues of G, satisfying $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, and $\psi_1(t), \psi_2(t), \ldots$ are the corresponding orthonormal eigenfunctions, which satisfy $\int \psi_j\psi_k\mathrm{d}t = 1$ if $j = k$ and 0 otherwise. Then the process $X(t)$ admits the

Karhunen–Loève expansion:

$$X(t) = m(t) + \sum_{k=1}^{\infty} \xi_k \psi_k(t), \tag{2}$$

where $\xi_k = \int_{\mathcal{I}} (X(t) - m(t))\psi_k(t)dt$ is the uncorrelated functional principal component (FPC) score. In addition, $\mathrm{E}(\xi_k \xi_{k'}) = \lambda_k$ if $k = k'$ and 0 otherwise. (See [14] for an interesting recent use of these ideas for processes on the sphere.) Replacing $X(t)$ in (1) with the expression given in (2), we have

$$\mathrm{E}(Y|X, \boldsymbol{z}) = b + \sum_{k=1}^{\infty} \xi_k b_k + \boldsymbol{z}^\top \boldsymbol{\alpha},$$

where $b = \int_{\mathcal{I}} \beta(t)m(t)\mathrm{d}t$ and $b_k = \int_{\mathcal{I}} \beta(t)\psi_k(t)\mathrm{d}t$.

To allow for greater flexibility, the additive components with respect to the FPC scores $\xi_1, \xi_2, \ldots$ in the above equation can take a more general form. Motivated by the idea of a generalized additive model [16], we consider

$$\mathrm{E}(Y|X, \boldsymbol{z}) = b + \sum_{k=1}^{\infty} f_k(\xi_k) + \boldsymbol{z}^\top \boldsymbol{\alpha}.$$

This model without scalar predictors is previously studied in [24,36] to describe the relationship between a scalar response and a functional predictor.

For convenience of regularization on each component f_k, we first scale the FPC scores to [0, 1]. One possible approach is to treat ξ_k as having a $\mathcal{N}(0, \lambda_k)$ distribution and apply the cumulative distribution function (cdf) of $\mathcal{N}(0, \lambda_k)$ to ξ_k, i.e., $\zeta_k = \Phi(\xi_k/\sqrt{\lambda_k})$, where Φ is the cdf of $\mathcal{N}(0, 1)$. Other cumulative distributions could be employed for scaling, but we focus solely on the Gaussian case here. The corresponding additive model becomes:

$$\mathrm{E}(Y|X, \boldsymbol{z}) = b + \sum_{k=1}^{\infty} f_k(\zeta_k) + \boldsymbol{z}^\top \boldsymbol{\alpha}. \tag{3}$$

In addition to making the following regularization scheme easier to implement, there are two main reasons why we consider transferring FPC scores to a compact domain. The first reason concerns theoretical derivations. When each function in a functional space can be represented in terms of spline basis functions like B-spline bases or reproducing kernel functions, assuming the domain is compact can simplify theoretical derivations. Such examples can be found, e.g., in [23,30].

Our second reason for focussing on a compact domain explains why that would be reasonable. Let h_j denote the transformation $h(\xi_j) = \zeta_j$ and g_j denote the function with the argument being ζ_j. Note that if h_j is any strictly monotone, continuous map from $\mathbb{R}$ to (0, 1), we may write $f_j = g_j \circ h_j$ with $g_j = f_j \circ h_j^{-1}$. We assume that there exists an integer, d, which is large enough that $f_k \equiv 0$ when $k > d$. This amounts to assuming that only some of the FPC scores of the functional predictor are relevant to the response. Our truncated model is then given as

$$\mathrm{E}(Y|X, \boldsymbol{z}) = b + \sum_{j=1}^{d} f_j(\zeta_j) + \boldsymbol{z}^\top \boldsymbol{\alpha}. \tag{4}$$

In practice we choose an initial value of d in such a way that at least 99.9% of variability in $X(t)$ can be explained by the first d FPCs. Let $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_d)^\top$.

We also assume that the effect of each transformed score, $f_1, \ldots, f_d$ is a smooth function. In this paper, we call the effect of $X(t)$, namely $f(\boldsymbol{\zeta}) = b + f_1(\zeta_1) + \cdots + f_d(\zeta_d)$, the nonparametric part of the model and the linear combination $\boldsymbol{z}^\top \boldsymbol{\alpha}$ the parametric part of the model. In addition, the f_js are called nonparametric components. Model (4) is called a functional semiparametric additive model (FSAM) in this article.

2.2. Estimation method

The objective of this paper is to propose an estimation method which can select and estimate nonparametric components that are relevant to the response while estimating the effects of scalar covariates in Model (4).

Zhu et al. [36] considered a special case when $\boldsymbol{\alpha}$ is known to be $\mathbf{0}$ in Model (4); to select and smooth non-vanishing nonparametric components in the estimation of the nonparametric part, they apply the COmponent Selection and Smoothing Operator (COSSO) proposed by [21]. We provide a brief review of COSSO next.

Let H be the ℓth-order Sobolev space on [0, 1], defined by

$$H([0, 1]) = \{h : h^{(\nu)} \text{ is absolutely continuous for all } \nu \in \{0, \ldots, \ell - 1\}; h^{(\ell)} \in L^2\}.$$

Then H is a reproducing kernel Hilbert space (RKHS) equipped with the squared norm

$$\|h\|^2 = \sum_{\nu=0}^{\ell-1} \left\{ \int_0^1 h^{(\nu)}(t)\mathrm{d}t \right\}^2 + \int_0^1 \{h^{(\ell)}(t)\}^2 \mathrm{d}t. \tag{5}$$

For a more detailed introduction to this RKHS, one can refer to Section 2.3 in [13]. One can decompose H as $H = \{1\} \oplus \bar{H}$, where elements of $\bar{H}$ have been centered. For example, take $h(t) = t$. Then $h(t) = 1/2 + t - 1/2$ and $t - 1/2 \in \bar{H}$. Assuming that $f_1, \ldots, f_d \in \bar{H}$, then $f(\boldsymbol{\zeta})$ lies in the subspace $\mathscr{F}^d = \{1\} \oplus \sum_{j=1}^{d} \bar{H}$, i.e., in the direct sum of the space of constant functions and d copies of $\bar{H}$. This assumption addresses the issue of identifiability for nonparametric components in Model (4). The COSSO regularization, applied to functions in the RKHS, is used to select and smooth non-vanishing components when estimating f. Suppose the data consist of n independent and identically distributed triples $(X_1, \boldsymbol{z}_1, y_1), \ldots, (X_n, \boldsymbol{z}_n, y_n)$. When $\boldsymbol{\alpha}$ in Model (4) are known to be $\mathbf{0}$, then the COSSO estimate of f is defined by minimizing

$$Q(f) = \frac{1}{n} \sum_{i=1}^{n} \{y_i - f(\boldsymbol{\zeta}_i)\}^2 + \tau^2 J(f), \tag{6}$$

where $J(f) = \|P^1 f\| + \cdots + \|P^j f\|$ with each $P^j f$ denoting the projection of f onto $\bar{H}$ with the argument being the jth component of $\boldsymbol{\zeta}$, and τ denotes a tuning parameter which controls the trade-off between fidelity to the data and complexity of the model.

If $P^1 f, \ldots, P^j f$ are linear, then the minimizer of $Q(f)$ is the LASSO estimate. However, the sum of the seminorms of $P^j f$, i.e., $J(f)$, rather than the L_1 norm of the coefficient vector, is penalized in $Q(f)$ in general scenarios. More specifically, if we represent each $P^j f$ with a linear combination of the reproducing kernel functions, the $\|P^j f\|$s are not differentiable with respect to the coefficients. This fact makes minimization of $Q(f)$ an intricate problem. Lin and Zhang [21] argued that introducing an ancillary parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top$, can ease the minimization task greatly. As shown in Lemma 2 of [21], minimization of (6) is equivalent to minimizing

$$H(f, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \{y_i - f(\boldsymbol{\zeta}_i)\}^2 + \lambda_0 \sum_{j=1}^{d} \theta_j^{-1} \|P^j f\|^2 + \lambda \sum_{j=1}^{d} \theta_j \tag{7}$$

with respect to f and $\boldsymbol{\theta}$, when $f \in \mathscr{F}^d$, $\theta_1 \geq 0, \ldots, \theta_d \geq 0$ and $\lambda = \tau^4/(4\lambda_0)$. In (7), both λ_0 and λ are nonnegative tuning parameters, which control the smoothness and selection of the estimated nonparametric part, respectively. If $\theta_j = 0$, then the minimizer satisfies $\|P^j f\| = 0$, indicating that f_j, the jth component in the nonparametric part, vanishes. The outline of the algorithm is given as follows.

