Statistica Sinica **29** (2019), 719-739
doi:https://doi.org/10.5705/ss.202017.0249

NONLINEAR FUNCTION-ON-FUNCTION ADDITIVE MODEL WITH MULTIPLE PREDICTOR CURVES

Xin Qi and Ruiyan Luo

Georgia State University

Abstract: We consider a nonlinear function-on-function additive regression model with multiple functional predictors. The forms of the nonlinear functions are unspecified, and offer great flexibility to model various relationships between the response curve and predictor curves. We clarify the identifiability issue of the model and identify the best decompositions of the nonlinear functions in the model in terms of prediction. To estimate this expansion, we solve a penalized functional generalized eigenvalue problem followed by a penalized least squares procedure. With the minimum prediction error of the proposed decomposition, our approach has good prediction accuracy. Our approach converts the estimation of three-dimensional nonlinear functions to the estimation of two- and one-dimensional functions, which considerably reduces computational costs. Asymptotic results are provided, and simulations and a data application show that the proposed method has good predictive performance and is efficient in dimension reduction and computation. This method is implemented in the **R** package `FRegSigCom`.

Key words and phrases: Additive model, function-on-function regression, generalized functional eigenvalue problem, nonlinear functional regression model, signal function, the Karhunen-Loève expansion.

1. Introduction

As a useful tool in functional data analysis, functional regression has gained increasing attention in recent years. Much effort has been made for linear regression models with functional predictors, such as Ramsay and Dalzell (1991), Cardot, Ferraty and Sarda (1999), Brown, Fearn and Vannucci (2001), Ratcliffe, Leader and Heller (2002), Reiss and Ogden (2007), Goldsmith et al. (2012) and Delaigle, Hall (2012) for linear scalar-on-function regression models, and Ramsay and Silverman (2005, Chap. 16), Yao, Müller and Wang (2005), Ivanescu et al. (2015), Meyer et al. (2015), Chiou, Yang and Chen (2016), Luo and Qi (2017) and Luo, Qi and Wang (2016) for linear function-on-function regression. There have also been numerous studies on nonlinear scalar-on-function regression models. For the functional version of the single-index model (Stoker (1986)),

$y = h\left(\int_{\mathcal{I}} X(s)\beta(s)ds\right) + \varepsilon$, the coefficient function $\beta(\cdot)$ and the unspecified function $h(\cdot)$ are typically estimated in an iterative way (Ait-Saïdi et al. (2008); Ferraty, Park and Vieu (2011)). The single-index model has been extended to a multiple-index model (James and Silverman (2005); Chen, Hall and Müller (2011); Ferraty et al. (2013)) with multiple linear functionals of the single predictor $y = \sum_{j=1}^{p} h_j\left(\int_{\mathcal{I}} X(s)\beta_j(s)ds\right) + \varepsilon$. Müller, Wu and Yao (2013) and McLean et al. (2014) proposed the continuously additive model $y = \mu + \int_{\mathcal{I}} F(X(s), s)ds + \varepsilon$, where $F(\cdot,\cdot)$ is a smooth function estimated by penalized tensor product B-splines.

When both response and predictors are functions, to study their nonlinear relationship, we consider a nonlinear function-on-function additive regression model

$$Y(t) = \mu(t) + \sum_{j=1}^{p} \int_{\mathcal{I}_j} F_j(X_j(s), s, t)ds + \varepsilon(t), \quad a \le t \le b, \tag{1.1}$$

where $X_1(\cdot), \ldots, X_p(\cdot)$ are functional predictors and $F_1(x, s, t), \ldots, F_p(x, s, t)$ are unspecified smooth functions. Without loss of generality, we assume that $\mathcal{I}_j = [0, 1]$. We allow within-function correlation in the noise function $\varepsilon(t)$. Model (1.1) extends the continuously additive model with scalar response and a single functional predictor to the model with functional response and multiple functional predictors. This model offers great flexibility in studying the relationship between the functional predictors and functional response. If each $F_j(x, s, t)$ is linear with respect to x, (1.1) is the usual linear function-on-function model. Model (1.1) has been explored by Scheipl, Staicu and Greven (2015) as an extension of a general frame work for functional additive mixed model. Scheipl, Staicu and Greven (2015) estimate the model with $p = 1$ by expanding the nonlinear function $F(x, s, t)$ (the subscript is omitted) using the tensor product of the basis for x, s and t. As $F(x, s, t)$ is trivariate, the number of the tensor product basis functions will be large even if the numbers of marginal basis functions are small. Scheipl, Staicu and Greven (2015) estimate the coefficients of all the tensor product basis functions simultaneously, which imposes heavy computational loads and may affect the estimation and prediction accuracy.

In this paper, we provide a novel approach to fit model (1.1). To briefly introduce our idea, we consider the model with one functional predictor: $Y(t) = \mu(t) + \int_0^1 F(X(s), s, t)ds + \varepsilon(t)$. We identify the best expansion $\sum_{k=1}^{\infty} G_k(x, s)\phi_k(t)$ of the nonlinear function $F(x, s, t)$ in terms of prediction among all possible expansions of the form $\sum_{k=1}^{\infty} H_k(x, s)\varphi_k(t)$, where $H_k(x, s)$'s and $\varphi_k(t)$'s are ar-

bitrary functions. Aiming to estimate this best expansion which has the minimum prediction error, our approach has good prediction accuracy. To estimate $\sum_{k=1}^{\infty} G_k(x,s)\phi_k(t)$, we first estimate $G_k(x,s)$'s sequentially by solving a penalized generalized functional eigenvalue problem. With the estimated $G_k(x,s)$'s, we transform the original model to a linear function-on-scalar regression model with scalar predictors, where $\mu(t)$ and $\phi_k(t)$'s are the coefficient functions. Then we estimate $\mu(t)$ and $\phi_k(t)$'s separately using a penalized least squares method. Our method breaks down the problem of estimating the trivariate function $F(x,s,t)$ to the problems of sequentially estimating the bivariate functions $G_k(x,s)$ and separately estimating the univariate functions $\mu(t)$, $\phi_k(t)$'s, which greatly improves the computational efficiency and can be easily extended to the model with multiple functional predictors.

The rest of the paper is organized as follows. We introduce our approach for one functional predictor in Sections 2 and extend it to multiple functional predictors in Section 3. In Sections 4 and 5, we report on simulation studies and a data analysis, respectively. We conclude the paper with a discussion in Section 6. We provide additional figures and tables, technical details and proofs in the online supplementary materials.

2. Nonlinear Regression with One Functional Predictor

To simplify notation and ease understanding, we introduce our method for model (1.1) with $p = 1$ in this section and extend it to $p > 1$ in Section 3. When $p = 1$, the model is

$$Y(t) = \mu(t) + \int_0^1 F(X(s), s, t)ds + \varepsilon(t), \quad a \le t \le b, \tag{2.1}$$

where $X(s)$ and $\varepsilon(t)$ are independent. Without loss of generality, we assume

$$E\{F(X(s), s, t)\} = 0, \quad \text{for all} \quad 0 \le s \le 1, a \le t \le b, \tag{2.2}$$

for otherwise we can replace $F(x,s,t)$ by $F(x,s,t) - E\{F(X(s),s,t)\}$ and $\mu(t)$ by $\mu(t) + \int_0^1 E\{F(X(s),s,t)\}\,ds$. We call $S(t) = \int_0^1 F(X(s),s,t)ds$ the signal function; it is a random function with zero mean and is crucial for predicting $Y(t)$.

2.1. Decomposition induced by signal compression (DISC)

Given $F(x,s,t)$, the distribution of $Y(t)$ is completely determined by (2.1) and the distribution of $X(s)$ and $\varepsilon(t)$. However, the function $F(x,s,t)$ may be unidentifiable. There may exist a function $\widetilde{F}(x,s,t) \neq F(x,s,t)$ which satisfies

both (2.2) and $\int_0^1 \widetilde{F}(X(s),s,t)ds = \int_0^1 F(X(s),s,t)ds$. Based on the distribution of $X(s)$ and $Y(t)$, or their random samples, we cannot differentiate $\widetilde{F}(x,s,t)$ from $F(x,s,t)$. Let

$$\mathcal{F} = \Bigg\{ F^* : \int_0^1 F^*(X(s),s,t)ds = \int_0^1 F(X(s),s,t)ds, \\ \text{and } F^* \text{ satisfies (2.2) and some regularity conditions} \Bigg\} \tag{2.3}$$

be the collection of all functions which lead to the same model as the true model and have the same regularity properties as possessed by $F(x,s,t)$, such as smoothness. Insteading of estimating the true function $F(x,s,t)$, we aim to estimate the signal function and predict the response curve by identifying and estimating a specific function in $\mathcal{F}$. If $F(x,s,t)$ is identifiable, the set $\mathcal{F}$ only contains the true function $F(x,s,t)$, and the function identified by our method is the same as the true function. Our approach does not need the identifiability of $F(x,s,t)$ and we do not assume it. In Section S2 of the supplementary material, we provide idenfiable conditions for $F(x,s,t)$.

