Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Estimation and Model Selection for Nonparametric Function-on-Function Regression

Zhanfeng Wang, Hao Dong, Ping Ma & Yuedong Wang

To cite this article: Zhanfeng Wang, Hao Dong, Ping Ma & Yuedong Wang (2022): Estimation and Model Selection for Nonparametric Function-on-Function Regression, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2022.2037434

To link to this article: https://doi.org/10.1080/10618600.2022.2037434

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 00, NO. 0, 1–11
https://doi.org/10.1080/10618600.2022.2037434

Estimation and Model Selection for Nonparametric Function-on-Function Regression

Zhanfeng Wang[a*], Hao Dong[b*], Ping Ma[c], and Yuedong Wang[b]

[a]International Institute of Finance, The School of Management, University of Science and Technology of China, Hefei, China; [b]Department of Statistics and Applied Probability, University of California, Santa Barbara, Santa Barbara, CA; [c]Department of Statistics, University of Georgia, Athens, GA

ABSTRACT

Regression models with a functional response and functional covariate have received significant attention recently. While various nonparametric and semiparametric models have been developed, there is an urgent need for model selection and diagnostic methods. In this article, we develop a unified framework for estimation and model selection in nonparametric function-on-function regression. We propose a general nonparametric functional regression model with the model space constructed through smoothing spline analysis of variance (SS ANOVA). The proposed model reduces to some of the existing models when selected components in the SS ANOVA decomposition are eliminated. We propose new estimation procedures under either L_1 or L_2 penalty and show that the combination of the SS ANOVA decomposition and L_1 penalty provides powerful tools for model selection and diagnostics. We establish consistency and convergence rates for estimates of the regression function and each component in its decomposition under both the L_1 and L_2 penalties. Simulation studies and real examples show that the proposed methods perform well. Technical details and additional simulation results are available in online supplementary materials.

ARTICLE HISTORY
Received July 2021
Revised October 2021

KEYWORDS
Convergence rate; Regularization; Reproducing kernel Hilbert space; Smoothing spline ANOVA

1. Introduction

With the advance of modern technology, it becomes increasingly common that data are in the form of functions. Rapid developments of new statistical methods for this new type of data has created the field of functional data analysis (FDA) (Yao, Müller, and Wang 2005b; Ramsay and Silverman 2006; Ferraty and Vieu 2006; Hsing and Eubank 2015; Kokoszka and Reimherr 2017; Lin, Müller, and Yao 2018; Lin and Yao 2019). Many semiparametric and nonparametric methods have been proposed for regression with functional response and/or covariates. Nevertheless, there is a lack of flexible model selection and diagnostics methods for functional regression (Wang, Chiou, and Müller 2015; Morris 2015). Ling and Vieu (2018) suggested model building and testing using nonparametric methods as two important future research areas.

In this article, we consider function-on-function regression where both the response Y and covariate X are functions. Denote the functional observations by $\{(Y_i(t), X_i(t)), i = 1, \ldots, n;\ t \in \mathcal{T}\}$, where $\mathcal{T}$ is an arbitrary set. We want to investigate the relationship between X and Y. Ramsay and Silverman (2006) proposed the concurrent linear model (CLM),

$$Y_i(t) = \alpha(t) + \beta(t)X_i(t) + \epsilon_i(t), \quad i = 1, \ldots, n, \qquad (1)$$

where $\alpha(t)$ and $\beta(t)$ are unknown functions to be estimated, $\epsilon_i(t)$, independent of $X_i(t)$, are iid random errors. Model (1) assumes that the value of Y at t depends on X at the same point t only. Ramsay and Silverman (2006) also proposed the functional linear model (FLM),

$$Y_i(t) = \alpha(t) + \int_{\mathcal{T}} \beta(s,t)X_i(s)ds + \epsilon_i(t), \quad i = 1, \ldots, n, \qquad (2)$$

where $Y(t)$ depends on the whole function of $X(\cdot)$, and $\alpha(t)$ and $\beta(s,t)$ are unknown functions to be estimated. Both models (1) and (2) assume that Y is linearly dependent on X, which could be restrictive for some applications. While many nonlinear models have been proposed for scalar-on-function regression (Reiss et al. 2017; Ling and Vieu 2018), nonlinear function-on-function regression has received considerably less attention (Morris 2015; Reimherr, Sriperumbudur, and Taoufik 2018). Müller and Yao (2008) considered an additive model where functional principal component analysis (FPCA) scores were used as predictors.

Different from aforementioned linear and nonlinear models, nonparametric methods, which provide more flexible regression relationship between the response and covariates, have been considered by many authors (Yuan and Cai 2010; Ferraty et al. 2011; Ferraty, Van Keilegom, and Vieu 2012; Lian 2007; Kardi et al. 2016; Ling and Vieu 2018). In particular, the nonparametric concurrent model (NCM) (Zhang, Park, and Wang 2013; Scheipl, Staicu, and Greven 2015; Kim, Maity, and Staicu 2018) assumes that

$$Y_i(t) = g(t, X_i(t)) + \epsilon_i(t), \quad i = 1, \ldots, n, \qquad (3)$$

CONTACT Yuedong Wang yuedong@pstat.ucsb.edu Department of Statistics and Applied Probability, University of California, Santa Barbara, Santa Barbara, CA 93106.

*These authors contributed equally to this work.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

where g is an unknown bivariate function; and an extension of functional linear model (EFLM) (Ma and Zhu 2016; Kim et al. 2018; Reimherr, Sriperumbudur, and Taoufik 2018) assumes that

$$Y_i(t) = \int_{\mathcal{T}} g(t, s, X_i(s))ds + \epsilon_i(t), \quad i = 1, \ldots, n, \qquad (4)$$

where g is an unknown function of three variables. In practice, $Y(t)$ may depend on both $X(t)$ and the whole function of X (see examples in Section 6). Furthermore, data-driven model diagnostic tools are still lacking. It is very difficult for practitioners to decide how to choose an adequate model for their data. Hypothesis testing could be used for the model selection (Xing et al. 2020). However, rigorous tests are only available for simple settings. In addition, since these models were fitted using different methods, the convergence rate of the estimated functions have not been analyzed and compared carefully.

To establish a unified framework for estimation, model selection, and theoretical study, we consider the following nonparametric functional regression model,

$$Y_i(t) = \int_{\mathcal{T}} g(t, s, X_i(t), X_i(s))ds + \epsilon_i(t), \quad i = 1, \ldots, n, \qquad (5)$$

where $g : \mathcal{T} \times \mathcal{T} \times \mathcal{R} \times \mathcal{R} \to \mathcal{R}$ is an unknown function of four variables to be estimated, $\mathcal{R}$ is the range of the functional covariate X, and $\epsilon_i(t)$s are iid random error functions in L^2 with mean zero and finite $\int_{\mathcal{T}} E(\epsilon_i^2(t))dt$. The proposed method can be extended to other nonparametric functional models, such as scalar-on-function and function-on-scalar regression models.

Model (5) is a flexible function-on-function model that has not been studied in the literature. Different from the methods used in Scheipl, Staicu, and Greven (2015), Ma and Zhu (2016), Kim, Maity, and Staicu (2018), and Reimherr, Sriperumbudur, and Taoufik (2018), in this article we construct model space for the multivariate function g through smoothing spline analysis of variance (SS ANOVA) based on the decomposition of tensor product of reproducing kernel Hilbert spaces (RKHS). Model (5) reduces to existing models when selected components in the SS ANOVA decomposition are eliminated. We propose penalized least squares methods for estimating g under either L_1 or L_2 penalty. We establish the representer theorem and develop computational methods for solving both L_1 and L_2-regularized least squares problems. The SS ANOVA model with L_1 penalty provides a systematic approach for model selection and diagnostics of existing models. We establish a coherent framework to study the convergence rates of the penalized least-square estimates as well as each component of the SS ANOVA model under both the L_1 and L_2 penalties. We note that Sun et al. (2018) constructed SS ANOVA for the bivariate function $\beta(s, t)$ in the FLM (2). However, there is an error in their derivation of the representer theorem. We will use a different approach to derive the representer theorem and our estimate is different from that in Sun et al. (2018).

Our methodological contribution for nonparametric function-on-function regression is to develop new estimation procedures under either a L_2 or a joint L_1 and L_2 penalty. Note that through eliminating selected components in a SS ANOVA model, our model (5) can be reduced to models (1)–(4) and other special cases. In practice, researchers may use a reduced model appropriate for their data rather than the general model (5). However, coherent model selection and model diagnostics tools for function-on-function regression are still lacking. Model selection and diagnostics have to be conducted manually, which is laborious, time-consuming, and error prone. The lack of model selection and diagnostics hinders the wide application of the aforementioned models. Our work is the first to provide model selection and diagnostic tools through the combination of SS ANOVA and a joint L_1 and L_2 penalty. In particular, model fitting, model selection and diagnostics can be conducted simultaneously and substantially alleviate the cost of manual model selection and diagnostics.

Our theoretical contribution is to establish a coherent framework to study the convergence rates of the penalized least-square estimates and components in the SS ANOVA model. In the literature, the convergence of the estimators and components are conducted through two technical routes, dimensionless approach (Gu and Qiu 1993) and tensor product approach (Lin 2000). In this article, we reconcile these two methods into one coherent framework. As far as we know, these convergence rates of the estimators and components are the first ones established for function-on-function regression models in a general framework. Moreover, the convergence rates of the estimators and components under a joint L_1 and L_2 penalty are still lacking in the literature even for simple SS ANOVA models. We bridge this gap by establishing the first convergence rates for function-on-function regression models in a general framework. Equipped with all these convergence rates, our theoretical work lays out the first full-fledged analysis framework in this line of research.

The remainder of the article is organized as follows. In Section 2, we present SS ANOVA model for the function g in model (5), and the estimation procedure under L_2 penalty. In Section 3, we present model selection method through SS ANOVA and estimation under L_1 penalty. In Section 4, consistence and convergence rates of estimators of the overall regression function and each component in the SS ANOVA decomposition are obtained. Numerical studies and real examples are given in Section 5 and 6. We conclude in Section 7. Additional technical details, proofs, and codes are included in the supplementary materials.