Generally speaking, the FPC scores cannot be observed directly; thus estimating the first d FPC scores for each trajectory X_i is indispensable to estimate the nonparametric part later. Trajectories are usually recorded at a grid of time points, which can be different across subjects, and they are often observed with measurement errors. To address these issues when estimating FPC scores, we can employ regularized FPCA, proposed by Ramsay and Silverman [25], or PACE, proposed by Yao et al. [35]. Then the estimated scaled FPC scores, denoted as $\hat{\boldsymbol{\zeta}}_i = (\hat{\zeta}_{i1}, \ldots, \hat{\zeta}_{id})$, can be obtained by applying the CDF of a normal distribution with specific variance to the estimated FPC scores.

Now we can implement COSSO. Let R_j denote the $n \times n$ matrix with the $(s,\ t)$ entry $R(\hat{\zeta}_{sj},\ \hat{\zeta}_{tj})$, where $R(\cdot, \cdot)$ is the reproducing kernel of $\bar{H}$, and $R_{\boldsymbol{\theta}}$ for the matrix $\theta_1 R_1 + \cdots + \theta_d R_d$. For fixed λ_0 and λ, the minimizer of (7) has the form

$$f(\boldsymbol{\zeta}) = b + \sum_{i=1}^{n} c_i \sum_{j=1}^{d} \theta_j R(\zeta_j, \hat{\zeta}_{ij}).$$

Thus $\boldsymbol{f} = (f(\hat{\boldsymbol{\zeta}}_1), \ldots, f(\hat{\boldsymbol{\zeta}}_n))^\top = \mathbf{1}_n b + R_{\boldsymbol{\theta}} \boldsymbol{c}$, where $\boldsymbol{c} = (c_1, \ldots, c_n)^\top$ and $\mathbf{1}_n$ is the vector of ones of length n. Then the penalty term has

$$\sum_{j=1}^{d} \theta_j^{-1} \|P^j f\|^2 = \sum_{j=1}^{d} \theta_j \boldsymbol{c}^\top R_j \boldsymbol{c} = \boldsymbol{c}^\top R_{\boldsymbol{\theta}} \boldsymbol{c}.$$

Now (7) becomes

$$\min_{b, \boldsymbol{c}, \boldsymbol{\theta} \geq \mathbf{0}_d} \frac{1}{n} (\boldsymbol{y} - \mathbf{1}_n b - R_{\boldsymbol{\theta}} \boldsymbol{c})^\top (\boldsymbol{y} - \mathbf{1}_n b - R_{\boldsymbol{\theta}} \boldsymbol{c}) + \lambda_0 \boldsymbol{c}^\top R_{\boldsymbol{\theta}} \boldsymbol{c} + \lambda \mathbf{1}_d^\top \boldsymbol{\theta}, \tag{8}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{0}_d$ denotes the vector consisting of d zeros.

To solve (8), we alternatively solve for the pair $(b, \boldsymbol{c})$ with $\boldsymbol{\theta}$ fixed and then solve for $\boldsymbol{\theta}$ with $(b, \boldsymbol{c})$ fixed. More specifically:

(1) When $\boldsymbol{\theta}$ is fixed, solving (8) is equivalent to solving the standard smoothing spline

$$\min_{b, \boldsymbol{c}} \|\boldsymbol{y} - \mathbf{1}_n b - R_{\boldsymbol{\theta}} \boldsymbol{c}\|^2 + n\lambda_0 \boldsymbol{c}^\top R_{\boldsymbol{\theta}} \boldsymbol{c}. \tag{9}$$

The solution of (9) is similar to a smoothing spline estimate and can be found in [33].

(2) With $(b, \boldsymbol{c})$ fixed, (8) becomes

$$\min_{\boldsymbol{\theta} \geq \mathbf{0}} (\boldsymbol{v} - G\boldsymbol{\theta})^\top (\boldsymbol{v} - G\boldsymbol{\theta}) + n\lambda \mathbf{1}_S \boldsymbol{\theta}, \tag{10}$$

Algorithm 1 Iterative updating for regularized functional semiparametric additive model

Step 1: Start with an initial value of $\boldsymbol{\alpha}$, say $\hat{\boldsymbol{\alpha}}^{(0)}$, and an initial value of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}^{(0)}$.

Step 2: Use the current estimate $\hat{\boldsymbol{\alpha}}^{(m)}$ and $\hat{\boldsymbol{\theta}}^{(m)}$ to obtain estimates $\hat{b}^{(m+1)}$ and $\hat{\boldsymbol{c}}^{(m+1)}$ by solving (9), in which $\boldsymbol{y}$ is replaced by $\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}^{(m)}$.

Step 3: Use the current estimate $\hat{\boldsymbol{\alpha}}^{(m)}$, $\hat{b}^{(m+1)}$ and $\hat{\boldsymbol{c}}^{(m+1)}$ to obtain an updated estimate $\hat{\boldsymbol{\theta}}^{(m+1)}$ by solving (11), in which $\boldsymbol{v}$ is replaced by $\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}^{(m)} - (1/2)n\lambda_0\hat{\boldsymbol{c}}^{(m+1)} - \mathbf{1}_n\hat{b}^{(m+1)}$.

Step 4: Use the estimate $\hat{b}^{(m+1)}$, $\hat{\boldsymbol{c}}^{(m+1)}$ and $\hat{\boldsymbol{\theta}}^{(m+1)}$ to obtain an updated estimate $\hat{\boldsymbol{\alpha}}^{(m+1)}$ by solving a least squares problem.

Step 5: Repeat Steps 2, 3 and 4 until $||\hat{\boldsymbol{\alpha}}^{(m+1)} - \hat{\boldsymbol{\alpha}}^{(m)}|| < \epsilon$, where ϵ is a pre-determined tolerance value.

where $\boldsymbol{v} = \boldsymbol{y} - (1/2)n\lambda_0\boldsymbol{c} - \mathbf{1}_n b$ and G is the $n \times d$ matrix with the jth column being $R_j\boldsymbol{c}$. Lin and Zhang [21] suggested considering an equivalent optimization problem: for some $M \geq 0$, find

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{v} - G\boldsymbol{\theta}\|^2, \quad \text{subject to } \mathbf{1}^\top\boldsymbol{\theta} \leq M \text{ and } \boldsymbol{\theta} \geq \mathbf{0}_d. \tag{11}$$

The tuning parameter M in (11) is equivalent to λ in (10). Alternatively, the optimization problem (10) can be addressed directly using `glmnet` in R with the lower bound of the parameters to be estimated set as 0.

Only when the effect of the scalar covariates $\boldsymbol{z}$ can be removed or is known, can the above algorithm be implemented. Now we take the unknown effect of $\boldsymbol{z}$ into consideration as well. The estimate of $g = f(\boldsymbol{\zeta}) + \boldsymbol{\alpha}^\top\boldsymbol{z}$ is defined as

$$\hat{g}_n \in \underset{g(\boldsymbol{\zeta},\boldsymbol{z})=b+\sum_{j=1}^d f_j(\zeta_j)+\boldsymbol{\alpha}^\top\boldsymbol{z}}{\arg\min} \left[\frac{1}{n}\sum_{i=1}^n \left\{y_i - g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i)\right\}^2 + \tau^2 J(g)\right],$$

where $f_j \in \bar{H}$ and $J(g)$ is set to be $J(f) = \|P^1 f\| + \cdots + \|P^d f\|$. We use "$\in$" rather than "$=$" since we do not know if the minimizer is unique in general; this does not affect the results which follow. Note that the regularization suggested above penalizes only the nonparametric part, while neglecting the effect of the parametric part $\boldsymbol{\alpha}^\top\boldsymbol{z}$.

Difficulties arise when we apply COSSO directly to estimate the nonparametric part f in (4) since the effect of scalar predictor $\boldsymbol{z}$ needs to be accounted for as well. If the coefficient, $\boldsymbol{\alpha}$, of $\boldsymbol{z}$ were known, then a slight modification of COSSO would suffice to deal with the estimation problem: replace $\boldsymbol{y}$ in (9) with $\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha}$ and $\boldsymbol{v}$ in (10) with $\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha} - n\lambda_0\boldsymbol{c}/2 - \mathbf{1}_n b$, where $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top$. However, $\boldsymbol{\alpha}$ is unknown as well and the estimate of the nonparametric part f depends on the value of $\boldsymbol{\alpha}$, which poses a bottleneck when implementing COSSO to estimate f. To deal with this problem, we propose an iterative updating algorithm to estimate both the nonparametric and parametric parts and to select and smooth the non-vanishing components in f.

