Journal of Machine Learning Research 27 (2026) 1-54 Submitted 5/25; Revised 12/25; Published 2/26

Nonlinear function-on-function regression by RKHS

Peijun Sang PEIJUN.SANG@UWATERLOO.CA
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, ON, N2L3G1, Canada

Bing Li BXL9@PSU.EDU
Department of Statistics
Pennsylvania State University
University Park, PA, 16802, USA

Editor: Shiqian Ma

Abstract

We propose a nonlinear function-on-function regression model where both the covariate and the response are random functions. The nonlinear regression is carried out in two steps: we first construct Hilbert spaces to accommodate the functional covariate and the functional response, and then build a second-layer Hilbert space for the covariate to capture nonlinearity. The second-layer space is assumed to be a reproducing kernel Hilbert space, which is generated by a positive definite kernel determined by the inner product of the first-layer Hilbert space for X–this structure is known as the nested Hilbert spaces. We develop estimation procedures to implement the proposed method, which allows the functional data to be observed at different time points for different subjects. Furthermore, we establish the convergence rate of our estimator as well as the weak convergence of the predicted response in the Hilbert space. Numerical studies including both simulations and a data application are conducted to investigate the performance of our estimator in finite sample.

Keywords: functional data analysis, linear operator, Tikhonov regularization, weak convergence, Hawaii Ocean Time-series

1 Introduction

With the development of techniques in data collection, functional data have become increasingly common in modern statistical applications. As a useful tool to treat such data, functional data analysis (FDA) has been widely applied to diverse fields such as neural science, chemometrics, environmetrics and finance. A comprehensive overview of FDA can be found in several monographs (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Kokoszka and Reimherr, 2017). An important problem in FDA is to study the relationship between a response, which can be a scalar, a vector or a function, and a functional covariate. The aforementioned references as well as other monographs like Horváth and Kokoszka (2012) and Hsing and Eubank (2015) describe many ideas and methods to tackle this problem. In this paper, we propose a general regression model that allows for flexible nonlinear relations between the functional covariate and the functional response.

The scalar-on-function regression, where a scalar response is regressed against a functional covariate, has been extensively studied. In particular, estimation and inference for linear scalar-on-function regression models have been one of the focal points of the FDA research in the past two decades. There are two mainstream approaches to fitting a functional linear model. The first represents the slope function in the linear model as a linear combination of a finite number of basis functions, so that fitting the model is reduced to estimating the linear coefficients. The basis can be either a pre-determined basis such as the B-spline basis or a data-driven basis such as the estimated eigenfunctions from the functional principal component analysis (FPCA). See Cardot et al. (2003), Hall and Horowitz (2007) and references therein. The second approach assumes that the slope function belongs to a reproducing kernel Hilbert space (RKHS) and resorts to the representer theorem to fit the model; see, for example, Yuan and Cai (2010) and Cai and Yuan (2012). In the asymptotic development, the aforementioned work mainly considered convergence rates of estimation and prediction in functional linear models. Shang and Cheng (2015) studied confidence intervals for regression mean and prediction intervals for a future response in generalized functional linear models. Cuesta-Albertos et al. (2019) considered goodness-of-fit of functional linear models to investigate whether they adequately characterize the relation between a scalar response and a functional covariate.

Unlike scalar-on-function regression, function-on-function regression has not yet been extensively developed due to its computational complexity. However, predicting a random function by another random function is an important problem in many applications. For instance, in the Canadian weather data set (Ramsay and Silverman, 2005, Chapter 12.4), the temperature profile is used to predict the annual profile of precipitation, rather than the total precipitation. As with scalar-on-function regression, linear models were most frequently investigated in the literature of function-on-function regression. Ramsay and Silverman (2005) used tensor products of spline basis functions and Yao et al. (2005b) and Crambes and Mas (2013) used estimated eigenfunctions through the FPCA to estimate the (bivariate) slope function in a linear function-on-function regression model. The idea of a regularized estimation through RKHS proposed by Yuan and Cai (2010) for a linear scalar-on-function regression model was extended to the function-on-function case by Sun et al. (2018). However, as pointed out by Müller and Yao (2008), linear relations may not be adequate to characterize dependence of one random function on another function in some applications. The Hawaii ocean data set studied in Qi and Luo (2019) justified this statement. Nonlinear function-on-function regression models have received substantially less attention in FDA (Reimherr et al., 2018), let alone statistical inference of such models. Rao and Reimherr (2023) developed a neural network framework to model the nonlinear relationship between multiple functional covariates and a functional response. But they did not establish theoretical guarantees for the proposed framework. In contrast, Luo and Qi (2024) established a functional universal approximation theorem, which enables a good approximation to the complex relationship between a functional covariate and a functional response. They further found the convergence rate for the prediction error for their estimate.

In this paper, we propose a nonlinear function-on-function regression model that relies on the structure of the nested Hilbert spaces. Specifically, we assume that the functional covariate X and the functional response Y each resides in a Hilbert space, denoted by $\mathcal{H}_X$ and $\mathcal{H}_Y$, respectively. To achieve a nonlinear relation, we build another Hilbert space of functions on $\mathcal{H}_X$, which is assumed to be an RKHS generated by a positive definite kernel κ determined by the inner product of $\mathcal{H}_X$. Following Li and Song (2017), we refer to this space as the second-layer space, and the two-space structure as the nested Hilbert spaces. Lastly, we construct a linear operator from the second-layer RKHS to the target space $\mathcal{H}_Y$, which gives rise to a nonlinear relation. This idea is similar in spirit to that of support vector machines for regression (Friedman et al., 2009, Chapter 5.8). A similar structure was adopted in Lee and Li (2022), and their main focus is to estimate directional relations from multivariate functional data. We then propose an implementation method based on this nonlinear function-on-function regression model, which allows for profiles with irregularly spaced time points and contaminated with measurement errors for both the covariate and the response. Numerical studies demonstrate that our proposed method can still achieve relatively good predictive performance when random functions are sparsely observed and are contaminated with measurement errors. We establish consistency with the convergence rate of our estimator under some mild conditions. In the ideal case where full trajectories of both the covariate and the response are available, we even find that the convergence rate of the estimated regression operator can be improved from $n^{-1/4}$ in Li and Song (2017) in function-on-function sufficient dimension reduction to $n^{-1/3}$ in the current setting under mild conditions. More importantly, unlike the previous theoretical work on function-on-function regression that was focused on convergence rates without an asymptotic distribution, we establish weak convergence of the predicted mean for a future observation in $\mathcal{H}_Y$ in the ideal case. This enables us to construct both pointwise confidence intervals and a simultaneous confidence band for the conditional mean.

The rest of the paper is organized as follows. In Section 2, we propose the nonlinear function-on-function regression model. In Section 3, we develop an algorithm to fit the model and propose suitable methods to select the tuning parameters that are involved in the estimation procedure. In the ideal case where full trajectories of both the covariate and the response are available, consistency and weak convergence of the estimator are studied in Sections 4 and 5, respectively. Pointwise confidence intervals and a simultaneous confidence band for the conditional mean are then constructed based on the weak convergence result. In Section 6, we develop consistency and convergence rate for our estimator proposed in Section 3 under the more realistic case where both X and Y are sparsely observed and contaminated with measurement errors. In Section 7, we conduct simulation studies to investigate the performance of our proposed model in finite samples. The new model is applied to a data set to further demonstrate its performance in Section 8. Some concluding remarks are made in Section 9. All technical proofs are delegated to the supplementary material.

We first introduce some notations about linear operators. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two generic separable Hilbert spaces and $A : \mathcal{G}_1 \to \mathcal{G}_2$ a linear operator. Then A is a Hilbert-Schmidt operator if $\sum_{i\in\mathbb{N}} \|Ae_i\|_{\mathcal{G}_2}^2 < \infty$, where $\mathbb{N} = \{1, 2, 3, \ldots\}$, and $\{e_i : i \in \mathbb{N}\}$ is any orthonormal basis (ONB) of $\mathcal{G}_1$. The square root of this finite number is the Hilbert-Schmidt norm, and is denoted by $\|A\|_{\mathrm{HS}}$. We will use $\|\cdot\|_{\mathrm{OP}}$ to denote the operator norm. Let $\mathcal{B}(\mathcal{G}_1, \mathcal{G}_2)$ denote the class of all bounded linear operators from $\mathcal{G}_1$ to $\mathcal{G}_2$; the special case $\mathcal{B}(\mathcal{G}, \mathcal{G})$ is abbreviated by $\mathcal{B}(\mathcal{G})$. For a linear operator $A \in \mathcal{B}(\mathcal{G}_1, \mathcal{G}_2)$, we use $\ker(A)$ to denote the kernel of A; that is, $\ker(A) = \{f \in \mathcal{H}_1 : Af = 0\}$; we use $\mathrm{ran}(A)$ to denote the range of A; that is, $\mathrm{ran}(A) = \{Af : f \in \mathcal{G}_1\}$; we use $\overline{\mathrm{ran}}(A)$ to denote the closure of $\mathrm{ran}(A)$ in $\mathcal{G}_2$. For a self-adjoint operator $A \in \mathcal{B}(\mathcal{G})$, we have $\ker(A)^{\perp} = \overline{\mathrm{ran}}(A)$ and $\overline{\mathrm{ran}}(A)^{\perp} = \mathrm{ran}(A)^{\perp} = \ker(A)$. Given two arbitrary positive sequences $\{a_n : n \in \mathbb{N}\}$ and $\{b_n : n \in \mathbb{N}\}$, we write $a_n \prec b_n$ if

$a_n/b_n \to 0$, write $a_n \succ b_n$ if $b_n \prec a_n$, write $a_n \preceq b_n$ if a_n/b_n is a bounded sequence and write $a_n \asymp b_n$ if $a_n \preceq b_n$ and $b_n \preceq a_n$. For two real numbers a and b, we use $a \wedge b$ to represent the minimum of a and b.

2 Model construction

In this section, we first introduce the concept of the nested Hilbert space, which plays an important role in our model construction, and then lay out the detailed steps to build our nonlinear function-on-function regression model.

2.1 Nested Hilbert space for predictor

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathbb{I}$ an interval in $\mathbb{R}$, $\mathcal{H}_X$ and $\mathcal{H}_Y$ Hilbert spaces of functions on $\mathbb{I}$. Let $X : \Omega \to \mathcal{H}_X$, $Y : \Omega \to \mathcal{H}_Y$ be random elements in $\mathcal{H}_X$ and $\mathcal{H}_Y$ measurable with respect to $\mathcal{F}/\mathcal{F}_X$ and $\mathcal{F}/\mathcal{F}_Y$, where $\mathcal{F}_X$ and $\mathcal{F}_Y$ denote the Borel σ-algebra generated by the open sets in $\mathcal{H}_X$ and $\mathcal{H}_Y$. Further, we assume that both $\mathcal{H}_X$ and $\mathcal{H}_Y$ are separable to ensure the existence of the conditional distribution of Y given X. Let P_X and P_Y denote the distributions of X and Y, and $P_{Y|X} : \mathcal{H}_X \times \mathcal{F}_Y \to \mathbb{R}$ the conditional distribution of Y given X.

Let $\kappa : \mathcal{H}_X \times \mathcal{H}_X \to \mathbb{R}$ be a positive definite kernel and $\mathfrak{M}_X$ be the RKHS generated by κ.

Assumption 1 *The reproducing kernel κ is induced by the inner product in $\mathcal{H}_X$; that is, there exists a function $\rho : \mathbb{R}^3 \to \mathbb{R}^+$, such that for any $f, g \in \mathcal{H}_X$,*

$$\kappa(f, g) = \rho(\langle f, f\rangle_{\mathcal{H}_X}, \langle f, g\rangle_{\mathcal{H}_X}, \langle g, g\rangle_{\mathcal{H}_X}).$$

An example of such a kernel is $\kappa(f, g) = \exp(-\gamma\|f - g\|^2_{\mathcal{H}_X})$, where $\gamma > 0$ is a tuning constant. This is an extension of the Gaussian radial basis function (GRB) with the Euclidean norm replaced by the $\mathcal{H}_X$-norm. Since the kernel of $\mathfrak{M}_X$ is determined by the inner product of $\mathcal{H}_X$, we refer to the $\mathfrak{M}_X$ the nested RKHS via ρ (Li and Song, 2017).

2.2 Nonlinear function-on-function regression

Let $L_2(P_X)$ denote the class of all measurable functions of X such that $\mathrm{E}\,[f^2(X)] < \infty$ under P_X. Let $L_2(P_Y)$ be defined in the same way for Y. In the following, $\mathfrak{M}_X + \mathbb{R}$ represents the space $\{f + c : f \in \mathfrak{M}_X, c \in \mathbb{R}\}$.

Assumption 2 $\mathfrak{M}_X + \mathbb{R}$ *is a dense subset of* $L_2(P_X)$ *and* $\mathrm{E}\,(\|Y\|_{\mathcal{H}_Y}) < \infty$.

Assumption 2 is essentially the same as Assumption (AS) in Fukumizu et al. (2009). It can be easily shown, using Proposition 2 and Theorem 1 of Zhang et al. (2024), that the RKHS generated GRB kernel satisfies this assumption. This assumption ensures that any function $f \in L_2(P_X)$ can be approximated by a sequence of functions $\{f_n\} \subseteq \mathfrak{M}_X$ in the sense that $\mathrm{var}(f - f_n) \to 0$.

We suggest choosing ρ such that the corresponding reproducing kernel κ is universal (Micchelli et al., 2006). Examples of universal kernels include the GRB kernel and the Laplace kernel, where $\kappa(f, g) = \exp\{-\gamma\|f - g\|_{\mathcal{H}_X}\}$ for any $f, g \in \mathcal{H}_X$. Furthermore, a bounded ρ is preferable to enhance robustness. To ensure Assumption 2, the RKHS generated by κ (or equivalently by ρ) should be dense in the ambient $L_2(P_X)$ space. Both the Gaussian kernel and the Laplace kernel satisfy these requirements.

Li and Solea (2018) and Lee et al. (2023) considered the polynomial kernel in their construction of the nested Hilbert spaces, where

$$\kappa(f,g) = \rho(\langle f,g\rangle_{\mathcal{H}_X}) = (1+\langle f,g\rangle_{\mathcal{H}_X})^\gamma$$

for any $f, g \in \mathcal{H}_X$ and some positive integer γ, to generate the second-layer RKHS. In the polynomial kernel, κ is still uniquely determined by ρ. However, none of the properties mentioned above is satisfied for this kernel. Therefore, we do not recommend the polynomial kernel as our choice of κ. The additional simulation studies in Appendix A.4 further demonstrate that the performance of our method with the linear kernel is inferior to that with either the GRB kernel or the Laplace kernel.

Assumption 3 *There exists a constant $C_0 > 0$ such that for any $f \in \mathfrak{M}_X$, $\mathrm{E}\,[f^2(X)] \leq C_0\|f\|^2_{\mathfrak{M}_X}$.*

Assumption 3 ensures that the inclusion mapping $\mathfrak{M}_X \to L_2(P_X)$, $f \mapsto f$ is a bounded linear operator. It also guarantees that the bilinear form $\mathfrak{M}_X \times \mathfrak{M}_X \to \mathbb{R}$, $(f,g) \mapsto \mathrm{cov}(f(X), g(X))$ is bounded. Therefore, there exists an operator $\Sigma_{XX} \in \mathcal{B}(\mathfrak{M}_X)$ such that $\langle f, \Sigma_{XX} g\rangle_{\mathfrak{M}_X} = \mathrm{cov}(f(X), g(X))$. Similarly, under Assumptions 2 and 3, the bilinear form $\mathfrak{M}_X \times \mathcal{H}_Y \to \mathbb{R}$, $(f,g) \mapsto \mathrm{cov}(f(X), \langle g, Y\rangle_{\mathcal{H}_Y})$ is bounded, and this implies that there is an operator $\Sigma_{XY} \in \mathcal{B}(\mathcal{H}_Y, \mathfrak{M}_X)$ such that $\langle f, \Sigma_{XY} g\rangle_{\mathfrak{M}_X} = \mathrm{cov}(f(X), \langle g, Y\rangle_{\mathcal{H}_Y})$ for any $f \in \mathfrak{M}_X$ and $g \in \mathcal{H}_Y$. Moreover, Assumption 3 also implies that the linear functional $f \mapsto \mathrm{E}\,[f(X)]$ on $\mathfrak{M}_X$ is bounded. Let μ_X denote the Riesz representation of this linear functional; that is, $\langle f, \mu_X\rangle_{\mathfrak{M}_X} = \mathrm{E}\,[f(X)]$ for all $f \in \mathfrak{M}_X$. By construction, for any $x \in \mathcal{H}_X$, $\mu_X(x) = \langle \mu_X, \kappa(\cdot, x)\rangle_{\mathfrak{M}_X} = \mathrm{E}\,[\kappa(X,x)]$. Similarly, μ_Y is defined as the Riesz representation of the bounded linear functional $f \mapsto \mathrm{E}\,[\langle Y, f\rangle_{\mathcal{H}_Y}]$ where $f \in \mathcal{H}_Y$, that is, $\langle f, \mu_Y\rangle_{\mathcal{H}_Y} = \mathrm{E}\,[\langle f, Y\rangle_{\mathcal{H}_Y}]$ for $f \in \mathcal{H}_Y$. the function μ_X is called the mean element of X in $\mathfrak{M}_X$.

Using the same argument as in Fukumizu et al. (2009), it can be shown that

$$\begin{aligned}&\Sigma_{XX} = \mathrm{E}\,\{(\kappa(\cdot, X) - \mu_X) \otimes (\kappa(\cdot, X) - \mu_X)\}, \quad \Sigma_{XY} = \mathrm{E}\,\{(\kappa(\cdot, X) - \mu_X) \otimes (Y - \mu_Y)\},\\ &\Sigma_{YX} = \Sigma_{XY}^* = \mathrm{E}\,\{(Y - \mu_Y) \otimes (\kappa(\cdot, X) - \mu_X)\},\end{aligned}$$

under Assumption 3. Here, the expectations can be understood as the Bochner integrals (see, for example, Hsing and Eubank (2015), Section 2.6). Since $\overline{\mathrm{ran}}(\Sigma_{XX})^\perp = \ker(\Sigma_{XX})$ consists of functions that are constants almost surely, $\overline{\mathrm{ran}}(\Sigma_{XX})$ is the effective domain of Σ_{XX}. As shown in Li and Song (2017), this effective domain can be represented explicitly using the kernel as

$$\overline{\mathrm{span}}\{\kappa(\cdot, x) - \mu_X : x \in \mathcal{H}_X\}.$$

We will use $\mathfrak{M}_X^0$ to denote the effective domain.

Assumption 4 $\mathrm{ran}(\Sigma_{XY}) \subseteq \mathrm{ran}(\Sigma_{XX})$

The justification of Assumption 4 under linear function-on-function regression (Yao et al., 2005b) is given in Appendix A.3. Consider the restricted operator $\Sigma_{XX}|\overline{\text{ran}}(\Sigma_{XX})$. This is a mapping from $\overline{\text{ran}}(\Sigma_{XX})$ to $\text{ran}(\Sigma_{XX})$, and its inverse function exists. In fact, for any $g \in \text{ran}(\Sigma_{XX})$, there exists some $f \in \mathfrak{M}_X$ such that $\Sigma_{XX} f = g$. By Theorem 3.3.7 of Hsing and Eubank (2015) and the fact that $\overline{\text{ran}}(\Sigma_{XX}) = \mathfrak{M}_X^0$, there exists a unique decomposition $f = f_1 + f_2$ such that $f_1 \in \ker(\Sigma_{XX})$ and $f_2 \in \overline{\text{ran}}(\Sigma_{XX}) = \mathfrak{M}_X^0$. Therefore, the mapping $g \mapsto f_2$ from $\text{ran}(\Sigma_{XX})$ to $\overline{\text{ran}}(\Sigma_{XX})$ is well-defined under Assumption 3. The inverse $[\Sigma_{XX}|\text{ran}(\Sigma_{XX})]^{-1}$ is called the Moore-Penrose inverse and is written as $\Sigma_{XX}^{\dagger}$. Under Assumption 4, the following operator is well defined

$$R_{XY} = \Sigma_{XX}^{\dagger}\Sigma_{XY},$$

and we call it the regression operator. We make the following assumption on R_{XY}.

Assumption 5 *R_{XY} is a compact operator.*

Generally speaking, $\Sigma_{XX}^{\dagger}$ is not bounded since Σ_{XX} is a Hilbert-Schmidt operator (Fukumizu et al., 2009). However, as argued in Li and Song (2017), it is reasonable to assume that $\Sigma_{XX}^{\dagger}\Sigma_{XY}$ is compact, which is determined by the interaction of these two operators. Under Assumptions 4 and 5, the regression oeprator satisfies the following property: for any $\alpha \in \mathcal{H}_Y$,

$$\mathrm{E}\left[\langle Y, \alpha\rangle_{\mathcal{H}_Y} \mid X\right] = (R_{XY}\alpha)(X) - \mathrm{E}\left[(R_{XY}\alpha)(X)\right] + \mathrm{E}\left[\langle Y, \alpha\rangle_{\mathcal{H}_Y}\right]. \tag{1}$$

We next introduce our function-on-function nonlinear regression model. Our construction is motivated by an alternative view of the classical multivariate linear regression. Suppose X and Y are p- and q-dimensional vectors, and we are interested in building a linear model between them, treating Y as the response and X as the predictor. The most straightforward construction is

$$Y = c + B^{\mathsf{T}}X + \epsilon, \tag{2}$$

where c is a nonrandom vector in $\mathbb{R}^q$, B is a $p \times q$ matrix, and ϵ is a q-dimensional random vector. While being the most direct, this construction is hard to extend to our current setting in an interpretable way. An alternative way to understand the above construction is through the linear model for the scalar response $\alpha^{\mathsf{T}}Y$ for any $\alpha \in \mathbb{R}^p$:

$$\alpha^{\mathsf{T}}Y = c_\alpha + \beta_\alpha^{\mathsf{T}}X + \epsilon_\alpha. \tag{3}$$

By the linearity of the function $\alpha \mapsto \alpha^{\mathsf{T}}Y$, it is easy to see that the mappings $\alpha \mapsto c_\alpha$, $\alpha \mapsto \beta_\alpha$, and $\alpha \mapsto \epsilon_\alpha$ are linear functions of α. Hence there exists $c \in \mathbb{R}^q$, $B \in \mathbb{R}^{p\times q}$, and $\epsilon \in \mathbb{R}^q$ such that $c_\alpha = c^{\mathsf{T}}\alpha$, $\beta_\alpha = B\alpha$, and $\epsilon_\alpha = \alpha^{\mathsf{T}}\epsilon$, leading to the equation

$$\alpha^{\mathsf{T}}Y = \alpha^{\mathsf{T}}c + \alpha^{\mathsf{T}}B^{\mathsf{T}}X + \alpha^{\mathsf{T}}\epsilon.$$

Since this holds for all $\alpha \in \mathbb{R}^q$, we have model (2). In this way, we arrive at the vector-on-vector linear model (2) from the scalar-on-vector linear model (3).

Now let's come back to the problem of constructing an RKHS model for Hilbert-space-valued random elements X and Y. We assume, for each $\alpha \in \mathcal{H}_Y$, there is a function $f_\alpha \in \mathfrak{M}_X$, a constant c_α, and a random variable $U_\alpha \in \mathbb{R}$, such that

$$\langle Y, \alpha \rangle_{\mathcal{H}_Y} = c_\alpha + f_\alpha(X) + U_\alpha, \tag{4}$$

where U_α is independent of X, $\mathrm{E}\,(U_\alpha) = 0$, and $\mathrm{E}\, f_\alpha(X) = 0$. Since $\langle f_\alpha, \mu_X \rangle_{\mathfrak{M}_X} = \mathrm{E}\, f_\alpha(X)$, assuming $\mathrm{E}\, f_\alpha(X) = 0$ is equivalent to assuming $f_\alpha \perp \mu_X$ in $\mathfrak{M}_X$, or equivalently, $f_\alpha \in \mathfrak{M}_X \ominus \mathrm{span}(\mu_X)$. Model (4) is simply a nonlinear scalar-on-function regression where the regression function f_α is an element of $\mathfrak{M}_X$. This type of problem has been studied in Preda (2007).

The next theorem shows that, under some mild additional assumptions, (4) implies a joint regression model in terms of the regression operator R_{XY}.

Theorem 1 *Suppose Assumptions 2-5 hold. If model* (4) *holds for every $\alpha \in \mathcal{H}_Y$, then there exists a random element $U : \Omega \to \mathcal{H}_Y$, $U \perp\!\!\!\perp X$, such that*

$$Y = \mu_Y + R^*_{XY}[\kappa(\cdot, X) - \mu_X] + U, \tag{5}$$

where $R^*_{XY} = \Sigma_{YX}\Sigma^\dagger_{XX} \in \mathcal{B}(\mathfrak{M}^0_X, \mathcal{H}_Y)$ is the adjoint operator of R_{XY}.

For convenience, in the following, we use Y_c to represent $Y - \mu_Y$ and use $\kappa_c(\cdot, x)$ to represent $\kappa(\cdot, x) - \mu_X$. Note that, for any $f \in \mathfrak{M}_X$, we have $\langle \kappa_c(\cdot, x), f \rangle_{\mathfrak{M}_X} = f(x) - \mathrm{E}\,[f(X)]$. Theorem 1 indicates that, for a given $x \in \mathcal{H}_X$, the predicted value of Y is given by

$$\mathrm{E}\,(Y|X = x) = \Sigma_{YX}\Sigma^\dagger_{XX}\kappa_c(\cdot, x) + \mu_Y = \mathrm{E}\,[\{(\Sigma^\dagger_{XX}\kappa_c(\cdot, x))(X)\}\, Y_c] + \mu_Y. \tag{6}$$

The conditional expectation $E(Y|X = x)$ on the left is defined as follows. Consider the function $T_x : \mathcal{H}_Y \to \mathbb{R}$ defined by $\alpha \mapsto E(\langle Y, \alpha \rangle_{\mathcal{H}_Y} | X = x)$. This is obviously a linear functional on $\mathcal{H}_Y$ and, under the assumption $E(\|Y\|^2_{\mathcal{H}_X}) < \infty$, it is also a bounded linear functional. The conditional expectation $E(Y|X = x)$ is defined as the Riesz representation of the bounded linear functional T_x. That is, it is the unique member of $\mathcal{H}_Y$ such that, for any $\alpha \in \mathcal{H}_Y$, $E(\langle Y, \alpha \rangle_{\mathcal{H}_Y} | X = x) = \langle E(Y|X = x), \alpha \rangle_{\mathcal{H}_Y}$.

3 Estimation

In the last section we have described the solution to the nonlinear function-on-function regression at the population level. In this section, we implement the regression at the sample level. The key step is to construct the sample estimate of the regression operator based on sparse and noisy observations on (X, Y) by representing relevant operators as $n \times n$ matrices with a coordinate representation system. See, for example, Johnson and Horn (1985) and Li (2018).

