Statistics and Computing (2025) 35:95
https://doi.org/10.1007/s11222-025-10633-7

ORIGINAL PAPER

Nonlinear Functional Sufficient Dimension Reduction via Principal Fitted Components

Minjee Kim[1] · Yujin Park[1] · Kyongwon Kim[2] · Jae Keun Yoo[1]

Received: 12 February 2025 / Accepted: 1 May 2025

Abstract
In this paper, we propose a novel functional nonlinear sufficient dimension reduction method based on the principal fitted component model. Our approach extends the concept of principal fitted components to functional data, covering the case where both the predictors and responses are functions. We consider a general framework in which the predictor and response can each be viewed as elements of potentially infinite dimensional Hilbert spaces. This includes the important scalar on function and function on function cases as special instances. We generalize a nonlinear principal fitted component model within the framework of reproducing kernel Hilbert space, leveraging the nested Hilbert spaces theory to characterize nonlinear structures in functional data. The first space accommodates functions of random curves and the second space captures their nonlinear relationships. To establish the theoretical validity of our approach, we develop an asymptotic theory that characterizes the convergence behavior of the proposed estimator under mild regularity conditions. Extensive simulation studies demonstrate that our method outperforms existing functional sufficient dimension reduction methods, particularly in scenarios with complex nonlinear dependencies. The effectiveness of the proposed method is further validated through real data analysis.

Keywords Principal fitted component model · Sufficient dimension reduction · Functional data analysis · Reproducing kernel Hilbert space · Kernel methods

1 Introduction

In modern statistical analysis, data that are collected in the form of functions or curves over a continuum, commonly referred to as functional data, have gained significant importance across a wide range of disciplines. Unlike traditional multivariate data, functional data are inherently infinite dimensional, making their analysis fundamentally different and more complex. Such data arise naturally in various scientific and applied fields, including growth curves in biology, where researchers track the developmental patterns of organisms over time, temperature profiles in climatology that capture variations in atmospheric conditions across different time scales, and spectral measurements in chemistry, which record the absorption or emission of light at different wavelengths. The infinite dimensional nature of functional data presents unique analytical challenges, as conventional statistical methods designed for finite dimensional data are often inadequate. Consequently, specialized statistical methods have been developed to extract meaningful patterns, reduce dimensionality, and model dependencies in functional data, enabling researchers to gain deeper insights and make more accurate predictions in their respective fields.

To address the complexity inherent in high-dimensional or infinite dimensional data, sufficient dimension reduction (SDR) methods have been extended to the functional data setting. SDR seeks to reduce the dimensionality of predictor variables while retaining essential information relevant

Minjee Kim and Yujin Park contributed equally to this work

✉ Kyongwon Kim
kimk@yonsei.ac.kr

✉ Jae Keun Yoo
peter.yoo@ewha.ac.kr

Minjee Kim
jee365@ewhain.net

Yujin Park
yujin.alex.p@gmail.com

[1] Department of Statistics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Republic of Korea

[2] Department of Applied Statistics, Department of Statistics and Data Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea

Published online: 15 May 2025

to the response variable. Classical SDR, including sliced inverse regression (Li 1991), sliced average variance estimation (Cook and Weisberg 1991), contour regression (Li et al. 2005), and directional regression (Li and Wang 2007), have been widely applied for this purpose. In linear SDR, the central subspace is defined as the smallest subspace spanned by vectors $\eta = (\eta_1, \cdots, \eta_d)$ such that $Y \perp\!\!\!\perp X \mid \eta^{\mathsf{T}} X$. However, these methods typically rely on linearity assumptions, which may not always hold in practice.

Extending this framework to nonlinear settings requires generalizing the conditional independence assumption to $Y \perp\!\!\!\perp X \mid f(X)$, where $f(X) = (f_1(X), \cdots, f_d(X))$ consists of nonlinear functions of X. However, this generalization introduces significant challenges, as $f(X)$ is not uniquely identifiable, making it difficult to rigorously define the central subspace. To address this issue, Lee et al. (2013) introduced the concept of the central σ-field, which captures the minimal sufficient information needed for dimension reduction in nonlinear settings.

