Interaction model and model selection for function-on-function regression

Ruiyan Luo
Division of Epidemiology and Biostatistics, School of Public Health,
Georgia State University
and
Xin Qi
Department of Mathematics and Statistics,
Georgia State University

May 8, 2018

Abstract

Regression models with interaction effects have been widely used in multivariate analysis to improve model flexibility and prediction accuracy. In functional data analysis, however, due to the challenges of estimating three-dimensional coefficient functions, interaction effects have not been considered for function-on-function linear regression. In this paper, we propose function-on-function regression models with interaction and quadratic effects. For a model with specified main and interaction effects, we propose an efficient estimation method which enjoys a minimum prediction error property and has good predictive performance in practice. Moreover, converting the estimation of three-dimensional coefficient functions of the interaction effects to the estimation of two- and one-dimensional functions separately, our method is computationally efficient. We also propose adaptive penalties to account for varying magnitudes and roughness levels of coefficient functions. In practice, the forms of the models are usually unspecified. We propose a stepwise procedure for model selection based on a predictive criterion. This method is implemented in our R package `FRegSigComp`.

Keywords: adaptive penalty; function-on-function regression; interaction and quadratic effects; model selection; stepwise procedure

1 Introduction

Functional data with multiple curves have been collected in various areas such as medical science, public health and environmental research. For example, ocean data can contain sample curves of salinity, density, temperature, oxygen, and chloropigment as functions of depth, and air pollution data can include daily curves of the concentration levels of multiple air pollutants. The function-on-function linear regression is a tool in functional data analysis to study the association between multiple curves. Various approaches have been proposed for function-on-function regression, such as Ramsay and Dalzell (1991), Ramsay and Silverman (2005), Yao *et al.* (2005), Wu and Müller (2011), Wang (2014), Ivanescu *et al.* (2014), Scheipl *et al.* (2015), Meyer *et al.* (2015), Luo and Qi (2017), Luo *et al.* (2016), and Qi and Luo (Accepted). All these existing methods only consider models with main effects. In an arXiv paper, Matsui (2017) considered a function-on-function regression model with one predictor curve $X(s)$ and its quadratic effect of the form $\int\int X(r)X(s)\gamma(r,s,t)drds$. However, to the best of our knowledge, the existing methods do not consider multiple predictor curves with their interaction effects, nor explore the model selection.

To improve model flexibility and prediction accuracy, interaction effects are very common in multiple regression models, and model selection is a crucial step in data analysis. In functional data analysis, interactions between functional predictors have been considered by a few authors in scalar-on-function regression, where the response is scalar and the predictors are functional. Li and Marx (2008) considered the quadratic term $\int\{X(t)\}^2\gamma(t)dt$, and Yao and Müller (2010) considered the full quadratic effect $\int\int X(s)X(t)\gamma(s,t)dsdt$, where $X(s)$ and $X(t)$ represents the same functional predictor. Usset *et al.* (2016) considered the interaction terms of the form $\int\int X_i(s)X_j(t)\gamma_{ij}(s,t)dsdt$ where $X_i(s)$ and $X_j(t)$ represent different functional predictors. Model selection for the scalar-on-function regression have been considered in Zhu *et al.* (2010), Matsui and Konishi (2011), Gertheiss *et al.* (2013), Swihart *et al.* (2014), Fuchs *et al.* (2015), Collazos *et al.* (2016), Kong *et al.* (2016) and so on.

In this paper, we consider function-on-function regression models with quadratic and interaction effects. Suppose that we have p functional predictors $X_1(s),\cdots,X_p(s)$ and a response curve $Y(t)$. Without loss of generality, we assume that both s and t belong to the

interval $[0,1]$, and $X_i(s)$'s all have mean zeros. We consider the following model:

$$Y(t) = \mu(t) + \sum_{i\in\mathcal{M}} \int_0^1 X_i(s)\beta_i(s,t)ds + \sum_{(i,j)\in\mathcal{I}} \int_0^1 \int_0^1 X_i(u)X_j(v)\gamma_{ij}(u,v,t)dudv + \varepsilon(t). \quad (1)$$

The index set $\mathcal{M}$ of the main effects can be any subset of $\{1,\cdots,p\}$ and the index set $\mathcal{I}$ of the quadratic and interaction effects can be any subset of $\{(i,j) : 1 \le i \le j \le p\}$. We require that the two index sets satisfy the *hierarchical principle*: if a pair (i,j) belongs to $\mathcal{I}$, then both i and j must belong to $\mathcal{M}$. The coefficient functions for the main and interaction effects are represented as $\beta_i(s,t)$'s and $\gamma_{ij}(u,v,t)$'s, respectively. The noise $\varepsilon(t)$ has mean zero and is independent of $X_i(s)$'s. So the quadratic model in Matsui (2017) is a special case of (2) with one predictor ($p = 1$) and its the quadratic term. For convenience, we center the quadratic and interaction terms, and express this model as

$$Y(t) = \mu_0(t) + S(t) + \varepsilon(t), \quad \text{where} \quad S(t) = \sum_{i\in\mathcal{M}} \int_0^1 X_i(s)\beta_i(s,t)ds + \quad (2)$$
$$\sum_{(i,j)\in\mathcal{I}} \int_0^1 \int_0^1 \{X_i(u)X_j(v) - \mathrm{E}[X_i(u)X_j(v)]\}\,\gamma_{ij}(u,v,t)dudv,$$

where $\mu_0(t) = \mu(t) + \sum_{(i,j)\in\mathcal{I}} \int_0^1 \int_0^1 \mathrm{E}[X_i(u)X_j(v)]\gamma_{ij}(u,v,t)dudv$. The centered regression function, $S(t)$, is a quadratic function of $X_i(s)$'s, and has mean zero. Then $\mu_0(t)$ is the mean function of $Y(t)$. In classic multivariate linear regression, where both predictors and response variables are scalars, the existence of interaction between the i-th and j-th predictors implies that the slope coefficient of the i-th predictor can be different for different values of the j-th predictor. A similar interpretation can be applied to model (2). The existence of the (i,j)-th interaction leads to a change of the coefficient function of $X_i(s)$ by $\int_0^1 X_j(v)\gamma_{ij}(s,v,t)dv$.

The purpose of this paper is twofold. First, when the index sets $\mathcal{M}$ and $\mathcal{I}$ are given, we propose a method to estimate $\mu_0(t)$ and the quadratic regression function $S(t)$ in model (2), and provide a prediction procedure. The coefficient functions in model (2) are two-dimensional for main effects and three-dimensional for interactions. Direct estimation of these coefficient functions using basis expansions in two- and three-dimensional spaces, as in Matsui (2017), requires estimation of a large number of basis coefficients simultaneously, which imposes challenges on both the efficiency of computation and the accuracy of estimation and prediction. We propose a special representation of $S(t)$, which has a property

of minimum prediction error, and leads to good predictive performance of our approach. Our method transforms the estimation of two- and three-dimensional coefficient functions to the estimation of one- and two-dimensional functions separately, which greatly improves computational efficiency. We also propose adaptive penalties to account for varying magnitudes and roughness levels of coefficient functions. Second, as the form of the true model is usually unspecified and $\mathcal{M}$ and $\mathcal{I}$ are unknown in practice, we propose a stepwise procedure for model selection based on a predictive criterion. To illustrate our proposed estimation method and model selection procedure, we apply them to the Hawaii ocean data and the air pollution data.

The rest of the paper is organized as below. In Section 2, we introduce our estimation procedure for model (2) with a given form. In Section 3, we describe our stepwise procedure for model selection. In Section 4, we provide the methods to choose the adaptive weights in adaptive penalties and tuning parameters. In Sections 5 and 6, we evaluate the performance of the proposed method via simulation and applications to two real data sets. We summarize this paper in Section 7. Additional figures, tables, computational details and proofs are provided in supplementary material.

2 Estimation of model (2) with a specified form

In this section, we assume that the index sets $\mathcal{M}$ and $\mathcal{I}$ in model (2) are specified and propose a method to estimate the quadratic regression function $S(t)$ in (2).

2.1 A representation of $S(t)$ induced by its KL expansion

We consider the Karhunen-Loève (KL) expansion of $S(t)$:

$$S(t) = \sum_{k=1}^{\infty} \mathbf{z}_k w_k(t), \tag{3}$$

where $w_k(t)$'s are orthogonal eigenfunctions of the covariance function of $S(t)$ with corresponding eigenvalues $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq 0$, and $\mathbf{z}_k$'s are uncorrelated random variables. Usually, $w_k(t)$'s are chosen to have unit L^2 norms and then $\mathbf{z}_k$'s have variances equal to σ_k^2's. In this paper, for technical reasons, we scale $\mathbf{z}_k$'s to have unit variances and correspondingly,

we scale $w_k(t)$ to have the L^2 norm equal to σ_k, for $k \geq 1$. Then $S(t) = \sum_{k=1}^{\infty} \mathbf{z}_k w_k(t)$ is still the KL expansion, and it can be shown that $\mathbf{z}_k = \int_0^1 S(t) w_k(t) dt / \sigma_k^2$. The truncated KL expansion is the best finite dimensional approximation to $S(t)$ in that, for any K, $\sum_{k=1}^{K} \mathbf{z}_k w_k(t)$ has the smallest expected integrated squared approximation error to $S(t)$ among all K-dimensional approximations.

To simplify notation, let $\mathbf{\Phi} = \{\psi_i(s), i \in \mathcal{M}; \phi_{ij}(u,v), (i,j) \in \mathcal{I}\}$ denote a collection of one- and two-dimensional functions indexed by $\mathcal{M}$ and $\mathcal{I}$, respectively, where $\psi_i(s)$ and $\phi_{ij}(u,v)$ are arbitrary square integrable functions. Let $\mathbf{X} = (X_1(s), \ldots, X_p(s))^\top$ denote the collection of p functional predictors. We introduce a notation

$$\langle \mathbf{X}, \mathbf{\Phi} \rangle_{\mathcal{M},\mathcal{I}} = \sum_{i \in \mathcal{M}} \int_0^1 X_i(s) \psi_i(s) ds + \sum_{(i,j) \in \mathcal{I}} \int_0^1 \int_0^1 \{X_i(u) X_j(v) - \mathrm{E}[X_i(u) X_j(v)]\} \phi_{ij}(u,v) du dv.$$

By the definition of $S(t)$ in model (2), we have

$$\mathbf{z}_k = \int_0^1 S(t) w_k(t) dt / \sigma_k^2 = \langle \mathbf{X}, \mathbf{\Phi}_k \rangle_{\mathcal{M},\mathcal{I}}, \tag{4}$$

where

$$\begin{aligned} &\mathbf{\Phi}_k = \{\psi_{i,k}(s), i \in \mathcal{M}; \phi_{ij,k}(u,v), (i,j) \in \mathcal{I}\} \text{ with} \\ &\psi_{i,k}(s) = \int_0^1 \beta_i(s,t) w_k(t) dt / \sigma_k^2 \quad \text{and} \quad \phi_{ij,k}(u,v) = \int_0^1 \gamma_{ij}(u,v,t) w_k(t) dt / \sigma_k^2, \end{aligned} \tag{5}$$

and the $\beta_i(s,t)$'s and $\gamma_{ij}(u,v,t)$'s are the coefficient functions in model (2). Let

$$\begin{aligned} \beta_i^{(\mathrm{KL})}(s,t) &= \sum_{k=1}^{\infty} \psi_{i,k}(s) w_k(t), \quad \text{for all } i \in \mathcal{M}, \\ \gamma_{ij}^{(\mathrm{KL})}(u,v,t) &= \sum_{k=1}^{\infty} \phi_{ij,k}(u,v) w_k(t), \quad \text{for all } (i,j) \in \mathcal{I}, \end{aligned} \tag{6}$$

be the functions induced by the KL expansion of $S(t)$. A little algebra leads to the following representation of $S(t)$ using $\beta_i^{(\mathrm{KL})}(s,t)$ and $\gamma_{ij}^{(\mathrm{KL})}(u,v,t)$

$$\begin{aligned} S(t) = &\sum_{i \in \mathcal{M}} \int_0^1 X_i(s) \beta_i^{(\mathrm{KL})}(s,t) ds \\ &+ \sum_{(i,j) \in \mathcal{I}} \int_0^1 \int_0^1 \{X_i(u) X_j(v) - \mathrm{E}[X_i(u) X_j(v)]\} \gamma_{ij}^{(\mathrm{KL})}(u,v,t) du dv. \end{aligned} \tag{7}$$

Therefore, using the estimates of $\beta_i^{(\mathrm{KL})}(s,t)$'s and $\gamma_{ij}^{(\mathrm{KL})}(u,v,t)$'s and the representation (7), we can obtain the estimate of $S(t)$. A discussion of the coefficient identifiability of model (2) and the relationship between $\beta_i^{(\mathrm{KL})}(s,t)$, $\gamma_{ij}^{(\mathrm{KL})}(u,v,t)$'s and the original coefficients $\beta_i(s,t)$, $\gamma_{ij}(u,v,t)$'s is provided in Section S.2 of the supplementary material. In the following, we focus on the estimation of $\beta_i^{(\mathrm{KL})}(s,t)$ and $\gamma_{ij}^{(\mathrm{KL})}(u,v,t)$ which are represented as infinite series in (6). In practice, we estimate the truncated expansions. Given K, let

$$\begin{aligned}\beta_i^{(K)}(s,t) &= \sum_{k=1}^{K} \psi_{i,k}(s) w_k(t), \quad i \in \mathcal{M}, \\ \gamma_{ij}^{(K)}(u,v,t) &= \sum_{k=1}^{K} \phi_{ij,k}(u,v) w_k(t), \quad (i,j) \in \mathcal{I} \end{aligned} \tag{8}$$

denote the truncated series of $\beta_i^{(\mathrm{KL})}(s,t)$ and $\gamma_{ij}^{(\mathrm{KL})}(u,v,t)$, respectively, after the first K terms. Given a new observation, $\mathbf{X}_{\mathrm{new}} = (X_{\mathrm{new},1}(s), \ldots, X_{\mathrm{new},p}(s))^\top$ and $Y_{\mathrm{new}}(t)$, we consider the predicted function based on (8): $Y_{\mathrm{pred}}(t) = \mu_0(t) + \sum_{i\in\mathcal{M}} \int_0^1 X_{\mathrm{new},i}(s)\beta_i^{(K)}(s,t)ds + \sum_{(i,j)\in\mathcal{I}} \int_0^1 \int_0^1 \{X_{\mathrm{new},i}(u)X_{\mathrm{new},j}(v) - \mathrm{E}[X_{\mathrm{new},i}(u)X_{\mathrm{new},j}(v)]\}\gamma_{ij}^{(K)}(u,v,t)dudv$. The following theorem shows that this predicted function has the minimum prediction error among a large family of predicted functions.

