Journal of Multivariate Analysis 181 (2021) 104675

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

# Nonlinear and additive principal component analysis for functional data

Jun Song [a,*], Bing Li [b]

[a] Department of Mathematics and Statistics, University of North Carolina at Charlotte, United States
[b] Department of Statistics, The Pennsylvania State University, United States

ARTICLE INFO

*Article history:*
Received 7 May 2019
Received in revised form 28 August 2020
Accepted 29 August 2020
Available online 9 September 2020

*AMS 2010 subject classifications:*
46E22
47B40
62H12

*Keywords:*
Functional data analysis
Handwritten digit classification
Nonlinear dimension reduction
Principal component analysis
Reproducing kernel Hilbert space

ABSTRACT

We introduce a nonlinear additive functional principal component analysis (NAFPCA) for vector-valued functional data. This is a generalization of functional principal component analysis and allows the relations among the random functions involved to be nonlinear. The method is constructed via two additively nested Hilbert spaces of functions, in which the first space characterizes the functional nature of the data, and the second space captures the nonlinear dependence. In the meantime, additivity is imposed so that we can avoid high-dimensional kernels in the functional space, which causes the curse of dimensionality. Along with the NAFPCA, we also develop a method of selection of the number of principal components and the tuning parameters that determines the degree of nonlinearity, as well as the asymptotic results for both the fully observed and the incompletely observed functional data. Simulation results show that the new method performs better than functional principal component analysis when the relations among random functions are nonlinear. We apply the new method to online handwritten digits and electroencephalogram (EEG) data sets.



## 1. Introduction

Modern technologies have made functional data increasingly prevalent in sciences and industries. For example, functional magnetic resonance imaging (fMRI) records brain activities as a collection of functions on a time interval; handwriting data, which can be regarded as two-dimensional vector-valued functions, are widely collected by electronic devices; health data such as blood pressures are routinely recorded by smart wearable devices. Since functional data are intrinsically infinite-dimensional, it is important to extract useful and interpretable information from them by suitable dimension reduction methods.

Principal component analysis (PCA) is one of the most popular methods in exploratory data analysis for extracting useful information from a sample of vectors. Intuitively, it seeks directions in the vectors that represent the greatest variation. Let $X$ be a $p$-dimensional random vector. At the population level, PCA solves the following problem

$$\text{maximize} \quad \text{var}(u^{\mathsf{T}}X) \quad \text{subject to} \quad \|u\| = 1. \tag{1}$$

This optimization is performed successively in a sequence of orthogonal spaces, resulting in a sequence of vectors $u_1, \ldots, u_d$. The projected random variables, ${u_1}^{\mathsf{T}}X, \ldots, {u_d}^{\mathsf{T}}X$, are then used as the principal components. See, for example, [15].

* Corresponding author.
*E-mail address:* Jun.Song@uncc.edu (J. Song).

https://doi.org/10.1016/j.jmva.2020.104675
0047-259X/

PCA was generalized to the nonlinear case by [28] by finding the nonlinear function of $X$ in a similar way. The idea is to replace the space of linear functions of $X$ with a Hilbert space of nonlinear functions of $X$, say $\mathfrak{M} = \{\phi : \mathbb{R}^p \to \mathbb{R}\}$, so as to capture the nonlinearity of the random vector $X$. Specifically, the first nonlinear principal component $\phi$ is an element in $\mathfrak{M}$ that satisfies

$$\text{var}[\phi(X)] = \max\{\text{var}[\psi(X)] : \ \psi \in \mathfrak{M}, \ \|\psi\|_{\mathfrak{M}} = 1\}. \tag{2}$$

Thus, the nonlinear PCA seeks the function of $X$ in a much larger space so that it captures more complex features than the linear PCA. The particular Hilbert space $\mathfrak{M}$ used by [28] is the reproducing Kernel Hilbert space (RKHS), which is computationally convenient. For this reason, the nonlinear-type PCA is called the kernel principal component analysis (KPCA).

Another useful generalization of PCA is to functional data, where $X$ is a function rather than a vector. The idea is to enlarge $\mathbb{R}^p$ to an infinite-dimensional Hilbert space to accommodate the random function $X$, and replace the Euclidean inner product by a functional inner product. More specifically, let $\mathscr{H}$ be a Hilbert space of functions defined on an interval and $X$ be a random element in $H$. In functional PCA (FPCA), we seek the member $f$ of $\mathscr{H}$ such that

$$\text{var}(\langle f, X\rangle_{\mathscr{H}}) = \max\{\text{var}(\langle g, X\rangle_{\mathscr{H}}) : g \in \mathscr{H}, \|g\|_{\mathscr{H}} = 1\}.$$

See Ramsay and Li [24], Ramsay and Silverman [25], and [21].

In this paper, we further extend the functional PCA to accommodate nonlinear functions of functional data. The nonlinear and functional nature of our problem requires two separable Hilbert spaces, say $\mathscr{H}$ and $\mathfrak{M}$; $\mathscr{H}$ is the space where $X$ resides, and $\mathfrak{M}$ includes nonlinear functions from $\mathscr{H}$ to $\mathbb{R}$, which captures the nonlinear feature of $X$. To avoid the curse of dimensionality, we assume $\mathfrak{M}$ to be an additive space; that is, each kernel only contains one component of $X$. We develop the numerical procedure and theoretical properties of this method, both at the population level and at the asymptotic level.

The rest of the paper is organized as follows. In Section 2, we propose the Nonlinear Additive Functional Principal Component Analysis (NAFPCA) at the population level. In Section 3, we present the sample-level estimation algorithm. In Section 4, we develop the asymptotic results of the NAFPCA. In Section 5, we conduct simulation comparisons between FPCA and NAFPCA. Some concluding remarks are made in Section 6. Real data applications to handwritten digits and EEG data sets are shown in the supplementary material.

## 2. Population-level development

Let $(\Omega, \mathcal{F}, P)$ be a probability space. For $i \in \{1, \ldots, p\}$, let $T_i$ be a compact set in $\mathbb{R}^{q_i}$ and let $\mathscr{H}_i$ be a Hilbert space of functions from $T_i$ to $\mathbb{R}$ with inner product $\langle\cdot,\cdot\rangle_{\mathscr{H}_i}$. Let $\mathscr{H} = \mathscr{H}_1 \times \cdots \times \mathscr{H}_p$ be endowed with the inner product

$$\langle f, g\rangle_{\mathscr{H}} = \langle f_1, g_1\rangle_{\mathscr{H}_1} + \cdots + \langle f_p, g_p\rangle_{\mathscr{H}_p},$$

for any $f = (f_1, \ldots, f_p)^\intercal \in \mathscr{H}$, and $g = (g_1, \ldots, g_p)^\intercal \in \mathscr{H}$. Let $X : \Omega \to \mathscr{H}$ be a random element in $\mathscr{H}$. Thus, $X = (X^1, \ldots, X^p)$, where each $X^i$ is a random element in $\mathscr{H}_i$. The additive structure for $\mathscr{H}$ has been imposed to reduce complexity. For example, see [10] and [14]. This structure, in effect, imposes geometric orthogonality among components, leaving any statistical dependence among them to be captured by the covariance operators. As soon to be shown, we also apply the additive structure to the second-level Hilbert space.

In the above and throughout the rest of this paper, we use superscript to identify the component of a vector, and subscript to identify a subject in a sample. Thus, $X^i$ is the $i$th component of the random vector $X$, and $X^i_a$ is the $i$th component in the $a$th subject in the sample of observations on $X$.

To characterize the nonlinear features of $X$, we need a second-level Hilbert space of functions defined on $\mathscr{H}$. To do so, we introduce a positive definite kernel $\kappa_i$ based on the inner product of $\mathscr{H}_i$. For $a, b \in \mathscr{H}_i$, let

$$\kappa_i(a, b) = \rho(\langle a, a\rangle_{\mathscr{H}_i}, \langle a, b\rangle_{\mathscr{H}_i}, \langle b, b\rangle_{\mathscr{H}_i}) \tag{3}$$

be a positive definite kernel in the sense that, for any finite subset $\{a_1, \ldots, a_m\}$ of $\mathscr{H}_i$, the matrix $\{\kappa(a_r, a_s) : r, s \in \{1, \ldots, m\}\}$ is positive definite. Here, $\rho$ is a known function. There are many ways to construct such kernels in the classical setting where $\mathscr{H}_i$ is a Euclidean space; see Berlinet and Thomas-Agnan [2] and Rasmussen and Williams [26, Chapter 4]. To adapt these kernels to the current setting, all we need to do is to replace the Euclidean inner product by the $\mathscr{H}_i$-inner product. For example, two of the most commonly used kernels are

$$\kappa_i(a, b) = \exp(-\gamma\|a - b\|^2_{\mathscr{H}_i}), \quad \kappa_i(a, b) = (c + \langle a, b\rangle_{\mathscr{H}_i})^k, \tag{4}$$

where, in the first expression, $\gamma > 0$, and in the second expression, $c > 0$ and $k$ is a positive integer. The first kernel is the Gaussian radial basis function (RBF), and the second kernel is the polynomial kernel, each adapted to functional context. One advantage of the Gaussian RBF is that the RKHS generated by it is dense in $L_2(\mathscr{H})$, which is rich enough to approximate any square-integrable nonlinear function on $\mathscr{H}$. See [23], [4, Chapter 4]. For this reason, we use the Gaussian RBF as the reproducing kernel for the second-level Hilbert space throughout the paper.

We use the kernel $\kappa_i$ to generate a reproducing kernel Hilbert space $\mathfrak{M}_i$; that is

$$\mathfrak{M}_i = \overline{\text{span}}\{\kappa_i(\cdot, a) : a \in \mathscr{H}_i\}.$$

This notation means $\mathfrak{M}_i$ is the completion of the space of all the functions of the form $c_1\kappa(\cdot, a_1) + \cdots + c_m\kappa(\cdot, a_m)$ where $c_1, \ldots, c_m \in \mathbb{R}$ and $a_1, \ldots, a_m \in \mathscr{H}_i$. The inner product in $\mathfrak{M}_i$ is uniquely determined by the relation

$$\langle \kappa_i(\cdot, a), \kappa_i(\cdot, b)\rangle_{\mathfrak{M}_i} = \kappa_i(a, b).$$

We call the pair $\{\mathscr{H}_i, \mathfrak{M}_i\}$ nested Hilbert spaces, because the kernel for $\mathfrak{M}_i$ is determined by the inner product of $\mathscr{H}_i$. We call $\mathscr{H}_i$ the first-level spaces and $\mathfrak{M}_i$'s the second-level spaces. The role played by $\mathscr{H}_i$ is to accommodate functional data; the role played by $\mathfrak{M}_i$ is to characterize the nonlinearity in the functional data.

Having constructed $\mathfrak{M}_1, \ldots, \mathfrak{M}_p$, we let $\mathfrak{M}$ be the direct sum of $\mathfrak{M}_1, \ldots, \mathfrak{M}_p$. That is, $\mathfrak{M}$ consists of the functions

$$\{\phi_1 + \cdots + \phi_p : \phi_1 \in \mathfrak{M}_1, \ldots, \phi_p \in \mathfrak{M}_p\},$$

and the inner product between two members of $\mathfrak{M}$, say $\phi = \phi_1 + \cdots + \phi_p$ and $\psi = \psi_1 + \cdots + \psi_p$, is defined by

$$\langle \phi_1 + \cdots + \phi_p, \psi_1 + \cdots + \psi_p\rangle_{\mathfrak{M}} = \langle \phi_1, \psi_1\rangle_{\mathfrak{M}_1} + \cdots + \langle \phi_p, \psi_p\rangle_{\mathfrak{M}_p}.$$

Note that, according to this definition, the $\mathfrak{M}_i$'s are subspaces of $\mathfrak{M}$. We denote this direct sum by $\mathfrak{M} = \oplus_{i=1}^p \mathfrak{M}_i$. We now introduce the Nonlinear Additive Functional Principal Component Analysis at the population level.

**Definition 1.** The population-level nonlinear additive functional principal components are defined through the following iterative maximization: at step $k$, $\phi^{(k)}$ is obtained by

$$\begin{aligned} &\text{maximizing} \quad \text{var}[\phi(X)] \\ &\text{subject to} \quad \phi \in \mathfrak{M},\ \langle \phi, \phi^{(1)}\rangle_{\mathfrak{M}} = \cdots = \langle \phi, \phi^{(k-1)}\rangle_{\mathfrak{M}} = 0,\ \|\phi\|_{\mathfrak{M}} = 1. \end{aligned}$$

The random variable

$$\phi_{(k)} = \phi_1^{(k)}(X^1) + \cdots + \phi_p^{(k)}(X^p)$$

is called the $k$th nonlinear additive functional principal component of $X$.

We next express the solutions in Definition 1 as eigenfunctions of a linear operator. We make the following assumption.

**Assumption 1.** For $i \in \{1, \ldots, p\}$, $E[\kappa_i(X^i, X^i)] < \infty$.

The function,

$$\mathscr{H}_i \to \mathbb{R}, \quad a \mapsto E[\kappa_i(a, X^i)],$$

is called the mean element in $\mathfrak{M}_i$ and is denoted by $E\kappa_i(\cdot, X^i)$. Let $\mathscr{B}(\mathfrak{M}_i, \mathfrak{M}_j)$ be the class of all bounded linear operators from $\mathfrak{M}_i$ to $\mathfrak{M}_j$. This is a Banach space in terms of the operator norm. A random operator $A$ in $\mathscr{B}(\mathfrak{M}_i, \mathfrak{M}_j)$ is a mapping from $\Omega$ to $\mathscr{B}(\mathfrak{M}_i, \mathfrak{M}_j)$ that is measurable with respect to the Borel $\sigma$-field generated by the open sets in $\mathscr{B}(\mathfrak{M}_i, \mathfrak{M}_j)$. For a random operator $A$ in $\mathscr{B}(\mathfrak{M}_i, \mathfrak{M}_j)$, if the bilinear form

$$(\phi_i, \phi_j) \mapsto E\langle A\phi_i, \phi_j\rangle_{\mathfrak{M}_j}$$

is bounded, then by Theorem 2.2 of [5], there exists a (nonrandom) operator $B \in \mathscr{B}(\mathfrak{M}_i, \mathfrak{M}_j)$ such that

$$\langle B\phi_i, \phi_j\rangle_{\mathfrak{M}_j} = E\langle A\phi_i, \phi_j\rangle_{\mathfrak{M}_j}$$

The operator $B$ is then defined as the expectation of $A$, and is written as $E(A)$. For two members $\phi$ and $\psi$ of $\mathfrak{M}$, the tensor product $\phi \otimes \psi$ is the operator

$$\phi \otimes \psi : \quad \mathfrak{M} \to \mathfrak{M}, \quad \eta \mapsto \phi\langle \psi, \eta\rangle_{\mathfrak{M}}.$$

Define a linear operator from $\mathfrak{M}_i$ to $\mathfrak{M}_j$,

$$C(X^j, X^i) = [\kappa_j(\cdot, X^j) - E\kappa_j(\cdot, X^j)] \otimes [\kappa_i(\cdot, X^i) - E\kappa_i(\cdot, X^i)].$$

Under Assumption 1, for each $(i, j) \in \{1, \ldots, p\} \times \{1, \ldots, p\}$, the bilinear form

$$\mathfrak{M}_i \times \mathfrak{M}_j \to \mathbb{R}, \quad (\phi_i, \phi_j) \mapsto E[\langle \phi_i, C(X^i, X^j)\phi_j\rangle_{\mathfrak{M}_j}]$$

is bounded, so that the expectation $E[C(X^j, X^i)]$ is a well defined operator in $\mathscr{B}(\mathfrak{M}_i, \mathfrak{M}_j)$. This operator is called the covariance operator from $\mathfrak{M}_i$ to $\mathfrak{M}_j$ (or simply from $X^i$ to $X^j$) and is written as $\Sigma_{X^jX^i}$.