We provide the explicit expression and an optimal prediction property of the specific function in $\mathcal{F}$ that we will estimate. We consider the model (2.1) at the population level in this section and provide our estimation method in Sections 2.2 and 2.3. Consider the Karhunen-Loève (KL) expansion $S(t) = \sum_{k=1}^{\infty} \mathbf{r}_k \phi_k(t)$, where $\phi_k(t)$'s are orthogonal (scaled) eigenfunctions of the covariance function of $S(t)$ corresponding to the eigenvalues $\sigma_1^2 \geq \sigma_2^2 \ldots$ with $\|\phi_k\|_{L^2} = \sigma_k$. The random variables $\mathbf{r}_k = \int_a^b S(t)\phi_k(t)dt/\sigma_k^2$, $k \geq 1$, are uncorrelated and have mean 0 and variance 1. For any function $G(x,s)$ of (x,s), let

$$\mathbf{r}(G) = \int_0^1 G(X(s),s)ds - \mathrm{E}\left\{ \int_0^1 G(X(s),s)ds \right\}, \tag{2.4}$$

which maps $G(x,s)$ to a random variable with mean zero. Define

$$G_k(x,s) = \int_a^b F(x,s,t)\phi_k(t)dt/\sigma_k^2, \quad \text{then by (2.2)} \quad \mathrm{E}\left\{ \int_0^1 G_k(X(s),s)ds \right\} = 0.$$

As $S(t) = \int_0^1 F(X(s),s,t)ds$ and $\mathbf{r}_k = \int_a^b S(t)\phi_k(t)dt/\sigma_k^2$, we have

$$\mathbf{r}_k = \int_0^1 G_k(X(s),s)ds = \mathbf{r}(G_k). \tag{2.5}$$

Thus $S(t) = \sum_{k=1}^{\infty} \mathbf{r}_k \phi_k(t) = \sum_{k=1}^{\infty} \mathbf{r}(G_k)\phi_k(t)$ which, together with (2.5), leads to

$$\int_0^1 F(X(s),s,t)ds = S(t) = \sum_{k=1}^{\infty} \mathbf{r}(G_k)\phi_k(t) = \int_0^1 \left\{ \sum_{k=1}^{\infty} G_k(X(s),s)\phi_k(t) \right\} ds. \tag{2.6}$$

Let $F^{(DISC)}(x,s,t) = \sum_{k=1}^{\infty} G_k(x,s)\phi_k(t)$, which we call as the *d*ecomposition *i*nduced by the *s*ignal *c*ompression (Luo and Qi (2017)). By (2.6), $F^{(DISC)}(x,s,t)$ leads to the same model as the true function $F(x,s,t)$. We will estimate the signal function and make prediction by estimating $F^{(DISC)}(x,s,t)$. In practice, we only need to estimate the first few terms in $F^{(DISC)}$. We find that the partial sum $\sum_{k=1}^{K} G_k(x,s)\phi_k(t)$, for any $K \geq 1$, has the minimum prediction error among all expansions of the form $\sum_{k=1}^{K} H_k(x,s)\varphi_k(t)$. The expansion $\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K} a_{ijk}\Psi_i(x)\Phi_j(s)\varphi_k(t)$ in Scheipl, Staicu and Greven (2015) by K^3 tensor product basis functions, where $\Psi_i(x)$'s, $\Phi_j(s)$'s and $\varphi_k(t)$'s are arbitrary marginal basis functions for x, s and t, respectively, has this form with $H_k(x,s) = \sum_{i=1}^{K}\sum_{j=1}^{K} a_{ijk}\Psi_i(x)\Phi_j(s)$.

Theorem 1. *Let $(X_{\text{new}}(s), Y_{\text{new}}(t))$ be a new observation with the same distribution as $(X(s), Y(t))$. For any $K \geq 1$, the mean squared prediction error for the partial sum $\sum_{k=1}^{K} G_k(x,s)\phi_k(t)$ satisfies*

$$\begin{aligned} \mathrm{E}\left(\|\varepsilon\|_{L^2}^2\right) &\leq \mathrm{E}\left\{ \left\| Y_{\text{new}} - \mu - \int_0^1 \sum_{k=1}^{K} G_k(X_{\text{new}}(s),s)\phi_k ds \right\|_{L^2}^2 \right\} = \sum_{k=K+1}^{\infty} \sigma_k^2 + \mathrm{E}\left(\|\varepsilon\|_{L^2}^2\right) \\ &= \min_{\substack{H_k(x,s),\varphi_k(t),\\ 1\leq k\leq K}} \left(\mathrm{E}\left\{ \left\| Y_{\text{new}} - \mu - \int_0^1 \sum_{k=1}^{K} H_k(X_{\text{new}}(s),s)\varphi_k ds \right\|_{L^2}^2 \right\} \right), \end{aligned} \tag{2.7}$$

where the minimum is taken over all possible $H_k(x,s)$ satisfying that $\int_0^1 H_k(X(s), s)ds$ has a finite second moment and all possible square integrable functions $\varphi_k(t)$, $1 \leq k \leq K$. Under Condition 1 *in Section* 2.4*, we have*

$$\begin{aligned} C_1 K^{-\theta+1} \leq \mathrm{E}&\left\{ \left\| Y_{\text{new}} - \mu - \int_0^1 \sum_{k=1}^{K} G_k(X_{\text{new}}(s),s)\phi_k ds \right\|_{L^2}^2 \right\} \\ &- \mathrm{E}\left(\|\varepsilon\|_{L^2}^2\right) \leq C_2 K^{-\theta+1}, \end{aligned} \tag{2.8}$$

where $0 < C_1 < C_2$ are constants not depending on K.

The prediction error of $\sum_{k=1}^{K} G_k(x,s)\phi_k(t)$ has a lower bound $\mathrm{E}\left(\|\varepsilon\|_{L^2}^2\right)$ because the noise function is independent of the predictor function and is completely unpredictable only based on predictor curves. So the prediction error due to the

noise function cannot be reduced. The part, $\sum_{k=K+1}^{\infty}\sigma_k^2$, in the prediction error is related to the bias caused by the truncation of $\sum_{k=1}^{\infty}G_k(X(s),s)\phi_k(t)$ after the first K terms. We can reduce this truncation bias by adding more terms in the partial sum. However, in practice, we have to estimate $G_k(x,s)$ and $\phi_k(t)$, which leads to additional prediction error. The prediction error due to estimation will increase with more terms added into the partial sum. A trade-off between the error due to truncation and the error due to estimation is a balance between bias and variance, and can be achieved by an appropriate choice of K, which we term the number of components.

Our estimation procedure consists of two steps. We first sequentially estimate $G_1(x,s),\ldots,G_K(x,s)$. Then we separately estimate $\mu(t),\phi_1(t),\ldots,\phi_K(t)$, by penalized least squares. For any $G(x,s)$ and $\widetilde{G}(x,s)$, let

$$\mathbf{\Lambda}(G,G)=\int_a^b\left[\mathrm{E}\left\{S(t)\mathbf{r}(G)\right\}\right]^2dt,\quad \mathbf{\Sigma}(G,\widetilde{G})=\mathrm{E}\left\{\mathbf{r}(G)\mathbf{r}(\widetilde{G})\right\},\tag{2.9}$$

where $\mathbf{r}(\cdot)$ is the function defined in (2.4). $\mathbf{\Sigma}(G,\widetilde{G})$ is the covariance of $\mathbf{r}(G)$ and $\mathbf{r}(\widetilde{G})$.

Theorem 2. *The $G_k(x,s)$'s are the solutions to the generalized eigenvalue problem,*

$$\begin{aligned}&\max_{G(x,s)}\quad \mathbf{\Lambda}(G,G),\\ &\text{subject to}\quad \mathbf{\Sigma}(G,G)=1\quad\text{and}\quad \mathbf{\Sigma}(G_{k'},G)=0,\quad 1\le k'\le k-1,\end{aligned}\tag{2.10}$$

where the maximum is taken over all possible $G(x,s)$ such that $\int_0^1 G(X(s),s)ds$ has a finite second moment. Moreover, the maximum value of (2.10) *is σ_k^2.*

Now (2.5) implies that $\mathbf{r}(G_k)$'s are uncorrelated random variables with means zero and variances one, so $\mathbf{\Sigma}(G_k,G_k)=\mathrm{E}\left\{\mathbf{r}(G_k)^2\right\}=1$ and $\mathbf{\Sigma}(G_k,G_{k'})=\mathrm{E}\left\{\mathbf{r}(G_k)\mathbf{r}(G_{k'})\right\}=0$ for any $k'\neq k$, which leads to the constraints in (2.10). To interpret the problem (2.10), we first consider $k=1$. Since $G_1(x,s)$ is the first generalized eigenfunction of the problem (2.10), it maximizes the following Rayleigh quotient,

$$\begin{aligned}\frac{\mathbf{\Lambda}(G,G)}{\mathbf{\Sigma}(G,G)}&=\int_a^b\frac{\left[\mathrm{E}\left\{S(t)\mathbf{r}(G)\right\}\right]^2}{\mathrm{E}\left\{\mathbf{r}(G)\mathbf{r}(G)\right\}}dt=\int_a^b\frac{\mathrm{Cov}^2\left(S(t),\mathbf{r}(G)\right)}{\mathrm{Var}(\mathbf{r}(G))}dt\\&=\int_a^b\mathrm{Corr}^2\left(S(t),\mathbf{r}(G)\right)\mathrm{Var}(S(t))dt.\end{aligned}$$

Solving (2.10) is equivalent to finding a function $G(x,s)$ to maximize the integral of the squared correlation between $S(t)$ and $\mathbf{r}(G)$ multiplied by the variance of

$S(t)$. When $k > 1$, we have the additional constraint that $\mathbf{r}(G_k)$ is uncorrelated with $\mathbf{r}(G_{k'})$ for $k' < k$.