2. Model Space and Estimation

We introduce the SS ANOVA model in Section 2.1 and present the estimation procedure with L_2 penalty in Section 2.2.

2.1. SS ANOVA Models

Different from the methods used in Scheipl, Staicu, and Greven (2015), Ma and Zhu (2016), Kim, Maity, and Staicu (2018), and Reimherr, Sriperumbudur, and Taoufik (2018), we use tensor product of RKHS's and the SS ANOVA decomposition to build a model space for the regression function g in (5). Denote $\mathcal{H}_1(\mathcal{T})$, $\mathcal{H}_2(\mathcal{T})$, $\mathcal{H}_3(\mathcal{R})$, and $\mathcal{H}_4(\mathcal{R})$ as RKHS's of functions on domains

$\mathcal{T}$ and $\mathcal{R}$, respectively. The choices of these RKHS's depend on the domain of these functions, prior knowledge, and the purpose of study (Wang 2011). For example, for functions on a compact interval, without loss of generality denote the interval as $[0, 1]$, we may consider the Sobolev space

$$W_2^m[0,1] = \{f : f, f', \ldots, f^{(m-1)} \text{ are absolutely continuous}, \int_0^1 (f^{(m)})^2 dx < \infty\}. \quad (6)$$

We consider an SS ANOVA decomposition of g in the tensor product RKHS $\mathcal{H}_1(\mathcal{T})\otimes\mathcal{H}_2(\mathcal{T})\otimes\mathcal{H}_3(\mathcal{R})\otimes\mathcal{H}_4(\mathcal{R})$ (Wang 2011; Gu 2013):

$$\begin{aligned} g(t,s,X(t),X(s)) &= \mu + g_1(t) + g_2(s) + g_3(X(t)) + g_4(X(s)) \\ &+ g_{12}(t,s) + g_{13}(t,X(t)) + g_{14}(t,X(s)) + g_{23}(s,X(t)) \\ &+ g_{24}(s,X(s)) + g_{34}(X(t),X(s)) + g_{123}(t,s,X(t)) \\ &+ g_{124}(t,s,X(s)) + g_{134}(t,X(t),X(s)) \\ &+ g_{234}(s,X(t),X(s)) + g_{1234}(t,s,X(t),X(s)), \end{aligned}$$

where μ is the grand mean, g_1 and other single subscript g are main effects, double subscript g are two-way interactions, and so on. For identifiability of model (5), we need side conditions,

$$\begin{aligned} g_2(s) &= g_{12}(t,s) = g_{23}(s,X(t)) = g_{123}(t,s,X(t)) \\ &= g_{134}(t,X(t),X(s)) = 0. \end{aligned}$$

Thus, g has a decomposition as follows,

$$\begin{aligned} g(t,s,X(t),X(s)) &= \mu + g_1(t) + g_3(X(t)) + g_4(X(s)) \\ &+ g_{13}(t,X(t)) + g_{14}(t,X(s)) + g_{24}(s,X(s)) + g_{34}(X(t),X(s)) \\ &+ g_{124}(t,s,X(s)) + g_{234}(s,X(t),X(s)) + g_{1234}(t,s,X(t),X(s)). \end{aligned} \quad (7)$$

The SS ANOVA decomposition (7) builds a hierarchical structure for function g and handles the side conditions in a natural manner. In addition to having nice interpretation as main effects and interactions, the SS ANOVA decomposition facilitates diagnostics and model selection. It is easy to see that models (1)–(3) are special cases of (7) with certain components equal zero. For example, the NCM (4) and EFLM (3) are special cases with

$$g_4 = g_{14} = g_{24} = g_{34} = g_{124} = g_{234} = g_{1234} = 0, \quad (8)$$

and

$$g_3 = g_{13} = g_{34} = g_{234} = g_{1234} = 0, \quad (9)$$

respectively. Conditions for the CLM (1) and FLM (2) are presented in the Supplemental Materials. Therefore, checking whether components in these conditions equal zero provides a diagnostic tool for existing models. In addition, we can fit existing models by removing components in these conditions. Thus, the proposed estimation methods provide a unified framework for fitting some of the existing models with theoretical guarantees.

We may model g using any subset of components in the SS ANOVA decomposition (7). Given an SS ANOVA model, we can regroup components and rewrite the model space for g as Section 4.5 of Wang (2011)

$$\mathcal{M} = \mathcal{H}_0 \oplus \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_q, \quad (10)$$

where $\mathcal{H}_0$ is a finite dimensional space with an orthonormal basis $\{\phi_1, \ldots, \phi_m\}$, $\mathcal{H}_1, \ldots, \mathcal{H}_q$ are orthogonal RKHS's with reproducing kernels (RK) $R_1, \ldots, R_q$, respectively, and q is the number of components in the SS ANOVA model. We note that different regrouping may be used for different purposes. In particular, we may consider different groups of components for model selection with L_1 penalty (see Section 3).

2.2. Penalized Least Squares with L_2 Penalty

We assume that $Y_i(t)$'s are stochastic processes that belongs to $L^2(\mathcal{T})$. We estimate g as a minimizer of the following penalized least squares (PLS):

$$\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} \left\{ Y_i(t) - \int_{\mathcal{T}} g(t,s,X_i(t),X_i(s))ds \right\}^2 dt + \frac{\lambda}{2}\sum_{j=1}^q \frac{1}{\theta_j}||P_j g||^2, \quad (11)$$

where P_j is a projection operator onto $\mathcal{H}_j$, λ and θ_j are smoothing parameters, and $||\cdot||$ is an induced norm in $\mathcal{M}$.

We shall now present the solution to the PLS (11). Let $\mathcal{H}_1^* = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_q$, and define a new inner product in $\mathcal{H}_1^*$ as

$$\langle f, g\rangle_* = \sum_{j=1}^q \frac{1}{\theta_j}\langle P_j f, P_j g\rangle, \quad (12)$$

where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{M}$. Under the new inner product, the RK of $\mathcal{M}$ is $R = R^0 + R^1$, where

$$\begin{aligned} R^0(t,s,x,z,t',s',x',z') &= R^0_{t,s,x,z}(t',s',x',z') \\ &= \sum_{k=1}^m \phi_k(t,s,x,z)\phi_k(t',s',x',z'), \\ R^1(t,s,x,z,t',s',x',z') &= R^1_{t,s,x,z}(t',s',x',z') \\ &= \sum_{j=1}^q \theta_j R_j(t,s,x,z,t',s',x',z'). \end{aligned}$$

Let $\{\nu_k(t), k = 1, 2, \ldots\}$ be an orthonormal basis of $L^2(\mathcal{T})$ where the first n basis functions are the empirical functional principal components (EFPC) of $Y_1, \ldots, Y_n$. See Hsing and Eubank (2015) for principal component analysis of stochastic processes defined on a general compact metric space. Our result is presented in the following representer theorem, which proof is in the supplemental materials.

Theorem 2.1 (Representer Theorem). (a) The solution to the PLS (11) is

$$\hat{g}(t,s,x,z) = \sum_{k=1}^m d_k\phi_k(t,s,x,z) + \sum_{i=1}^n\sum_{j=1}^n c_{ij}\xi_{ij}(t,s,x,z), \quad (13)$$

where $\xi_{ij}(t,s,x,z) = \int_{\mathcal{T}}\int_{\mathcal{T}} R^1_{t,s,x,z}(t',s',X_i(t'),X_i(s'))\nu_j(t')\,dt'ds'$.

(b) Consider the following standard SS ANOVA model with a scalar response variable,

$$Y_{ij} = L_{ij}g + \epsilon_{ij}, \quad (14)$$

where $Y_{ij} = \langle Y_i, v_j \rangle$, $g \in \mathcal{M}$, $L_{ij}g = \int_{\mathcal{T}}\int_{\mathcal{T}} g(t, s, X_i(t), X_i(s))v_j(t)dsdt$ which is assumed to be a bounded linear functional, and ϵ_{ij} are iid random errors with mean zero. Then the estimate $\hat{g}$ in (11) is the same as the solution to the following PLS

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{Y_{ij} - L_{ij}g\right\}^2 + \frac{\lambda}{2}||P_1^* g||_*^2, \tag{15}$$

where P_1^* is the projection operator onto $\mathcal{H}_1^*$ with the inner product $\langle f, g\rangle_*$ in (12), and $||\cdot||_*$ is the induced norm in $\mathcal{H}_1^*$ by $\langle\cdot,\cdot\rangle_*$.

The representer theorem states that the task of fitting a function-on-function regression model reduces to the task of fitting an SS ANOVA model with a scalar response variable. We now present the computation details. Note that $R = R^0 + R^1$ and it can be shown that

$$\langle g, \int_{\mathcal{T}}\int_{\mathcal{T}} R^0_{t,s,X_i(t),X_i(s)}v_j(t)dsdt\rangle = \sum_{k=1}^{m} a_{ijk}d_k,$$
$$\langle g, \int_{\mathcal{T}}\int_{\mathcal{T}} R^1_{t,s,X_i(t),X_i(s)}v_j(t)dsdt\rangle = \sum_{k=1}^{n}\sum_{l=1}^{n} c_{kl}b_{ijkl},$$

where

$$a_{ijk} = \int_{\mathcal{T}}\int_{\mathcal{T}} \phi_k(t, s, X_i(t), X_i(s))v_j(t)dsdt,$$
$$b_{ijkl} = \int_{\mathcal{T}}\int_{\mathcal{T}} \xi_{kl}(t, s, X_i(t), X_i(s))v_j(t)dsdt.$$

Let $\boldsymbol{Y}_j = (Y_{1j}, \ldots, Y_{nj})^\top$, $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_n^\top)^\top$, $\boldsymbol{d} = (d_1, \ldots, d_m)^\top$, $\boldsymbol{c} = (c_{11}, c_{21}, \ldots, c_{nn})^\top$, $\boldsymbol{T}$ be an $n^2 \times m$ matrix with the $(i + (j-1)n, k)$th element as a_{ijk}, and $\boldsymbol{\Sigma}$ be an $n^2 \times n^2$ matrix with the $(i + (j-1)n, k + (l-1)n)$th element as b_{ijkl}. Then the PLS (15) reduces to

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(Y_{ij} - \sum_{k=1}^{m} a_{ijk}d_k - \sum_{k=1}^{n}\sum_{l=1}^{n} c_{kl}b_{ijkl}\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} c_{ij}b_{ijkl}c_{kl}$$
$$= \frac{1}{n}||\boldsymbol{Y} - \boldsymbol{T}\boldsymbol{d} - \boldsymbol{\Sigma}\boldsymbol{c}||^2 + \frac{\lambda}{2}\boldsymbol{c}^\top\boldsymbol{\Sigma}\boldsymbol{c}. \tag{16}$$

Following the above discussion, one may use existing software to compute the estimate. Details for solving the PLS (16) with data-driven smoothing parameters selection methods such as the generalized cross-validation (GCV) and restricted maximum likelihood (REML), also known as generalized maximum likelihood (GML), can be found in Wang (2011). The recent development in selecting smoothing parameters can be found in Sun, Zhong, and Ma (2020). Since the reduced models have n^2 observations, existing procedures may be infeasible when n is large. One may consider an approximate solution by using the first p ($p \ll n$) EFPCs. Furthermore, to save computational time, we propose a backfitting procedure to estimate smoothing parameters. Details about the backfitting procedure can be found in the Supplemental Materials.