Let $\Sigma_{XX}$ be the $p \times p$ matrix of operators $\{\Sigma_{X^iX^j} : i, j \in \{1, \ldots, p\}\}$, by which we mean $\Sigma_{XX} : \mathfrak{M} \to \mathfrak{M}$ maps a function $\phi = \phi_1 + \cdots + \phi_p \in \mathfrak{M}$ to the function $\psi_1 + \cdots + \psi_p \in \mathfrak{M}$ where

$$\psi_i = \sum_{j=1}^{p} \Sigma_{X^iX^j}\phi_j.$$

Then,

$$\langle \phi, \Sigma_{XX}\phi\rangle_{\mathfrak{M}} = \sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{cov}[\phi_i(X^i), \phi_j(X^j)] = \mathrm{var}[\phi(X)].$$

Using these operators, we can state the iterative maximization in Definition 1 equivalently as: at the $k$th step, obtain $\phi^{(k)}$ by

$$\begin{aligned} &\text{maximizing} \quad \langle \phi, \Sigma_{XX}\phi\rangle_{\mathfrak{M}}, \\ &\text{subject to} \quad \phi \in \mathfrak{M},\ \langle \phi, \phi^{(1)}\rangle_{\mathfrak{M}} = \cdots = \langle \phi, \phi^{(k-1)}\rangle_{\mathfrak{M}} = 0,\ \|\phi\|_{\mathfrak{M}} = 1. \end{aligned} \tag{5}$$

In other words, $\phi^{(k)}$ is the $k$th eigenfunction of the operator $\Sigma_{XX}$.

We can further simplify the iterative maximization in (5). For each $i = 1, \ldots, p$, let

$$\ker(\Sigma_{X^iX^i}) = \{\phi_i \in \mathfrak{M}_i :\ \Sigma_{X^iX^i}\phi_i = 0\}, \quad \mathrm{ran}(\Sigma_{X^iX^i}) = \{\Sigma_{X^iX^i}\phi_i : \phi_i \in \mathfrak{M}_i\}$$

be the kernel and range of the operator $\Sigma_{X^iX^i}$, respectively. Let $\overline{\mathrm{ran}}(\Sigma_{X^iX^i})$ denote the closure of $\mathrm{ran}(\Sigma_{X^iX^i})$. Since $\phi_i \in \ker(\Sigma_{X^iX^i})$ if and only if $\mathrm{var}[\phi_i(X^i)] = 0$, we can assume, without loss of generality, that all the maximizers in (5) are contained in

$$\oplus_{i=1}^{p} \ker(\Sigma_{X^iX^i})^{\perp} = \oplus_{i=1}^{p} \overline{\mathrm{ran}}(\Sigma_{X^iX^i}) \equiv \mathfrak{M}_0.$$

In fact, it is easy to see that, for any $\phi \in \mathfrak{M}$, there exists a $\phi' \in \mathfrak{M}_0$ such that $\mathrm{var}[\phi(X)] = \mathrm{var}[\phi'(X)]$. Thus, the iterative procedure in (5) can be further restated as: at the $k$th step, obtain $\phi^{(k)}$ by

$$\begin{aligned} &\text{maximizing} \quad \langle \phi, \Sigma_{XX}\phi\rangle_{\mathfrak{M}}, \\ &\text{subject to} \quad \phi \in \mathfrak{M}_0,\ \langle \phi, \phi^{(1)}\rangle_{\mathfrak{M}} = \cdots = \langle \phi, \phi^{(k-1)}\rangle_{\mathfrak{M}} = 0,\ \|\phi\|_{\mathfrak{M}} = 1. \end{aligned}$$

By Lemma 1 of Li and Song [22], each $\overline{\mathrm{ran}}(\Sigma_{X^iX^i})$ is, in fact, the following space

$$\overline{\mathrm{span}}\{\kappa_i(\cdot, a) - E\kappa_i(\cdot, X^i) :\ a \in \mathfrak{M}_i\}.$$

Hence the subspace $\mathfrak{M}_0$ can be explicitly written as

$$\mathfrak{M}_0 = \oplus_{i=1}^{p} \overline{\mathrm{span}}\{\kappa_i(\cdot, a) - E\kappa_i(\cdot, X^i) :\ a \in \mathfrak{M}_i\}.$$

## 3. Sample-level implementation

In this section, we implement the population-level NAFPCA defined in the last section as a sample-level algorithm. Let $X_1, \ldots, X_n$ be an i.i.d. sample of $X$. The algorithm hinges on representing relevant linear operators as $n \times n$ matrices in a finite sample. Throughout the rest of the paper, we assume each $T_i$ to be a closed interval in $\mathbb{R}R$.

### 3.1. Coordinate system

The following notation for coordinate representation and various related results stated without proof are taken from [11,19,20,22]. Let $\mathscr{H}_1$ be a generic finite-dimensional Hilbert space with spanning system $\mathscr{B} = \{b_1, \ldots, b_n\}$. Then for any member $f \in \mathscr{H}_1$, there is a vector $\alpha = (\alpha_1, \ldots, \alpha_n)^{\mathsf{T}} \in \mathbb{R}^n$ such that $f = \sum_{i=1}^{n} \alpha_i b_i$. We call the vector $\alpha$ the coordinate of $f$ with respect to $\mathscr{B}$, and denote it $[f]_{\mathscr{B}}$. Let $\mathscr{H}_2$ be another Hilbert space with spanning system $\mathscr{C} = \{c_1, \ldots, c_m\}$, and let $A$ be a linear operator in $\mathscr{B}(\mathscr{H}_1, \mathscr{H}_2)$. Then we define the coordinate of $A$ with respect to $\mathscr{B}$ and $\mathscr{C}$ by the $m \times n$ matrix, denote ${}_{\mathscr{C}}[A]_{\mathscr{B}}$, whose $(i, j)$-th entry is $([Ab_j]_{\mathscr{C}})_i$, the $i$th component of the vector $[Ab_j]_{\mathscr{C}}$. Then it is easy to see that for any $f \in \mathscr{H}_1$, $[Af]_{\mathscr{C}} = ({}_{\mathscr{C}}[A]_{\mathscr{B}})[f]_{\mathscr{B}}$. Furthermore, if $\mathscr{H}_3$ is a third Hilbert space with spanning system $\mathscr{D} = \{d_1, \ldots, d_\ell\}$ and $B$ is a linear operator in $\mathscr{B}(\mathscr{H}_2, \mathscr{H}_3)$, then ${}_{\mathscr{D}}[BA]_{\mathscr{B}} = ({}_{\mathscr{D}}[B]_{\mathscr{C}})({}_{\mathscr{C}}[A]_{\mathscr{B}})$. For convenience, when the spanning systems are obvious from the context, we drop the subscript and write the coordinate of $A$ simply as $[A]$.

### 3.2. Construction of first-level function space

At the sample level, we can only observe the random function $X_a^i$ at a finite set of time points, say $S_a^i = \{t_{a1}^i, \ldots, t_{ak_a^i}^i\}$ for $i \in \{1, \ldots, p\}$ and $a \in \{1, \ldots, n\}$. Define $\tau^i = \text{sort}(\cup_{a=1}^n S_a^i) = (\tau_1^i, \ldots, \tau_{u_i}^i)^\top$ to be a vector consisting of all observed time points with $\tau_1^i < \cdots < \tau_{u_i}^i$, where $u_i$ is the cardinality of $\cup_{a=1}^n S_a^i$. Let $J_a^i$ be the set of indices of the members of $\tau^i$ that are the observed time points for $X_a^i$.

We now illustrate two ways to construct the first-level function space: one with a set of given functions as a basis and the other with the RKHS over the observed time points. In both cases, the coordinates are found by each functional component.

Here we assume that $\mathscr{H}_i$ is spanned by a given set of functions, $\mathscr{B}^i = \{b_1^i, \ldots, b_{m_i}^i\}$. We further assume that the inner product in $\mathscr{H}_i$ is the $L_2$-inner product with respect to the Lebesgue measure; that is,

$$\langle f, g\rangle_{\mathscr{H}_i} = \int_{T_i} f(t)g(t)dt.$$

To evaluate the above inner product for $f$ or $g$ equal to $X_a^i$, we need to estimate the entire function $X_a^i$ for all $t \in T_i$. We do so by finding the member of $\mathscr{H}_i$ closest to $X_a^i$ at the observed points with a roughness penalty. Let $\hat{X}_a^i(t) = C^i{}_{a,1}b_1^i(t) + \cdots + C^i{}_{a,m}b_{m_i}^i(t) = [\hat{X}_a^i]^\top b_{1:m_i}^i(t)$, where $b_{1:m_i}^i(t) = (b_1^i(t), \ldots, b_{m_i}^i(t))^\top$. We minimize the penalized least-squares criterion,

$$\sum_{j=1}^{k_a^i}\{\hat{X}_a^i(t_{aj}) - X_a^i(t_{aj})\}^2 + \epsilon_a^i \left\| \frac{d^2}{dt^2}\hat{X}_a^i(t)\right\|_{\mathscr{H}_i}^2. \tag{6}$$

In terms of coordinates, the first term of (6) can be written as

$$\|X_a^i(J_a^i) - B_a^i[\hat{X}_a^i]\|^2,$$

where $\|\cdot\|$ is the Euclidean norm, $X_a^i(J_a^i)$ is the vector $\{X_a^i(\tau_r) : r \in J_a^i\}$, and $B_a^i$ is $\text{card}(J_a^i) \times m_i$ matrix whose rows are ${b_{1:m_i}}^\top(\tau_r)$, $r \in J_a^i$. The derivative in the second term is

$$\frac{d^2}{dt^2}\hat{X}_a^i(t) = [\hat{X}_a^i]^\top \ddot{b}_{1:m_i}(t).$$

It follows that

$$\left\| \frac{d^2}{dt^2}\hat{X}_a^i(t)\right\|_{\mathscr{H}_i}^2 = [\hat{X}_a^i]^\top \int_{T_i} \ddot{b}_{1:m_i}(t)\ddot{b}_{1:m_i}^\top(t)dt[\hat{X}_a^i] \equiv [\hat{X}_a^i]^\top R^i[\hat{X}_a^i],$$

where

$$R^i = \int_{T_i} \ddot{b}_{1:m_i}(t)\ddot{b}_{1:m_i}^\top(t)dt.$$

The objective function (6) can now be rewritten as

$$\|X_a^i(J_a^i) - B_a^i[\hat{X}_a^i]\|^2 + \epsilon_a^i[\hat{X}_a^i]^\top R^i[\hat{X}_a^i].$$

This is a quadratic equation in $[\hat{X}_a^i]$, and has the explicit solution

$$[\hat{X}_a^i] = (B_a^{i\top}B_a^i + \epsilon_a^i R^i)^{-1}B_a^\top X_a^i(J_a^i).$$

The tuning parameter $\epsilon_a^i > 0$ is chosen by Generalized Cross-Validation (GCV, Golub et al. [8]), that is, $\epsilon_a^i$ is the minimizer of

$$GCV(\epsilon_a^i) = \frac{\sum_{j=1}^{k_a^i}(X_a^i(t_{aj}^i) - \hat{X}_a^i(t_{aj}^i))^2}{\left[\text{tr}(I - B^i(B^{i\top}B^i + \epsilon_a^i R^i)^{-1}B^{i\top})/m_i\right]^2},$$

over a grid.

The function $X_a^i(t)$, $t \in T$, is then estimated by $\hat{X}_a^i(t) = [\hat{X}_a^i]^\top b_{1:m_i}(t)$. These functions are used in computing the inner products $\langle \hat{X}_a^i, \hat{X}_b^i\rangle_{\mathscr{H}_i}$, which are then substituted into (3) to construct the $n \times n$ kernel matrix whose $(a, b)$-th entry is $\kappa_i(\hat{X}_a^i, \hat{X}_b^i)$. In terms of coordinate for $[\hat{X}_a^i]$, the inner product can be expressed as

$$\langle X_a^i, X_b^i\rangle_{\mathscr{H}_i} = [\hat{X}_a^i]^\top G_{B^i}[\hat{X}_a^i],$$

where $G_{B^i}$ is the $m_i \times m_i$ Gram matrix of $B^i$ whose $(r, s)$-th element is $\langle b_r^i, b_s^i\rangle_{\mathscr{H}_i}$.

Alternatively, we can estimate $X_a^i$ based on RKHS. Let $\kappa_{T_i} : T_i \times T_i \to \mathbb{R}$ be a positive definite kernel and let $\mathscr{H}_i$ be the RKHS generated by $\{\kappa_{T_i}(\cdot, \tau_j^i) : j \in \{1, \ldots, u_i\}\}$. Note that if we assume the kernel is universal – which is the case

for the Gaussian RBF, for example – then $\mathscr{H}_i$ is a large enough subset to approximate any functions in $L_2(T_i)$. See, for example, [29] and [3].

Let $K_{T_i}$ be the $u_i \times u_i$ Gram matrix with $(K_{T_i})_{ab} = \kappa_T(\tau_a^i, \tau_b^i)$. Let $b_{T_i}(\cdot)$ be the vector-valued function from $T_i$ to $\mathbb{R}^{u_i}$

$$b_{T_i}(t) = (\kappa_{T_i}(t, \tau_1^i), \ldots, \kappa_{T_i}(t, \tau_{u_i}^i))^{\mathsf{T}}.$$

Correspondingly, let $\mathscr{B}^i$ be the set of functions $\{\kappa_{T_i}(\cdot, \tau_1^i), \ldots, \kappa_{T_i}(\cdot, \tau_{u_i}^i)\}$. Then the inner product in $\mathscr{H}_i$ can be expressed as

$$\langle \phi, \psi \rangle_{\mathscr{H}_i} = \langle [\phi]^{\mathsf{T}} b_{T_i}, [\psi]^{\mathsf{T}} b_{T_i} \rangle_{\mathscr{H}_i} = [\phi]^{\mathsf{T}} K_{T_i} [\psi], \quad \text{for } \phi, \psi \in \mathscr{H}_i.$$

It is natural to use only those functions $\kappa_{T_i}(\cdot, t)$ with $t \in S_a^i$ to estimate $X_a^i$, which means that the entries of $[\hat{X}_a^i]$ are 0 except its components with indices in $J_a^i$, and $\hat{X}_a^i(\cdot) = \sum_{k \in J_a} [\hat{X}_a^i]_k \kappa_{T_i}(\cdot, \tau_k^i)$. Let $[\hat{X}_a^i]^0$ be the $k_a^i$-dimensional sub-vector of $[\hat{X}_a^i]$ whose elements are nonzero. Define $k_{T_i}^{(a,b)}$ be a sub-matrix of $K_{T_i}$ with indices in $J_a \times J_b$. Since, for any $\ell \in J_a^i$, we have $X_a^i(\tau_\ell^i) = \langle X_a^i, \kappa_{T_i}(\cdot, \tau_\ell^i) \rangle_{\mathscr{H}} = \sum_{k \in J_a^i} [X_a^i]_k \kappa_{T_i}(\tau_\ell^i, \tau_k^i)$, the vector $[\hat{X}_a^i]^0$ is the solution to the following equation,

$$X_a^i(J_a^i) = {K_{T_i}}^{(a,a)} [\hat{X}_a^i]^0.$$

To enhance model parsimony, we use Tychonoff regularization when solving the equation, which gives us

$$[\hat{X}_a^i]^0 = ({K_{T_i}}^{(a,a)} + \epsilon_{T_i} I_{k_a^i})^{-1} X_a^i(J_a^i).$$

Equivalently, the above process can be obtained by minimizing the objective function

$$\sum_{j=1}^{k_a^i} (x(t_{aj}^i) - X_a^i(t_{aj}^i))^2 + \epsilon_{T_i} \|x\|_{\mathscr{H}_i}^2,$$
$$\text{subject to} \quad x \in \mathscr{H}_i.$$

An advantage of using RKHS to construct the function space is that the optimization over the finite-dimensional space, $\mathscr{H}_i$, is equivalent to that over the infinite-dimensional RKHS generated by the kernel $\kappa_{T_i}$; that is, the space $\overline{\text{span}}\{\kappa_{T_i}(\cdot, t) : t \in T_i\}$. This result is known as the representer theorem: see Schölkopf et al. [27]. The penalty term $\|x\|_{\mathscr{H}_i}^2$ depends on the choice of reproducing kernel. For example, if we use Brownian covariance function as the reproducing kernel, then $\|x\|_{\mathscr{H}_i}^2 = \|\dot{x}\|_{L_2(T_i)}^2$, where $\dot{x}$ is the derivative of $x$. In many cases, the construction with RKHS tends to find smoother functions than those given by the b-spline basis.