Theorem 2.1 *The expected prediction error* $\mathrm{E}[\int_0^1 \{Y_{\mathrm{pred}}(t) - Y_{\mathrm{new}}(t)\}^2 dt]$ *is the smallest among all* $\mathrm{E}[\int_0^1 \{\widetilde{Y}_{\mathrm{pred}}(t) - Y_{\mathrm{new}}(t)\}^2 dt]$ *for* $\widetilde{Y}_{\mathrm{pred}}(t) = \mu_0(t) + \sum_{i\in\mathcal{M}} \int_0^1 X_{\mathrm{new},i}(s)\widetilde{\beta}_i^{(K)}(s,t)ds + \sum_{(i,j)\in\mathcal{I}} \int_0^1 \int_0^1 \{X_{\mathrm{new},i}(u)X_{\mathrm{new},j}(v) - \mathrm{E}[X_{\mathrm{new},i}(u)X_{\mathrm{new},j}(v)]\}\widetilde{\gamma}_{ij}^{(K)}(u,v,t)dudv$, *where* $\widetilde{\beta}_i^{(K)}(s,t) = \sum_{l=1}^K \sum_{k=1}^K a_{lk}^{(i)} \alpha_l(s)\delta_k(t)$ *and* $\widetilde{\gamma}_i^{(K)}(u,v,t) = \sum_{l=1}^K \sum_{k=1}^K b_{lk}^{(ij)} \eta_l(u,v)\delta_k(t)$ *are any basis expansions with arbitrary basis functions* $\alpha_l(s)$*'s,* $\eta_l(u,v)$*'s and* $\delta_k(t)$*'s, and arbitrary coefficients* $a_{lk}^{(i)}$*'s and* $b_{lk}^{(ij)}$*'s,* $1 \le k, l \le K$.

To estimate the truncated expansions in (8), we first estimate the $\mathbf{\Phi}_k$'s in (5), and then estimate $\mu_0(t)$ and the $w_k(t)$'s in (8). The estimation of $\mathbf{\Phi}_k$'s is based on the following Theorem 2.2 which characterizes $\mathbf{\Phi}_k$'s as solutions to generalized eigenvalue problems.

Theorem 2.2 *The* $\mathbf{\Phi}_k$*,* $k \ge 1$*, defined in* (5)*, is the solution to the following generalized eigenvalue problem,*

$$\begin{aligned} \max_{\mathbf{\Phi}} \quad & \int_0^1 \left\{\mathrm{E}\left[\langle \mathbf{X}, \mathbf{\Phi}\rangle_{\mathcal{M},\mathcal{I}} S(t)\right]\right\}^2 dt, \\ \textit{subject to} \quad & \mathrm{E}\left[\{\langle \mathbf{X}, \mathbf{\Phi}\rangle_{\mathcal{M},\mathcal{I}}\}^2\right] = 1, \quad \mathrm{E}\left[\langle \mathbf{X}, \mathbf{\Phi}\rangle_{\mathcal{M},\mathcal{I}} \langle \mathbf{X}, \mathbf{\Phi}_{k'}\rangle_{\mathcal{M},\mathcal{I}}\right] = 0 \textit{ for } 1 \le k' < k. \end{aligned} \tag{9}$$

The maximum is taken over all $\mathbf{\Phi} = \{\psi_i(s), i \in \mathcal{M}; \phi_{ij}(u,v), (i,j) \in \mathcal{I}\}$*, where* $\psi_i(s)$*'s and* $\phi_{ij}(u,v)$*'s are arbitrary functions satisfying the constraints in (9). Moreover, the maximum value of (9) is equal to the k-th eigenvalue* σ_k^2 *of the covariance function of* $S(t)$*.*

To estimate $w_k(t)$'s, we consider a transformation of model (2),

$$\begin{aligned} Y(t) &= \mu_0(t) + S(t) + \varepsilon(t) = \mu_0(t) + \sum_{k=1}^{\infty} \mathbf{z}_k w_k(t) + \varepsilon(t) \\ &= \mu_0(t) + \sum_{k=1}^{\infty} \langle \mathbf{X}, \mathbf{\Phi}_k \rangle_{\mathcal{M},\mathcal{I}} w_k(t) + \varepsilon(t), \end{aligned} \tag{10}$$

where the second and the third equalities respectively follow from the KL expansion (3) and the equation (4). With the estimates $\widehat{\mathbf{\Phi}}_k$'s of $\mathbf{\Phi}_k$'s, $\langle \mathbf{X}, \mathbf{\Phi}_k \rangle_{\mathcal{M},\mathcal{I}}$ can be estimated by $\langle \mathbf{X}, \widehat{\mathbf{\Phi}}_k \rangle_{\mathcal{M},\mathcal{I}}$. Then we estimate $w_k(t)$'s by regressing $Y(t)$ on $\langle \mathbf{X}, \widehat{\mathbf{\Phi}}_k \rangle_{\mathcal{M},\mathcal{I}}$'s. We provide the details in the following section.

2.2 Estimation procedure

Given a set of independent samples $\{Y_l(t), \mathbf{X}_l = (X_{l,1}(s), \ldots, X_{l,p}(s))^\top : 1 \le l \le n\}$ of $\{Y(t), \mathbf{X}\}$, let $\overline{Y}(t) = \sum_{l=1}^n Y_l(t)/n$ denote the mean sample response. For any $\mathbf{\Phi}$, define $T_l(\mathbf{\Phi}) = \langle \mathbf{X}_l, \mathbf{\Phi} \rangle_{\mathcal{M},\mathcal{I}} - \overline{\langle \mathbf{X}_\bullet, \mathbf{\Phi} \rangle_{\mathcal{M},\mathcal{I}}}$, where $\overline{\langle \mathbf{X}_\bullet, \mathbf{\Phi} \rangle_{\mathcal{M},\mathcal{I}}} = \sum_{l=1}^n \langle \mathbf{X}_l, \mathbf{\Phi} \rangle_{\mathcal{M},\mathcal{I}}/n$. By the law of large numbers and noting that $\mathbf{X}$ has mean zero, we have that $\sum_{l=1}^n T_l(\mathbf{\Phi})\{Y_l(t) - \overline{Y}(t)\}/n$ converges to $\mathrm{E}[\langle \mathbf{X}, \mathbf{\Phi} \rangle_{\mathcal{M},\mathcal{I}} \{S(t) + \varepsilon(t)\}]$ which is equal to $\mathrm{E}[\langle \mathbf{X}, \mathbf{\Phi} \rangle_{\mathcal{M},\mathcal{I}} S(t)]$ as $\mathbf{X}$ and $\varepsilon(t)$ are independent. Moreover, $\sum_{l=1}^n T_l(\mathbf{\Phi})^2/n$, the sample variance of $\{\langle \mathbf{X}_l, \mathbf{\Phi} \rangle_{\mathcal{M},\mathcal{I}}, 1 \le l \le n\}$, converges to $\mathrm{E}\left[\langle \mathbf{X}, \mathbf{\Phi} \rangle_{\mathcal{M},\mathcal{I}}^2\right]$. These results, together with the fact that solving the generalized eigenvalue problem (9) is equivalent to maximizing the Rayleigh quotient $\int_0^1 \mathrm{E}\left[\{\langle \mathbf{X}, \mathbf{\Phi} \rangle_{\mathcal{M},\mathcal{I}} S(t)\}^2\right] dt / \mathrm{E}\left[\langle \mathbf{X}, \mathbf{\Phi} \rangle_{\mathcal{M},\mathcal{I}}^2\right]$ with the same constraints, motivate us to propose the following penalized optimization problem. Our estimates $\widehat{\mathbf{\Phi}}_k$ of $\mathbf{\Phi}_k$, $k \ge 1$, are the solutions to

$$\begin{aligned} &\max_{\mathbf{\Phi}} \quad \frac{\int_0^1 \left[\frac{1}{n}\sum_{l=1}^n T_l(\mathbf{\Phi})\{Y_l(t) - \overline{Y}(t)\}\right]^2 dt}{\frac{1}{n}\sum_{l=1}^n T_l(\mathbf{\Phi})^2 + P(\mathbf{\Phi})}, \\ &\text{subject to} \quad \frac{1}{n}\sum_{l=1}^n T_l(\mathbf{\Phi})^2 = 1, \quad \text{and} \quad \frac{1}{n}\sum_{l=1}^n T_l(\mathbf{\Phi}) T_l(\widehat{\mathbf{\Phi}}_{k'}) = 0, \quad 1 \le k' < k, \end{aligned} \tag{11}$$

where the maximum is taken over all possible $\mathbf{\Phi} = \{\psi_i(s), i \in \mathcal{M}; \phi_{ij}(u,v), (i,j) \in \mathcal{I}\}$ satisfying the constraints in (11), and $P(\mathbf{\Phi})$ denotes penalties. Extending the penalty in

Luo and Qi (2017) for the function-on-function model with only main effects, we introduce the following (*nonadaptive*) penalty

$$P(\mathbf{\Phi}) = \lambda \sum_{i\in\mathcal{M}} \left\{ \|\psi_i\|_{L^2}^2 + \tau \|\psi_i''\|_{L^2}^2 \right\} + \lambda \sum_{(i,j)\in\mathcal{I}} \left\{ \|\phi_{ij}\|_{L^2}^2 + \tau \left(\|\partial_{uu}\phi_{ij}\|_{L^2}^2 + \|\partial_{uv}\phi_{ij}\|_{L^2}^2 + \|\partial_{vv}\phi_{ij}\|_{L^2}^2 \right) \right\}, \tag{12}$$

where $\lambda > 0$ and $\tau > 0$ are tuning parameters, $\|\cdot\|_{L^2}$ is the usual L^2 norm of a function, and $\partial_{uu}\phi_{ij}$, $\partial_{uv}\phi_{ij}$ and $\partial_{vv}\phi_{ij}$ denote the second order partial derivatives of ϕ_{ij}. The penalty $P(\mathbf{\Phi})$ consists of both the L^2 norms and the roughness of the functions $\psi_i(s)$'s and $\phi_{ij}(u,v)$'s. As $P(\mathbf{\Phi})$ is in the denominator of the objective function in (11), functions with larger norms or being less smooth will lead to smaller values of the objective function. So the penalty $P(\mathbf{\Phi})$ can control the magnitude of the norms and roughness of the estimated functions. In Luo and Qi (2017), we have shown that the control of the L^2 norms of $\psi_i(s)$'s is essential for the convergence of the prediction error to the smallest one, the expected L^2 norm of the noise $\varepsilon(t)$, as $n \to \infty$. For the same reason, we control the norms of both $\psi_i(s)$'s and $\phi_{ij}(u,v)$'s. This also guarantees the existence and uniqueness of the solution in practice (see Section S.2 in supplementary material for details). The smoothness penalty helps improving the estimation and prediction accuracy. With the tuning parameter τ, we allow different magnitudes of penalties on the L^2 norms and the smoothness, which increases the flexibility of regularization.