3. Model Selection

In this section, we consider penalized least squares with L_1 penalty for model selection and diagnostics. Given an SS ANOVA model, we can regroup and rewrite the model space in the form (10). Different regrouping may be considered for different purposes. For estimation with L_2 penalty, we usually set $\mathcal{H}_0$ as the space collecting all functions that are not penalized, and other spaces contain main effects and interactions subject to penalties. For model selection, we are interested in whether certain components in the SS ANOVA decomposition can be eliminated. We may use the same regrouping as (10) with L_1 penalties to all components including functions in the space $\mathcal{H}_0$ except for the constant functions. For diagnostics of a specific model, we are interested in whether a group of components in the SS ANOVA decomposition can be eliminated. For example, the NCM is equivalent to all components in (8) equal zero. We may regroup such that one subspace collects all components in (8). Consequently, the L_1 penalty to functions in this subspace encourages all components in (8) to be simultaneously zero. For generality, for a given SS ANOVA model, we rewrite the regrouped model space as

$$\mathcal{M} = \tilde{\mathcal{H}}_0 \oplus \tilde{\mathcal{H}}_1 \oplus \cdots \oplus \tilde{\mathcal{H}}_{\tilde{q}}, \tag{17}$$

where $\tilde{\mathcal{H}}_0$ is a finite dimensional space with an orthonormal basis $\{\tilde{\phi}_1, \ldots, \tilde{\phi}_{\tilde{m}}\}$, and $\tilde{\mathcal{H}}_1, \ldots, \tilde{\mathcal{H}}_{\tilde{q}}$ are orthogonal RKHS's with RKs $\tilde{R}_1, \ldots, \tilde{R}_{\tilde{q}}$, respectively. Note that, as discussed above, each space $\tilde{\mathcal{H}}_j$ may include parametric components in the space $\mathcal{H}_0$ and multiple components of $\mathcal{H}_1, \ldots, \mathcal{H}_q$.

Denote the projection of g onto $\tilde{\mathcal{H}}_0$ as $\sum_{k=1}^{\tilde{m}} d_k\tilde{\phi}_k(t, s, x, z)$ and the RK of $\mathcal{M}$ as $\tilde{R}$. Following the same arguments in Section 2.2, we estimate g via the following PLS with L_1 penalties:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{Y_{ij} - \langle g, \int_{\mathcal{T}}\int_{\mathcal{T}} \tilde{R}_{t,s,X_i(t),X_i(s)}v_j(t)dsdt\rangle\right\}^2 + \lambda_1\sum_{k=1}^{\tilde{m}} w_{1k}|d_k| + \lambda_2\sum_{v=1}^{\tilde{q}} w_{2v}||\tilde{P}_v g||, \tag{18}$$

where $\tilde{P}_v$ is the projection operator onto $\tilde{\mathcal{H}}_v$, λ_1 and λ_2 are tuning parameters, and $0 \le w_{1k}, w_{2v} < \infty$ are prespecified weights that can be selected based on some initial estimates. One may set $w_{11} = 0$ when $\tilde{\phi}_1 = 1$ to avoid penalty to the constant functions. Note that if one subspace, say $\tilde{\mathcal{H}}_1$, collects all components in (8), then $||\tilde{P}_1 g|| = (||g_4||^2 + ||g_{14}||^2 + ||g_{24}||^2 + ||g_{34}||^2 + ||g_{124}||^2 + ||g_{234}||^2 + ||g_{1234}||^2)^{1/2}$ is a group penalty that encourages all components in (8) to be simultaneously zero.

As in Zhang, Cheng, and Liu (2011), instead of (18), Lemma 3.1 indicates that we can solve the following equivalent but more convenient minimization problem:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{Y_{ij} - \langle g, \int_{\mathcal{T}}\int_{\mathcal{T}} \tilde{R}_{t,s,X_i(t),X_i(s)}v_j(t)dsdt\rangle\right\}^2 + \lambda_1\sum_{k=1}^{\tilde{m}} w_{1k}|d_k| + \tau_0\sum_{v=1}^{\tilde{q}} w_{2v}\theta_v^{-1}||\tilde{P}_v g||^2 + \tau_1\sum_{v=1}^{\tilde{q}} w_{2v}\theta_v, \tag{19}$$

subject to $\theta_v \ge 0$ for $1 \le v \le \tilde{q}$, where λ_1, τ_0, and τ_1 are tuning parameters.

Lemma 3.1. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{\tilde{q}})^T$, and $\tau_1 = \lambda_2^2/4\tau_0$. If $\hat{g}$ minimizes (18), set $\hat{\theta}_\nu = \tau_0^{1/2}\tau_1^{-1/2}||\tilde{P}_\nu \hat{g}||$, then $(\hat{\boldsymbol{\theta}}, \hat{g})$ minimizes (19). On the other hand, if $(\hat{\boldsymbol{\theta}}, \hat{g})$ minimizes (19), then $\hat{g}$ minimizes (18).

We choose to solve (19) since it is similar to the PLS with L_2 penalty in (16). The additional penalties on θ_ν's control the sparsity of components in the SS ANOVA model for the purpose of model selection. Specifically, $\theta_\nu = 0$ implies $||\tilde{P}_\nu g|| = 0$. It is easy to show that the solution to (19) is

$$\check{g}(t,s,x,z) = \sum_{k=1}^{\tilde{m}} d_k \tilde{\phi}_k(t,s,x,z) + \sum_{\nu=1}^{\tilde{q}} w_{2\nu}^{-1}\theta_\nu \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\tilde{\xi}_{ij}(t,s,x,z), \tag{20}$$

where $\tilde{\xi}_{ij}(t,s,x,z) = \int_{\mathcal{T}}\int_{\mathcal{T}} \tilde{R}_\nu(t,s,x,z,t',s',X_i(t'),X_i(s'))$ $v_j(t')ds'dt'$ and $\tilde{R}_\nu$ is the RK in $\tilde{\mathcal{H}}_\nu$. Hence, (19) reduces to

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big(Y_{ij} - \sum_{k=1}^{\tilde{m}}\tilde{a}_{ijk}d_k - \sum_{\nu=1}^{\tilde{q}} w_{2\nu}^{-1}\theta_\nu \sum_{k=1}^{n}\sum_{l=1}^{n} c_{kl}\tilde{b}_{ijkl}\Big)^2 \\ &\quad + \lambda_1\sum_{k=1}^{\tilde{m}} w_{1k}|d_k| + \tau_0\sum_{\nu=1}^{\tilde{q}} w_{2\nu}^{-1}\theta_\nu \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} c_{ij}\tilde{b}_{ijkl}c_{kl} \\ &\quad + \tau_1\sum_{\nu=1}^{\tilde{q}} w_{2\nu}\theta_\nu,\end{aligned} \tag{21}$$

where

$$\tilde{a}_{ijk} = \int_{\mathcal{T}}\int_{\mathcal{T}} \tilde{\phi}_k(t,s,X_i(t),X_i(s))v_j(t)dsdt,$$
$$\tilde{b}_{ijkl} = \int_{\mathcal{T}}\int_{\mathcal{T}} \tilde{\xi}_{kl}(t,s,X_i(t),X_i(s))v_j(t)dsdt.$$

Let $\boldsymbol{d} = (d_1, \ldots, d_{\tilde{m}})^\top$, $\boldsymbol{c} = (c_{11}, c_{21}, \ldots, c_{nn})^\top$, $\boldsymbol{w}_2 = (w_{21}, \ldots, w_{2\tilde{q}})^T$, $\tilde{\boldsymbol{T}}$ be an $n^2 \times \tilde{m}$ matrix with the $(i+(j-1)n, k)$-th element as $\tilde{a}_{ijk}$, $\tilde{\boldsymbol{\Sigma}}_\nu$ be an $n^2 \times n^2$ matrix with the $(i+(j-1)n, k+(l-1)n)$-th element as $\tilde{b}_{ijkl}$, and $\tilde{\boldsymbol{\Sigma}} = \sum_{\nu=1}^{\tilde{q}} w_{2\nu}^{-1}\theta_\nu\tilde{\boldsymbol{\Sigma}}_\nu$. Then (21) is simplified as

$$\frac{1}{n}||\boldsymbol{Y} - \tilde{\boldsymbol{T}}\boldsymbol{d} - \tilde{\boldsymbol{\Sigma}}\boldsymbol{c}||^2 + \lambda_1\sum_{k=1}^{\tilde{m}} w_{1k}|d_k| + \tau_0\boldsymbol{c}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{c} + \tau_1\boldsymbol{w}_2^T\boldsymbol{\theta} \tag{22}$$

subject to $\theta_\nu \geq 0, \nu = 1, \ldots, \tilde{q}$.