Among various SDR methods, the principal fitted component (PFC) model, introduced by Cook (2007), stands out as an effective model-based approach for dimension reduction from the perspective of inverse regression. The PFC model assumes that the conditional distribution of the predictor $X \in \mathbb{R}^p$ given the response Y follows a specific structure, facilitating dimension reduction while preserving essential information. Let $\mathbf{X}_y$ denote a random vector distributed as $X \mid (Y = y)$, and assume that $\mathbf{X}_y$ follows a normal distribution with mean μ. The PFC model is formulated as

$$\mathbf{X}_y = \mu + \Gamma\beta \mathrm{g}_y + \sigma\epsilon,$$

where $\mu \in \mathbb{R}^p$ represents the mean vector, $\Gamma \in \mathbb{R}^{p\times d}$ is an orthonormal basis satisfying $\Gamma^{\mathsf{T}}\Gamma = I_d$, and $\beta \in \mathbb{R}^{d\times r}$ is a coefficient matrix with $d \le r$. The term $\mathrm{g}_y \in \mathbb{R}^r$ is a known vector-valued function of y satisfying $\Sigma_y \mathrm{g}_y = \mathbf{0}_{r\times 1}$. The error vector $\epsilon \in \mathbb{R}^p$ is assumed to be independent of Y with mean zero and covariance matrix I_p. The likelihood function of the PFC model can be written as

$$\log L(\mu;\Gamma;\beta;\sigma) = -\frac{pn}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_y \parallel \mathbf{X}_y - \mu - \Gamma\beta \mathrm{g}_y \parallel^2 . \tag{1}$$

To estimate the central subspace, the partial maximum likelihood approach is applied to (1) with respect to Γ and β. This optimization problem reduces to finding the first d eigenvectors of the covariance matrix of $\mathbf{X}_y$, which span the dimension reduced subspace. For further theoretical details and implementation, see Cook and Forzani (2008).

As an extension of the PFC model, Song et al. (2023) introduced the kernel principal fitted component (KPFC) model, which captures the nonlinear relationship between $\mathbf{X}_y$ and Y. This approach formulates linear functions of $\mathbf{X}_y \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ in a reproducing kernel Hilbert space (RKHS), leveraging the kernel trick. The kernel trick enables nonlinear data to be mapped into a higher dimensional feature space, where the data exhibit a linear structure, making it possible to apply linear methodologies in this transformed space. Let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be the RKHS spaces associated with $\mathbf{X}_y$ and Y, respectively, and define the corresponding feature mappings as $\Phi(\mathbf{X}_y) \in \mathcal{H}_X$ and $\Psi(y) \in \mathcal{H}_Y$. The objective function of the KPFC model is given by

$$\arg\min_{\Gamma_k,\beta_k} \sum_y \parallel \Phi(\mathbf{X}_y) - E\Phi(\mathbf{X}_y) - \sum_{k=1}^{d} (\Gamma_k \otimes \beta_k)\Psi(y) \parallel^2_{\mathcal{H}},$$

subject to the orthonormality constraint $\langle \Gamma_k, \Gamma_j\rangle_{\mathcal{H}} = \delta_{kj}$, where $\Gamma_k, \beta_k \in \mathcal{H}$. Here, $\delta_{kj} = 1$ when $k = j$ and $\delta_{kj} = 0$ otherwise. This indicates the orthogonality of components. The operator $\otimes$ represents the tensor product, which is computed as $(f \otimes g)h = f\langle g, h\rangle_{\mathcal{H}}$ for all $h \in \mathcal{H}$.

Here, we extend the PFC model to the functional setting, where both $\mathbf{X}_y$ and Y are random curves, denoted as $\mathbf{X}_y(t)$ and $y(t)$ for $t \in [a, b]$. This extension generalizes the Euclidean space $\mathbb{R}^p$ to an infinite dimensional Hilbert space and replaces the standard Euclidean inner product with a functional inner product defined as

$$\langle f, g\rangle = \int_a^b f(t)g(t)dt, \quad \forall f, g \in L_2[a, b].$$

Let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be the Hilbert spaces associated with the functional predictors $\mathbf{X}_y$ and the functional response y. The functional principal fitted component (FPFC) model extends the PFC framework to the functional domain, and its objective function is formulated as

$$\arg\min_{\Gamma_k,\beta_k} \sum_y \parallel \mathbf{X}_y - E(\mathbf{X}_y) - \sum_{k=1}^{d} (\Gamma_k \otimes \beta_k)y \parallel^2_{\mathcal{H}}, \tag{2}$$

subject to the constraint $\langle \Gamma_k, \Gamma_j\rangle_{\mathcal{H}} = \delta_{kj}$, where $\Gamma_k, \beta_k \in \mathcal{H}$. This formulation ensures the orthogonality of functional components, enabling effective dimension reduction in infinite dimensional spaces.