In practice, coefficient functions of different effects may have quite different L^2 norms and smoothness levels, so do the component functions $\psi_{i,k}(s)$'s and $\phi_{ij,k}(u,v)$'s for different i, (i,j) and k. In addition, given (i,j) and k, the smoothness of $\phi_{ij,k}(u,v)$'s along different axises (u or v) can be different. So using the same tuning parameters for $\psi_{i,k}(s)$'s and $\phi_{ij,k}(u,v)$'s and their derivatives or partial derivatives may not be efficient to control the magnitudes and smoothness of the estimated coefficient functions. Motivated by the adaptive Lasso (Zou, 2006), we propose the following *adaptive* penalty

$$P(\Phi) = \lambda \sum_{i\in\mathcal{M}} \left\{ \omega_{i,k}^{(0)} \|\psi_i\|_{L^2}^2 + \tau \omega_{i,k}^{(2)} \|\psi_i''\|_{L^2}^2 \right\} + \lambda \sum_{(i,j)\in\mathcal{I}} \left\{ \omega_{ij,k}^{(00)} \|\phi_{ij}\|_{L^2}^2 + \tau \left(\omega_{ij,k}^{(20)} \|\partial_{uu}\phi_{ij}\|_{L^2}^2 + \omega_{ij,k}^{(11)} \|\partial_{uv}\phi_{ij}\|_{L^2}^2 + \omega_{ij,k}^{(02)} \|\partial_{vv}\phi_{ij}\|_{L^2}^2 \right) \right\}. \tag{13}$$

For any $i \in \mathcal{M}$, $\{\omega_{i,k}^{(0)}, \omega_{i,k}^{(2)}\}$ are the adaptive weights for the magnitude and smoothness

of $\psi_{i,k}(s)$'s, respectively. For any $(i,j) \in \mathcal{I}$, $\{\omega_{ij,k}^{(00)}, \omega_{ij,k}^{(20)}, \omega_{ij,k}^{(11)}, \omega_{ij,k}^{(02)}\}$, are the adaptive weights for $\phi_{ij,k}(u,v)$'s and its three second order partial derivatives. The general idea of the adaptive penalties is that functions with smaller magnitudes or being smoother will be penalized more by using larger adaptive weights. The choice of these adaptive weights is provided in Section 4.1. The nonadaptive penalty (12) is a special case of (13) with all weights equal to one, and will be used in a preliminary fitting of model (2) to calculate the adaptive weights. As an example, we compare the prediction accuracy of the fitted models using adaptive penalty to those with nonadaptive penalty for the Hawaii ocean data in Section 6.1. The results suggest that for a model with a large number of effects, such as the full model with all main, quadratic and interaction effects, using adaptive penalty can greatly improve the prediction accuracy.

Using basis expansions for $\psi_i(s)$'s and $\phi_{ij}(u,v)$'s, we express (11) as an optimization problem of the expansion coefficients. Let $\mathbf{b}(s) = (b_1(s), \ldots, b_L(s))^\top$ denote a set of L basis functions in $[0,1]$, and $\mathbf{h}(u) = (h_1(u), \ldots, h_J(u))^\top$ denote another set of J basis functions in $[0,1]$. Then the vector $\mathbf{\Psi}(u,v) = \mathbf{h}(u) \otimes \mathbf{h}(v) = (h_1(u)h_1(v), h_1(u)h_2(v), \ldots, h_J(u)h_J(v))^\top$ is a set of J^2 tensor product basis functions in the two-dimensional region $[0,1] \times [0,1]$. Let $\psi_i(s) = \mathbf{c}_i^\top \mathbf{b}(s)$ and $\phi_{ij}(u,v) = \mathbf{d}_{ij}^\top \mathbf{\Psi}(u,v)$ be the basis expansion representations of the functions in $\mathbf{\Phi}$, where $\mathbf{c}_i$ and $\mathbf{d}_{ij}$ are the L-dimensional and J^2-dimensional coefficient vectors, respectively. Let $\mathbf{v} = (\mathbf{c}_i^\top, i \in \mathcal{M}; \mathbf{d}_{ij}^\top, (i,j) \in \mathcal{I})^\top$ be the $(L|\mathcal{M}| + J^2|\mathcal{I}|)$-dimensional concatenated vector of all coefficients, where $|\mathcal{M}|$ and $|\mathcal{I}|$ denote the cardinality of $\mathcal{M}$ and $\mathcal{I}$, respectively. The optimization problem (11) can be expressed as the following multivariate generalized eigenvalue problem of the expansion coefficients, with the solution denoted as $\widehat{\mathbf{v}}_k = (\widehat{\mathbf{c}}_{i,k}^\top, i \in \mathcal{M}; \widehat{\mathbf{d}}_{ij,k}^\top, (i,j) \in \mathcal{I})^\top$,

$$\max_{\mathbf{v}} \frac{\mathbf{v}^\top \Xi \mathbf{v}}{\mathbf{v}^\top \mathbf{Q} \mathbf{v}}, \quad \text{subject to} \quad \mathbf{v}^\top \mathbf{H} \mathbf{v} = 1, \quad \widehat{\mathbf{v}}_{k'}^\top \mathbf{H} \mathbf{v} = 0, \quad 1 \le k' \le k-1, \tag{14}$$

where the maximum is taken over all $(L|\mathcal{M}| + J^2|\mathcal{I}|)$-dimensional vector $\mathbf{v}$. The Ξ, $\mathbf{Q}$ and $\mathbf{H}$ in (14) are nonnegative definite symmetric matrices and their explicit forms are provided in Section S.1.1 of the supplementary material. We can solve (14) using the same method as in Luo and Qi (2017). Then $\mathbf{\Phi}_k$ is estimated by $\widehat{\mathbf{\Phi}}_k = \{\widehat{\psi}_{i,k}(s), i \in \mathcal{M}; \widehat{\phi}_{ij,k}(u,v), (i,j) \in \mathcal{I}\}$, where $\widehat{\psi}_{i,k}(s) = \widehat{\mathbf{c}}_{i,k}^\top \mathbf{b}(s)$ and $\widehat{\phi}_{ij,k}(u,v) = \widehat{\mathbf{d}}_{ij,k}^\top \mathbf{\Psi}(u,v)$.

Suppose that we have obtained $\widehat{\mathbf{\Phi}}_1, \ldots, \widehat{\mathbf{\Phi}}_K$, where K is the number of selected com-

ponents K and the details of selecting K are provided in Section 4.2. Then we estimate $\mu_0(t), w_1(t), \ldots, w_K(t)$ from the transformed model (10). The value of the predictor $\mathbf{z}_k = \langle \mathbf{X}, \mathbf{\Phi}_k \rangle_{\mathcal{M},\mathcal{I}}$ in (10) for the l-th observation is $z_{lk} = \langle \mathbf{X}_l, \mathbf{\Phi}_k \rangle_{\mathcal{M},\mathcal{I}}$ which can be estimated by $\widehat{z}_{lk} = T_l(\widehat{\mathbf{\Phi}}_k) = \langle \mathbf{X}_l, \widehat{\mathbf{\Phi}}_k \rangle_{\mathcal{M},\mathcal{I}} - \overline{\langle \mathbf{X}_\bullet, \widehat{\mathbf{\Phi}}_k \rangle_{\mathcal{M},\mathcal{I}}}$. Let $\widehat{Z}_k = (\widehat{z}_{1k}, \ldots, \widehat{z}_{nk})^\top$ for $1 \le k \le K$. Then we estimate $\mu_0(t), w_1(t), \ldots, w_K(t)$ by regressing $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))^\top$ on $\widehat{Z}_1, \ldots, \widehat{Z}_K$ using the penalized least square method. Our estimates $\{\widehat{\mu}_0(t), \widehat{w}_k(t), 1 \le k \le K\}$ of $\{\mu_0(t), w_k(t), 1 \le k \le K\}$ are the solution to

$$\min_{\nu_k(t), 0 \le k \le K} \int_0^1 \frac{1}{n} \sum_{l=1}^n \left\{ Y_l(t) - \nu_0(t) - \sum_{k=1}^K \widehat{z}_{lk} \nu_k(t) \right\}^2 dt + \eta \left\{ \omega_0^{(t)} \|\nu_0''\|_{L^2}^2 + \sum_{k=1}^K \omega_k^{(t)} \|\nu_k''\|_{L^2}^2 \right\}, \quad (15)$$

where the minimum is taken over all functions $\nu_k(t), 0 \le k \le K$, with square integrable second derivatives, and $\eta > 0$ is the tuning parameter of smoothness penalty. Here we use adaptive weights $\omega_k^{(t)}$ to account for different smoothness levels of $\widehat{w}_k(t)$'s. The form and estimates of these adaptive weights are provided in Section 4.1. To solve (15), we use basis functions $\mathbf{d}(t) = (d_1(t), \ldots, d_M(t))^\top$ to expand $\nu_k(t)$'s. In Section S.1.2 of the supplementary material, we show that the solutions to (15) have the following explicit forms:

$$\widehat{\mu}_0(t) = \mathbf{d}(t)^\top \left[\int_0^1 \mathbf{d}(t)\mathbf{d}^\top(t)dt + \eta\omega_0^{(t)} \int_0^1 \{\mathbf{d}''(t)\}\{\mathbf{d}''(t)\}^\top dt \right]^{-1} \int_0^1 \overline{Y}(t)\mathbf{d}(t)dt, \quad (16)$$

$$\widehat{w}_k(t) = \mathbf{d}(t)^\top \left[\int_0^1 \mathbf{d}(t)\mathbf{d}^\top(t)dt + \eta\omega_k^{(t)} \int_0^1 \{\mathbf{d}''(t)\}\{\mathbf{d}''(t)\}^\top dt \right]^{-1} \int_0^1 \frac{1}{n} \sum_{l=1}^n Y_l(t)\widehat{z}_{lk}\mathbf{d}(t)dt. \quad (17)$$

Then the coefficient functions $\beta_i^{(\mathrm{KL})}(s,t)$ and $\gamma_{ij}^{(K)}(u,v,t)$ given in (6) are estimated by

$$\widehat{\beta}_i^{(\mathrm{KL})}(s,t) = \sum_{k=1}^K \widehat{\psi}_{i,k}(s)\widehat{w}_k(t), \quad \widehat{\gamma}_{ij}^{(K)}(u,v,t) = \sum_{k=1}^K \widehat{\phi}_{ij,k}(u,v)\widehat{w}_k(t). \quad (18)$$

Given a set of new predictor curves $\mathbf{X}_{\mathrm{new}} = (X_{\mathrm{new},1}(s), \ldots, X_{\mathrm{new},p}(s))^\top$, we predict the response by $Y_{\mathrm{pred}}(t) = \widehat{\mu}_0(t) + \widehat{S}_{\mathrm{new}}(t)$, where

$$\begin{aligned} \widehat{S}_{\mathrm{new}}(t) &= \sum_{k=1}^K \left\{ \langle \mathbf{X}_{\mathrm{new}}, \widehat{\mathbf{\Phi}}_k \rangle_{\mathcal{M},\mathcal{I}} - \overline{\langle \mathbf{X}_\bullet, \widehat{\mathbf{\Phi}}_k \rangle_{\mathcal{M},\mathcal{I}}} \right\} \widehat{w}_k(t) \quad (19) \\ &= \sum_{i \in \mathcal{M}} \int_0^1 \left\{ X_{\mathrm{new},i}(s) - \overline{X}_i(s) \right\} \widehat{\beta}_i^{(\mathrm{KL})}(s,t)ds + \\ &\quad \sum_{(i,j) \in \mathcal{I}} \int_0^1 \int_0^1 \left\{ X_{\mathrm{new},i}(u)X_{\mathrm{new},j}(v) - \overline{X_iX_j}(u,v) \right\} \widehat{\gamma}_{ij}^{(\mathrm{KL})}(u,v,t)dudv, \end{aligned}$$

is the estimate of the regression function, $\overline{X}_i(s) = \sum_{l=1}^{n} X_{l,i}(s)/n$, and $\overline{X_i X_j}(u,v) = \sum_{l=1}^{n} X_{l,i}(u)X_{l,j}(v)/n$. The details of choosing the tuning parameters, λ and τ in (11), and η in (15), are provided in Section 4.2.

3 Model selection

In practice, the form of model (2) is unspecified, and a model selection procedure is necessary. In multiple linear regression with scalar response and predictors, various criteria have been proposed for model selection, such as the AIC, BIC, C_p, and PRESS. Some of these criteria are based on the goodness of fit and model simplicity, and some are based on prediction. One wants to find the best model under certain criteria. However, with the interaction effects considered, the number of all possible models is typically large and it is computationally expensive to explore all possible models. Therefore, various selection procedures, such as backward elimination, forward selection, and stepwise selection, have been proposed. Hastie *et al.* (2017) conducted extensive comparisons by simulations and showed that the best subset selection and (forward) stepwise perform quite similarly in linear regression with scalar response and predictors, and generally perform better than the penalized method Lasso in high signal-to-noise ratio regimes (with Lasso better in low signal-to-noise ratio regimes). In this paper, we propose a stepwise procedure for function-on-function interaction model based on prediction accuracy.

Let $(\mathcal{M}^{(0)}, \mathcal{I}^{(0)})$ denote the index set for the main effects and interaction effects in the initial model, and $(\mathcal{M}^{(m)}, \mathcal{I}^{(m)})$ denote the index set for the model selected after the m-th iteration, $m \geq 1$. We choose the main effect model ($\mathcal{M}^{(0)} = \{1, \ldots, p\}$ and $\mathcal{I}^{(0)}$ is empty) as our initial model because usually the main effects are more important than the interaction and quadratic effects, and moreover, a quadratic or interaction term needs much more basis functions than a main effect term. This choice of initial model includes potentially important terms and meanwhile makes our procedure computational efficient.

We first randomly split the n observations into five cross-validation (CV) subsets with roughly the same size and keep these CV subsets unchanged during the stepwise selection procedure. Then we iteratively select a variable to enter or remove from the model in a stepwise manner based on prediction accuracy. Specifically, in the m-th iteration ($m \geq 1$),

calling the model obtained after the $(m-1)$-th step as the "*current model*" with index set $(\mathcal{M}^{(m-1)}, \mathcal{I}^{(m-1)})$, we take the following steps.

Step 1. Using the five CV subsets, we apply the CV procedure described in Section 4.2 to the current model to choose the optimal number of components and tuning parameters. Let $V^{(m-1)}$ denote the smallest validation error and correspondingly, the optimal number of components and tuning parameters are $\{K^{(m-1)}, \lambda^{(m-1)}, \tau^{(m-1)}, \eta^{(m-1)}\}$.