Using the individual Hilbert spaces $\mathscr{H}_i$ for $X^i$, $i = 1, \ldots, p$, we then construct the joint Hilbert space $\mathscr{H}$ for $(X^1, \ldots, X^p)$ as $\mathscr{H}_1 \times \cdots \times \mathscr{H}_p$ with its inner product defined by the sum of the inner products for $\mathscr{H}_1, \ldots, \mathscr{H}_p$.

### 3.3. NAFPCA for scalar-valued functional data

With the first-level space of $\mathscr{H}$ constructed, we now construct the second-level space that captures nonlinearity in the data. This space is also called the feature space in the machine learning literature. In this subsection, we first deal with scalar-valued function $X$. Thus, in this case, $\mathscr{H} = \mathscr{H}_0$. For simplicity, we denote $\hat{X}_a$ by $X_a$. As described in Section 2, the crucial step is to construct the variance operator, $\hat{\Sigma}_{XX}$, in the feature space, and its coordinate representation. Let

$$\mathfrak{M} = \text{span}\{\kappa_X(\cdot, X_a) : \ a \in \{1, \ldots, n\}\}.$$

Again, it is sufficient to use this finite-dimensional space due to the representer theorem for kernel PCA. See Example 3 of [6]. Let $K_X$ be the $n \times n$ Gram matrix whose $(a, b)$-th element is $\kappa_X(X_a, X_b)$. Let

$$b_X(\cdot) = (\kappa_X(\cdot, X_1), \ldots, \kappa_X(\cdot, X_n))^{\mathsf{T}},$$

and let $Q = I_n - n^{-1} 1_n 1_n^{\mathsf{T}}$. At the sample level, the centered space $\mathfrak{M}_0$ is

$$\mathfrak{M}_0 = \text{span}\{\kappa_X(\cdot, X_a) - E_n \kappa_X(\cdot, X) : \ a \in \{1, \ldots, n\}\}.$$

**Lemma 1.** *Let* $\mathscr{B}_X = \{\kappa_X(\cdot, X_a) - E_n \kappa_X(\cdot, X) : \ a \in \{1, \ldots, n\}\}$, *and let*

$$\hat{\Sigma}_{XX} = E_n[(\kappa_X(\cdot, X) - E_n \kappa_X(\cdot, X)) \otimes (\kappa_X(\cdot, X) - E_n \kappa_X(\cdot, X))].$$

*Then* ${}_{\mathscr{B}_X}[\hat{\Sigma}_{XX}]_{\mathscr{B}_X} = n^{-1} Q K_X Q \equiv n^{-1} G_X$.

The proof is similar to that used in [7], and is omitted. Since the Gram matrix of $\mathscr{B}_X$ is $G_X = Q K_X Q$, the inner product $\langle \phi, \hat{\Sigma}_{XX} \phi \rangle$ is $[\phi]^{\mathsf{T}} G_X [\hat{\Sigma}_{XX}][\phi] = n^{-1} [\phi]^{\mathsf{T}} G_X^2 [\phi]$. Similarly, the inner product $\langle \phi, \psi \rangle$ is $[\phi]^{\mathsf{T}} G_X [\psi]$. Hence the eigenvalue

problem in (5) in Section 2 becomes

$$\begin{aligned}&\text{maximizing} \quad [\phi]^{\mathsf{T}} G_X^2 [\phi],\\ &\text{subject to} \quad [\phi]^{\mathsf{T}} G_X [\phi] = 1, [\phi]^{\mathsf{T}} G_X [\phi_1] = \cdots = [\phi]^{\mathsf{T}} G_X [\phi_{k-1}] = 0.\end{aligned}$$

If we let $v$ be the $n$-dimensional vector $v = G_X^{1/2}[\phi]$, then the above problem is converted to the following eigenvalue problem:

$$\begin{aligned}&\text{maximizing} \quad v^{\mathsf{T}} G_X^{-1/2} G_X^2 G_X^{-1/2} v = v^{\mathsf{T}} G_X v,\\ &\text{subject to} \quad v^{\mathsf{T}} v = 1, v^{\mathsf{T}} v_1 = \cdots = v^{\mathsf{T}} v_{k-1} = 0.\end{aligned}$$

Let $\lambda_1 \geq \cdots \geq \lambda_n$ and $v_1, \ldots, v_n$ be the eigenvalues and eigenvectors of $G_X$, and let $\phi_k$ be the $k$th eigenfunction of the operator $\hat{\Sigma}_{XX}$. Then the coordinates $[\phi_k]$ is given by $G_X^{-1/2} v_k$. Hence the $k$th eigenfunction of $\hat{\Sigma}$ is $\phi_k(x) = v_k{}^{\mathsf{T}} G_X^{-1/2} c_X(x)$, where

$$c_X(\cdot) = (\kappa_X(\cdot, X_1) - E_n(\kappa_X(\cdot, X)), \ldots, \kappa_X(\cdot, X_n) - E_n(\kappa_X(\cdot, X)))^{\mathsf{T}}.$$

The $k$th principal component is defined as the vector $(\phi_k(X_1), \ldots, \phi_k(X_n))^{\mathsf{T}}$.

If we take first $d$ principal components, then the dimension of observed data in function space is reduced to $d$:

$$X \in \mathscr{H} \mapsto X^* \equiv (\phi_1(X), \ldots, \phi_d(X))^{\mathsf{T}} \in \mathbb{R}^d,$$

where $X^*$ contains the amount of information of $X$ of proportional $\sum_{i=1}^d \lambda_i / \sum_{i=1}^n \lambda_i$. In practice, the eigenvalues of $G_X$ often decreases sharply after the first few eigenvalues.

### 3.4. NAFPCA for vector-valued functional data

We now consider the case of $p > 1$. For each $i = 1, \ldots, p$, let $\kappa_i : \mathscr{H}_i \times \mathscr{H}_i \to \mathbb{R}$ be a positive definite kernel and let $\mathscr{B}_{X^i} = \{\kappa_i(\cdot, X_a^i) - E_n \kappa_i(\cdot, X^i) : \ a = 1, \ldots, n\}$. Let $\mathfrak{M}_i$ be the Hilbert space spanned by $\mathscr{B}_{X^i}$ with the inner product determined by the kernel $\kappa_i$. For each $i, j = 1, \ldots, p$, let

$$\hat{\Sigma}_{X^j X^i} = E_n[(\kappa_j(\cdot, X^j) - E_n \kappa_j(\cdot, X^j)) \otimes (\kappa_i(\cdot, X^i) - E_n \kappa_i(\cdot, X^i))].$$

Then $\hat{\Sigma}_{X^j X^i}$ is an operator from $\mathfrak{M}_i$ to $\mathfrak{M}_j$ with the coordinate representation

$$_{\mathscr{B}_{X^j}}[\hat{\Sigma}_{X^j X^i}]_{\mathscr{B}_{X^i}} = n^{-1} G_{X^i} = n^{-1} Q K_{X^i} Q. \tag{7}$$

Again, the derivation is similar to that given in [7], and is omitted.

Let $\mathfrak{M}$ be the direct sum $\bigoplus_{i=1}^p \mathfrak{M}_i$ in the sense explained in Section 2. Let

$$\mathscr{B}_X = \{\phi_1 + \cdots + \phi_p : \ \phi_1 \in \mathscr{B}_{X^1}, \ldots, \phi_p \in \mathscr{B}_{X^p}\}.$$

Then $\mathfrak{M}$ is spanned by $\mathscr{B}_X$. For a function $\phi = \phi_1 + \cdots + \phi_p$ in $\mathfrak{M}$, its coordinate with respect to $\mathscr{B}_X$ can be expressed as

$$[\phi]_{\mathscr{B}_X} = ([\phi]_{\mathscr{B}_{X^1}}{}^{\mathsf{T}}, \ldots, [\phi]_{\mathscr{B}_{X^p}}{}^{\mathsf{T}})^{\mathsf{T}}.$$

That is,

$$\phi = [\phi]_{\mathscr{B}_{X^1}}{}^{\mathsf{T}} c_{X^1} + \cdots + [\phi]_{\mathscr{B}_{X^p}}{}^{\mathsf{T}} c_{X^p},$$

where $c_{X^i}$ is the vector-valued function

$$((\kappa_i(\cdot, X_1^i) - E_n \kappa_i(\cdot, X^i)), \ldots, (\kappa_i(\cdot, X_n^i) - E_n \kappa_i(\cdot, X^i)))^{\mathsf{T}}.$$

Let $\hat{\Sigma}_{XX}$ be the $p \times p$ matrix of operators whose $(i, j)$-th entry is the operator $\hat{\Sigma}_{X^i X^j}$.

Equipped with the coordinate representation, we now express the eigenvalue problem (5) at the sample level. First, the inner product $\langle \phi, \hat{\Sigma}_{XX} \phi \rangle_{\mathfrak{M}}$ can be reexpressed as

$$\langle \phi, \hat{\Sigma}_{XX} \phi \rangle_{\mathfrak{M}} = \sum_{i=1}^p \sum_{j=1}^p \langle \phi_i, \hat{\Sigma}_{X^i X^j} \phi_j \rangle_{\mathfrak{M}_i},$$

where the summand has coordinate representation

$$\langle \phi_i, \hat{\Sigma}_{X^i X^j} \phi_j \rangle_{\mathfrak{M}_i} = [\phi_i]_{\mathscr{B}_{X^i}}{}^{\mathsf{T}} G_{X^i} (_{\mathscr{B}_{X^i}}[\hat{\Sigma}_{X^i X^j}]_{\mathscr{B}_{X^j}}) [\phi_j]_{\mathscr{B}_{X^j}}.$$

By (7), the right hand side is

$$n^{-1} [\phi_i]_{\mathscr{B}_{X^i}}{}^{\mathsf{T}} G_{X^i} G_{X^j} [\phi_j]_{\mathscr{B}_{X^j}}.$$

Thus, if we let $M = (G_{X^1}, \ldots, G_{X^p})^{\mathsf{T}} (G_{X^1}, \ldots, G_{X^p})$, then we can express $\langle \phi, \hat{\Sigma}_{XX} \phi \rangle_{\mathfrak{M}} = n^{-1} [\phi]_{\mathscr{B}_X}{}^{\mathsf{T}} M [\phi]_{\mathscr{B}_X}$.

Second, the inner product $\langle \phi, \phi^{(\ell)} \rangle_{\mathfrak{M}}$ in (5) has the following coordinate representation:

$$\langle \phi, \phi^{(\ell)} \rangle_{\mathfrak{M}} = \sum_{i=1}^{n} \langle \phi_i, \phi_i^{(\ell)} \rangle_{\mathfrak{M}_i} = \sum_{i=1}^{n} [\phi_i]_{\mathscr{B}_{X^i}}{}^{\mathsf{T}} G_{X^i} [\phi_i^{(\ell)}]_{\mathscr{B}_{X^i}}.$$

Thus, if we let $D = \operatorname{diag}(G_{X^1}, \ldots, G_{X^p})$, then $\langle \phi, \phi^{(\ell)} \rangle_{\mathfrak{M}} = [\phi]_{\mathscr{B}_X}{}^{\mathsf{T}} D [\phi^{(\ell)}]_{\mathscr{B}_X}$. As in the scalar-valued function case, we make the transformation $v = D^{1/2}[\phi]_{\mathscr{B}_X}$. Then the eigenvalue problem becomes: at the $k$th step,

$$\begin{aligned} &\text{maximizing} \quad v^{\mathsf{T}} D^{-1/2} M D^{-1/2} v, \\ &\text{subject to} \quad v^{\mathsf{T}} v = 1, v^{\mathsf{T}} v^{(\ell)} = 0, \ell \in \{1, \ldots, k-1\}. \end{aligned}$$

Since $G_{X^i} = Q K_{X^i} Q$ is of rank $n-1$, the $np \times np$ matrix $D^{-1/2} M D^{-1/2}$ also has rank $n-1$. So it has $n-1$ eigenvectors, say $v^{(1)}, \ldots, v^{(n-1)}$, with nonzero eigenvalues. The corresponding eigenfunctions of $\hat{\Sigma}_{XX}$ are $\phi^{(\ell)} = (v^{(\ell)})^{\mathsf{T}} D^{-1/2} c_X$, where

$$c_X = (c_{X^1}{}^{\mathsf{T}}, \ldots, c_{X^p}{}^{\mathsf{T}})^{\mathsf{T}}.$$

The vector $(\phi^{(\ell)}(X_1), \ldots, \phi^{(\ell)}(X_n))^{\mathsf{T}}$ is the $\ell$-th functional additive principal component.

### 3.5. Tuning parameters selection

The NAFPCA requires selection of the following items:

(1) A set of basis functions for the first-level Hilbert space $\mathscr{H}$ if we use basis expansion method to estimate $X$, or a reproducing kernel and its associated tuning parameters if we use the RKHS method to estimate $X$.
(2) The tuning parameter $\gamma$ for the second-level RKHS $\mathfrak{M}$.
(3) The number of significant NAF-principal components $d$.

For the choices in (1), we have experimented with spline basis and the Gaussian RBF-based RKHS. The tuning parameters for Gaussian RBF can be chosen using the method described in [22]. Since, in the applications we considered, the functional data are observed in relatively densely placed intervals, the choices in (1) do not affect the result significantly.

An important choice is that of $\gamma$ in (2), which directly determines the degree of nonlinearity in the NAFPCA. In particular, with a large $\gamma$, the resulting principal components tend to be highly nonlinear; whereas with a small $\gamma$, they tend to be approximately linear. This feature can be explained by the geometry of the RKHS with the Gaussian RBF kernel. See [1]. For $X_a, X_b \in \mathscr{H}$, the squared distance between two elements in $\mathfrak{M}$ is

$$\|\kappa_X(\cdot, X_a) - \kappa_X(\cdot, X_b)\|_{\mathfrak{M}}^2 = \kappa_X(X_a, X_a) + \kappa_X(X_b, X_b) - 2\kappa_X(X_a, X_b) = 2\{1 - \kappa_X(X_a, X_b)\} = 2\{1 - \exp(-\gamma \|X_a - X_b\|_{\mathscr{H}}^2)\}.$$

By Taylor approximation,

$$\|\kappa_X(\cdot, X_a) - \kappa_X(\cdot, X_b)\|_{\mathfrak{M}}^2 = 2\{1 - 1 - \gamma \|X_a - X_b\|_{\mathscr{H}}^2 + O(\gamma^2)\} = \gamma \|X_a - X_b\|_{\mathscr{H}}^2 + O(\gamma^2).$$

Consequently, when $\gamma$ is small, the NAFPCA can be approximated by the linear FPCA, which is based on the distance $\|X_a - X_b\|_{\mathscr{H}}$.