3.1 Coordinate representation system

Suppose that $\mathcal{L}_1$ is a finite-dimensional linear space with basis $\mathcal{B} = \{\xi_1, \ldots, \xi_p\}$. Then for any $\xi \in \mathcal{L}_1$, there is a unique vector $(a_1, \ldots, a_p)^\intercal \in \mathbb{R}^p$ such that $\xi = \sum_{i=1}^p a_i \xi_i$. The vector $(a_1, \ldots, a_p)^\intercal$ is called the coordinate of ξ with respect to $\mathcal{B}$, and denoted as $[\xi]_{\mathcal{B}}$. Throughout this section we will reserve the square brackets $[\cdot]$ exclusively for coordinate representation. Next, we introduce the coordinate representation of a linear operator between two (finite-dimensional) linear spaces. Suppose $\mathcal{L}_2$ is another linear space with basis $\mathcal{C} = \{\eta_1, \ldots, \eta_q\}$ and A is a linear operator from $\mathcal{L}_1$ to $\mathcal{L}_2$. Then for any $\xi \in \mathcal{L}_1$, we have

$$A\xi = A\left(\sum_{i=1}^{p}([\xi]_{\mathcal{B}})_i \xi_i\right) = \sum_{i=1}^{p}([\xi]_{\mathcal{B}})_i (A\xi_i) = \sum_{i=1}^{p}([\xi]_{\mathcal{B}})_i \sum_{j=1}^{q}([A\xi_i]_{\mathcal{C}})_j \eta_j.$$

By the law of matrix multiplication, we can rewrite the right-hand side of the above equation as

$$\sum_{j=1}^{q}\sum_{i=1}^{p}([A\xi_i]_{\mathcal{C}})_j([\xi]_{\mathcal{B}})_i\eta_j = \sum_{j=1}^{q}\{({}_{\mathcal{C}}[A]_{\mathcal{B}})([\xi]_{\mathcal{B}})\}_j\eta_j,$$

where ${}_{\mathcal{C}}[A]_{\mathcal{B}}$ is the $q \times p$ matrix with (i,j)th entry being $([A\xi_j]_{\mathcal{C}})_i$. This equation indicates that $[A\xi]_{\mathcal{C}} = ({}_{\mathcal{C}}[A]_{\mathcal{B}})([\xi]_{\mathcal{B}})$. We therefore call the matrix ${}_{\mathcal{C}}[A]_{\mathcal{B}}$ the coordinate of the linear operator A with respect to bases $\mathcal{B}$ and $\mathcal{C}$. If we have a third linear space $\mathcal{L}_3$ with basis $\mathcal{D} = \{\zeta_1, \ldots, \zeta_l\}$ and another linear operator $B : \mathcal{L}_2 \to \mathcal{L}_3$, then it is straightforward to show that ${}_{\mathcal{D}}[BA]_{\mathcal{B}} = ({}_{\mathcal{D}}[B]_{\mathcal{C}})({}_{\mathcal{C}}[A]_{\mathcal{B}})$. When the relevant bases are clear from the context and no confusion will be caused, we will drop subscripts and write ${}_{\mathcal{C}}[A]_{\mathcal{B}}$ and $[\xi]_{\mathcal{B}}$ as $[A]$ and $[\xi]$, respectively.

3.2 Recovering trajectories of the functional predictor

While in the ideal situation each pair of the random functions (X_i, Y_i) are observed in their entirety, in practice we only observe them on a finite set of time points in $\mathbb{I}$. Without loss of generality, assume $\mathbb{I} = [0,1]$, and let $t_{i1}, \ldots, t_{im_i}$ be the points in $\mathbb{I}$ where we observe (X_i, Y_i), which may differ from subject to subject. For simplicity, we assume $m_1 = \cdots = m_n = N_X$. In addition, these discretely observed functions maybe contaminated with a random error. These considerations lead to the following model for the observed data:

$$X_{ij} = X_i(t_{ij}) + e_{ij}, \quad j = 1, \ldots, N_X,\ i = 1, \ldots, n,$$

where e_{ij} denotes the random noise. We assume the observation times are i.i.d. copies of a random variable T uniformly distributed on $\mathbb{I}$, and e_{ij}'s are i.i.d. copies of a random variable e with $E(e) = 0$ and $\mathrm{var}(e) = \sigma_e^2 < \infty$. Moreover, we assume, for each i,

$$X_i \perp\!\!\!\perp \{e_{ij} : j = 1, \ldots, N\} \perp\!\!\!\perp \{t_{ij} : j = 1, \ldots, N\}.$$

Such assumptions are common in functional data analysis; see, for example, Yao et al. (2005a), Li and Hsing (2010), and Zhang and Wang (2016).

Without loss of generality, we assume $\mathcal{H}_X$ and $\mathcal{H}_Y$ are subspaces of $L_2(\mathbb{I})$, which is the Hilbert space consisting of functions that satisfy $\int_{\mathbb{I}} f^2(t)dt < \infty$ with inner product $\int_{\mathbb{I}} f(t)g(t)dt$. We use $\|f\|$ and $\langle f, g\rangle$ to denote the norm and inner product in $L_2(\mathbb{I})$. There are two main approaches for recovering the full trajectories of X_i and Y_i based on their discrete observations. For ultra-dense observations, for example when $N_X \succeq n^{5/4}$, applying nonparametric pre-smoothing to $\{X_{ij}, j = 1, \ldots, N_X\}$ to recover X_i, as described in (Ramsay and Silverman, 2005, Chapter 4 and 5), can ensure desirable theoretical properties; see Li and Hsing (2010) and Zhang and Wang (2016). For the less dense or sparse cases, pre-smoothing would lead to unreliable results, and functional principal component analysis (FPCA) is widely used, as outlined below (see, for example, Yao et al. 2005a).

For any $t, s \in \mathbb{I}$, let $q(t) = \mathrm{E}\left[X(t)\right]$ and $C(s,t) = \mathrm{cov}[X(s), X(t)]$. By Mercer's theorem (Hsing and Eubank (2015), Theorem 4.6.5), $C(s,t)$ admits the spectral decomposition: $C(s,t) = \sum_{k=1}^{\infty} \theta_k \phi_k(s)\phi_k(t)$, where $\theta_1 > \theta_2 > \cdots > 0$ are the eigenvalues and $\{\phi_k\}_{k=1}^{\infty}$ the corresponding eigenfunctions of C. Correspondingly, X_i admits the Karhunen-Loève expansion

$$X_i(t) = q(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k, \tag{7}$$

where $\xi_{ik} = \int_{\mathbb{I}} [X_i(t) - q(t)]\phi_k(t)dt$, $k = 1, 2, \ldots$, are functional principal component (FPC) scores that satisfy $\mathrm{cov}(\xi_{ik}, \xi_{i\ell}) = \theta_k \delta_{k\ell}$, with $\delta_{k\ell}$ being the Kronecker δ-function. We assume $q(t) = 0$ without loss of generality. We follow the strategy of pooling information from all subjects proposed by Yao et al. (2005a) to carry out the FPCA. Specifically, let K be a symmetric and Lipschitz continuous density function on $[-1, 1]$ and let $h_X > 0$ the bandwidth. We estimate C by

$$\hat{C}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{N_X(N_X - 1)} \sum_{1 \le j \ne \ell \le N_X} \frac{1}{h_X^2} K\left(\frac{t_{ij} - s}{h}\right) K\left(\frac{t_{i\ell} - t}{h_X}\right) X_{ij}X_{i\ell},$$

Note that, for $j \ne \ell$, $E(X_{ij}X_{i\ell} \mid t_{ij}, t_{i\ell}) = C(t_{ij}, t_{i\ell})$, but for $j = \ell$, $E(X_{ij}X_{ij} \mid t_{ij}) \ne C(t_{ij}, t_{ij})$. So, intuitively, $\hat{C}(s,t)$ is close to $C(s,t)$ when $s \ne t$, but needs not be close to $C(s,t)$ when $s = t$. Nevertheless, what is important for us is to estimate the eigenfunctions of $C(s,t)$, and this is not affected by perturbations in the diagonal elements of $C(s,t)$. Solving equations $\int_{\mathbb{I}} \widehat{C}(s,t)\phi_k(t)dt = \theta_k\phi_k(s)$ for $k = 1, \ldots, m_X$ numerically leads to estimates of θ_k and $\phi_k, k = 1, \ldots, m_X$, which are denoted by $\hat{\theta}_k$ and $\hat{\phi}_k$, respectively.

We employ the above kernel estimate of $C(s,t)$ in the following sample-splitting procedure to produce an estimate of the entire function X_i for every $t \in \mathbb{I}$. The sample-splitting is designed to remove dependence. As shown in Zhou et al. (2023), sample splitting both simplifies theoretical analysis and enhances estimation accuracy. Let $\tau_1 = \{(i,j) : i = 1, \ldots, n/2, j = 1, \ldots, N_X\}$ and $\tau_2 = \{(i,j) : i = n/2+1, \ldots, n, j = 1, \ldots, N_X\}$ be the disjoint index sets for sample splitting. For $r = 1, 2$, let $\widehat{C}_{(r)}(s,t)$ and $\hat{\phi}_{(r),k}$ be the estimated covariance function and its eigenfunctions based on the observations in τ_r. For each $k = 1, \ldots, m_X$, let

$$\hat{\xi}_{ik} = \begin{cases} \frac{1}{N_X}\sum_{j=1}^{N_X} X_{ij}\hat{\phi}_{(2),k}(t_{ij}) & i = 1, \ldots, n/2, \\ \frac{1}{N_X}\sum_{j=1}^{N_X} X_{ij}\hat{\phi}_{(1),k}(t_{ij}) & i = n/2+1, \ldots, n. \end{cases}$$

Mimicking (7), the trajectory of X_i is estimated by

$$\hat{X}_i(t) = \begin{cases} \sum_{k=1}^{m_X} \hat{\xi}_{ik}\hat{\phi}_{(2),k}(t) & i = 1, \ldots, n/2, \\ \sum_{k=1}^{m_X} \hat{\xi}_{ik}\hat{\phi}_{(1),k}(t) & i = n/2+1, \ldots, n. \end{cases} \tag{8}$$

The truncation integer m_X plays a critical role as often seen in the tuning parameters in nonparametric estimation: a large m_X would help to reduce the bias but tends to increase the variance of the estimate (Hall and Hosseini-Nasab, 2006).

We assume the following regularity conditions on X to ensure desirable estimation accuracy of $\hat{X}_i(t)$ defined in (8). Let $\xi_k = \int_{\mathbb{I}} X(t)\phi_k(t)dt$, and let $\phi_k^{(j)}$ denote the j-th derivative of ϕ_k.

Assumption C1 *X has finite fourth moment, i.e., $\int_{\mathbb{I}} E\{X^4(t)\}dt < \infty$, the function $t \mapsto \mathrm{E}\{X^2(t)\}$ is continuous, and $E(\xi_k^4) \preceq \theta_k^2$ for any $k \in \mathbb{N}$.*

Assumption C2 *The eigenvalues of C satisfy $D^{-1}k^{-a-1} \le \theta_{k+1} < \theta_k \le Dk^{-a}$ for any $k \in \mathbb{N}$ and some $D > 1$ and $a > 1$.*

Assumption C3 *The eigenfunctions of C satisfy $\sup_{k\ge 1}\sup_{t\in[0,1]}|\phi_k(t)| < \infty$ and*

$$\sup_{t\in[0,1]}|\phi_k^{(j)}(t)| \preceq k^{c/2}\sup_{t\in[0,1]}|\phi_k^{(j-1)}(t)|, \quad \text{for } j = 1, 2 \text{ and any } k \in \mathbb{N},$$

where $c > 0$ is a constant. Additionally, $\phi_k(0) = \phi_k(1)$ and $\phi_k^{(1)}(0) = \phi_k^{(1)}(1)$ for any $k \in \mathbb{N}$.

Assumptions C1 to C3 are essentially the same as Assumptions A.1 - A.3 in Zhou et al. (2025). In fact, Assumptions C1 and C2 are commonly adopted in the literature of functional data analysis; see Yao et al. (2005a) and Hall and Horowitz (2007), for example. The following theorem characterizes the discrepancy between $\hat{X}_i$ and X_i.

Theorem 2 *Suppose Assumptions C1 to C3 are satisfied. If, as $n \to \infty$,*

$$\textbf{(M1)}\quad \frac{m_X^{a-1}}{N_X} \to 0,\ m_X \to \infty,\ \frac{m_X^{2a+2}}{n} \to 0,\ \frac{m_X^{2a+2}}{nN_X^2h_X^2} \to 0, h_X^4 m_X^{2a+2} \to 0 \ \text{ and } \ h_X^4 m_X^{2a+2c} = O(1),$$

then

$$\|\hat{X}_i - X_i\|^2 = O_P\left(m_X^{-(a-1)} + \frac{m_X^{2a+3}}{nN_X^2} + \frac{m_X^{a+1}}{nN_Xh_X} + \frac{m_X^{2a+1}}{nN_X^2h_X} + h_X^4 m_X^{2c+3}\right) \tag{9}$$

for $i = 1, \ldots, n$.

We can make a meaningful characterization of the observation schedule of the functional data X_i from "sparse" to "dense" according to the estimation error in recovering trajectories from noisy observations.

Corollary 3 *Suppose Assumptions C1 to C3 are satisfied, and $m_X \in \mathbb{N}$ satisfies (M1). Let the optimal bandwidth be chosen as $h_X = (nN_X)^{-1/5}m_X^{(a-2c-2)/5}$. We have*

1. *if $m_X^a \preceq N_X$,*

$$\|\hat{X}_i - X_i\|^2 = O_P\left(m_X^{-(a-1)} + \frac{m_X^{(4a+2c+7)/5}}{(nN_X)^{4/5}}\right).$$

   Additionally, if $nm_X^{(9a+2c+2)/4} \preceq N_X$, then $\|\hat{X}_i - X_i\|^2 = O_P(m_X^{-(a-1)})$.

2. *if $N_X = o(m_X^a)$,*

$$\|\hat{X}_i - X_i\|^2 = O_P\left(m_X^{-(a-1)} + \frac{m_X^{2a+3}}{nN_X^2} + \frac{m_X^{(9a+2c+7)/5}}{(nN_X^2)^{4/5}}\right).$$

To facilitate the following analysis, we focus on the dense regime: $m_X^a \preceq N_X$. Next, we study coordinate representation for sample-level quantities in $\mathfrak{M}_X$, which is an RKHS generated by the positive definite kernel κ on $\mathcal{H}_X$. By Assumption 1, for any $u, v \in \mathcal{H}_X$,

$$\kappa(u,v) = \rho\left(\langle u,u\rangle_{\mathcal{H}_X}, \langle u,v\rangle_{\mathcal{H}_X}, \langle v,v\rangle_{\mathcal{H}_X}\right).$$

In the following implementation, κ is taken as the GRB function. At the sample level, we consider a subspace of $\mathfrak{M}_X$ spanned by $\{\kappa(\cdot,\hat{X}_i) : i = 1,\ldots,n\}$ with inner product

$$\langle f,g\rangle_{\mathfrak{M}_X} = [f]^\intercal K_{\hat{X}}[g],$$

for any $f, g \in \mathfrak{M}_X$, where $K_{\hat{X}}$ is the $n \times n$ Gram matrix whose (i,j)th entry is $\kappa(\hat{X}_i,\hat{X}_j)$.

The justification for switching from $\mathfrak{M}_X$l, defined as the infinite dimensional RKHS spanned by $\{\kappa(\cdot,x) : x \in \mathcal{H}_X\}$, to $\mathfrak{M}_X$, defined as the finite-dimensional RKHS spanned by $\{\kappa(\cdot,\hat{X}_i) : i = 1,\ldots,n\}$, is the so called representer theorem for RKHS (Scholkopf et al., 2001), which implies that, for the optimization problem in our context, the solution always resides in the latter RKHS. Thus, it is equivalent to consider the finite dimensional RKHS.

3.3 Recovering response trajectories

We employ the same strategy in the last subsection to recover each Y_i from its noisy observations. Specifically, consider the following model for the noisy observations on Y_i:

$$Y_{i\ell} = Y_i(s_{ij}) + \varepsilon_{i\ell}, \quad l = 1,\ldots,N_Y, i = 1,\ldots,n,$$

where $\{s_{i\ell} : l = 1,\ldots,N_Y\}$ are the observation times of Y_i, and $\varepsilon_{i\ell}$ is the random noise. We assume $\{s_{i\ell} : l = 1,\ldots,N_Y, i = 1,\ldots,n\}$ are i.i.d copies of S uniformly distributed on $[0,1]$, and $\varepsilon_{i\ell}$'s are i.i.d copies of ε with mean 0 and variance $\sigma_\varepsilon^2 < \infty$. Moreover, we assume that Y, S and ε are mutually independent.

Following the notations introduced in Section 2.2, we have $\mu_Y(t) = E\{Y(t)\}$ and $\Sigma_{YY}(s,t) = E[\{Y(s) - \mu_Y(s)\}\{Y(t) - \mu_Y(t)\}]$. For $i = 1,\ldots,n$, let $\hat{Y}_i(t)$ denote the estimated trajectory via the FPCA. Under certain regularity conditions, which are essentially the same as Assumptions C1 to C3, we establish the following result for the estimation error of $\hat{Y}_i$. Let m_Y and h_Y denote the truncation level in recovering each Y_i and the bandwidth used in the FPCA on the noisy observations of Y, respectively.

Corollary 4 *Suppose Assumptions C4 to C6 (in the supplementary material) are satisfied, and $m_Y \in \mathbb{N}$ satisfies (M1) with N_X and h_X replaced by N_Y and h_Y, respectively. If $h_Y = (nN_Y)^{-1/5} m_Y^{(a-2c-2)/5}$ and $m_Y^a \preceq N_Y$, we have*

$$\|\hat{Y}_i - Y_i\|^2 = O_P\left(m_Y^{-(a-1)} + \frac{m_Y^{(4a+2c+7)/5}}{(nN_Y)^{4/5}}\right).$$

3.4 Model fitting

As indicated in (6), to estimate $\mathrm{E}\,(Y|X = x)$, we need to estimate the regression operator R_{XY}. Having constructed the subspace of $\mathfrak{M}_X$, we define $\hat{\mu}_X$ as the Riesz representation of the linear functional $f \mapsto \mathrm{E}_n[f(X)]$, if full trajectories of X_i's are available. By Proposition

2 of Li and Song (2017), $\hat{\mu}_X$ is the function $n^{-1}\sum_{i=1}^n \kappa(\cdot, X_i)$ in $\mathfrak{M}_X$. Since the trajectory of X_i is not fully observed, we define a more realistic estimator $\hat{\mu}_{\hat{X}}$ as $n^{-1}\sum_{i=1}^n \kappa(\cdot, \hat{X}_i)$ in $\mathfrak{M}_X$, and consider a subspace of $\mathfrak{M}_X^0$ spanned by $\{\kappa(\cdot, \hat{X}_i) - \mu_{\hat{X}} : i = 1, \ldots, n\}$. To reflect the fact that the estimates discussed below are constructed from the recovered trajectories, we use $\hat{\Sigma}_{\hat{X}\hat{X}}$ and $\hat{\Sigma}_{\hat{X}\hat{Y}}$ to denote the estimates of Σ_{XX} and Σ_{XY}, i.e., $\hat{\Sigma}_{\hat{X}\hat{X}} = 1/n\sum_{i=1}^n \{\kappa(\cdot, \hat{X}_i) - \hat{\mu}_{\hat{X}}\} \otimes \{\kappa(\cdot, \hat{X}_i) - \hat{\mu}_{\hat{X}}\}$ and $\hat{\Sigma}_{\hat{X}\hat{Y}} = 1/n\sum_{i=1}^n \{\kappa(\cdot, \hat{X}_i) - \hat{\mu}_{\hat{X}}\} \otimes (\hat{Y}_i - \hat{\mu}_{\hat{Y}})\}$.

We estimate $\mathrm{E}\,(Y|X = x)$ by mimicking (6) at the sample level. To do so, we next derive the coordinates of relevant operators therein. Here, we omit the associated bases from the notation of coordinate representation as they are obvious from the context. Let $Q = I_n - 1_n 1_n^\intercal/n$, where 1_n denotes the column vector of length n with each component being 1. Let $G_{\hat{X}} = QK_{\hat{X}}Q$. Then, by Proposition 3 of Li and Song (2017),

$$[\hat{\Sigma}_{\hat{X}\hat{X}}] = n^{-1}G_{\hat{X}}, \qquad [\hat{\Sigma}_{\hat{Y}\hat{X}}] = n^{-1}G_{\hat{X}}, \qquad [\hat{\Sigma}^\dagger_{XX}] = nG^\dagger_{\hat{X}},$$

with respect to the spanning system $\{\kappa(\cdot, \hat{X}_i) - \mu_{\hat{X}} : i = 1, \ldots, n\}$. Let $h_{\hat{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\intercal$. Then, given $x \in \mathcal{H}_X$,

$$\begin{aligned}\hat{R}^*_{\hat{X}\hat{Y}}\{\kappa(\cdot, x) - \hat{\mu}_{\hat{X}}\} &= h_{\hat{Y}}^\intercal Q\{[\hat{R}^*_{\hat{X}\hat{Y}}][\kappa(\cdot, x) - \hat{\mu}_{\hat{X}}]\} \\ &= h_{\hat{Y}}^\intercal Q\left\{[\hat{\Sigma}_{\hat{Y}\hat{X}}][\hat{\Sigma}^\dagger_{\hat{X}\hat{X}}][\kappa(\cdot, x) - \hat{\mu}_{\hat{X}}]\right\} \\ &= h_{\hat{Y}}^\intercal\{G_{\hat{X}}G^\dagger_{\hat{X}}[\kappa(\cdot, x) - \hat{\mu}_{\hat{X}}]\},\end{aligned}$$

where the last equality holds because $G_{\hat{X}} = QK_{\hat{X}}Q$ and $Q^2 = Q$. We estimate the Moore-Penrose inverse of $G_{\hat{X}}$ by the Tikhonov-regularized inverse $(G_{\hat{X}} + \epsilon_X I_n)^{-1}$ to prevent overfitting, where $\epsilon_X > 0$ is a tuning constant. It remains to figure out the coordinate of $\kappa(\cdot, x) - \hat{\mu}_{\hat{X}}$ with respect to the spanning system $\{\kappa(\cdot, \hat{X}_i) - \hat{\mu}_{\hat{X}} : i = 1, \ldots, n\}$. Suppose that x is observed at time points $t_1, \ldots, t_{N_X}$, and $x_{t_j} = x(t_j) + \epsilon(t_j)$ for $j = 1, \ldots, N_X$. We employ the FPCA in last subsection to recover the trajectory of x. That is, $x(t)$ is estimated by

$$\hat{x}(t) = \sum_{k=1}^{m_X} \hat{\xi}_k^x \hat{\phi}_k(t), \tag{10}$$

where $\hat{\phi}_k(t) = \{\hat{\phi}_{(1),k}(t) + \hat{\phi}_{(2),k}(t)\}/2$ and $\hat{\xi}_k^x = \sum_{j=1}^{N_X} \hat{\phi}_k(t_j)x_{t_j}/N_X$.

Having recovered the trajectory of x, we next identify the coordinate of $\kappa(\cdot, \hat{x}) - \hat{\mu}_{\hat{X}}$, which will be used as the surrogate of that of $\kappa(\cdot, x) - \mu_X$. Suppose that $[\kappa(\cdot, \hat{x}) - \hat{\mu}_{\hat{X}}] = c_{\hat{x}}$ for some $c_{\hat{x}} \in \mathbb{R}^n$. Then

$$\langle \kappa(\cdot, \hat{x}) - \hat{\mu}_{\hat{X}}, \kappa(\cdot, \hat{X}_i)\rangle_{\mathfrak{M}_X} = e_i^\intercal K_{\hat{X}} c_{\hat{x}} \; - \frac{1}{n}(e_i^\intercal K_{\hat{X}} 1_n)(1_n^\intercal c_{\hat{x}}) = e_i^\intercal K_{\hat{X}} Q c_{\hat{x}},$$

where e_i denotes the vector whose ith component is 1 and all others are 0. Taking $i = 1, \ldots, n$, we have $d_{\hat{x}} = K_{\hat{X}}Qc_{\hat{x}}$, where $d_{\hat{x}}$ is a vector of length n with ith component $\kappa(\hat{X}_i, \hat{x}) - \mathrm{E}_n\kappa(\hat{X}_i, \hat{X})$. With the Tikhonov regularization, we obtain the solution $c_{\hat{x}} = Q(K_{\hat{X}} + \epsilon_X I_n)^{-1}d_{\hat{x}}$. Lastly, by (6), the predicted value of y is

$$\hat{y}(\hat{x}) = h_Y^\intercal G_{\hat{X}}(G_{\hat{X}} + \epsilon_X I_n)^{-1}c_{\hat{x}} + \frac{1}{n}h_Y^\intercal 1_n. \tag{11}$$

3.5 Tuning parameter selection

This section is concerned with tuning parameters. If we use the GRB as the kernel to construct $\mathfrak{M}_X$, then we have tuning parameters: m_X, m_Y, and (ϵ_X, γ_X) for $\mathcal{H}_X$, $\mathcal{H}_Y$, and $\mathfrak{M}_X$, respectively.

As suggested in Yao et al. (2005a), we can use either the leave-one-curve-out cross validation or the information-based criteria such as AIC or BIC to select m_X and m_Y. Moreover, Li et al. (2013) found that m_X (or m_Y) selected through AIC or BIC is consistent estimator of m if the covariance operator only has m has nonzero eigenfunctions. To estimate R_{XY}, we choose the tuning parameters (ϵ_X, γ_X) by generalized cross-validation (GCV). By (11), the fitted value of Y_i at $\hat{X}_i$ is

$$\hat{Y}(\hat{X}_i) = [Q_i^{\mathsf{T}} G_{\hat{X}}(G_{\hat{X}} + \epsilon_X I_n)^{-1} + 1_n^{\mathsf{T}}/n] h_{\hat{Y}},$$

where $Q_i = Qe_i$ is the ith column of the projection matrix Q. Therefore, the GCV score in this case is defined as

$$\mathrm{GCV}(\epsilon_X, \gamma_X) = \frac{1}{n} \sum_{i=1}^{n} \frac{\|\hat{Y}_i - \hat{Y}(\hat{X}_i)\|^2_{\mathcal{H}_Y}}{\{1 - \mathrm{trace}[QG_{\hat{X}}(G_{\hat{X}} + \epsilon_X I_n)^{-1} + 1_n 1_n^{\mathsf{T}}/n]/n\}^2},$$

where $\hat{Y}_i$ and $\hat{Y}(\hat{X}_i)$ denote the recovered trajectory of Y_i from noisy observations and the predicted trajectory from model (5), respectively. The optimal (ϵ_X, γ_X) is chosen by minimizing GCV over a grid of (ϵ_X, γ_X).

4 Convergence rates for fully observed data

In this section we develop the convergence rates of our nonparametric regression in the setting where the full trajectories of each X_i and Y_i are available for $i = 1, \ldots, n$. In particular, we are interested in the the following two rates:

(1) the convergence rate of the estimated regression operator $\hat{R}_{XY}$;

(2) the convergence rate of the regression estimate $\widehat{\mathrm{E}}\,(Y|x_0)$ at a new predictor and at any time point t, where x_0 is a fully observed sample path of X.

We will also derive the optimal tuning parameter $\epsilon_n = \epsilon_X$ that makes these rates the fastest. The theories developed in this section lay the foundation for establishing convergence rate for the estimator defined in Section 3 for the more realistic case, where only sparse and noisy observations are available for each X_i and Y_i for $i = 1, \ldots, n$.

4.1 Some preliminary lemmas

Let $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{XY}$ be the plug-in estimates of Σ_{XX} and Σ_{XY}, i.e., $\hat{\Sigma}_{XX} = 1/n \sum_{i=1}^{n} \{\kappa(\cdot, X_i) - \hat{\mu}_X\} \otimes \{\kappa(\cdot, X_i) - \hat{\mu}_X\}$ and $\hat{\Sigma}_{XY} = 1/n \sum_{i=1}^{n} \{\kappa(\cdot, X_i) - \hat{\mu}_X\} \otimes (Y_i - \hat{\mu}_Y)\}$ where $\hat{\mu}_X = 1/n \sum_{i=1}^{n} \kappa(\cdot, X_i)$ and $\hat{\mu}_Y = 1/n \sum_{i=1}^{n} Y_i$. We make the following assumption.

Assumption 6

(i) $\mathrm{E}\,[\kappa(X, X)] < \infty$, *and* $\mathrm{E}\,(\|Y\|^2_{\mathcal{H}_Y}) < \infty$;

(ii) there is a $\beta > 0$ such that $\Sigma_{XY} = \Sigma_{XX}^{1+\beta} S_{XY}$ for some bounded linear operator $S_{XY} : \mathcal{H}_Y \to \mathfrak{M}_X$.