Step 2. We pick one main effect which leads to the greatest improvement of the predictive performance of the current model. Specifically, for the i-th main effect, $1 \le i \le p$, let

$$\mathcal{M}_{(i)}^{(m-1)} = \begin{cases} \mathcal{M}^{(m-1)} \bigcup \{i\} & \text{if } i \notin \mathcal{M}^{(m-1)} \\ \mathcal{M}^{(m-1)} \backslash \{i\} & \text{if } i \in \mathcal{M}^{(m-1)} \end{cases}. \tag{20}$$

be the index set of main effects by adding i to $\mathcal{M}^{(m-1)}$ if $X_i(s)$ is not in the current model, or removing it from $\mathcal{M}^{(m-1)}$ if $X_i(s)$ is already in the current model. Let $\mathcal{F}_1$ denote the collection of $1 \le i \le p$ such that $(\mathcal{M}_{(i)}^{(m-1)}, \mathcal{I}^{(m-1)})$ satisfies the hierarchical principle. We will compare the predictive performance of all the models with index sets in $\{(\mathcal{M}_{(i)}^{(m-1)}, \mathcal{I}^{(m-1)}) : i \in \mathcal{F}_1\}$.

To evaluate the predictive performance of these models, we conduct a partial CV using the chosen $\{K^{(m-1)}, \lambda^{(m-1)}, \tau^{(m-1)}, \eta^{(m-1)}\}$ rather than a complete CV procedure to each model for the following two reasons. First, the computational cost will be much higher if we consider all the choices of the number of components and tuning parameters for each model. Second, because each of these models differs from the current model only by one term, it is very possible that the optimal choice of the number of components and tuning parameters for each of these models is the same as or close to the optimal choice $\{K^{(m-1)}, \lambda^{(m-1)}, \tau^{(m-1)}, \eta^{(m-1)}\}$ for the current model. For each $i \in \mathcal{F}_1$, let $V_{(i)}^{(m-1)}$ be the validation error obtained in the partial CV procedure above for the model indexed by $(\mathcal{M}_{(i)}^{(m-1)}, \mathcal{I}^{(m-1)})$. Let $V_*^{(m)} = \min_{i \in \mathcal{F}_1}\{V_{(i)}^{(m)}\}$ and suppose that the minimum is taken at the index $i_0 \in \mathcal{F}_1$.

Step 3. We pick one quadratic or interaction effect. For the (i,j)-th interaction effect, as in Step 2, we define

$$\mathcal{I}_{(i,j)}^{(m-1)} = \begin{cases} \mathcal{I}^{(m-1)} \bigcup \{(i,j)\} & \text{if } (i,j) \notin \mathcal{I}^{(m-1)} \\ \mathcal{I}^{(m-1)} \backslash \{(i,j)\} & \text{if } (i,j) \in \mathcal{I}^{(m-1)} \end{cases}, \qquad 1 \le i \le j \le p. \tag{21}$$

Let $\mathcal{F}_2$ denote the collection of pairs (i,j) such that $(\mathcal{M}^{(m-1)}, \mathcal{I}^{(m-1)}_{(i,j)})$ satisfies the hierarchical principle. We will compare the predictive performance of all models with index sets in $\{(\mathcal{M}^{(m-1)}, \mathcal{I}^{(m-1)}_{(i,j)}) : (i,j) \in \mathcal{F}_2\}$ using a partial CV as in Step 2 with tuning parameters $\{K^{(m-1)}, \lambda^{(m-1)}, \tau^{(m-1)}, \eta^{(m-1)}\}$ and calculate the validation errors $V^{(m)}_{(i,j)}$'s. Let $V^{(m)}_{**} = \min_{(i,j)\in\mathcal{F}_2}\{V^{(m)}_{(i,j)}\}$ and suppose that the minimum is taken at $(i_1, j_1) \in \mathcal{F}_2$.

Step 4. We determine the selected model in this iteration. Let

$$(\mathcal{M}^{(m)}, \mathcal{I}^{(m)}) = \begin{cases} (\mathcal{M}^{(m-1)}_{i_0}, \mathcal{I}^{(m-1)}), & \text{if } V^{(m)}_{*} \le V^{(m)}_{**}, \\ (\mathcal{M}^{(m-1)}, \mathcal{I}^{(m-1)}_{(i_1,j_1)}), & \text{if } V^{(m)}_{*} > V^{(m)}_{**} \end{cases} \tag{22}$$

be the index set of the selected model in this iteration. Therefore, we update the current model with the i_0-th main effect added (or removed) if it leads to a greater improvement in prediction than adding (or removing) the (i_1, j_1)-th interaction effect; otherwise, we add or remove the (i_1, j_1)-th interaction effect.

Stop Criterion. After the m-th iteration, we go back to Step 1 to determine the optimal choice $\{K^{(m)}, \lambda^{(m)}, \tau^{(m)}, \eta^{(m)}\}$ and calculate the validation error $V^{(m)}$. If $V^{(m)} \le V^{(m-1)}$, we proceed to Steps 2$\sim$4 to obtain the model for the $(m+1)$-th iteration. If $V^{(m)} > V^{(m-1)}$, we stop the procedure and choose the model indexed by $(\mathcal{M}^{(m-1)}, \mathcal{I}^{(m-1)})$ as our final model.

In the whole stepwise procedure, we use our nonadaptive estimation procedure, (that is, all adaptive weights in both (13) and (15) are set equal to one), to fit all models. That is because using adaptive penalties for each candidate model in each step leads to over-flexibility and can make the selection procedure unstable. After the final model is determined, we will fit the final model using the adaptive procedure described below in Section 4.1.

4 Computational issues

4.1 Calculation of adaptive weights

Suppose that the main and interaction effects in model (2) are given or have been selected using our stepwise procedure as described in Section 3. In order to fit model (2) using adaptive penalties, we need to calculate two sets of adaptive weights. The first set is

$\{\omega_{i,k}^{(0)}, \omega_{i,k}^{(2)}, \omega_{ij,k}^{(00)}, \omega_{ij,k}^{(20)}, \omega_{ij,k}^{(11)}, \omega_{ij,k}^{(02)} : i \in \mathcal{M}, (i,j) \in \mathcal{I}, k \geq 1\}$ for $\{\psi_{i,k}(s)\}$ and $\{\phi_{ij,k}(u,v)\}$ in (13), and the second set is $\{\omega_k^{(t)} : k \geq 1\}$ for $\mu(t)$ and $\{w_k(t)\}$ in (15). These weights tune the penalties on the magnitudes and smoothness of target functions so that more penalties are imposed on functions with smaller magnitudes or being smoother. Let $\widehat{\psi}_{i,k}^{(\text{non})}(s)$, $\widehat{\phi}_{ij,k}^{(\text{non})}(u,v)$ and $\widehat{w}_k^{(\text{non})}(t)$ respectively be the initial estimates of $\psi_{i,k}^{(\text{non})}(s)$, $\phi_{ij,k}^{(\text{non})}(u,v)$ and $w_k^{(\text{non})}(t)$ obtained from the nonadaptive procedure (including the CV procedure as described below in Section 4.2). A direct extension of the idea in adaptive lasso is to take the adaptive weights as the reciprocals of the squared norms or roughness of these estimated functions. However, the random fluctuations of the estimates $\widehat{\psi}_{i,k}^{(\text{non})}(s)$'s and $\widehat{\phi}_{ij,k}^{(\text{non})}(u,v)$'s can bring large variations to the estimates of these adaptive weights, which consequently lead to large variations of the adaptive estimates of the coefficient functions, and affect the predictive performance. In the following, we introduce some constraints on these adaptive weights to reduce the variations.

By (5), we have $\psi_{i,k}(s) = \int_0^1 \beta_i(s,t) w_k(t) dt / \sigma_k^2$ for any $i \in \mathcal{M}$ and $k \geq 1$ with $\sigma_k \neq 0$. As $w_k(t)$ is set to have the L^2 norm equal to σ_k (Section 2.1), $w_k(t)/\sigma_k$ has unit norm. So given i, $\psi_{i,k}(s)$'s can be viewed as the weighted averages of $\beta_i(s,t)$, with the weight function $w_k(t)/\sigma_k$ up to a scaling factor $1/\sigma_k$. Since $\beta_i(s,t)$ is smooth in both s and t, we assume that given i, for different k, the weighted average, $\int_0^1 \beta_i(s,t) w_k(t) dt/\sigma_k$, does not differ greatly and hence $\|\psi_{i,k}\|_{L^2}^2$ is roughly proportional to $1/\sigma_k^2$. Then, in practice, given i, we choose the adaptive weight $\omega_{i,k}^{(0)}$ for $\psi_{i,k}$ proportional to $\widehat{\sigma}_k^2$, the maximum value of the optimization problem (11) in the nonadaptive procedure, which is an estimate of σ_k^2. Specifically, we choose the weight

$$\omega_{i,k}^{(0)} = \frac{c_0 \widehat{\sigma}_k^2}{\sum_{m=1}^{\widehat{K}_{opt}^{(\text{non})}} \widehat{\sigma}_{k'}^2 \|\widehat{\psi}_{i,k'}^{(\text{non})}\|_{L^2}^2 / \widehat{K}_{opt}^{(\text{non})}}, \tag{23}$$

for any $1 \leq k \leq \widehat{K}_{opt}^{(\text{non})}$ and $i \in \mathcal{M}$, where $\widehat{K}_{opt}^{(\text{non})}$ is the optimal number of components selected in the nonadaptive procedure, and c_0 is a scaling factor to be explained later. The denominator in (23) is the average of $\{\widehat{\sigma}_{k'}^2 \|\widehat{\psi}_{i,k'}^{(\text{non})}\|_{L^2}^2, 1 \leq k' \leq \widehat{K}_{opt}^{(\text{non})}\}$, the estimates of $\{\sigma_{k'}^2 \|\psi_{i,k'}\|_{L^2}^2, 1 \leq k' \leq \widehat{K}_{opt}^{(\text{non})}\}$. As discussed above, $\sigma_{k'}^2 \|\psi_{i,k'}\|_{L^2}^2$, $1 \leq k' \leq \widehat{K}_{opt}^{(\text{non})}$, are roughly the same and can be estimated by the average in the denominator of (23), which can reduce the variations due to estimation. Similarly, we choose the other adaptive weights

for $\psi_{i,k}(s)$'s and $\phi_{ij,k}(u,v)$'s in (13) as follows,

$$\omega_{i,k}^{(2)} = \frac{c_2\widehat{\sigma}_k^2}{\sum_{k'=1}^{\widehat{K}_{opt}^{(\mathrm{non})}} \widehat{\sigma}_{k'}^2 \left\| \widehat{\psi}_{i,k'}^{(\mathrm{non})\prime\prime} \right\|_{L^2}^2}, \quad i \in \mathcal{M},$$

$$\omega_{ij,k}^{(00)} = \frac{c_0\widehat{\sigma}_k^2}{\sum_{k'=1}^{\widehat{K}_{opt}^{(\mathrm{non})}} \widehat{\sigma}_{k'}^2 \|\widehat{\phi}_{ij,k'}^{(\mathrm{non})}\|_{L^2}^2}, \quad \omega_{ij,k}^{(20)} = \frac{c_2\widehat{\sigma}_k^2}{\sum_{k'=1}^{\widehat{K}_{opt}^{(\mathrm{non})}} \widehat{\sigma}_{k'}^2 \|\partial_{uu}\widehat{\phi}_{ij,k'}^{(\mathrm{non})}\|_{L^2}^2},$$

$$\omega_{ij,k}^{(11)} = \frac{c_2\widehat{\sigma}_k^2}{\sum_{k'=1}^{\widehat{K}_{opt}^{(\mathrm{non})}} \widehat{\sigma}_{k'}^2 \|\partial_{uv}\widehat{\phi}_{ij,k'}^{(\mathrm{non})}\|_{L^2}^2}, \quad \omega_{ij,k}^{(02)} = \frac{c_2\widehat{\sigma}_k^2}{\sum_{k'=1}^{\widehat{K}_{opt}^{(\mathrm{non})}} \widehat{\sigma}_{k'}^2 \|\partial_{vv}\widehat{\phi}_{ij,k'}^{(\mathrm{non})}\|_{L^2}^2}, \quad (i,j) \in \mathcal{I},$$

where c_0 and c_2 are scaling constants so that $\omega_{i_1,1}^{(0)} = 1$ and $\omega_{i_1,1}^{(2)} = 1$, and i_1 is the index of the first main effect.

For the second set of adaptive weights, we choose

$$\omega_k^{(t)} = \left\| \widehat{w}_0^{(\mathrm{non})\prime\prime} \right\|_{L^2}^2 \Big/ \left\| \widehat{w}_k^{(\mathrm{non})\prime\prime} \right\|_{L^2}^2 \tag{24}$$

to penalize more on smoother functions and restrict $\omega_0^{(t)} = 1$. With all these adaptive weights, we perform the CV procedure described in Section 4.2 to choose the optimal number of components and tuning parameters, and then estimate the model as described as in Section 2.2.