We use a backfitting algorithm to solve (22). With fixed $\boldsymbol{d}$ and $\boldsymbol{\theta}$, (22) reduces to (16) with a working response $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \tilde{\boldsymbol{T}}\boldsymbol{d}$. Therefore, we can update $\boldsymbol{c}$ using existing methods with τ_0 chosen by the GCV or the REML method. With fixed $\boldsymbol{c}$ and τ_0, noting that $\tilde{\boldsymbol{\Sigma}}$ is linear in $\boldsymbol{\theta}$, (22) reduces to the standard LASSO for a linear regression model subject to constraints on $\boldsymbol{\theta}$. With fixed $\boldsymbol{\theta}$ and τ_1, we apply an LASSO regularization procedure to update $\boldsymbol{d}$ with λ_1 chosen by k-fold cross-validation. Then with fixed $\boldsymbol{d}$ and λ_1, we apply a constrained quadratic programming procedure to update $\boldsymbol{\theta}$. The selection of τ_1 is equivalent to finding M such that $\boldsymbol{w}_2^T\boldsymbol{\theta} \leq M$. We apply the k-fold cross-validation or the BIC method to select M. The complete algorithm can be found in the supplemental materials.

4. Statistical Properties

In this section, we study consistence and convergence rate of the regression function estimates $\hat{g}$ and $\check{g}$ as well as their components in the SS ANOVA decompositions under both the L_1 and L_2 penalties.

4.1. Loss Function and Regularity Conditions

We will first define a loss function similar to that in Gu (2013). Denote the least squares as $l(g) = \int_{\mathcal{T}}\{Y(t) - \int_{\mathcal{T}} g(t,s,X(t),X(s))ds\}^2\,dt = \sum_{j=1}^{\infty} Q_j(g,Y)$ where $Q_j(g,Y) = \left\{\langle Y, v_j\rangle - \langle g, \int_{\mathcal{T}}\int_{\mathcal{T}} R_{t,s,x(t),x(s)}v_j(t)dsdt\rangle_{\mathcal{M}}\right\}^2$. Let D be the Fréchet derivative with respect to g, $u_j(g;Y) = DQ_j(g;Y)$, $w_j(g;Y) = D^2Q_j(g;Y)$, $u(g;Y) = \sum_{j=1}^{n} u_j(g;Y)$, and $w(g;Y) = \sum_{j=1}^{n} w_j(g;Y)$. It is easy to check that the first and second Fréchet derivatives of $Q_j(g,Y)$ at directions h_1 and h_2 are

$$\begin{aligned}u_j(g;Y)(h_1) &= -2\left\{\langle Y, v_j\rangle - \langle g, \int_{\mathcal{T}}\int_{\mathcal{T}} R_{t,s,x(t),x(s)}v_j(t)dsdt\rangle_{\mathcal{M}}\right\} \\ &\quad\times \langle h_1, \int_{\mathcal{T}}\int_{\mathcal{T}} R_{t,s,x(t),x(s)}v_j(t)dsdt\rangle_{\mathcal{M}}, \\ w_j(g;Y)(h_1,h_2) &= 2\langle h_1, \int_{\mathcal{T}}\int_{\mathcal{T}} R_{t,s,x(t),x(s)}v_j(t)dsdt\rangle_{\mathcal{M}} \\ &\quad\times \langle h_2, \int_{\mathcal{T}}\int_{\mathcal{T}} R_{t,s,x(t),x(s)}v_j(t)dsdt\rangle_{\mathcal{M}}.\end{aligned}$$

Let g_0 be the true function of g and $w_j(g;Y) = w_j(g;Y)(g,g)$. We denote

$$V(g_1,g_2) = 2E[g_1g_2E\{D^2l(g_0)\}] = 2E[g_1g_2E\{\sum_{j=1}^{\infty} w_j(g_0;Y)\}]$$

as a quadratic functional, and define the loss function

$$V(g) = V(g,g) = 2E[g^2E\{\sum_{j=1}^{\infty} w_j(g_0;Y)\}]. \tag{23}$$

Note that $\langle g_0, \int_{\mathcal{T}}\int_{\mathcal{T}} R_{t,s,x(t),x(s)}v_j(t)dsdt\rangle_{\mathcal{M}} = 0$ when $j \geq n+1$. Consequently

$$V(g) = 2E[g^2E\{w(g_0;Y)\}]. \tag{24}$$

We shall show that the loss function $V(g)$ is equivalent to the integrated mean square loss function $E(g^2)$ under some conditions. We denote $J(g_1,g_2) = \langle P_1^*g_1, P_1^*g_2\rangle_*$ as another quadratic functional, and define the penalty function $J(g) = J(g,g) = \langle P_1^*g, P_1^*g\rangle_* = ||P_1^*g||_*^2$. A quadratic functional V is said to be completely continuous with respect to another quadratic functional J, if for any $\epsilon > 0$, there exist a finite number of linear functionals $L_1, \ldots, L_k$ such that $L_jf = 0, j = 1, \ldots, k$, implies that $V(g) \leq \epsilon J(g)$. To derive the convergence rate, we need the following conditions.

Condition 1. V is completely continuous with respect to $J = ||P_1^*g||_*^2$.

Condition 1 is satisfied when $c_1 \leq w(g_0;y) \leq c_2$ holds for some positive constants c_1 and c_2. From Theorem 3.1 of Weinberger (1974), under Condition 1, there exist eigenvalues

ρ_k and eigenfunctions ψ_k such that $V(\psi_k, \psi_j) = \delta_{k,j}$ and $J(\psi_k, \psi_j) = \rho_k \delta_{k,j}$ where $\delta_{k,j}$ is the Kronecker delta and $0 \leq \rho_k \uparrow \infty$.

Define a quadratic functional as

$$\frac{1}{n}\sum_{i=1}^{n} u(g_{0i}; Y_i)g_i + \frac{1}{2}V(g - g_0) + \frac{\lambda}{2}J(g), \tag{25}$$

where $g_i = g(t, s, X_i(t), X_i(s))$ and $g_{0i} = g_0(t, s, X_i(t), X_i(s))$. By the Fourier series expansion with basis $\{\psi_k\}$, we have $g = \sum_k \zeta_k \psi_k$ and $g_0 = \sum_k \zeta_{k,0}\psi_k$. It is easy to show that the minimizer $\tilde{g}$ of (25) has Fourier coefficients $\tilde{\zeta}_k = (\beta_k + \zeta_{k,0})/(1 + \lambda\rho_k)$, where $\beta_k = -n^{-1}\sum_{i=1}^{n} u(g_{0i}; Y_i)\psi_k(t, s, X_i(t), X_i(s))$.

We need the following conditions for the eigenvalues ρ_k and the eigenfunctions ψ_k.

Condition 2. For k large enough and some $\beta > 0$, the eigenvalues ρ_k satisfy $\rho_k > \beta k^r$ with $r > 1$ and $\sum_k \rho_k \zeta_{k,0}^2 < \infty$.

Condition 3. For any k and j, $\text{var}\{\psi_k(X)\psi_j(X)w(g_0(X), Y)\} \leq c_3$ with some $c_3 < \infty$, where $\psi_k(X) = \psi_k(t, s, X(t), X(s))$ and $g_0(X) = g_0(t, s, X(t), X(s))$.

Condition 4. For g in a convex set B_0 around g_0 containing $\hat{g}$ and $\tilde{g}$, $c_4 w(g_0, Y) \leq w(g, Y) \leq c_5 w(g_0, Y)$ holds uniformly for some $0 < c_4 < c_5 < \infty$.

Conditions 1–4 are common assumptions for convergence rate analysis of the SS ANOVA estimates, which were also made in Gu (2013). The $\sum_k \rho_k \zeta_{k,0}^2 < \infty$ in Condition 2 is a special case in Theorem 9.15 with $p = 1$ in Gu (2013), and states that the growth rate of the eigenvalues ρ_k is at k^r, which controls how fast λ approaches zero. Condition 3 requires the fourth moment of ψ_i is bounded. Condition 4 bounds $w(g, Y)$ at g in a convex set B_0 around g_0.

4.2. Convergence Results Under L_2 Penalty

In order to present our theorem, we need the O_p notation. For a sequence of random variables, $\{A_n\}$, and a sequence of constants, $\{a_n\}$, the notation $A_n = O_p(a_n)$, means that $\{A_n/a_n\}$ is stochastically bounded (or bounded in probability). That is, for any $\tau > 0$, there exit a constant $K(\tau)$ and an integer $n(\tau)$ such that if $n \geq n(\tau)$, then

$$P(|A_n/a_n| \leq K(\tau)) \geq 1 - \tau.$$

More details and examples of this can be found in Section 1.2 of Serfling (1980).

Theorem 4.1. Assume that $g_0 \in \mathcal{M}$. Under Conditions 1–4, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$,

$$(V + \lambda J)(\hat{g} - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda).$$

For the penalized least-square estimation, we transform the integration in the objection function (11) to a summation criterion by using EFPCs of $Y(t)$. Different from the proofs of consistence and convergence rate of function estimation in Gu (2013), we need to deal with an infinite summation for EFPC decomposition of $Y(t)$. The proof of Theorem 4.1 is presented in the Supplemental Materials. In addition, we derive convergence rate for each term in the SS ANOVA model as follows.