It can be shown that, under the assumption $\mathrm{E}\left[\kappa(X,X)\right] < \infty$, Σ_{XX} is a trace-class operator. As argued in Li and Song (2017) and Li (2018), Assumption 6(ii) represents a degree of smoothness in the relation between X and Y. It requires the output functions of R_{XY} to be sufficiently concentrated on the low-frequency components of Σ_{XX}. Indeed, if $\{(\lambda_j, \varphi_j) : j \in \mathbb{N}\}$ is the eigenvalue-eigenfunction sequence of Σ_{XX} with $\lambda_1 \geq \lambda_2 \geq \cdots$, then $\Sigma_{XY} = \Sigma_{XX}^{1+\beta} S_{XY}$ implies that, for any $g \in \mathcal{H}_Y$, $\sum_{j\in\mathbb{N}} \lambda_j^{-2\beta} \langle R_{XY} g, \varphi_j \rangle_{\mathfrak{M}_X}^2 < \infty$. Thus, the Fourier coefficients of the output of R_{XY} with respect to the orthonormal basis $\{\varphi_i\}$ has to decay sufficiently faster that the decaying rate of the eigenvalues λ_i of the covariance operator Σ_{XX}. The following lemma gives the convergence rates of $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{XY}$, whose proof is similar to Lemma 5 of Fukumizu et al. (2007) and is omitted.

Lemma 5 *Under Assumption 6(i), Σ_{XX} and Σ_{XY} are Hilbert-Schmidt operators and*

$$\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{\mathrm{HS}} = O_P(n^{-1/2}), \quad \|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_{\mathrm{HS}} = O_P(n^{-1/2}).$$

Let $\hat{R}_{XY} = (\hat{\Sigma}_{XX} + \epsilon_n I)^{-1} \hat{\Sigma}_{XY}$ denote another estimator of R_{XY}, where we have used ϵ_n to replace ϵ_X to highlight the dependence on the sample size n. Under Assumption 6, the best convergence rate of $\hat{R}_{XY}$ to R_{XY} developed by Li and Song (2017) is $n^{-\beta/[2(\beta+1)]}$. If $\beta = 1$, this rate reaches its fastest possible level $n^{-1/4}$. In the next subsection we will show that, in our regression setting and with an additional assumption on Σ_{XX}, the convergence rate of $\hat{R}_{XY}$ can approach $n^{-1/3}$.

Based on model (5), let $U = Y - \mathrm{E}(Y|X)$ be the population-level residual, which is a random element in $\mathcal{H}_Y$. Let $\Sigma_{XU} = \mathrm{E}\left[(\kappa(\cdot, X) - \mu_X) \otimes U\right]$. Let $\hat{\mu}_U$ and $\hat{\Sigma}_{XU}$ be the sample estimates of μ_U and Σ_{XU} defined by

$$\hat{\mu}_U = \mathrm{E}_n(U), \quad \hat{\Sigma}_{XU} = \mathrm{E}_n[(\kappa(\cdot, X) - \hat{\mu}_X) \otimes (U - \hat{\mu}_U)].$$

Lemma 6 *Under Assumption 6(i),*

1. $\Sigma_{XU} = 0$*;*

2. $\hat{\Sigma}_{XY} = \hat{\Sigma}_{XU} + \hat{\Sigma}_{XX} R_{XY}$*.*

Let $\tilde{\Sigma}_{XU} = E_n[(\kappa(\cdot, X) - \mu_X) \otimes U]$, which is an intermediate operator between $\hat{\Sigma}_{XU}$ and Σ_{XU}.

Lemma 7 *Under Assumption 6(i), we have $\|\hat{\Sigma}_{XU} - \tilde{\Sigma}_{XU}\|_{\mathrm{HS}} = O_P(n^{-1})$.*

Since Σ_{XX} is a trace-class operator under Assumption 6(i), we have $\sum_{j\in\mathbb{N}} \lambda_j < \infty$. The next assumption strengthens this condition.

Assumption 7 *$\lambda_j \asymp j^{-\alpha}$ for some $\alpha > 1$ and all $j \in \mathbb{N}$.*

By Theorem 1, the functional covariate X is independent of the population-level residual U in model (5). Assumption 7 is about the niceness of the random function X: its variation is concentrated on the low-frequency domain of the spectrum of the covariance operator Σ_{XX}. The next lemma reveals how Assumption 7 interacts with Tychonoff regularization.

Lemma 8 *Under Assumption 7, if $\epsilon_n \prec 1$, then $\sum_{j\in\mathbb{N}} \lambda_j (\lambda_j + \epsilon_n)^{-2} = O(\epsilon_n^{-(\alpha+1)/\alpha})$.*

4.2 Convergence rate for estimated regression operator

For convenience, we abbreviate $(\hat{\Sigma}_{XX}+\epsilon_n I)^{-1}$, $(\Sigma_{XX}+\epsilon_n I)^{-1}$, and $\Sigma_{XX}^{\dagger}$ by $\hat{V}$, V_n and V, respectively. The following Fourier expansion of $\kappa(\cdot, X)-\mu_X$ with respect to the eigenfunction orthonormal basis (ONB) $\{\varphi_j : j \in \mathbb{N}\}$ will be useful:

$$\kappa(\cdot, X) - \mu_X = \textstyle\sum_{j\in\mathbb{N}} \langle \kappa(\cdot, X) - \mu_X, \varphi_j \rangle_{\mathfrak{M}_X} \varphi_j \equiv \sum_{j\in\mathbb{N}} \zeta_j \varphi_j, \tag{12}$$

where $\zeta_1, \zeta_2, \ldots$ are uncorrelated variables with $\mathrm{E}(\zeta_j) = 0$ and $\mathrm{var}(\zeta_j) = \lambda_j$.

Theorem 9 *Suppose Assumptions 2 through 5 hold; Assumption 6 holds for some $\beta > 0$; Assumption 7 holds for some $\alpha > 1$; $\epsilon_n \prec 1$. Then*

1. $$\|\hat{R}_{XY} - R_{XY}\|_{\mathrm{OP}} = O_P(n^{-1/2}\epsilon_n^{(\beta\wedge 1)-1} + \epsilon_n^{\beta\wedge 1} + n^{-1}\epsilon_n^{-(3\alpha+1)/(2\alpha)} + n^{-1/2}\epsilon_n^{-(\alpha+1)/(2\alpha)}). \tag{13}$$
2. *If $\epsilon_n \succ \max(n^{-1/[2\{1-(\beta\wedge 1)\}]}, n^{-2\alpha/(3\alpha+1)})$, then the right-hand side of (13) tends to 0.*

4.3 Optimal turning and convergence

Next, we derive the optimal convergence rate of (13) where ϵ_n is of the form $\epsilon_n \asymp n^{-\delta}$ for some $\delta > 0$. With ϵ_n in this form, the four terms in (13) reduce to

$$n^{-1/2+\delta\{1-(\beta\wedge 1)\}}, \quad n^{-\delta(\beta\wedge 1)}, \quad n^{-1+\delta(3\alpha+1)/(2\alpha)}, \quad n^{-1/2+\delta(\alpha+1)/(2\alpha)}.$$

Let $\ell_1, \ldots, \ell_4$ be the linear functions of δ in the exponents; that is,

$$\ell_1(\delta) = -1/2 + \delta\{1-(\beta\wedge 1)\}, \quad \ell_2(\delta) = -\delta(\beta\wedge 1), \quad \ell_3(\delta) = -1 + \delta(3\alpha+1)/(2\alpha),$$
$$\ell_4(\delta) = -1/2 + \delta(\alpha+1)/(2\alpha).$$

Let $m(\delta) = \max\{\ell_1(\delta), \ldots, \ell_4(\delta)\}$. Then the rate in (13) can be rewritten as $n^{m(\delta)}$. Letting δ_{opt} be the δ that minimizes $m(\delta)$, the optimal tuning parameter is $\epsilon_n = n^{-\delta_{\mathrm{opt}}}$, and the corresponding convergence rate is $n^{m(\delta_{\mathrm{opt}})} \equiv \rho_{\mathrm{opt}}$.

Theorem 10 *Suppose the conditions in Theorem 9 hold for some $\alpha > 1$, $\beta > 0$.*

(1) if $\beta > (\alpha-1)/(2\alpha)$, then $\delta_{\mathrm{opt}} = \frac{\alpha}{2\alpha(\beta\wedge 1)+\alpha+1}$, $\rho_{\mathrm{opt}} = n^{-\frac{\alpha(\beta\wedge 1)}{2\alpha(\beta\wedge 1)+\alpha+1}}$.

(2) if $\beta \leq (\alpha-1)/(2\alpha)$, then $\delta_{\mathrm{opt}} = \frac{1}{2}$, $\rho_{\mathrm{opt}} = n^{-\frac{\beta}{2}}$.

The best rate for the regression operator reported in Li and Song (2017) is

$$\rho_{\mathrm{LS}} = n^{-(\beta\wedge 1)/[2\{1+(\beta\wedge 1)\}]}.$$

It is easy to check that ρ_{opt} converges to 0 faster than ρ_{LS} in both scenarios of β; that is,

$$n^{-(\alpha\beta\wedge\alpha)/(2\alpha(\beta\wedge 1)+\alpha+1)} \prec n^{-(\beta\wedge 1)/[2\{1+(\beta\wedge 1)\}]}, \quad n^{-\beta/2} \prec n^{-\beta/[2(1+\beta)]}$$

for all $\beta > 0$ and $\alpha > 1$. The reason for this improvement is twofold: first, we are dealing with the more specific regression problem (5), whereas Li and Song (2017) dealt with a general problem where the regression operator corresponds directly to a conditional distribution, without any regression structure; second, we have made Assumption 7, which was not made in Li and Song (2017). Note that, when $\beta = 1$, Li and Song's rate is $n^{-1/4}$, whereas our current rate is always faster than $n^{-1/4}$ regardless of the value of α, and approaches $n^{-1/3}$ when α is sufficiently large.

4.4 Convergence rate for regression estimate

In this section we develop the convergence rate of our nonparametric regression estimate $\widehat{\mathrm{E}}\,(Y|x_0)$ to the true mean response $\mathrm{E}\,(Y|x_0)$ at any given time point $t \in \mathbb{I}$. We will use $\mathrm{E}\,(Y|x_0)(t)$ to denote the function $\mathrm{E}\,(Y|X = x_0)$, which is a member of $\mathcal{H}_Y$, evaluated at time t; the same applies to $\widehat{\mathrm{E}}\,(Y|x_0)(t)$. Assuming $\mathcal{H}_Y$ is an RKHS with kernel υ, the conditional mean $\mathrm{E}\,(Y|x_0)(t)$ can be written as $\langle \upsilon(\cdot, t), \mathrm{E}\,(Y|x_0)\rangle_{\mathcal{H}_Y}$. Since $\mathrm{E}\,(Y|x_0) = R^*_{XY}(\kappa(\cdot, x_0) - \mu_X) + \mu_Y$, we have

$$\mathrm{E}\,(Y|x_0)(t) = \langle R_{XY}\,\upsilon(\cdot, t), \kappa(\cdot, x_0) - \mu_X\rangle_{\mathfrak{M}_X} + \mu_Y(t). \tag{14}$$

The estimate of the above is

$$\widehat{\mathrm{E}}\,(Y|x_0)(t) = \langle \hat{R}_{XY}\,\upsilon(\cdot, t), \kappa(\cdot, x_0) - \hat{\mu}_X\rangle_{\mathfrak{M}_X} + \hat{\mu}_Y(t). \tag{15}$$

The next corollary shows that $\widehat{\mathrm{E}}\,(Y|x_0)(t)$ has the same convergence rate as $\hat{R}_{XY}$.

Corollary 11 *Suppose*

(1) the conditions in Theorem 9 hold for some $\alpha > 1$, $\beta > 0$,

(2) $\max(n^{-1/[2\{1-(\beta\wedge 1)\}]}, n^{-2\alpha/(3\alpha+1)}) \prec \epsilon_n \prec 1$;

(3) $\mathcal{H}_Y$ is an RKHS generated by a kernel υ.

Then $\widehat{\mathrm{E}}\,(Y|x_0)(t)$ is consistent with convergence rate

$$\widehat{\mathrm{E}}\,(Y|x_0)(t) - \mathrm{E}\,(Y|x_0)(t) = O_P(n^{-1/2}\epsilon_n^{(\beta\wedge 1)-1} + \epsilon_n^{\beta\wedge 1} + n^{-1}\epsilon_n^{-(3\alpha+1)/(2\alpha)} + n^{-1/2}\epsilon_n^{-(\alpha+1)/(2\alpha)}). \tag{16}$$

Furthermore, the conclusions of Theorem 10 also hold.

Remark 12 *If we only want to study the convergence rate of $\widehat{\mathrm{E}}\,(Y|x_0)$ alone, we do not need to assume that $\mathcal{H}_Y$ is an RKHS. By Equation (6) and the preceding results, the condition that $\mathcal{H}_Y$ is a Hilbert space suffices to develop the convergence rate of $\widehat{\mathrm{E}}\,(Y|x_0)$, which is the same as $\hat{R}_{XY}$. However, to study the convergence rate of $\widehat{\mathrm{E}}\,(Y|x_0)(t)$, we need the assumption that $\mathcal{H}_Y$ is an RKHS, since it ensures the evaluation functional $f \mapsto f(t)$ is continuous for any $f \in \mathcal{H}_Y$ and $t \in \mathbb{I}$. Developing the convergence rate and further a limiting distribution in next section for the pointwise prediction is crucial to construct pointwise confidence intervals for $\mathrm{E}\,(Y|x_0)$.*

5 Central limit theorem

In this section we develop limiting distributions for the predicted response given a new functional covariance in the setting where the full trajectories of each X_i and Y_i are available for $i = 1, \ldots, n$.

5.1 Pointwise central limit theorem

We first develop the central limit theorem of the regression estimate $\widehat{\mathrm{E}}\,(Y|x_0)(t)$; the result is useful for constructing the confidence interval for the mean response $\mathrm{E}\,(Y|x_0)(t)$. We will only consider the case $\beta > (\alpha-1)/(2\alpha)$ and $\delta > \alpha/(2\alpha\beta+\alpha+1)$, which means the relation between Y and X is relatively smooth and ϵ_n is chosen so that the bias term is of a smaller order than the dominating term. More specifically, recall that

$$\begin{aligned}\hat{R}_{XY} - R_{XY} &= \hat{R}_{\mathrm{res}} + (\hat{R}_{\mathrm{reg}} - R_n) + (R_n - R_{XY}), \quad \text{where}\\ \hat{R}_{\mathrm{res}} &= (\hat{V}\hat{\Sigma}_{XU} - \hat{V}\tilde{\Sigma}_{XU}) + (\hat{V}\tilde{\Sigma}_{XU} - V_n\tilde{\Sigma}_{XU}) + V_n\tilde{\Sigma}_{XU}.\end{aligned}$$

Let $B_{n,4}$ and $B_{n,5}$ be the last two terms of the first equation, and $B_{n,1}, B_{n,2}, B_{n,3}$ be the three terms of $\hat{R}_{\mathrm{res}}$ in the second equation. Let

$$A_{n,r} = \langle B_{n,r}\upsilon(\cdot,t), \kappa(\cdot,x_0) - \mu_X\rangle_{\mathfrak{M}_X}, \quad r = 1,\ldots,5.$$

Note that $A_{n,5}$ is a nonrandom number. By Theorem 9, when $\beta > (\alpha-1)/(2\alpha)$ and $\alpha/(2\alpha\beta+\alpha+1) < \delta < 1/2$, $B_{n,3}$ is the dominating term among all the other terms. Hence it is reasonable to expect that $A_{n,3}$ is also the dominating term. Our central limit theorem is based on this assumption.

Assumption 8 *$A_{n,1},\ldots,A_{n,4}$ have finite variances $\sigma^2_{n,1},\ldots,\sigma^2_{n,4}$ and*

$$\sigma_{n,3} \succ \max(\sigma_{n,1}, \sigma_{n,2}, \sigma_{n,4}, |A_{n,5}|).$$

Theorem 13 *Suppose the conditions in Theorem 9 are satisfied for some $\alpha > 1$ and $\beta > (\alpha-1)/(2\alpha)$, and Assumption 8 is satisfied. Furthermore, suppose that the kernel κ is bounded. Then the following statements hold true:*

(1) $\sigma^2_{n,3} = n^{-1}\mathrm{E}\,[U^2(t)]\sum_{j\in\mathbb{N}}(\lambda_j+\epsilon_n)^{-2}\lambda_j[\varphi_j(x_0)]^2$;

(2) if $\sigma_{n,3} \succ \epsilon_n^{\beta\wedge 1}$ and $\epsilon_n \succ n^{-1/2}$, then for any fixed $x_0 \in \mathcal{H}_X$ and $t \in \mathbb{I}$,

$$\sigma^{-1}_{n,3}[\widehat{\mathrm{E}}\,(Y|x_0)(t) - \mathrm{E}\,(Y|x_0)(t)] \xrightarrow{\mathcal{D}} N(0,1).$$

To use this theorem to construct confidence intervals, we need to have an estimate of $\sigma^2_{n,3}$. As will be discussed later, we can substitute the estimates of λ_j, φ_j and $\mathrm{E}\,[U(t)^2]$ to estimate $\sigma^2_{n,3}$ for constructing the confidence interval.

5.2 Uniform central limit theorem

Following the idea of Cardot et al. (2007), we now study the weak convergence of the regression estimate as a random function in the Hilbert space $\mathcal{H}_Y$. With a slight abuse of notation, we denote the Riesz representation of T_x defined in Section 2 by $M(x)$ given $X = x$ in $\mathcal{H}_X$, which is actually $\mathrm{E}\,(Y|X=x) \in \mathcal{H}_Y$. Let $\widehat{M}(x)$ denote the predicted value in $\mathcal{H}_Y$ for a new value x obtained by means of the estimation method introduced in Section 3. We are interested in the following problem. Given a new random element $X_{n+1} \in \mathcal{H}_X$ that is a copy of X and independent of $X_1, \ldots, X_n$, we aim to investigate the weak convergence of $a_n[\widehat{M}(X_{n+1}) - \mathrm{E}\,(Y_{n+1}|X_{n+1})]$ in $\mathcal{H}_Y$ for some normalizing constant a_n. The following lemma illustrates the stochastic order of the crucial term in establishing weak convergence of $\widehat{M}(X_{n+1})$.

Lemma 14 *Suppose the conditions in Theorem 9 are satisfied for some $\alpha > 1$, $\beta \geq 1$, and $n^{-1/2} \prec \epsilon_n \prec 1$. We further assume that S_{XY} in Assumption 6(ii) is a Hilbert-Schmdit operator. Let $W_n = \sum_{i=1}^n Z_{i,n}$, where $Z_{i,n} = \frac{1}{n} U_i \langle \hat{V} G_i, G_{n+1} \rangle_{\mathfrak{M}_X}$ and $G_i = \kappa(\cdot, X_i) - \mu_X$. Then the following statements hold true:*

1. $\widehat{M}(X_{n+1}) - \mathrm{E}\,(Y_{n+1}|X_{n+1}) = W_n + O_P\,(n^{-1/2}\epsilon_n^{(\alpha-1)/(2\alpha)} + \epsilon_n + n^{-1}\epsilon_n^{-(\alpha+1)/(2\alpha)})$;
2. $\mathrm{E}\,(\|W_n\|^2_{\mathcal{H}_Y}) = O(n^{-1}\epsilon_n^{-1})$.

Remark 15 *We impose the condition $\beta \geq 1$ to facilitate the analysis of the stochastic order of $\hat{\Sigma}^{\beta}_{XX} - \Sigma^{\beta}_{XX}$. Without this assumption, even though we can still prove that it is $o_p(1)$, determining its convergence rate is quite complicated.*

By (A33) (in the supplementary material) in the proof of Lemma 14 part *1*, after ignoring $O_P(n^{-1/2})$ terms, we have

$$\begin{aligned}
&\widehat{M}(X_{n+1}) - \mathrm{E}\,(Y_{n+1}|X_{n+1}) \\
&= (\hat{\Sigma}_{YX}\hat{V} - \Sigma_{YX}\Sigma^{\dagger}_{XX})\{\kappa(\cdot, X_{n+1}) - \mu_X\} \\
&= R^*_{XY}(\hat{\Sigma}_{XX}\hat{V} - \Sigma_{XX}V_n)G_{n+1} + R^*_{XY}(\Sigma_{XX}V_n - I)G_{n+1} + \frac{1}{n}\sum_{i=1}^n U_i \langle \hat{V}\tilde{G}_i, G_{n+1}\rangle_{\mathfrak{M}_X} \\
&= F_{1n} + F_{2n} + W_n + H_n,
\end{aligned}$$

where $H_n = \bar{U}\langle \hat{V}(\mu_X - \hat{\mu}_X), G_{n+1}\rangle_{\mathfrak{M}_X} = O_P(n^{-1}\epsilon_n^{-(\alpha+1)/(2\alpha)})$. It is straightforward to check that $n^{-1/2}\epsilon_n^{-1/2} \succ \max(n^{-1/2}\epsilon_n^{(\alpha-1)/(2\alpha)}, \epsilon_n, n^{-1}\epsilon_n^{-(\alpha+1)/(2\alpha)})$ if $n^{-1/2} \prec \epsilon_n \prec n^{-1/3}$ for $\alpha > 1$. Thus W_n is the dominating term among all the other terms when $n^{-1/2} \prec \epsilon_n \prec n^{-1/3}$. Let $s_n^2 = n^{-1}\mathrm{E}\,[\langle \hat{V}G_i, G_{n+1}\rangle^2_{\mathfrak{M}_X}]$. The weak convergence of $\widehat{M}(X_{n+1})$ in $\mathcal{H}_Y$ is based on the assumption that W_n is the dominating term.

Assumption 9

$$s_n^2 \succ \max(\mathrm{var}(F_{1n}), \mathrm{var}(F_{2n}), \mathrm{var}(H_n)).$$

Theorem 16 *Suppose assumptions in Lemma 14 are met and Assumption 9 is satisfied. Then*

$$s_n^{-1}[\widehat{M}(X_{n+1}) - \mathrm{E}\,(Y_{n+1}|X_{n+1})] \xrightarrow{d} \mathcal{N}, \tag{17}$$

where $\mathcal{N}$ is a centered Gaussian element taking values in $\mathcal{H}_Y$ with covariance operator Σ_{UU}.

Remark 17 *By* (17) *and the continuous mapping theorem, we have*

$$\sup_{t\in\mathbb{I}}\left| s_n^{-1}\left[\{\widehat{M}(X_{n+1})\}(t) - \mathrm{E}\,(Y_{n+1}|X_{n+1})(t)\right]\right| \xrightarrow{\mathcal{D}} \sup_{t\in T}|\mathcal{N}(t)|$$

as $n \to \infty$. Therefore, if we are able to find $C(\alpha)$ that satisfies $\Pr(\sup_{t\in T} |\mathcal{N}(t)| \leq C(\alpha)) = 1-\alpha$, then a $(1-\alpha)$ simultaneous confidence band for $\mathrm{E}\,(Y_{n+1}|X_{n+1})$ can be constructed as

$$\Big(\{\widehat{M}(X_{n+1})\}(t) - s_n C(\alpha), \{\widehat{M}(X_{n+1})\}(t) + s_n C(\alpha)\Big). \tag{18}$$

The determination of $C(\alpha)$ is illustrated in one of our simulation studies near the end of Section 7.2.

6 Convergence rates for partially observed data

In this section we develop the convergence rates of our nonparametric regression in the setting considered in Section 3, where only discretized and noisy observations on each X_i and Y_i are available for $i = 1, \ldots, n$. Similar to Section 4, we are interested in the following two rates:

(1) the convergence rate of the estimated regression operator $\hat{R}_{\hat{X}\hat{Y}}$;

(2) the convergence rate of the regression estimate $\widehat{\mathrm{E}}\,(Y|x_0)$ at a new predictor and at any time point $t \in \mathbb{I}$, where x_0 is not fully observed.

In Section 4 we have established consistency and convergence rate for the estimator in the ideal case. To address the two problems above, it suffices to find the discrepancy between the ideal estimator and the more realistic estimator defined in Section 3 for discretely observed functional data, which are assumed to be contaminated with measurement error.

Assumption 10 *The reproducing kernel κ of $\mathfrak{M}_X$ is Lipschitz continuous; namely, there exists a constant L such that for any $f, g \in \mathcal{H}_X$, one has*

$$\|\kappa(f,\cdot) - \kappa(g,\cdot)\|_{\mathfrak{M}_X} \leq L\|f-g\|.$$

Remark 18 *Note that Assumption 10 is met for the GRB kernel. In fact, if κ is the GRB kernel, $\|\kappa(f,\cdot) - \kappa(g,\cdot)\|^2_{\mathfrak{M}_X} = 2 - 2\kappa(f,g)$. By the Taylor expansion on $\kappa(f,g)$, we can easily show that $\|\kappa(f,\cdot) - \kappa(g,\cdot)\|^2_{\mathfrak{M}_X} \leq 2\gamma\|f-g\|^2$. Hence Assumption 10 is met if we take $L = (2\gamma)^{1/2}$. This assumption also explains why we want to achieve a desirable convergence rate in the L_2-norm when recovering the trajectory of each X_i from its sparse and noisy observations.*

For ease of notation, we assume $m_X \asymp m_Y \asymp m$ and $N_X \asymp N_Y \asymp N$. The following lemma establishes the estimation errors of $\hat{\Sigma}_{\hat{X}\hat{X}}$ and $\hat{\Sigma}_{\hat{X}\hat{Y}}$, respectively.

Lemma 19 *We assume that the assumptions in Corollaries 3 and 4 hold. Under Assumptions 6(i) and 10, we have*

$$\|\hat{\Sigma}_{\hat{X}\hat{X}} - \Sigma_{XX}\|_{\mathrm{HS}} = O_P\left(m^{-(a-1)} + \frac{m^{(4a+2c+7)/5}}{(nN)^{4/5}}\right),$$
$$\|\hat{\Sigma}_{\hat{X}\hat{Y}} - \Sigma_{XY}\|_{\mathrm{HS}} = O_P\left(m^{-(a-1)} + \frac{m^{(4a+2c+7)/5}}{(nN)^{4/5}}\right).$$

Recall in Section 3 the estimator of the regression operator R_{XY} is defined as $\hat{R}_{\hat{X}\hat{Y}} = (\hat{\Sigma}_{\hat{X}\hat{X}} + \epsilon_n I)^{-1}\hat{\Sigma}_{\hat{X}\hat{Y}}$. Based on Lemma 19, we establish the convergence rate of $\hat{R}_{\hat{X}\hat{Y}}$, as stated in the following theorem.

Theorem 20 *Suppose assumptions in Theorem 9 and Lemma 19 hold. If $n^{-1/2} \prec \epsilon_n \prec 1$, then*

(1.)

$$\|\hat{R}_{\hat{X}\hat{Y}} - R_{XY}\|_{\mathrm{OP}} = O_P(a_n + \epsilon_n^{-2} b_n^2 + \epsilon_n^{-1} b_n), \tag{19}$$

where $a_n = \epsilon_n^{\beta\wedge 1} + n^{-1/2}\epsilon_n^{-(\alpha+1)/(2\alpha)}$ and $b_n = m^{-(a-1)} + m^{(4a+2c+7)/5}(nN)^{-4/5}$.

(2.) *If $\epsilon_n \succ \max(b_n, (b_n n^{-1})^{2\alpha/(5\alpha+1)})$, then the right-hand side of (19) becomes $O_P(a_n + \epsilon_n^{-1} b_n)$, which tends to 0 as $n \to \infty$.*

Lastly, we consider the convergence rate of our nonparametric regression estimate $\widehat{\mathrm{E}}\,(Y|x_0)$ to the true mean response $\mathrm{E}\,(Y|x_0)$ at any given time point $s \in \mathbb{I}$. Instead of observing the true trajectory of x_0, we assume that we only make noisy observations on x_0 at $t_1, \ldots, t_{N_X}$, which are independent copies of T. A similar set-up for the prediction analysis was considered in Yao et al. (2005b).