4.2 Choice of the number of components and tuning parameters

Two tuning parameters λ and τ are involved in the optimization problem (11), and one tuning parameter η is involved in (15). We choose these tuning parameters from a three dimensional grid determined by $\lambda \in \{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2\}$, $\tau \in \{10^{-3}, 10^{-1}, 10, 10^3\}$, and $\eta \in \{10^{-11}, 10^{-9}, 10^{-7}, 10^{-5}, 10^{-3}, 10^{-1}, 10, 10^3\}$ using a CV procedure described below. We will calculate the CV error for each of the 192 combinations of candidate values of the three tuning parameters. We note that the smoothness of $\psi_{i,k}(s)$'s and $\phi(u,v)$'s is tuned by the multiplication $\lambda\tau$, so a small λ and a large τ can have a similar effect on smoothness as a small λ and a large τ. Here we choose a relative small number of candidate value for each of λ and τ, but the number of their combinations is not small, which provides enough choices to tune the smoothness of $\psi_{i,k}(s)$'s and $\phi_{ij,k}(u,v)$'s.

We first determine the maximum number of components we need to calculate for each combination of candidate values of the three tuning parameters. Specifically, for any $1 \leq$

$i \leq 6$, $1 \leq j \leq 4$ and $1 \leq \ell \leq 8$, the maximum number of components for the i-th value of λ, the j-th value of τ, the ℓ-th value of η, is given by

$$\widehat{K}_{max}^{(i,j,\ell)} = \min\left\{k > 1 : \frac{\widehat{\sigma}_k^2}{\widehat{\sigma}_1^2 + \cdots + \widehat{\sigma}_k^2} \leq 0.01\right\}, \tag{25}$$

where $\widehat{\sigma}_k^2$ is the maximum value of the optimization problem (11) using either the penalty (12) in the nonadaptive procedure or the penalty (13) in the adaptive procedure, with λ taking its i-th value, τ taking its j-th value and η taking its ℓ-th value. The $\widehat{\sigma}_k^2$ is the estimate of σ_k^2, which is both the k-th eigenvalue of the covariance function of the signal function $S(t)$ and the maximum value of the generalized eigenvalue problem (9). σ_k^2 also measures the magnitude of the signal in the k-th component. Therefore, we stop solving the sequential problems (11) when $\widehat{\sigma}_k^2$ does not exceed 1% of the cumulative sum. These maximum numbers $\widehat{K}_{max}^{(i,j,\ell)}$'s will serve as upper bounds for choosing the optimal number of components.

Next, we use a CV method to determine the optimal tuning parameters and the optimal number of components, $\widehat{K}_{opt}$, simultaneously. We provide the details in the following algorithm.

Algorithm 4.1 *1. We determine $\widehat{K}_{max}^{(i,j,\ell)}$'s using the whole data set and* (25) *for all 192 combinations of the candidate values of the three tuning parameters.*

2. We use the five-fold CV to determine the optimal number of components and the tuning parameters simultaneously. Specifically, we randomly split the whole data set into five subsets. For each $1 \leq v \leq 5$, we use the v-th subset as the v-th validation set and all other observations as the v-th training set. For the i-th value of λ, the j-th value of τ, the ℓ-th value of η and the v-th training set,

(a) we estimate $\widehat{\mathbf{\Phi}}_k^{(v,i,j,\ell)}$ for $1 \leq k \leq \widehat{K}_{max}^{(i,j,\ell)}$ based on (11)*, and then calculate $\widehat{w}_k^{(v,i,j,\ell)}(t)$ and $\widehat{\mu}_0^{(v,i,j,\ell)}(t)$ based on* (15)*.*

(b) For each $1 \leq K \leq \widehat{K}_{max}^{(i,j,\ell)}$, we use $\widehat{\mu}_0^{(v,i,j,\ell)}(t)$, $\widehat{\mathbf{\Phi}}_k^{(v,i,j,\ell)}$ and $\widehat{w}_k^{(v,i,j,\ell)}(t)$, $1 \leq k \leq K$, to calculate the predicted responses $\{\widehat{y}_m^{(v,i,j,,\ell,K)}(t), 1 \leq m \leq n_v\}$ for observations in the v-th validation data set based on formula (19)*. Then we obtain the validation error $e_{v,i,j,,\ell,K} = \sum_{m=1}^{n_v} \int_0^1 \{\widehat{y}_m^{(v,i,j,,\ell,K)}(t) - y_m^{(v)}(t)\}^2 dt$, where $\{y_m^{(v)}(t), 1 \leq m \leq n_v\}$ are the observed responses in the v-th validation set.*

After we repeat (a)-(b) for all $1 \le v \le 5$, $1 \le i \le 6$, $1 \le j \le 4$ *and* $1 \le \ell \le 8$, *we calculate the average validation error,* $\bar{e}_{i,j,\ell,K} = \sum_{v=1}^{5} e_{v,i,j,\ell,K}/5$.

3. *Let* $\bar{e}_{i_0,j_0,\ell_0,K_0} = \min\{\bar{e}_{i,j,\ell,K} : 1 \le i \le 6, 1 \le j \le 4, 1 \le \ell \le 8, 1 \le K \le \widehat{K}_{max}^{(i,j,\ell)}\}$. *Then we choose the* i_0*-th value of* λ, *the* j_0*-th value of* τ, *the* ℓ_0*-th value of* η *as optimal tuning parameters, and* K_0 *as the optimal number of components* $\widehat{K}_{opt}$.

5 Simulations

In this section and the following section of real data analysis, without further emphasis, all the specified and selected models will be fitted using our adaptive estimation method.

We consider two sets of simulations, each of which has $p = 5$ predictor curves simulated from a Gaussian process as follows. We first generate independent curves $V_j(s)$, $j = 1, \ldots, 9$, from the Gaussian process with mean zero and a positive definite covariance function $\mathbf{\Sigma}_V(s, s') = \exp\{-100(s - s')^2\}$. Then we define $X_j(s) = \sum_{i=0}^{Lag} V_{j+i}(s)/\sqrt{Lag + 1}$, $1 \le j \le 5$, where the parameter Lag is a positive integer controlling the correlation between predictor curves. A larger Lag leads to a stronger correlation between predictor curves. We consider two correlation levels with $Lag = 2$ and 4, respectively. We show twenty sample curves for $X_j(s)$, $1 \le j \le 5$, with $Lag = 4$, in Figure 1. In both simulations, we generate the noise $\epsilon(t)$ independently from $N(0, \sigma_\epsilon^2)$. We will consider two noise levels, $\sigma_\epsilon^2 = 0.1$ and 1. In both simulations, we generate response functions using $Y(t) = \tilde{S}(t) + \varepsilon(t)$, where $\tilde{S}(t) = \sum_{i \in \mathcal{M}} \int_0^1 X_i(s)\beta_i(s, t)ds + \sum_{(i,j) \in \mathcal{I}} \int_0^1 \int_0^1 X_i(u)X_j(v)\gamma_{ij}(u, v, t)dudv$ is the uncentered quadratic regression function. The index sets of $\mathcal{M}$ and $\mathcal{I}$ for the true model and the forms of the coefficient functions, $\beta_i(s, t)$'s and $\gamma_{ij}(u, v, t)$'s, for each simulation are given as follows. When $\sigma_\epsilon^2 = 1$, the signal-to-noise ratio ($\int_0^1 \mathrm{Var}(S(t))dt / \int_0^1 \mathrm{Var}(\varepsilon(t))dt$) ranges around $0.3 \sim 0.8$ for different settings.

<u>*Simulation 1*</u>. The index sets are $\mathcal{M} = \{2, 3, 4, 5\}$ and $\mathcal{I} = \{(2, 2), (3, 4), (4, 5)\}$. The corresponding coefficient functions are

$$\beta_2(s, t) = e^{-3(s-1)^2-5(t-0.5)^2}, \quad \beta_3(s, t) = e^{-5(s-0.5)^2-5(t-0.5)^2} + 8e^{-5(s-1.5)^2-5(t-0.5)^2},$$
$$\beta_4(s, t) = \sin(1.5\pi s)\sin(\pi t), \quad \beta_5(s, t) = \sqrt{st},$$

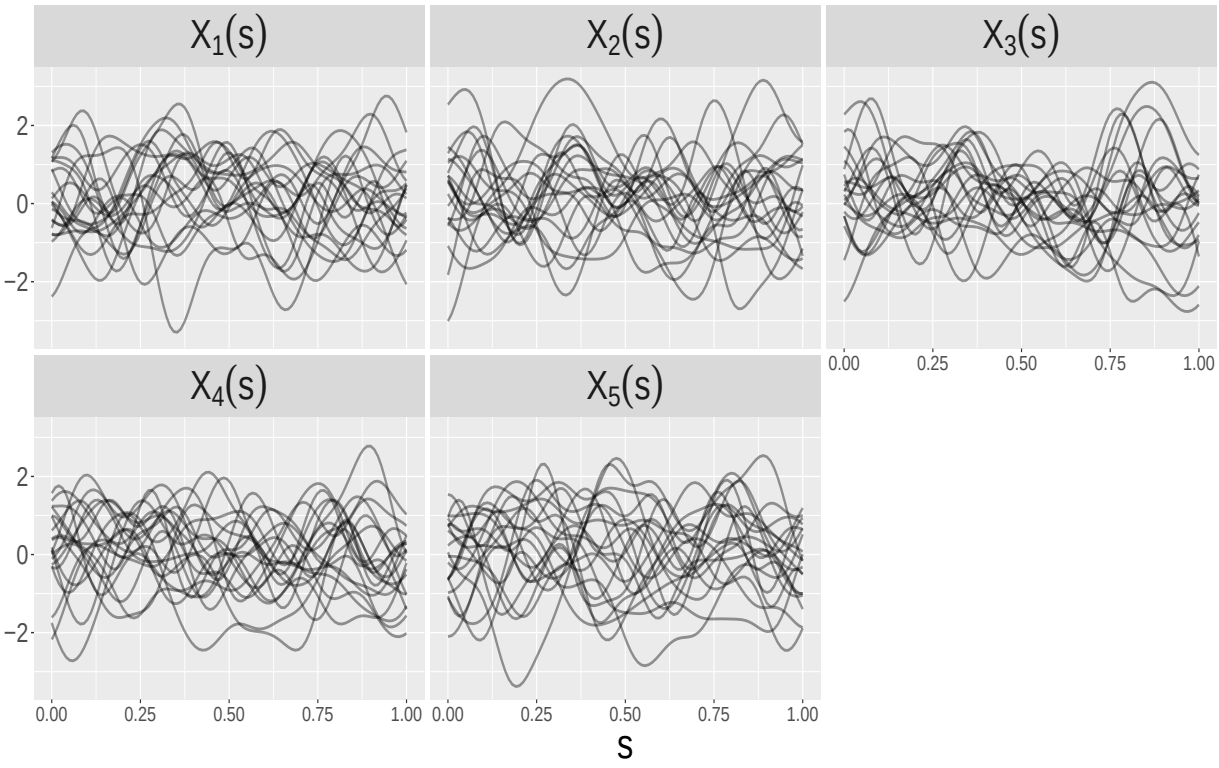

Figure 1: Twenty sample curves for each of the five functional predictors, $X_j(s)$, $1 \leq j \leq 5$, with $Lag = 4$ in our simulation study. All the sample curves are defined in the interval $0 \leq s \leq 1$, and the ranges of the y-axis of all the five plots are the same.

for the main effects, and

$$\gamma_{22}(u,v,t) = 5uv\sqrt{t}, \qquad \gamma_{34}(u,v,t) = 5\cos(\pi u)\sin(2\pi v)\cos(2\pi t),$$
$$\gamma_{45}(u,v,t) = 0.5e^{u+2v-t},$$

for the interaction effects.

Simulation 2. The index sets are $\mathcal{M} = \{1,2,4,5\}$ and $\mathcal{I} = \{(1,1),(1,2),(1,5),(2,4),(4,5),(5,5)\}$ with the corresponding coefficient functions

$$\beta_1(s,t) = (s-2t)^2/3, \qquad \beta_2(s,t) = 2\{\ln(1+s)\}^2\sin\{2\pi(t-0.5)\},$$
$$\beta_4(s,t) = \{\cos(1-s)+\sqrt{t}\}/3, \quad \beta_5(s,t) = (1+s^2)/\{3(1+t^2)\},$$

for the main effects, and

$$\gamma_{11}(u,v,t) = 2(u+v)t^2, \quad \gamma_{12}(u,v,t) = 0.01(u^2 - v^3 + t),$$
$$\gamma_{15}(u,v,t) = 0.01e^{2u-v+3t}, \quad \gamma_{24}(u,v,t) = 0.01(2u - v + 3t),$$
$$\gamma_{45}(u,v,t) = 0.01\tfrac{\ln(1+2u)}{1+t}, \quad \gamma_{55}(u,v,t) = \cos\{\pi(u+v)\} + 3\sqrt{t},$$

for the interaction effects.

All the predictor and response curves are observed at 50 equally spaced points in $[0,1]$. In each case, we repeat the following procedure 100 times. In each repeat, we generate a training data of size 100 and a test data of size 500. Then we apply our stepwise procedure to select variables, and then fit the final model. Applying the fitted selected model to the test data, $\{y_l^{test}(t), 1 \le l \le 500\}$, we get the predicted response curves $\{y_l^{pred}(t), 1 \le l \le 500\}$. To evaluate the performance of prediction and estimation, we calculate the mean squared prediction error MSPE $= \sum_{l=1}^{500} \|y_l^{pred} - y_l^{test}\|_{L^2}^2/500$, and the mean squared estimation error MSEE $= \sum_{l=1}^{500} \|y_l^{pred} - \tilde{S}_l^{test}\|_{L^2}^2/500$ for the regression functions $\tilde{S}_l^{test}(t) = y_l^{test}(t) - \epsilon_l^{test}(t)$, $1 \le l \le 500$, in the test data set. Here $\|\cdot\|_{L^2}^2$ represents the squared L^2 norm which is approximated by the Riemann sum. For comparison, we also fit the *main effect model* with all five main effects but without any interaction effects, the *full model* with all main, quadratic and two-way interaction terms, and the *true model*. In our estimation procedure for all models, to solve (11) in Section 2.2, we use 40 B-spline basis functions for estimating $\psi_i(s)$'s and $20 \times 20 = 400$ tensor product B-spline basis functions for estimating $\phi_{ij}(u,v)$'s. To solve (16) and (17), we use 40 B-spline basis functions. All the B-spline basis functions have equally spaced knots.