To study the convergence rate of each component in the decomposition of g, as Lin (2000), we consider the special case with $\mathcal{T} = \mathcal{R} = [0, 1]$ and $g \in \otimes^4 W_2^m([0, 1])$ where $\otimes^k W_2^m([0, 1])$ denotes the tensor product of k Sobolev spaces $W_2^m([0, 1])$ defined in (6). Then the main effects lies in $W_2^m([0, 1])$ and the kth order interactions lies in a tensor product space of $\otimes^k W_2^m([0, 1])$. From the common norm $||\cdot||_{W_2^m}$ in $W_2^m([0, 1])$, we can deduce a norm $||\cdot||_{\otimes^q W_2^m}$ in the tense product space of $\otimes^q W_2^m([0, 1])$. For example with $\otimes^q W_2^0([0, 1])$, the norm is $||g(z)||^2_{\otimes^q W_2^0} = \int g^2(z)dz$. Following Lin (2000), under the condition $c_1 \leq w(g_0; y) \leq c_2$ for some positive constants c_1 and c_2, we can show that with $m > 0$ the norms $||\cdot||^2_{\otimes^q W_2^m}$ and $|\cdot|^2_{\otimes^q W_2^m} = V(\cdot) + ||P_1^*\cdot||_*^2$ are equivalent on the tensor product space $\otimes^q W_2^m([0, 1])$, and when $m = 0$ the norms $||\cdot||^2_{\otimes^q W_2^0}$ and $|\cdot|^2_{\otimes^q W_2^0} = V(\cdot)$ are equivalent on the tensor product space $\otimes^q W_2^0([0, 1])$, denoted by $||\cdot||^2_{\otimes^q W_2^m} \sim |\cdot|^2_{\otimes^q W_2^m}$ and $||\cdot||^2_{\otimes^q W_2^0} \sim |\cdot|^2_{\otimes^q W_2^0} = V(\cdot)$. From the SS ANOVA decomposition (7), it is easy to see that

$$\begin{aligned}||g||^2_{\otimes^q W_2^m} &= \mu^2 + ||g_1||^2_{W_2^m} + ||g_3||^2_{W_2^m} + ||g_4||^2_{W_2^m} \\ &\quad + ||g_{13}||^2_{\otimes^2 W_2^m} + ||g_{14}||^2_{\otimes^2 W_2^m} \\ &+ ||g_{24}||^2_{\otimes^2 W_2^m} + ||g_{34}||^2_{\otimes^2 W_2^m} + ||g_{124}||^2_{\otimes^3 W_2^m} + ||g_{234}||^2_{\otimes^3 W_2^m} \\ &\quad + ||g_{1234}||^2_{\otimes^4 W_2^m}.\end{aligned}$$

Thence, we have

$$\begin{aligned}|g|^2_{\otimes^q W_2^m} &\sim \mu^2 + ||g_1||^2_{W_2^m} + ||g_3||^2_{W_2^m} + ||g_4||^2_{W_2^m} + ||g_{13}||^2_{\otimes^2 W_2^m} \\ &\quad + ||g_{14}||^2_{\otimes^2 W_2^m} + ||g_{24}||^2_{\otimes^2 W_2^m} + ||g_{34}||^2_{\otimes^2 W_2^m} \\ &\quad + ||g_{124}||^2_{\otimes^3 W_2^m} + ||g_{234}||^2_{\otimes^3 W_2^m} + ||g_{1234}||^2_{\otimes^4 W_2^m}.\end{aligned} \tag{26}$$

Since $|g|^2_{\otimes^q W_2^0} = V(g)$, we have $|\hat{g} - g_0|^2_{\otimes^q W_2^0} = O_p(n^{-1}\lambda^{-1/r} + \lambda)$. Combined with (26), we have the following corollary.

Corollary 4.1. Assume that $g \in \otimes^4 W_2^m([0, 1])$, the conditions in Theorem 4.1 hold, and $c_1 \leq w(g_0; y) \leq c_2$ holds for some positive constants c_1 and c_2. As $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$, we have

$$\begin{aligned}&||\hat{g}_I - g_{0I}||^2_{W_2^0} = O_p(n^{-1}\lambda^{-1/r} + \lambda), \\ &I = 1,\ 3,\ 4,\ (1,3),\ (1,4),\ (2,4),\ (3,4),\ (1,2,4), \\ &\qquad (2,3,4),\ (1,2,3,4),\end{aligned} \tag{27}$$

where indices inside each index set I indicate the component in the SS ANOVA decomposition.

Corollary 4.1 provides convergence rates of the component estimators in the SS ANOVA decomposition under the integrated mean square loss function. If we set the tuning parameter λ as $\lambda = [n(\log n)^{-a}]^{-\frac{m}{2m+1}}$ where $a > 0$, it satisfies that $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$. Hence, we have for n larger enough,

$$n^{-1}\lambda^{-1/r} = n^{-(1-\frac{m}{(2m+1)r})}(\log n)^{-\frac{am}{(2m+1)r}} < [n(\log n)^{-a}]^{-\frac{m}{2m+1}}.$$

From Theorem 4.1 and Corollary 4.1, it follows that

$$O_p(n^{-1}\lambda^{-1/r}+\lambda) = O_p([n(\log n)^{-a}]^{-\frac{m}{2m+1}}).$$

Furthermore, when $r \geq \max\{2, 2m\}$, we set $\lambda = [n(\log n)^{-a}]^{-\frac{2m}{2m+1}}$, then $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$. We have faster convergence rates for $\hat{g}$ and its components as

$$O_p(n^{-1}\lambda^{-1/r}+\lambda) = O_p([n(\log n)^{-a}]^{-\frac{2m}{2m+1}}),$$

which is similar to the convergence result in Lin (2000).

4.3. Convergence Results Under L_1 penalty

We consider model (5) with $g \in \mathcal{M}$ where the model space is given in (17). Write $g = g^{(0)} + g^{(1)}$ where $g^{(0)} \in \tilde{\mathcal{H}}_0$ and $g^{(1)} \in \tilde{\mathcal{H}}_1 \oplus \cdots \oplus \tilde{\mathcal{H}}_{\tilde{q}}$. Then

$$\begin{aligned} w_j(g;Y)(h_1,h_2) =& 2\langle h_1, \int_{\mathcal{T}}\int_{\mathcal{T}} \tilde{R}_{t,s,x(t),x(s)} v_j(t)dsdt\rangle_{\mathcal{M}} \\ &\times \langle h_2, \int_{\mathcal{T}}\int_{\mathcal{T}} \tilde{R}_{t,s,x(t),x(s)} v_j(t)dsdt\rangle_{\mathcal{M}} \\ =& 2\langle h_1, \int_{\mathcal{T}}\int_{\mathcal{T}} \tilde{R}^0_{t,s,x(t),x(s)} v_j(t)dsdt\rangle_{\mathcal{M}} \\ &\times \langle h_2, \int_{\mathcal{T}}\int_{\mathcal{T}} \tilde{R}^0_{t,s,x(t),x(s)} v_j(t)dsdt\rangle_{\mathcal{M}} \\ &+2\langle h_1, \int_{\mathcal{T}}\int_{\mathcal{T}} \tilde{R}^1_{t,s,x(t),x(s)} v_j(t)dsdt\rangle_{\mathcal{M}} \\ &\times \langle h_2, \int_{\mathcal{T}}\int_{\mathcal{T}} \tilde{R}^1_{t,s,x(t),x(s)} v_j(t)dsdt\rangle_{\mathcal{M}} \\ =& w_j^0(g^{(0)};Y) + w_j^1(g^{(1)};Y), \end{aligned}$$

where $\tilde{R}^0$ and $\tilde{R}^1$ are the RKs of $\tilde{\mathcal{H}}_0$ and $\tilde{\mathcal{H}}_1 \oplus \cdots \oplus \tilde{\mathcal{H}}_{\tilde{q}}$. The cross term is zero because $\tilde{\mathcal{H}}_0$ and $\tilde{\mathcal{H}}_1 \oplus \cdots \oplus \tilde{\mathcal{H}}_{\tilde{q}}$ are orthogonal to each other. Let $w^0(g^{(0)};Y) = \sum_{j=1}^{n} w_j^0(g^{(0)};Y)$ and $w^1(g^{(1)};Y) = \sum_{j=1}^{n} w_j^1(g^{(1)};Y)$.

Define $V^*(g) = V_0(g^{(0)}) + V_1(g^{(1)})$ and $J^*(g) = J_0(g^{(0)}) + J_1(g^{(1)})$, where $V_0(g^{(0)}) = \sqrt{2E[g^{(0)2}E\{w^0(g^{(0)};Y)\}]}$, $V_1(g^{(1)}) = \sqrt{2E[g^{(1)2}E\{w^1(g^{(1)};Y)\}]}$, $J_0(g^{(0)}) = \sum_{k=1}^{\tilde{m}} ||d_k\phi_k|| = \sum_{k=1}^{\tilde{m}} |d_k|$, and $J_1(g^{(1)}) = \sum_{j=1}^{\tilde{q}} ||\tilde{P}_v g||$. Let $\check{g}$ and $\check{d}_k$ $(k = 1,\ldots,\tilde{m})$ be the minimizer of (18), and $\check{\lambda} = \max\{\lambda_1,\lambda_2\}$. Then, we have the following convergence results.

Theorem 4.2. Assume that $g_0 \in \mathcal{M}$. Under Conditions 1–4 with $\hat{g}$ in Condition 4 being replaced by $\check{g}$, as $\check{\lambda} \to 0$ and $n\check{\lambda}^{2/r} \to \infty$,

$$(V^* + \check{\lambda}J^*)(\check{g} - g_0) = O_p(n^{-1/2}\check{\lambda}^{-1/2r} + \check{\lambda}^{1/2}).$$

Corollary 4.2. Assume conditions in Theorem 4.2 hold and $\tilde{c}_1 \leq w(g_0;y) \leq \tilde{c}_2$ holds for some positive constants $\tilde{c}_1$ and $\tilde{c}_2$, as $\check{\lambda} \to 0$ and $n\check{\lambda}^{2/r} \to \infty$, we have

$$|\check{d}_k - d_{0,k}| = O_p(n^{-1/2}\check{\lambda}^{-1/2r} + \check{\lambda}^{1/2}) \qquad k = 1,\ldots,\tilde{m},$$

$$||\tilde{P}_v\check{g} - \tilde{P}_v g_0||_{W_2^0} = O_p(n^{-1/2}\check{\lambda}^{-1/2r} + \check{\lambda}^{1/2}) \qquad v = 1,\ldots,\tilde{q},$$

where $\{d_{0,k}\}$ are coefficients of parametric functions in the true function g_0.

Proofs of Theorem 4.2 and Corollary 4.2 are given in the supplementary materials.

Remark. The convergence rate in Theorem 4.2 and the component convergence rate in Corollary 4.2 are the square root of the rate in Theorem 4.1 and Corollary 4.1. This is due to the fact that the square of L_2 norm was used in the Section 4.2 while the L_2 norm was used in the Section 4.3.