2.2. Estimation of $G_k(x, s)$

Let $(X_1(s), Y_1(t)), \ldots, (X_n(s), Y_n(t))$ denote n independent observations from the model (2.1). Then we have

$$Y_l(t) = \mu(t) + \int_0^1 F(X_l(s), s, t)ds + \varepsilon_l(t), \quad 1 \le l \le n, \tag{2.11}$$

where $\varepsilon_l(t)$ denotes the l-th noise function. For any function $G(x, s)$, we define

$$r_l(G) = \int_0^1 G(X_l(s), s)ds - \mathrm{E}\left\{\int_0^1 G(X_l(s), s)ds\right\}, \quad \text{and} \quad \overline{r}(G) = \frac{1}{n}\sum_{l=1}^n r_l(G). \tag{2.12}$$

Then $r_1(G), \ldots, r_n(G)$ are i.i.d. samples of $\mathbf{r}(G)$, and $\overline{r}(G)$ is their sample mean. Given any functions $G(x, s)$ and $\widetilde{G}(x, s)$, we estimate $\mathbf{\Lambda}(G, G)$ and $\mathbf{\Sigma}(G, \widetilde{G})$ by

$$\widehat{\mathbf{\Lambda}}(G, G) = \frac{1}{n^2}\int_a^b \left[\sum_{l=1}^n \{r_l(G) - \overline{r}(G)\}\{Y_l(t) - \overline{Y}(t)\}\right]^2 dt, \tag{2.13}$$

$$\widehat{\mathbf{\Sigma}}(G, \widetilde{G}) = \frac{1}{n}\left[\sum_{l=1}^n \{r_l(G) - \overline{r}(G)\}\left\{r_l(\widetilde{G}) - \overline{r}(\widetilde{G})\right\}\right],$$

where $r_l(G) - \overline{r}(G) = \int_0^1 \left\{G(X_l(s), s) - \overline{G}(s)\right\} ds$ can be calculated from sample curves, and $\overline{G}(s) = \sum_{l=1}^n G(X_l(s), s)ds/n$. To impose the smoothness penalty, we introduce some notation. For functions $G(x, s)$ and $\widetilde{G}(x, s)$, let $\langle G, \widetilde{G}\rangle_{L^2} = \int\int G(x, s)\widetilde{G}(x, s)dxds$ and $\|G\|_{L^2}$ denote the usual L^2 inner product and L^2 norm, respectively. Let

$$\begin{aligned}\langle G, \widetilde{G}\rangle_{H^2} &= \langle G, \widetilde{G}\rangle_{L^2} + \langle \partial_{xx}G(x, s), \partial_{xx}\widetilde{G}(x, s)\rangle_{L^2} + \langle \partial_{xs}G(x, s), \partial_{xs}\widetilde{G}(x, s)\rangle_{L^2} \\ &\quad + \langle \partial_{ss}G(x, s), \partial_{ss}\widetilde{G}(x, s)\rangle_{L^2}, \\ \|G\|_{H^2} &= \sqrt{\|G\|_{L^2}^2 + \|\partial_{xx}G\|_{L^2}^2 + \|\partial_{xs}G\|_{L^2}^2 + \|\partial_{ss}G\|_{L^2}^2}\end{aligned}$$

denote the Sobolev inner product and the Sobolev norm, respectively, where $\partial_{xx}G$, $\partial_{xs}G$ and $\partial_{ss}G$ are the second order partial derivatives. Our estimate $\widehat{G}_k(x, s)$ of $G_k(x, s)$ is obtained sequentially by solving the penalized optimization problem

$$\max_{G(x,s)} \frac{\widehat{\mathbf{\Lambda}}(G, G)}{\widehat{\mathbf{\Sigma}}(G, G) + \lambda\|G\|_{H^2}^2}, \quad \text{subject to} \quad \widehat{\mathbf{\Sigma}}(\widehat{G}_{k'}, G) + \lambda\langle \widehat{G}_{k'}, G\rangle_{H^2} = 0, \tag{2.14}$$

for all $1 \le k' \le k - 1$, where $\lambda\|G\|_{H^2}^2$ is the smoothness penalty imposed on

$G(x,s)$. Because the penalty is imposed on the denominator of the objective function in (2.14), to maximize this objective function, the solutions $\widehat{G}_k(x,s)$ tend to have small $\|\cdot\|_{H^2}$-norms, especially for a large tuning parameter λ. Therefore, the proposed penalty encourages the smoothness of the estimated functions.

In practice, we use tensor product B-spline basis functions to expand $G(x,s)$ and express (2.14) as an optimization problem of the expansion coefficients. Without loss of generality, we suppose that the sample predictor curves, $X_l(s)$, $1 \le l \le n$, have been scaled such that their values are between 0 and 1. Let $\mathbf{b}(s) = (b_1(s), b_2(s), \ldots, b_L(s))^\top$ and $\mathbf{h}(x) = (h_1(x), h_2(x), \ldots, h_J(x))^\top$ be the vectors basis functions for $s \in [0,1]$ and $x \in [0,1]$, respectively. Let $\mathbf{\Psi}(x,s) = \mathbf{h}(x) \otimes \mathbf{b}(s) = (h_1(x)b_1(s), h_1(x)b_2(s), \ldots, h_J(x)b_L(s))^\top$ denote the vector of tensor product basis functions in the two-dimensional region $[0,1] \times [0,1]$.

In this paper, we assume that the sample predictor curves $X_l(s)$, $1 \le l \le n$, are densely observed on a common grid $0 = s_1 < \cdots < s_{N_x} = 1$, and the sample response curves $Y_l(t)$, $1 \le l \le n$, are densely observed in a common grid $a = t_1 < \cdots < t_{N_y} = b$, where N_x and N_y are the number of observation points. For any continuous functions $g(s)$ observed at $\{s_1, \ldots s_{N_x}\}$ and $f(t)$ observed at $\{t_1, \ldots t_{N_y}\}$, we use the approximation $\int_0^1 g(s)ds \approx \sum_{m=1}^{N_x} \delta_m^x f(s_m)$ and $\int_a^b f(t)dt \approx \sum_{m=1}^{N_y} \delta_m^y f(t_m)$, where $\{\delta_m^x : 1 \le m \le N_x\}$ and $\{\delta_m^y : 1 \le m \le N_y\}$ are weights. For equally spaced observation points of $X_l(s)$'s, we choose $\delta_1^x = \cdots = \delta_{N_x}^x = 1/N_x$; for unequally spaced observation points, we choose $\delta_1^x = (s_2 - s_1)/2$, $\delta_m^x = (s_{m+1} - s_{m-1})/2$ for $1 < m < N_x$, and $\delta_{N_x}^x = (s_{N_x} - s_{N_x-1})/2$, based on the trapezoidal formula. δ_m^y's are chosen in a similar way.

Let $G(x,s) = \mathbf{z}^\top \mathbf{\Psi}(x,s)$ be a linear combination of the tensor product basis functions $\mathbf{\Psi}(x,s)$, where $\mathbf{z}$ is the JL dimensional coefficient vector. Then the numerator of the objective function in (3.2) can be expressed as

$$\begin{aligned}
\widehat{\mathbf{\Lambda}}(G,G) &= \frac{1}{n^2} \int_a^b \left(\sum_{l=1}^n \left[\int_0^1 \mathbf{z}^\top \left\{ \mathbf{\Psi}(X_l(s), s) - \overline{\mathbf{\Psi}}(s) \right\} ds \right] \{Y_l(t) - \overline{Y}(t)\} \right)^2 dt \\
&\approx \frac{1}{n^2} \sum_{v=1}^{N_y} \left(\mathbf{z}^\top \sum_{l=1}^n \left[\sum_{u=1}^{N_x} \delta_u^x \left\{ \mathbf{\Psi}(X_l(s_u), s_u) - \overline{\mathbf{\Psi}}(s_u) \right\} \right] \{Y_l(t_v) - \overline{Y}(t_v)\} \right)^2 \delta_v^y \\
&= \sum_{l=1}^n \sum_{l'=1}^n \mathbf{z}^\top (\mathbf{g}_l - \bar{\mathbf{g}}) \mathbf{\Pi}_{ll'} (\mathbf{g}_{l'} - \bar{\mathbf{g}})^\top \mathbf{z} = \mathbf{z}^\top \mathbf{\Xi} \mathbf{z}, \qquad (2.15)
\end{aligned}$$

where $\overline{\mathbf{\Psi}}(s) = \sum_{l=1}^n \mathbf{\Psi}(X_l(s), s)/n$, $\mathbf{g}_l = \sum_{u=1}^{N_x} \delta_u^x \mathbf{\Psi}(X_l(s_u), s_u)/\sqrt{n}$ and $\bar{\mathbf{g}} = \sum_{l=1}^n \mathbf{g}_l/n$ are all JL-dimensional vectors. $\mathbf{\Pi}$ is an $n \times n$ matrix with the