In the rest of this subsection, we propose a two-step procedure to select the tuning parameter $\gamma$ in (2) and the number of principal components $d$ in (3) as follows.

1. Find the number of NAF-principal components by the following BIC-type criterion:

$$d = \operatorname{argmax}\{G_n(k) = \sum_{i=1}^{k} \hat{\lambda}_i - \log(n+1)^{-1} k : k \in \{1, 2, \ldots\}\}, \tag{8}$$

   where $\hat{\lambda}_i$ are the eigenvalues in the optimization problem in Sections 3.3 and 3.4.
2. Find $\gamma_1, \ldots, \gamma_p$ which maximize

$$\operatorname{var}_n\{\hat{\phi}^{(1)}(X) + \cdots + \hat{\phi}^{(d)}(X)\}, \tag{9}$$

   where $\operatorname{var}_n$ is the sample variance based on the sample $X_1, \ldots, X_n$ and $\hat{\phi}^{(1)}(\cdot), \ldots, \hat{\phi}^{(d)}(\cdot)$ are the first $d$-principal components computed with the reproducing kernel, $\kappa_j(a, b) = \exp(-\gamma_j \|a - b\|_{\mathscr{H}_j}^2), j \in \{1, \ldots, p\}$.

Since the goal of NAFPCA is to search for a nonlinear function of $X$ which maximizes the variance of predictors, it is natural to let (9) be the objective function for choosing $\gamma$ as well. The proposed estimator of $d$ in the first step is similar to the BIC-type criteria in [18] and [22]. It can be justified by the following lemma.

**Lemma 2.** *Suppose that $\hat{\Sigma}_{XX}$ is a random operator that converges to $\Sigma_{XX}$ in probability, $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots \geq 0$ are the eigenvalues of $\hat{\Sigma}_{XX}$ and $\lambda_1 > \lambda_2 > \cdots > \lambda_d > \lambda_{d+1} = 0$ are the eigenvalues of $\Sigma_{XX}$. Suppose that $\hat{\lambda}_k = \lambda_k + O_p(a_n)$ for some positive sequence $\{a_n\}$ such that $a_n = o(\frac{1}{\log(n+1)})$. Let $G_n(k) = \sum_{i=1}^{k} \hat{\lambda}_i - \log(n+1)^{-1} k$ and $\hat{d} = \operatorname{argmax}\{G_n(k) : k \in \{1, 2, \ldots, \}\}$. Then*

$$P(\hat{d} = d) \to 1.$$

**Proof of Lemma 2.** When $k < d$,

$$G_n(k) - G_n(d) = -\sum_{i=k+1}^{d} \hat{\lambda}_i - \log(n+1)^{-1}(k-d) = -\sum_{i=k+1}^{d} \{\lambda_i + O_p(a_n)\} - o(1) = -\sum_{i=k+1}^{d} \lambda_i + o_p(1).$$

Since the eigenvalues $\lambda_{k+1}, \ldots, \lambda_d$ are positive, we have $P(G_n(k) < G_n(d)) \to 1$.

When $k > d$,

$$G_n(k) - G_n(d) = \sum_{i=d+1}^{k} \hat{\lambda}_i - log(n+1)^{-1}(k-d) = 0 + O_p(a_n) - \log(n+1)^{-1}(k-d),$$

where, because $a_n = o(\frac{1}{\log(n+1)})$, the negative term $-\log(n+1)^{-1}(k-d)$ dominates the right hand side. Hence $P(G_n(k) < G_n(d)) \to 1$. □

In practice, we find the $\gamma$ in Step 2 by maximizing (9) via a grid search over $\rho \in [10^{-8}, 10^2]$ defined by

$$\gamma = \rho/(2\sigma^2), \quad \sigma^2 = \binom{n}{2}^{-1} \sum_{1 \le a < b \le n} \|X_a - X_b\|^2. \tag{10}$$

The simulation results in Section 5 show that this tuning method is very effective. We should mention that the above estimator of $d$ can also be applied to nonlinear sufficient dimension reduction in [17] and [22] to determine the number of sufficient predictors. In that context, we replace step 2 by the k-fold cross-validation as used in [22].

## 4. Asymptotic results

Since our method involves two layers of function spaces, $\mathscr{H}$ and $\mathfrak{M}$, if we can assume that the functions, $X_1, \ldots, X_n$, are observed in their entirety, then $\kappa_X(\cdot, X_1), \ldots, \kappa_X(\cdot, X_n)$ are an i.i.d. sample of random elements in $\mathfrak{M}$. In the functional data analysis literature, asymptotic theories of random elements in a generic Hilbert space have been developed and can be adapted to the current setting. See Hall et al. [9], Horváth and Kokoszka [12], and Hsing and Eubank [13]. Our asymptotic development consists of two steps. First, we show the asymptotic normality assuming each $X_a$ is fully observed for all $t$. We then derive the convergence rate of NAFPCA allowing $X_a$ to be partially observed at a set of sampled time points, and the rate is adapted to the convergence rate of the estimate of $X_a$.

### 4.1. Asymptotic distribution of estimated covariance operator

Assume first $p = 1$. Suppose $X_1, \ldots, X_n$ are fully observed as i.i.d. random elements in $\mathscr{H}$. Since the reproducing kernel $\kappa_X(\cdot, \cdot)$ is a known function, $\kappa_X(\cdot, X_1), \ldots, \kappa_X(\cdot, X_n)$ are i.i.d. copies of the random element $\kappa_X(\cdot, X)$ in the separable Hilbert space $\mathfrak{M}$. The sample mean element, $\hat{\mu}_X \equiv n^{-1} \sum_{a=1}^{n} \kappa_X(\cdot, X_a)$, and the sample covariance operator, $\hat{\Sigma}_{XX}$, are random elements in $\mathfrak{M}$ and $\mathscr{B}(\mathfrak{M}, \mathfrak{M})$ respectively. Let $\mu_X$ denote the mean element, $E\kappa_X(\cdot, X)$, in $\mathfrak{M}$. Since $E\|\kappa_X(\cdot, X)\|_{\mathfrak{M}}^2 = E[\langle \kappa_X(\cdot, X), \kappa_X(\cdot, X)\rangle_{\mathfrak{M}}] = E[\kappa_X(X, X)]$ and $E\|\kappa_X(\cdot, X)\|^4 = E[\kappa_X(X, X)^2]$, if we assume $E[\kappa_X(X, X)]$ or $E[\kappa_X(X, X)^2]$ to be finite (the Gaussian RBF satisfies both conditions), then the following theorems are immediately obtained by the central limit theorem in a separable Hilbert space (see [13]).

**Theorem 1** (*CLT in a Hilbert Space*). *If $U_1, \ldots, U_n$ are independent and identically distributed random elements in a separable Hilbert space $\mathscr{H}$ with mean 0 and $E\|U_a\|_{\mathscr{H}}^2 < \infty$, then*

$$\sqrt{n} \sum_{a=1}^{n} U_a \xrightarrow{\mathscr{D}} F,$$

*where $F$ is a Gaussian random elements in $\mathscr{H}$ with mean zero and covariance operator equal to $E(U_1 \otimes U_1)$.*

In the following, for two Hilbert spaces, say $\mathfrak{M}_1$, $\mathfrak{M}_2$, $\mathscr{B}_2(\mathfrak{M}_1, \mathfrak{M}_2)$ denote the collection of all Hilbert space–Schmidt operators from $\mathfrak{M}_1$ to $\mathfrak{M}_2$. Note that $\mathscr{B}_2(\mathfrak{M}_1, \mathfrak{M}_2)$ is a Hilbert space, and let us denote its inner product by $\langle \cdot, \cdot \rangle_{HS}$. For two operators $A, B \in \mathscr{B}_2(\mathfrak{M}_1, \mathfrak{M}_2)$, we define their tensor product exactly as we define the tensor product between two elements in any Hilbert space; that is, $A \otimes B$ is the linear operator from $\mathscr{B}_2(\mathfrak{M}_1, \mathfrak{M}_2)$ to $\mathscr{B}_2(\mathfrak{M}_1, \mathfrak{M}_2)$ such that, for any $C \in \mathscr{B}_2(\mathfrak{M}_1, \mathfrak{M}_2)$, $(A \otimes B)C = A\langle B, C\rangle_{HS}$. We next apply Theorem 1 to $U = \kappa_X(\cdot, X) - \mu_X$ and $U = [\kappa_X(\cdot, X) - \mu_X] \otimes [\kappa_X(\cdot, X) - \mu_X] - \Sigma_{XX}$ to obtain the following result.

**Theorem 2.** *If $E[\kappa_X(X, X)] < \infty$, then*

$$\sqrt{n}(\hat{\mu}_X - \mu_X) \xrightarrow{\mathscr{D}} N(0, \Sigma_{XX}).$$

*In addition, if* $E[\kappa_X(X,X)]^2 < \infty$, *then* $\hat{\Sigma}_{XX} - \Sigma_{XX}$ *is a Hilbert–Schmidt operator in* $\mathscr{B}_2(\mathfrak{M},\mathfrak{M})$, *and*

$$\sqrt{n}(\hat{\Sigma}_{XX} - \Sigma_{XX}) \xrightarrow{\mathscr{D}} N(0, \Gamma),$$

*where*

$$\Gamma = E\{[(\kappa_X(\cdot,X) - \mu_X) \otimes (\kappa_X(\cdot,X) - \mu_X) - \Sigma_{XX}] \otimes [(\kappa_X(\cdot,X) - \mu_X) \otimes (\kappa_X(\cdot,X) - \mu_X) - \Sigma_{XX}]\}.$$

Next, we consider the case $p > 1$. The following lemma can be proved by straightforward calculation.

**Lemma 3.** *Let* $\phi(\cdot,X) = \sum_{i=1}^{p} \kappa_i(\cdot,X^i)$, $\hat{\mu}_X = E_n\phi(\cdot,X)$, *and* $\mu_X = E\phi(\cdot,X)$. *Then* $\phi(\cdot,X)$ *is a random element in* $\mathfrak{M}$ *and*

$$\Sigma_{XX} = E[(\phi(\cdot,X) - \mu_X) \otimes (\phi(\cdot,X) - \mu_X)], \quad \hat{\Sigma}_{XX} = E_n[(\phi(\cdot,X) - \hat{\mu}_X) \otimes (\phi(\cdot,X) - \hat{\mu}_X)].$$

*where* $\Sigma_{XX}$ *and* $\hat{\Sigma}_{XX}$ *are defined in Sections 2 and 3 respectively.*

The next theorem is a generalization of Theorem 2 when $p > 1$.

**Theorem 3.** *If* $E[\kappa_i(X^i,X^i)]^2 < \infty$ *for* $i \in \{1,\dots,p\}$, *then*

$$\sqrt{n}(\hat{\mu}_X - \mu_X) \xrightarrow{\mathscr{D}} N(0, \Sigma_{XX}), \quad \sqrt{n}(\hat{\Sigma}_{XX} - \Sigma_{XX}) \xrightarrow{\mathscr{D}} N(0, \Gamma),$$

*where* $\Gamma = E[(\Sigma_0(X) - \Sigma_{XX}) \otimes (\Sigma_0(X) - \Sigma_{XX})]$ *and* $\Sigma_0(X) = \{\phi(\cdot,X) - E(\phi(\cdot,X))\} \otimes \{\phi(\cdot,X) - E(\phi(\cdot,X))\}$.

**Proof of Theorem 3.** By Theorem 1, it suffices to show that

$$E\|[\phi(\cdot,X) - E(\phi(\cdot,X))] \otimes [\phi(\cdot,X) - E(\phi(\cdot,X))] - \Sigma_{XX}\|_{HS}^2 < \infty, \tag{11}$$

where $\|\cdot\|_{HS}$ is the norm induced by the Hilbert–Schmidt inner product. Since $E\|Y - E(Y)\|^2 = E\|Y\|^2 - \|EY\|^2 \le E\|Y\|^2$ for any random element $Y$, the left hand side of (11) is bounded by

$$\begin{aligned} E\|[\phi(\cdot,X) - E(\phi(\cdot,X))] \otimes [\phi(\cdot,X) - E(\phi(\cdot,X))]\|_{HS}^2 &= E\|\{\phi(\cdot,X) - E(\phi(\cdot,X))\}\|_{\mathfrak{M}}^4 \le E\|\{\phi(\cdot,X)\}\|_{\mathfrak{M}}^4 \\ &= \sum_{i=1}^{p} E[\kappa_i(X^i,X^i)]^2 < \infty, \end{aligned}$$

as desired. □

*4.2. Asymptotic distribution of the nonlinear additive functional principal components*

We now develop the asymptotic normality of the eigenvalues and eigenfunctions of $\hat{\Sigma}_{XX}$. Let $(\lambda_1,\phi_1),\dots,(\lambda_d,\phi_d)$ be the first $d$-pairs of eigenvalue and eigenfunctions of $\Sigma_{XX}$ with $\lambda_1 > \cdots > \lambda_d$. Let $\langle\cdot,\cdot\rangle_{\otimes}$ be the inner product in the tensor product space, $\mathfrak{M} \otimes \mathfrak{M}$, and $\langle\cdot,\cdot\rangle_{\otimes^2}$ be the inner product in the tensor product space, $(\mathfrak{M} \otimes \mathfrak{M}) \otimes (\mathfrak{M} \otimes \mathfrak{M})$. See [16].

**Theorem 4.** *If* $E[\kappa_X(X,X)^2] < \infty$ *and the first* $d$ *eigenvalues of* $\Sigma_{XX}$ *are distinct, then for each* $j \in \{1,\dots,d\}$,

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \xrightarrow{\mathscr{D}} N(0, \langle\Gamma, \phi_j \otimes \phi_j \otimes \phi_j \otimes \phi_j\rangle_{\otimes^2}),$$

*and*

$$\sqrt{n}(\hat{c}_j\hat{\phi}_j - \phi_j) \xrightarrow{\mathscr{D}} N(0, C_j),$$

*where* $\hat{c}_j = \operatorname{sign}(\langle\phi_j,\hat{\phi}_j\rangle_{\mathfrak{M}})$, *and*

$$C_j = \sum_{k\neq j}\sum_{\ell\neq j}(\lambda_j - \lambda_k)^{-1}(\lambda_j - \lambda_\ell)^{-1}\langle\Gamma, \phi_k \otimes \phi_j \otimes \phi_\ell \otimes \phi_j\rangle_{\otimes^2}(\phi_k \otimes \phi_\ell).$$

**Proof of Theorem 4.** Let $\Sigma_F$ be a random element in $\mathscr{B}_2(\mathfrak{M},\mathfrak{M})$ having the limiting distribution $N(0,\Gamma)$ in Theorem 3. Then, by Theorem 9.1.3 of [13],

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \xrightarrow{\mathscr{D}} \langle\Sigma_F\phi_j, \phi_j\rangle_{\mathfrak{M}},$$

and

$$\sqrt{n}(\hat{\phi}_j - \phi_j) \xrightarrow{\mathscr{D}} \sum_{k\neq j}(\lambda_j - \lambda_k)^{-1}P_k\Sigma_F\phi_j,$$

where $P_k$ is the orthogonal projection operator onto $k$th eigenspace which is the space spanned by the $k$th eigenfunction $\phi_k$.