Table 1 summarizes the MSPEs and MSEEs from 100 iterations for all settings in the two simulations. Table S.1 in Section S.4.2 of supplementary material provides the selection frequency for each of 20 effects (5 main effects and 15 quadratic and interaction effects) by our stepwise procedure in 100 iterations. The averaged running time of our model selection procedure over 100 repeats is provided in Table S.3 in Section S.4.2 of supplementary material.

In all cases, the *selected model* has much better estimation and prediction performance than the *main effect model* and the *full model*. The averaged MSEEs of the *main effect model* and the *full model* are $1.3 \sim 48$ times larger than those of the *selected model* in

Table 1: The averages (and standard deviations) of MSEEs and MSPEs from 100 replicates in the simulation studies. The main effect model contains all five main effects and no interaction effects. The full model contains all main effects, quadratic and interaction terms.

			Estimation Error (MSEE)			
Simulation	σ_ϵ^2	Lag	Selected model	Main effect model	Full model	True model
1	0.1	2	0.011(0.008)	0.398(0.041)	0.032(0.007)	0.010(0.003)
		4	0.012(0.004)	0.492(0.068)	0.026(0.006)	0.010(0.003)
	1	2	0.068(0.032)	0.411(0.040)	0.104(0.017)	0.058(0.011)
		4	0.061(0.019)	0.490(0.059)	0.082(0.015)	0.051(0.012)
2	0.1	2	0.006(0.001)	0.270(0.030)	0.009(0.003)	0.006(0.002)
		4	0.006(0.001)	0.282(0.041)	0.008(0.002)	0.006(0.001)
	1	2	0.030(0.007)	0.274(0.036)	0.044(0.012)	0.036(0.009)
		4	0.029(0.006)	0.296(0.042)	0.041(0.011)	0.034(0.008)
			Prediction Error (MSPE)			
Simulation	σ_ϵ^2	Lag	Selected model	Main effect model	Full model	True model
1	0.1	2	0.111(0.008)	0.497(0.041)	0.133(0.008)	0.110(0.003)
		4	0.112(0.004)	0.592(0.054	0.126(0.006)	0.110(0.003)
	1	2	1.068(0.032)	1.411(0.043)	1.105(0.019)	1.058(0.013)
		4	1.061(0.020)	1.490(0.059)	1.082(0.016)	1.050(0.013)
2	0.1	2	0.106(0.002)	0.371(0.030)	0.109(0.003)	0.106(0.002)
		4	0.106(0.002)	0.382(0.041)	0.108(0.002)	0.106(0.002)
	1	2	1.029(0.013)	1.274(0.038)	1.043(0.015)	1.035(0.014)
		4	1.028(0.011)	1.295(0.044)	1.040(0.013)	1.033(0.011)

different settings. The estimation and prediction errors of the *selected model* increase as σ_ϵ^2 increases, but the correlation between the predictor curves, controlled by Lag, has little effect on these errors. We also observe that the *selected model* can even perform better than the *true model* in Simulation 2. This is because that in the true model of Simulation 2, the interaction effects, $X_1(u)X_2(v)$, $X_2(u)X_4(v)$, and $X_4(u)X_5(v)$, have coefficient functions with much smaller magnitudes than others such as $X_1(u)X_1(v)$ and $X_5(u)X_5(v)$ (see Table S.2 in supplementary material). So these three interaction terms have little effect on the response function. But to fit the true model, we have to estimate these three interaction effects which dramatically increases the number of basis coefficients to be estimated, and reduces the estimation and prediction accuracy. On the contrary, Table S.1 of supplementary material shows that in most of repeats of all settings in Simulation 2, our selection procedure drops these three interaction terms from the selected model, which reduces the number of parameters and hence improves the estimation and prediction accuracy.

In Simulation 1, none of the true main and interaction effects are too small to be ignorable. When σ_ϵ^2 is small, these true effects are selected in almost all repeats with some false positive effects (Table S.1 of supplementary material). When σ_ϵ^2 is larger, the selection frequency of less strong true interaction effects $X_2(u)X_2(v)$ and $X_3(u)X_4(v)$ decreases, whereas the false effects $X_4(u)X_4(v)$ and $X_5(u)X_5(v)$ are selected more often. These patterns explain the phenomenon that the selected models have similar estimation and prediction performance as the true models when $\sigma_\epsilon^2 = 0.1$, and are slightly higher when $\sigma_\epsilon^2 = 1$.

To study the coefficient estimation accuracy, we calculate the relative mean squared estimation errors

$$\text{RelMSEE} = \frac{\|\widehat{\beta} - \beta\|_{L^2}^2}{\|\beta\|_{L^2}^2} \qquad \text{and} \qquad \text{RelMSEE} = \frac{\|\widehat{\gamma} - \gamma\|_{L^2}^2}{\|\gamma\|_{L^2}^2}$$

for the coefficients of main effects and interaction effects in the *true model*. We provide the magnitudes of true coefficient functions, $\|\beta\|_{L^2}^2$ and $\|\gamma\|_{L^2}^2$, and summarize the averages and standard deviations of RelMSEEs of coefficient functions in the true models over 100 iterations for each setting in Table S.2 in Section S.4.2 of supplementary material. It is obvious that we have smaller estimation error when σ_ϵ is smaller. For simulation 2, we have very large relative mean squared estimation error for the coefficients of $X_1(u)X_2(v)$,

$X_2(u)X_4(v)$, $X_4(u)X_5(v)$. This is due to the very small magnitudes of true coefficient functions, and the mean squared estimation errors ($\|\widehat{\beta} - \beta\|^2_{L^2}$ and $\|\widehat{\gamma} - \gamma\|^2_{L^2}$) for these terms have similar magnitudes as the other effects.

We also compare our approach to the ordinary linear regression model with interaction effects, by viewing the discrete observations of sample curves as usual scalar variables. Details are presented in Section S.4.3 of supplementary material. The results in Table S.5 show that ignoring the functional structure can dramatically worsen the estimation and prediction accuracies.

6 Applications

6.1 Hawaii ocean data

The Hawaii ocean time-series program has been making repeated observations of various hydrographic, chemical and biological properties of the water column at a station north of Oahu, Hawaii since October, 1988. In the CTD data set (http://hahana.soest.hawaii.edu/hot/hot-dogs/cextraction.html) of this program, various variables were measured every two meters between 0 and 200 meters below the sea surface. These variables are viewed as functions of depth. The layer between 0 and 200 meters below the sea surface is called the epipelagic zone (or sunlight zone), where enough light is available for photosynthesis. Therefore, in this zone, the primary production in the ocean occurs, and plants and animals are largely concentrated. The measurements were repeated at different dates and we view observations from different dates as different sample curves. We remove variables with a large proportion of missing values and consider five functional variables: *Salinity* ($Y(t)$), *Potential Density* ($X_1(s)$), *Temperature* ($X_2(s)$), *Oxygen* ($X_3(s)$), and *Chloropigment* ($X_4(s)$), where $0 \leq s, t \leq 200$. After removing observations with missing values for these five functional variables, we obtain 116 observations totally. We take the centered *Salinity* as the response curve, take the centered curves of the other four variables as functional predictors, and study their relationship. Each curve is observed at 101 equally spaced points in $[0, 200]$. We plot 50 (centered) sample curves in Figure 2 and the original uncentered curves in Figure S.1 of supplementary material.

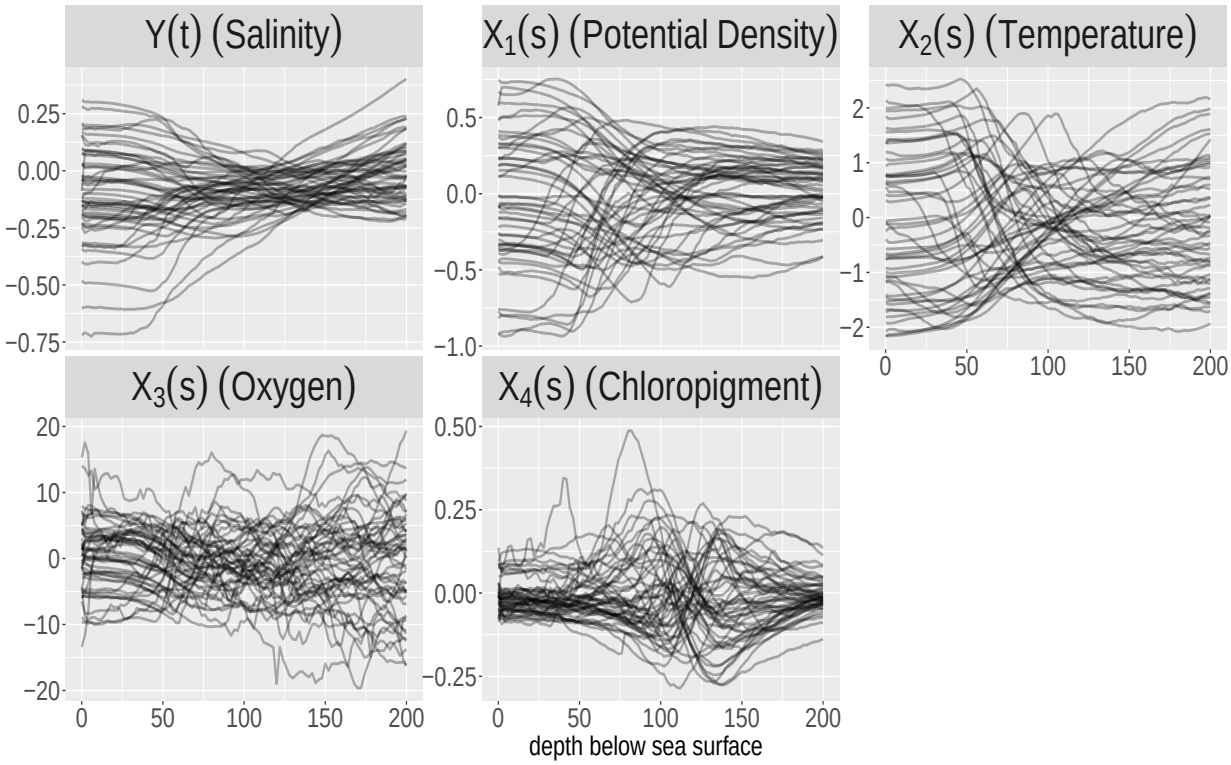

Figure 2: Fifty sample curves of the functional response variable, *Salinity*, and four functional predictor variables, *Potential Density*, *Temperature*, *Oxygen*, and *Chloropigment*, in the Hawaii ocean data set. All curves are functions of depth below the sea surface. The range of depth is $[0, 200]$. All the sample curves have been centered using the mean curves of corresponding variables.

We consider interaction models and perform a model selection using our stepwise procedure. We first compare the *selected model* to the *main effect model* and the *full model*. To evaluate the predictive performance of these models, we repeat the following procedure 100 times. In each repeat, we randomly split the 116 observations into a training set with 90 observations and a test set with 26 observations. Based on the training set, we conduct model selection and obtain the *selected model*. Then we fit the *selected model*, *main effect model* and *full model*, respectively. Finally, we apply the three fitted models to the test data to calculate the MSPEs as in the simulation study. The average (and standard deviation) of the MSPEs over 100 repeats is 2.04×10^{-4} (0.78×10^{-4}) for the *selected model*, 2.67×10^{-4} (0.73×10^{-4}) for the *main effect model*, and 2.89×10^{-4} (1.02×10^{-4}) for the *full model*. The averaged MSPEs for the *main effect model* and the *full model* are respectively 31% and 42% higher than that of the *selected model*. As a comparison, we also use the nonadaptive

estimation procedure to fit three models in each iteration. The average (and standard deviation) of the MSPEs over 100 repeats is 2.06×10^{-4} (0.74×10^{-4}) for the *selected model*, 2.82×10^{-4} (0.74×10^{-4}) for the *main effect model*, and 24.15×10^{-4} (6.01×10^{-4}) for the *full model*. So for a model with a large number of effects, compared to the nonadaptive method, using adaptive method can greatly improve the prediction accuracy.

To measure the goodness of fit and the relative predictive performance with respect to the total variation in the test set, we calculate the following R^2 values and predictive R^2 values

$$R^2 = 1 - \frac{\sum_{l=1}^{90} \|Y_l^{\text{fit}} - Y_l^{\text{train}}\|_{L^2}^2}{\sum_{l=1}^{90} \|Y_l^{\text{train}} - \overline{Y}^{\text{train}}\|_{L^2}^2},$$
$$R^2_{\text{pred}} = 1 - \frac{\sum_{l=1}^{26} \|Y_l^{\text{pred}} - Y_l^{\text{test}}\|_{L^2}^2}{\sum_{l=1}^{26} \|Y_l^{\text{test}} - \overline{Y}^{\text{test}}\|_{L^2}^2},$$

for the selected models, where $\{Y_l^{\text{train}}(t), Y_l^{\text{fit}}(t)\}$, $1 \le l \le 90$, are the sample response curves and the corresponding fitted curves in the training set, and $\{Y_l^{\text{test}}(t), Y_l^{\text{pred}}(t)\}$, $1 \le l \le 26$, are the sample response curves and the corresponding predicted curves in the test set. The averages (and standard deviations) of the R^2 and R^2_{pred} of the selected models over 100 repeats are 99.3% (0.1%) and 99.1% (0.5%), respectively. Both of them are very close to 1.