5. Simulation Results

Simulation studies are conducted to evaluate the performance of the proposed model selection and estimation methods. We generate data based on model (5) with $\mathcal{T} = \mathcal{R} = [0,1]$. For convenience of presentation, let $f(t) = \int_0^1 g(t,s,X(t),X(s))ds$. We consider a factorial design with the following three choices of f:

M1: $f(t) = 1 + 0.5\cos(2\pi t) + 3X(t) + 3(X(t)-0.5)(2t-1)^2$,
M2: $f(t) = 1 + 3\int_0^1 (2X(s)-1)ds + 3(t-0.5)^2 + 10\int_0^1 (X(s)-0.5)((2s-1)^2 - 1/3)ds$,
M3: $f(t) = 1 + 5(2X(t)-1)^3 + 2\int_0^1 (2X(s)-1)ds + 5(t-0.5)^2$.

M1 is an NCM model with $g_1(t) = 0.5\cos(2\pi t)$, $g_3(X(t)) = 3X(t)$, and $g_{13} = 3(X(t)-0.5)(2t-1)^2$. M2 is an EFLM model with $g_1(t) = 3(t-0.5)^2$, $g_4(X(s)) = 3(2X(s)-1)$, and $g_{24} = 10(X(s)-0.5)((2s-1)^2 - 1/3)$. M3 is neither an NCM nor an EFLM model with $g_1(t) = 5(t-0.5)^2$, $g_3(X(t)) = 5(2X(t)-1)^3$, and $g_4(X(s)) = 2(2X(s)-1)$. All other terms in the SS ANOVA decomposition not mentioned above are set to be zero. Two choices of sample size $n = 40$ and $n = 80$; and two choices of error standard deviation: $\sigma = 0.2$ and $\sigma = 0.5$. For M1, we generate $X_i(t) = a_{i1}(\cos\{2\pi(t + a_{i2})\} + 1)/2$ where $a_{i1}, a_{i2} \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$. For M2 and M3, $X_i(t) = \exp(X_i^*(t))/(1+\exp(X_i^*(t)))$ where $X_i^*(t) \overset{\text{iid}}{\sim} \text{GP}(0,k)$, and $\text{GP}(0,k)$ is the Gaussian process with mean 0 and Matérn kernel $k(t,s) = (1+8|t-s|)\exp(-8|u_l - v_l|)$. We generate observations of $Y_i(t)$ on 20 equally spaced grid points in $[0,1]$. All simulation are repeated 100 times. In all simulations in this section and real data examples in the next section, we compute the approximate estimates using the first $p = 10$ EFPCs and estimate smoothing parameters using the GCV method.

We assume that $g \in \mathcal{M} = \otimes^4 W_2^2([0,1])$. It is known that $W_2^2[0,1] = \mathcal{H}^{(0)} \oplus \mathcal{H}^{(1)} \oplus \mathcal{H}^{(2)}$ which corresponds to constant, linear, and smooth (nonparametric) functions (Wahba 1990; Wang 2011). Then

$$\begin{aligned} &\otimes^4 W_2^2([0,1]) \\ =&(\mathcal{H}_1^{(0)} \oplus \mathcal{H}_1^{(1)} \oplus \mathcal{H}_1^{(2)}) \otimes (\mathcal{H}_2^{(0)} \oplus \mathcal{H}_2^{(1)} \oplus \mathcal{H}_2^{(2)}) \\ &\otimes (\mathcal{H}_3^{(0)} \oplus \mathcal{H}_3^{(1)} \oplus \mathcal{H}_3^{(2)}) \otimes (\mathcal{H}_4^{(0)} \oplus \mathcal{H}_4^{(1)} \oplus \mathcal{H}_4^{(2)}) \end{aligned}$$

$$= \mathcal{H}_\emptyset \oplus \sum_{k=1}^{4}\sum_{j=1}^{2} \mathcal{H}_k^{(j)} \oplus \sum_{1\le k_1<k_2\le 4}\sum_{j_1,j_2=1}^{2} \mathcal{H}_{k_1k_2}^{(j_1j_2)} \tag{28}$$

$$\oplus \sum_{1\le k_1<k_2<k_3\le 4}\sum_{j_1,j_2,j_3=1}^{2} \mathcal{H}_{k_1k_2k_3}^{(j_1j_2j_3)}$$

$$\oplus \sum_{j_1,j_2,j_3,j_4=1}^{2} \mathcal{H}_{1234}^{(j_1j_2j_3j_4)}, \tag{29}$$

where $\mathcal{H}_\emptyset$ contains constant functions; $\mathcal{H}_k^{(1)}$ and $\mathcal{H}_k^{(2)}$ contain the linear and smooth main effects of the kth variable; $\mathcal{H}_{k_1k_2}^{(11)}$, $\mathcal{H}_{k_1k_2}^{(12)}$, $\mathcal{H}_{k_1k_2}^{(21)}$, and $\mathcal{H}_{k_1k_2}^{(22)}$ contain the linear-linear, linear-smooth, smooth-linear, and smooth-smooth two-way interactions between the k_1th and the k_2th variables; and so on. When fitting the SS ANOVA model under the L_2 penalty, we regroup the model space $\mathcal{M}$ with

$$\mathcal{H}_0 = \mathcal{H}_\emptyset \oplus \sum_{k\in I_1} \mathcal{H}_k^{(1)} \oplus \sum_{(k_1,k_2)\in I_2} \mathcal{H}_{k_1k_2}^{(11)} \oplus \sum_{(k_1,k_2,k_3)\in I_3} \mathcal{H}_{k_1k_2k_3}^{(111)} \oplus \mathcal{H}_{1234}^{(1111)}$$

collecting all parametric components for $I_1 = \{1,3,4\}, I_2 = \{(1,3),(1,4),(2,4),(3,4)\}, I_3 = \{(1,2,4),(2,3,4)\}$, and each of the remaining subspaces constitute $\mathcal{H}_1$ to $\mathcal{H}_q$ with $q = 10$ in (10). Specifically, $\mathcal{H}_1$ to $\mathcal{H}_3$ collect smooth main effects with $\mathcal{H}_1 = \mathcal{H}_1^{(2)}, \mathcal{H}_2 = \mathcal{H}_3^{(2)}$, and $\mathcal{H}_3 = \mathcal{H}_4^{(2)}$; $\mathcal{H}_l = \mathcal{H}_{k_1k_2}^{(12)} \oplus \mathcal{H}_{k_1k_2}^{(21)} \oplus \mathcal{H}_{k_1k_2}^{(22)}$ for $l = 4,5,6,7$ collects the linear-smooth, smooth-linear, and smooth-smooth two-way interactions between the k_1th and the k_2th variables for $(k_1,k_2) \in I_2$; $\mathcal{H}_l$ for $l = 8,9$ collects three-way interactions that involving nonparametric components for variables in I_3, and $\mathcal{H}_{10}$ collects four-way interactions that involving nonparametric components. Notice that subspaces containing g_2, g_{12}, g_{23}, g_{123} and g_{134} are removed for identifiablity. When fitting the SS ANOVA model with L_1 penalty, we use the same regrouping for comparison with the L_2 estimates. That is, we set $\tilde{\mathcal{H}}_l = \mathcal{H}_l$ for $l = 0,1,\ldots,10$ and enforce L_1 penalties to all elements in the subspace $\tilde{\mathcal{H}}_0$.

For comparison, we also fit the NCM (4) and EFLM (3) with reduced SS ANOVA models under conditions (8) and (9), respectively. We evaluate the performance using the following integrated mean square errors (IMSE),

$$\text{IMSE} = \int_0^1 (\hat{f}(t) - f(t))^2 dt.$$

We apply both the L_2 and L_1 penalized methods to fit the SS ANOVA model with all main effects and two-way interactions, and apply the L_2 penalized method to fit the NCM and EFLM. Table 1 presents IMSEs under different simulation settings. The IMSEs are generally smaller when the error variance is smaller or the sample size is larger. In general fitting correct models leads to smaller IMSEs. The proposed general model has comparable IMSEs with NCM under M1 and comparable IMSEs with EFLM under M2. Under M3 when neither NCM nor EFLM holds, the general model has much smaller IMSEs. IMSEs under L_1 penalty are larger than that under L_2, confirming the theoretical results in Theorems 4.1 and 4.2 that the PLS estimate under L_1 has a slower convergence rate.

For the purpose of selection, we are interested in the nonzero entries in $\boldsymbol{d}$ and $\boldsymbol{\theta}$. To evaluate the model selection performance,

Table 1. Averages and standard deviations (in parentheses) of the integrated mean square error (IMSE) for M1, M2 and M3 with sample sizes $n = 40$ and $n = 80$ and error standard deviations $\sigma = 0.2$ and $\sigma = 0.5$.

n	σ	Method	M1	M2	M3
40	0.2	L_2	0.006(0.001)	0.003(0.001)	**0.006(0.001)**
		L_1	0.080(0.019)	0.011(0.008)	0.019(0.006)
		NCM	**0.003(0.007)**	0.321(0.039)	0.152(0.019)
		EFLM	0.011(0.002)	**0.002(0.001)**	0.096(0.049)
40	0.5	L_2	0.017(0.006)	0.010(0.003)	**0.017(0.004)**
		L_1	0.080(0.032)	0.022(0.010)	0.034(0.007)
		NCM	**0.016(0.036)**	0.323(0.039)	0.158(0.019)
		EFLM	0.034(0.011)	**0.006(0.003)**	0.172(0.064)
80	0.2	L_2	0.005(0.001)	0.003(0.000)	**0.005(0.001)**
		L_1	0.081(0.012)	0.004(0.001)	0.012(0.004)
		NCM	**0.003(0.017)**	0.324(0.025)	0.155(0.013)
		EFLM	0.009(0.002)	**0.002(0.000)**	0.093(0.012)
80	0.5	L_2	**0.014(0.007)**	0.006(0.002)	**0.013(0.002)**
		L_1	0.086(0.021)	0.012(0.005)	0.021(0.006)
		NCM	0.021(0.050)	0.329(0.025)	0.158(0.013)
		EFLM	0.022(0.009)	**0.004(0.002)**	0.146(0.018)

The smallest IMSE among four methods for each data example is presented in bold.