After estimating the $\boldsymbol{\zeta}_i$s via regularized FPCA or PACE, the target function to be minimized can be written as

$$Q^\star(f, \boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^n \{y_i - f(\hat{\boldsymbol{\zeta}}_i) - \boldsymbol{\alpha}^\top\boldsymbol{z}_i\}^2 + \tau^2\sum_{j=1}^d \|P^j f\|,$$

where

$$f(\hat{\boldsymbol{\zeta}}_i) = b + \sum_{j=1}^d f_j(\hat{\zeta}_{ij})$$

denotes the nonparametric part evaluated at the estimated transformation $\hat{\boldsymbol{\zeta}}_i$. We aim to look for $\hat{f} \in \mathscr{F}^d$ and $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^p$, which can minimize the target function $Q^\star$. As illustrated above, minimizing $Q^\star$ is equivalent to another minimization problem, viz.

$$\min_{\boldsymbol{\alpha},b,\boldsymbol{c},\boldsymbol{\theta}\geq\mathbf{0}_d} \frac{1}{n}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha} - \mathbf{1}_n b - R_{\boldsymbol{\theta}}\boldsymbol{c})^\top(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha} - \mathbf{1}_n b - R_{\boldsymbol{\theta}}\boldsymbol{c}) + \lambda_0\boldsymbol{c}^\top R_{\boldsymbol{\theta}}\boldsymbol{c} + \lambda\mathbf{1}_d^\top\boldsymbol{\theta}. \tag{12}$$

Algorithm 1 outlines the steps to solve (12). The fitting method presented above is called Functional Semiparametric Additive Model via COmponent Selection and Smoothing Operator (FSAM-COSSO) in this paper. Minimization of (12) turns out to be a convex problem. Our numerical studies show that this algorithm can converge in a few steps with reasonable initial estimates for both $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$; these are taken to be the ordinary least squares estimate and $\mathbf{0}_d$, respectively.

2.3. Tuning parameter selection

Cross-validation (CV) or generalized cross-validation (GCV) can be employed to choose the tuning parameters. The following adaptive tuning scheme is a slight modification of the proposal by [21]:

(1) In step 1 of Algorithm 1, the initial value of $\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(0)}$ is chosen as $\mathbf{1}_d$. We employ GCV or CV to choose the tuning parameter λ_0 when addressing the smoothing spline problem. In the following updating steps, λ_0 is fixed to be the chosen value.
(2) A grid of points in a reasonable range are chosen as candidates for M. CV is employed to choose the "optimal" value of M. More specifically, the whole data set is randomly split into G folds. The optimal M is chosen as the value which can minimize

$$CV(M) = \frac{1}{n}\sum_{g=1}^{G}(\hat{\boldsymbol{y}}_g^{(-g)} - \boldsymbol{y}_g)^\top(\hat{\boldsymbol{y}}_g^{(-g)} - \boldsymbol{y}_g),$$

where $\hat{\boldsymbol{y}}_g^{(-g)}$ denotes the predicted values for the gth fold of the data when it is removed and the model is fitted using the other $G-1$ folds of the data.

3. Simulation studies

In this section, two simulation studies are conducted to evaluate the finite-sample performance of our proposed approach and compare it with other alternative methods.

3.1. FSAM (4) with scalar covariates

A functional covariate is generated from the first 20 Fourier basis functions with eigenvalues defined, for each $k \in \{1, \ldots, 20\}$, by $\lambda_k = ab^k$; we take $a = 31.6$ and $b = 0.5$. More specifically,

$$X(t) = \mu(t) + \sum_{k=1}^{20}\xi_k\psi_k(t) + e(t),$$

where $\mu(t) = t + \sin(t)$ denotes the mean function of $X(t)$, $\xi_k \sim \mathcal{N}(0, \lambda_k)$, the ψ_ks denote the Fourier basis functions and the measurement error $e(t)$ follows $\mathcal{N}(0, 0.01)$, independently of all the ξ_ks. We generate $n = 1000$ independent curves in total; each curve is sampled at 200 equally spaced points between 0 and 10. The corresponding scaled FPC scores, the ζ_{ik}s, are defined as $\zeta_{ik} = \Phi(\xi_{ik}/\sqrt{\lambda_k})$ for all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, 20\}$. Then the response variable y is generated from the model defined, for all $i \in \{1, \ldots, n\}$, by

$$y_i = 1.2 + f_1(\zeta_{i1}) + f_2(\zeta_{i2}) + f_4(\zeta_{i4}) + \boldsymbol{z}_i^\top\boldsymbol{\alpha}_0 + \epsilon_i.$$

In this model,

(a) $f_1(x) = xe^x - 1$, $f_2(x) = \cos(2\pi x)$ and $f_4(x) = 3(x-1/4)^2 - 7/16$. They have a common domain $[0, 1]$. The nonparametric part is $f(\boldsymbol{\zeta}) = 1.2 + f_1(\zeta_1) + f_2(\zeta_2) + f_4(\zeta_4)$, where $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_{21})^\top$; in other words, the non-vanishing nonparametric components are f_1, f_2 and f_4.
(b) $\boldsymbol{z}_i = (z_{i1}, z_{i2})^\top$ is independent of $X_i(t)$; the two components of $\boldsymbol{z}_i$ are independently generated from the $\mathcal{U}[0, 1]$ uniform distribution; $\boldsymbol{\alpha}_0 = (-1, 2)^\top$.
(c) ϵ_i is independent of both $X_i(t)$ and $\boldsymbol{z}_i$ and is generated from $\mathcal{N}(0, 1)$.

The signal-to-noise ratio, defined as $\mathrm{var}\{f(\boldsymbol{\zeta})\}/\mathrm{var}(\epsilon)$, is around 1.75 under this setup.

Among the 1000 data points $(X_1(t), \boldsymbol{z}_1, y_1), \ldots, (X_{1000}(t), \boldsymbol{z}_{1000}, y_{1000})$, 200 are randomly selected as the training set and the remaining 800 data points are treated as the test set. We used PACE to estimate FPC scores and then choose so that the first d fitted scores explain 99.9% of the variability in sample curves of the variability of the training set. We find d is around 20 in all simulation replicates. Let $\hat{\boldsymbol{\zeta}}_i = (\hat{\zeta}_{i1}, \ldots, \hat{\zeta}_{id})^\top$ denote the estimate of $\boldsymbol{\zeta}_i$. Then different methods are fitted to the triple $(\hat{\boldsymbol{\zeta}}_i, \boldsymbol{z}_i, y_i)$, where $i \in$ training set. The proposed method in this paper, FSAM-COSSO, is implemented to fit Model (4) to estimate and select non-vanishing nonparametric components as well as estimating the coefficient vector of the scalar covariate $\boldsymbol{z}$. In simulation studies and real data applications presented in Section 4, we take the order of the Sobolev space to be $\ell = 2$. But the algorithm proposed below can be extended to more general cases.

MARS [11] fits an additive model for $(\hat{\boldsymbol{\zeta}}_i, y_i - \boldsymbol{z}_i^\top\boldsymbol{\alpha}_0)$, assuming that the coefficients in the parametric part are known to be $(-1, 2)^\top$ and $y_i - \boldsymbol{z}_i^\top\boldsymbol{\alpha}_0$ is the new response. As a comparison, two types of extended FAMs are considered as well. The FSAM-GAMS model denotes a saturated model in which $(\hat{\zeta}_{i1}, \ldots, \hat{\zeta}_{id})^\top$ are fitted by a generalized additive model (GAM) while z_{i1}, z_{i2} are fitted in a linear form. FSAM-COSSO differs from FSAM-GAMS in that the latter does not take component selection into consideration. In the second extended FAM, assuming that ζ_1, ζ_2 and ζ_4 are known to be the only non-vanishing features and the expressions of f_1, f_2, f_4 are known as well, a multiple linear regression is fitted on $(f_1(\hat{\zeta}_{i1}), f_2(\hat{\zeta}_{i2}), f_4(\hat{\zeta}_{i4}), \boldsymbol{z}_i, y_i)$, in which y_i denotes the response and the explanatory variables consist of $f_1(\hat{\zeta}_{i1}), f_2(\hat{\zeta}_{i2}), f_4(\hat{\zeta}_{i4}), z_{i1}, z_{i2}$; this model is called FSAM-GAM1 in this paper. The FSAM-PFLR model employs a linear combination of $\hat{\zeta}_{i1}, \ldots, \hat{\zeta}_{im}$, where m denotes the number of retained FPCs, to represent the effect of $X_i(t)$ on y_i. It is a modified version of the partial functional linear regression proposed by [28], where the effect of a functional predictor is represented by a linear combination of the original FPC scores. The tuning parameter m is chosen based on AIC, as suggested in [28]. To investigate the effect of using the $\hat{\boldsymbol{\zeta}}_i$s on estimation

Table 1
Summary of the number of selected nonparametric components over the 1000 simulations for each model. Model size indicates the number of nonparametric components selected in the model. In FSAM-GAMS we only retain the significant nonparametric components (p-value less than 0.05). Here we implement the function `gam` in the R package `mgcv` to fit FSAM-GAMS. The corresponding p-values of nonparametric components are available from the function `summary.gam`. This selection rule applies to FSAM-PFLR as well, where the p-value is available from the function `lm`.