Let $\hat{x}_0$ denote the recovered trajectory of x_0 from its noisy observations using (10). Then based on (14), we have

$$\hat{\mathrm{E}}\,(Y|\hat{x}_0)(s) = \langle \hat{R}_{\hat{X}\hat{Y}}\, \upsilon(\cdot, s), \kappa(\cdot, \hat{x}_0) - \hat{\mu}_{\hat{X}} \rangle_{\mathfrak{M}_X} + \hat{\mu}_{\hat{Y}}(s), \tag{20}$$

under the assumption that $\mathcal{H}_Y$ is an RKHS generated by a kernel υ. As with Corollary 11 for fully observed data, the following corollary states that $\widehat{\mathrm{E}}\,(Y|\hat{x}_0)(s)$ has the same convergence rate as $\hat{R}_{\hat{X}\hat{Y}}$.

Corollary 21 *Suppose*

(1) the conditions in Theorem 20 hold;

(2) $\max(b_n, (b_n n^{-1})^{2\alpha/(5\alpha+1)}) \prec \epsilon_n \prec 1$;

(3) $\mathcal{H}_Y$ is an RKHS generated by a kernel υ.

Then $\widehat{\mathrm{E}}\,(Y|\hat{x}_0)(s)$ is consistent with convergence rate

$$\widehat{\mathrm{E}}\,(Y|\hat{x}_0)(s) - \mathrm{E}\,(Y|x_0)(s) = O_P(a_n + \epsilon_n^{-1} b_n), \tag{21}$$

where $a_n = \epsilon_n^{\beta\wedge 1} + n^{-1/2}\epsilon_n^{-(\alpha+1)/(2\alpha)}$ and $b_n = m^{-(a-1)} + m^{(4a+2c+7)/5}(nN)^{-4/5}$.

Based on Corollaries 3, 4 and 21, if

$$\|\hat{X}_i - X_i\|^2 = O_p(n^{-1/2}\epsilon_n^{(\beta\wedge 1)} + n^{-1}\epsilon_n^{-\frac{\alpha+1}{2\alpha}}), \tag{22}$$

the convergence rate of the prediction error based on partially observed data is identical to that from the fully observed trajectories. Furthermore, by Corollary 3, if $nm^{(9a+2c+2)/4} \preceq N$, (22) is equivalent to $m^{-(a-1)} \preceq n^{-1/2}\epsilon_n^{(\beta\wedge 1)} + n^{-1}\epsilon_n^{-\frac{\alpha+1}{2\alpha}}$.

If functional data are ultra densely observed, say $N \succeq n^{5/4}$, we may obtain recovered trajectories through local linear smoothing and they satisfy

$$\|\hat{X}_i - X_i\|^2 \lesssim n^{-1}, \quad i = 1, \ldots, n,$$

by Theorem 2 of Zhang and Chen (2007). Then (22) is satisfied as $\epsilon_n = o(1)$. But this requires the sample path of the functional data has a continuous second-order derivative almost surely; other relevant regularity conditions can be found from Condition A of Zhang and Chen (2007). In FDA, random functions are usually assumed to be mean-square continuous rather than have differentiable sample paths almost surely. The standard Brownian motion is a prominent example whose sample paths are mean-square continuous but are nowhere differentiable almost surely (Hsing and Eubank, 2015, Chapter 7). On the other hand, it is worth noting that the proposed method can accommodate sparsely observed functional data, where N has an upper fixed bound or has a finite expectation. Applying a local linear smoother to sparse observations to recover trajectories is inappropriate. The numerical studies in Section 7.3 demonstrate the strong performance of our method in the sparse setting.

7 Simulation studies

In this section, we investigate the performance of the proposed methodology in prediction under different simulation scenarios. We compare our nonlinear function-on-function regression (to be abbreviated by NLFFR) method with several alternative methods: optimal penalized linear function-on-function regression (to be abbreviated by PLFFR) proposed by Sun et al. (2018), linear function-on-function regression estimated via functional principal component analysis (to be abbreviated by FPCA) proposed by Yao et al. (2005b) and Crambes and Mas (2013), and nonlinear function-on-functional regression via functional universal approximation (to be abbreviated FUA) developed by Luo and Qi (2024). In addition, we evaluate the finite-sample performances of both the pointwise confidence interval and the simultaneous confidence band developed in Section 5.

7.1 Dense design

We adopt a similar strategy in Li and Song (2017) to generate functional covariates. Specifically, we construct $\mathcal{H}_X$ as the RKHS induced by two kernels: the GRB and the Brownian motion covariance function (BMC). When the GRB kernel is employed, the functional covariate X is generated by $X(t) = \sum_{k=1}^{5} a_k \exp\{-\gamma_T (t - t_k)^2\}$, where $a_1, \ldots, a_5$ are independently sampled from $N(0,1)$, $t_1, \ldots, t_5$ are independently sampled from $U[0,1]$ and $\gamma_T = 7$. When the BMC kernel is employed, X is generated as

$$X(t) = \sum_{j=1}^{100} \sqrt{2}[(j - 1/2)\pi]^{-1} a_j \sin[(j - 1/2)\pi t],$$

where a_j's are independently sampled from $N(0,1)$. In the dense design, we choose 50 equally spaced points in $[0,1]$ as the observed time points of X for each subject. The left panel of Figure 1 depicts 10 sample paths of X generated by these two kernels in the dense design.

Two models are then used to generate the functional response Y:

$$\begin{aligned} \text{Model} \quad 1: \quad & Y(t) = \left(\frac{1}{1 + e^{\langle X, b_1 \rangle_{\mathcal{H}_X}}} + \langle X, b_2 \rangle^2_{\mathcal{H}_X} \right) \rho(t) + \sigma\epsilon(t), \\ \text{Model} \quad 2: \quad & Y(t) = \left\{ \cos\left(\langle X, b_3 \rangle_{\mathcal{H}_X} \right) \right\} \rho(t) + \sigma\epsilon(t). \end{aligned}$$

In both models, when the GRB kernel is used, $b_j(t) = \exp\{-\gamma_T(t - t_j)^2\}$ with $t_1 = 0.6,\ t_2 = 0.9$ and $t_3 = 0.1$; when the BMC kernel is used, for $j = 1, 3$, $b_j(t) = \nu_j(t) := \sqrt{2}a_j \sin[(j - 1/2)\pi t]$, which is actually the jth eigenfunction of the covariance operator of the standard Brownian motion, and $b_2(t) = 0$. Regardless of the choice of kernel, $\rho(t) = \sum_{j=1}^{5} \nu_j(t)$ and $\epsilon(t)$ is generated from the standard Brownian motion. The choices of ρ and ϵ ensure that the true conditional mean $E(Y|X)$ resides in the RKHS generated by the BMC kernel. The right panel of Figure 1 shows the shape of $\rho(t)$, which indicates that the (true) conditional mean has a relative large fluctuation around 0.18. We consider two different values of σ: 0.1 and 2, to deliver different signal-to-noise ratios.

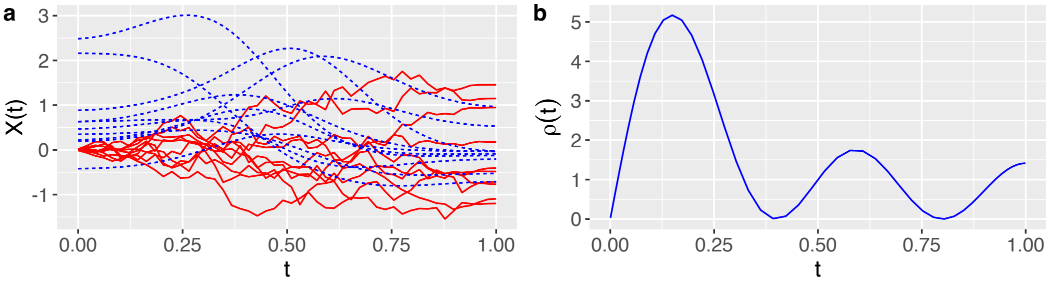

Figure 1: (a): trajectories of the functional covariate X generated by the GRB kernel (blue dotted lines) and the BMC kernel (red solid lines). (b): function $\rho(t)$ in models 1 and 2.

In each simulation scenario, we randomly generate 100 pairs of (X_i, Y_i)'s as the training set and 500 pairs as the test set. For comparing the predictive performance of these three methods, the prediction error is defined as the median of the integrated squared errors

$$\text{ISE} = \int_0^1 \{\hat{Y}(t) - \mathrm{E}\left(Y(t) \mid X\right)\}^2 dt \tag{23}$$

calculated on the test set. The inner products in $\mathcal{H}_X$ and $\mathcal{H}_Y$ are always calculated using the same kernel: either GRB or BMC, and GRB is always chosen as the reproducing kernel in $\mathfrak{M}_X$. To better assess the performance of our proposed method in comparison with other methods, each simulation scenario is repeated 200 times.

7.2 Results for dense design

In the dense design, X_i and Y_i are observed at 50 equally spaced time points in [0, 1]. Table 1 summarizes the medians and the inter quartiles of the prediction errors for each method across the 200 simulation runs. The FUA method dominates the other methods in all simulation settings, but its advantage over our method becomes less evident for lower signal-to-noise levels. Meanwhile, it should be noted that the FUA method is considerably more computationally intensive than its competitors because of the tedious tuning parameter selection procedure. Our method has much better prediction accuracy than its linear competitors regardless of the signal-to-noise level. Moreover, even when we used the wrong kernel, for instance when X is generated by the BMC kernel but we use the GRB kernel to calculate the inner product in both $\mathcal{H}_X$ and $\mathcal{H}_Y$, our method still achieves satisfactory prediction accuracy. This demonstrates the robustness of our method against the choice of the kernel when computing the inner product in $\mathcal{H}_X$ and $\mathcal{H}_Y$.

Model	X	σ	Methods				
			FPCA	PLFFR	FUA	NLFFR (GRB)	NLFFR (BMC)
1	GRB	0.1	13.09 (4.76)	6.79 (1.55)	9.53e-4 (7.80e-4)	3.46 (4.98)	1.73 (1.28)
		2	13.20 (4.78)	6.83 (1.54)	0.13 (0.08)	3.49 (4.92)	1.74 (1.31)
	BMC	0.1	0.48 (0.06)	1.25 (0.11)	1.56e-3 (7.86e-4)	0.56 (0.08)	0.43 (0.06)
		2	0.66 (0.12)	1.39 (0.15)	0.23 (0.10)	0.58 (0.09)	0.45 (0.09)
2	GRB	0.1	1.31 (0.10)	3.01 (0.23)	6.59e-4 (4.77e-4)	0.19 (0.08)	0.19 (0.12)
		2	1.35 (0.15)	3.02 (0.25)	0.15 (0.10)	0.25 (0.09)	0.25 (0.12)
	BMC	0.1	1.45 (0.14)	2.43 (0.21)	1.38e-3 (1.50e-3)	0.47 (0.33)	0.45 (0.30)
		2	1.57 (0.18)	2.51 (0.23)	0.21 (0.16)	0.51 (0.37)	0.51 (0.34)

Table 1: Summary of the medians and the interquartile ranges (in parentheses) of the prediction errors across the 200 simulation runs under different simulation scenarios for each method in the dense design. The column of X indicates which kernel is used to generate X in model 1 or 2, and the columns of NLFFR (GRB) and NLFFR (BMC) indicate which kernel is used to calculate the inner product in $\mathcal{H}_X$ and $\mathcal{H}_Y$ when using the proposed NLFFR.

Method	Model			
	1		2	
	X: GRB	X: BMC	X: GRB	X: BMC
GRB	0.896	0.976	0.924	0.946
BMC	0.898	0.972	0.916	0.962

Table 2: Summary of the mean of the coverage probability of simultaneous confidence bands over the 200 simulation runs with $\sigma = 0.1$. The first column indicates the kernel used to calculate the inner product in $\mathcal{H}_X$ in model fitting, and the row with $X : (\cdots)$ indicates the kernel used to generate X in either model 1 or 2.

We next construct the pointwise confidence intervals described in Theorem 13. In particular, we randomly selected one subject from model 2 with $\sigma = 0.1$, and constructed a confidence interval for $\mathrm{E}\,(Y|x_0)(t)$ at any $t \in [0, 1]$. Figure 2 displays the pointwise 95% confidence intervals. Regardless of the choice of the kernel used to generate X or calculate the inner product in $\mathcal{H}_X$ in model fitting, the intervals cover the true conditional mean reasonably well. In particular, the estimated conditional mean shows a relatively large fluctuation around 0.18 due to the shape of ρ shown in the right panel Figure 1. After around $t = 0.25$, the magnitude of ρ becomes relatively smaller; it implies smaller variability of the true conditional mean at $t > 0.25$. Consequently, the pointwise confidence intervals become considerably narrower in this region, which is consistent with the shape of ρ.

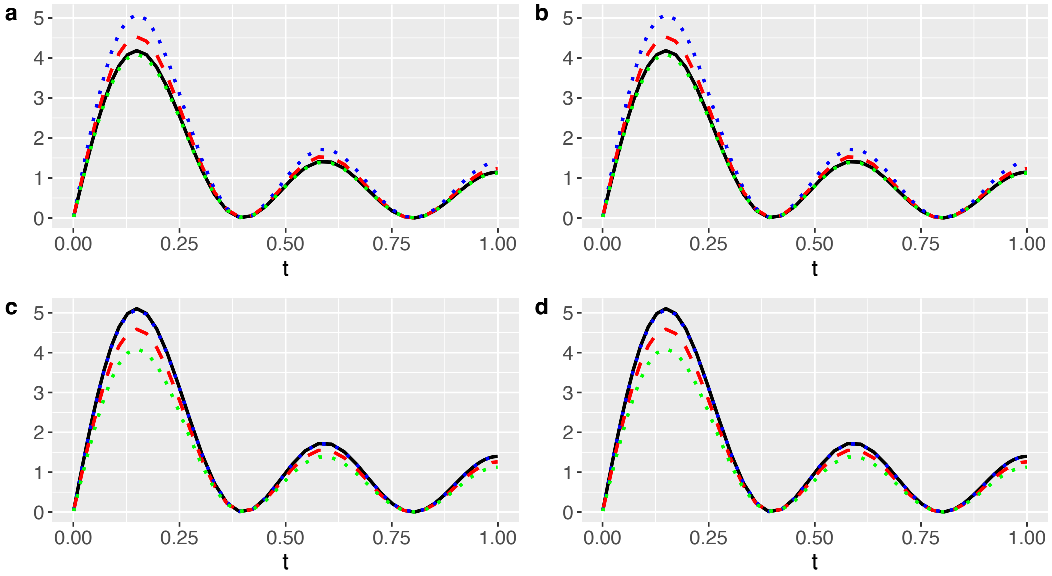

Figure 2: Pointwise confidence intervals for one randomly sampled subject in the test set from model 2. (a) & (b): X is generated by the GRB kernel, while in model fitting $\mathcal{H}_X$ is constructed via the the GRB and the BMC kernel, respectively. (c) & (d): X is generated by the BMC kernel, while in model fitting the inner product in $\mathcal{H}_X$ is calculate via the the GRB and the BMC kernel, respectively. In each panel, the solid black line represents the true conditional mean function $\mathrm{E}\left(Y|x_0\right)(t)$, the red dashed line represents the estimated mean, and the blue and green dotted lines represent the upper and the lower bounds of 95% pointwise confidence intervals, respectively.

We further study the simultaneous confidence band of $\mathrm{E}\left(Y_{n+1}|X_{n+1}\right)$ given by (18). Estimation of s_n in (18) is straightforward. To determine the value of $C(\alpha)$, we first calculate $\hat{U}_i$'s on the training set based on the observed Y_i and the estimated mean. Then a plugged-in estimate of Σ_{UU} is available. We generated a large number of sample paths of a centered Gaussian process with the estimated Σ_{UU} as the covariance function. Let $Z_i(t), i = 1, \ldots, N$ denote the randomly generated sample paths. For each of them, $\sup_{t\in\mathbb{I}} |Z_i(t)|$ is approximated by evaluating $|Z_i(t)|$ on a dense grid of $\mathbb{I}$ and then taking the maximum. The value of $C(\alpha)$ is taken as the α-upper empirical quantile of $\sup_{t\in\mathbb{I}} |Z_i(t)|$'s. Table 2 presents the average of the true coverage probabilities of the 95% simultaneous confidence bands across the 200 simulation runs with $\sigma = 0.1$. The true coverage probabilities for both models 1 and 2 are close to the nominal level (95%) in most cases. Note that the coverage probability for the design when X is generated from the GRB kernel in model 1 is slightly lower than the nominal level. This result is consistent with as is shown in Table 1: compared with other designs, the prediction accuracy of the proposed method is slightly worse in this design.

$N_X(N_Y)$	σ	Methods			
		FPCA	PLFFR	FUA	NLFFR
$\{4,5,6\}$	0.1	102.66 (30.13)	253.51 (1028.32)	13.58 (6.60)	17.91 (16.68)
	2	103.56 (34.55)	313.19 (1258.95)	14.56 (7.87)	17.96 (16.64)
$\{16,18,20\}$	0.1	99.14 (28.07)	201.10 (762.72)	5.69 (2.03)	8.15 (7.86)
	2	99.25 (27.70)	210.76 (773.12)	5.74 (2.30)	8.30 (10.05)
$\{5,10,20\}$	0.1	104.73 (29.82)	230.51 (872.70)	6.95 (2.52)	9.00 (8.04)
	2	104.66 (32.22)	289.88 (858.86)	7.11 (2.83)	9.64 (13.97)

Table 3: Summary of the medians and the interquartile ranges (in parentheses) of the ISE on the test set across the 200 simulation runs under different simulation scenarios for each method in the sparse design.

7.3 Sparse design and results

To study the performance of our method in a sparse design, we generate X as $X(t) = t + \sin(t) + \xi_1\sqrt{2}\cos(\pi t) + \xi_2\sqrt{2}\cos(2\pi t) + e(t)$ for $t \in [0,1]$, where ξ_1 and ξ_2 are independently generated from $N(0,4)$ and $N(0,1)$, respectively, and $e(t) \sim N(0, 0.25)$ denotes the measurement error of X. For each subject, the number of observations of X, N_X, is randomly sampled from one of the three sets, $\{4,5,6\}$, $\{16,18,20\}$ and $\{5,10,20\}$, to generate sparse functional data with different sampling frequencies, and then we randomly generate N_X observation times from the uniform distribution on [0, 1]. To reflect the nonlinear relation between X and Y, we combine the idea of functional quadratic regression (Yao and Müller, 2010) and the polynomial kernel function in RKHS regression. In particular, the response in generated as

$$Y(s) = \left\{\int_0^1 X^c(t)\beta(t)dt + \int_0^1\int_0^1 X^c(t)X^c(s)\gamma(s,t)dsdt\right\}\rho_1(s) + \varepsilon(s),$$

where $X^c(t) = X(t) - (t + \sin(t))$ denotes centered X, $\beta(t) = \sqrt{2}\cos(\pi t) + \sqrt{2}\cos(2\pi t)$, $\gamma(s,t) = 2\cos(\pi t)\cos(\pi s) + \cos(2\pi t)\cos(\pi s) + 2\cos(2\pi t)\cos(2\pi s)$. The projection direction $\rho_1(s) = \sum_{k=1}^5 \sqrt{2}\cos(k\pi s)$. For each subject, we make N_Y, which is randomly sampled from the same set as N_X, observations at uniformly distributed times on [0, 1], and $\varepsilon(s) \sim N(0, \sigma^2)$ denotes the measurement error of Y. Next we explain why the above model satisfies our definition of nonlinear function-on-function regression by RKHS. Note that for a fixed $\alpha \in L_2[0,1]$, one can consider a nonlinear scalar-on-function regression between $\langle Y, \alpha\rangle$ and X:

$$f(X) = \left\{\int_0^1 X^c(t)\beta(t)dt + \int_0^1\int_0^1 X^c(t)X^c(s)\gamma(s,t)dsdt\right\}\langle\rho_1, \alpha\rangle,$$

which is a functional analogue of RKHS regression with a polynomial kernel of degree two for scalar or vector-valued covariates.

As with the dense design, we conduct 200 independent simulation runs, and in each simulation run, we randomly generate 100 pairs of (X, Y) as the training set and 500 pairs as the test set. For our method, the PLFFR method and the FUA method, we employ the principal component analysis through conditional expectation (PACE) method proposed by Yao et al. (2005a) to recover each sparse trajectory first, as described in Section 3.2. In contrast, the FPCA-based method proposed by Yao et al. (2005b) only entails estimating FPCs of X and Y as well as their scores through PACE. Interested readers can refer to Yao et al. (2005b) for more details. Prediction errors on the test set by each method are summarized in Table 3. In comparison with the dense case, the FUA method outperforms the other methods, and our proposed method displays a similar advantage over the two linear competitors in terms of prediction accuracy. However, the advantage of the FUA method over ours is less evident compared with the dense case.

8 Data application

In this section, we apply our proposed method and the aforementioned competitors to a data application. We are not only interested in predication accuracy of our method in real applications, but also the pointwise confidence intervals and the simultaneous confidence band introduced in Section 5.

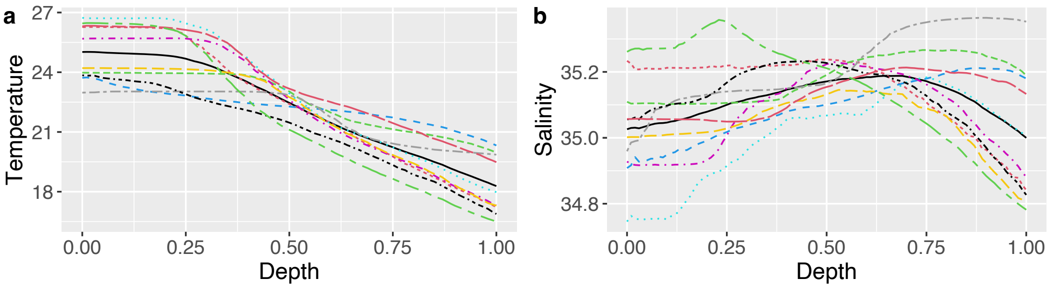

Figure 3: (a): ten sample curves of Temperature and the sample mean curve of them (black solid line). (b): ten sample curves of Salinity and the sample mean curve of them (black solid line).

As indicated by `http://hahana.soest.hawaii.edu/hot/hot-dogs/cextraction.html`, the Hawaii Ocean Time-series (HOT) program has been collecting time course observations on the hydrography, chemistry and biology of the water column at a station north of Oahu, Hawaii since October 1988. One goal of this program is to learn about concentrations of some materials in the upper water column (0 - 200 m below the sea surface). With the aid of CTD sampling support, profiles of temperature, salinity, oxygen and potential density as a function of pressure (or equivalently depth) are available. In our study, we took a portion of the whole data set. The data set has five variables: Temperature, Salinity, Potential Density, Oxygen and Chloropigment and, in a single day, each of them has 101 measurements, one per two meters from 0 to 200 meters. They can be treated as a function of depth, and trajectories collected from different days are viewed as different sample curves. There are 116 sample curves in total for each variable.

In this study, we are interested in using the trajectories of Temperature to predict those of Salinity. As indicated by Good et al. (2013), Temperature is strongly associated with Salinity and there exists a nonlinear relationship between them. This assertion can be further justified by Figure 3, which shows the trajectories of Temperature and Salinity of 10 randomly selected samples, where the depth was rescaled to [0, 1] from [0, 200]. The trends of these two groups of mean curves suggest that Temperature decreases as the depth increases, whereas Salinity goes up first and then drops down as depth increases. Additionally, the response variable, Salinity, displays more variability near the boundary than in the interior region.

To evaluate prediction accuracy of each method, we randomly and evenly split the whole data set into a training set and a test set. Each method was fitted to the training set and then the fitted function-on-function regression was used to predict the response in the test set. It is more reasonable to assume that the observations of Temperature and Salinity are contaminated with measurement error. We take this into consideration when fitting each model. This process was repeated $M = 200$ times to assess variability in predictions. The medians and the interquartile ranges of the prediction errors across the 200 splits are shown in Table 4. Our proposed method performs similarly to the FUA method and greatly outperforms the two linear competitors. The poor performances of the FPCA and PLFFR methods indicate that the relationship between Salinity and Temperature cannot be adequately fitted by a linear function-one-function regression model.

	FPCA	PLFFR	FUA	NLFFR
median	1.24×10^3	1.24×10^3	1.16×10^{-2}	1.21×10^{-2}
IQR	(0.51)	(38.78)	(2.13×10^{-3})	(1.55×10^{-3})

Table 4: Summary of the averages and standard errors (in parentheses) of the prediction errors across the 200 random splits.

We also constructed the pointwise confidence intervals defined by Theorem 13 and the simultaneous confidence band by (18) for this regression problem, by assuming that full trajectories of Temperature and Salinity without any measurement error are available. Figure 4 shows both the 95% pointwise confidence intervals and the 95% simultaneous confidence band constructed by the GRB and the BMC kernels for one randomly selected sample from the test set. The shapes of both the pointwise confidence intervals and the simultaneous bands are similar under these two kernels. It implies that pointwise confidence intervals and simultaneous confidence bands are robust to the choice of the kernel used to calculate the inner product in $\mathcal{H}_X$ and $\mathcal{H}_Y$. Not surprisingly, the simultaneous confidence band is wider than the pointwise confidence intervals for both kernels. Furthermore, the two left panels of Figure 4 indicate that the predicted mean response tends to be more variable near the boundary in comparison with the interior region. This finding is consistent with what we have seen from the right panel of Figure 3.

In this analysis, we ignored the potential day-to-day dependence of the functional data. Thus, we have treated our model as a working model for this data. Developing a nonlinear function-on-function regression model for functional time series data is an useful and challenging problem, and we will leave it to future research.

9 Conclusions

In this paper we have proposed a nonlinear function-on-function regression model based on a linear operator in RKHS. Compared with the current linear function-on-function regression approaches, our approach shows a remarkable improvement in prediction accuracy for both densely and sparsely observed data. In addition, with the aid of nested Hilbert spaces, our method avoids the large number of parameters that need to be estimated when the tensor products of spline basis functions or the eigenfunctions of the predictor and response are deployed in linear function-on-function regression (Ramsay and Silverman, 2005; Yao et al., 2005b; Sun et al., 2018). The estimation procedure can accommodate irregularly and sparsely observed functional predictor and response.

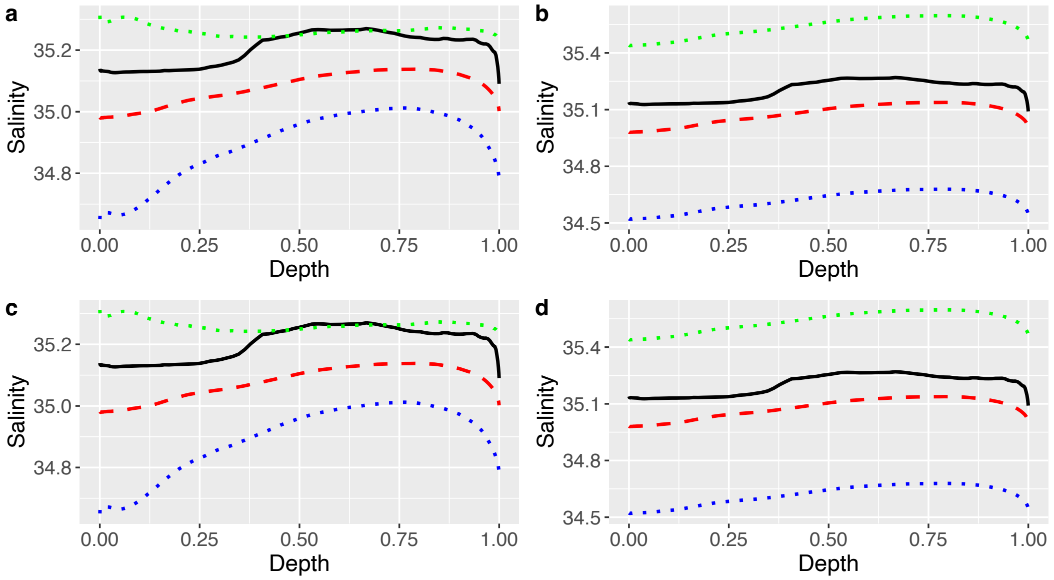

Figure 4: (a) & (b): Pointwise confidence intervals and simultaneous confidence bands constructed from the GRB kernel for one random sample in the test set. (c) & (d): Pointwise confidence intervals and simultaneous confidence bands constructed from the BMC kernel for one random sample in the test set. In each panel, the solid black line represents the observed trajectory of Y, the red dashed line represents the estimated mean, and the blue and green dotted lines represent the upper and the lower bounds of 95% pointwise (or simulataneous) confidence intervals, respectively.