Since $\Sigma_F \sim N(0, \Gamma)$, $\langle \Sigma_F \phi_j, \phi_j \rangle_{\mathfrak{M}} = \langle \Sigma_F, \phi_j \otimes \phi_j \rangle_{\otimes}$ is a normal random variable with mean zero and variance

$$\begin{aligned}\operatorname{var}[\langle \Sigma_F, \phi_j \otimes \phi_j \rangle_{\otimes}] &= E[\langle \Sigma_F, \phi_j \otimes \phi_j \rangle_{\otimes}^2] = E[\langle \Sigma_F \otimes \Sigma_F, (\phi_j \otimes \phi_j) \otimes (\phi_j \otimes \phi_j) \rangle_{\otimes^2}] \\ &= \langle E[\Sigma_F \otimes \Sigma_F], (\phi_j \otimes \phi_j) \otimes (\phi_j \otimes \phi_j) \rangle_{\otimes^2} \\ &= \langle \Gamma, \phi_j \otimes \phi_j \otimes \phi_j \otimes \phi_j \rangle_{\otimes^2},\end{aligned}$$

where the third equality follows from the definition of expectation in a generic Hilbert space. This completes the proof of the asymptotic normality of eigenvalues. Similarly, $\sum_{k \neq j} (\lambda_j - \lambda_k)^{-1} P_k \Sigma_F \phi_j$ is a Gaussian element in $\mathfrak{M}$ with mean 0 and variance operator

$$\begin{aligned}\operatorname{var}\left[\sum_{k \neq j} (\lambda_j - \lambda_k)^{-1} P_k \Sigma_F \phi_j\right] &= E\left[\left\{\sum_{k \neq j} (\lambda_j - \lambda_k)^{-1} P_k \Sigma_F \phi_j\right\} \otimes \left\{\sum_{\ell \neq j} (\lambda_j - \lambda_\ell)^{-1} P_\ell \Sigma_F \phi_j\right\}\right] \\ &= \sum_{k \neq j} \sum_{\ell \neq j} (\lambda_j - \lambda_k)^{-1} (\lambda_j - \lambda_\ell)^{-1} E[(\phi_k \otimes \phi_k \Sigma_F \phi_j) \otimes (\phi_\ell \otimes \phi_\ell \Sigma_F \phi_j)] \\ &= \sum_{k \neq j} \sum_{\ell \neq j} (\lambda_j - \lambda_k)^{-1} (\lambda_j - \lambda_\ell)^{-1} E[(\phi_k \langle \phi_k, \Sigma_F \phi_j \rangle) \otimes (\phi_\ell \langle \phi_\ell, \Sigma_F \phi_j \rangle)] \\ &= \sum_{k \neq j} \sum_{\ell \neq j} (\lambda_j - \lambda_k)^{-1} (\lambda_j - \lambda_\ell)^{-1} E[\langle \phi_k, \Sigma_F \phi_j \rangle \langle \phi_\ell, \Sigma_F \phi_j \rangle] \phi_k \otimes \phi_\ell \\ &= \sum_{k \neq j} \sum_{\ell \neq j} (\lambda_j - \lambda_k)^{-1} (\lambda_j - \lambda_\ell)^{-1} E[\langle \Sigma_F \otimes \Sigma_F, \phi_k \otimes \phi_j \otimes \phi_\ell \otimes \phi_j \rangle] \phi_k \otimes \phi_\ell.\end{aligned}$$

This completes the proof. □

### 4.3. Incompletely observed functional data

Our asymptotic development so far is for the ideal case where the random functions $X_a(t)$, $a \in \{1, \ldots, n\}$, are observed for all $t \in T$. In practice, they are observed on a set of regularly or irregularly placed time points, say $S_a^i = \{t_{a_j}^i : j \in \{1, \ldots, k_a^i\}\}$, and we must first estimate $X_a^i$ based on their observed values at $S_a^i$, and the error incurred by this estimation should be taken into account in a careful asymptotic analysis. For notational simplicity, in this section we assume $p = 1$, and write $k_a^i$, $S_a^i$, $t_{a_j}^i$, $X_a^i$ as $k_a$, $S_a$, $t_{a_j}$, $X_a$. Wang et al. [30] reviewed a variety of methods for estimating $X_a$ under different measurement schedules, such as the dense schedule, under which $X_a$ can be estimated at the $\sqrt{n}$ rate, and the sparse schedule, under which $X_a$ can be estimated at a slower-than-$\sqrt{n}$ rate. They also discussed different convergence rates of the functional estimates under these various scenarios. In this subsection we assume that $X_a$ has been estimated by some method at a general rate $0 < \delta_n \to 0$, and investigate how this error propagates into our final results.

Henceforth, for two positive sequences $c_n$ and $d_n$, we write $c_n \prec d_n$ if $c_n/d_n \to 0$, write $c_n \succ d_n$ if $d_n \prec c_n$, and write $c_n \asymp d_n$ if there exist numbers $0 < C_1 < C_2 < \infty$ such that $C_1 < c_n/d_n < C_2$.

A reasonable asymptotic regime is to assume $k_a$ to be a function of $n$ that goes to infinity, that is, $k_a = m_n(a)$, with $\lim_{n\to\infty} m_n(a) = \infty$. For simplicity, we assume $m_n(a)$ to be the same for all $a$, and denote this common number by $m_n$. There are three asymptotic scenarios $m_n \prec n$, $m_n \asymp n$, and $m_n \succ n$, all of which are possible in practice: the first is appropriate for the situations where the number of observations on each $X_a$ is much larger than the number of subjects; the last for the opposite situations; the second for the situations where the two numbers are similarly. Depending on the estimators used and the relation between $m_n$ and $n$, the convergence rate $\delta_n$ may also have three scenarios:

$$\delta_n \prec \sqrt{n}, \ \ \delta_n \succ \sqrt{n}, \ \text{ or } \delta_n \asymp \sqrt{n},$$

which will determine the net asymptotic behavior of the NAFPCA.

Suppose, then, for each $a \in \{1, \ldots, n\}$, an estimate $\hat{X}_a$ of $X_a$ converges to $X_a$ at a rate $\delta_n$. In principle, this can be formulated as $\|\hat{X}_a - X_a\|_{\mathscr{H}} = O_P(\delta_n)$, but to simplify the theory, we make the slightly stronger assumption

$$E(\|\hat{X}_a - X_a\|_{\mathscr{H}}) = O(\delta_n).$$

Since $X_1, \ldots, X_n$ are i.i.d. random elements, it is also reasonable to assume that $\hat{X}_1, \ldots, \hat{X}_n$ are i.i.d. random elements. This is true, for example, if we use the same method to construct each $\hat{X}_a$, and only use the observations on $X_a$ to construct $\hat{X}_a$.

The following identities regarding the Hilbert–Schmidt inner product of two tensor products will be used repeatedly: if $f_1, \ldots, f_4 \in \mathfrak{M}$, then $\langle f_1 \otimes f_2, f_3 \otimes f_4 \rangle_{\mathrm{HS}} = \langle f_1, f_3 \rangle_{\mathfrak{M}} \langle f_2, f_4 \rangle_{\mathfrak{M}}$. Applying this to $f_3 = f_1$ and $f_4 = f_2$, we have

$$\|f_1 \otimes f_2\|_{\mathrm{HS}} = \|f_1\|_{\mathfrak{M}} \|f_2\|_{\mathfrak{M}}. \tag{12}$$

Let

$$\hat{\Sigma}_{XX} = E_n\{[\kappa(\cdot, \hat{X}) - E_n\kappa(\cdot, \hat{X})] \otimes [\kappa(\cdot, \hat{X}) - E_n\kappa(\cdot, \hat{X})]\}, \quad \tilde{\Sigma}_{XX} = E_n\{[\kappa(\cdot, X) - E_n\kappa(\cdot, X)] \otimes [\kappa(\cdot, X) - E_n\kappa(\cdot, X)]\}.$$

The operator $\hat{\Sigma}_{XX}$ is based on the estimated functional data, whose asymptotic behavior is our objective, whereas $\tilde{\Sigma}_{XX}$ is an intermediate operator introduced only for the proof. The next theorem establishes the convergence rate of $\hat{\Sigma}_{XX}$.

**Theorem 5.** *Suppose*

1. *(boundedness) there is a constant* $0 < C < \infty$ *such that* $\kappa(x, x) \le C$ *for each* $x \in \mathscr{H}$;
2. *(Lipschitz) there is a constant* $0 < C_1 < \infty$ *such that, for any* $x_1, x_2 \in \mathscr{H}$, $\|\kappa(\cdot, x_1) - \kappa(\cdot, x_2)\|_{\mathfrak{M}} \le C_1 \|x_1 - x_2\|_{\mathscr{H}}$.
3. *(curve estimate)* $\hat{X}_1, \ldots, \hat{X}_n$ *are i.i.d. random elements in* $\mathscr{H}$ *with* $E\|\hat{X}_a - X_a\|_{\mathscr{H}} = O(\delta_n)$ *for some* $0 < \delta_n \to 0$.

*Then*

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

It is easy to verify that the boundedness condition and the Lipschitz kernel condition are satisfied by the Gaussian radial basis function. As will be discussed later in this section, these uniform conditions can be relaxed to moment conditions.

**Proof of Theorem 5.** By the triangular inequality,

$$\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{\mathrm{HS}} \le \|\hat{\Sigma}_{XX} - \tilde{\Sigma}_{XX}\|_{\mathrm{HS}} + \|\tilde{\Sigma}_{XX} - \Sigma_{XX}\|_{\mathrm{HS}}.$$

Using essentially the same argument as that for Lemma 5 of Fukumizu, Bach, and Gretton (2007), we can prove that

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

So we only need to prove

$$\|\hat{\Sigma}_{XX} - \tilde{\Sigma}_{XX}\|_{\mathrm{HS}} = O_P(\delta_n). \tag{13}$$

First note that

$$\tilde{\Sigma}_{XX} = E_n[\kappa(\cdot, X) \otimes \kappa(\cdot, X)] - E_n[\kappa(\cdot, X)] \otimes E_n[\kappa(\cdot, X)], \quad \hat{\Sigma}_{XX} = E_n[\kappa(\cdot, \hat{X}) \otimes \kappa(\cdot, \hat{X})] - E_n[\kappa(\cdot, \hat{X})] \otimes E_n[\kappa(\cdot, \hat{X})].$$

Hence

$$\tilde{\Sigma}_{XX} - \hat{\Sigma}_{XX} = E_n[\kappa(\cdot, X) \otimes \kappa(\cdot, X)] - E_n[\kappa(\cdot, \hat{X}) \otimes \kappa(\cdot, \hat{X})] + E_n[\kappa(\cdot, \hat{X})] \otimes E_n[\kappa(\cdot, \hat{X})] - E_n[\kappa(\cdot, X)] \otimes E_n[\kappa(\cdot, X)].$$

The above can be written as the sum of the following six terms:

$$\begin{aligned}
&I_1 = E_n\{[\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)] \otimes [\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)]\}, \quad I_2 = E_n\{[\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)] \otimes \kappa(\cdot, X)\}, \\
&I_3 = E_n\{\kappa(\cdot, X) \otimes [\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)]\}, \quad I_4 = E_n[\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)] \otimes E_n[\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)], \\
&I_5 = E_n[\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)] \otimes E_n[\kappa(\cdot, X)], \quad I_6 = E_n[\kappa(\cdot, X)] \otimes E_n[\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)].
\end{aligned} \tag{14}$$

It is easy to see that $\|I_2\|_{\mathrm{HS}} = \|I_3\|_{\mathrm{HS}}$ and $\|I_5\|_{\mathrm{HS}} = \|I_6\|_{\mathrm{HS}}$. So it suffices to derive the order of magnitudes of the norms of $I_1, I_3, I_4, I_6$.

For $I_3$, we have, by the triangular inequality, identity (12), and conditions *1* and *2*,

$$\begin{aligned}
\|E_n\{\kappa(\cdot, X) \otimes [\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)]\}\|_{\mathrm{HS}} &\le E_n\|\kappa(\cdot, X) \otimes [\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)]\|_{\mathrm{HS}} = E_n\|\kappa(\cdot, X)\|_{\mathfrak{M}} \|\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)\|_{\mathfrak{M}} \\
&\le \sqrt{C}\, C_1 E_n\|\hat{X} - X\|_{\mathscr{H}}.
\end{aligned}$$

By Markov's inequality, for any $K > 0$,

$$P\left(\delta_n{}^{-1}\|E_n\{\kappa(\cdot, X) \otimes [\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)]\}\|_{\mathrm{HS}} > K\right) \le P\left(\delta_n{}^{-1}\sqrt{C}\, C_1 E_n\|\hat{X} - X\|_{\mathscr{H}} > K\right) \le K^{-1}\delta_n^{-1}\sqrt{C}C_1 E\|\hat{X} - X\|_{\mathscr{H}}.$$

By condition *3*, $\delta_n^{-1} E\|\hat{X} - X\|_{\mathscr{H}} < C_2$ for some $0 < C_2 < \infty$. Hence the right-hand side above is bounded by $K^{-1}\sqrt{C}C_1C_2$, which can be made arbitrarily small by choosing a sufficiently large $K$. Thus $\|I_3\|_{\mathrm{HS}} = O_P(\delta_n)$.

For $I_1$, we have, again by the triangular inequality and identity (12),

$$\begin{aligned}
\|E_n\{[\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)] \otimes [\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)]\}\|_{\mathrm{HS}} &\le E_n\|[\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)] \otimes [\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)]\|_{\mathrm{HS}} \\
&= E_n\|\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)\|_{\mathfrak{M}}^2.
\end{aligned}$$

Furthermore,

$$\|\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)\|_{\mathfrak{M}}^2 \le (\|\kappa(\cdot, \hat{X})\|_{\mathfrak{M}} + \|\kappa(\cdot, X)\|_{\mathfrak{M}})\, \|\kappa(\cdot, \hat{X}) - \kappa(\cdot, X)\|_{\mathfrak{M}} \le 2\sqrt{C}C_1\, \|\hat{X} - X\|_{\mathscr{H}},$$

where the second inequality follows from conditions *1* and *2*. Apply Markov's inequality in the same way as before to obtain $I_1 = O_P(\delta_n)$.

Similarly, for $I_6$, we have, by the triangular inequality, identity (12), and conditions *1*, *2*,

$$\|E_n[\kappa(\cdot,X)]\otimes E_n[\kappa(\cdot,\hat{X})-\kappa(\cdot,X)]\|_{\mathrm{HS}} \le \frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n}\|\kappa(\cdot,X_a)\otimes[\kappa(\cdot,\hat{X}_b)-\kappa(\cdot,X_b)]\|_{\mathrm{HS}}$$
$$= \frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n}\|\kappa(\cdot,X_a)\|_{\mathfrak{M}}\,\|\kappa(\cdot,\hat{X}_b)-\kappa(\cdot,X_b)\|_{\mathfrak{M}} \le \sqrt{C}C_1\frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n}\|\hat{X}_b-X_b\|_{\mathscr{H}}.$$

Consequently,

$$E(\|E_n[\kappa(\cdot,X)]\otimes E_n[\kappa(\cdot,\hat{X})-\kappa(\cdot,X)]\|_{\mathrm{HS}}) \le \sqrt{C}C_1E\|\hat{X}-X\|_{\mathscr{H}}.$$

Apply Markov's inequality again to obtain $\|I_6\|_{\mathrm{HS}} = O_P(\delta_n)$.