In this paper, we further extend the FPFC model to accommodate nonlinear transformations of functional data. This generalization is achieved by introducing a nested Hilbert space framework, incorporating a second level Hilbert space $\mathfrak{M}_X$ and $\mathfrak{M}_Y$ to capture nonlinear structures more effectively. Specifically, the functional nature of our problem necessitates

the construction of two Hilbert spaces. The first represents the space in which the functional predictor X resides, while the second, assumed to be an RKHS, characterizes the underlying nonlinear relationships. By leveraging this nested Hilbert space structure, our approach enables a more flexible modeling of nonlinear dependencies among random functions. This extension provides a systematic way to generalize the PFC model to nonlinear functional settings while preserving dimension reduction properties. In the finite dimensional setting, the PFC model of Cook (2007) naturally arises from the likelihood function under a multivariate normal assumption, as shown in (1). However, extending this likelihood based approach directly to infinite dimensional functional spaces is challenging because defining probability density functions for functional data is fundamentally problematic (Dai et al. 2017; Delaigle and Hall 2010). Therefore, the likelihood formulation used in the multivariate context does not straightforwardly generalize to the functional setting. To circumvent this issue, we propose an alternative objective function in (3), which can capture the core idea of minimizing squared deviations from a lower dimensional mean structure within an appropriate Hilbert space framework. This approach does not require the existence of an infinite dimensional density function and thus preserves the statistical essence of the original PFC model. Figure 1 illustrates the conceptual progression from the standard PFC model to the proposed nonlinear functional principal fitted component (NFPFC) model. As we can see from Figure 1, the PFC model extends to the KPFC to capture nonlinearity and to the FPFC to handle functional data. The NFPFC model unifies these extensions, effectively modeling nonlinear functions of functional data.

The remainder of the paper is structured as follows. In Section 2, we establish the theoretical foundation of functional SDR and introduce the NFPFC model at the population level. Section 3 presents the estimation algorithm for the sample level implementation. In Section 4, we provide the asymptotic theory of the proposed estimator. Section 5 evaluates the performance of our method through simulation studies, followed by a real world data application in Section 6. Finally, Section 7 summarizes our work and potential directions for future research. The Appendix includes detailed proofs, additional simulation studies, and an illustration of the parameter associated with the convergence rate.

2 Population-level development

2.1 Nonlinear functional sufficient dimension reduction

Ferré and Yao (2003) and Ferré and Yao (2005) introduced a framework for functional linear SDR, which is formulated as follows

$$Y \perp\!\!\!\perp X \mid \langle f_1, X\rangle_{\mathcal{H}}, \cdots, \langle f_d, X\rangle_{\mathcal{H}},$$

where $f_1, \cdots, f_d$ are elements of the Hilbert space $\mathcal{H} = L_2[a, b]$, and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product in $\mathcal{H}$. The objective of functional linear SDR is to identify the subspace of $\mathcal{H}$ spanned by $f_1, \cdots, f_d$, which captures the essential information of X relevant to Y while reducing dimensionality.

Let Ω be a probability space equipped with a σ-algebra $\mathcal{F}$ and a probability measure P. Suppose $\mathcal{T}_X$ and $\mathcal{T}_Y$ are intervals in $\mathbb{R}$, and let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be Hilbert spaces of functions defined on $\mathcal{T}_X$ and $\mathcal{T}_Y$, respectively. We define the Borel σ-algebras $\mathcal{F}_X$ and $\mathcal{F}_Y$ as those generated by the open sets in $\mathcal{H}_X$ and $\mathcal{H}_Y$. For any general random element S, let $\sigma(S)$ denote the σ-field generated by S. As defined in Billingsley (2012), a 'random element' refers to any measurable function mapping from a probability space Ω to another measurable space $(\Omega_S, \mathcal{F}_S)$. This broad definition includes various cases, such as random variables in $\mathbb{R}$, random vectors in $\mathbb{R}^k$, random functions in a Hilbert space, and even collections of random functions forming a direct sum of Hilbert spaces, which aligns with our setting. Here, a mapping $X : \mathbb{R} \to \mathcal{H}_X$ is called a random element in $\mathcal{H}_X$, and similarly, a mapping $Y : \mathbb{R} \to \mathcal{H}_Y$ is a random element in $\mathcal{H}_Y$. These random elements are measurable with respect to the σ-algebras $\mathcal{F}/\mathcal{F}_X$ and $\mathcal{F}/\mathcal{F}_Y$, securing well defined probability distributions. The distributions of X and Y are denoted by P_X and P_Y, respectively.