(l,l')-th entry $\mathbf{\Pi}_{ll'} = \sum_{v=1}^{N_y} \delta_v^y \{Y_l(t_v) - \overline{Y}(t_v)\}\{Y_{l'}(t_v) - \overline{Y}(t_v)\}/n$ and $\mathbf{\Xi} = \sum_{l=1}^n \sum_{l'=1}^n (\mathbf{g}_l - \bar{\mathbf{g}})\mathbf{\Pi}_{ll'}(\mathbf{g}_{l'} - \bar{\mathbf{g}})^\top$ is a $JL \times JL$ matrix. Similarly, the first term in the denominator of the objective function in (3.2) can be expressed as $\widehat{\mathbf{\Sigma}}(G,G) \approx \mathbf{z}^\top \mathbf{H}\mathbf{z}$, where $\mathbf{H} = \sum_{l=1}^n \sum_{l'=1}^n (\mathbf{g}_l - \bar{\mathbf{g}})(\mathbf{g}_{l'} - \bar{\mathbf{g}})^\top$ is a $JL \times JL$ matrix. The penalty term in (2.14) can be expressed as $\lambda \|G\|_{H^2}^2 = \mathbf{z}^\top \mathbf{K}\mathbf{z}$, where $\mathbf{K} = \lambda \int_0^1 \int_0^1 \big[\mathbf{\Psi}(x,s)\mathbf{\Psi}(x,s)^\top + \{\partial_{xx}\mathbf{\Psi}(x,s)\}\{\partial_{xx}\mathbf{\Psi}(x,s)\}^\top + \{\partial_{xs}\mathbf{\Psi}(x,s)\}\{\partial_{xs}\mathbf{\Psi}(x,s)\}^\top + \{\partial_{ss}\mathbf{\Psi}(x,s)\}\{\partial_{ss}\mathbf{\Psi}(x,s)\}^\top\big]\, dxds$ is a $JL \times JL$ matrix.

The optimization problem (2.14) can be expressed as the eigenvalue problem for the coefficient vector $\mathbf{z}$,

$$\max_{\mathbf{z} \in \mathbb{R}^{JL}} \frac{\mathbf{z}^\top \mathbf{\Xi}\mathbf{z}}{\mathbf{z}^\top \mathbf{Q}\mathbf{z}}, \quad \text{subject to} \quad \widehat{\mathbf{z}}_{k'}^\top \mathbf{Q}\mathbf{z} = 0, \quad 1 \le k' \le k-1, \tag{2.16}$$

where $\mathbf{Q} = \mathbf{H} + \mathbf{K}$. Let $\widehat{\mathbf{z}}_k$ be the solution to (2.16). Then $G_k(x,s)$ is estimated by $\widehat{G}_k(x,s) = \widehat{\mathbf{z}}_k^\top \mathbf{\Psi}(x,s)$.

2.3. Estimation of $\mu(t)$ and $\phi_k(t)$

With estimates $\widehat{G}_k(x,s)$, $1 \le k \le K$, we next estimate $\mu(t)$ and $\phi_k(t)$, $1 \le k \le K$, by a transformation of the model (2.11). Since $X_1(s), \ldots, X_n(s)$ have the same distribution as $X(s)$, by (2.6), we have

$$\int_0^1 F(X_l(s), s, t)ds = \int_0^1 \left\{ \sum_{k=1}^\infty G_k(X_l(s), s)\phi_k(t) \right\} ds = \sum_{k=1}^\infty r_l(G_k)\phi_k(t),$$

for all $1 \le l \le n$, where $r_l(\cdot)$ is defined in (2.12) and we use $\mathrm{E}\big\{ \int_0^1 G_k(X_l(s), s)ds \big\} = 0$. Let $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))^\top$, $\mathbf{E}(t) = (\varepsilon_1(t), \ldots, \varepsilon_n(t))^\top$ and $\mathbf{R}_k = (r_1(G_k), \ldots, r_n(G_k))^\top$. Then the nonlinear function-on-function additive model (2.11) can be transformed to a linear function-on-scalar regression model

$$\mathbf{Y}(t) = \mathbf{1}_n \mu(t) + \sum_{k=1}^\infty \mathbf{R}_k \phi_k(t) + \mathbf{E}(t),$$

where $\mathbf{R}_k$'s can be viewed as new scalar predictors and $\phi_k(t)$'s are the corresponding coefficient functions. Here $\mathbf{R}_k$ is not observed, but can be estimated by $\widehat{\mathbf{R}}_k = \{r_1(\widehat{G}_k) - \overline{r}(\widehat{G}_k), \ldots, r_n(\widehat{G}_k) - \overline{r}(\widehat{G}_k)\}^\top$ for $1 \le k \le K$, where $\overline{r}(\cdot)$ is defined in (2.12). If the remainder term $\sum_{k=K+1}^\infty \mathbf{R}_k \phi_k(t)$ is small enough, we have $\mathbf{Y}(t) \approx \mathbf{1}_n \mu(t) + \sum_{k=1}^K \widehat{\mathbf{R}}_k \phi_k(t) + \mathbf{E}(t)$. Thus, to estimate $\mu(t), \phi_1(t), \ldots, \phi_K(t)$, we regress $\mathbf{Y}(t)$ on $\mathbf{1}_n, \widehat{\mathbf{R}}_1, \ldots, \widehat{\mathbf{R}}_K$ using the penalized least squares method in (Ramsay and Silverman (2005, Chap. 13)), which solves

$$\min_{\substack{v_0(t),\\ v_1(t),\ldots,v_K(t)}} \left(\frac{1}{n}\int_a^b \left\| \mathbf{Y}(t) - v_0(t) - \sum_{k=1}^{K} \widehat{\mathbf{R}}_k v_k(t) \right\|_2^2 dt \right.$$
$$\left. + \kappa \int_a^b \left\{ v_0''(t)^2 + \sum_{k=1}^{K} v_k''(t)^2 \right\} dt \right), \tag{2.17}$$

and the minimum is taken over all possible functions $v_0(t)$ and $v_k(t)$'s with square-integrable second derivatives in $[a, b]$. The details for solving (2.17) are provided in Section S.3.1 of the supplementary material. The estimates $\widehat{\mu}(t)$ and $\widehat{\phi}_k(t)$'s of $\mu(t)$ and $\phi_k(t)$'s are the solution to (2.17).

Given a new functional predictor $X_{\text{new}}(s)$ (which has been scaled in the same way as the $X_l(s)$'s), the response function is predicted by

$$Y_{\text{pred}}(t) = \widehat{\mu}(t) + \int_0^1 \sum_{k=1}^{K} \left\{ \widehat{G}_k(X_{\text{new}}(s), s) - \overline{\widehat{G}}_k(s) \right\} \widehat{\phi}_k(t) ds, \tag{2.18}$$

where $\overline{\widehat{G}}_k(s) = \sum_{l=1}^{n} \widehat{G}_k(X_l(s), s)/n$. One practical issue is that $X_{\text{new}}(s)$ may take values outside of $[0, 1]$. In this case, we extend $\widehat{G}_k(x, s)$ by letting $\widehat{G}_k(x, s) = \widehat{G}_k(0, s)$ if $x < 0$ and $\widehat{G}_k(x, s) = \widehat{G}_k(1, s)$ if $x > 1$, for any $0 \le s \le 1$.

The theoretical choices of the number K of components and the tuning parameters λ and κ are provided in Section 2.4. Their choices in practice are given in Section S.3.2 of the supplementary material.

2.4. Asymptotic theory

Let $\widehat{F}(x, s, t) = \sum_{k=1}^{K} \{\widehat{G}_k(x, s) - \overline{\widehat{G}}_k(s)\}\widehat{\phi}_k(t)$. We provide a convergence rate of the estimation errors $\int_0^1 \widehat{F}(X_l(s), s, t) ds - \int_0^1 F(X_l(s), s, t) ds$, $1 \le l \le n$, for the signal functions as $n \to \infty$. For a new observation $(X_{\text{new}}(s), Y_{\text{new}}(t))$ independent of $\mathbf{X}(s) = (X_1(s), \ldots, X_n(s))^\top$ and $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))^\top$, we provide lower and upper bounds for the prediction error $Y_{\text{pred}}(t) - Y_{\text{new}}(t)$, where $Y_{\text{pred}}(t)$ is given in (2.18). Here we consider the case where the predictor curve $X(s)$ is bounded. Recall that σ_k^2 is the kth eigenvalue of the covariance function of the signal function $S(t)$, and it is also the maximum value of the generalized eigenvalue problem (2.10) (see Theorem 2). We introduce a regularity condition on the decay rates of the σ_k^2's and $(\sigma_k^2 - \sigma_{k+1}^2)$'s, which is similar to the regularity conditions in Cai and Hall (2006) and Hall and Horowitz (2007) on the decay rates of the eigenvalues of the covariance function of the predictive curve.