The low MSPE and high R^2_{pred} imply that the fitted model based on the proposed procedure has good prediction performance. To further quantify the uncertainty of estimation and prediction, we apply the bootstrap method, which has been used to calculate point-wise confidence bands for coefficient functions in functional regression models, such as in Malfait and Ramsay (2003). Here we describe our construction of the point-wise prediction bands in details. We fix a partition of training data and test data. We first apply our stepwise procedure to the training data to choose a model denoted by $(\mathcal{M}_0, \mathcal{I}_0)$, then we fit this model and calculate the predicted response curves for observations in the test data. Then we will construct two types of point-wise prediction bands for the predicted response curves. In the first one, we do not consider the model selection and use the fixed model $(\mathcal{M}_0, \mathcal{I}_0)$, and in the second one, we will consider the model selection. Specifically, we repeat the following procedure 1000 times. In each iteration, from the training data, we draw a bootstrap sample with replacement that has the same size as the training data. We conduct two procedures on the bootstrap sample. (a) We fit the model $(\mathcal{M}_0, \mathcal{I}_0)$ using

the bootstrap sample and calculate the predicted curves for the test data. (b) We apply our stepwise procedure to the bootstrap sample to select a model to fit and calculate the predicted curves for the test data. The procedure (a) captures the uncertainty due to randomness and choice of optimal tuning parameters. The procedure (b) captures additional uncertainty of variable selection. For each of these two procedures and each observation in the test data, we obtain 1000 predicted curves, from which we get the point-wise 95% prediction limits based on the 2.5% and 97.5% percentiles of the predicted values at each observation point. Other bootstrap intervals such as the BC_a method can also be used (Efron and Tibshirani, 1994). We draw the point-wise prediction bands for the first two observations in the test data in Figure 3. As procedure (b) captures more uncertainty, it is expected that the prediction bands from (b) are wider than those from procedure (a). From Figure 3, we observe that the two prediction bands are very close, with the bands (b) slightly wider at certain regions. This is because that the number of different models selected in the 1000 bootstrap iterations is small. Actually the main effects selected in all iterations are X_1 and X_2, and the selected quadratic and interactions effects are (X_1, X_1), or (X_1, X_2), or (X_2, X_2). So the variations due to model selections are small.

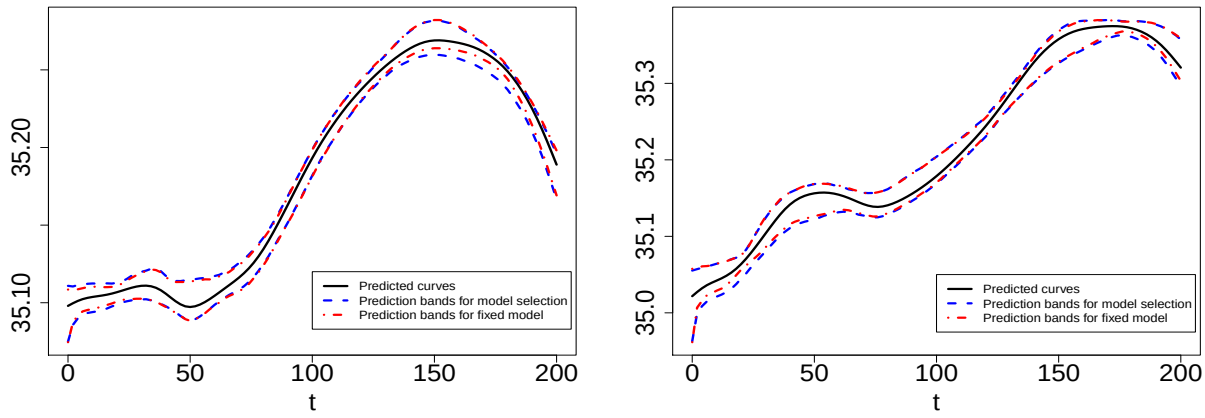

Figure 3: The predicted curves and the 95% bootstrap point-wise prediction bands for the first two observations in the test data for the Hawaii ocean data set. In both plots, the black curves are the predicted curves based on the selected model using the training data and our stepwise procedure; the red curves are the bootstrap point-wise prediction bands without considering model selection; the blue curves are the bootstrap point-wise prediction bands with model selection.

Finally, we use all the 116 observations to conduct a model selection. We list the effects added or removed, and the CV error of the model obtained in each step of our stepwise procedure in Table S.4 of the supplementary material. The final selected model has two main effect terms: Potential Density ($X_1(s)$), Temperature ($X_2(s)$), and one quadratic term $X_2(u)X_2(v)$. The R^2 for the fitted final model is 99.3% which implies that the fitted response curves are very close to the observed response curves. We present the estimated $\widehat{\psi}_{i,k}(s)$'s, $\widehat{w}_k(t)$'s and $\widehat{\phi}_{ij,k}(u,v)$'s for the selected main effects and quadratic effect in the final model in Figures S.2 and S.3 of supplementary material. The estimated coefficient functions $\widehat{\beta}_1^{(\mathrm{KL})}(s,t)$ and $\widehat{\beta}_2^{(\mathrm{KL})}(s,t)$ for the selected main effects are shown in Figure 4. The first heatmap in Figure 4 shows that $\widehat{\beta}_1^{(\mathrm{KL})}(s,t)$ takes large positive values near the diagonal line ($t = s$), and take small or negative values when $|t - s|$ is relatively large. This pattern and the term $\int_0^{200} X_1(s)\widehat{\beta}_1^{(\mathrm{KL})}(s,t)ds$ in the equation (19) imply that given the depth $0 \leq t \leq 200$, the salinity $Y(t)$ is positively associated with the potential density $X_1(s)$ with $|t-s|$ less than 25 meters and the association gradually decays when $|t-s|$ increases. Similarly, the second heatmap in Figure 4 implies that $Y(t)$ is positively associated with the temperature $X_2(s)$ when $|t-s|$ is less than 25 meters. Moreover, these positive associations are strongest near the ocean surface ($t = s = 0$) and about 200 meters depth ($t = s = 200$). It is known that 200 meters below sea surface is the depth separating the epipelagic zone from the mesopelagic zone. The epipelagic zone, extending from the sea surface to 200 meters, is the layer of ocean where most of the visible light exists. The mesopelagic zone, extending from 200 meters to 1,000 meters, is the layer where majority of light comes from bioluminescence, light that is generated by chemical reactions in bacteria, animals, and plants.

Since the quadratic term $X_2(u)X_2(v)$ is included in the final model, the coefficient function of $X_2(s)$ becomes $\widehat{\beta}_2^{(\mathrm{KL})}(s,t)+\int_0^{200} X_2(v)\widehat{\gamma}_{22}^{(\mathrm{KL})}(s,v,t)dv$, where $\int_0^{200} X_2(v)\widehat{\gamma}_{22}^{(\mathrm{KL})}(s,v,t)dv$ can be considered as the effect of $X_2(v)$ on this coefficient function. We illustrate how the change of $X_2(v)$ in different regions in $[0, 200]$ affects this coefficient functions using the "probe function" (Section 5.5.1 of Ramsay and Silverman (2005)), which highlights interesting features over a subregion of an interval. We consider two probe functions: $g_1(s)$ and $g_2(s)$, which are density functions of the normal distributions with means 15 and

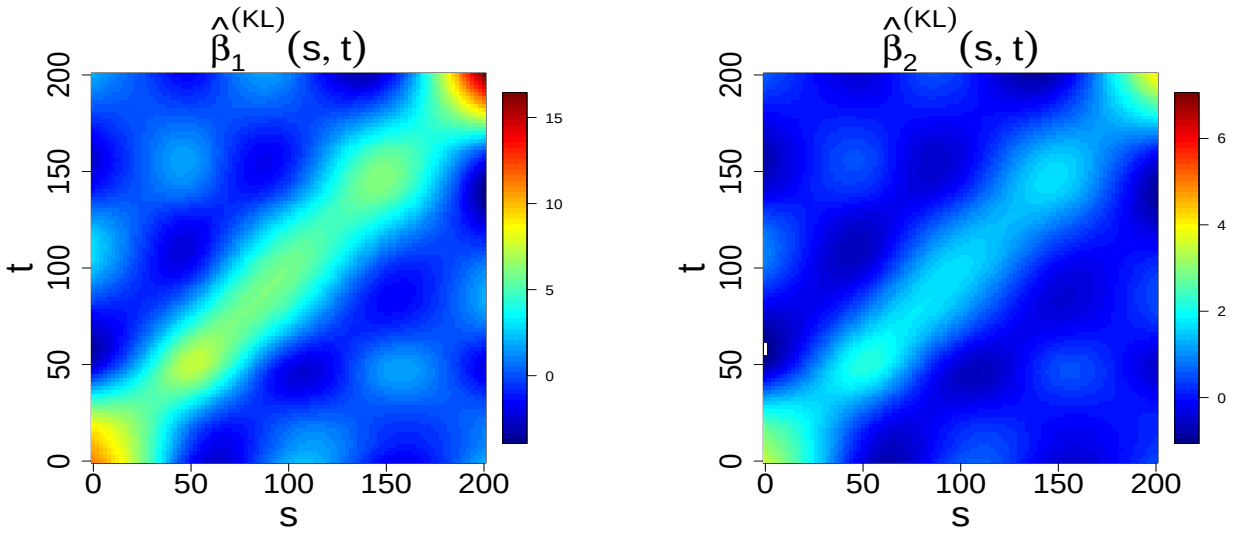

Figure 4: The plots of the estimated coefficient functions $\widehat{\beta}_1^{(\mathrm{KL})}(s,t)$ and $\widehat{\beta}_2^{(\mathrm{KL})}(s,t)$ for the two selected main effects: Potential Density ($X_1(s)$) and Temperature ($X_2(s)$), in the final model for the Hawaii ocean data set.

185, respectively, and the same standard deviation 5. We use $\int_0^{200} g_i(v)\widehat{\gamma}_{22}^{(\mathrm{KL})}(s,v,t)dv$, $i=1,2$, to illustrate the change of $\int_0^{200} X_2(v)\widehat{\gamma}_{22}^{(\mathrm{KL})}(s,v,t)dv$ when $X_2(v)$ increases around $v=15,185$, respectively. We draw the functions $g_i(s)$ and $\int_0^{200} g_i(v)\widehat{\gamma}_{22}^{(\mathrm{KL})}(s,v,t)dv$, $i=1,2$, in Figures S.4 and S.5 in supplementary material, respectively. These plots show that the changes of the temperature ($X_2(v)$) over the two subregions have different effects on $\int_0^{200} X_2(v)\widehat{\gamma}_{22}^{(\mathrm{KL})}(s,v,t)dv$. An increase of temperature around $s=185$ leads to a greater increase of the coefficient function around the corner $(s,t)=(200,200)$ than in other regions, whereas an increase of temperature around $v=15$ leads to a greater increase of the coefficient function near the ocean surface $(s,t)=(0,0)$.

6.2 Air quality data

The air quality data were recorded by an array of five metal oxide chemical sensors embedded in an air quality chemical multi-sensor device that is located in a significantly polluted area, at road level, within an Italian city (De Vito *et al.*, 2008). This data contain 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 percentages) were also recorded hourly in each day. Therefore, we have seven functional variables with sample curves observed at 24 discrete time points. With the removal of missing values, we have 355 observations. We plot all these (uncentered) curves in Figure S.6 of supplementary material. We use the centered curves of NO_2 as the response curve and the centered curves of the other six variables as predictor curves, which are shown in Figure 5.