We assume there exists a sub-σ-field $\mathcal{G}$ within $\sigma(X)$ that satisfies

$$Y \perp\!\!\!\perp X \mid \mathcal{G}.$$

The process of identifying $\mathcal{G}$ is referred to as nonlinear functional SDR (Lee et al. 2013; Li and Song 2017). Following Lee et al. (2013), Li and Song (2017), we define $\mathcal{G}$ as a sufficient σ-field within $\sigma(X)$ for predicting Y. Under a mild condition, as shown in Lee et al. (2013), Li and Song (2017), the intersection of all such sub-σ-fields remains a sufficient σ-field. This intersection is termed the central σ-field for $Y \mid X$. Throughout this paper, we assume that this condition is satisfied. This guarantees the existence of the central σ-field. For simplicity, we redefine $\mathcal{G}$ to represent this central σ-field for $Y \mid X$. Intuitively, this represents the minimal amount of information in X that retains all predictive power for Y. The central σ-field generalizes the concept of a central subspace in classical SDR to settings where both X and Y are functional.

Tan et al. (2024) propose a nonlinear dimension reduction framework for functional data that extends manifold learning to accommodate intricate data structures such as signifi-

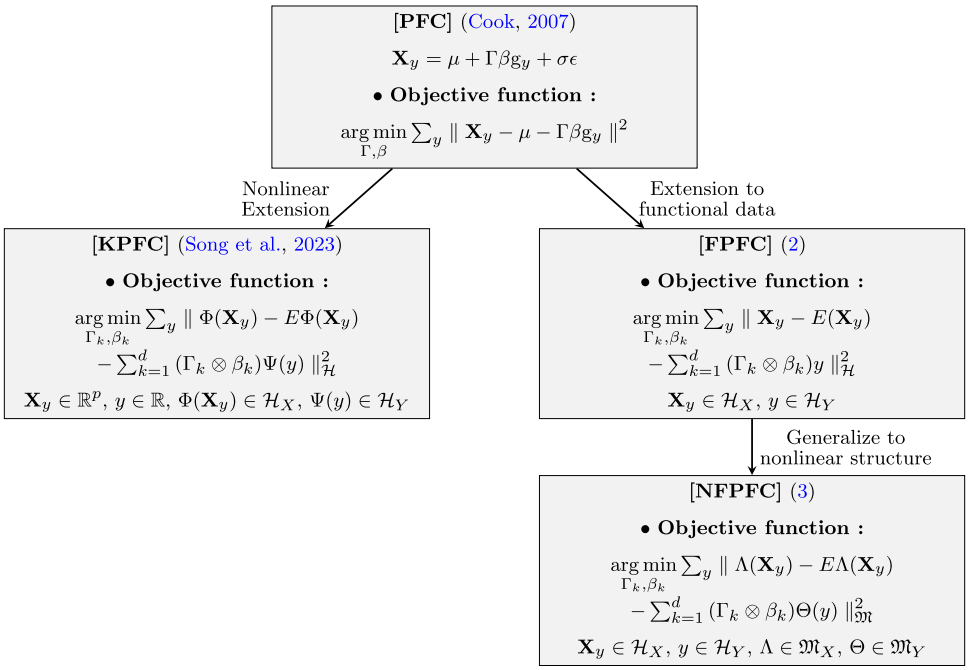

Fig. 1 Flowchart illustrating the progression toward the NFPFC model

cant phase variation. Their methodology includes theoretical guarantees and practical strategies for handling measurement errors in functional data. By focusing on the intrinsic geometry of functional data assumed to lie on an unknown manifold, Tan et al. (2024) demonstrate how their manifold based approach can outperform traditional methods, particularly for clustering tasks where data exhibit nonlinear structures.

To effectively capture the nonlinear relationships between X and Y, Li and Song (2017) introduced a second-level Hilbert space of functions defined on $\mathcal{H}_X$ and $\mathcal{H}_Y$. This framework leverages RKHS to model complex dependencies beyond linear structures.

Let $\kappa : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ be a positive definite kernel, and let $\mathfrak{M}_X$ and $\mathfrak{M}_Y$ denote the corresponding RKHS induced by κ. The kernel function is assumed to take the form as

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

for any $f, g \in \mathcal{H}$, where $\rho : \mathbb{R}^3 \to \mathbb{R}^+$ is a predefined function. This formulation extends the standard kernel trick by replacing the Euclidean inner product with the inner product in the first-level Hilbert space $\mathcal{H}$, thereby adapting the kernel to the functional setting.