Model	Counts With the Model Size												
	1	2	3	4	5	6	7	8	9	10	11	12	13
MARS	0	0	20	71	140	170	197	181	127	69	16	6	3
FSAM-GAMS	0	0	169	261	238	161	98	42	19	9	2	1	0
FSAM-PFLR	5	533	312	98	40	8	4	0	0	0	0	0	0
FSAM-COSSO	0	5	593	202	113	54	26	7	0	0	0	0	0
FSAM-COSSO1	0	1	709	149	73	39	22	6	1	0	0	0	0

Table 2
Summary of frequency of each nonparametric component selected over the 1000 simulations for each model. In FSAM-GAMS we only retain the significant nonparametric components (p-value less than 0.05). This selection rule applies to FSAM-PFLR as well.

Model	Frequency of Each Nonparametric Factor									
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_4$	$\hat{f}_5$	$\hat{f}_6$	$\hat{f}_7$	$\hat{f}_8$	$\hat{f}_9$	$\hat{f}_{10}$
MARS	1000	1000	356	1000	274	226	233	235	211	244
FSAM-GAMS	1000	1000	232	997	155	100	120	106	103	114
FSAM-PFLR	1000	981	231	1000	154	100	121	98	92	83
FSAM-COSSO	1000	1000	100	993	59	30	36	22	30	31
FSAM-COSSO1	1000	1000	34	999	26	29	40	32	32	25
	$\hat{f}_{11}$	$\hat{f}_{12}$	$\hat{f}_{13}$	$\hat{f}_{14}$	$\hat{f}_{15}$	$\hat{f}_{16}$	$\hat{f}_{17}$	$\hat{f}_{18}$	$\hat{f}_{19}$	$\hat{f}_{20}$
MARS	210	208	196	226	222	240	236	228	217	249
FSAM-GAMS	105	108	105	106	96	101	126	111	114	125
FSAM-PFLR	63	75	65	84	70	60	63	59	18	42
FSAM-COSSO	31	39	38	36	29	40	50	48	57	55
FSAM-COSSO1	31	25	29	50	26	33	32	29	31	32

of each f_j, we also implement the proposed method with true scores, the $\boldsymbol{\zeta}_i$s, to fit the model. This method is denoted by FSAM-COSSO1 in the paper.

To assess the performance of the above methods, 1000 simulation replicates are conducted to estimate the mean squared predicted errors (MSPE) on the test set, which is defined as $\sum_i (y_i - \hat{y}_i)^2/n_0$ with n_0 equal to the size of the test set. Besides prediction accuracy, we can also compare the performance of the methods from the perspective of model fitting; particularly, the number of selected nonparametric components, the frequency with which each $\hat{\zeta}_k$ is selected, and the bias and standard error (SE) of the estimates of $\boldsymbol{\alpha}$ are reported for each method as well.

Tables 1 and 2 summarize the number and frequency of nonparametric components selected in each method over the 1000 simulations, respectively. FSAM-COSSO in most cases selects the correct number of nonparametric components. In contrast, FSAM-GAMS and MARS are prone to retain some irrelevant nonparametric components, which results in more complex models and hence greater variance. Since AIC is employed to select the number of retained FPCs, FSAM-PFLR tends to yield a model with a relatively small size. As a result, even though it is less likely for irrelevant features to be selected, FSAM-PFLR suffers from frequently ignoring relevant features. Furthermore, FSAM-COSSO not only selects relevant factors $(\zeta_1, \zeta_2, \zeta_4)$ in almost every simulation but also retains irrelevant features considerably less often compared with MARS and FSAM-GAMS. The similarity between FSAM-COSSO and FSAM-COSSO1 suggests that replacing the true scores with estimates would make little difference in component selection. Tables 1 and 2 therefore demonstrate that FSAM-COSSO enables us to better discover the nonparametric relationship between the functional covariate $X(t)$ and the response when a model is given in the form of (4).

Table 3 compares the above methods in terms of the estimated bias and SE of the estimates of $\boldsymbol{\alpha}$ and of the prediction accuracy which is represented by MSPE. To be specific we estimate the bias of an estimate $\hat{\theta}$ of a parameter θ with true value θ_0 by

$$\text{bias}(\hat{\theta}) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\theta}_i - \theta_0) \quad \text{and} \quad \text{SE}(\hat{\theta}) = \sqrt{\frac{1}{999}\sum_{i=1}^{1000}(\hat{\theta}_i - \bar{\hat{\theta}})^2},$$

in which $\hat{\theta}_i$ denotes the estimated θ in the ith simulation and $\bar{\hat{\theta}}$ is average of $\hat{\theta}$ over the 1000 simulations. FSAM-COSSO compares favorably with the other three competitors except FSAM-GAM1 and FSAM-COSSO1 in terms of prediction accuracy, even under the assumption that $\boldsymbol{\alpha}$ are known to be $\boldsymbol{\alpha}_0$ in MARS. In addition, the point estimator of $\boldsymbol{\alpha}$ obtained from FSAM-COSSO is more stable than its counterparts from the other three competitors. Even though FSAM-GAM1 outperforms FSAM-COSSO with respect to prediction accuracy and/or the bias and SE of estimated $\boldsymbol{\alpha}$, in practice we usually have no sufficient evidence to point out non-vanishing nonparametric components in advance, let alone the closed forms of these components.

Table 3
Summary of estimated bias and standard error (SE) of estimated $\boldsymbol{\alpha}$ using each method, and mean squared prediction errors (MSPE). The above statistics are calculated over the 1000 simulations. Note that the column of MSPE corresponds to average of MSPE over the 1000 simulations.

Model	Bias	SE	MSPE
MARS	–	–	1.33
FSAM-GAM1	(−0.025, −0.021)	(0.255, 0.272)	1.15
FSAM-GAMS	(0.032, −0.047)	(0.282, 0.308)	1.41
FSAM-PFLR	(0.034, −0.031)	(0.303, 0.319)	1.67
FSAM-COSSO	(−0.026, −0.028)	(0.262, 0.283)	1.20
FSAM-COSSO1	(−0.022, −0.018)	(0.249, 0.250)	1.11

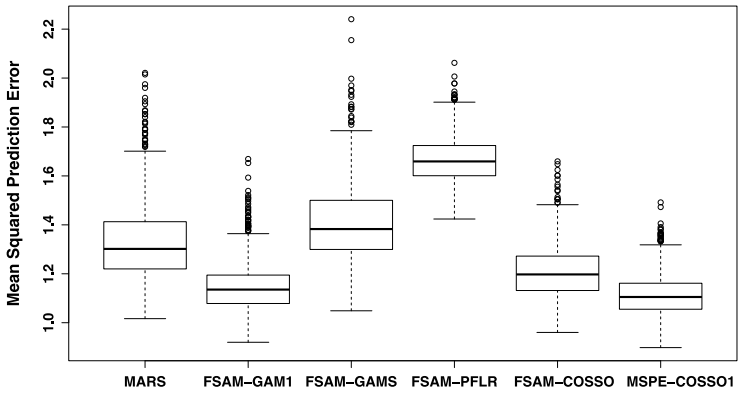

Fig. 1. Mean squared prediction errors of each method over 1000 simulations.

The boxplot in Fig. 1 provides a more detailed comparison of prediction errors among the six methods over the 1000 simulations; it shows that FSAM-COSSO has a substantial advantage in prediction when the underlying model is given in the form of (4) but unknown. The fact that FSAM-COSSO1 outperforms FSAM-GAM1 in prediction further indicates that the proposed algorithm is effective in discovering predictive features of the response.