For $I_4$, we have

$$\|E_n[\kappa(\cdot,\hat{X})-\kappa(\cdot,X)]\otimes E_n[\kappa(\cdot,\hat{X})-\kappa(\cdot,X)]\|_{\mathrm{HS}} \le n^{-2}\sum_{a=1}^{n}\sum_{b=1}^{n}\|[\kappa(\cdot,\hat{X}_a)-\kappa(\cdot,X_a)]\otimes[\kappa(\cdot,\hat{X}_b)-\kappa(\cdot,X_b)]\|_{\mathrm{HS}}$$
$$\le n^{-2}\sum_{a=1}^{n}\sum_{b=1}^{n}\|\kappa(\cdot,\hat{X}_a)-\kappa(\cdot,X_a)\|_{\mathfrak{M}}\|\kappa(\cdot,\hat{X}_b)-\kappa(\cdot,X_b)\|_{\mathfrak{M}} \le n^{-2}\sum_{a=1}^{n}\sum_{b=1}^{n}(\|\kappa(\cdot,\hat{X}_a)\|_{\mathfrak{M}}+\|\kappa(\cdot,X_a)\|_{\mathfrak{M}})\|\kappa(\cdot,\hat{X}_b)-\kappa(\cdot,X_b)\|_{\mathfrak{M}}$$
$$\le n^{-2}\sum_{a=1}^{n}\sum_{b=1}^{n}2\sqrt{C}C_1\|\hat{X}_b-X_b\|_{\mathfrak{M}}$$

Therefore,

$$E(\|E_n[\kappa(\cdot,\hat{X})-\kappa(\cdot,X)]\otimes E_n[\kappa(\cdot,\hat{X})-\kappa(\cdot,X)]\|_{\mathrm{HS}}) \le 2\sqrt{C}C_1E(\|\hat{X}-X\|_{\mathscr{H}}).$$

Applying Markov's inequality as before, we have $\|I_4\|_{\mathrm{HS}} = O_P(\delta_n)$. □

Conditions *1* and *2* of Theorem 5 are uniform in nature, but they can be relaxed in terms of moments. Let $\alpha > 1$, $\beta > 1$ be a conjugate pair; that is, $\alpha^{-1}+\beta^{-1}=1$. By Holder's inequality,

$$E(\|\kappa(\cdot,X)\|_{\mathfrak{M}}\,\|\kappa(\cdot,\hat{X})-\kappa(\cdot,X)\|_{\mathfrak{M}}) \le [E(\|\kappa(\cdot,X)\|_{\mathfrak{M}}^{\alpha})]^{1/\alpha}\,[E(\|\kappa(\cdot,\hat{X})-\kappa(\cdot,X)\|_{\mathfrak{M}}^{\beta})]^{1/\beta}.$$

Using this inequality we can derive the same asymptotic rate, as shown in the next theorem. The proof is similar, and therefore omitted.

**Theorem 6.** *The conclusion of Theorem 5 still holds if the first two conditions therein are replaced by the following conditions: for some conjugate pair* $(\alpha,\beta)\in(1,\infty)\times(1,\infty)$,

1. $E(\|\kappa(\cdot,X)\|_{\mathfrak{M}}^{\alpha})<\infty$ *and* $E(\|\kappa(\cdot,\hat{X})\|_{\mathfrak{M}}^{\alpha})<\infty$;
2. *there is a constant* $0<C_1<\infty$ *such that* $E(\|\kappa(\cdot,X)-\kappa(\cdot,\hat{X})\|_{\mathfrak{M}}^{\beta}) \le C_1(E\|X-\hat{X}\|_{\mathscr{H}})^{\beta}$.

The two theorems immediately imply that, if $\delta_n \prec n^{-1/2}$, then the asymptotic distributions developed in Sections 4.1 and 4.2 are still valid. If, depending on the method used and the $(n, m_n)$ asymptotic regime, $\delta_n \succ n^{-1/2}$ and ${\delta_n}^{-1}(\hat{\Sigma}_{XX}-\tilde{\Sigma}_{XX})$ converges in distribution to some random element $F$, then ${\delta_n}^{-1}(\hat{\Sigma}_{XX}-\Sigma_{XX})$ would converge to the same random element. Finally, under some mild conditions, the eigenfunctions of $\hat{\Sigma}_{XX}$ have the same convergence rate $O_P(\delta_n+n^{-1/2})$, and their asymptotic distributions remain the same as derived in Sections 4.1 and 4.2 if $\delta_n \prec n^{-1/2}$, or can be derived from the limiting random element $F$ if $\delta_n \succ n^{-1/2}$.

## 5. Simulation studies

In this section, we investigate the performance of our method under different scenarios for dimension reduction of functional data. In particular, we applied our NAFPCA and the multivariate functional PCA in [10] to one-dimensional and two-dimensional functional data to compare their performances and to demonstrate how the nonlinearity is captured by our method. In the one-dimensional functional data, each random function is observed at a set of equally-spaced time points; in the two-dimensional functional data, each random function is observed at a set of randomly spaced time points. The simulation studies consist of visualization via NAFPCA and comparisons of classification performance with the linear functional PCA.

### 5.1. Behavior of NAFPCA

We first consider two models (Model I-1 and Model I-2) with regularly observed univariate functional data. Model I-1 consists of two clusters of random functions with highly nonlinear features. First, we generate a random sample $Y_1,\ldots,Y_n$,

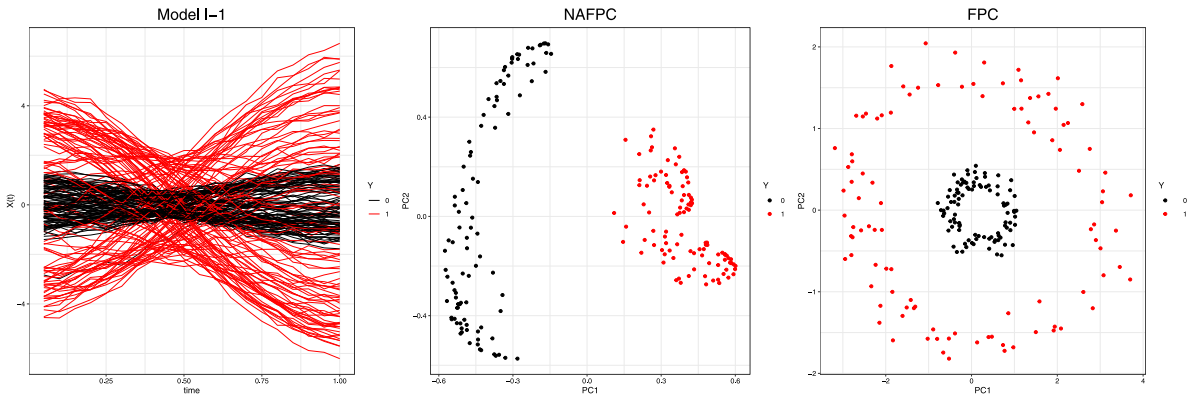


**Fig. 1.** (Model I-1) Left: observed curves; Middle: first two PCs from NAFPCA; Right: first two PCs from the FPCA. Black colored curves and points are $X_a$'s with $Y_a = 0$ and the red ones are those with $Y_a = 1$. The nonlinear features of the Model I-1 is well captured by the NAFPCA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where $n = 200$, from the Bernoulli distribution with $p = 0.5$. Then we generate random functions $X_1, \ldots, X_n$ from the following model:

$$\text{Model I-1}: \begin{cases} X_a(t)|Y_a = 0 & \sim Z_{1a}\cos(\theta_{1a})\cos(\pi t) + Z_{1a}\sin(\theta_{1a})\sin(t) + \epsilon_a(t), \\ X_a(t)|Y_a = 1 & \sim Z_{2a}\cos(\theta_{2a})\cos(\pi t) + Z_{2a}\sin(\theta_{2a})\sin(t) + \epsilon_a(t), \end{cases}$$

where $Z_{1a} \sim N(1, 0.2^2)$, $\theta_{1a} \sim U(0, 2\pi)$, $Z_{2a} \sim N(4, 0.5^2)$, $\theta_{2a} \sim U(0, 2\pi)$, and $\epsilon_a(t_j) \sim N(0, 0.1^2)$, and $Z_{1a}$, $\theta_{1a}$, $Z_{2a}$, $\theta_{2a}$, $\epsilon_a(t_j)$ are independently sampled. Of course, the labels $Y_a$ are not used in the analysis.

Each curve $X_a$ is sampled at 20 equally-spaced time points in $[0, 1]$. The simulated functions are shown in the left panel of Fig. 1, where the black curves are $X_a$'s with $Y_a = 0$ and the red ones are those with $Y_a = 1$. We apply our method and the linear FPCA to the data set. The middle and right panels in Fig. 1 show the scatter plots of the first two nonlinear principal components and the first two linear principal components, respectively. It is clear that the nonlinear feature in $X$ is well captured by our method.

A more direct representation of the results is to color each curve $X_a$ according to its PC score. In the upper-left panel of Fig. 2, we present the curves $X_a$ with their intensity of red colors scaled by the first PC scores of the NAFPCA. Thus, if the first PC score of $X_a$ is high, then the color of $X_a$ is close to red; if the score is low, then the color is close to black. The upper-right panel is the same representation for the second PC score by the NAFPCA. The lower panels are the corresponding representations for the FPCA. Comparing with FPCA, our method, NAFPCA, characterizes the shapes of the curves in different clusters more clearly than FPCA: the combination of the upper panels more closely resembles the left plot in Fig. 1 than does the combination of the lower panels of Fig. 2.

For a more comprehensive and compact visualization, in Fig. 3 we represent the first three PC scores by the RGB color scaling. In particular, we represent the first PC score by Red color scaling, the second by Green color, and the third by Blue color. The NAFPCA is in the left panel, and the FPCA is in the right panel. The left panel shows that the first NAFPC represents the curves with large amplitude, the second NAFPC represents the curves with small amplitude and upward trend, and the third NAFPC represents the curves with small amplitude and a downward trend. The right panel shows that linear index is insufficient to represent the complexity in the functional data.

Model I-2 also consists of two clusters of random functions, but this time they are both linear random elements in $\mathscr{H}$. Such a model favors the linear FPCA; our goal is to see how much efficiency is lost by the NAFPCA under this circumstance. Similar to Model I-1, we first generate a random sample $Y_1, \ldots, Y_n$ with $n = 200$ from the Bernoulli distribution with $p = 0.5$. We then generate $X_1, \ldots, X_n$ by

$$\text{Model I-2}: \begin{cases} X_a(t)|Y_a = 0 & \sim Z_{1a}(b_1(t) + b_2(t)) + Z_{2a}b_2(t) + \epsilon_a(t), \\ X_a(t)|Y_a = 1 & \sim (Z_{1a} + Z_{2a})(b_5(t) + b_6(t)) + \epsilon_a(t), \end{cases}$$

where $b_1(t), \ldots, b_6(t)$ are the 6 B-spline basis functions defined on $[0, 1]$, $Z_{1a}, Z_{2a}$ are i.i.d. $N(0, 2^2)$, $\epsilon_a(t_j)$ follows $N(0, 0.1^2)$ and is independent of $Z_{1a}$ and $Z_{2a}$. Fig. 4 shows that the linear FPCA perfectly captures the linear features of the two clusters. Meanwhile, the NAFPCA is also able to separate the two different clusters even though not as clearly as the linear FPCA.

Next, we investigate the effects of the tuning parameters $\gamma$ used in the kernel for the RKHS $\mathfrak{M}$, which determines the complexity of the space. Figs. 5 and 6 show the shape of the first two NAFPC's for a wider range of $\rho$, where $\rho$ is proportional to $\gamma$ as defined by (10): Fig. 5 for Model I-1 and Fig. 6 for Model I-2. We can see that with the small $\gamma$ ($\rho = 1$), they are very close to their linear FPC counterparts (the right panels of Figs. 1 and 4) and our tuning procedure performs adequately.

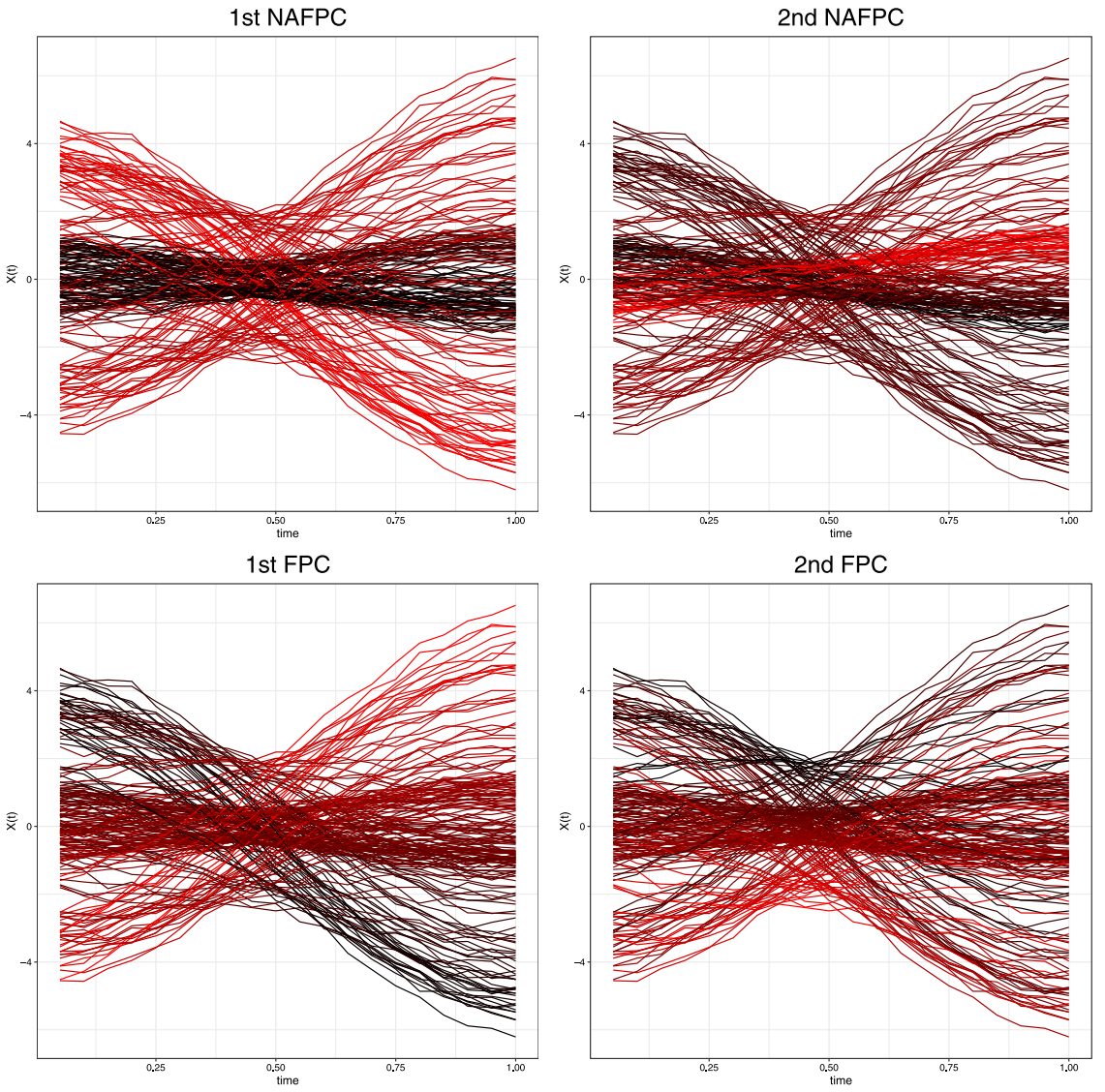


**Fig. 2.** (Model I-1) Observed curves colored with their intensity of red colors scaled by principal component scores. As the PC score is high, the color of the curve is close to red; if the score is low, then the color is close to black. The upper panels are for NAFPCA and the lower panels are for linear FPCA. The left panels are colored with the first PC scores and the right panels are colored with the second PC scores. NAFPCA characterizes the shape of the curves in different clusters more clearly than FPCA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We next consider a model (Model II) with bivariate functional data, also forming two clusters. We generate $Y_1, \ldots, Y_n$ with $n = 200$ as before and then generate $X_a = (X_a^1, X_a^2)$, $a = 1, \ldots, n$, according to

$$\text{Model II} \ : \begin{cases} (X_a^1(t), X_a^2(t))|Y_a = 0 \sim (Z_{1a}\cos(2\pi t), Z_{1a}\sin(2\pi t)), \\ (X_a^1(t), X_a^2(t))|Y_a = 1 \sim (2Z_{2a}\cos(2\pi t), Z_{2a}\sin(4\pi t)), \end{cases}$$

where $Z_{1a} \sim N(1, 0.4^2)$, $Z_{2a} \sim N(0.7, 0.2^2)$, and they are independent. Again, the functional data $X_a$ are observed on 20 time points in $[0, 1]$, but this time, the time points are different for different curves. For each curve, we randomly choose 20 time points from 200-equally spaced time points in $[0, 1]$. The left panel of Fig. 7 shows the observed points of the image of the realized functions with the black color representing the $Y_a = 0$ group and the red color representing the $Y_a = 1$ group. The middle panel shows the first two principal components by NAFPCA, and the right panel shows the first two principal components by the linear FPCA. The plots show that NAFPCA provides nearly perfect separation of the two clusters; whereas the linear FPCA hardly separates the two clusters at all.