Common choices for ρ yield well-known nested kernels, such as

$$\kappa(f, g) = \exp(\langle f, g\rangle_{\mathcal{H}}),$$
$$\kappa(f, g) = (\langle f, g\rangle_{\mathcal{H}} + c)^m, \quad c \geq 0,$$
$$\kappa(f, g) = \exp(-\gamma \|f - g\|^2_{\mathcal{H}}), \quad \gamma = \frac{1}{2\sigma^2},$$

where the second expression represents the polynomial kernel with a constant shift c, and the third expression corresponds to the Gaussian radial basis function (RBF) kernel with bandwidth parameter σ.

Since the inner product in the second-level RKHS $\mathfrak{M}$ is uniquely determined by κ, and κ itself is fully characterized by the inner product in $\mathcal{H}$, it follows that the inner product in $\mathfrak{M}$ is inherently dependent on the structure of $\mathcal{H}$. Consequently, the second-level Hilbert space $\mathfrak{M}$ can be viewed as a nested RKHS induced by the function ρ, providing a systematic framework for capturing nonlinear structures in functional data.

2.2 Nonlinear functional principal fitted component model

In this section, we extend the FPFC model to the nonlinear setting. To formalize this extension, let $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ be the space of bounded linear operators mapping from a Hilbert space $\mathcal{H}_1$ to another Hilbert space $\mathcal{H}_2$. For any linear operator $J \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$, we define J^* as its adjoint operator, $\ker(J)$ as its kernel (null space), $\mathrm{ran}(J)$ as its range (image), and $\overline{\mathrm{ran}}(J)$ as the closure of its range. Further, $(\cdot)^\dagger$ represents the Moore-Penrose inverse. These operator-theoretic concepts provide a fundamental framework for analyzing and formulating the nonlinear extension of the FPFC model.

Let $L_2(P_X)$ denote the space of all square-integrable functions of X, defined as the set of functions satisfying $E[f^2(X)] < \infty$. The following assumptions are fundamental in establishing the theoretical framework of our proposed method.

Assumption 1 $\mathfrak{M}_X$ is a dense subset of $L_2(P_X)$ modulo constants if, for any $f \in L_2(P_X)$, there is a sequence $f_n \subseteq \mathfrak{M}_X$ such that $\text{var}[f_n(X) - f(X)] \to 0$ as $n \to \infty$.

Under Assumption 1, Li (2018) demonstrated that verifying the independence of two functions from $\mathcal{H}_X$ and $\mathcal{H}_Y$ is sufficient, without requiring a Gaussian assumption, given the covariance operator Σ_{XY}. Since the RKHS induced by the Gaussian RBF kernel is dense in $L_2(P)$, we adopt the Gaussian RBF kernel in this paper to ensure a rich function space for modeling dependencies.

The central class, denoted by $\mathfrak{S}_{Y|X}$, consists of all functions in the second-level RKHS $\mathfrak{M}_X$ that are measurable with respect to $\mathcal{G}$. In other words, the central class captures the set of nonlinear functions of X that fully describe the dependence between X and Y, serving as a concrete function space representation of the central σ-field.

A key result from Li and Song (2017) states that under mild assumptions, the central class can be characterized using covariance operators in the RKHS framework. It is related to the range space of a regression operator constructed from these covariance operators. If the central class is complete, meaning it contains all necessary nonlinear transformations of X, then the estimated central class is exhaustive, which guarantees that no relevant information is lost in the reduction process.

Thus, the central class provides a functional counterpart to the central sub-σ-field by offering a practical approach to estimating and recovering the information-preserving subspace for nonlinear functional SDR. Therefore, our objective is to estimate the central class $\mathfrak{S}_{Y|X}$ based on a random sample of (X, Y).

Assumption 2 There exist constants $C_1 > 0$ and $C_2 > 0$ such that, for all $f \in \mathfrak{M}_X$ and $g \in \mathfrak{M}_Y$, $\text{var}[f(X)] \le C_1 \|f\|^2_{\mathfrak{M}_X}$, $\text{var}[g(Y)] \le C_2 \|g\|^2_{\mathfrak{M}_Y}$.