Condition 1. *There exist constants $\theta > 1$ and $C > 1$ such that for any $k \ge 1$, we have $C^{-1}k^{-\theta} \le \sigma_k^2 \le Ck^{-\theta}$ and $\sigma_k^2 - \sigma_{k+1}^2 \ge C^{-1}k^{-(\theta+1)}$.*

Theorem 3. *Suppose that Condition* 1 *holds and* $F(x,s,t)$ *has continuous second partial derivatives. Suppose that* $0 \le X(s) \le 1$ *for all* $0 \le s \le 1$ *and* $\mathrm{E}(\|\varepsilon\|_{L^2}^4) < \infty$. *If the tuning parameters are chosen to satisfy* $\lambda = C_\lambda n^{-1/2}$ *and* $\kappa = C_\kappa n^{-1/2}$ *and the number of components satisfies* $K = C_K n^{1/\{2(3\theta+2)\}}$, *where* C_λ, C_κ *and* C_K *are constants not depending on* n, *for any* n *and* $\epsilon > 0$, *there exists an event* $\Omega_{n,\epsilon}$ *with* $P(\Omega_{n,\epsilon}) \ge 1-\epsilon$ *such that in* $\Omega_{n,\epsilon}$, *we have*

$$\frac{1}{n}\sum_{l=1}^{n}\left\|\int_0^1 \widehat{F}(X_l(s),s,\cdot)ds - \int_0^1 F(X_l(s),s,\cdot)ds\right\|_{L^2}^2 \le D_1 n^{-(\theta-1)/2(3\theta+2)}, \quad (2.19)$$

$$\mathrm{E}\left(\|\varepsilon\|_{L^2}^2\right) \le \mathrm{E}\left\{\|Y_{\mathrm{pred}} - Y_{\mathrm{new}}\|_{L^2}^2 \,\middle|\, \mathbf{X}(s), \mathbf{Y}(t)\right\} \le \mathrm{E}\left(\|\varepsilon\|_{L^2}^2\right) + D_2 n^{-(\theta-1)/2(3\theta+2)}, \quad (2.20)$$

where D_1 *and* D_2 *are constants which depend on* ϵ, C_λ, C_κ, m *and* δ, *but not on* n.

The inequalities in (2.20) show that a lower bound of the prediction error is $\mathrm{E}\left(\|\varepsilon\|_{L^2}^2\right)$ that is due to the noise function. As $n \to 0$, the prediction error converges to the lower bound $\mathrm{E}\left(\|\varepsilon\|_{L^2}^2\right)$ at the same rate as that of the estimation error in (2.19). In our method, we do not make specific assumptions on the correlation structure of the error function $\epsilon(t)$. In Theorem 3, we only require that $\mathrm{E}(\|\varepsilon\|_{L^2}^4) < \infty$. Therefore, various structures of the within-function correlation $\mathrm{Corr}(\varepsilon(t), \varepsilon(t'))$ are allowed for $t \ne t'$ and the variance $\mathrm{Var}(\varepsilon(t))$ can vary with t, which is one aspect of the flexibility of our method.

3. Nonlinear Regression with Multiple Functional Predictors

We extend the method in Section 2 to the model (1.1) with multiple functional predictors. Without loss of generality, we assume that $\mathrm{E}\{F_j(X(s),s,t)\} = 0$, for all $0 \le s \le 1$, $a \le t \le b$ and $1 \le j \le p$. Let $S(t) = \int_0^1 \sum_{j=1}^p F_j(X_j(s),s,t)ds$ be the signal function with KL expansion $S(t) = \sum_{k=1}^\infty \mathbf{r}_k \phi_k(t)$, where the random variables $\mathbf{r}_k$'s are uncorrelated with zero mean and unit variance, and $\|\phi_k\|_{L^2} = \sigma_k$. Let $\mathbf{X}(s) = (X_1(s), \ldots, X_p(s))$ be the vector of p functional predictors (at the population level). As an analogue to $G_k(x,s)$ when $p = 1$, we define the p-dimensional vector $\mathbf{G}_k(\mathbf{x},s) = (G_{k1}(x_1,s), \ldots, G_{kp}(x_p,s))^\top$, where $G_{kj}(x_j,s) = \int_a^b F_j(x_j,s,t)\phi_k(t)dt/\sigma_k^2$ for $k \ge 1$ and $1 \le j \le p$, and $\mathbf{x} = (x_1, \ldots, x_p)$. For any $\mathbf{G}(\mathbf{x},s) = (G_1(x_1,s), \ldots, G_p(x_p,s))^\top$, we extend the definition of $\mathbf{r}(G)$ in (2.4) to $\mathbf{r}(\mathbf{G}) = \sum_{j=1}^p \left[\int_0^1 G_j(X_j(s),s)ds - \mathrm{E}\{\int_0^1 G_j(X_j(s),s)ds\}\right]$. Then as $\mathrm{E}\{\int_0^1 G_{kj}(X_j(s),s)ds\} = 0$, $\mathbf{r}_k = \int_a^b S(t)\phi_k(t)dt/\sigma_k^2 = \sum_{j=1}^p \int_0^1 G_{kj}(X_j(s),s)ds = \mathbf{r}(G_k)$.

To estimate $\mathbf{G}_k(\mathbf{x},s)$, let $Y_l(t)$ and $\mathbf{X}_l(s) = (X_{l1}(s), \ldots, X_{lp}(s))$, $1 \le l \le n$,

be the i.i.d samples. For any $\mathbf{G}(\mathbf{x}, s) = (G_1(x_1, s), \ldots, G_p(x_p, s))^\top$ and $\widetilde{\mathbf{G}}(\mathbf{x}, s) = (\widetilde{G}_1(x_1, s), \ldots, \widetilde{G}_p(x_p, s))^\top$, let

$$\widehat{\mathbf{\Lambda}}(\mathbf{G}, \mathbf{G}) = \frac{1}{n^2} \int_a^b \left(\sum_{l=1}^{n} \left[\sum_{j=1}^{p} \int_0^1 \left\{ G_j(X_{lj}(s), s) - \overline{G}_j(s) \right\} \right] \left\{ Y_l(t) - \overline{Y}(t) \right\} \right)^2 dt, \tag{3.1}$$

$$\widehat{\mathbf{\Sigma}}(\mathbf{G}, \widetilde{\mathbf{G}}) = \frac{1}{n} \sum_{l=1}^{n} \left[\sum_{j=1}^{p} \int_0^1 \left\{ G_j(X_{lj}(s), s) - \overline{G}_j(s) \right\} \right] \left[\sum_{j=1}^{p} \int_0^1 \left\{ \widetilde{G}_j(X_{lj}(s), s) - \overline{\widetilde{G}}_j(s) \right\} \right],$$

where $\overline{G}_j(s) = \sum_{l=1}^{n} G_j(X_{lj}(s), s)/n$ and $\overline{\widetilde{G}}_j(s) = \sum_{l=1}^{n} \widetilde{G}_j(X_{lj}(s), s)/n$. We get the estimate $\widehat{\mathbf{G}}_k(\mathbf{x}, s) = (\widehat{G}_{k1}(x_1, s), \ldots, \widehat{G}_{kp}(x_p, s))$ of $\mathbf{G}_k(\mathbf{x}, s)$ by solving

$$\max_{\mathbf{G}=(G_1,\ldots,G_p)^\top} \frac{\widehat{\mathbf{\Lambda}}(\mathbf{G}, \mathbf{G})}{\widehat{\mathbf{\Sigma}}(\mathbf{G}, \mathbf{G}) + \lambda \sum_{j=1}^{p} \|G_j\|_{H^2}^2}, \quad \text{subject to}$$

$$\widehat{\mathbf{\Sigma}}(\mathbf{G}_{k'}, \mathbf{G}) + \lambda \sum_{j=1}^{p} \langle G_j, \widehat{G}_{k'j} \rangle_{H^2} = 0, \tag{3.2}$$

for all $1 \le k' \le k-1$. Using the tensor product basis functions, we can express (3.2) as a multivariate generalized eigenvalue problem in the same way as in Section 2.2. The estimates $\widehat{\mu}(t)$ and $\widehat{\phi}_k(t)$ of $\mu(t)$ and $\phi_k(t)$ are obtained as in Section 2.3. Given a new observed functional predictor vector $\mathbf{X}_{\text{new}}(s) = (X_{\text{new},1}(s), \ldots, X_{\text{new},p}(s))$, the predicted response function is $Y_{\text{pred}}(t) = \widehat{\mu}(t) + \sum_{k=1}^{K} \int_0^1 \sum_{j=1}^{p} \left\{ \widehat{G}_{kj}(X_{\text{new},j}(s), s) - \overline{\widehat{G}}_{kj}(s) \right\} \widehat{\phi}_k(t) ds$.