Table 2. Averages and standard deviations (in parentheses) of specificity (SPE), sensitivity (SEN), and F_1 score for M1, M2 and M3 models with sample sizes $n = 40$ and $n = 80$ and error stadard deviations $\sigma = 0.2$ and $\sigma = 0.5$.

n	σ		SPE	SEN	F_1
40	0.2	*M1*	0.929(0.064)	0.997(0.033)	0.849(0.126)
		M2	0.928(0.053)	0.977(0.085)	0.828(0.093)
		M3	0.872(0.065)	1.000(0.000)	0.806(0.086)
	0.5	*M1*	0.895(0.049)	0.997(0.033)	0.777(0.084)
		M2	0.932(0.054)	0.98(0.080)	0.841(0.107)
		M3	0.842(0.068)	1.000(0.000)	0.768(0.082)
80	0.2	*M1*	0.944(0.061)	1.000(0.000)	0.881(0.122)
		M2	0.938(0.067)	1.000(0.000)	0.870(0.130)
		M3	0.972(0.039)	1.000(0.000)	0.951(0.065)
	0.5	*M1*	0.907(0.065)	1.000(0.000)	0.808(0.115)
		M2	0.920(0.070)	0.997(0.033)	0.833(0.127)
		M3	0.908(0.049)	1.000(0.000)	0.850(0.070)

we compute three criteria: specificity (SPE), sensitivity (SEN) and F_1 scores:

$$\text{SPE} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{F}_1 = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}},$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives. Here we consider the nonzero parametric or nonparametric components in (19) as true positives and the corresponding nonzero entries in the estimates $\boldsymbol{d}$ and $\boldsymbol{\theta}$ as identified positives.

Table 2 presents sensitivities, specificities, and F_1 scores. Overall, the proposed method performed very well under all simulation settings. The selection performance improves as sample size increases or error variance decreases.

6. Real Examples

In this section, we illustrate the proposed method using four real data examples.

Example 1. (Canada weather). Weekly temperature and precipitation were collected from 35 Canadian weather observation stations, and the goal is to study the relationship between temperature ($X(t)$) and precipitation ($Y(t)$). Ramsay, Hooker, and Graves (2010) fitted the CLM and FLM models.

Example 2. (Gait curve). The gait curve data consist of movement cycle curve of hip and knee angles in degrees of 39 boys (Ramsay and Silverman 2006). Relationship of knee angle curve ($Y(t)$) and hip angle curve ($X(t)$) is studied. We fit the proposed model and check the NCM model.

Example 3. (Growth curve). The growth curve data consist of heights of 39 boys and 54 girls from age 1 to 18 (Jones and Bayley 1941). Regarding the first derivative of growth curve as the response $Y(t)$ and the growth curve as the covariate $X(t)$, Verzelen, Tao, and Müller (2012) studied the dynamics of girl growth by a nonlinear dynamic systems. We fit the proposed model and check the NCM model in Verzelen, Tao, and Müller (2012). We use functions in the `fda` R package, `smooth.basis` and `deriv.fd`, to smooth the growth curves and then compute their first derivatives.

Table 3. Model selection results from the PLS with L_1 penalty.

	Function components									
Data	g_1	g_3	g_4	g_{13}	g_{14}	g_{34}	g_{24}	g_{124}	g_{234}	g_{1234}
Weather	1	0	1	0	1	1	1	0	0	0
Gait	1	1	0	1	0	0	0	0	0	0
Growth	1	1	0	1	1	0	1	0	0	1
GE/JPM	1	1	0	0	0	0	0	1	1	1
IBM/JPM	0	1	0	1	0	0	0	0	1	0

NOTE: The selected and removed components are denoted by 1s and 0s, respectively.

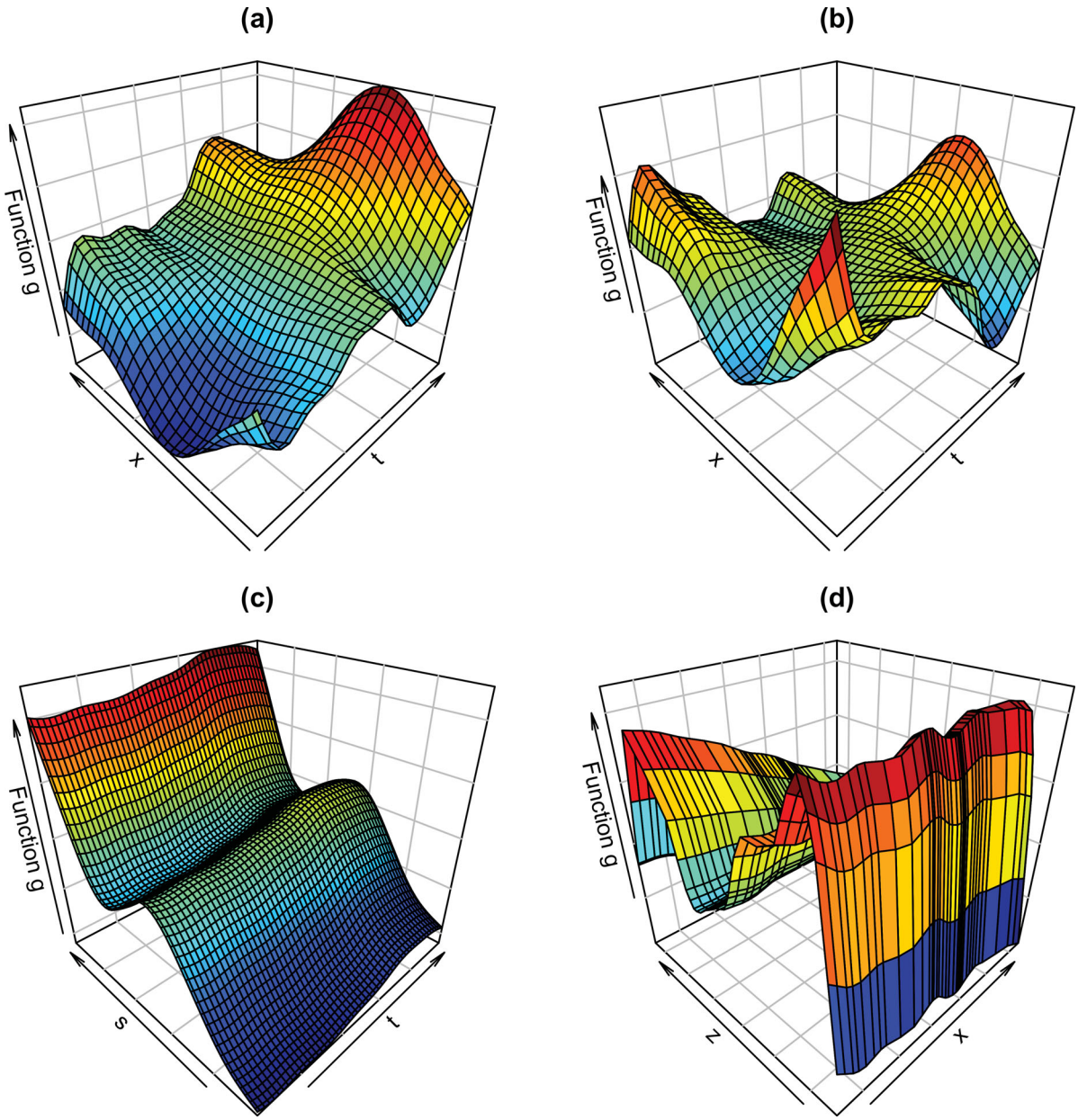

Figure 1. 3D-plots for estimates of the function $g(t, s, x, z)$ or interaction. Gait data: (a) g as a function of t and x, (b) the interaction g_{13} as a function of t and x; Weather data: (c) g as a function of t and s with x and z fixed at their mean values, (d) g as a function of x and z with t and s set to be the value of the 10th observation time.

Table 4. Mean square prediction error (MSPE) and their standard deviations (in parentheses) based on 10-fold cross-validation for each real dataset.

	Weather	Gait	Growth	GE/JPM	IBM/JPM
$L1+L2$	0.537(0.737)	0.106(0.029)	**0.204(0.08)**	**0.470(0.445)**	**0.509(0.341)**
$L1$	0.592(0.373)	0.261(0.034)	0.245(0.097)	0.553(0.561)	0.552(0.375)
$L2$	0.540(0.694)	1.797(1.083)	0.620(0.053)	0.496(0.477)	0.561(0.371)
EFLM	**0.371(0.502)**	1.043(0.592)	0.659(0.103)	0.64(0.663)	0.577(0.362)
NCM	0.669(0.512)	**0.079(0.017)**	0.208(0.081)	0.588(0.473)	0.516(0.353)

The smallest MSPE among four methods for each data example is presented in bold.

Example 4. (Stock price). We collect weekly stock prices of GE (General Electric), IBM (International Business Machines), and JPM (JP Morgan Chase & Co.) from 1996 to 2015. The goal is to study the relationship between two stock prices. For illustration, we build two models to investigate how weekly GE prices or IBM prices ($Y(t)$) depend on weekly JPM price ($X(t)$).

Data in Example 1–3 can be found in the `fda` R package, and stock data in Example 4 will be provided with code in the supplements. We transform the domains and ranges for all real data examples such that $\mathcal{T} = \mathcal{R} = [0,1]$. We consider model (5) for all real data examples with $g \in \mathcal{M} = \otimes^4 W_2^2([0,1])$ and the same SS ANOVA regrouping as discussed in Section 5 including all high-order interactions.

We first apply our model selection method with L_1 penalty. We avoid the selection of the constant function by setting $w_{11} = 0$ where $\tilde{\phi}_1 = 1$. Table 3 lists the selected components. For Gait data, the selected model reduces to an NCM model which indicates that the knee angle depends on the current time and current hip angle only. The other selected models contain components that are functions of $X(t)$ and $X(s)$, indicating that $Y(t)$ depends on both current values and the whole function.