Given Assumption 2, the mapping $\mathfrak{M}_X \to L_2(P_X)$, defined by $f \mapsto f$, is a bounded linear operator. Furthermore, the bilinear form $\mathfrak{M}_X \times \mathfrak{M}_X \to \mathbb{R}$, given by $(f, g) \mapsto \text{cov}(f(X), g(X))$, is also bounded. Consequently, we define the variance and covariance operators as

$$\Sigma_{XX} \in \mathcal{B}(\mathfrak{M}_X), \quad \Sigma_{YY} \in \mathcal{B}(\mathfrak{M}_Y),$$
$$\Sigma_{XY} \in \mathcal{B}(\mathfrak{M}_Y, \mathfrak{M}_X), \quad \Sigma_{YX} \in \mathcal{B}(\mathfrak{M}_X, \mathfrak{M}_Y),$$

where the inner product relation

$$\langle f, \Sigma_{XX} g\rangle_{\mathfrak{M}_X} = \text{cov}(f(X), g(X))$$

holds. By definition, the operators Σ_{XX} and Σ_{YY} are self-adjoint, and the adjoint property $\Sigma^*_{XY} = \Sigma_{YX}$ holds.

Assumption 2 further ensures the existence of positive constants $C_1 > 0$ and $C_2 > 0$ such that, for all $f \in \mathfrak{M}_X$ and $g \in \mathfrak{M}_Y$, we have

$$E|f(X)| \le C_1 \|f\|_{\mathfrak{M}_X}, \quad E|g(Y)| \le C_2 \|g\|_{\mathfrak{M}_Y}.$$

As a result, the linear functionals $f \mapsto E[f(X)] : \mathfrak{M}_X \to \mathbb{R}$ and $g \mapsto E[g(Y)] : \mathfrak{M}_Y \to \mathbb{R}$ are bounded. Let μ_X and μ_Y be the Riesz representations of these bounded linear functionals.

Moreover, by Assumption 2, we obtain

$$\overline{\text{ran}}(\Sigma_{XX}) = \mathfrak{M}^0_X, \quad \overline{\text{ran}}(\Sigma_{YY}) = \mathfrak{M}^0_Y.$$

Thus, the spaces $\mathfrak{M}^0_X$ and $\mathfrak{M}^0_Y$ can be expressed as

$$\mathfrak{M}^0_X = \overline{\text{span}}\{\kappa_X(\cdot, x) - \mu_X : x \in \mathcal{H}_X\},$$
$$\mathfrak{M}^0_Y = \overline{\text{span}}\{\kappa_Y(\cdot, y) - \mu_Y : y \in \mathcal{H}_Y\},$$

where $\overline{\text{span}}$ denotes the closure of the linear span.

Assumption 3

(i) $\text{ran}(\Sigma_{XY}) \subseteq \text{ran}(\Sigma_{XX})$ and $\Sigma^\dagger_{XX}\Sigma_{XY}$ is a bounded operator.

(ii) $\text{ran}(\Sigma_{YX}) \subseteq \text{ran}(\Sigma_{YY})$ and $\Sigma^\dagger_{YY}\Sigma_{YX}$ is a bounded operator.

Assumption 3 imposes collective smoothness, which secures that the operators $\Sigma^\dagger_{XX}\Sigma_{XY}$ and $\Sigma^\dagger_{YY}\Sigma_{YX}$ are well defined.

Based on these assumptions, the objective function of the FPFC model in (2) can be reformulated to accommodate nonlinear structures. The objective function of the NFPFC model is given by

$$\arg\min_{\Gamma_k, \beta_k} \sum_y \Big\| \Lambda(\mathbf{X}_y) - E\Lambda(\mathbf{X}_y) - \sum_{k=1}^{d} (\Gamma_k \otimes \beta_k)\Theta(y) \Big\|^2_{\mathfrak{M}}, \tag{3}$$

subject to the orthogonality constraint $\langle \Gamma_k, \Gamma_j\rangle_{\mathfrak{M}} = \delta_{kj}$, where $\Gamma_k, \beta_k \in \mathfrak{M}$, and $\delta_{kj} = 1$ if $k = j$ and $\delta_{kj} = 0$ otherwise. Here, $\mathbf{X}_y \in \mathcal{H}_X$ and $y \in \mathcal{H}_Y$ represent the functional predictor and response, respectively, while $\Lambda \in \mathfrak{M}_X$ and $\Theta \in \mathfrak{M}_Y$ are transformation operators that map functional data into an RKHS. By solving this optimization problem, we estimate the central class $\mathfrak{S}_{Y|X}$ using the first d eigenfunctions of the objective operator.

Up to this point, we have used $\mathbf{X}_y$ to explicitly denote a random vector drawn from the conditional distribution of the predictor X given $Y = y$, emphasizing how the distribution of X is generated from y. However, for notational simplicity, we will use $X \in \mathcal{H}_X$ in place of $\mathbf{X}_y$. This convention does not affect the underlying conditional structure. Unless

explicitly stated otherwise, all subsequent statements involving X implicitly assume the conditional dependence on Y established earlier.