4. Simulation Studies

We evaluated the predictive performance of the proposed method in three sets of simulation studies. In the first study, there was only one functional predictor and we considered both linear (in x) and nonlinear forms of $F(x, s, t)$. When $F(x, s, t)$ was linear in x, we compared our method (denoted by *SigComp.nonlinear*) with the nonlinear method in Scheipl, Staicu and Greven (2015) which is based on simultaneous basis expansion on x, s, and t (denoted by *pffr.nonlinear*), and three methods for linear function-on-function regression models: the linear regression based on signal compression (*SigComp.linear*) in Luo and Qi (2017), the penalized function-on-function regression (*pffr*) in Ivanescu et al. (2015) and the pffr with eigenbasis (*pffr.pc*) in Scheipl, Staicu and Greven (2015). When $F(x, s, t)$ was nonlinear, the three linear methods served as benchmarks. In the last two studies, we considered multiple functional predictors.

To solve the optimization problems (2.14) and (3.2), we used 30 cubic spline basis functions in $[0,1]$ as $\mathbf{b}(s)$, and 30 cubic spline basis functions in $[0,1]$ as $\mathbf{h}(x)$, both with equally spaced knots. To solve (2.17), we used 30 cubic spline basis functions for $0 \le t \le 1$ with equally spaced knots. The tuning parameters λ in (2.14) and (3.2) and κ in (2.17) were both chosen from the set $\{10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$ using the cross-validation method described in Section S.2.2 in the supplementary material. The *SigComp.linear* was implemented in R and we used 30 cubic B-spline basis functions for both s and t, and chose both smoothness parameters λ and κ from $\{10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$ by five-fold cross-validation. For *pffr.nonlinear*, *pffr* and *pffr.pc*, we used the R package `refund`. For *pffr* and *pffr.pc*, we used their default settings except that we used 30 basis functions. Due to the computational cost of the *pffr.nonlinear*, it was impossible to use 30 basis functions for each of x, s and t, and the R function for the *pffr.nonlinear* does not provide options for changing the number of basis functions. Hence, we used its default setting. In all studies, both the functional predictors and the response curves were observed at 100 equally spaced time points in [0,1]. We ran 100 simulations for each setting with the sample size fixed at 100 for training data and 500 for test data in each run.

4.1. The case of single functional predictor

Simulation 1: We generated the single predictor $X(s)$ from the Gaussian process with mean zero and covariance function $\mathbf{\Sigma}_X(s,s') = e^{-\{10|s-s'|\}^2}$. We considered four models with different $\mu(t)$ and $F(x,s,t)$, as follows

$$
\begin{aligned}
&(1).\ \mu(t) = 3(t+1)^2, \quad F(x,s,t) = 3x\{s\sin(6\pi t) + 3\cos(\pi s t^2)\},\\
&(2).\ \mu(t) = \sin(\pi t), \quad F(x,s,t) = 27\cos\left((x+s^2)\sqrt{t}\right),\\
&(3).\ \mu(t) = \log(1+t), \quad F(x,s,t) = \frac{32(xs+t)}{1+(x^2+s^2)e^{-t}},\\
&(4).\ \mu(t) = e^{-t}, \quad F(x,s,t) = 12\log\left(1+5x^2se^t+3e^{-(x+s)}\cos^2(t)\right).
\end{aligned}
$$

The first $F(x,s,t)$ is linear in x and leads to a linear function-on-function model. The other three $F(x,s,t)$'s are fully nonlinear functions for all arguments x,s,t. The noise function $\varepsilon(t)$ was generated from a Gaussian process with covariance function $\mathbf{\Sigma}_\epsilon(t_1,t_2) = \sigma^2\rho^{\{30(t_1-t_2)\}^2}$ for $0 \le t_1, t_2 \le 1$, where σ^2 is the variance of $\varepsilon(t)$, and ρ controls the correlation between $\varepsilon(t_1)$ and $\varepsilon(t_2)$, $0 \le t_1 \ne t_2 \le 1$. We considered noise levels $\sigma^2 = 0.1, 10$ and correlation levels $\rho = 0, 0.7$. When $\sigma^2 = 10$, the signal to noise ratio is about one for all four models. The sample

Table 1. The averages (standard deviations) of MSMEs of 100 replicates for Simulation 1.

Model	σ^2	ρ	SigComp.nonlin	pffr.nonlin	SigComp.lin	pffr.pc	pffr
Estimation Error (MSME)							
1	0.1	0	0.003(0.001)	0.206(0.014)	0.001(0)	0.003(0)	0.004(0)
		0.7	0.010(0.002)	0.212(0.013)	0.006(0.001)	0.016(0.002)	0.017(0.002)
	10	0	0.052(0.010)	0.228(0.017)	0.048(0.012)	0.113(0.016)	0.055(0.009)
		0.7	0.266(0.073)	1.208(0.375)	0.271(0.068)	1.196(0.165)	1.250(0.198)
2	0.1	0	0.009(0.004)	0.022(0.018)	7.111(0.711)	7.961(0.791)	8.913(1.020)
		0.7	0.026(0.008)	0.036(0.018)	6.956(0.694)	7.777(0.752)	8.785(1.019)
	10	0	0.084(0.023)	0.078(0.023)	7.015(0.721)	7.778(0.809)	8.213(1.083)
		0.7	0.317(0.106)	1.075(0.372)	7.106(0.801)	8.876(1.024)	9.283(1.156)
3	0.1	0	0.004(0.001)	0.265(0.134)	1.972(0.172)	2.224(0.223)	2.496(0.303)
		0.7	0.014(0.005)	0.283(0.132)	2.012(0.196)	2.288(0.232)	2.596(0.329)
	10	0	0.087(0.019)	0.222(0.077)	2.010(0.182)	2.278(0.215)	2.286(0.250)
		0.7	0.430(0.125)	1.290(0.448)	2.094(0.187)	3.395(0.338)	3.588(0.402)
4	0.1	0	0.010(0.002)	0.527(0.329)	9.217(0.956)	10.349(1.008)	12.337(1.461)
		0.7	0.029(0.006)	0.568(0.433)	9.088(0.783)	10.166(0.927)	12.197(1.365)
	10	0	0.172(0.040)	0.459(0.264)	9.289(0.885)	10.495(0.966)	11.685(1.669)
		0.7	0.586(0.121)	1.693(0.717)	9.176(0.712)	11.072(1.009)	12.185(1.579)

curves of $X(s)$ and $Y(t)$ are plotted in Figures S.1 and S.2 in the supplementary material, respectively.

We considered 16 combinations of the four models, two σ^2 values and two ρ values. For each method, applying the final model obtained from the training data to the predictive curves $X_l^{\text{test}}(t)$, $1 \le l \le 500$ in the test data, we obtained 500 predicted response curves $Y_l^{\text{predict}}(t)$, $1 \le l \le 500$. We calculated the mean squared prediction error (MSPE) $\text{MSPE} = (1/500)\sum_{l=1}^{500}\big[(1/100)\sum_{m=1}^{100}\big\{Y_l^{\text{predict}}(t_m) - Y_l^{\text{test}}(t_m)\big\}^2\big]$, and the mean squared model estimation error (MSME) $\text{MSME} = (1/500)\sum_{l=1}^{500}\big[(1/100)\sum_{m=1}^{100}\big\{Y_l^{\text{predict}}(t_m) - \mu(t) - \int_0^1 F(X_l^{\text{test}}(s), s, t_m)(s)ds\big\}^2\big] = (1/500)\sum_{l=1}^{500}\big[(1/100)\sum_{m=1}^{100}\big\{Y_l^{\text{predict}}(t_m) - Y_l^{\text{test}}(t_m) + \varepsilon_l^{\text{test}}(t_m)\big\}^2\big]$, where t_m, $1 \le m \le 100$, were 100 equally spaced observation points for $Y(t)$.

We report the averages and standard deviations of the MSMEs and MSPEs in 100 iterations in Tables 1 and 2, respectively. For model 1, a linear function-on-function regression model, when $\rho = 0$, *SigComp.nonlinear* and all the linear methods have very close prediction and estimation errors that are less than *pffr.nonlin*. The *pffr.pc* and *pffr* are sensitive to the within function correlation in $\varepsilon(t)$ and have larger prediction and estimation errors than *SigComp.linear* and *SigComp.nonlinear* when $\rho = 0.7$. When $F(x, s, t)$ is nonlinear with respect to x as in Models $2 \sim 4$, it is not surprising that the three linear methods have

Table 2. The averages (standard deviations) of MSPEs of 100 replicates for Simulation 1.