Existing asymptotic development on function-on-function regression was focused on consistency and convergence rates. For instance, Sun et al. (2018) studied the minimax rate in mean prediction using an RKHS-based approach. Both consistency and the convergence rate were established by Luo and Qi (2017) in a linear function-on-function regression model. But it should be noted that these theoretical properties were developed for fully observed functional data without any measurement error. In contrast, we establish consistency and convergence rate for our nonlinear regression estimate with irregularly and sparsely observed functional predictor and response, which are even contaminated with measurement error. Moreover, little work has been done to develop statistical inferences for function-on-function regression. Though there were several precursors in this regard [see Yao et al. (2005b) and Crambes and Mas (2013) for example], they were mainly concerned with linear models. In comparison, our theoretical development includes both convergence rate, pointwise and uniform central limit theorem of the regression estimate, when the full trajectory of functional covariate and response are available.

In this Appendix we provide technical assumptions for Y and the proofs of the theorems, lemmas, and corollaries in the manuscript. The equation labels such as (1) and (2) are for the equations in the manuscript; equation labels such as (A1) and (A2) are for the equations in this Appendix.

Appendix A.1. Technical assumptions for the response

We assume the following regularity conditions on Y to ensure desirable estimation accuracy of $\hat{Y}_i(t)$ when we use the FPCA method as outlined in Section 3.2. Let $\{(\omega_k, \psi_k) : k \in \mathbb{N}\}$ denote the eigen-pairs of the covariance function of Y. Then $\eta_k = \int_{\mathbb{I}} \{Y(t) - \mu_Y(t)\}\psi_k(t)dt$ is the kth FPC score of Y for $k \in \mathbb{N}$.

Assumption C4 *Y has finite fourth moment, i.e., $\int_{\mathbb{I}} E\{Y^4(t)\}dt < \infty$, the function $t \mapsto \mathrm{E}\{Y^2(t)\}$ is continuous, and $E(\eta_k^4) \preceq \omega_k^2$ for any $k \in \mathbb{N}$.*

Assumption C5 *$D^{-1}k^{-a-1} \leq \omega_{k+1} < \omega_k \leq Dk^{-a}$ for any $k \in \mathbb{N}$ and some $D > 1$ and $a > 1$.*

Assumption C6 *$\sup_{k\geq 1} \sup_{t\in[0,1]} |\psi_k(t)| < \infty$ and*

$$\sup_{t\in[0,1]} |\psi_k^{(j)}(t)| \preceq k^{c/2} \sup_{t\in[0,1]} |\psi_k^{(j-1)}(t)|, \quad \textit{for } j = 1, 2 \textit{ and any } k \in \mathbb{N},$$

where $c > 0$ is a constant. Additionally, $\psi_k(0) = \psi_k(1)$ and $\psi_k^{(1)}(0) = \psi_k^{(1)}(1)$ for any $k \in \mathbb{N}$.

These assumptions are essentially the same as Assumptions C1-C3 in the main text. Here we use the same constants, i.e., D, a, c, as in Assumptions C1-C3 for ease of notation. In general, these constants for Y can be different from their counterparts for X.

Appendix A.2. Theoretical proofs

Proof. [Proof of Theorem 1] We first establish a relation between the regression function f_α and the regression operator R_{XY}. By (4), $\mathrm{E}(U_\alpha) = 0$, $\mathrm{E}_\alpha f_\alpha(X) = 0$, and $U_\alpha \perp\!\!\!\perp X$, we have

$$\mathrm{E}[\langle Y, \alpha\rangle_{\mathcal{H}_Y} \mid X] = c_\alpha + f_\alpha(X), \quad \mathrm{E}[\langle Y, \alpha\rangle_{\mathcal{H}_Y}] = c_\alpha.$$

So

$$\mathrm{E}[\langle Y, \alpha\rangle_{\mathcal{H}_Y} \mid X] = \langle \mu_Y, \alpha\rangle_{\mathcal{H}_Y} + f_\alpha(X).$$

Comparing this with (1), we see that

$$f_\alpha(X) = (R_{XY}\alpha)(X) - \mathrm{E}[(R_{XY}\alpha)(X)].$$

By the reproducing property,

$$(R_{XY}\alpha)(X) = \langle R_{XY}\alpha, \kappa(\cdot, X)\rangle_{\mathfrak{M}_X} = \langle \alpha, R_{XY}^* \kappa(\cdot, X)\rangle_{\mathcal{H}_Y}.$$

Similarly,

$$\mathrm{E}[(R_{XY}\alpha)(X)] = \langle R_{XY}\alpha, \mathrm{E}[\kappa(\cdot, X)]\rangle_{\mathfrak{M}_X} = \langle \alpha, R_{XY}^* \mu_X\rangle_{\mathcal{H}_Y}.$$

So

$$f_\alpha(X) = \langle \alpha, R^*_{XY}(\kappa(\cdot, X) - \mu_X) \rangle_{\mathcal{H}_Y}. \tag{A1}$$

Substituting this into (4), we have

$$\langle Y, \alpha \rangle_{\mathcal{H}_Y} = c_\alpha + \langle \alpha, R^*_{XY}(\kappa(\cdot, X) - \mu_X) \rangle_{\mathcal{H}_Y} + U_\alpha. \tag{A2}$$

Secondly, we show that the mappings

$$\alpha \mapsto c_\alpha, \quad \alpha \mapsto f_\alpha, \quad \alpha \mapsto U_\alpha$$

from $\mathcal{H}_Y$ to $\mathcal{H}_Y$, from $\mathcal{H}_Y$ to $\mathfrak{M}_X$, and from $\mathcal{H}_Y$ to $\mathcal{H}_Y$, respectively, are linear in α. That $\alpha \mapsto f_\alpha(x)$ is a linear function follows obviously from (A1). Note that, for any $\alpha_1, \alpha_2 \in \mathcal{H}_Y$ and $\lambda_1, \lambda_2 \in \mathbb{R}$, we have

$$\langle \lambda_1\alpha_1 + \lambda_2\alpha_2, Y \rangle_{\mathcal{H}_Y} = \lambda_1 \langle \alpha_1, Y \rangle_{\mathcal{H}_Y} + \lambda_2 \langle \alpha_2, Y \rangle_{\mathcal{H}_Y}. \tag{A3}$$

Applying the scalar model (4) to the left-hand side of (A3), we have

$$\langle \lambda_1\alpha_1 + \lambda_2\alpha_2, Y \rangle_{\mathcal{H}_Y} = c_{\lambda_1\alpha_1+\lambda_2\alpha_2} + f_{\lambda_1\alpha_1+\lambda_2\alpha_2}(X) + U_{\lambda_1\alpha_1+\lambda_2\alpha_2}.$$

Applying the scalar model to the right-hand side of (A3), we have

$$\begin{aligned} &c_{\lambda_1\alpha_1+\lambda_2\alpha_2} + f_{\lambda_1\alpha_1+\lambda_2\alpha_2}(X) + U_{\lambda_1\alpha_1+\lambda_2\alpha_2} \\ &\quad = (\lambda_1 c_{\alpha_1} + \lambda_2 c_{\alpha_2}) + (\lambda_1 f_{\alpha_1} + \lambda_2 f_{\alpha_2})(X) + (\lambda_1 U_{\alpha_1} + \lambda_2 U_{\alpha_2}). \end{aligned} \tag{A4}$$

Taking expectation of this equation with respect to (X, U_α), we have

$$c_{\lambda_1\alpha_1+\lambda_2\alpha_2} = \lambda_1 c_{\alpha_1} + \lambda_2 c_{\alpha_2}, \tag{A5}$$

which means $\alpha \mapsto c_\alpha$ is a linear mapping. Thus we now have

$$f_{\lambda_1\alpha_1+\lambda_2\alpha_2}(X) + U_{\lambda_1\alpha_1+\lambda_2\alpha_2} = (\lambda_1 f_{\alpha_1} + \lambda_2 f_{\alpha_2})(X) + (\lambda_1 U_{\alpha_1} + \lambda_2 U_{\alpha_2}).$$

By the linearity of $\alpha \mapsto f_\alpha(x)$, the first terms on both sides of the equation cancel out, leaving us

$$U_{\lambda_1\alpha_1+\lambda_2\alpha_2} = \lambda_1 U_{\alpha_1} + \lambda_2 U_{\alpha_2},$$

which means $\alpha \mapsto U_\alpha$ is a linear map.

Thirdly, we show that c_α and U_α can be written as inner products. Note that

$$|c_\alpha| = |\mathrm{E}\,[\langle Y, \alpha \rangle_{\mathcal{H}_Y}]| \leq (\mathrm{E}\,\|Y\|_{\mathcal{H}_Y})\|\alpha\|_{\mathcal{H}_Y}.$$

Since $\mathrm{E}\,(\|Y\|_{\mathcal{H}_Y}) < \infty$ under Assumption 2, c_α is bounded linear functional. By Riesz representation theorem, there exists $c \in \mathcal{H}_Y$ such that $c_\alpha = \langle c, \alpha \rangle_{\mathcal{H}_Y}$ fro all $\alpha \in \mathcal{H}_Y$. Then, for all $\alpha \in \mathcal{H}_Y$,

$$\langle \alpha, \mu_Y \rangle_{\mathcal{H}_Y} = \langle \alpha, c \rangle_{\mathcal{H}_Y},$$

which implies $c = \mu_Y$. Substituting (A2) and $c_\alpha = \langle c, \alpha \rangle_{\mathcal{H}_Y}$ into (4), we have

$$\langle Y, \alpha \rangle_{\mathcal{H}_Y} = \langle \alpha, \mu_Y \rangle_{\mathcal{H}_Y} + \langle \alpha, R^*_{XY}(\kappa(\cdot, X) - \mu_X) \rangle_{\mathcal{H}_Y} + U_\alpha, \tag{A6}$$

which implies

$$U_\alpha = \langle \alpha, Y - \mu_Y - R^*_{XY}(\kappa(\cdot, X) - \mu_X) \rangle_{\mathcal{H}_Y}.$$

So, for each $\omega \in \Omega$ and each $\alpha \in \mathcal{H}_Y$,

$$|U_\alpha(\omega)| \leq \|\alpha\|_{\mathcal{H}_Y} \cdot \|Y(\omega) - \mu_Y - R^*_{XY}(\kappa(\cdot, X(\omega)) - \mu_X)\|_{\mathcal{H}_Y}.$$

Since $Y(\omega) - \mu_Y - R^*_{XY}(\kappa(\cdot, X(\omega)) - \mu_X)$ is a member of $\mathcal{H}_Y$, its norm is finite. Hence, by Riesz representation theorem, there exists an $U(w) \in \mathcal{H}_Y$ such that

$$U_\alpha(\omega) = \langle U(\omega), \alpha \rangle_{\mathcal{H}_Y}$$

for any $\alpha \in \mathcal{H}_Y$. Let U be the random element $\omega \mapsto U(w)$. Then $U_\alpha = \langle U, \alpha \rangle_{\mathcal{H}_Y}$. Since $\langle U, \alpha \rangle_{\mathcal{H}_Y} \perp\!\!\!\perp X$ for all α, we have $U \perp\!\!\!\perp X$.

Finally, substituting $U_\alpha = \langle U, \alpha \rangle_{\mathcal{H}_Y}$ into (A6), we have

$$\langle Y, \alpha \rangle_{\mathcal{H}_Y} = \langle \alpha, \mu_Y \rangle_{\mathcal{H}_Y} + \langle \alpha, R^*_{XY}(\kappa(\cdot, X) - \mu_X) \rangle_{\mathcal{H}_Y} + \langle U, \alpha \rangle_{\mathcal{H}_Y}.$$

Since the above holds for all $\alpha \in \mathcal{H}_Y$, we have the desired equation (4). ■

Proof. [Proof of Theorem 2] By symmetry, we only need to consider $i = 1, \ldots, n/2$. By (8), we have, for any $t \in [0, 1]$,

$$\begin{aligned}\hat{X}_i(t) - X_i(t) &= \sum_{k=1}^{m_X} \hat{\xi}_{ik} \hat{\phi}_{(2),k}(t) - \sum_{k=1}^{m_X} \xi_{ik} \phi_k(t) - \sum_{k=m_X+1}^{\infty} \xi_{ik} \phi_k(t) \\ &= \sum_{k=1}^{m_X} (\hat{\xi}_{ik} - \xi_{ik}) \hat{\phi}_{(2),k}(t) + \sum_{k=1}^{m_X} \xi_{ik} [\hat{\phi}_{(2),k}(t) - \phi_{(2),k}(t)] - \sum_{k=m_X+1}^{\infty} \xi_{ik} \phi_k(t) \\ &\equiv A_1(t) + A_2(t) + A_3(t).\end{aligned}$$

By the triangular inequality and the inequality $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$, we have

$$\begin{aligned}E\{\|\hat{X}_i - X_i\|^2\} &\leq \mathrm{E}\,(\|A_1\| + \|A_1\| + \|A_1\|)^2 \\ &\preceq E(\|A_1\|^2) + E(\|A_2\|^2) + E(\|A_3\|^2).\end{aligned} \tag{A7}$$

The first term on the right is

$$\sum_{k=1}^{m_X} \sum_{\ell=1}^{m_X} \mathrm{E}\,[(\hat{\xi}_{ik} - \xi_{ik})(\hat{\xi}_{i\ell} - \xi_{i\ell}) \langle \hat{\phi}_{(2),k}, \hat{\phi}_{(2),\ell} \rangle] = \sum_{k=1}^{m_X} \mathrm{E}\,(\hat{\xi}_{ik} - \xi_{ik})^2,$$

where the equality holds because $\{\hat{\phi}_{(2),k}\}_{k=1}^{m_X}$ is an orthonormal set. The second term on the right-hand side of (A7) is

$$\sum_{k=1}^{m_X} \sum_{\ell=1}^{m_X} \mathrm{E}\,[\xi_{ik} \xi_{i\ell} \langle \hat{\phi}_{(2),k} - \phi_{(2),k}, \hat{\phi}_{(2),\ell} - \phi_{(2),\ell} \rangle]. \tag{A8}$$

By the iterative expectation law, we rewrite the expectation in the summand as

$$\begin{aligned}
&\mathrm{E}\left\{\mathrm{E}\left\{\xi_{ik}\xi_{i\ell}\langle\hat{\phi}_{(2),k}-\phi_{(2),k},\hat{\phi}_{(2),\ell}-\phi_{(2),\ell}\rangle\right\}|\hat{\phi}_{(2),k}(t),\ \hat{\phi}_{(2),\ell}(t)\right\}\\
&=\mathrm{E}\left\{\mathrm{E}\left(\xi_{ik}\xi_{i\ell}\right)\langle\hat{\phi}_{(2),k}-\phi_{(2),k},\hat{\phi}_{(2),\ell}-\phi_{(2),\ell}\rangle\right\}\\
&=\delta_{k,\ell}\mathrm{E}\left(\xi_{ik}^2\right)\mathrm{E}\left\{\langle\hat{\phi}_{(2),k}-\phi_{(2),k},\hat{\phi}_{(2),\ell}-\phi_{(2),\ell}\rangle\right\},
\end{aligned}$$

where the first equality holds because $\xi_{ik}\xi_{i\ell}\perp\!\!\!\perp(\hat{\phi}_{(2),k},\ \hat{\phi}_{(2),\ell}(t))$, and the second holds because ξ_{ik} and $\xi_{i\ell}$ are uncorrelated. Then, (A8) reduces to

$$\sum_{k=1}^{m}\mathrm{E}\left(\xi_{ik}^2\right)\mathrm{E}\left\|\hat{\phi}_{(2),k}-\phi_{(2),k}\right\|^2 \tag{A9}$$

Combining (A7), (A8), and (A9), we have

$$E\{\|\hat{X}_i-X_i\|^2\}\preceq\sum_{k=1}^{m_X}E(\hat{\xi}_{ik}-\xi_{ik})^2+\sum_{k=1}^{m_X}\mathrm{E}\left(\xi_{ik}^2\right)\mathrm{E}\left\|\hat{\phi}_{(2),k}-\phi_{(2),k}\right\|^2+\sum_{k=m_X+1}^{\infty}E(\xi_{ik}^2).$$

Further decompose $\hat{\xi}_{ik}-\xi_{ik}$ as $(\hat{\xi}_{ik}-\tilde{\xi}_{ik})+(\tilde{\xi}_{ik}-\xi_{ik})$, where $\tilde{\xi}_{ik}=N_X^{-1}\sum_{j=1}^{N_X}X_{ij}\phi_k(t_{ij})$. Then, by the inequality $(a+b)^2\le 2(a^2+b^2)$,

$$\begin{aligned}
&E\{\|\hat{X}_i-X_i\|^2\}\\
&\preceq\sum_{k=1}^{m_X}E(\hat{\xi}_{ik}-\tilde{\xi}_{ik})^2+\sum_{k=1}^{m_X}E(\tilde{\xi}_{ik}-\xi_{ik})^2+\sum_{k=1}^{m_X}\mathrm{E}\left(\xi_{ik}^2\right)\mathrm{E}\left\|\hat{\phi}_{(2),k}-\phi_{(2),k}\right\|^2+\sum_{k=m_X+1}^{\infty}E(\xi_{ik}^2)\\
&\equiv B_1+B_2+B_3+B_4.
\end{aligned} \tag{A10}$$

We next analyze the four terms on the right-hand side of (A10). In the first term,

$$\begin{aligned}
E(\hat{\xi}_{ik}-\tilde{\xi}_{ik})^2&=E\left(\left[\frac{1}{N_X}\sum_{j=1}^{N_X}X_{ij}\{\hat{\phi}_{(2),k}(t_{ij})-\phi_k(t_{ij})\}\right]^2\right)\\
&=\frac{1}{N_X^2}\sum_{j\neq\ell}E[X_{ij}X_{i\ell}\{\hat{\phi}_{(2),k}(t_{ij})-\phi_k(t_{ij})\}\{\hat{\phi}_{(2),k}(t_{i\ell})-\phi_k(t_{i\ell})\}]\\
&\quad+\frac{1}{N_X^2}\sum_{j=1}^{N_X}E[X_{ij}^2\{\hat{\phi}_{(2),k}(t_{ij})-\phi_k(t_{ij})\}^2]\equiv\frac{1}{N_X^2}\sum_{j\neq\ell}C_{ik,j\ell}+\frac{1}{N_X^2}\sum_{j=1}^{N_X}D_{ik,j}.
\end{aligned} \tag{A11}$$

For the $D_{ik,j}$ terms in (A11), by the law of iterated expectations,

$$D_{ik,j}=\mathrm{E}\left\{\mathrm{E}\left[(X_i(t_{ij})+e_{ij})^2\{\hat{\phi}_{(2),k}(t_{ij})-\phi_k(t_{ij})\}^2|X_i,\hat{\phi}_{(2),k},e_{ij}\right]\right\}.$$

Since $(X_i,\hat{\phi}_{(2),k},e_{ij})\perp\!\!\!\perp t_{ij}$ and $t_{ij}\sim U[0,1]$, we have

$$\mathrm{E}\left[(X_i(t_{ij})+e_{ij})^2\{\hat{\phi}_{(2),k}(t_{ij})-\phi_k(t_{ij})\}^2|X_i,\hat{\phi}_{(2),k},e_{ij}\right]=\int_{\mathbb{I}}(X_i(t)+e_{ij})^2\{\hat{\phi}_{(2),k}(t)-\phi_k(t)\}^2dt.$$

Hence

$$\begin{aligned} D_{ik,j} =& E\left[\int_{\mathbb{I}}(X_i(t)+e_{ij})^2\{\hat{\phi}_{(2),k}(t)-\phi_k(t)\}^2 dt\right] \\ =& \int_{\mathbb{I}} \mathrm{E}\,[(X_i(t)+e_{ij})^2]\mathrm{E}\,\{[\hat{\phi}_{(2),k}(t)-\phi_k(t)]\}^2 dt \\ =& \int_{\mathbb{I}}(\mathrm{E}\,X_i^2(t)+\sigma_e^2)\mathrm{E}\,\{[\hat{\phi}_{(2),k}(t)-\phi_k(t)]\}^2 dt \\ \preceq& E\|\hat{\phi}_{(2),k}-\phi_k\|^2, \end{aligned}$$

where the second equality holds because $(X_i(t)+e_{ij}) \perp\!\!\!\perp \hat{\phi}_{(2),k}(t)$, the third holds because $X_i(t) \perp\!\!\!\perp e_{ij}$, and $\preceq$ holds because, by condition C1, $EX_i^2(t)$ is continuous and therefore bounded. Then

$$\frac{1}{N_X^2}\sum_{j=1}^{N_X} D_{ik,j} \preceq N_X^{-1}E\|\hat{\phi}_{(2),k}-\phi_k\|^2. \tag{A12}$$

For the terms $C_{ik,j\ell}$ in (A11), since, for any $j \neq \ell$,

$$(X_i(t_{ij})+e_{ij})[\hat{\phi}_{(2),k}(t_{ij})-\phi_k(t_{ij})] \perp\!\!\!\perp (X_i(t_{i\ell})+e_{i\ell})[\hat{\phi}_{(2),k}(t_{i\ell})-\phi_k(t_{i\ell})]|(X_i,\hat{\phi}_{(2),k}),$$

we have

$$\begin{aligned} &E[X_{ij}X_{i\ell}(\hat{\phi}_{(2),k}(t_{ij})-\phi_k(t_{ij}))(\hat{\phi}_{(2),k}(t_{il})-\phi_k(t_{il})) \mid X_i,\hat{\phi}_{(2),k}] \\ &= E[(X_i(t_{ij})+e_{ij})(X_i(t_{i\ell})+e_{i\ell})(\hat{\phi}_{(2),k}(t_{ij})-\phi_k(t_{ij}))(\hat{\phi}_{(2),k}(t_{il})-\phi_k(t_{il})) \mid X_i,\hat{\phi}_{(2),k}] \\ &= E[(X_i(t_{ij})+e_{ij})(\hat{\phi}_{(2),k}(t_{ij})-\phi_k(t_{ij})) \mid X_i,\hat{\phi}_{(2),k}] \\ &\quad E[(X_i(t_{i\ell})+e_{i\ell})(\hat{\phi}_{(2),k}(t_{i\ell})-\phi_k(t_{i\ell})) \mid X_i,\hat{\phi}_{(2),k}]. \end{aligned}$$

Since $e_{ij} \perp\!\!\!\perp (X_i,\hat{\phi}_{(2),k})$, $E(e_{ij})=0$, and $t_{ij}\sim U[0,1]$, and since the same is true if we replace j by ℓ, the right-hand side above further reduces to

$$\left[\int_0^1 X_i(t)\{\hat{\phi}_{(2),k}(t)-\phi_k(t)\}dt\right]^2 \le \|X_i\|^2\,\|\phi_{(2),k}-\phi_k\|^2$$

Then

$$C_{ik,j\ell} \le E(\|X_i\|^2)E(\|\hat{\phi}_{(2),k}-\phi_k\|^2) \preceq E(\|\hat{\phi}_{(2),k}-\phi_k\|^2).$$

Consequently,

$$\frac{1}{N_X^2}\sum_{j\neq\ell}^{N_X} C_{ik,j\ell} \preceq \frac{N_X^2-N_X}{N_X^2}E(\|\hat{\phi}_{(2),k}-\phi_k\|^2) \preceq E(\|\hat{\phi}_{(2),k}-\phi_k\|^2). \tag{A13}$$

In the second term in (A11), from the proof of Lemma S1 in Zhou et al. (2023).

$$E(\tilde{\xi}_{ik}-\xi_{ik})^2 = E\left(\left[\frac{1}{N_X}\sum_{j=1}^{N_X} X_{ij}\phi_k(t_{ij})-\langle X_i,\phi_k\rangle\right]^2\right) \le C_0 N_X^{-1} \tag{A14}$$

for some constant $C_0 > 0$.

Combining (A12), (A13), and (A14), we have

$$\begin{aligned} B_1 + B_2 &\preceq N_X^{-1} \sum_{k=1}^{m_X} E\|\hat{\phi}_{(2),k} - \phi_k\|^2 + \sum_{k=1}^{m_X} E\|\hat{\phi}_{(2),k} - \phi_k\|^2 + N_X^{-1} \\ &\preceq \sum_{k=1}^{m_X} E\|\hat{\phi}_{(2),k} - \phi_k\|^2 + N_X^{-1}. \end{aligned} \tag{A15}$$

Next, consider the last two terms in (A10). Since, by Assumption C2, $E(\xi_{ik}^2) \preceq k^{-a}$, we have

$$B_3 \preceq \sum_{k=1}^{m_X} k^{-a} E(\|\hat{\phi}_{(2),k} - \phi_k\|^2). \tag{A16}$$

By Assumption C2,

$$B_4 = \sum_{k=m_X+1}^{\infty} \theta_k \leq \sum_{k=m_X+1}^{\infty} Dk^{-a} \preceq m_X^{-(a-1)}. \tag{A17}$$

From (A15), (A16), and (A17) we see that

$$\begin{aligned} E\{\|\hat{X}_i - X_i\|^2\} &\preceq \sum_{k=1}^{m_X} E\|\hat{\phi}_{(2),k} - \phi_k\|^2 + N_X^{-1} + \sum_{k=1}^{m_X} k^{-a} E(\|\hat{\phi}_{(2),k} - \phi_k\|^2) + m_X^{-(a-1)} \\ &\preceq \sum_{k=1}^{m_X} E\|\hat{\phi}_{(2),k} - \phi_k\|^2 + N_X^{-1} + m_X^{-(a-1)}, \end{aligned} \tag{A18}$$

where the second $\preceq$ holds because $\sum_{k=1}^{m_X} k^{-a} E(\|\hat{\phi}_{(2),k} - \phi_k\|^2) \leq \sum_{k=1}^{m_X} E\|\hat{\phi}_{(2),k} - \phi_k\|^2$. By Theorem 2 of Zhou et al. (2025), we have, under Assumptions C1 to C3 and M1,

$$E\|\hat{\phi}_{(2),k} - \phi_k\|^2 \preceq \frac{k^2}{n}\left[1 + \left(\frac{k^a}{N_X}\right)^2\right] + \frac{k^a}{nN_X h_X}\left(1 + \frac{k^a}{N_X}\right) + h_X^4 k^{2c+2}.$$

Hence the first term on the right-hand side of (A18) satisfies

$$\begin{aligned} \sum_{k=1}^{m_X} E\|\hat{\phi}_{(2),k} - \phi_k\|^2 &\preceq \sum_{k=1}^{m_X} \left[\frac{k^2}{n}\left\{1 + \left(\frac{k^a}{N_X}\right)^2\right\} + \frac{k^a}{nN_X h_X}\left(1 + \frac{k^a}{N_X}\right) + h_X^4 k^{2c+2}\right] \\ &\preceq \frac{m_X^3}{n} + \frac{m_X^{2a+3}}{nN_X^2} + \frac{m_X^{a+1}}{nN_X h_X} + \frac{m_X^{2a+1}}{nN_X^2 h_X} + h_X^4 m_X^{2c+3}. \end{aligned} \tag{A19}$$

Substituting this into (A18), we have

$$E\{\|\hat{X}_i - X_i\|^2\} \preceq \frac{m_X^3}{n} + \frac{m_X^{2a+3}}{nN_X^2} + \frac{m_X^{a+1}}{nN_X h_X} + \frac{m_X^{2a+1}}{nN_X^2 h_X} + h_X^4 m_X^{2c+3} + N_X^{-1} + m_X^{-(a-1)},$$

Using condition (M1), we can show that

$$\frac{m_X^3}{n} \preceq m_X^{-(a-1)}, \quad N_X^{-1} \preceq m_X^{-(a-1)}.$$

Therefore,

$$\|\hat{X}_i - X_i\|^2 \preceq O_P\left(m_X^{-(a-1)} + \frac{m_X^{2a+3}}{nN_X^2} + \frac{m_X^{a+1}}{nN_X h_X} + \frac{m^{2a+1}}{nN^2 h} + h^4 m^{2c+3}\right).$$

We establish the desired equation (9).