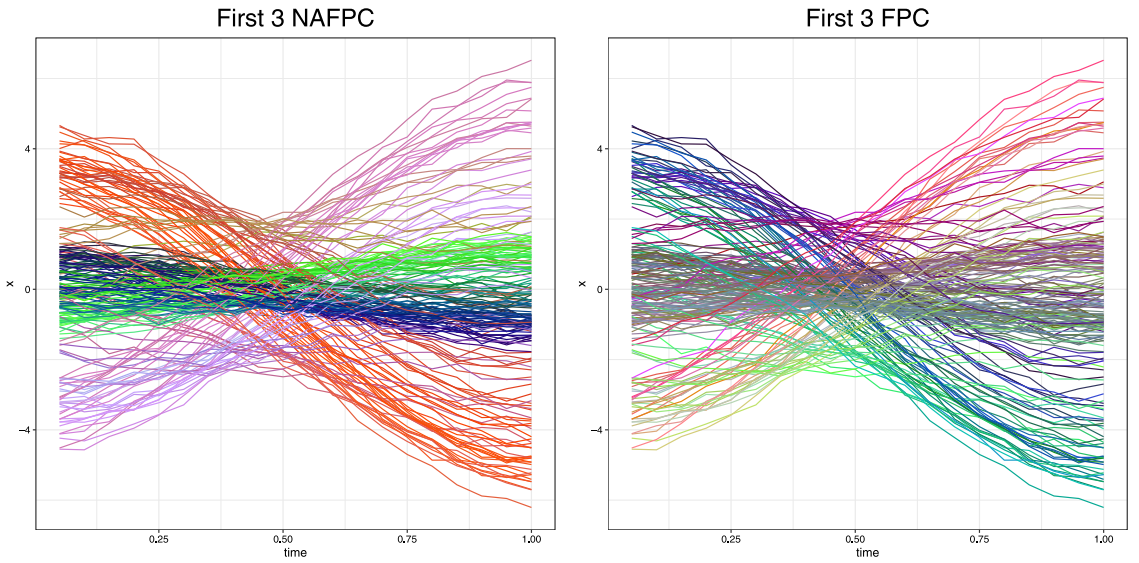


**Fig. 3.** (Model I-1) The first three PC scores, as represented by the RGB scales by NAFPCA (left panel) and FPCA (right panel). The left panel shows that the first NAFPC represents the curves with large amplitude, the second NAFPC represents the curves with small amplitude and upward trend, and the third NAFPC represents the curves with small amplitude and a downward trend. The right panel shows that linear index is insufficient to represent the complexity in the functional data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

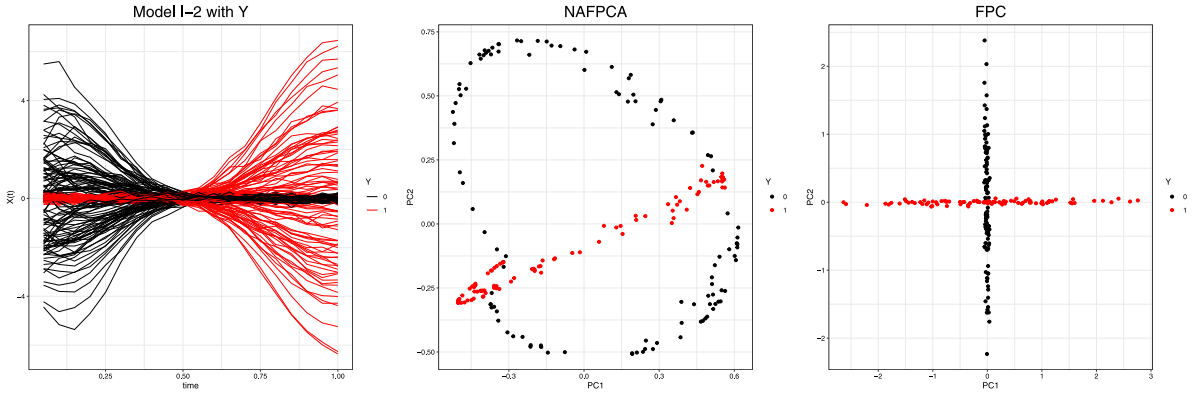


**Fig. 4.** (Model I-2) Left: observed curves; Middle: first two PCs from the NAFPCA; Right: first two PCs from the FPCA. When the model is linear, the FPCA perfectly captures each cluster. Meanwhile, the NAFPCA also separate the two clusters well.

### *5.2. Simulation in classification problems*

We next consider classification problems, where we use NAFPCA and FPCA as the preprocessing dimension reduction step, followed by a classification step with classifiers such as the linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and the support vector machines (SVM). In each classification model, we use three different sample sizes, 200, 400, 800, and three different numbers of observed time points, 20, 50, 100. We divide each sample into a training set and a test set of equal sample sizes. We first apply NAFPCA and FPCA to the training set to extract principal components, and then apply the three classifiers to the principal components to develop the classification rules. Finally, we apply the classification rules to the test set and record the percentages of correct classifications. The numbers of principal components are determined by five-fold cross-validation applied to the training set. To ensure the reliability of the results, we generate 100 samples for each model, and present the means and standard deviations of the percentages.

The first model we applied is Model I-2 in Section 5.1.1. Since the model consists of linearly related random elements, our method has a disadvantage. Nevertheless, the performances of classification are comparable. As shown in Table 1, NAFPCA works very well with all the classifiers and FPCA works well with QDA but does not work well with the other classifiers. In both cases, QDA is the best classifier and FPCA has a slightly better result. In all combinations of scenarios,

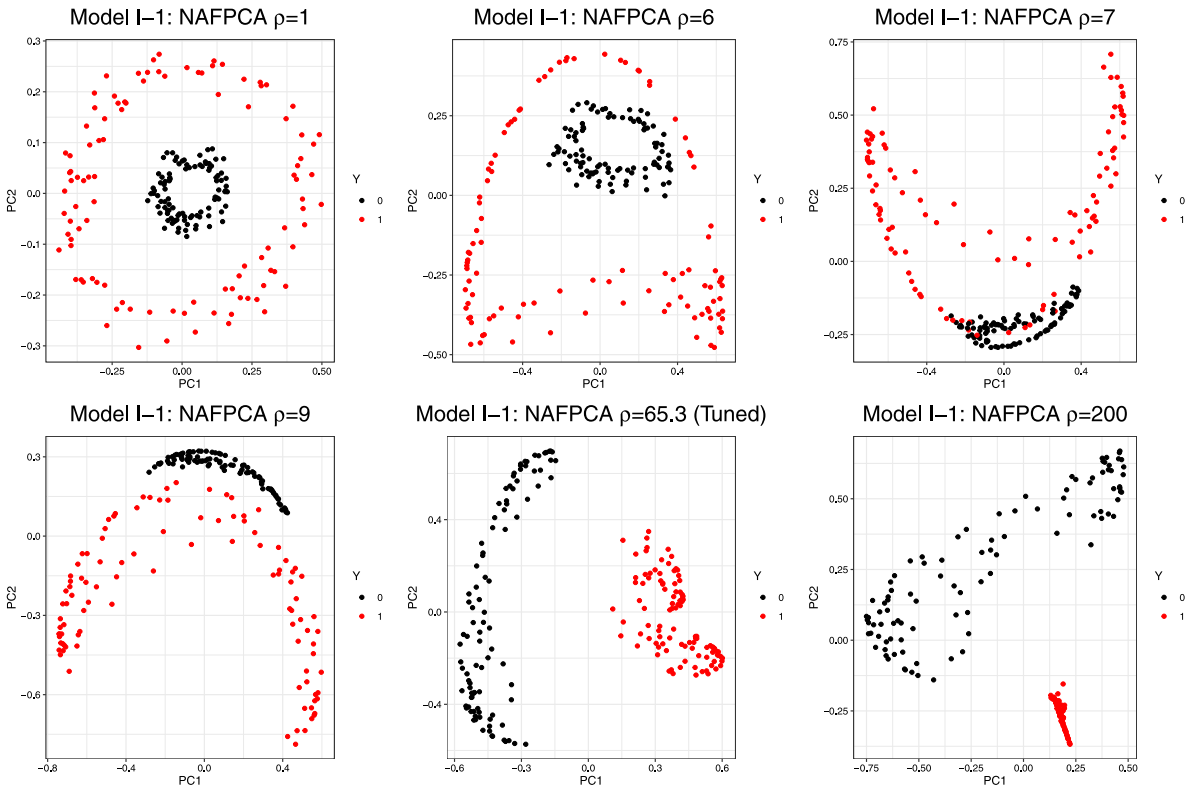


**Fig. 5.** (Model I-1) Effects of tuning parameter $\gamma$ used in the kernel for the RKHS $\mathfrak{M}$. The plots show first two PCs from NAFPCA with different $\rho$, where $\rho$ is proportional to $\gamma$. Top: $\rho = 1, 6, 7$, bottom: $\rho = 9, 65.3, 200$, where 65.3 is the tuned value from the selection procedure.

**Table 1**
Percentages of correct classifications for Model I-2 where $X_a$'s are linear random elements. Entries are the means and standard deviations (in parentheses) of the percentages calculated from 100 samples. In both cases, QDA performs the best. In all combinations of scenarios, the results for NAFPCA and FPCA are similar. The results get better as the sample size increases.

| Sample size | # Time points | NAFPCA | | | FPCA | | |
|---|---|---|---|---|---|---|---|
| | | LDA | QDA | SVM | LDA | QDA | SVM |
| 200 | 20 | 0.926 (0.043) | 0.977 (0.019) | 0.93 (0.038) | 0.569 (0.101) | 0.988 (0.015) | 0.904 (0.039) |
| | 50 | 0.922 (0.04) | 0.987 (0.013) | 0.944 (0.029) | 0.565 (0.097) | 0.993 (0.01) | 0.914 (0.038) |
| | 100 | 0.928 (0.033) | 0.991 (0.012) | 0.948 (0.028) | 0.572 (0.094) | 0.996 (0.007) | 0.908 (0.039) |
| 400 | 20 | 0.942 (0.023) | 0.982 (0.011) | 0.958 (0.023) | 0.569 (0.086) | 0.99 (0.008) | 0.918 (0.023) |
| | 50 | 0.946 (0.024) | 0.99 (0.009) | 0.965 (0.019) | 0.572 (0.079) | 0.997 (0.005) | 0.933 (0.021) |
| | 100 | 0.947 (0.023) | 0.991 (0.009) | 0.963 (0.018) | 0.588 (0.086) | 0.997 (0.005) | 0.93 (0.022) |
| 800 | 20 | 0.956 (0.014) | 0.981 (0.008) | 0.964 (0.014) | 0.563 (0.075) | 0.992 (0.005) | 0.939 (0.015) |
| | 50 | 0.953 (0.017) | 0.988 (0.006) | 0.97 (0.012) | 0.572 (0.084) | 0.997 (0.004) | 0.947 (0.013) |
| | 100 | 0.953 (0.015) | 0.992 (0.005) | 0.968 (0.011) | 0.558 (0.074) | 0.998 (0.003) | 0.947 (0.015) |

the results for NAFPCA and FPCA are similar. As expected, the percentages of correct classification increase as the sample size and the number of observed time points increase.

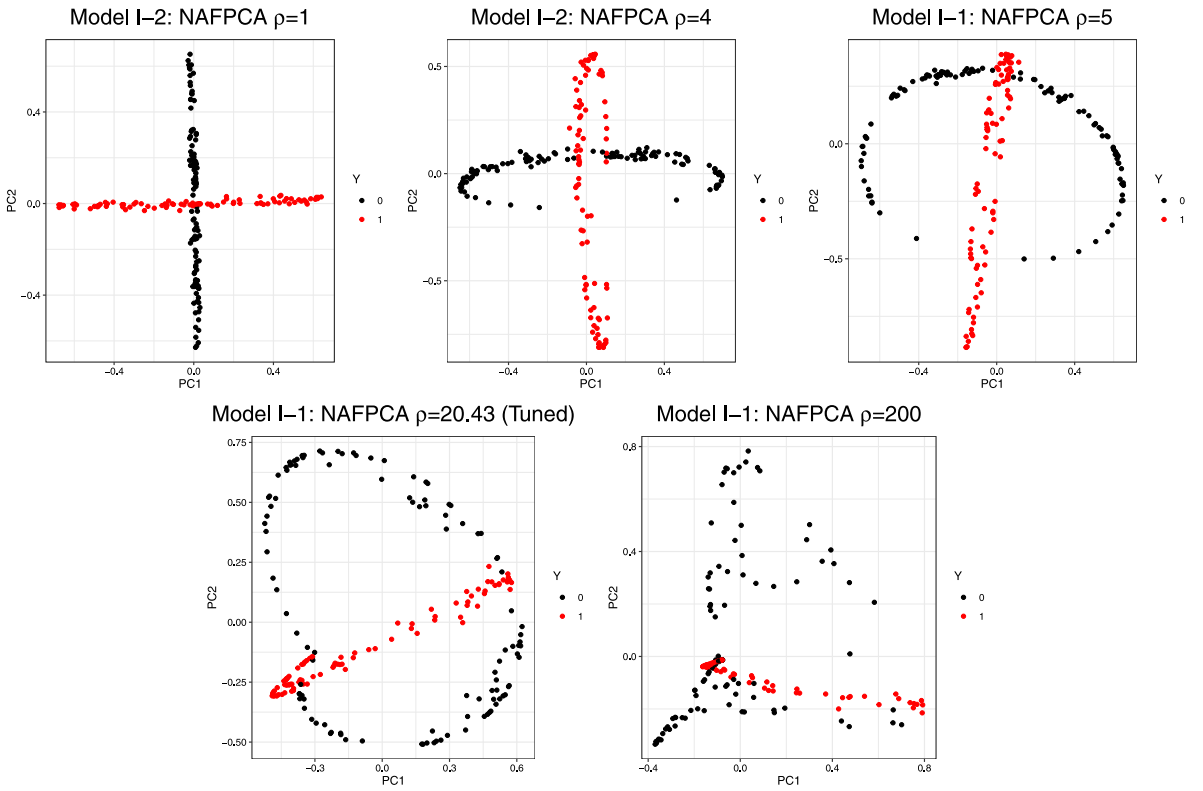


**Fig. 6.** (Model I-2) Effects of tuning parameter $\gamma$ used in the kernel for the RKHS $\mathfrak{M}$. The plots show first two PCs from NAFPCA with different $\rho$, where $\rho$ is proportional to $\gamma$. Top: $\rho = 1, 4, 5$, bottom: $\rho = 20.43, 200$, where 20.43 is the tuned value from the selection procedure.