Theorem 1 *Under Assumption 3, the NFPFC model is formulated as an eigenvalue problem, where the central class $\mathfrak{S}_{Y|X}$ is estimated by solving for the first d eigenfunctions of the operator*

$$\Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YX},$$

where d is a known parameter specifying the dimensionality of the reduced subspace.

Since Σ_{XX} and Σ_{YY} are Hilbert-Schmidt operators, their inverse operators, Σ_{XX}^{-1} and Σ_{YY}^{-1}, are unbounded (Fukumizu et al. 2007). To circumvent this issue, we employ the regularized operators $\Sigma_{XX}^{\dagger}\Sigma_{XY}$ and $\Sigma_{YY}^{\dagger}\Sigma_{YX}$ as surrogates for Σ_{XX}^{-1} and Σ_{YY}^{-1} in the estimation of the central class $\mathfrak{S}_{Y|X}$. Following the results of Li and Song (2017), under Assumptions 1, 2, and 3, the inclusion

$$\overline{\text{ran}}(\Sigma_{XY}\Sigma_{YY}^{\dagger}) \subseteq \text{cl}(\Sigma_{XX}\mathfrak{S}_{Y|X})$$

holds, where $\text{cl}(\cdot)$ denotes the closure. Furthermore, if $\mathfrak{S}_{Y|X}$ is complete, then equality is attained as

$$\overline{\text{ran}}(\Sigma_{XY}\Sigma_{YY}^{\dagger}) = \text{cl}(\Sigma_{XX}\mathfrak{S}_{Y|X}).$$

For any invertible operator U mapping $\mathfrak{M}_Y^0$ onto itself, we establish the equivalence

$$\overline{\text{ran}}(\Sigma_{XY}\Sigma_{YY}^{\dagger}) = \overline{\text{ran}}(\Sigma_{XY}\Sigma_{YY}^{\dagger}U\Sigma_{YY}^{\dagger}\Sigma_{YX}).$$

By setting $U = \Sigma_{YY}$, it follows that

$$\begin{aligned}\overline{\text{ran}}(\Sigma_{XY}\Sigma_{YY}^{\dagger}) &= \overline{\text{ran}}(\Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YY}\Sigma_{YY}^{\dagger}\Sigma_{YX}) \\ &= \overline{\text{ran}}(\Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YX}).\end{aligned}$$

Consequently, we conclude that

$$\overline{\text{ran}}(\Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YX}) = \text{cl}(\Sigma_{XX}\mathfrak{S}_{Y|X}).$$

Proposition 1 *Assume Assumptions 1, 2, and 3 hold. If $\mathcal{H}_X$ is dense in $L_2(P_X)$ and $\mathcal{H}_Y$ is dense in $L_2(P_Y)$, then we have*

$$\overline{\text{ran}}(\Sigma_{XX}^{\dagger}\Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YX}\Sigma_{XX}^{\dagger}) = \mathfrak{S}_{NFPFC} \subseteq \mathfrak{S}_{Y|X}.$$

Furthermore, if $\mathfrak{S}_{Y|X}$ is complete, then we have

$$\mathfrak{S}_{NFPFC} = \mathfrak{S}_{Y|X}.$$

The inclusion of $\Sigma_{XX}^{\dagger}$ ensures that the resulting operator remains well-defined and bounded, thereby maintaining stability in the estimation process. Moreover, this formulation aligns with Assumption 4 of Li and Song (2017), reinforcing the theoretical foundation of the proposed approach.

Corollary 1 *Suppose Assumptions 2 and 3 hold. Let $f_1, \ldots, f_d$ be solutions to the following generalized eigenfunction problem*

$$\begin{aligned}&\textit{maximize} && \langle f, \Sigma_{XX}^{\dagger}\Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YX}\Sigma_{XX}^{\dagger}f\rangle_{\mathfrak{M}_X}, \\ &\textit{subject to} && f \in \mathfrak{M}_X^0,\ \langle f, f\rangle_{\mathfrak{M}_X} = 1, \\ &&& \langle f, f_1\rangle_{\mathfrak{M}_X} = \cdots = \langle f, f_{k-1}\rangle_{\mathfrak{M}_X} = 0, \\ &&& k \in \{2, \ldots, d\}.\end{aligned}$$

Then, the set of functions $\{f_1(X), \ldots, f_d(X)\}$ spans a subspace of $\mathfrak{S}_{Y|X}$. Furthermore, if $\mathfrak{S}_{Y|X}$ is complete, these functions fully characterize the central class.