■

Proof. [Proof of Lemma 6] *1.* Since, for any $f_1, f_2 \in \mathfrak{M}_X$, $f_1 \otimes (R_{XY}^* f_2) = (f_1 \otimes f_2) R_{XY}$, we have

$$\begin{aligned}\Sigma_{XU} =& \mathrm{E}\left[(\kappa(\cdot, X) - \mu_X) \otimes \{Y - \mu_Y - R_{XY}^*(\kappa(\cdot, X) - \mu_X)\}\right] \\ =& \mathrm{E}\left[(\kappa(\cdot, X) - \mu_X) \otimes (Y - \mu_Y)\right] - \mathrm{E}\left[(\kappa(\cdot, X) - \mu_X) \otimes (R_{XY}^*\{\kappa(\cdot, X) - \mu_X\})\right] \\ =& \Sigma_{XY} - \Sigma_{XX} R_{XY} = 0.\end{aligned}$$

2. By definition,

$$\begin{aligned}\hat{\Sigma}_{XU} =& \mathrm{E}_n[(\kappa(\cdot, X) - \hat{\mu}_X) \otimes \{Y - \mu_Y - R_{XY}^*(\kappa(\cdot, X) - \mu_X)\}] \\ =& \mathrm{E}_n[(\kappa(\cdot, X) - \hat{\mu}_X) \otimes (Y - \mu_Y)] - \mathrm{E}_n[(\kappa(\cdot, X) - \hat{\mu}_X) \otimes \{R_{XY}^*(\kappa(\cdot, X) - \mu_X)\}].\end{aligned}$$

The first term on the right is $\hat{\Sigma}_{XY}$. Since $\mathrm{E}_n[\kappa(\cdot, X) - \hat{\mu}_X] = 0$, the second term is unchanged if we replace μ_X in $\kappa(\cdot, X) - \mu_X$ by $\hat{\mu}_X$. Thus it can be rewritten as

$$\mathrm{E}_n[(\kappa(\cdot, X) - \hat{\mu}_X) \otimes (R_{XY}^*(\kappa(\cdot, X) - \hat{\mu}_X))] = \hat{\Sigma}_{XX} R_{XY},$$

as desired. ■

Proof. [Proof of Lemma 7] By the definitions of $\hat{\Sigma}_{XU}$ and $\tilde{\Sigma}_{XU}$ and some simple calculation, we have

$$\hat{\Sigma}_{XU} - \tilde{\Sigma}_{XU} = (\hat{\mu}_X - \mu_X) \otimes \hat{\mu}_U. \tag{A20}$$

Hence

$$\|\hat{\Sigma}_{XU} - \tilde{\Sigma}_{XU}\|_{\mathrm{HS}} = \|(\hat{\mu}_X - \mu_X) \otimes \hat{\mu}_U\|_{\mathrm{HS}} = \|\hat{\mu}_X - \mu_X\|_{\mathfrak{M}_X} \|\hat{\mu}_U\|_{\mathcal{H}_Y}.$$

By Chebychev's inequality, it can be easily shown that $\|\hat{\mu}_X - \mu_X\|_{\mathfrak{M}_X} = O_P(n^{-1/2})$ and $\|\hat{\mu}_U\|_{\mathcal{H}_Y} = O_P(n^{-1/2})$, which imply the asserted result. ■

Proof. [Proof of Lemma 8] Let $m_n = \lfloor \epsilon_n^{-1/\alpha} \rfloor$. Then, by Assumption 7(ii),

$$\begin{aligned}\textstyle\sum_{j\in\mathbb{N}} (\lambda_i + \epsilon_n)^{-2} \lambda_j \leq& \textstyle\sum_{j=1}^{m_n} \lambda_i^{-1} + \epsilon_n^{-2} \sum_{j=m_n+1}^{\infty} \lambda_j \\ &\asymp \int_1^{m_n} x^{\alpha} dx + \epsilon_n^{-2} \int_{m_n}^{\infty} x^{-\alpha} dx \asymp \epsilon_n^{-(\alpha+1)/\alpha},\end{aligned} \tag{A21}$$

as desired. ■

Proof. [Proof of Theorem 9] *1.* Using Lemma 6, we decompose $\hat{R}_{XY}$ as $\hat{R}_{\text{reg}} + \hat{R}_{\text{res}}$, where

$$\hat{R}_{\text{reg}} = \hat{V}\hat{\Sigma}_{XX}R_{XY}, \quad \hat{R}_{\text{res}} = \hat{V}\hat{\Sigma}_{XU}.$$

As suggested by the notation, $\hat{R}_{\text{reg}}$ represents the regression part of $\hat{R}_{XY}$, whereas $\hat{R}_{\text{res}}$ the residual part. Let $R_n = V_n\Sigma_{XY}$, which represents the population-level approximation of R_{XY} via Tychonoff regularization, and further decompose $\hat{R}_{\text{reg}}$ as $\hat{R}_{\text{reg}} - R_n + R_n$. We have

$$\hat{R}_{XY} - R_{XY} = \hat{R}_{\text{res}} + (\hat{R}_{\text{reg}} - R_n) + (R_n - R_{XY}).$$

We first analyze the regression term $\hat{R}_{\text{reg}} - R_n$. By construction,

$$\hat{R}_{\text{reg}} - R_n = \hat{V}\hat{\Sigma}_{XX}R_{XY} - R_n = \hat{V}\hat{\Sigma}_{XX}R_{XY} - V_n\Sigma_{XX}R_{XY} = (\hat{V}\hat{\Sigma}_{XX} - V_n\Sigma_{XX})R_{XY}.$$

Since $\hat{V}$ and $\hat{\Sigma}_{XX}$ commute, and V_n and Σ_{XX} commute, we can rewrite

$$\hat{V}\hat{\Sigma}_{XX} - \Sigma_{XX}V_n = \hat{V}(\hat{\Sigma}_{XX}V_n^{-1} - \hat{V}^{-1}\Sigma_{XX})V_n = \epsilon_n\hat{V}(\hat{\Sigma}_{XX} - \Sigma_{XX})V_n.$$

Therefore,

$$\|\hat{R}_{\text{reg}} - R_n\|_{\text{OP}} \leq \|\epsilon_n\hat{V}\|_{\text{OP}}\,\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{\text{OP}}\,\|V_nR_{XY}\|_{\text{OP}} = O_P(n^{-1/2})\,\|V_nR_{XY}\|_{\text{OP}},$$

where the second equality holds because $\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{\text{OP}} = O_P(n^{-1/2})$ (Lemma 5) and $\|\epsilon_n\hat{V}\|_{\text{OP}} \leq \|I\|_{\text{OP}} = 1$. By Assumption 6(ii), $V_nR_{XY} = V_nV\Sigma_{XX}^{1+\beta}S_{XY}$ for a bounded operator S_{XY}. Hence if $\beta \in (0,1]$,

$$\begin{aligned}\|V_nR_{XY}\|_{\text{OP}} &= \|V_nV\Sigma_{XX}^{1+\beta}S_{XY}\|_{\text{OP}} = \|V_n\Sigma_{XX}^{\beta}\|_{\text{OP}}\|S_{XY}\|_{\text{OP}}\\ &\leq \|(\Sigma_{XX} + \epsilon_nI)^{-1+\beta}\|_{\text{OP}}\|S_{XY}\|_{\text{OP}}\\ &= \epsilon_n^{\beta-1}\|(\Sigma_{XX} + \epsilon_nI)^{-1+\beta}(\epsilon_nI)^{1-\beta}\|_{\text{OP}}\|S_{XY}\|_{\text{OP}}\\ &= O(\epsilon_n^{\beta-1}).\end{aligned}$$

If $\beta > 1$, one has

$$\|V_n\Sigma_{XX}^{\beta}\|_{\text{OP}} \leq \|(\Sigma_{XX} + \epsilon_nI)^{-1}\Sigma_{XX}\|_{\text{OP}}\|\Sigma_{XX}\|_{\text{OP}} = O(1).$$

It follows that

$$\|V_nR_{XY}\|_{\text{OP}} = O(\epsilon_n^{\beta\wedge1-1}). \tag{A22}$$

Consequently,

$$\|\hat{R}_{\text{reg}} - R_n\|_{\text{OP}} = O_P(n^{-1/2}\epsilon_n^{\beta\wedge1-1}). \tag{A23}$$

Secondly, we analyze the bias term $R_n - R_{XY}$. Since

$$(V_n - V)\Sigma_{XY} = V_n[\Sigma_{XX} - (\Sigma_{XX} + \epsilon_nI)]V\Sigma_{XY} = -\epsilon_nV_nR_{XY},$$

we have, by (A22),

$$\|R_n - R_{XY}\|_{\text{OP}} = O(\epsilon^{\beta\wedge1}). \tag{A24}$$

Thirdly, we analyze $\hat{R}_{\mathrm{res}}$, which can be further decomposed as

$$(\hat{V}\hat{\Sigma}_{XU} - \hat{V}\tilde{\Sigma}_{XU}) + (\hat{V}\tilde{\Sigma}_{XU} - V_n\tilde{\Sigma}_{XU}) + V_n\tilde{\Sigma}_{XU}.$$

Since $\|\hat{V}\|_{\mathrm{OP}} \leq \epsilon_n^{-1}\|I\|_{\mathrm{OP}} = \epsilon_n^{-1}$ and, by Lemma 7, $\|\hat{\Sigma}_{XU} - \tilde{\Sigma}_{XU}\|_{\mathrm{OP}} = O_P(n^{-1})$, we have

$$\|\hat{V}\hat{\Sigma}_{XU} - \hat{V}\tilde{\Sigma}_{XU}\|_{\mathrm{OP}} = O_P(n^{-1}\epsilon_n^{-1}). \tag{A25}$$

Since $\hat{V} - V_n = \hat{V}(\Sigma_{XX} - \hat{\Sigma}_{XX})V_n$, we have

$$\begin{aligned}\|\hat{V}\tilde{\Sigma}_{XU} - V_n\tilde{\Sigma}_{XU}\|_{\mathrm{OP}} &\leq \|\hat{V}\|_{\mathrm{OP}}\,\|\Sigma_{XX} - \hat{\Sigma}_{XX}\|_{\mathrm{OP}}\,\|V_n\tilde{\Sigma}_{XU}\|_{\mathrm{OP}} \\ &= O_P(n^{-1/2}\epsilon_n^{-1})\,\|V_n\tilde{\Sigma}_{XU}\|_{\mathrm{OP}}.\end{aligned} \tag{A26}$$

The term $\|V_n\tilde{\Sigma}_{XU}\|_{\mathrm{OP}}$ is bounded by $\|V_n\tilde{\Sigma}_{XU}\|_{\mathrm{HS}}$, whose square is

$$\|V_n\tilde{\Sigma}_{XU}\|_{\mathrm{HS}}^2 = \|n^{-1}\textstyle\sum_{i=1}^n V_n[(\kappa(\cdot, X_i) - \mu_X) \otimes U_i]\|_{\mathrm{HS}}^2.$$

Since $\mathrm{E}\,[(\kappa(\cdot, X_i) - \mu_X) \otimes U_i] = \Sigma_{XU} = 0$ and $(X_1, U_1), \ldots, (X_n, U_n)$ are i.i.d., we have

$$\begin{aligned}\mathrm{E}\,(\|V_n\tilde{\Sigma}_{XU}\|_{\mathrm{HS}}^2) &= n^{-2}\textstyle\sum_{a=1}^n\sum_{b=1}^n \mathrm{E}\,(\langle V_n[(\kappa(\cdot, X_a) - \mu_X) \otimes U_a],\, V_n[(\kappa(\cdot, X_b) - \mu_X) \otimes U_b]\rangle_{\mathrm{HS}}^2) \\ &= n^{-2}\textstyle\sum_{a=1}^n \mathrm{E}\,(\|V_n[(\kappa(\cdot, X_a) - \mu_X) \otimes U_a\|_{\mathrm{HS}}^2) \\ &= n^{-1}\mathrm{E}\,(\|V_n[(\kappa(\cdot, X) - \mu_X) \otimes U\|_{\mathrm{HS}}^2).\end{aligned}$$

The squared Hilbert-Schmidt norm on the right-hand side is

$$\begin{aligned}&\|V_n[(\kappa(\cdot, X) - \mu_X) \otimes U\|_{\mathrm{HS}}^2 \\ &\qquad= \textstyle\sum_{j\in\mathbb{N}} \langle \varphi_j, \|U\|^2\langle \varphi_j, [V_n(\kappa(\cdot, X) - \mu_X)] \otimes [V_n(\kappa(\cdot, X) - \mu_X)]\varphi_j\rangle_{\mathfrak{M}_X} \\ &\qquad= \|U\|^2 \textstyle\sum_{j\in\mathbb{N}} (\lambda_j + \epsilon_n)^{-2}\zeta_j^2,\end{aligned}$$

where, for the last equality, we have used the expansion (12). Taking expectation on both sides, and invoking the condition $X \perp\!\!\!\perp U$, we have

$$\mathrm{E}\,(\|V_n[(\kappa(\cdot, X) - \mu_X) \otimes U\|_{\mathrm{HS}}^2) = \mathrm{E}\,(\|U\|^2)\textstyle\sum_{j\in\mathbb{N}} (\lambda_j + \epsilon_n)^{-2}\lambda_j.$$

By Lemma 8, the right hand side is of the order $O(\epsilon_n^{-(\alpha+1)/\alpha})$. Hence $\mathrm{E}\,\|V_n\tilde{\Sigma}_{XU}\|_{\mathrm{HS}}^2$ is of the order $O(n^{-1}\epsilon_n^{-(\alpha+1)/\alpha})$, which, by Chebychev's inequality, implies

$$\|V_n\tilde{\Sigma}_{XU}\|_{\mathrm{HS}} = O_P(n^{-1/2}\epsilon_n^{-(\alpha+1)/(2\alpha)}) \;\Rightarrow\; \|V_n\tilde{\Sigma}_{XU}\|_{\mathrm{OP}} = O_P(n^{-1/2}\epsilon_n^{-(\alpha+1)/(2\alpha)}).$$

Combining this with (A26) we have

$$\|\hat{V}\tilde{\Sigma}_{XU} - V_n\tilde{\Sigma}_{XU}\|_{\mathrm{OP}} = O_P(n^{-1/2}\epsilon_n^{-1}n^{-1/2}\epsilon_n^{-(\alpha+1)/(2\alpha)}) = O_P(n^{-1}\epsilon_n^{-(3\alpha+1)/(2\alpha)}). \tag{A27}$$

So

$$\begin{aligned}\hat{R}_{\mathrm{res}} &= O_P(n^{-1}\epsilon_n^{-1} + n^{-1}\epsilon_n^{-(3\alpha+1)/(2\alpha)} + n^{-1/2}\epsilon_n^{-(\alpha+1)/(2\alpha)}) \\ &= O_P(n^{-1}\epsilon_n^{-(3\alpha+1)/(2\alpha)} + n^{-1/2}\epsilon_n^{-(\alpha+1)/(2\alpha)}).\end{aligned} \tag{A28}$$

Combining (A23), (A24), and (A28), we have (13).

2. For the right-hand side of (13) to go to 0 we need

$$n^{-1/2}\epsilon_n^{\beta\wedge 1-1} \prec 1, \quad n^{-1}\epsilon_n^{-(3\alpha+1)/(2\alpha)} \prec 1, \quad n^{-1/2}\epsilon_n^{-(\alpha+1)/(2\alpha)}) \prec 1,$$

which are satisfied if

$$\epsilon_n \succ n^{-1/[2\{1-(\beta\wedge 1)\}]}, \quad \epsilon_n \succ n^{-(2\alpha)/(3\alpha+1)}, \quad \epsilon_n \succ n^{-\alpha/(\alpha+1)}.$$

It is easy to check that, for $\alpha > 1$, we have $-(2\alpha)/(3\alpha+1) > -\alpha/(\alpha+1)$. Therefore, if the first two relations above hold, then the right-hand side of (13) tends to 0. ■

Proof. [Proof of Theorem 10] *1.* If $\beta > (\alpha-1)/(2\alpha)$, then $\ell_1(\delta) < \ell_4(\delta)$ for all $\delta > 0$, and consequently

$$m(\delta) = \max\{\ell_2(\delta), \ell_3(\delta), \ell_4(\delta)\}.$$

By computation, the intersection of ℓ_2 and ℓ_4 occurs at $\delta_{2,4} = \alpha/(2\alpha\beta+\alpha+1)$, and the intersection of ℓ_3 and ℓ_4 occurs at $\delta_{3,4} = 1/2$. Moreover, $\beta > (\alpha-1)/(2\alpha)$ implies $\delta_{2,4} < 1/2 = \delta_{3,4}$. Hence the relative positions of the three lines ℓ_2, ℓ_3, ℓ_4 are as depicted in Figure 1, left panel, and the minimum of $\max\{\ell_2(\delta), \ell_3(\delta), \ell_4(\delta)\}$ is achieved at $\delta_{\text{opt}} = \delta_{2,4}$, with $m(\delta_{\text{opt}}) = \ell_2(\delta_{2,4}) = -\alpha(\beta\wedge 1)/\{2\alpha(\beta\wedge 1)+\alpha+1\}$.

2. If $\beta \le (\alpha-1)/(2\alpha)$, the $\ell_1(\delta) \ge \ell_4(\delta)$ for all $\delta > 0$, and

$$m(\delta) = \max\{\ell_1(\delta), \ell_2(\delta), \ell_3(\delta)\}.$$

The intersection of ℓ_1 and ℓ_2 occurs at $\delta_{1,2} = 1/2$, and the intersection of ℓ_1 and ℓ_3 occurs at $\delta_{1,3} = \alpha/(2\alpha\beta+\alpha+1)$. Moreover, $\beta < (\alpha-1)/(2\alpha)$ implies $\delta_{1,3} > 1/2 = \delta_{1,2}$. Hence the relative positions of ℓ_1, ℓ_2 and ℓ_3 are as shown in right plot of Figure A1, and the minimum of $\max\{\ell_1(\delta), \ell_2(\delta), \ell_3(\delta)\}$ is achieved at $\delta_{\text{opt}} = \delta_{1,2} = 1/2$ with $m(\delta_{\text{opt}}) = \ell_2(\delta_{1,2}) = -\beta/2$. ■

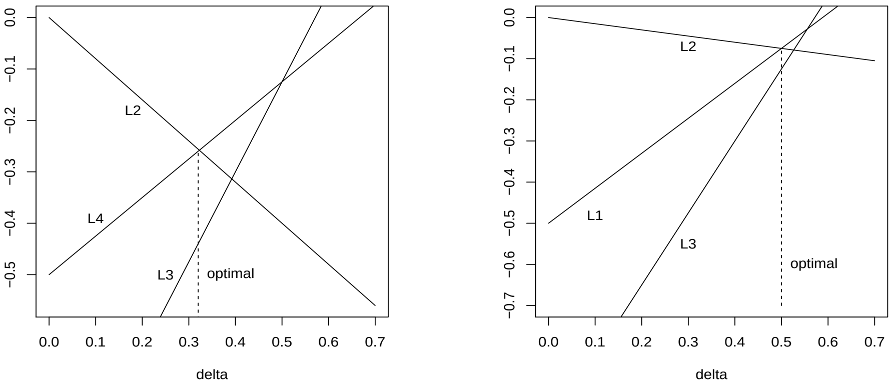

Figure A1: Optimal tuning parameter in two scenarios of β. Left panel: L2, L3, L4 represent the lines ℓ_2, ℓ_3, ℓ_4 with $\beta > (\alpha-1)/(2\alpha)$. Right panel: L1, L2, L3 represent the lines ℓ_1, ℓ_2, ℓ_3 with $\beta < (\alpha-1)/(2\alpha)$.

Proof. [Proof of Corollary 11] By (14) and (15), we have

$$\begin{aligned}
&\widehat{\mathrm{E}}\,(Y|x_0)(t) - \mathrm{E}\,(Y|x_0)(t) \\
&= \langle \upsilon(\cdot,t), [\hat{R}^*_{XY}(\kappa(\cdot,x_0) - \hat{\mu}_X) - R^*_{XY}(\kappa(\cdot,x_0) - \mu_X)]\rangle_{\mathcal{H}_Y} + \hat{\mu}_Y(t) - \mu_Y(t) \\
&= \langle (\hat{R}_{XY} - R_{XY})\,\upsilon(\cdot,t), \kappa(\cdot,x_0) - \mu_X\rangle_{\mathfrak{M}_X} + \langle (\hat{R}_{XY} - R_{XY})\,\upsilon(\cdot,t), \mu_X - \hat{\mu}_X\rangle_{\mathfrak{M}_X} \\
&\quad + \langle R_{XY}\,\upsilon(\cdot,t), \mu_X - \hat{\mu}_X\rangle_{\mathfrak{M}_X} + [\hat{\mu}_Y(t) - \mu_Y(t)].
\end{aligned} \tag{A29}$$

Hence

$$\begin{aligned}
&|\widehat{\mathrm{E}}\,(Y|x_0)(t) - \mathrm{E}\,(Y|x_0)(t)| \\
&\leq \|\hat{R}_{XY} - R_{XY}\|_{\mathrm{OP}}\,\|\upsilon(\cdot,t)\|_{\mathcal{H}_Y}\,\|\kappa(\cdot,x_0) - \mu_X\|_{\mathfrak{M}_X} + \|\hat{R}_{XY} - R_{XY}\|_{\mathrm{OP}}\,\|\upsilon(\cdot,t)\|_{\mathcal{H}_Y} \\
&\quad \|\mu_X - \hat{\mu}_X\|_{\mathfrak{M}_X} + \|R_{XY}\|_{\mathrm{OP}}\,\|\upsilon(\cdot,t)\|_{\mathcal{H}_Y}\,\|\mu_X - \hat{\mu}_X\|_{\mathfrak{M}_X} + |\hat{\mu}_Y(t) - \mu_Y(t)|.
\end{aligned} \tag{A30}$$

Since $\hat{\mu}_Y$ and $\hat{\mu}_X$ are sample averages, by Chebychev's inequality,

$$\|\mu_X - \hat{\mu}_X\|_{\mathfrak{M}_X} = O_P(n^{-1/2}), \quad \hat{\mu}_Y(t) - \mu_Y(t) = O_P(n^{-1/2}).$$

Hence the right-hand side of (A30) is dominated by $\|\hat{R}_{XY} - R_X\|_{\mathrm{OP}}$, which proves (16). The rest of the corollary is obvious. ■

Proof. [Proof of Theorem 13] *1.* For convenience, let g and f denote the functions $\upsilon(\cdot,t)$ and $\kappa(\cdot,x_0) - \mu_X$. Then we can reexpress $A_{n,3}$ as $n^{-1}\sum_{i=1}^n Z_{ni}$ where

$$Z_{ni} = \langle g, U_i\rangle_{\mathcal{H}_Y}\,\langle V_n f, \kappa(\cdot,X_i) - \mu_X\rangle_{\mathfrak{M}_X}.$$

Note that $Z_{n1},\ldots,Z_{nn}$ are i.i.d. random variables and, since $X_i \perp\!\!\!\perp U_i$, we have $\mathrm{E}\,Z_{ni} = 0$. Hence

$$\sigma^2_{n,3} = \mathrm{E}\,[\langle f, V_n\tilde{\Sigma}_{XU}g\rangle^2_{\mathfrak{M}_X}] = n^{-1}\mathrm{E}\,(Z^2_{n1}) = n^{-1}\,\mathrm{E}\,[\langle g, U\rangle^2_{\mathcal{H}_Y}\,\langle V_n f, \kappa(\cdot,X) - \mu_X\rangle^2_{\mathfrak{M}_X}]. \tag{A31}$$

By $U \perp\!\!\!\perp X$ and (12), the right-hand side is

$$\begin{aligned}
n^{-1}\,&\mathrm{E}\,[\langle g, U\rangle^2_{\mathcal{H}_Y}]\,\mathrm{E}\,[\langle V_n f, \kappa(\cdot,X) - \mu_X\rangle^2_{\mathfrak{M}_X}] \\
&= n^{-1}\,\mathrm{E}\,[U(t)^2]\,\mathrm{E}\,[\langle V_n f, \textstyle\sum_{j\in\mathbb{N}}\zeta_j\varphi_j\rangle_{\mathfrak{M}_X})^2] \\
&= n^{-1}\,\mathrm{E}\,[U(t)^2]\,\mathrm{E}\,[(\textstyle\sum_{j\in\mathbb{N}}\zeta_j(\lambda_j + \epsilon_n)^{-1}\langle f, \varphi_j\rangle_{\mathfrak{M}_X})^2].
\end{aligned}$$

Since $\zeta_1, \zeta_2, \ldots$ are uncorrelated, we have

$$\begin{aligned}
\mathrm{E}\,[(\textstyle\sum_{j\in\mathbb{N}}\zeta_j(\lambda_j + \epsilon_n)^{-1}\langle f, \varphi_j\rangle^2_{\mathfrak{M}_X}] &= \textstyle\sum_{j\in\mathbb{N}}\mathrm{E}\,[(\zeta_j^2(\lambda_j + \epsilon_n)^{-2}\langle f, \varphi_j\rangle^2_{\mathfrak{M}_X}] \\
&= \textstyle\sum_{j\in\mathbb{N}}\lambda_j(\lambda_j + \epsilon_n)^{-2}\langle f, \varphi_j\rangle^2_{\mathfrak{M}_X}.
\end{aligned}$$

Note that for any $j \geq 1$, φ_j is a member of $\mathfrak{M}^0_X$, the effective domain of Σ_{XX}. Since $\langle f, \varphi_j\rangle_{\mathfrak{M}_X} = \varphi_j(x_0) - \mathrm{E}\,\varphi_j(X) = \varphi_j(x_0)$, we have the desired equality in part *1.*

2. As argued in the proof of Corollary 11, the last three terms on the right-hand side of (A29) are all of the parametric order $O_P(n^{-1/2})$ or smaller. Therefore we only need to consider the term

$$\langle (\hat{R}_{XY} - R_{XY})\,g, f\rangle_{\mathcal{H}_Y} = A_{n,1} + \cdots + A_{n,5}.$$

By Assumption 8, $A_{n,3}$ is the dominating term, and so we only need to derive its asymptotic distribution. Since $A_{n,3} = n^{-1}\sum_{i=1}^{n} Z_{ni}$, where $\{n^{-1}Z_{ni} : i = 1, \dots, n, n \in \mathbb{N}\}$ is a triangular array, we use Lyapounov's central limit theorem. Thus, for a $d > 0$, let

$$L_n(d) = \sigma_{n,3}^{-2-d} \sum_{i=1}^{n} \mathrm{E}\,|n^{-1}Z_{ni}|^{2+d}.$$

We need to verify $L_n(d) \to 0$ as $n \to \infty$ for some $d > 0$. Take $d = 2$. Then

$$L_n(2) = \sigma_{n,3}^{-4} \sum_{i=1}^{n} \mathrm{E}\,(|n^{-1}Z_{ni}|^4) = n^{-1}[\mathrm{E}\,(Z_{ni}^2)]^{-2}\mathrm{E}\,(Z_{ni}^4).$$

By $U \perp\!\!\!\perp X$,

$$\mathrm{E}\,(Z_{ni}^4) = \mathrm{E}\,\langle U, g\rangle_{\mathcal{H}_Y}^4 \,\mathrm{E}\,\langle V_n f, \kappa(\cdot, X) - \mu_X\rangle_{\mathfrak{M}_X}^4 \asymp \mathrm{E}\,\langle V_n f, \kappa(\cdot, X) - \mu_X\rangle_{\mathfrak{M}_X}^4.$$