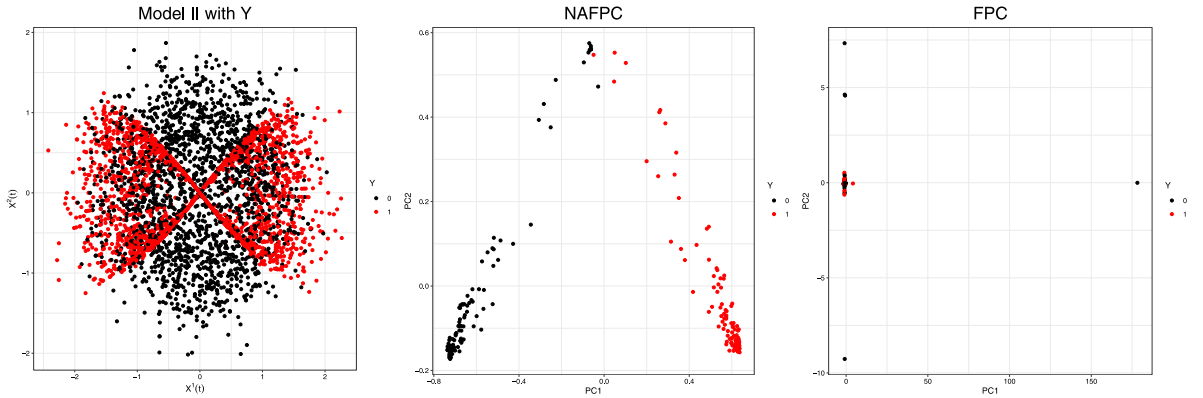


**Fig. 7.** (Model II) Left: observed data without lines of bivariate functional data from two clusters; Middle: first two PCs of NAFPCA; Right: first two PCs of FPCA. NAFPCA separates the two clusters near perfectly while the linear FPCA does not work well. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To compare the two methods in more challenging setting, we next consider the following model:

$$\text{Model III-1}: \begin{cases} X_a(t)|Y_a = 1 & \sim \{Z_{1a}(b_1(t)+b_2(t)) + Z_{2a}b_2(t) + \epsilon_{1a}(t)\}\{(Z_{3a}+Z_{4a})(b_5(t)+b_6(t)) + \epsilon_{2a}(t)\}, \\ X_a(t)|Y_a = 2 & \sim Z_{5a}\cos(\theta_{1a})\cos(\pi t) + Z_{5a}\sin(\theta_{1a})\sin(t) + \epsilon_{3a}(t), \\ X_a(t)|Y_a = 3 & \sim Z_{6a}\cos(\theta_{2a})\cos(\pi t) + Z_{6a}\sin(\theta_{2a})\sin(t) + \epsilon_{4a}(t), \end{cases}$$

where $Y_a \sim Unif(\{1, 2, 3\})$, $Z_{1a}, Z_{2a}, Z_{3a}, Z_{4a} \sim N(0, 2^2)$, $Z_{5a} \sim N(1, 0.2^2)$, $\theta_{1a} \sim U(0, 2\pi)$, $Z_{6a} \sim N(4, 0.5^2)$, $\theta_{2a} \sim U(0, 2\pi)$, $\epsilon_{1a}(t_j)$, $\epsilon_{2a}(t_j)$, $\epsilon_{3a}(t_j)$, $\epsilon_{4a}(t_j) \sim N(0, 0.1^2)$, and $b_1(t), \ldots, b_6(t)$ are the 6 B-spline basis functions defined on $[0, 1]$. All these random variables are independent. The Model III-1 consists of both of linear and nonlinear relations.

**Table 2**
Percentages of correct classifications for Model III-1 where $X_a$'s are combined with linear and nonlinear relations. Entries are the means and standard deviations (in parentheses) of the percentages calculated from 100 samples. NAFPCA works well with all the classifiers, whereas FPCA works well for SVM, which indicates that our method effectively captures the nonlinear characteristic beyond linear and quadratic features.

| Sample size | # Time points | NAFPCA | | | FPCA | | |
|---|---|---|---|---|---|---|---|
| | | LDA | QDA | SVM | LDA | QDA | SVM |
| 200 | 20 | 0.879 (0.036) | 0.885 (0.044) | 0.914 (0.033) | 0.411 (0.067) | 0.879 (0.04) | 0.922 (0.028) |
| | 50 | 0.878 (0.034) | 0.899 (0.037) | 0.916 (0.031) | 0.412 (0.086) | 0.881 (0.043) | 0.926 (0.026) |
| | 100 | 0.868 (0.037) | 0.896 (0.033) | 0.91 (0.037) | 0.412 (0.074) | 0.884 (0.044) | 0.923 (0.026) |
| 400 | 20 | 0.893 (0.024) | 0.913 (0.025) | 0.922 (0.021) | 0.405 (0.069) | 0.899 (0.03) | 0.929 (0.021) |
| | 50 | 0.89 (0.024) | 0.923 (0.02) | 0.929 (0.019) | 0.433 (0.073) | 0.913 (0.026) | 0.938 (0.018) |
| | 100 | 0.887 (0.024) | 0.927 (0.023) | 0.928 (0.021) | 0.41 (0.057) | 0.916 (0.025) | 0.941 (0.016) |
| 800 | 20 | 0.899 (0.016) | 0.927 (0.015) | 0.934 (0.014) | 0.396 (0.063) | 0.917 (0.016) | 0.937 (0.013) |
| | 50 | 0.893 (0.016) | 0.934 (0.013) | 0.938 (0.014) | 0.41 (0.063) | 0.925 (0.018) | 0.945 (0.011) |
| | 100 | 0.89 (0.018) | 0.936 (0.012) | 0.938 (0.014) | 0.409 (0.063) | 0.928 (0.015) | 0.946 (0.012) |

**Table 3**
Percentages of correct classifications for Model III-2 where the labels are related with $X$ in a non-additive way. Entries are the means and standard deviations (in parentheses) of the percentages calculated from 100 samples. Overall, both NAFPCA and FPCA perform reasonably well. When combined with LDA, their performances seem to be worse compared with the additive models.

| Sample size | # Time points | NAFPCA | | | FPCA | | |
|---|---|---|---|---|---|---|---|
| | | LDA | QDA | SVM | LDA | QDA | SVM |
| 200 | 20 | 0.521 (0.058) | 0.797 (0.048) | 0.783 (0.056) | 0.506 (0.05) | 0.798 (0.048) | 0.796 (0.045) |
| | 50 | 0.515 (0.063) | 0.81 (0.047) | 0.789 (0.052) | 0.497 (0.057) | 0.799 (0.051) | 0.794 (0.056) |
| | 100 | 0.521 (0.062) | 0.802 (0.049) | 0.785 (0.044) | 0.502 (0.06) | 0.794 (0.048) | 0.789 (0.058) |
| 400 | 20 | 0.514 (0.048) | 0.844 (0.032) | 0.825 (0.032) | 0.504 (0.038) | 0.833 (0.036) | 0.823 (0.035) |
| | 50 | 0.515 (0.05) | 0.835 (0.037) | 0.817 (0.032) | 0.508 (0.045) | 0.831 (0.036) | 0.818 (0.032) |
| | 100 | 0.514 (0.048) | 0.839 (0.031) | 0.822 (0.034) | 0.51 (0.048) | 0.834 (0.032) | 0.827 (0.033) |
| 800 | 20 | 0.511 (0.034) | 0.868 (0.039) | 0.846 (0.03) | 0.505 (0.039) | 0.859 (0.022) | 0.852 (0.025) |
| | 50 | 0.515 (0.037) | 0.875 (0.028) | 0.853 (0.024) | 0.506 (0.038) | 0.856 (0.024) | 0.853 (0.022) |
| | 100 | 0.516 (0.04) | 0.872 (0.031) | 0.851 (0.029) | 0.511 (0.037) | 0.859 (0.024) | 0.85 (0.025) |

Table 2 shows that NAFPCA works well with all the classifiers, whereas FPCA works well for SVM, which indicates that our method effectively captures the nonlinear characteristic beyond linear and quadratic features.

Lastly, we consider the situations where the class labels are related with the functional predictors in a non-additive way. To see how the methods work in this case, instead of generating an inversed model, we generate the model in a forward way as follows:

$$\text{Model III-2} : \begin{cases} X_a^1(t) & \sim \sum_{m=1}^{6} Z_{1am} b_m(t) + Z_{2a}(t^3/2 + t^2), \\ X_a^2(t) & \sim \sum_{m=1}^{6} Z_{3am} b_m(t) + Z_{4a}(2\sqrt{t} + (1-t)^2), \\ Y_a & \sim \langle X_a^1(t) X_a^2(t) + \epsilon_a(t), b_3(t) + b_4(t) \rangle_{L_2}, \end{cases}$$

where $Z_{1am}, Z_{3am} \sim N(0, 2^2)$, $Z_{2a}, Z_{4a} \sim N(0, 1)$, $\epsilon_a(t_j) \sim N(0, 0.1^2)$, and $b_1(t), \ldots, b_6(t)$ are the 6 B-spline basis functions defined on $[0, 1]$. All the random variables are independent. We then define $Y_a^*$ as the indicator function that takes the value 1 if $Y_a$ is higher than the median of $Y_1, \ldots, Y_n$. We apply the dimension reduction methods to $X = (X^1, X^2)$ and classification methods to the reduced predictors and $Y^*$.

Table 3 summarizes the percentages of correct classifications. The results show that both NAFCA and FPCA still perform reasonably well, and their overall performances are similar. Furthermore, when combined with LDA, their performances seem to be worse compared with the additive models.

## 6. Conclusion

In this paper, we introduce a nonlinear and additive functional principal component analysis for multivariate functional data, which is capable of capturing nonlinear variations in functional data. We developed the population-level as well as asymptotic properties of this method. We showed by simulations and real-data analysis that our method effectively captures nonlinear feature that is missed by the linear functional PCA method, and achieves better classification when combined with a majority of the classifiers we used.

Along with NAFPCA we also proposed a two-step tuning method to determine a tuning parameter in the kernel and the number of significant principal components, and established the consistency of the estimation of the number of principal components. In addition, we proposed a novel way to represent functional principal components using the intensity of three prime colors, which make functional PCA directly interpretable from the spaghetti plots. Finally, this work has also raised several new problems. For example, it is plausible that the order determination method we developed for NAFPCA here can be extended nonlinear sufficient dimension reduction, and our analysis of the handwritten digit data raises the importance of normalization. We hope to tackle these problems in a future research.

## CRediT authorship contribution statement

**Jun Song:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization, Supervision, Project administration, Funding acquisition. **Bing Li:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

## Acknowledgments

We thank the Editor, an Associate Editor, and two referees for their insightful and constructive comments and suggestions, which helped us greatly in revising this work. Jun Song's research is supported, in part, by funds provided by the University of North Carolina at Charlotte. Bing Li's research is supported in part by the National Science Foundation Grant DMS-1713078.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2020.104675. The supplementary file shows the real data analysis of NAFPCA.

## References

[1] J. Ahn, A stable hyperparameter selection for the Gaussian RBF kernel for discrimination, Stat. Anal. Data Min.: ASA Data Sci. J. 3 (3) (2010) 142–148.
[2] A. Berlinet, C. Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics, Springer-Verlag, New York, 2004.
[3] C. Carmeli, E. De Vito, A. Toigo, V. Umanitá, Vector valued reproducing kernel Hilbert spaces and universality, Anal. Appl. 8 (01) (2010) 19–61.
[4] A. Christmann, I. Steinwart, Support Vector Machines, Springer-Verlag, New York, 2008.
[5] J.B. Conway, A Course in Functional Analysis, second ed., Springer-Verlag, New York, 1990.
[6] F. Dinuzzo, B. Schölkopf, The representer theorem for Hilbert spaces: a necessary and sufficient condition, Adv. Neural Inf. Process. Syst. (2012) 189–196.
[7] K. Fukumizu, F.R. Bach, M.I. Jordan, Kernel dimension reduction in regression, Ann. Statist. 37 (4) (2009) 1871–1905.
[8] G.H. Golub, M. Heath, G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics 21 (2) (1979) 215–223.
[9] P. Hall, H. Müller, J. Wang, Properties of principal component methods for functional and longitudinal data analysis, Ann. Statist. 34 (3) (2006) 1493–1517.
[10] C. Happ, S. Greven, Multivariate functional principal component analysis for data observed on different (dimensional) domains, J. Amer. Statist. Assoc. 113 (522) (2018) 649–659.
[11] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, 1985.
[12] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Springer-Verlag, New York, 2012.
[13] T. Hsing, R. Eubank, Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators, John Wiley & Sons, West Sussex, UK, 2015.
[14] J. Jacques, C. Preda, Model-based clustering for multivariate functional data, Comput. Statist. Data Anal. 71 (2014) 92–106.
[15] I. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 2011.
[16] P. Kokoszka, M. Reimherr, Introduction to Functional Data Analysis, CRC Press, Boca Raton, FL, 2017.
[17] K.-Y. Lee, B. Li, F. Chiaromonte, A general theory for nonlinear sufficient dimension reduction: Formulation and estimation, Ann. Statist. 41 (1) (2013) 221–249.
[18] B. Li, A. Artemiou, L. Li, Principal support vector machines for linear and nonlinear sufficient dimension reduction, Ann. Statist. 39 (6) (2011) 3182–3210.

[19] B. Li, H. Chun, H. Zhao, Sparse estimation of conditional graphical models with application to gene networks, J. Amer. Statist. Assoc. 107 (497) (2012) 152–167.
[20] B. Li, H. Chun, H. Zhao, On an additive semigraphoid model for statistical networks with application to pathway analysis, J. Amer. Statist. Assoc. 109 (507) (2014) 1188–1204.
[21] Y. Li, Y. Guan, Functional principal component analysis of spatiotemporal point process with applications in disease surveillance, J. Amer. Statist. Assoc. 109 (507) (2014) 1205–1215.
[22] B. Li, J. Song, Nonlinear sufficient dimension reduction for functional data, Ann. Statist. 45 (3) (2017) 1059–1095.
[23] H.Q. Minh, Some properties of gaussian reproducing kernel hilbert space and their implications for function approximation and learning theory, Constr. Approx. 32 (2010) 307–338.
[24] J. Ramsay, X. Li, Curve registration, J. R. Stat. Soc. Ser. B 60 (1998) 351–363.
[25] J. Ramsay, B. Silverman, Functional Data Analysis, second ed., Springer-Verlag, New York, 2005, p. 430.
[26] C.E. Rasmussen, C.K. Williams, Gaussian Processes for Machine Learning, MIT Press, Cambridge, MA, 2006, pp. 79–104.
[27] B. Schölkopf, R. Herbrich, A.J. Smola, A generalized representer theorem, Comput. Learn. Theory (2001) 416–426.
[28] B. Schölkopf, A. Smola, K. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.
[29] B.K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, G.R. Lanckriet, Hilbert space embeddings and metrics on probability measures, J. Mach. Learn. Res. 11 (Apr) (2010) 1517–1561.
[30] J.-L. Wang, J.-M. Chiou, H.-G. Müller, Functional data analysis, Annu. Rev. Stat. Appl. 3 (2016) 257–295.