In a randomly selected trial, 20 FPCs are retained initially such that over 99.9% of the variability in the curves can be captured. Fig. 2 illustrates how cross-validation is employed to choose the tuning parameter M. Choosing the value of M which can minimize the cross-validation error, FSAM-COSSO correctly selects the three non-vanishing nonparametric components. In addition, the other three panels in Fig. 2 display the estimated nonparametric components obtained from using the estimated scores and the true scores, as well as the true nonparametric components. It shows that estimates from these two methods are close to the true nonparametric functions and there is little disagreement between the two. This observation demonstrates that replacing true scores with the estimates has little impact on estimation of the nonparametric components.

3.2. FSAM (4) without scalar covariates

We also generate data in the same set up as in Section 3.1, except that the coefficient vector for the scalar covariate $\boldsymbol{z}$, $\boldsymbol{\alpha}_0$, is now set to $(0, 0)^\top$. This is essentially the model discussed in [36]. Besides the methods employed in Section 3.1, we also apply a method which regresses the scalar response y against $\hat{\boldsymbol{\zeta}}$ with COSSO regularization. This method is called FSAM-GAM2 in this paper. We also fit the data using the FSAM-GAM1 method, which estimates FSAM (4) by assuming that ζ_1, ζ_2 and ζ_4 are known to be the only non-vanishing features, the parametric expressions of f_1, f_2, f_4 are known, and the coefficients of the scalar covariate $\boldsymbol{z}$ are known to be 0. In other words, the FSAM-GAM1 method is essentially a multiple linear regression model with y_i as the response and $f_1(\hat{\zeta}_{i1}), f_2(\hat{\zeta}_{i2}), f_4(\hat{\zeta}_{i4})$ as the explanatory variables. Results comparing these methods are presented in the supplementary document.

Table S1 in the Online Supplement summarizes the number of nonparametric components selected by each method over the 1000 Monte Carlo runs. There is only a slight difference between FSAM-COSSO and FSAM-GAM2 in terms of selecting relevant components, which suggests that our proposed method can still perform well in component selection even if there is actually no effect from scalar covariates. Table S2 in the Online Supplement further compares these methods in a more delicate way by providing the frequency of each component selected across these 1000 Monte Carlo runs. Likewise in the scenario when there are scalar covariates involved in the model, FSAM-COSSO shows great advantages in retaining irrelevant components far less often compared with MARS and FSAM-PFLR. In addition, the reason why FSAM-PFLR compares favorably

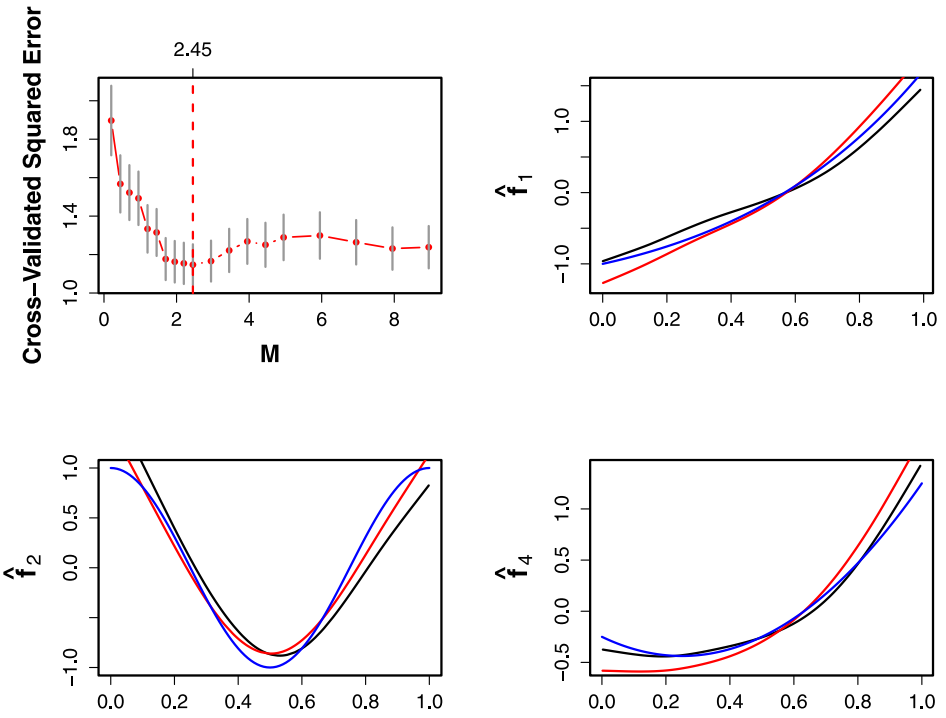

Fig. 2. The top left panel shows how the cross validation errors change across a range of plausible values for the tuning parameter M. The other 3 panels compare the estimated nonparametric components and the true underlying nonparametric components (f_1, f_2, f_4). The blue lines denote the true nonparametric components, while the black and red lines represent the estimated nonparametric components from FSAM-COSSO and FSAM-COSSO1, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with FSAM-COSSO in retaining irrelevant components is that AIC tends to select a relatively small number of scaled FPC scores.

More remarkable distinctions among these methods can be found in Figure S1 in the Online Supplement, which depicts the mean squared predictions on the test data over these 1000 Monte Carlo runs. The performance of the proposed method, FSAM-COSSO is slightly inferior to those of FSAM-GAM1 and FSAM-GAM2, but much better than the other methods in prediction accuracy. FSAM-GAM1 and FSAM-GAM2 each know the model structure to some extent in advance. That is why they can achieve greater prediction accuracy. Our method, however, does not assume that the linear part or the nonparametric part is known. Thus our proposed method is highly competitive in prediction compared with other fitting methods.

4. Real data applications

In this section, the proposed method (FSAM-COSSO) and several alternative methods are applied to analyze two real datasets: the Tecator data and attention deficit hyperactivity disorder (ADHD) data.

4.1. Tecator data

The Tecator data are recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 85–1050 nm by the Near Infrared Transmission (NIT) principle. The dataset consists of 240 meat samples; a 100-channel spectrum of absorbance (negative base 10 logarithms of the transmittance measured by the spectrometer) is recorded for each sample along with the percentages of three components of the meat: moisture (water), fat and protein. The three contents are determined by analytic chemistry. There has been extensive research on how to predict the contents using the spectrum of absorbance; see, e.g., [6,12,31,36]. The objective of this study is to examine the effect of the spectral trajectories and the fat and water content of the meat sample on the protein contents by fitting Model (1).

In Model (1), the response of primary interest, Y, is the protein content; both the spectral trajectories denoted as $X(t)$, and the fat and water contents denoted as $\boldsymbol{z}$, are considered as explanatory variables for predicting the protein content, differing from the method called the component selection and estimation for functional additive model (CSEFAM) in [36] where only spectral trajectories were taken into consideration for predicting the protein content. After applying PACE to the spectral trajectories to obtain estimated $\boldsymbol{\zeta}_i$s, FSAM-COSSO is implemented to fit Model (4) to estimate and select non-vanishing