Since the kernel κ is bounded, the right-hand side is upper bounded by

$$\begin{aligned} &\mathrm{E}\,(\langle V_n f, \kappa(\cdot, X) - \mu_X\rangle_{\mathfrak{M}_X}^2 \,\|V_n\|_{\mathrm{OP}}^2 \,\|f\|_{\mathfrak{M}_X}^2 \,\|\kappa(\cdot, X) - \mu_X\|_{\mathfrak{M}_X}^2) \\ &\qquad \le M\epsilon_n^{-2}\mathrm{E}\,(\langle V_n f, \kappa(\cdot, X) - \mu_X\rangle_{\mathfrak{M}_X}^2) \end{aligned}$$

for some $M > 0$. Also, recall that

$$\mathrm{E}\,(Z_{ni}^2) = \mathrm{E}\,\langle U, g\rangle_{\mathcal{H}_Y}^2 \,\mathrm{E}\,(\langle f, V_n(\kappa(\cdot, X) - \mu_X)\rangle_{\mathfrak{M}_X}^2). \tag{A32}$$

Consequently,

$$L_n(2) = \frac{O(n^{-1}\epsilon_n^{-2})}{\mathrm{E}\,(\langle f, V_n(\kappa(\cdot, X) - \mu_X)\rangle_{\mathfrak{M}_X}^2)}.$$

By expansion (12), the denominator above can be bounded below as follows:

$$\mathrm{E}\,(\langle f, V_n(\kappa(\cdot, X) - \mu_X)\rangle_{\mathfrak{M}_X}^2) = \sum_{j\in\mathbb{N}} (\lambda_j + \epsilon_n)^{-2}\lambda_j f_j^2 \ge (\lambda_1 + \epsilon_n)^{-2}\lambda_1 f_1^2 \ge (\lambda_1 + \epsilon_1)^{-2}\lambda_1 f_1^2.$$

Hence $L_n(2) = O(n^{-1}\epsilon_n^{-2})$. Since $\epsilon_n \succ n^{-1/2}$, we have $L_n(2) \to 0$. ■

Proof. [Proof of Lemma 14] *1.* By Theorem 1, recall that

$$\begin{aligned} &\widehat{M}(X_{n+1}) - \mathrm{E}\,(Y_{n+1}|X_{n+1}) \\ &= \hat{\Sigma}_{YX}\hat{V}[\kappa(\cdot, X_{n+1}) - \widehat{\mu}_X] - \Sigma_{YX}\Sigma_{XX}^{\dagger}[\kappa(\cdot, X_{n+1}) - \mu_X] + \widehat{\mu}_Y - \mu_Y \\ &= (\hat{\Sigma}_{YX}\hat{V} - \Sigma_{YX}\Sigma_{XX}^{\dagger})[\kappa(\cdot, X_{n+1}) - \mu_X] + \hat{\Sigma}_{YX}\hat{V}(\mu_X - \widehat{\mu}_X) + \widehat{\mu}_Y - \mu_Y \\ &:= D_{1n} + D_{2n} + \widehat{\mu}_Y - \mu_Y \end{aligned} \tag{A33}$$

in obvious correspondence. We study the second term first. By Lemma 6 (2.) and Lemma 7, we have

$$\begin{aligned} \|D_{2n}\| &= \|\hat{\Sigma}_{YX}(\hat{\Sigma}_{XX} + \epsilon_n I)^{-1}(\mu_X - \widehat{\mu}_X)\|_{\mathfrak{M}_X} \\ &= \|(\hat{\Sigma}_{UX} + R_{XY}^{*}\hat{\Sigma}_{XX})(\hat{\Sigma}_{XX} + \epsilon_n I)^{-1}(\mu_X - \widehat{\mu}_X)\|_{\mathfrak{M}_X} \\ &= O_P(\epsilon_n^{-1}n^{-1/2})[\|\hat{\Sigma}_{UX} - \tilde{\Sigma}_{UX}\|_{\mathrm{OP}} + \|\tilde{\Sigma}_{UX}\|_{\mathrm{OP}}] + O_P(n^{-1/2}) \\ &= O_P(\epsilon_n^{-1}n^{-1}) + O_P(n^{-1/2}) = O_P(n^{-1/2}). \end{aligned}$$

The third relation holds since $\|\tilde{\Sigma}_{UX}\|_{\mathrm{OP}} = O_P(n^{-1/2})$. For $i = 1, \dots, n+1$, let $G_i := \kappa(\cdot, X_i) - \mu_X \in \mathfrak{M}^0_X$ (P_X- a.s.), and $\tilde{G}_i = \kappa(\cdot, X_i) - \hat{\mu}_X$. Then D_{1n} can be expressed as:

$$\begin{aligned} D_{1n} &= \left(\hat{\Sigma}_{YX}\hat{V} - \Sigma_{YX}V\right) G_{n+1} \\ &= R^*_{XY}(\hat{\Sigma}_{XX}\hat{V} - \Sigma_{XX}V_n)G_{n+1} + R^*_{XY}(\Sigma_{XX}V_n - I)G_{n+1} + \frac{1}{n}\sum_{i=1}^n U_i\langle \hat{V}\tilde{G}_i, G_{n+1}\rangle_{\mathfrak{M}_X} \\ &- \frac{\bar{U}}{n}\sum_{i=1}^n \langle \hat{V}\tilde{G}_i, G_{n+1}\rangle_{\mathfrak{M}_X}. \end{aligned}$$

Since $\sum_{i=1}^n \tilde{G}_i = 0$, the last term is 0. Denote the first three terms by E_{1n}, E_{2n} and E_{3n}, respectively.

We first deal with the bias term, E_{2n}. By the Karhunen-Loève theorem, we can rewrite $G_{n+1} = \kappa(\cdot, X_{n+1}) - \mu_X$ as $\sum_{j=1}^\infty \zeta_j\varphi_j$, where $\zeta_1, \zeta_2, \cdots$ are mean 0, uncorrelated random variables with variance $\lambda_1, \lambda_2, \cdots$. We now derive the form of $\mathrm{E}\,\|E_{2n}\|^2_{\mathcal{H}_Y}$ to determine stochastic order of E_{2n}. Let $\{\psi_k : k \in \mathbb{N}\}$ be an ONB of $\mathcal{H}_Y$. Under Assumption 6(*ii*), we have $\langle R^*_{XY}\varphi_j, \psi_k\rangle_{\mathcal{H}_Y} = \langle \Sigma^\beta_{XX}\varphi_j, S_{XY}\psi_k\rangle_{\mathfrak{M}_X} = \lambda_j^\beta s_{jk}$, where $s_{jk} = \langle \varphi_j, S_{XY}\psi_k\rangle_{\mathfrak{M}_X}$. Since S_{XY} is a Hilbert-Schmidt operator, $\sum_j\sum_k s^2_{jk} < \infty$ holds, and hence

$$\begin{aligned} \mathrm{E}\,\|E_{2n}\|^2_{\mathcal{H}_Y} &= \mathrm{E}\left\{\|R^*_{XY}\left[\Sigma_{XX}(\Sigma_{XX} + \epsilon_n I)^{-1} - I\right]G_{n+1}\|^2_{\mathcal{H}_Y}\right\} \\ &= \mathrm{E}\left\{\left\|R^*_{XY}\sum_{j=1}^\infty \frac{\epsilon_n\zeta_j}{\epsilon_n + \lambda_j}\varphi_j\right\|^2_{\mathcal{H}_Y}\right\} \\ &= \mathrm{E}\left\{\sum_{k=1}^\infty \left\langle R^*_{XY}\sum_{j=1}^\infty \frac{\epsilon_n\zeta_j}{\epsilon_n + \lambda_j}\varphi_j, \psi_k\right\rangle^2_{\mathcal{H}_Y}\right\} \\ &= \epsilon_n^2\sum_{j=1}^\infty \frac{\lambda_j^{2\beta+1}}{(\epsilon_n + \lambda_j)^2}\sum_{k=1}^\infty s^2_{jk}. \end{aligned}$$

By straightforward calculation, we can verify that, when $\beta \geq 1/2$, $\sup_{j\geq 1}\frac{\lambda_j^{2\beta+1}}{(\epsilon_n+\lambda_j)^2} = O(1)$ as $n \to \infty$ regardless of the decaying rate of ϵ_n. As a result, $E_{2n} = O_P(\epsilon_n)$.

Next, we consider E_{1n}. Denote $\hat{\Sigma}_{XX} - \Sigma_{XX}$ by Δ. Since $\hat{\Sigma}_{XX}\hat{V} - \Sigma_{XX}V_n = \epsilon_n\hat{V}\Delta V_n$, we have

$$E_{1n} = \epsilon_n R^*_{XY}\hat{V}\Delta V_n G_{n+1} = (\epsilon_n S_{YX}\Sigma^\beta_{XX}\hat{V}\Delta)(V_n G_{n+1}) := E_{11n} \times E_{12n}.$$

By Lemma 5, $\|\Delta\| = O_P(n^{-1/2})$. If we assume that $\beta \geq 1$,

$$\begin{aligned} \|E_{11n}\| &\leq \epsilon_n\|S_{YX}\|_{\mathrm{OP}} \times \|\Sigma^{\beta-1}_{XX}\Sigma_{XX}(\hat{\Sigma}_{XX} + \epsilon_n I)^{-1}\|_{\mathrm{OP}} \times \|\Delta\|_{\mathrm{OP}} \\ &= O_P(n^{-1/2}\epsilon_n) \times \|\Sigma^{\beta-1}_{XX}\| \times \|[(\Sigma_{XX} - \hat{\Sigma}_{XX}) + \hat{\Sigma}_{XX}](\hat{\Sigma}_{XX} + \epsilon_n I)^{-1}\|_{\mathrm{OP}} \\ &= O_P(n^{-1/2}\epsilon_n)O_P(n^{-1/2}\epsilon_n^{-1} + 1) = O_P(n^{-1/2}\epsilon_n), \end{aligned}$$

where the last relation holds since $\epsilon_n \succ n^{-1/2}$. By Lemma 8, we have

$$\mathrm{E}\left[\|E_{12n}\|^2_{\mathfrak{M}_X}\right] = \sum_{j=1}^\infty \frac{\lambda_j}{(\epsilon_n + \lambda_j)^2} = O(\epsilon_n^{-(\alpha+1)/\alpha}).$$

It follows that $E_{1n} = O_P(n^{-1/2}\epsilon_n) \times O_P(\epsilon_n^{-(\alpha+1)/(2\alpha)}) = O_P(n^{-1/2}\epsilon_n^{(\alpha-1)/(2\alpha)})$.

Lastly, we consider

$$E_{3n} = n^{-1}\sum_{i=1}^{n} U_i \langle \hat{V}\tilde{G}_i, G_{n+1}\rangle_{\mathfrak{M}_X} = \sum_{i=1}^{n} Z_{i,n} + \bar{U}\langle \hat{V}(\mu_X - \hat{\mu}_X), G_{n+1}\rangle_{\mathfrak{M}_X},$$

where $Z_{i,n} = \frac{1}{n}U_i\langle \hat{V}G_i, G_{n+1}\rangle_{\mathfrak{M}_X}$. The remainder term satisfies that

$$\|\bar{U}\langle \hat{V}(\mu_X - \hat{\mu}_X), G_{n+1}\rangle_{\mathfrak{M}_X}\|_{\mathcal{H}_Y} \leq \|\bar{U}\|_{\mathcal{H}_Y} \cdot \|\mu_X - \hat{\mu}_X\|_{\mathfrak{M}_X} \cdot \Big\{\|V_n G_{n+1}\|_{\mathfrak{M}_X} + \|(\hat{V} - V_n)G_{n+1}\|_{\mathfrak{M}_X}\Big\}.$$

Based our previous calculations of $\mathrm{E}\,(\|E_{12n}\|^2)$, we have $\|V_n G_{n+1}\|_{\mathfrak{M}_X} = O_P(\epsilon_n^{-(\alpha+1)/(2\alpha)})$. Note that $\hat{V} - V_n = -\hat{V}\Delta V_n$. Hence

$$\begin{aligned} ||(\hat{V} - V_n)G_{n+1}\|_{\mathfrak{M}_X} &= \|\hat{V}\Delta V_n G_{n+1}\| \\ &\leq \|\hat{V}\Delta\|\|V_n G_{n+1}\| \\ &= O_P(\epsilon_n^{-1}n^{-1/2})O_P(\epsilon_n^{-(\alpha+1)/(2\alpha)}) \\ &= o_P(\epsilon_n^{-(\alpha+1)/(2\alpha)}), \end{aligned}$$

where the last equality holds because $n^{-1/2} \prec \epsilon_n$. It follows that the remainder is $O_P(n^{-1}\epsilon_n^{-(\alpha+1)/(2\alpha)})$.

2. By definition, we have

$$\begin{aligned} \|W_n\|^2_{\mathcal{H}_Y} &= \frac{1}{n^2}\sum_{i=1}^{n}\|U_i\|^2_{\mathcal{H}_Y}\langle \hat{V}G_i, G_{n+1}\rangle^2_{\mathfrak{M}_X} \\ &\quad + \frac{1}{n^2}\sum_{i\neq i'}^{n}\langle U_i, U_{i'}\rangle_{\mathcal{H}_Y}\langle \hat{V}G_i, G_{n+1}\rangle_{\mathfrak{M}_X}\langle \hat{V}G_{i'}, G_{n+1}\rangle_{\mathfrak{M}_X}. \end{aligned}$$

By $(U_1, \ldots, U_n) \perp\!\!\!\perp (X_1, \ldots, X_n)$ and $U_i \perp\!\!\!\perp U_{i'}$ if $i \neq i'$, the expectation of the second term is 0. Moreover, $\mathrm{E}\,\|W_n\|^2_{\mathcal{H}_Y} = n^{-1}\mathrm{E}\,[\|U_1\|^2_{\mathcal{H}_Y}]\mathrm{E}\,[\langle \hat{V}G_i, G_{n+1}\rangle^2_{\mathfrak{M}_X}] = n^{-1}\sigma^2_U\mathrm{E}\,[\langle \hat{V}G_i, G_{n+1}\rangle^2_{\mathfrak{M}_X}]$, where $\sigma^2_U = \mathrm{trace}(\Sigma_{UU}) < \infty$ since $\mathrm{E}\,\|Y\|^2_{\mathcal{H}_Y} < \infty$. By the law of iterated expectations,

$$\begin{aligned} \mathrm{E}\,[\langle \hat{V}G_i, G_{n+1}\rangle^2_{\mathfrak{M}_X}] &= \mathrm{E}\,\Big[\mathrm{E}\,\Big(\langle \hat{V}G_i, G_{n+1}\rangle^2_{\mathfrak{M}_X} \mid X_1, \ldots, X_n\Big)\Big] \\ &= \mathrm{E}\,\Big[\mathrm{E}\,\Big\{\langle \hat{V}G_i, (G_{n+1} \otimes G_{n+1})\hat{V}G_i\rangle_{\mathfrak{M}_X} \mid X_1, \ldots, X_n\Big\}\Big] \\ &= \mathrm{E}\,[\langle \hat{V}\Sigma_{XX}\hat{V}G_i, G_i\rangle_{\mathfrak{M}_X}] \\ &= \mathrm{E}\,[\mathrm{trace}(\hat{V}\Sigma_{XX}\hat{V}(G_i \otimes G_i))], \end{aligned}$$

where the third equality holds because $\mathrm{E}\,(G_{n+1} \otimes G_{n+1}|X_1, \ldots, X_n) = \mathrm{E}\,(G_{n+1} \otimes G_{n+1}) = \Sigma_{XX}$, and the last equation holds because $\mathrm{trace}(A(g \otimes h)) = \langle Ag, h\rangle_{\mathcal{H}}$ for any $g, h \in \mathcal{H}$ and $A \in \mathcal{B}(\mathcal{H})$.

Since $\hat{V}\Sigma_{XX}\hat{V}(G_1 \otimes G_1), \ldots, \hat{V}\Sigma_{XX}\hat{V}(G_n \otimes G_n)$ have the same distribution, they have the same expectation. Let $\tilde{\Sigma}_{XX} = n^{-1}\sum_{i=1}^{n}(\kappa(\cdot, X_i) - \mu_X) \otimes (\kappa(\cdot, X_i) - \mu_X)$. Hence

$$\mathrm{E}\,[\mathrm{trace}(\hat{V}\Sigma_{XX}\hat{V}(G_i \otimes G_i))] = \mathrm{E}\,[\mathrm{trace}(\hat{V}\Sigma_{XX}\hat{V}\tilde{\Sigma}_{XX})].$$

By the properties of the trace of linear operators, we have

$$\mathrm{E}\,[\mathrm{trace}(\hat{V}\Sigma_{XX}\hat{V}\tilde{\Sigma}_{XX})] = \mathrm{trace}(\Sigma_{XX}\mathrm{E}\,(\hat{V}\tilde{\Sigma}_{XX}\hat{V})) \leq \mathrm{trace}(\Sigma_{XX})\|\mathrm{E}\,(\hat{V}\tilde{\Sigma}_{XX}\hat{V})\|_{\mathrm{OP}}.$$

Rewriting $\hat{V}\tilde{\Sigma}_{XX}\hat{V}$ as $\hat{V}[(\tilde{\Sigma}_{XX} - \hat{\Sigma}_{XX}) + \hat{\Sigma}_{XX}]\hat{V}$, we have

$$\begin{aligned} \|\mathrm{E}\,(\hat{V}\tilde{\Sigma}_{XX}\hat{V})\|_{\mathrm{OP}} &\leq \mathrm{E}\,(\|\hat{V}\tilde{\Sigma}_{XX}\hat{V}\|_{\mathrm{OP}}) \leq \mathrm{E}\,(\|\hat{V}\|_{\mathrm{OP}}^2\|\tilde{\Sigma}_{XX} - \hat{\Sigma}_{XX}\|_{\mathrm{OP}}) + \mathrm{E}\,(\|\hat{V}\|_{\mathrm{OP}} \cdot \|\hat{\Sigma}_{XX}\hat{V}\|_{\mathrm{OP}}) \\ &\leq C\epsilon_n^{-2}n^{-1} + \epsilon_n^{-1} \end{aligned}$$

for some constant $C > 0$. The last inequality holds since $\mathrm{E}\,(\|\tilde{\Sigma}_{XX} - \hat{\Sigma}_{XX}\|_{\mathrm{OP}}) \leq Cn^{-1}$ by the proof of Lemma 5 in Fukumizu et al. (2007). Since $\epsilon_n \succ n^{-1/2}$, we have $\epsilon_n^{-2}n^{-1} \prec \epsilon_n^{-1}$, proving Part *2*. ■

Proof. [Proof of Theorem 16] As argued following the proof of Lemma 14,

$$\widehat{M}(X_{n+1}) - \mathrm{E}\,(Y_{n+1}|X_{n+1}) = F_{1n} + F_{2n} + W_n + H_n + O_P(n^{-1/2}).$$

Since, by Assumption 9, W_n is the dominating term. We focus on the weak convergence of W_n in $\mathcal{H}_Y$.

We first show the finite-dimensional convergence of W_n; that is, for any deterministic $y \in \mathcal{H}_Y$,

$$s_n^{-1}\langle W_n, y\rangle_{\mathcal{H}_Y} \xrightarrow{\mathcal{D}} N(0, \sigma_{U,y}^2), \tag{A34}$$

where $\sigma_{U,y} = \langle y, \Sigma_{UU}y\rangle_{\mathcal{H}_Y}$. Let $\mathcal{F}_i$ denote the σ-algebra generated by $\{X_1, U_1, \ldots, X_i, U_i\}$ (or equivalently by $\{X_1, Y_1, \ldots, X_i, Y_i\}$). Let $H_i(y) := \langle Z_{i,n}, y\rangle_{\mathcal{H}_Y}$. Obviously $\mathrm{E}\,[H_i(y)]$ is 0, and $H_i(y)$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_i$. To find its variance, we employ the law of iterated expectations:

$$\begin{aligned} \mathrm{E}\,\{H_i^2(y)|\mathcal{F}_i\} &= \mathrm{E}\,\{\langle Z_{i,n}, y\rangle_{\mathcal{H}_Y}^2|\mathcal{F}_i\} \\ &= \mathrm{E}\,\left\{\left\langle \frac{1}{n}U_i\langle\hat{V}G_i, G_{n+1}\rangle_{\mathfrak{M}_X}, y\right\rangle^2 |\mathcal{F}_i\right\} \\ &= \frac{1}{n^2}\mathrm{E}\,\left\{\langle\hat{V}G_i, G_{n+1}\rangle_{\mathfrak{M}_X}^2\langle U_i, y\rangle^2|\mathcal{F}_i\right\} \\ &= \frac{\langle y, U_i\rangle_{\mathcal{H}_Y}^2}{n^2}\mathrm{E}\,\left\{\langle\hat{V}G_i, G_{n+1}\rangle_{\mathfrak{M}_X}\langle\hat{V}G_i, G_{n+1}\rangle_{\mathfrak{M}_X}|\mathcal{F}_i\right\} \\ &= \frac{\langle y, U_i\rangle_{\mathcal{H}_Y}^2}{n^2}\mathrm{E}\,\left\{\langle\hat{V}G_i, (G_{n+1}\otimes G_{n+1})\hat{V}G_i\rangle_{\mathfrak{M}_X}|\mathcal{F}_i\right\} \\ &= \frac{\langle y, U_i\rangle_{\mathcal{H}_Y}^2}{n^2}\langle\hat{V}G_i, \Sigma_{XX}\hat{V}G_i\rangle_{\mathfrak{M}_X} \\ &= \frac{\langle y, U_i\rangle_{\mathcal{H}_Y}^2}{n^2}\|\Sigma_{XX}^{1/2}\hat{V}G_i\|_{\mathfrak{M}_X}^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{E}\,\{H_i^2(y)\} &= \mathrm{E}\,[\mathrm{E}\,\{H_i^2(y)|\mathcal{F}_i\}] \\ &= \frac{1}{n^2}\mathrm{E}\,\left[\|\Sigma_{XX}^{1/2}\hat{V}G_i\|^2\mathrm{E}\,\left\{\langle y, U_i\rangle_{\mathcal{H}_Y}^2|X_1, \ldots, X_i\right\}\right] \\ &= \frac{\sigma_{U,y}^2}{n^2}\mathrm{E}\,\left(\|\Sigma_{XX}^{1/2}\hat{V}G_i\|_{\mathfrak{M}_X}^2\right) = \frac{\sigma_{U,y}^2 s_n^2}{n}, \end{aligned}$$

where, for the third equality, we used $U_i \perp\!\!\!\perp (X_1, \ldots, X_n)$. The convergence in (A34) then follows from the central limit theorem for martingale difference arrays in McLeish (1974).

Next, we show that the sequence $s_n^{-1} W_n$ is asymptotically tight. By Lemma 1.8.1 of Van Der Vaart and Wellner (1996), it suffices to show that for any $\eta > 0$,

$$\limsup_{J\to\infty} \limsup_{n\to\infty} \Pr\left(\sum_{j>J} \langle s_n^{-1} W_n, \psi_j \rangle_{\mathcal{H}_Y}^2 > \eta \right) = 0, \tag{A35}$$

where $\{\psi_j : j \in \mathbb{N}\}$ is any ONB of $\mathcal{H}_Y$. For any $j \in \mathbb{N}$, we have

$$\begin{aligned}
\mathrm{E}\left(\langle s_n^{-1} W_n, \psi_j \rangle_{\mathcal{H}_Y}^2 \right) &= \mathrm{E}\left(\left\langle \frac{s_n^{-1}}{n} \sum_{i=1}^n U_i \langle \hat{V} G_i, G_{n+1} \rangle_{\mathfrak{M}_X}, \psi_j \right\rangle^2 \right) \\
&= \frac{s_n^{-2}}{n^2} \mathrm{E}\left\{ \left(\sum_{i=1}^n \langle \hat{V} G_i, G_{n+1} \rangle_{\mathfrak{M}_X} \langle U_i, \psi_j \rangle \right)^2 \right\} \\
&= \frac{s_n^{-2}}{n^2} \sum_{i=1}^n \sum_{i'=1}^n \mathrm{E}\left\{ \langle \hat{V} G_i, G_{n+1} \rangle_{\mathfrak{M}_X} \langle \hat{V} G_{i'}, G_{n+1} \rangle_{\mathfrak{M}_X} \langle U_i, \psi_j \rangle \langle U_{i'}, \psi_j \rangle \right\} \\
&= \frac{s_n^{-2}}{n^2} \sum_{i=1}^n \sum_{i'=1}^n \mathrm{E}\left\{ \langle \hat{V} G_i, G_{n+1} \rangle_{\mathfrak{M}_X} \langle \hat{V} G_{i'}, G_{n+1} \rangle_{\mathfrak{M}_X} \right\} \mathrm{E}\left\{ \langle U_i, \psi_j \rangle \langle U_{i'}, \psi_j \rangle \right\} \\
&= \frac{s_n^{-2}}{n^2} \sum_{i=1}^n \mathrm{E}\left\{ \langle \hat{V} G_i, G_{n+1} \rangle_{\mathfrak{M}_X}^2 \right\} \mathrm{E}\left\{ \langle U_i, \psi_j \rangle^2 \right\} \\
&= \mathrm{E}\left\{ \langle U_i, \psi_j \rangle^2 \right\},
\end{aligned}$$

where, for the fourth and fifth equalities, we used $(U_1, \ldots, U_n) \perp\!\!\!\perp (X_1, \ldots, X_n)$, $U_i \perp\!\!\!\perp U_{i'}$ if $i \neq i'$ and $\mathrm{E}(U_i) = 0$. Therefore, as $J \to \infty$,

$$\mathrm{E}\left(\sum_{j>J} \langle s_n^{-1} W_n, \psi_j \rangle_{\mathcal{H}_Y}^2 \right) = \mathrm{E} \sum_{j>J} \langle U, \psi_j \rangle_{\mathcal{H}_Y}^2 \to 0,$$

by the dominated convergence theorem, because the right-hand side is bounded by $\|U\|_{\mathcal{H}_Y}^2$, which has a finite expectation. Thus the sequence $s_n^{-1} W_n$ is asymptotically tight.