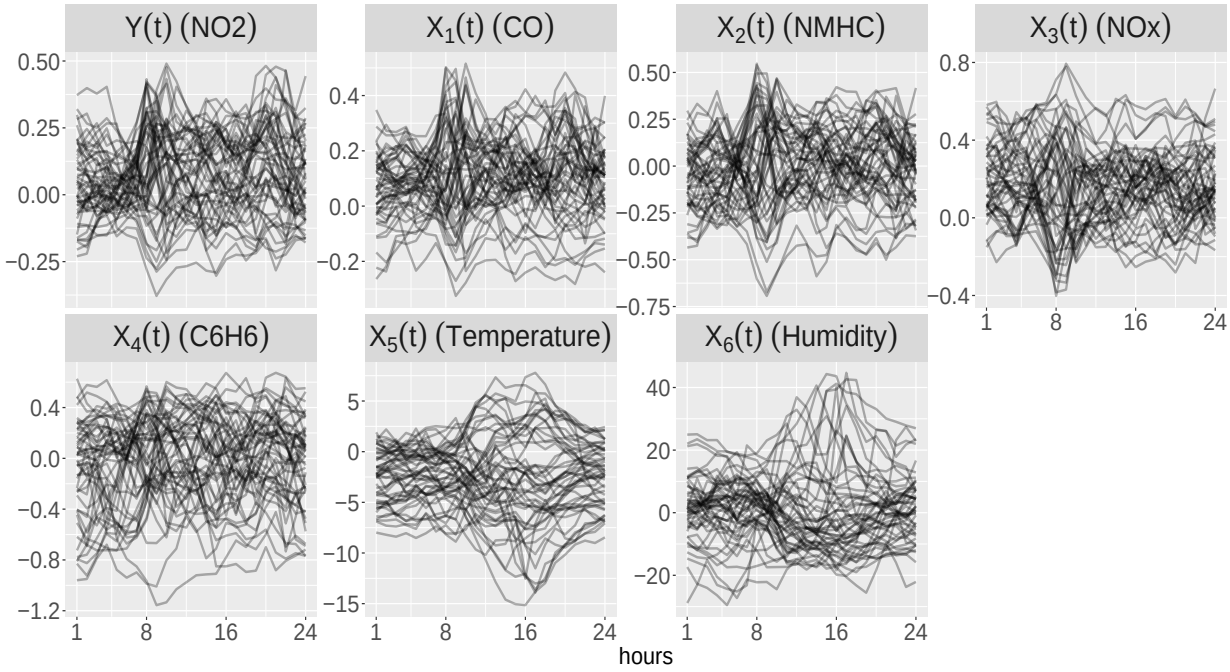

Figure 5: Fifty sample curves of the functional response variable, NO_2, and six functional predictor variables, *CO*, *NMHC*, NO_x, C_6H_6, *Temperature* and *Relative Humidity*, in the air quality data. All curves are functions of hours, and have been centered using the mean curves of corresponding variables.

We repeat a similar procedure 100 times as in the analysis of the Hawaii ocean data. In each repeat, we randomly split the 355 observations into a training set with 150 observations and a test set with 205 observations. Noticing that the variables in this data set are all periodic with period 24 hours, we use the Fourier basis functions. The averages (and standard deviations) of the MSPEs over 100 repeats are 4.55×10^{-3} (0.64×10^{-3}) for the *selected model*, 5.39×10^{-3} (0.32×10^{-3}) for the *main effect model*, and 5.33×10^{-2} (0.36×10^{-3}) for the *full model*. The averages (and standard deviations) of the R^2 and R^2_{pred} of the selected models over the 100 repeats are 91.1% (1.5%) and 82.1% (2.6%),

respectively. Fixing a training set and a test set, we also use the bootstrap method described in Section 6.1 to construct the 95% prediction bands for observations in the test set. From Figure 6 which shows the two types of bands for two observations in the test set, we observe a different pattern from those in the Hawaii ocean data. For this data set, the prediction bands with model selection are obviously wider than those without model selection due to relatively large random fluctuations generated by variable selection.

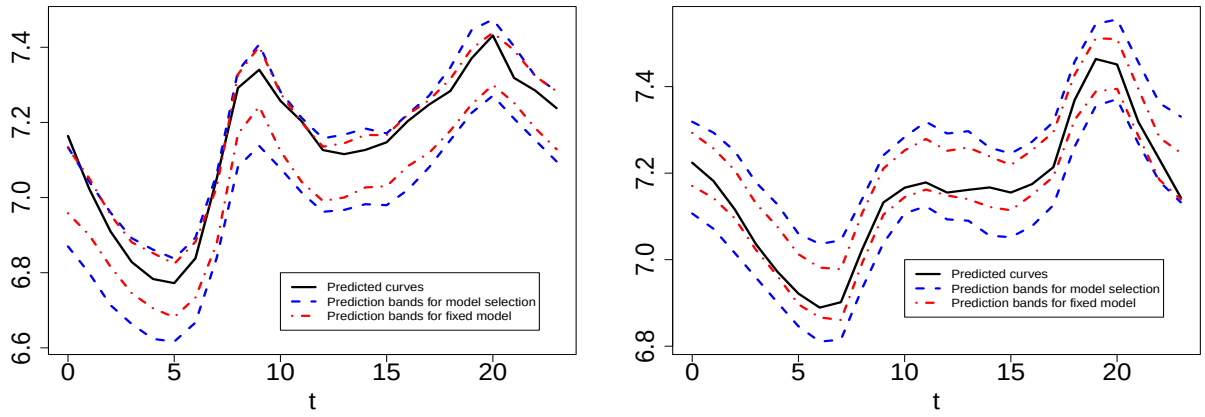

Figure 6: The predicted curves and the 95% bootstrap point-wise prediction bands for two observations in the test data for the air quality data set. In both plots, the black curves are the predicted curves based on the selected model using the training data and our stepwise procedure; the red curves are the bootstrap point-wise prediction bands without considering model selection; the blue curves are the bootstrap point-wise prediction bands with model selection.

Finally, we use all the 355 observations to conduct variable selection. The final model contains all six main effects: $X_1(s)$ (CO), $X_2(s)$ (NMHC), $X_3(s)$ (NO_x), $X_4(s)$ (C_6H_6), $X_5(s)$ (temperature) and $X_6(s)$ (relative humidity), one quadratic effect $X_1(u)X_1(v)$, and two interaction effects $X_1(u)X_3(v)$ and $X_1(u)X_4(v)$. The estimated $\widehat{\psi}_{i,k}(s)$'s, $\widehat{w}_k(t)$'s and $\widehat{\phi}_{ij,k}(u,v)$'s for the selected effects in the final model are presented in Figures S.7∼S.10 of supplementary material. The estimates $\widehat{\beta}_i^{(\mathrm{KL})}(s,t)$'s for the main effects are shown in Figure 7. In the first plot of Figure 7, large positive values appear around the diagonal line with $5 \leq t \leq 20$, which indicates that strong positive contemporaneous association exists between NO_2 and CO from 5am to 8pm, especially 4pm∼8pm. On the contrary, the third plot in Figure 7 shows a negative contemporaneous association between NO_2 and NO_x in all day. The associations between NO_2 and both temperature and relative humidity have

stripe patterns. For example, the concentration of NO_2 at any time tends to be positively associated with the temperature at 5am, but negatively associated with the temperature at 8pm.

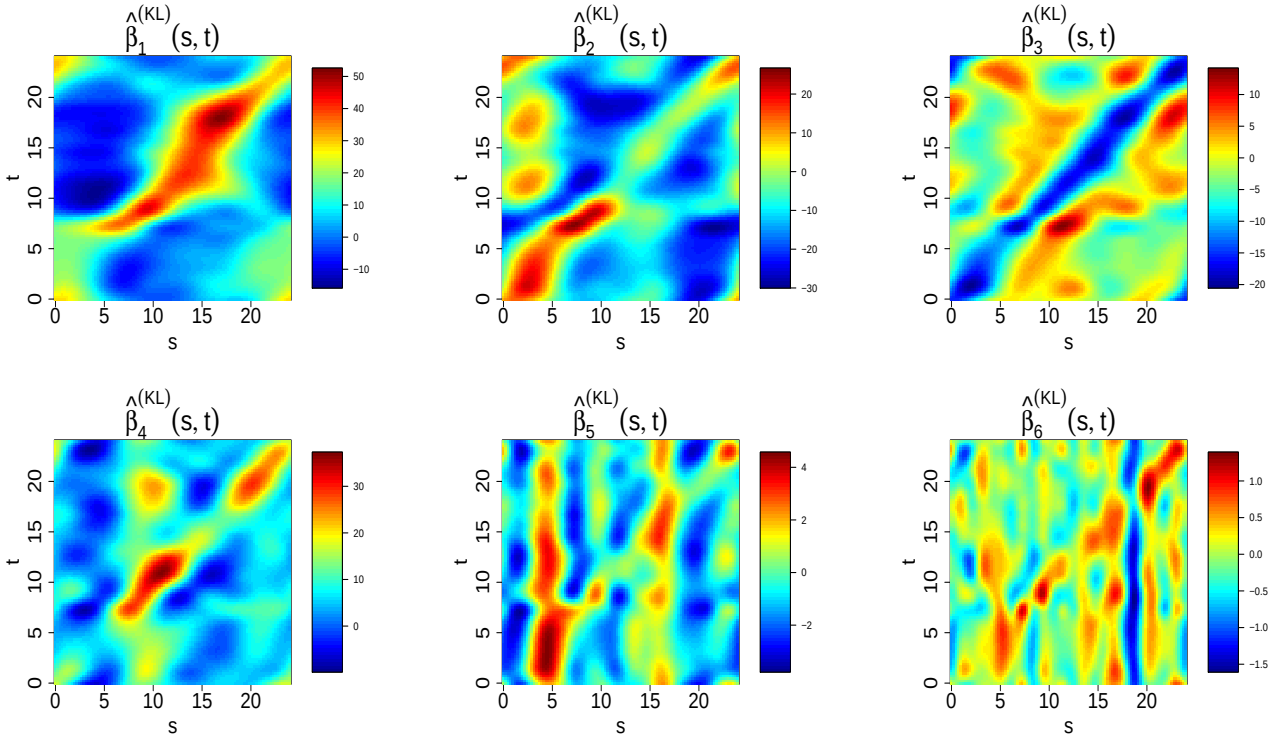

Figure 7: The plots of the estimated coefficient functions $\widehat{\beta}_i^{(\mathrm{KL})}(s,t)$'s for all six main effects: $X_1(s)$ (CO), $X_2(s)$ (NMHC), $X_3(s)$ (NO_x), $X_4(s)$ (C_6H_6), $X_5(s)$ (temperature) and $X_6(s)$ (relative humidity) in the final model for the air quality data.

The existence of the interaction terms means that the slope coefficient function of $X_1(s)$ is affected by other variables. With three probe functions, $g_1(s)$, $g_2(s)$ and $g_3(s)$, which are densities of normal distributions with mean 4, 12, and 20, respectively, and variance 1, we investigate how the change of $X_3(v)$ (NO_x) in different times affects the term $\int_0^1 X_1(v)\widehat{\gamma}_{13}^{(\mathrm{KL})}(s,v,t)dv$ and hence affects the slope coefficient function of $X_1(s)$. We draw the functions $g_i(s)$ and $\int_0^1 g_i(v)\widehat{\gamma}_{13}^{(\mathrm{KL})}(s,v,t)dv$, $1 \le i \le 3$, in Figures S.11 and S.12 in supplementary material. Figure S.12 indicates that the change of the concentration of $X_3(v)$ (NO_x) at 12pm leads to a bigger change of the coefficient function of $X_1(s)$ than the same magnitude change of $X_3(v)$ at 4am and 8pm. An increase of the concentration of $X_3(v)$ at 12pm generally leads to a decrease of the coefficient function around $12 \le s \le 20$.

7 Discussion

We consider the function-on-function regression model with interaction and quadratic effects. For a model with given form, we propose a method to estimate the quadratic regression function and conduct prediction. The key idea of our approach is to estimate a special representation of the regression function induced by its KL decomposition. We show that this representation has the minimum prediction error. Our approach converts the estimation of two- and three- dimensional coefficient functions to the estimation of one- and two- dimensional functions separately, which greatly reduces the computational load. We also propose adaptive penalties to account for varying magnitudes and roughness of coefficient functions. Since the forms of models are usually unspecified in practice, we propose a stepwise procedure for variable selection based on the predictive performance of models. We demonstrate by simulation and real data applications that the selected models by our stepwise procedure have improved estimation and prediction performance than the full models with all main effects and interaction terms and the models with main effects only. When the true model contains unimportant terms, the selected model can even outperform the true model by excluding the terms with little contributions.

Supplementary Materials

The Web Supplementary Material contains proofs of theorems, computational details and additional figures and tables in simulations and application.

References

Collazos, J. A., Dias, R. and Zambom, A. Z. (2016) Consistent variable selection for functional regression models. *Journal of Multivariate Analysis*, **146**, 63–71.

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.

Efron, B. and Tibshirani, R. J. (1994) *An introduction to the bootstrap*. CRC press.

Fuchs, K., Scheipl, F. and Greven, S. (2015) Penalized scalar-on-functions regression with interaction term. *Computational Statistics & Data Analysis*, **81**, 38–51.

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

Hastie, T., Tibshirani, R. and Tibshirani, R. J. (2017) Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*.

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

Kong, D., Staicu, A.-M. and Maity, A. (2016) Classical testing in functional linear models. *Journal of Nonparametric Statistics*, **28**, 813–838.

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

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., Wang, Y. *et al.* (2016) Functional wavelet regression for linear function-on-function models. *Electronic Journal of Statistics*, **10**, 3179–3216.

Malfait, N. and Ramsay, J. O. (2003) The historical functional linear model. *Canadian Journal of Statistics*, **31**, 115–128.

Matsui, H. (2017) Quadratic regression for functional response models. *arXiv:1702.02009*.

Matsui, H. and Konishi, S. (2011) Variable selection for functional regression models via the l1 regularization. *Computational Statistics & Data Analysis*, **55**, 3304–3310.

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.

Qi, X. and Luo, R. (Accepted) Nonlinear function-on-function additive model. *Statistica Sinica*.

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

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

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

Swihart, B. J., Goldsmith, J. and Crainiceanu, C. M. (2014) Restricted likelihood ratio tests for functional effects in the functional linear model. *Technometrics*, **56**, 483–493.

Usset, J., Staicu, A.-M. and Maity, A. (2016) Interaction models for functional regression. *Computational statistics & data analysis*, **94**, 317–330.

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

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

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

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

Zhu, H., Vannucci, M. and Cox, D. D. (2010) A bayesian hierarchical model for classification with selection of functional predictors. *Biometrics*, **66**, 463–473.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American statistical association*, **101**, 1418–1429.