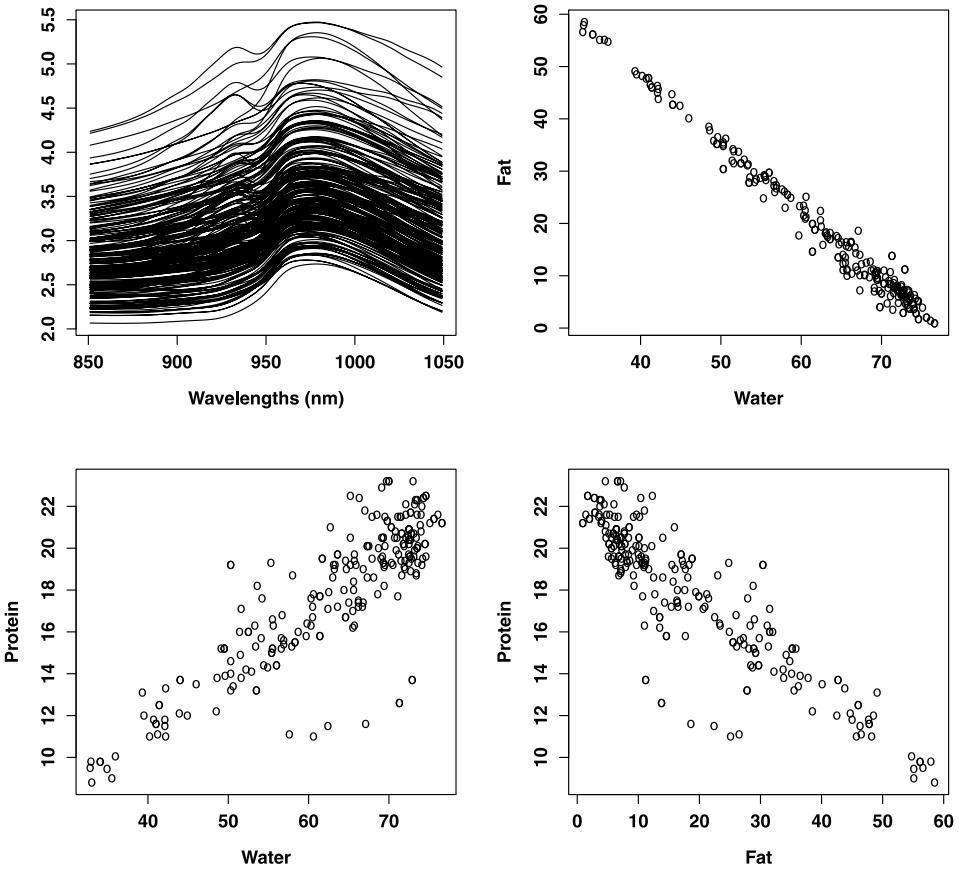

Fig. 3. The top left panel: the spectral trajectories recorded on the Tecator Infratec Food and Feed Analyzer. The other three panels depict the scatter plots among the three contents.

nonparametric components for the response as well as estimating the effect of scalar covariates on the response. The top left panel of Fig. 3 presents the spectral trajectories of the 240 meat samples. To assess the performance of each method, 185 out of the 240 meat samples are randomly selected from the training set and the remaining 55 samples constitute the test set.

As suggested by Zhu et al. [36], the first 20 FPCs, accounting for over 99.9% of the total variability in the spectral trajectories, are initially retained to avoid neglecting some relevant FPCs. In addition, pairwise scatter plots among the three contents, illustrated in Fig. 3, suggest a substantial multicollinearity between the fat and water contents and a linear relationship between the protein content and the fat and/or water content. Therefore, only the fat content is used in the parametric part when predicting the protein content. We then apply FSAM-COSSO to estimate and select nonparametric components while estimating the effect of the fat content on the prediction of the protein content. The tuning parameter λ_0 in the iterative updating algorithm is selected by sixfold cross-validation, which gives $\lambda_0 = 2.57 \times 10^{-5}$. Fivefold cross-validation suggests that 13 is an optimal choice for M.

The estimated nonparametric components are displayed in Figure S2 in the Online Supplement, which shows the 15 nonparametric components $\hat{f}_1, \ldots, \hat{f}_8, \hat{f}_{11}, \hat{f}_{13}, \ldots, \hat{f}_{18}$ are selected from the 20 components. The estimated coefficient of the fat content is -0.19, corresponding to the fact that the protein content is negatively correlated with the fat content indicated in the bottom right panel of Fig. 3.

MSPE and quasi-R^2 ([36]) are calculated on the test data set to compare various methods; the latter is defined by

$$R^2 = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

The last 10 FPCs actually explain less than 0.01% of the total variability in the spectral curves. However, they play a critical role in predicting the protein content, justifying why a sufficiently large number of principal components should be retained initially. To demonstrate the importance of the last 10 FPCs, the same model (FSAM-COSSO) is fit where only the first 10 FPCs are initially retained. The model neglecting the last 10 FPCs appears to be considerably inferior to that model retaining 20 FPCs initially in terms of both MPSE and R^2.

Table 4
Summary of prediction error and proportion of variance explained on the test set of each model. FAM represents the functional additive model [24] where only the $\hat{\boldsymbol{\zeta}}_i$s are considered as explanatory variables. MARS_0 denotes the MARS model considering only the $\hat{\boldsymbol{\zeta}}_i$s as explanatory variables while neglecting the effect of the fat content. $d = 10$ and $d = 20$ indicate that 10 and 20 leading FPCs are initially retained, respectively.

	$d = 20$					
	FSAM-COSSO	CSEFAM	FSAM-GAMS	FAM	MARS	MARS_0
MSPE	0.52	0.71	0.84	0.73	0.83	1.18
R^2	0.97	0.96	0.95	0.96	0.95	0.93
	$d = 10$					
	FSAM-COSSO	CSEFAM	FSAM-GAMS	FAM	MARS	MARS_0
MSPE	0.92	1.99	1.35	1.42	0.97	1.01
R^2	0.95	0.88	0.92	0.92	0.94	0.94

We also fit several alternative methods such as MARS and FSAM-FAMS to discover the relationship between the protein content and the explanatory variables and compare these models with FSAM-COSSO in prediction accuracy and the ability to explain the variability in the response. The effect of FPCs retained initially on the prediction of the protein content is examined as well in the alternative methods. Furthermore, to investigate how the prediction accuracy can be enhanced by incorporating the fat content in prediction models, the models are fitted by regressing the protein content only on the initially retained $\hat{\boldsymbol{\zeta}}_i$s and then compared to the corresponding models where both $\hat{\boldsymbol{\zeta}}_i$'s and the fat content are considered as explanatory variables.

Table 4 summarizes prediction errors and proportions of variance explained on the test set of all methods. It can be observed from the table that retaining a sufficiently large number of FPCs initially can ameliorate prediction accuracy to a great extent, even though the last 10 FPCs only make a negligible contribution to capturing the variance of the spectral curves. In addition, accounting for the effect of the fat content in each of the above models outperforms its counterpart, which does not incorporate the effect of the fat content, in terms of prediction accuracy and explaining variability in the response variable. Last but not least, the proposed method demonstrates its far superior ability to predict the response and explain its variance compared with other methods.

4.2. ADHD data

Attention deficit hyperactivity disorder (ADHD) is the most prevalent neurodevelopmental disorder in school-age children and adolescents ([9]). The key symptoms of ADHD comprise inattention, hyperactivity and impulsivity. Due to lack of objective measurements in diagnosis, there have been critical concerns for appropriate diagnosis of ADHD, which is associated with substantial social and economic costs ([10]). The data were obtained from the ADHD-200 Sample Initiative Project, which aimed to seek objective biological tools like neuroimaging signals to aid diagnosis of ADHD. Our analysis is based on the data collected from the New York University (NYU) Child Study Center, one of the eight sites in the project. The dataset consists of two main parts. The first part is filtered preprocessed resting state data using an anatomical automatic labeling atlas, which parcellates the brain into 116 Regions of Interests (ROIs). In each region, the mean blood-oxygen-level dependent (BOLD) signal was recorded at 172 equally spaced time points. The second part is composed of phenotypic features like gender, handedness, diagnosis of ADHD, medication status, ADHD index and IQ measures. A more detailed description of the data can be found in Bellec et al. [2]. Our objective is to use the BOLD signals and phenotypic features to predict the ADHD index, a measure which can reflect the overall level of ADHD symptoms ([5]).

We focus on 120 subjects in our analysis after removing measurements which failed in quality control. The functional predictor is taken as the BOLD signals of 91st–108th regions, because they are parcellations of the cerebellum region, which was found to play a role in the development of ADHD [3]. To compare the prediction performance of each method, these 120 subjects are randomly divided into a training set with 100 subjects and a test set with the other 20 subjects. Following this rule, we randomly split the data to the training and test set for 100 times.

Table 5 summarizes the mean squared prediction error across the 100 splits for each method. FSAM-COSSO turns out to be substantially superior to other methods in terms of prediction accuracy. In addition, accounting for effects of phenotypic features is able to improve prediction accuracy greatly for each method. Moreover, for methods other than FSAM-COSSO, retaining a large number of FPC scores initially may impair prediction of the ADHD index as may be seen by comparing the upper part with the lower part of Table 5. The primary reason for this might be that the BOLD signal is not a strong predictor of the ADHD index; thus incorporating more FPC scores would add considerable prediction variabilities while making little contribution to reducing bias. However, the negligible difference in prediction accuracy of FSAM-COSSO between these two scenarios suggests that the proposed method manages to reduce variances via component selection and thus achieve a better trade-off between bias and variance. As a result, the proposed method can still achieve a satisfactory performance in prediction even though a large number of irrelevant FPC scores are retained initially.