Combining the above results, we obtain (17). ■

Proof. [Proof of Lemma 19]

First, we show

$$\|\hat{\mu}_X - \hat{\mu}_{\hat{X}}\|_{\mathfrak{M}_X} = O_P\left(m^{-(a-1)} + \frac{m^{(4a+2c+7)/5}}{(nN)^{4/5}} \right). \tag{A36}$$

Under Assumption 10 , Corollary 3 yields

$$\begin{aligned}
&\|\hat{\mu}_X - \hat{\mu}_{\hat{X}}\|_{\mathfrak{M}_X} \\
&= \|n^{-1}\sum_{i=1}^{n}\kappa(X_i,\cdot) - n^{-1}\sum_{i=1}^{n}\kappa(\hat{X}_i,\cdot)\|_{\mathfrak{M}_X} \\
&\le n^{-1}\sum_{i=1}^{n}\|\kappa(X_i,\cdot) - \kappa(\hat{X}_i,\cdot)\|_{\mathfrak{M}_X} \\
&\le n^{-1}\sum_{i=1}^{n} L\|X_i - \hat{X}_i\| \\
&= O_P\left(m^{-(a-1)} + \frac{m^{(4a+2c+7)/5}}{(nN)^{4/5}}\right).
\end{aligned}$$

Let $F_i = \kappa(X_i,\cdot) - \hat{\mu}_X$ and $\widehat{F}_i = \kappa(\hat{X}_i,\cdot) - \widehat{\mu}_{\hat{X}}$ for $i = 1,\ldots,n$. Define $\mathcal{F} = \mathfrak{M}_X \otimes \mathfrak{M}_X$. By Lemma 4 of Fukumizu et al. (2007), one has

$$\|\hat{\Sigma}_{XX}\|_{\mathrm{HS}}^2 = \left\|\frac{1}{n}\sum_{i=1}^{n}F_i\otimes F_i\right\|_{\mathcal{F}}^2.$$

Similar arguments yield

$$\begin{aligned}
\|\hat{\Sigma}_{\hat{X}\hat{X}}\|_{\mathrm{HS}}^2 &= \left\|\frac{1}{n}\sum_{i=1}^{n}\hat{F}_i\otimes \hat{F}_i\right\|_{\mathcal{F}}^2, \\
\langle\hat{\Sigma}_{XX},\hat{\Sigma}_{\hat{X}\hat{X}}\rangle_{\mathrm{HS}} &= \left\langle\frac{1}{n}\sum_{i=1}^{n}F_i\otimes F_i, \frac{1}{n}\sum_{i=1}^{n}\hat{F}_i\otimes \hat{F}_i\right\rangle_{\mathcal{F}}.
\end{aligned}$$

Then it follows that

$$\begin{aligned}
&\|\hat{\Sigma}_{XX} - \hat{\Sigma}_{\hat{X}\hat{X}}\|_{\mathrm{HS}}^2 \\
&= \|\hat{\Sigma}_{XX}\|_{\mathrm{HS}}^2 - 2\langle\hat{\Sigma}_{XX}\hat{\Sigma}_{\hat{X}\hat{X}}\rangle_{\mathrm{HS}} + \|\hat{\Sigma}_{\hat{X}\hat{X}}\|_{\mathrm{HS}}^2 \\
&= \left\|n^{-1}\sum_{i=1}^{n}F_i\otimes F_i - n^{-1}\sum_{i=1}^{n}\hat{F}_i\otimes\hat{F}_i\right\|_{\mathcal{F}}^2 \\
&\le 2\left\|n^{-1}\sum_{i=1}^{n}F_i\otimes F_i - n^{-1}\sum_{i=1}^{n}F_i\otimes\hat{F}_i\right\|_{\mathcal{F}}^2 + 2\left\|n^{-1}\sum_{i=1}^{n}F_i\otimes \hat{F}_i - n^{-1}\sum_{i=1}^{n}\hat{F}_i\otimes\hat{F}_i\right\|_{\mathcal{F}}^2 \\
&= 2n^{-1}\sum_{i=1}^{n}\|F_i\otimes(F_i-\hat{F}_i)\|_{\mathcal{F}}^2 + 2n^{-1}\sum_{i=1}^{2}\|(F_i-\hat{F}_i)\otimes\hat{F}_i\|_{\mathcal{F}}^2 \\
&= 2n^{-1}\sum_{i=1}^{n}\left\{\left(\|F_i\|_{\mathfrak{M}_X}^2 + \|\hat{F}_i\|_{\mathfrak{M}_X}^2\right)\|F_i-\hat{F}_i\|_{\mathfrak{M}_X}^2\right\}
\end{aligned}$$

With similar arguments to verify (A36), one can show that

$$\|F_i - \hat{F}_i\|_{\mathfrak{M}_X}^2 = O_P\left(m^{-2(a-1)} + \frac{m^{(8a+4c+14)/5}}{(nN)^{8/5}}\right)$$

for $i = 1, \dots, n$. By rewriting $\hat{F}_i$ as $\hat{F}_i - F_i + F_i$, we obtain

$$\begin{aligned}
&\|\hat{\Sigma}_{XX} - \hat{\Sigma}_{\hat{X}\hat{X}}\|_{\mathrm{HS}}^2 \\
&\leq \left(6n^{-1}\sum_{i=1}^{n}\|F_i\|_{\mathfrak{M}_X}^2\right) \times O_P\left(m^{-2(a-1)} + \frac{m^{(8a+4c+14)/5}}{(nN)^{8/5}}\right) + O_P\left(m_X^{-4(a-1)} + \frac{m^{(16a+8c+28)/5}}{(nN)^{16/5}}\right).
\end{aligned}$$

From the proof of Lemma 5 of Fukumizu et al. (2007), one has

$$n^{-1}\sum_{i=1}^{n}\|F_i\|_{\mathfrak{M}_X}^2 = O_P(1)$$

under Assumption 6(i). As $n^{-1/2} \preceq m_X^{-(a-1)}$, we have

$$\|\hat{\Sigma}_{\hat{X}\hat{X}} - \Sigma_{XX}\|_{\mathrm{HS}} = O_P\left(m^{-(a-1)} + \frac{m^{(4a+2c+7)/5}}{(nN)^{4/5}}\right).$$

A similar argument can be employed to show the second equality in the corollary. The proof is completed. ■

Proof. [Proof of Theorem 20] (1). Since we have found the convergence rate of $\hat{R}_{XY}$ in Theorem 9, we only need to study the discrepancy between $\hat{R}_{\hat{X}\hat{Y}}$ and $\hat{R}_{XY}$.

We still abbreviate $(\hat{\Sigma}_{XX} + \epsilon_n I)^{-1}$, $(\Sigma_{XX} + \epsilon_n I)^{-1}$, and $\Sigma_{XX}^{\dagger}$ by $\hat{V}$, V_n and V, respectively. Note that

$$\begin{aligned}
&(\hat{\Sigma}_{\hat{X}\hat{X}} + \epsilon_n I)^{-1} - \hat{V} \\
&= (\hat{\Sigma}_{\hat{X}\hat{X}} + \epsilon_n I)^{-1}(\hat{\Sigma}_{XX} + \epsilon_n I)\hat{V} - (\hat{\Sigma}_{\hat{X}\hat{X}} + \epsilon_n I)^{-1}(\hat{\Sigma}_{\hat{X}\hat{X}} + \epsilon_n I)\hat{V} \\
&= (\hat{\Sigma}_{\hat{X}\hat{X}} + \epsilon_n I)^{-1}(\hat{\Sigma}_{XX} - \hat{\Sigma}_{\hat{X}\hat{X}})\hat{V},
\end{aligned} \tag{A37}$$

and

$$\hat{R}_{\hat{X}\hat{Y}} - \hat{R}_{XY} = \{(\hat{\Sigma}_{\hat{X}\hat{X}} + \epsilon_n I)^{-1} - \hat{V}\}\hat{\Sigma}_{\hat{X}\hat{Y}} + \hat{V}(\hat{\Sigma}_{\hat{X}\hat{Y}} - \hat{\Sigma}_{XY}).$$

Plugging (A37) into the above equation, it follows that

$$\hat{R}_{\hat{X}\hat{Y}} - \hat{R}_{XY} = (\hat{\Sigma}_{\hat{X}\hat{X}} + \epsilon_n I)^{-1}(\hat{\Sigma}_{XX} - \hat{\Sigma}_{\hat{X}\hat{X}})\hat{V}\hat{\Sigma}_{\hat{X}\hat{Y}} + \hat{V}(\hat{\Sigma}_{\hat{X}\hat{Y}} - \hat{\Sigma}_{XY}). \tag{A38}$$

We deal with the second term on the right-hand side of (A38) first. Since $\|\hat{V}\|_{\mathrm{OP}} \leq \epsilon_n^{-1}\|I\|_{\mathrm{OP}} = \epsilon_n^{-1}$, by Lemma 19, we have

$$\|\hat{V}(\hat{\Sigma}_{\hat{X}\hat{Y}} - \hat{\Sigma}_{XY})\|_{\mathrm{OP}} = O_P(\epsilon_n^{-1}b_n). \tag{A39}$$

Next we deal with the first term on the right-hand side of (A38). Write $\hat{V}$ and $\hat{\Sigma}_{\hat{X}\hat{Y}}$ as $\hat{V} - V_n + V_n$ and $\hat{\Sigma}_{\hat{X}\hat{Y}} - \hat{\Sigma}_{XY} + \hat{\Sigma}_{XY}$, respectively. Then

$$\begin{aligned}
\hat{V}\hat{\Sigma}_{\hat{X}\hat{Y}} &= (\hat{V} - V_n + V_n)(\hat{\Sigma}_{\hat{X}\hat{Y}} - \hat{\Sigma}_{XY} + \hat{\Sigma}_{XY}) \\
&= (\hat{V} - V_n)(\hat{\Sigma}_{\hat{X}\hat{Y}} - \hat{\Sigma}_{XY}) + (\hat{V} - V_n)\hat{\Sigma}_{XY} \\
&\quad + V_n(\hat{\Sigma}_{\hat{X}\hat{Y}} - \hat{\Sigma}_{XY}) + V_n\hat{\Sigma}_{XY} \\
&:= T_1 + T_2 + T_3 + T_4.
\end{aligned}$$

As shown in the proof of Theorem 9,

$$\hat{V} - V_n = \hat{V}(\Sigma_{XX} - \hat{\Sigma}_{XX})V_n.$$

Consequently, $\|T_1\|_{\mathrm{OP}} = O_P(\epsilon_n^{-2}n^{-1/2}b_n) = O_P(\epsilon_n^{-1}b_n)$ if $\epsilon_n \succ n^{-1/2}$. From equations (A23) and (A27) in the proof of Theorem 9, we have

$$\|T_2\|_{\mathrm{OP}} = O_P(n^{-1/2}\epsilon_n^{-1}). \tag{A40}$$

It is obvious that $\|T_3\|_{\mathrm{OP}} = O_P(\epsilon_n^{-1}b_n)$. For T_4, we have

$$\begin{aligned} T_4 &= V_n\hat{\Sigma}_{XY} \\ &= V_n(\hat{\Sigma}_{XY} - \Sigma_{XY} + \Sigma_{XY}) \\ &= V_n(\hat{\Sigma}_{XY} - \Sigma_{XY}) + V_n\Sigma_{XX}R_{XY}. \end{aligned}$$

Therefore, $\|T_4\|_{\mathrm{OP}} = O_P(1)$ given $\epsilon_n \succ n^{-1/2}$. Note that since $m^{-(a-1)} \succ n^{-(a-1)/(2a+2)} \succ n^{-1/2}$, $b_n \succ n^{-1/2}$. Combining the above results,

$$\|\hat{V}\hat{\Sigma}_{\hat{X}\hat{Y}}\|_{\mathrm{OP}} = O_P(\epsilon_n^{-1}b_n + n^{-1}\epsilon_n^{-(3\alpha+1)/(2\alpha)} + 1) = O_P(\epsilon_n^{-1}b_n + 1), \tag{A41}$$

where the last equality holds since $n^{-1/2} \prec \epsilon_n$.

Plugging (A41) into (A38), we obtain the desirable equation (19).

(2). If $\max(b_n, (b_n n^{-1})^{2\alpha/(5\alpha+1)}) \prec \epsilon_n \prec 1$, then $\epsilon_n^{-2}b_n^2 \prec \epsilon_n^{-1}b_n$ and $n^{-1}\epsilon_n^{-(5\alpha+1)/(2\alpha)}b_n \prec \epsilon_n^{-1}b_n$, which lead to the desirable equation. ■

Proof. [Proof of Corollary 21] Using the same arguments for proving Theorem 2, we can easily show that

$$\|\hat{x}_0 - x_0\| = O_P(b_n). \tag{A42}$$

Moreover, by equations (14) and (20), we have for any $s \in [0,1]$,

$$\begin{aligned} &\widehat{\mathrm{E}}\,(Y|\hat{x}_0)(s) - \mathrm{E}\,(Y|x_0)(s) \\ &= \langle \upsilon(\cdot,s), [\hat{R}^*_{\hat{X}\hat{Y}}(\kappa(\cdot,\hat{x}_0) - \hat{\mu}_{\hat{X}}) - R^*_{XY}(\kappa(\cdot,x_0) - \mu_X)]\rangle_{\mathcal{H}_Y} + \hat{\mu}_{\hat{Y}}(s) - \mu_Y(s) \\ &= \langle (\hat{R}_{\hat{X}\hat{Y}} - R_{XY})\,\upsilon(\cdot,s), \kappa(\cdot,x_0) - \mu_X\rangle_{\mathfrak{M}_X} + \langle \hat{R}_{\hat{X}\hat{Y}}\,\upsilon(\cdot,t), \mu_X - \hat{\mu}_{\hat{X}}\rangle_{\mathfrak{M}_X} \\ &\quad + \langle \hat{R}_{\hat{X}\hat{Y}}\,\upsilon(\cdot,t), \kappa(\cdot,\hat{x}_0) - \kappa(\cdot,x_0)\rangle_{\mathfrak{M}_X} + [\hat{\mu}_{\hat{Y}}(s) - \mu_Y(s)]. \end{aligned}$$

Hence

$$\begin{aligned} &|\widehat{\mathrm{E}}\,(Y|\hat{x}_0)(s) - \mathrm{E}\,(Y|x_0)(s)| \\ &\le \|\hat{R}_{\hat{X}\hat{Y}} - R_{XY}\|_{\mathrm{OP}}\,\|\upsilon(\cdot,s)\|_{\mathcal{H}_Y}\,\|\kappa(\cdot,x_0) - \mu_X\|_{\mathfrak{M}_X} \\ &\quad + \|\hat{R}_{\hat{X}\hat{Y}} - R_{XY} + R_{XY}\|_{\mathrm{OP}}\,\|\upsilon(\cdot,s)\|_{\mathcal{H}_Y}\,\|\mu_X - \hat{\mu}_{\hat{X}}\|_{\mathfrak{M}_X} \\ &\quad + \|\hat{R}_{\hat{X}\hat{Y}} - R_{XY} + R_{XY}\|_{\mathrm{OP}}\,\|\upsilon(\cdot,s)\|_{\mathcal{H}_Y}\,\|\kappa(\cdot,\hat{x}_0) - \kappa(\cdot,x_0)\|_{\mathfrak{M}_X} + |\hat{\mu}_{\hat{Y}}(s) - \mu_Y(s)|. \end{aligned} \tag{A43}$$

From previous proofs, we know

$$\begin{aligned} &\|\mu_X - \hat{\mu}_{\hat{X}}\|_{\mathfrak{M}_X} = O_P(b_n), \quad \hat{\mu}_{\hat{Y}}(s) - \mu_Y(s) = \langle \upsilon(\cdot,s), \hat{\mu}_{\hat{Y}} - \mu_Y\rangle = O_P(b_n), \\ &\|\kappa(\cdot,\hat{x}_0) - \kappa(\cdot,x_0)\|_{\mathfrak{M}_X} \le L\|\hat{x}_0 - x_0\| = O_P(b_n), \end{aligned}$$

where the second equality holds by the Cauchy inequality and the last one by Assumption 10. Hence the right-hand side of (A43) is dominated by $\|\hat{R}_{\hat{X}\hat{Y}} - R_{XY}\|_{\text{OP}}$, which proves (21).

■

Appendix A.3. Justification of Assumption 4

Consider the following linear function-on-function regression:

$$Y(t) = \int_{\mathbb{I}} \beta(s,t)X(s)ds + \epsilon(t), \quad t \in \mathbb{I},$$

where $X(s)$ and $Y(t)$ are centered random function defined on the interval $\mathbb{I}$. We further assume that $X(s) = \sum_{j=1}^{\infty} \zeta_j \psi_j(s)$, where ζ_j's are uncorrelated random variables with mean 0 and variance ρ_j, and $\{\psi_j : j = 1, 2, \ldots\}$ are an orthonormal basis of $L^2(\mathbb{I})$. For the response Y, we make similar assumptions: $Y(t) = \sum_{j=1}^{\infty} \xi_j \phi_j(t)$, where ξ_j's are uncorrelated random variables with mean 0 and variance λ_j, and $\{\phi_j : j = 1, 2, \ldots\}$ are an orthonormal basis of $L^2(\mathbb{I})$. This model was considered in Yao et al. (2005b). Under mild conditions, we have

$$\beta(s,t) = \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \frac{E(\zeta_m \xi_k)}{E(\zeta_m^2)} \psi_m(s)\phi_k(t). \tag{A44}$$

By Lemma A.2 of Yao et al. (2005b), the L^2 convergence of the right-hand side of (A44) is ensured by $\sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \sigma_{km}^2/\rho_m^2 < \infty$, where $\sigma_{km} = E(\zeta_m \xi_k)$ and $\rho_m = E(\zeta_m^2)$.

Furthermore, $\Sigma_{XY} = E(X \otimes Y)$ is a bounded linear operator from $L^2(\mathbb{I})$ to $L^2(\mathbb{I})$ satisfying

$$(\Sigma_{XY} f)(s) = E\left\{X(s) \cdot \int_{\mathbb{I}} Y(t)f(t)dt\right\}$$

for any $f \in L^2(\mathbb{I})$. Suppose $f(t) = \sum_{k=1}^{\infty} f_k \phi_k(t)$ with $\sum_k f_k^2 < \infty$. Then it follows

$$(\Sigma_{XY} f)(s) = \sum_{m=1}^{\infty} \sum_{k=1}^{\infty} f_k \sigma_{km} \psi_m(s).$$

Define $g(s) = \sum_{m=1}^{\infty} (a_m/\rho_m)\psi_m(s)$, where $a_m = \sum_{k=1}^{\infty} f_k \sigma_{km}$ for $m \in \mathbb{N}$. Note that

$$\begin{aligned}
\int_{\mathbb{I}} g^2(s)ds &= \sum_{m=1}^{\infty} \frac{a_m^2}{\rho_m^2} \\
&= \sum_{m=1}^{\infty} \frac{\left(\sum_{k=1}^{\infty} f_k \sigma_{km}\right)^2}{\rho_m^2} \\
&\leq \left(\sum_{k=1}^{\infty} f_k^2\right) \left(\sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \frac{\sigma_{km}^2}{\rho_m^2}\right) < \infty,
\end{aligned}$$

where the last line holds by the Cauchy-Schwarz inequality and (A44). Therefore, $g \in L^2(\mathbb{I})$. Moreover, since $\Sigma_{XX} = \sum_{m=1}^{\infty} \rho_m(\psi_m \otimes \psi_m)$, $\Sigma_{XX} g = \Sigma_{XY} f$. Thus, Assumption 4 is satisfied if

$$\sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \frac{\sigma_{km}^2}{\rho_m^2} < \infty.$$

This is a reasonable assumption because it holds if the dependence of Y on X, as measured by σ_{km}^2 dies out faster than ρ_m^2, which measures the variation in X, as $m \to \infty$. That is, the dependence is, to a degree, concentrated on the low-frequency region of the random function X.

Appendix A.4. Additional simulation studies

To implement our method for discretized observations, we provide some regularization conditions in Section 3.2 to ensure that the recovered trajectories are close to the underlying true ones, as outlined in Corollaries 3 and 4. However, these assumptions, i.e., Assumptions C1-C3, might be violated. To demonstrate the robustness of our method, we consider simulation designs where these assumptions do not hold.

In particular, the functional X is generated as follows: $X(t) = \sum_{k=1}^{50} \xi_k \phi_k(t)$ for $t \in [0,1]$, where $\phi_k(t) = \sqrt{2}\cos(k\pi t)$ for $k \geq 1$, $\nu_k = 1/k$ and $(3\nu_k)^{1/2}\xi_k$'s are independent student's t random variables with 3 degrees of freedom. Therefore, Assumptions C1 and C2 are violated. The response variable Y is then generated in the same manner as in Section 7.3. Moreover, in terms of the sampling frequency for X and Y, we adopt the same strategy as in Section 7.3. This entire simulation design is referred to as “the second sparse design", and the corresponding results are summarized in Table A1. It turns out that our proposed method can still perform reasonably well under various sampling frequencies of X and Y. These numerical results justify the robustness of our method to a violation of these regularity conditions in the sparse setting.

$N_X(N_Y)$	σ	Methods			
		FPCA	PLFFR	FUA	NLFFR
$\{4,5,6\}$	0.1	8.95 (7.80)	78.62 (369.42)	6.38 (3.68)	6.40 (5.46)
	2	9.00 (8.32)	71.16 (311.80)	6.52 (3.32)	6.58 (5.72)
$\{16,18,20\}$		FPCA	PLFFR	FUA	NLFFR
	0.1	8.73 (3.84)	3.56 (152.52)	4.31 (2.39)	3.34 (9.69)
	2	8.82 (4.16)	5.04 (279.78)	4.31 (2.79)	3.51 (9.08)
$\{5,10,20\}$		FPCA	PLFFR	FUA	NLFFR
	0.1	7.93 (5.14)	14.23 (128.75)	5.99 (4.16)	6.09 (5.77)
	2	8.11 (4.93)	12.26 (129.66)	6.00 (4.23)	6.22 (6.28)

Table A1: Summary of the medians and the interquartile ranges (in parentheses) of the ISE on the test set across the 200 simulation runs under different simulation scenarios for each method in the second sparse design.

By Assumption 1, the reproducing kernel κ in constructing $\mathfrak{M}_X$ is uniquely determined by ρ. To investigate the effect of the choice of ρ on the performance of our proposed method, we repeat the simulation studies in Sections 7.1 and 7.3 with different κ's. In particular, we consider the same designs as in these two sections and employ the same method to construct $\mathcal{H}_X$ and $\mathcal{H}_Y$, but use the Laplace kernel and/or the linear kernel to construct $\mathfrak{M}_X$ in the numerical implementation. Tables A2 and A3 summarize the performance of our method with different κ's under the dense design (in Section 7.1 and the sparse design (in Section 7.3), respectively. Both the GRB kernel and the Laplace kernel show a remarkable advantage over the linear kernel in terms of prediction accuracy when implementing our method under both the dense and sparse scenarios.

Model	X	σ	κ			
			Laplace		Linear	
			NLFFR (GRB)	NLFFR (BMC)	NLFFR (GRB)	NLFFR (BMC)
1	GRB	0.1	3.50 (1.58)	6.29 (2.51)	41.29 (75.17)	108.22 (201.34)
		2	3.48 (1.68)	6.39 (2.50)	42.04 (72.94)	115.18 (203.97)
	BMC	0.1	0.72 (0.08)	0.70 (0.08)	0.24 (0.06)	0.13 (0.07)
		2	0.72 (0.12)	0.70 (0.12)	0.33 (0.13)	0.31 (0.21)
2	GRB	0.1	0.57 (0.21)	0.64 (0.22)	1.36 (1.21)	4.04 (8.51)
		2	0.60 (0.25)	0.66 (0.26)	2.58 (2.15)	8.47 (11.77)
	BMC	0.1	0.51 (0.25)	0.74 (0.27)	1.57 (0.31)	1.52 (0.32)
		2	0.53 (0.27)	0.76 (0.30)	1.64 (0.37)	1.69 (0.42)

Table A2: Summary of the medians and the interquartile ranges (in parentheses) of the prediction errors across the 200 simulation runs under different simulation scenarios for our methods with different kernels κ in the dense design in Section 7.1. The column of X indicates which kernel is used to generate X in model 1 or 2, and the columns of NLFFR (GRB) and NLFFR (BMC) indicate which kernel is used to calculate the inner product in $\mathcal{H}_X$ and $\mathcal{H}_Y$ when using the proposed NLFFR.

$N_X(N_Y)$	σ	κ	
		Laplace	Linear
$\{4, 5, 6\}$	0.1	31.00 (15.97)	66.63 (24.90)
	2	33.08 (16.90)	67.63 (25.10)
$\{16, 18, 20\}$	0.1	22.95 (12.06)	67.74 (25.34)
	2	23.21 (12.08)	68.40 (26.00)
$\{5, 10, 20\}$	0.1	23.98 (9.69)	69.21 (19.25)
	2	24.60 (9.44)	70.43 (20.00)

Table A3: Summary of the medians and the interquartile ranges (in parentheses) of the ISE on the test set across the 200 simulation runs under different simulation scenarios for our methods with different kernels κ in the sparse design in Section 7.3.

References

T Tony Cai and Ming Yuan. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216, 2012.

Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591, 2003.

Hervé Cardot, André Mas, and Pascal Sarda. CLT in functional linear regression models. *Probability Theory and Related Fields*, 138:325–361, 2007.

Christophe Crambes and André Mas. Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli*, 19(5B):2627–2651, 2013.

Juan A Cuesta-Albertos, Eduardo García-Portugués, Manuel Febrero-Bande, and Wenceslao González-Manteiga. Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes. *The Annals of Statistics*, 47(1):439–467, 2019.

Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York, 2006.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning, 2nd edition*. Springer, New York, 2009.

Kenji Fukumizu, Francis R Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

Simon A Good, Matthew J Martin, and Nick A Rayner. EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research: Oceans*, 118(12):6704–6716, 2013.

Peter Hall and Joel L Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91, 2007.

Peter Hall and Mohammad Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126, 2006.

Lajos Horváth and Piotr Kokoszka. *Inference for Functional Data with Applications*. Springer, New York, 2012.

Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley, Chichester, 2015.

Charles R Johnson and Roger A Horn. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

Piotr Kokoszka and Matthew Reimherr. *Introduction to Functional Data Analysis*. CRC Press, London, 2017.

Kuang-Yao Lee and Lexin Li. Functional structural equation model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):600–629, 2022.

Kuang-Yao Lee, Lexin Li, Bing Li, and Hongyu Zhao. Nonparametric functional graphical modeling through functional additive regression operator. *Journal of the American Statistical Association*, 118(543):1718–1732, 2023.

B. Li and E. Solea. A nonparametric graphical model for functional data with application to brain networks based on fMRI. *Journal of the American Statistical Association*, 113: 1637–1655, 2018.

Bing Li. Linear operator-based statistical analysis: A useful paradigm for big data. *Canadian Journal of Statistics*, 46(1):79–103, 2018.

Bing Li and Jun Song. Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics*, 45(3):1059–1095, 2017.

Yehua Li and Tailen Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6):3321–3351, 2010.

Yehua Li, Naisyin Wang, and Raymond J Carroll. Selecting the number of principal components in functional data. *Journal of the American Statistical Association*, 108(504): 1284–1294, 2013.

Ruiyan Luo and Xin Qi. Function-on-function linear regression by signal compression. *Journal of the American Statistical Association*, 112(518):690–705, 2017.

Ruiyan Luo and Xin Qi. General nonlinear function-on-function regression via functional universal approximation. *Journal of Computational and Graphical Statistics*, 33(2):578–587, 2024.

Donald L McLeish. Dependent central limit theorems and invariance principles. *The Annals of Probability*, 2(4):620–628, 1974.

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

Hans-Georg Müller and Fang Yao. Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544, 2008.

Cristian Preda. Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of Statistical Planning and Inference*, 137(3):829–840, 2007.

Xin Qi and Ruiyan Luo. Nonlinear function-on-function additive model with multiple predictor curves. *Statistica Sinica*, 29:719–739, 2019.

Jim O Ramsay and Bernard W Silverman. *Functional Data Analysis (2nd edition)*. Springer-Verlag, New York, 2005.

Aniruddha Rajendra Rao and Matthew Reimherr. Modern non-linear function-on-function regression. *Statistics and Computing*, 33(6):130, 2023.

Matthew Reimherr, Bharath Sriperumbudur, and Bahaeddine Taoufik. Optimal prediction for additive function-on-function regression. *Electronic Journal of Statistics*, 12(2):4571–4601, 2018.

B. Scholkopf, R. Herbrich, and A. N. Smola. A generalized representer theorem. *Computational Learning Theory. Lecture Notes in Computer Science.*, 2111:416—-426, 2001.

Zuofeng Shang and Guang Cheng. Nonparametric inference in generalized functional linear models. *The Annals of Statistics*, 43(4):1742–1773, 2015.

Xiaoxiao Sun, Pang Du, Xiao Wang, and Ping Ma. Optimal penalized function-on-function regression under a reproducing kernel Hilbert space framework. *Journal of the American Statistical Association*, 113(524):1601–1611, 2018.

Aad W Van Der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes.* Springer, New York, 1996.

Fang Yao and Hans-Georg Müller. Functional quadratic regression. *Biometrika*, 97(1): 49–64, 2010.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005a.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005b.

Ming Yuan and T Tony Cai. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

Jin-Ting Zhang and Jianwei Chen. Statistical inferences for functional data. *The Annals of Statistics*, 35(3):1052–1079, 2007.

Q. Zhang, L. Xue, and B. Li. Dimension reduction for Fréchet regression. *Journal of the American Statistical Association*, 119:2733–2747, 2024.

Xiaoke Zhang and Jane-Ling Wang. From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321, 2016.

Hang Zhou, Fang Yao, and Huiming Zhang. Functional linear regression for discretely observed data: from ideal to reality. *Biometrika*, 110(2):381–393, 2023.

Hang Zhou, Dongyi Wei, and Fang Yao. Theory of functional principal component analysis for discretely observed data. *The Annals of Statistics*, 53(5):2103–2127, 2025.