After selection with the L_1 penalty, we then use the PLS with L_2 penalty to estimate the selected models. Figure 1 presents the 3D-plots for Gait and Weather data. Note that the selected model for Gait data is an NCM model, which only depends on current time t and x. Therefore, to visualize the estimate of function $g(t,s,x,z)$ and two-way interaction g_{13}, we construct 3D-plots of g against (t,x) and g_{13} against (t,x). For Weather data, we construct 3D-plots of $g(t,s,x,z)$ against (t,s) with x and z fixed at their mean values, and 3D-plots of $g(t,s,x,z)$ against (x,z) with t and s set to be the value of the tenth observation time. The 3D plots show complex interactions between variables.

To evaluate the prediction performance, we consider three approaches to fit model (5): PLS with L_2 penalty (denoted as L_2), PLS with L_1 penalty (denoted as L_1), and PLS with L_2 penalty after selection with the L_1 penalty (denoted as $L_1 + L_2$). For comparison, we consider EFLM and NCM with PLS subject to the L_2 penalty. We use 10-fold cross-validation to show performance of prediction from the five fitted methods and average the prediction errors on the test fold. For each real data example, we compute the mean square prediction error (MSPE)

$$\text{MSPE} = \frac{1}{n}\sum_{j=1}^{10}\sum_{i\in j\text{th fold}}\int_0^1 (Y_i(t) - \hat{Y}_i^{(-j)}(t))^2 dt,$$

where $\hat{Y}_i^{(-j)}(t)$ is the prediction of $Y_i(t)$ without using data in the jth fold.

MSPE's are listed in Table 4. For the Weather data, the EFLM has the smallest MSPE. For the Gait data, the NCM has the smallest MSPE, which agrees with model selection result. For the Growth and Stock price data, the propose general model has smaller MSPEs. Overall, the $L_1 + L_2$ approach performs well.

7. Conclusion and Discussion

In this article, we developed a unified framework for estimation, model selection, and theoretical study for function-on-function nonparametric regression. We proposed a flexible function-on-function nonparametric regression model that includes concurrent linear model, function linear model, extensions of function linear model, and nolinear concurrent model as special cases. We modeled the regression function using SS ANOVA decomposition of a tensor product of RKHS's. We fitted the model using penalized least squares with L_1 or L_2 penalty. The L_1 regularization can be used to select important components in an SS ANOVA model and check the adequacy of existing models. Consistency and convergence rates of regression function estimation were obtained. We note that the proposed estimation and model selection methods apply to existing models which are special cases of the proposed general model, and the theoretical results fill the gaps in the literature.

It is not difficult to extend the proposed methods to fit nonparametric scalar-on-function and function-on-scalar regression models. For simplicity, we considered the case where functional data were observed at equally spaced points in our simulation. When functional data are observed at irregular points, we can estimate each function first using any existing smoothing methods, and then apply the proposed method (Yao, Müller, and Wang 2005a; Zhang and Chen 2007). Future research topics include extending the proposed model and methods to more complex data such as repeated measures, developing methods and theories with sparse and irregular functional data, and developing computational methods for big data.

Supplemental Materials

Supplementary Materials: Relationship between the proposed model and some existing models, backfitting and model selection algorithm and some proofs. (.pdf file)

Computational R code and dataset: Computational R code for simulation and real examples. This file also contains the dataset used in one of the real examples. (.zip file)

Acknowledgments

The authors thank the editor, the associate editor, and the two referees for their constructive comments and suggests that have led to significant improvement in this article.

Funding

Zhanfeng Wang's research was partially supported by research grants from National Natural Science Foundation of China (No. 11971457), and Anhui Provincial Natural Science Foundation (No. 1908085MA06). Ping Ma's research was partially supported by U.S. National Science Foundation under grants DMS-1903226, DMS-1925066, DMS-2124493, the U.S. National Institute of Health under grant R01GM122080. Yuedong Wang's research was partially supported by U.S. National Science Foundation under grant DMS-1507620, the U.S. National Institute of Health under grant R01DK130067.

ORCID

Ping Ma http://orcid.org/0000-0002-5728-3596

References

Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. (2011), "Kernel Regression with Functional Response," *Electronic Journal of Statistics*, 5, 159–171. [1]

Ferraty, F., Van Keilegom, I., and Vieu, P. (2012), "Regression When Both Response and Predictor are Functions," *Journal of Multivariate Analysis*, 109, 10–28. [1]

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis*, New York: Springer. [1]

Gu, C. (2013), *Smoothing Spline ANOVA Models* (Vol. 297), New York: Springer. [3,5,6]

Gu, C., and Qiu, C. (1993), "Smoothing Spline Density Estimation: Theory," *The Annals of Statistics*, 21, 217–234. [2]

Hsing, T., and Eubank, R. (2015), *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, West Sussex: Wiley. [1,3]

Jones, H., and Bayley, N. (1941), "The Berkeley Growth Study," *Child Development*, 12, 167–173. [9]

Kardi, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A. and Audiffren, J. (2016), "Operator-Valued Kernels for Learning from Functional Response Data," *Journal of Machine Learning Research*, 17, 1–54. [1]

Kim, J. S., Maity, A., and Staicu, A.-M. (2018), "Additive Nonlinear Functional Concurrent Model," *Statistics and its interface*, 11, 669–685. [1,2]

Kim, J. S., Staicu, A. M., Maity, A., Carroll, R. J., and Ruppert, D. (2018), "Additive Function-on-Function Regression," *Journal of Computational and Graphical Statistics*, 27, 234–244. [2]

Kokoszka, P., and Reimherr, M. (2017), *Introduction to Functional Data Analysis*, Boca Raton, FL: Chapman and Hall/CRC. [1]

Lian, H. (2007), "Nonlinear Functional Models for Functional Responses in Reproducing Kernel Hilbert Spaces," *The Canadian Journal of Statistics*, 35, 597–606. [1]

Lin, Y. (2000), "Tensor Product Space Anova Models," *The Annals of Statistics*, 28, 734–755. [2,6,7]

Lin, Z., Müller, H., and Yao, F. (2018), "Mixture Inner Product Spaces and their Application to Functional Data Analysis," *The Annals of Statistics*, 46, 370–400. [1]

Lin, Z., and Yao, F. (2019), "Intrinsic Riemannian Funcitonal Data Analysis," *The Annals of Statistics*, 47, 3533–3577. [1]

Ling, N., and Vieu, P. (2018), "Nonparametric Modelling for Functional Data: Selected Survey and Tracks for Future," *Statistics: A Journal of Theoretical and Applied Statistics*, 52, 1–16. [1]

Ma, H., and Zhu, Z. (2016), "Continuously Dynamic Additive Models for Functional Data," *Journal of Machine learning Research*, 150, 1–13. [2]

Morris, J. S. (2015), "Functional Regression," *Annual Review of Statistics and Its Application*, 2, 321–359. [1]

Müller, H.-G., and Yao, F. (2008), "Functional Additive Models," *Journal of the American Statistical Association*, 103, 1534–1544. [1]

Ramsay, J. O., Hooker, G., and Graves, S. (2010), *Functional Data Analysis with R and Matlab*, New York: Springer. [9]

Ramsay, J. O., and Silverman, B. W. (2006), *Functional Data Analysis* (2nd ed.), New York: Springer. [1,9]

Reimherr, M., Sriperumbudur, B., and Taoufik, B. (2018), "Optimal Prediction for Additive Function-on-Function Regression," *Electronic Journal of Statistics*, 12, 4571–4601. [1,2]

Reiss, P. T., Goldsmith, J., Shang, H., and Ogden, R. T. (2017), "Methods for Scalar-on-Function Regression," *International Statistical Review*, 85, 228–249. [1]

Scheipl, F., Staicu, A. M., and Greven, S. (2015), "Functional Additive Mixed Models," *Journal of Computational and Graphical Statistics*, 24, 477–501. [1,2]

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley. [6]

Sun, X., Du, P., Wang, X., and Ma, P. (2018), "Optimal Penalized Function-on-Function Regression Under a Reproducing Kernel Hilbert Space Framework," *Journal of the American Statistical Association*, 113, 1601–1611. [2]

Sun, X., Zhong, W., and Ma, P. (2020), "An Asymptotic and Empirical Smoothing Parameters Selection Method for Smoothing Spline ANOVA Models in Large Samples," *Biometrika*, 108, 1–8. [4]

Verzelen, N., Tao, W., and Müller, H.-G. (2012), "Inferring Stochastic Dynamics from Functional Data," *Biometrika*, 99, 533–550. [9]

Wahba, G. (1990), *Spline Models for Observational Data* (Vol. 59), CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia: SIAM. [7]

Wang, J. L., Chiou, J. M., and Müller, H. G. (2015), "Functional Data Analysis," *Annual Review of Statistics and Its Application*, 3, 257–295. [1]

Wang, Y. (2011), *Smoothing Splines: Methods and Applications*, Boca Raton: CRC Press. [3,4,7]

Weinberger, H. (1974), *Variational Methods for Eigenvalue Approximation* (Vol. 15), CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia: SIAM. [5]

Xing, X., Liu, M., Ma, P., and Zhong, W. (2020), "Minimax Nonparametric Parallelism Test," *Journal of Machine Learning Research*, 21, 1–47. [2]

Yao, F., Müller, H.-G., and Wang, J.-L. (2005a), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590. [10]

——— (2005b), "Functional linear regression analysis for longitudinal data," *The Annals of Statistics*, 33, 2873–2903. [1]

Yuan, M., and Cai, T. T. (2010), "A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression," *The Annals of Statistics*, 38, 3412–3444. [1]

Zhang, H. H., Cheng, G., and Liu, Y. (2011), "Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models," *Journal of the American Statistical Association*, 106, 1099–1112. [4]

Zhang, J.-T., and Chen, J. (2007), "Statistical Inferences for Functional Data," *The Annals of Statistics*, 35, 1052–1079. [10]

Zhang, X., Park, B., and Wang, J. (2013), "Time-Varying Additive Models for Longitudinal Data," *Journal of the American Statistical Association*, 108, 983–998. [1]