5. Asymptotic properties

The following theorem is given only in the case when the true scaled FPC scores $\boldsymbol{\zeta}$ are known. It would be desirable to establish the corresponding theorem with the estimated scaled FPC scores, when the functional data are densely observed

Table 5
Summary of prediction error on the test set of each model. FLR represents the functional linear model where only the $\hat{\boldsymbol{\zeta}}_i$s are considered as explanatory variables. For both FSAM-PFLR and FLR, the number of retained FPCs is chosen via AIC. MARS_0 denotes the MARS model considering only the $\hat{\boldsymbol{\zeta}}_i$s as explanatory variables while neglecting other phenotypic features. 99.9% and 85% indicate that the first d FPCs initially retained can explain 99.9% and 85% of variability of curves in the training set, respectively.

	99.9%					
	FSAM-COSSO	CSEFAM	FSAM-PFLR	FLR	MARS	MARS_0
MSPE	49.24	196.66	84.52	216.78	88.67	332.51
	85%					
	FSAM-COSSO	CSEFAM	FSAM-PFLR	FLR	MARS	MARS_0
MSPE	49.27	194.07	84.52	216.78	67.44	247.48

and may even be contaminated with measurement errors. This problem was considered by Zhu et al. [36], where there were no scalar covariates. Thus it would be natural for us to follow their ideas to develop the theorem. Unfortunately we are not able to follow the proof of their Lemma 2, where a dual problem is used to show that the penalty term $J(f)$ is bounded by a constant independent of sample size. This statement is essential to show that the derivative of the estimated function is uniformly bounded for each argument. We are thus not able to derive their subsequent results without a bounded derivative.

For this reason we present the following theorem assuming the true scaled scores are known. We observe, however, that simulation studies show that there would not be remarkable differences in performances when the true scores are replaced by estimates. Thus, although we expect that our theorem should extend to estimated scores, we have not succeeded in doing so.

We now set out conditions to establish Theorem 1 given that the true scaled FPC scores are known. Suppose that n iid observations are generated from the following model

$$y_i = f_0(\boldsymbol{\zeta}_i) + \boldsymbol{\alpha}_0^\top \boldsymbol{z}_i + \epsilon_i,$$

where $\boldsymbol{\zeta}_i \in [0,1]^d$, $\boldsymbol{\alpha}_0 \in \mathbb{R}^p$, and f_0 is assumed to be an element of $\mathscr{F}^d = \{1\} \oplus \sum_{j=1}^d \bar{H}$ (again the sum indicates a direct sum of d copies of $\bar{H}$) with $H = \{1\} \oplus \bar{H}$ being the ℓth-order Sobolev space on $[0,1]$ with the norm defined in (5). Note that the estimated nonparametric part $\hat{f}_n$ and the estimated coefficient $\hat{\boldsymbol{\alpha}}$ in the parametric part are defined as the solution of the following minimization problem:

$$(\hat{f}_n, \hat{\boldsymbol{\alpha}}) = \underset{f \in \mathscr{F}^d, \boldsymbol{\alpha} \in \mathbb{R}^p}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^n \{y_i - f(\boldsymbol{\zeta}_i) - \boldsymbol{\alpha}^\top \boldsymbol{z}_i\}^2 + \tau_n^2 J(f) \right].$$

Consequently, the estimate of the conditional expectation function of y, $g_0(\boldsymbol{\zeta}, \boldsymbol{z}) = f_0(\boldsymbol{\zeta}) + \boldsymbol{\alpha}_0^\top \boldsymbol{z}$, is defined as

$$\hat{g}_n = \hat{f}_n + \hat{\boldsymbol{\alpha}}^\top \boldsymbol{z}.$$

The empirical norm of g is defined as

$$\|g\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^n g^2(\boldsymbol{\zeta}_i, \boldsymbol{z}_i)}$$

for $g \in \mathscr{G} = \{g : g(\boldsymbol{\zeta}, \boldsymbol{z}) = f(\boldsymbol{\zeta}) + \boldsymbol{\alpha}^\top \boldsymbol{z}, f \in \mathscr{F}^d, \boldsymbol{\alpha} \in \mathbb{R}^p\}$. Define $h(\boldsymbol{\zeta}) = \mathrm{E}(\boldsymbol{z}|\boldsymbol{\zeta})$ and $\boldsymbol{z}^\star = \boldsymbol{z} - h(\boldsymbol{\zeta})$. Let $\Lambda_{\min}(\boldsymbol{A})$ and $\Lambda_{\max}(\boldsymbol{A})$ denote the minimal and maximal eigenvalues of a matrix $\boldsymbol{A}$, respectively.

The following assumptions are needed.

(A.1) Both $\boldsymbol{\zeta}$ and $\boldsymbol{z}$ are statistically independent of ϵ. Furthermore, $\mathrm{E}(\epsilon) = 0$ and $\max\{\mathrm{E}(|\boldsymbol{z}_{(1)}|), \ldots, \mathrm{E}(|\boldsymbol{z}_{(p)}|)\} < \infty$, where $\boldsymbol{z}_{(j)}$ denotes the jth component of $\boldsymbol{z}$.
(A.2) $\Lambda_{\max}[\mathrm{var}\{h(\boldsymbol{\zeta})\}] < \infty$ and $0 < \Lambda_{\min}\{\mathrm{var}(\boldsymbol{z}^\star)\} \le \Lambda_{\max}\{\mathrm{var}(\boldsymbol{z}^\star)\} < \infty$. Obviously $0 < \Lambda_{\min}\{\mathrm{var}(\boldsymbol{z})\} \le \Lambda_{\max}\{\mathrm{var}(\boldsymbol{z})\} < \infty$ under (A.2).
(A.3) The ϵ_is are (uniformly) sub-Gaussian, i.e., there exist some constant K and σ_0^2, such that $K^2(\mathrm{E}e^{\epsilon_i^2/K^2} - 1) \le \sigma_0^2$.
(A.4) The support of $\boldsymbol{z}$ is compact in $\mathbb{R}^p$.

The main result is the following.

Theorem 1. *Assume that Assumptions (A.1)–(A.3) hold.*

(i) *If* $0 < J(f_0) < \infty$, *and* $\tau_n^{-1} = n^{\ell/(2\ell+1)}\{J(f_0)\}^{(2\ell-1)/(4\ell+2)}$, *we have* $\|\hat{g}_n - g_0\|_n = \{J(f_0)\}^{1/(2\ell+1)} \times O_P\{n^{-\ell(2\ell+1)}\}$ *and* $J(\hat{f}_n) = J(f_0)O_P(1)$.
(ii) *If* f_0 *is a constant, i.e.,* $J(f_0) = 0$, *and* $\tau_n^{-1} = n^{1/4}$, *then we have* $\|\hat{g}_n - g_0\|_n = O_P(n^{-1/2})$ *and* $J(\hat{f}_n) = O_P(n^{-1/2})$.

Corollary 1. *If, in addition to Assumptions (A.1)–(A.3), Assumption (A.4) holds as well, then in either case (i) or (ii), we have* $\|\hat{f}_n - f_0\|_n = O_P\{n^{-\ell(2\ell+1)}\}$ *and* $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|_E = O_P\{n^{-\ell/(2\ell+1)}\}$*, where* $\|\cdot\|_E$ *denotes the Euclidean norm of a vector.*

Proofs of Theorem 1 and Corollary 1 are given in the Online Supplement.

6. Conclusions and discussions

Semiparametric additive models are known to possess the flexibility of a nonparametric model and the interpretability of a parametric model. In this paper, we propose a functional semiparametric additive model in which a scalar response is regressed on a functional covariate and finite-dimensional scalar covariates. To achieve flexibility and interpretability simultaneously, the effect of the functional covariate on the mean response is modeled in the framework of FAM, where the additive components are functions of scaled FPC scores, and a linear relationship is assumed between the mean response and the scalar covariates. We also develop an estimation method to estimate both the nonparametric and parametric parts in the proposed model.

The estimation procedure consists of three important steps. First, FPCA or PACE is employed to estimate FPC scores of the functional covariate which may be subject to measurement errors. Second, we adopt a special regularization scheme (COSSO) to penalize the additive components to smooth and select non-vanishing components. Third, to address the issue of interdependence between the estimated nonparametric part and parametric part, we propose an iterative updating algorithm, which is similar in spirit to the EM algorithm.

We show that choosing a sufficiently large number of FPCs is essential. On the one hand, this can account for a great proportion of variability in the functional covariate. On the other hand, retaining a sufficiently large number of FPCs can to a great extent circumvent neglecting predictive FPC scores with small variances, since there is no guarantee that leading FPC scores are necessarily more relevant to the response. The importance of retaining a sufficiently large number of FPCs is demonstrated via the application to the Tecator data, where retaining a smaller number of FPCs results in substantially greater prediction errors. The applications also show that incorporating the effect of scalar covariates can enhance prediction accuracy compared with models that only account for the effect of the functional covariate when the scalar covariates are predictive of the response variable.