Model	σ^2	ρ	SigComp.nonlin	pffr.nonlin	SigComp.lin	pffr.pc	pffr
Prediction Error (MSPE)							
1	0.1	0	0.103(0.001)	0.306(0.014)	0.101(0.001)	0.103(0.001)	0.104(0.001)
		0.7	0.110(0.002)	0.312(0.013)	0.105(0.002)	0.116(0.002)	0.117(0.003)
	10	0	10.047(0.062)	10.224(0.06)	10.043(0.063)	10.108(0.065)	10.049(0.063)
		0.7	10.279(0.167)	11.212(0.383)	10.281(0.158)	11.217(0.251)	11.271(0.266)
2	0.1	0	0.109(0.005)	0.123(0.018)	7.211(0.712)	8.061(0.791)	9.013(1.02)
		0.7	0.126(0.008)	0.136(0.018)	7.056(0.699)	7.877(0.755)	8.885(1.021)
	10	0	10.093(0.067)	10.086(0.064)	17.033(0.755)	17.796(0.84)	18.231(1.105)
		0.7	10.324(0.185)	11.091(0.397)	17.121(0.862)	18.888(1.064)	19.299(1.173)
3	0.1	0	0.104(0.001)	0.365(0.134)	2.072(0.171)	2.324(0.223)	2.596(0.303)
		0.7	0.114(0.005)	0.384(0.132)	2.11 (0.199)	2.387(0.234)	2.694(0.332)
	10	0	10.086(0.068)	10.22(0.108)	12.012(0.198)	12.282(0.228)	12.289(0.27)
		0.7	10.410(0.227)	11.284(0.501)	12.065(0.287)	13.373(0.404)	13.566(0.459)
4	0.1	0	0.110(0.002)	0.628(0.329)	9.318(0.957)	10.449(1.009)	12.438(1.462)
		0.7	0.129(0.007)	0.668(0.432)	9.195(0.784)	10.272(0.925)	12.304(1.369)
	10	0	10.172(0.077)	10.461(0.274)	19.288(0.875)	20.496(0.96)	21.681(1.667)
		0.7	10.558(0.209)	11.658(0.706)	19.204(0.766)	21.117(1.051)	22.225(1.61)

much larger mean MSPE and MSME than the nonlinear methods. Our method is not very sensitive to the form of the function $F(x, s, t)$ and has the smallest mean MSPE and MSME in all settings of Models 2 $\sim$ 4, except in the setting of $\sigma^2 = 10$ and $\rho = 0$ for Model 2, where *pffr.nonlin* is slightly better than *SigComp.nonlinear*. A larger noise level or a stronger within-function correlation in $\varepsilon(t)$ leads to larger MSPEs and MSMEs of our method for all the four models.

We also compared the estimated function $\widehat{F}(x, s, t)$ of the function $F(x, s, t)$ for different methods. Due to the identifiability problem of $F(x, s, t)$ and other technical issues, we provide the details of calculating the estimation error for $F(x, s, t)$ and present the results in Section S4 of the supplementary material.

We summarize the numbers of selected components and running time (in seconds) in Table S.1, and draw the boxplots of selected tuning parameters λ and κ by our method in Figures S.3$\sim$S.6 in the supplementary material. For all the four models, with the increase of the noise level and correlation in $\varepsilon(t)$, our method tends to choose larger λ and κ.

4.2. The case of multiple functional predictors

To model the correlation between functional predictors, we took $V_j(s)$, $1 \leq j \leq p$, to be identical and independent Gaussian processes with covariance function $\mathbf{\Sigma}_X(s, s') = e^{-\{10|s-s'|\}^\gamma}$, where $\gamma = 1.5$ or 2. The corresponding Gaussian

Table 3. The averages (standard deviations) of MSPEs of 100 replicates for Simulations 2 and 3.

			$\gamma = 2$		$\gamma = 1.5$	
ρ_{curve}	σ^2	ρ	SigComp.nonlinear	SigComp.linear	SigComp.nonlinear	SigComp.linear
Simulation 2						
0	0.1	0	0.136(0.019)	4.359(0.508)	0.128(0.012)	3.985(0.469)
		0.7	0.170(0.029)	4.467(0.496)	0.160(0.029)	3.928(0.433)
	10	0	10.205(0.075)	14.369(0.418)	10.188(0.068)	13.887(0.418)
		0.7	10.687(0.224)	14.578(0.471)	10.696(0.271)	14.201(0.544)
0.7	0.1	0	0.139(0.022)	5.012(0.589)	0.129(0.015)	4.551(0.525)
		0.7	0.169(0.023)	4.965(0.511)	0.162(0.022)	4.450(0.441)
	10	0	10.206(0.079)	14.946(0.689)	10.190(0.078)	14.381(0.492)
		0.7	10.801(0.267)	15.026(0.640)	10.882(0.276)	14.744(0.532)
Simulation 3						
0	0.1	0	0.203(0.023)	2.805(0.319)	0.184(0.015)	2.575(0.327)
		0.7	0.238(0.027)	2.788(0.277)	0.224(0.019)	2.515(0.239)
	10	0	10.368(0.088)	12.744(0.298)	10.357(0.085)	12.486(0.232)
		0.7	11.424(0.465)	13.120(0.341)	11.345(0.426)	12.903(0.379)
0.7	0.1	0	0.195(0.026)	3.074(0.319)	0.180(0.019)	2.801(0.304)
		0.7	0.228(0.024)	3.084(0.338)	0.218(0.021)	2.817(0.329)
	10	0	10.327(0.087)	13.063(0.364)	10.323(0.075)	12.766(0.297)
		0.7	11.235(0.386)	13.265(0.426)	11.234(0.277)	13.018(0.435)

process with $\gamma = 2$ has smoother sample curves than those with $\gamma = 1.5$. Let $\mathbf{S}$ be the $p \times p$ matrix with diagonal entries equal to 1 and off-diagonal entries equal to $\rho_{\text{curve}} \in (0, 1)$ to control the correlation between functional predictors. We decompose $\mathbf{S} = \mathbf{\Delta}\mathbf{\Delta}^\top$, where $\mathbf{\Delta}$ is a $p \times p$ matrix. Then the predictor curves were obtained as $(X_1(s), X_2(s), \ldots, X_p(s)) = (V_1(s), V_2(s), \ldots, V_p(s))\mathbf{\Delta}^\top$. Given any s, $(X_1(s), X_2(s), \ldots, X_p(s))$ has a multivariate normal distribution with covariance matrix $\mathbf{S}$. Moreover, each $X_j(s)$ is a Gaussian process with covariance function $\mathbf{\Sigma}_X(s, s')$. We considered correlation levels: $\rho_{\text{curve}} = 0$ and 0.7. Because the function for *pffr.nonlinear* in the R package `refund` does not provide the options for multiple functional predictors, we only include the method *SigComp.linear* as a benchmark. We conducted two simulations for this case.

Simulation 2: We took $p = 3$ and with $F_1(x, s, t) = 4(2s - x - 2\sqrt{t})^2$, $F_2(x, s, t) = 2\sqrt{(s + x^2 t)/\{1 + \sin^2(x + st)\}}$ and $F_3(x, s, t) = 2\log(1 + (xt)^2\sqrt{s})$.

Simulation 3: We considered $p = 6$ and took the nonlinear functions

$$F_1(x, s, t) = 2.7(s + x - 3t)^2, \qquad F_2(x, s, t) = 2.7\cos(2xs) + 2.7\sin(x\sqrt{t}),$$
$$F_3(x, s, t) = \frac{2.7}{1 + x^2 + st}, \qquad F_4(x, s, t) = 2.7e^{-\cos(x+s-2t^2)},$$
$$F_5(x, s, t) = 2.7(s^2 + xt^2)(sx + t^3), \; F_6(x, s, t) = 2.7\log(1 + s^2 + 0.5x^2 + t^2).$$

The noise functions $\varepsilon(t)$ in both simulations were the same as in Simulation 1, with $\sigma^2 = 0.1$ or 10, $\rho = 0$ or 0.7. The signal to noise ratio is about one when $\sigma^2 = 10$. We report the MSPEs in Table 3. It is unsurprising that the linear method fails in both simulations. The increase of σ^2 or ρ leads to the increase of MSPE of *SigComp.nonlinear*. For Simulation 2, the correlation between functional predictors ρ_{curve} has little effect on MSPE. For Simulation 3, the correlation between functional predictors slightly decreases the MSPE, and smaller prediction errors are observed for less smooth functional predictors ($\gamma = 1.5$). We summarize the numbers of selected components K and running time in Table S.2 in the supplementary material. The means of K are almost the same for all settings in Simulation 2 except the case of $\sigma^2 = 10$ and $\rho = 0.7$, where more components are chosen. For Simulation 3, more components are also chosen for $\sigma^2 = 10$ and $\rho = 0.7$.

5. Application to a Daily Air Quality Dataset

The *Air Quality* dataset, available in UCI Machine Learning Repository (Lichman (2013)), was recorded by an array of five metal oxide chemical sensors embedded in an air quality chemical multisensor device located in a significantly polluted area, at road level, within an Italian city (De Vito et al. (2008)). This dataset contains the hourly averages of the concentration values of five different atmospheric pollutants in each day. The five pollutants are the nitrogen dioxide (NO_2), carbon monoxide (CO), non-methane hydrocarbons (NMHC), total nitrogen oxides (NO_x), and benzene (C_6H_6). In addition, the temperature (in Celsius) and relative humidity (in percentage) were also recorded hourly in each day. Therefore, we have seven functional variables with sample curves observed at 24 discrete time points. Removing missing values, we have 355 sample curves for each of the seven functional variables. We plot all sample curves in Figure S.7 in the supplementary material.