The asymptotic theory in our article is based on the assumption that the true scaled FPC scores are known, but in practice these are unavailable. We provide an algorithm with respect to estimating FPC scores from observed curves which may be subject to measurement errors and then estimating both nonparametric and parametric parts in the model using estimated FPC scores. It would be nice to extend the theory to this case where true scaled FPC scores are not observable. The simulation study suggests that the estimates are still quite close to the true nonparametric and parametric parts when FPC scores are estimated.

Even though this work focuses on regressing a scalar response on a functional covariate and another finite-dimensional covariate, the methodology can be extended to accommodate other scenarios. For example, the framework may be extended to fit a generalized functional semiparametric additive model in which the distribution of the response variable of interest belongs to an exponential family. In addition, more than one functional covariate can be investigated in future work, in which we are more concerned about choosing relevant ones from multiple functional covariates. The model may be further extended to allow for high-dimensional scalar covariates. In the model proposed in this article, the dimension of scalar covariates p is fixed. If $p = p_n$ is allowed to increase when sample size n increases, we may also need to penalize the parametric part to choose relevant scalar covariates. In this case, regularization of both the nonparametric part and the parametric part is necessary to improve interpretability and reduce variability.

The proposed functional semiparametric additive model can be regarded as a natural extension of the work by Shin [28], which represented the effect of a functional covariate by a linear combination of leading FPC scores. We, however, employ an additive structure of FPC scores to model the effect of a functional covariate. As pointed out by one referee, there might be concerns in both the additive structure and the method of dimension reduction associated with this model. Wang and Ruppert [34] considered a special case of functional generalized additive models, which are easy to interpret since they do not impose the condition that a scalar response depends on a functional covariate via an additive form of FPC scores. Scheipl et al. [27] extended the idea to function-on-function regression models. Another additive model was proposed to model the relationship between a functional response and high-dimensional scalar covariates in [1]. Concerning dimension reduction, we project a functional covariate onto the directions of FPCs and then retain the leading FPC scores in the model. A possible concern is that this procedure does not take into account the response information. To address this issue, Li and Song [20] proposed nonlinear sufficient dimension reduction based on the relationship between the functional covariate and the scalar response. It would be worthwhile to explore both the generalized functional additive model and sufficient dimension reduction in functional regression models with scalar covariates.

Acknowledgments

The authors are grateful for the constructive comments and suggestions from the Editor-in-Chief, Prof. Christian Genest, an Associate Editor, and three reviewers. Thanks also to Prof. Linglong Kong and Dengdeng Yu for sharing the ADHD data with us. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grants of Lockhart (RGPIN-4023-2014) and Cao (RGPIN-2018-06008).

Appendix A. Supplementary data

The Online Supplement contains the additional results in Section 3.2 and Section 4.1, and detailed proofs for Theorem 1 and Corollary 1. We also provide computing code for all simulation studies and the real data applications at https://github.com/caojiguo/FSAM. Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2018.06.010.

References

[1] R.F. Barber, M. Reimherr, T. Schill, The function-on-scalar LASSO with applications to longitudinal GWAS, Electron. J. Stat. 11 (2017) 1351–1389.
[2] P. Bellec, C. Chu, F. Chouinard-Decorte, Y. Benhajali, D.S. Margulies, R.C. Craddock, The Neuro Bureau ADHD-200 preprocessed repository, Neuroimage 144 (2017) 275–286.
[3] P. Berquin, J. Giedd, L. Jacobsen, S. Hamburger, A. Krain, J. Rapoport, F. Castellanos, Cerebellum in attention-deficit hyperactivity disorder: A morphometric MRI study, Neurology 50 (1998) 1087–1093.
[4] H. Chen, Convergence rates for parametric components in a partly linear model, Ann. Statist. 16 (1988) 136–146.
[5] C.K. Conners, D. Erhardt, E.P. Sparrow, Conners' Adult ADHD Rating Scales (CAARS): Technical Manual, Multi-Health Systems, Toronto, ON, Canada, 1999.
[6] O. Devos, L. Duponchel, Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression, Chemom. Intel. Lab. Syst. 107 (2011) 50–58.
[7] R.F. Engle, C.W. Granger, J. Rice, A. Weiss, Semiparametric estimates of the relation between weather and electricity sales, J. Amer. Statist. Assoc. 81 (1986) 310–320.
[8] J. Fan, I. Gijbels, Local Polynomial Modelling and its Applications, Chapman & Hall/CRC, London, 1996.
[9] H.M. Feldman, M.I. Reiff, Attention deficit–hyperactivity disorder in children and adolescents, New Engl. J. Med. 370 (2014) 838–846.
[10] P.C. Ford-Jones, Misdiagnosis of attention deficit hyperactivity disorder: 'Normal behaviour' and relative maturity, PæDiatrics & Child Health 20 (2015) 200–202.
[11] J. Friedman, T. Hastie, R.J. Tibshirani, The Elements of Statistical Learning, Springer, Berlin, 2001.
[12] J. Goldsmith, F. Scheipl, Estimator selection and combination in scalar-on-function regression, Comput. Statist. Data Anal. 70 (2014) 362–372.
[13] C. Gu, Smoothing Spline ANOVA Models, Springer, New York, 2013.
[14] J.C. Guella, V.A. Menegatto, E. Porcu, Strictly positive definite multivariate covariance functions on spheres, J. Multivariate Anal. 166 (2018) 150–159.
[15] W.K. Härdle, H. Liang, J. Gao, Partially Linear Models, Springer, New York, 2000.
[16] T. Hastie, R.J. Tibshirani, Generalized additive models, Stat. Sci. 1 (1986) 297–310.
[17] N.E. Heckman, Spline smoothing in a partly linear model, J. R. Stat. Soc. Ser. B Stat. Methodol. 48 (1986) 244–248.
[18] A.E. Ivanescu, A.-M. Staicu, F. Scheipl, S. Greven, Penalized function-on-function regression, Comput. Statist. 30 (2015) 539–568.
[19] D. Kong, K. Xue, F. Yao, H.H. Zhang, Partially functional linear regression in high dimensions, Biometrika 103 (2016) 147–159.
[20] B. Li, J. Song, Nonlinear sufficient dimension reduction for functional data, Ann. Statist. 45 (2017) 1059–1095.
[21] Y. Lin, H.H. Zhang, Component selection and smoothing in multivariate nonparametric regression, Ann. Statist. 34 (2006) 2272–2297.
[22] Y. Lu, J. Du, Z. Sun, Functional partially linear quantile regression model, Metrika 77 (2014) 317–332.
[23] E. Mammen, S. van de Geer, Penalized quasi-likelihood estimation in partial linear models, Ann. Statist. 25 (1997) 1014–1035.
[24] H.-G. Müller, F. Yao, Functional additive models, J. Amer. Statist. Assoc. 103 (2008) 1534–1544.
[25] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, New York, 2005.
[26] D. Ruppert, M.P. Wand, R.J. Carroll, Semiparametric Regression, Cambridge University Press, Cambridge, 2003.
[27] F. Scheipl, A.-M. Staicu, S. Greven, Functional additive mixed models, J. Comput. Graph. Statist. 24 (2015) 477–501.
[28] H. Shin, Partial functional linear regression, J. Statist. Plann. Inference 139 (2009) 3405–3418.
[29] P. Speckman, Kernel smoothing in partial linear models, J. R. Stat. Soc. Ser. B Stat. Methodol. 50 (1988) 413–436.
[30] C.J. Stone, Additive regression and other nonparametric models, Ann. Statist. 13 (1985) 689–705.
[31] J.-P. Vila, V. Wagner, P. Neveu, Bayesian nonlinear model selection and neural networks: A conjugate prior approach, IEEE Trans. Neural Netw. 11 (2000) 265–278.
[32] G. Wahba, Cross Validated Spline Methods for the Estimation of Multivariate Functions from Data on Functionals, Department of Statistics, University of Wisconsin, Madison, WI, 1984.
[33] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, PA, 1990.
[34] X. Wang, D. Ruppert, Optimal prediction in an additive functional model, Statist. Sinica 25 (2015) 567–589.
[35] F. Yao, H.-G. Müller, J.-L. Wang, Functional data analysis for sparse longitudinal data, J. Amer. Statist. Assoc. 100 (2005) 577–590.
[36] H. Zhu, F. Yao, H.H. Zhang, Structured functional additive regression in reproducing kernel hilbert spaces, J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (2014) 581–603.