NO_2 is a gaseous pollutant commonly used in health effects assessments. Researchers are interested in the relationship between NO_2 and other traffic pollutants (Beckerman et al. (2008)). We investigated to what extent we can predict the daily curve of NO_2 by the daily curves of the other four pollutants together with the temperature and relative humidity. We first considered some evidence on nonlinear relationship between the NO_2 curves and the six predictor curves. Due to the lack of hypothesis testing method on nonlinearity for function-on-function regression in the literature, we cannot test the nonlinearity as in

model (1.1) directly. However, if the true relationship between the response and predictor curves is linear, the value of the response function at each individual observation point is also linearly related to the predictor curves. Based on this fact, we performed a restricted likelihood ratio test (Crainiceanu and Ruppert (2004)) on the linear relationship between each individual value of NO_2 and the predictor curves. The restricted likelihood ratio test was implemented in the R package `RLRsim` (Scheipl, Greven and Kuechenhoff (2008)). The p-values of these tests for 24 individual observations of NO_2 were all very close to zero ($< 2.2 \times 10^{-16}$), which provides strong evidence to reject the null hypothesis of linearity and indicates nonlinear relationship between NO_2 and other curves.

In addition to the nonlinear additive model with six predictor curves, we also fit the linear function-on-function model using *SigComp.linear*. To evaluate the importance of each individual functional predictor, we removed one predictor at a time and fit a reduced nonlinear model using the other five predictors. To evaluate the predictive performance of these models, we randomly split the data into the training set of size 250 and the test set of size 105, and repeated this 100 times. The average MSPE in 100 repeats was 4.9×10^{-3} (standard deviation 0.66×10^{-3}) for the full nonlinear model, 5.1×10^{-3} (0.84×10^{-3}) for the linear model, 5.3×10^{-3} (0.56×10^{-3}) for the reduced nonlinear model without CO, 6.4×10^{-3} (0.78×10^{-3}) without NMHC, 6.7×10^{-3} (0.56×10^{-3}) without NO_x, 5.1×10^{-3} (0.64×10^{-3}) without C_6H_6, 11.3×10^{-3} (1.55×10^{-3}) without temperature, and 7.34×10^{-3} (0.83×10^{-4}) without relative humidity. The full nonlinear model had the smallest mean MSPE and was chosen as the final model. For the full nonlinear model, in each of the 100 repeats, we calculated the averaged functional R^2 (Meyer et al. (2015))

$$R^2_{\text{ave}} = \int_0^1 R^2(t)dt \approx \frac{1}{24}\sum_{m=1}^{24}\left[1 - \frac{\sum_{i=1}^{250}\{Y_{\text{train},i}(t_m) - \widehat{Y}_i(t_m)\}^2}{\sum_{i=1}^{250}\{Y_{\text{train},i}(t_m) - \overline{Y}_{\text{train}}(t_m)\}^2}\right],$$

where $Y_{\text{train},i}(t)$ is the i-th response curve in the training set, $\widehat{Y}_i(t)$ is the corresponding fitted curve, $\overline{Y}_{\text{train}}(t)$ is the sample mean of the response curves in the training set and $0 = t_1 < \cdots < t_{24} = 1$ are 24 equally spaced observation time points. The average and standard deviation of the R^2_{ave} over 100 repeats were 94.8% and 0.4%, respectively.

For the full nonlinear model, our method chose $\lambda = \kappa = 10^{-6}$ in all 100 replicates, and six components in 90 replicates. So we used all the data to build the final model with $\lambda = \kappa = 10^{-6}$ and 6 components. Due to the possible unidentifiable problem of F_j, we used $\widehat{F}_j^{(DISC)}$ to emphasize that we estimate

$F_j^{(DISC)}$ (Section 2.1) which may differ from the true nonlinear function. We plot the estimated $\widehat{\mu}(t)$ in Figure S.8, and the estimated $\widehat{F}_j^{(DISC)}(x, s, 12)$, $1 \leq j \leq 6$, with $t = 12$, in Figure S.9 of the supplementary material, which shows the nonlinear patterns of these estimated functions with respect to (x, s). To take a closer look at the nonlinear relationship between NO_2 and other pollutants, we considered NO_x (X_3) and studied the nonlinear pattern of $\widehat{F}_3^{(DISC)}(x, s, t)$ with respect to x. We draw the curves of $\widehat{F}_3^{(DISC)}(x, s, t)$ versus x for different values of s and t in Figure S.10 of the supplementary material. Most of these curves increase for relatively smaller x (lower concentration of NO_x), reach a maximum at a value between 7 and 7.5 and then decrease for higher concentration of NO_x. These nonlinear patterns may reflect the change of the relationship between NO_2 and NO_x for different concentration levels of NO_x. To see this, we pooled all the pairs of hourly concentration values of NO_2 and NO_x, and calculated the correlation coefficient using all the pairs with the concentration value of NO_2 less than 7 and the correlation coefficient using all the pairs with the concentration value of NO_2 greater than or equal to 7, respectively. The two correlation coefficients were -0.49 and -0.26, respectively, implying a great change of the correlation between NO_2 and NO_x with the increase of concentration level of NO_x.

6. Discussion

We consider an additive nonlinear function-on-function regression model with multiple functional predictors. The unspecified forms of the nonlinear functions offer great flexibility to model various relationships between the functional response and predictors. Without an identifiability assumption, we introduce specific nonlinear functions which lead to the same model as the true model and possess a minimum prediction error property. We focus on estimating these specific nonlinear functions, which leads to high accuracy in estimating the signal function and predicting the response function.

Our two-step estimation procedure consists of a penalized functional generalized eigenvalue problem and a penalized least squares procedure. Our approach breaks down the estimation of the trivariate nonlinear functions into problems of estimating bivariate functions and univariate functions separately, which considerably reduces computational costs in each step and allows multiple predictor curves. This generalizes the signal compression method in Luo and Qi (2017) from linear function-on-function regression to nonlinear function-on-function re-

gression.

As mentioned by a referee, an important issue is to test whether there is a significant nonlinear relationship between the response curve and the predictor curves. This has not been studied in the literature and deserves investigation.

Supplementary Materials

The supplementary material includes additional figures, tables and proofs.

References

Ait-Saïdi, A., Ferraty, F., Kassa, R. and Vieu, P. (2008). Cross-validated estimations in the single-functional index model. *Statistics* **42**, 475–494.

Beckerman, B., Jerrett, M., Brook, J. R., Verma, D. K., Arain, M. A. and Finkelstein, M. M. (2008). Correlation of nitrogen dioxide with other traffic pollutants near a major expressway. *Atmospheric Environment* **42**, 275–290.

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

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics* **34**, 2159–2179.

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

Chen, D., Hall, P. and Müller, H. G. (2011). Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics* **39**, 1720–1747.

Chiou, J.-M., Yang, Y.-F. and Chen, Y.-T. (2016). Multivariate functional linear regression and prediction. *Journal of Multivariate Analysis* **146**, 301–312.

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **66**, 165–185.

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

Delaigle, A., Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics* **40**, 322–352.

Ferraty, F., Goia, A., Salinelli, E. and Vieu, P. (2013). Functional projection pursuit regression. *Test* **22**, 293–320.

Ferraty, F., Park, J. and Vieu, P. (2011). Estimation of a Functional Single Index Model. In: Recent Advances in Functional Data Analysis and Related Topics. Springer, 111–116.

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

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics* **35**, 70–91.

Ivanescu, A. E., Staicu, A.-M., Scheipl, F. and Greven, S. (2015). Penalized function-on-function regression. *Computational Statistics* **30**, 539–568.

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

Lichman, M. (2013). UCI machine learning repository.

Luo, R. and Qi, X. (2017). Function-on-function linear regression by signal compression. *Journal of the American Statistical Association* **112**, 690–705.

Luo, R., Qi, X. and Wang, Y. (2016). Functional wavelet regression for linear function-on-function models. *Electronic Journal of Statistics* **10**, 3179–3216.

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

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

Müller, H.-G., Wu, Y. and Yao, F. (2013). Continuously additive models for nonlinear functional regression. *Biometrika* **100**, 607–622.

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B. (Statistical Methodological)* **53**, 539–572.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd Edition. Springer: New York.

Ratcliffe, S. J., Leader, L. R. and Heller, G. Z. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. i: Functional regression. *Statistics in Medicine* **21**, 1103–1114.

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

Scheipl, F., Greven, S. and Kuechenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis* **52**, 3283–3299.

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

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54**, 1461–1481.

Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics* **33**, 2873–2903.

Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, USA.

E-mail: xqi3@gsu.edu

Division of Epidemiology and Biostatistics, School of Public Health, Georgia State University, Atlanta, GA 30303, USA.

E-mail: rluo@gsu.edu

(Received June 2017; accepted August 2017)