Our approach shares with Li and Song (2017) the fundamental goal. In particular, both methods operate under analogous assumptions to induce well defined covariance operators and make use of the Moore Penrose inverses in the regularization process. However, a fundamental difference between our approach and that of Li and Song (2017) is that our method is conceptually based on the principal fitted component methodology (Cook 2007; Cook and Forzani 2008). A key difference lies in the specific operator we employ to capture the dependence between X and Y. While Li and Song (2017) focus on $\Sigma_{XX}^{\dagger}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{\dagger}$ we instead introduce $\Sigma_{XX}^{\dagger}\Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YX}\Sigma_{XX}^{\dagger}$ so that the cross covariance is weighted by $\Sigma_{YY}^{\dagger}$. This modification normalizes the variation of Y in the operator, leading to a more canonical correlation type criterion. From a stability standpoint, inserting $\Sigma_{YY}^{\dagger}$ can also mitigate issues arising when Y is high dimensional or exhibits strong collinearity. By adjusting for the covariance structure of Y, directions with disproportionately large variances do not overshadow the rest.

In the NFPFC framework, the eigenvalues corresponding to the leading d eigenfunctions of the operator $\Sigma_{XX}^{\dagger}\Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YX}\Sigma_{XX}^{\dagger}$ can be seen as quantifying the amount of response relevant variation captured by each dimension. This concept parallels principal component analysis, in which eigenvalues represent the proportion of variance explained by each principal component. Hence, examining both the magnitudes and the decay pattern of these eigenvalues offers insights for determining the effective dimension d. For further methodological details in the context of linear SDR, see Li (2018). Li et al. (2011) and Li and Song (2017) propose a cross validated Bayesian Information Criterion (CVBIC), explicitly designed to balance model complexity against predictive performance systematically using eigenvalues.

CVBIC penalizes the inclusion of additional dimensions based on incremental predictive gains relative to the corresponding increase in model complexity, thus providing a rigorous criterion for dimension selection.

3 Sample-level implementation

In this section, we develop a sample-level algorithm to implement the population-level formulation of the NFPFC model.

3.1 Coordinate representation

For the development of the sample-level procedure, we adopt the coordinate notation system introduced in Lee et al. (2013) and Li and Song (2017). Let $Q = I_n - \frac{1_n 1_n^\mathsf{T}}{n}$, where 1_n denotes an n-dimensional column vector with each component equal to 1, and let K_X be the $n \times n$ Gram matrix defined as $(K_X)_{ij} = \kappa_X(X_i, X_j)$. Then, we define $G_X = QK_XQ$ and $G_Y = QK_YQ$.

Proposition 2 *At the sample level, the covariance operators can be expressed in matrix form as follows:*

$$
\begin{aligned}
&[\hat{\Sigma}_{XX}] = n^{-1}G_X, \quad [\hat{\Sigma}_{YY}] = n^{-1}G_Y,\\
&[\hat{\Sigma}_{YX}] = n^{-1}G_X, \quad [\hat{\Sigma}_{XY}] = n^{-1}G_Y,\\
&[\hat{\Sigma}^\dagger_{XX}] = nG^\dagger_X, \qquad [\hat{\Sigma}^\dagger_{YY}] = nG^\dagger_Y.
\end{aligned}
$$

3.2 Implementation of the NFPFC

Using Proposition 2, we can express the quantities in Corollary 1 in matrix form. The operator can be written as

$$
\begin{aligned}
nG^\dagger_X(n^{-1}G_Y)nG^\dagger_Y(n^{-1}G_X)nG^\dagger_X\\
= G^\dagger_X G_Y(n^{-1}G_Y)^\dagger G_X G^\dagger_X
\end{aligned}
$$

Hence, the inner product becomes

$$
\begin{aligned}
&\langle f, \hat{\Sigma}^\dagger_{XX}\hat{\Sigma}_{XY}\hat{\Sigma}^\dagger_{YY}\hat{\Sigma}_{YX}\hat{\Sigma}^\dagger_{XX} f\rangle_{\mathfrak{M}_X}\\
&\quad = \langle f, G^\dagger_X G_Y(n^{-1}G_Y)^\dagger G_X G^\dagger_X f\rangle_{\mathfrak{M}_X}\\
&\quad = [f]^\mathsf{T} G_X G^\dagger_X G_Y(n^{-1}G_Y)^\dagger G_X G^\dagger_X [f],
\end{aligned}
$$

where the second equality follows from $[f] = Q[f]$ and $G^\dagger_X = QG^\dagger_X$. To mitigate overfitting, we replace the Moore-Penrose inverses $G^\dagger_X$ and $G^\dagger_Y$ with the Tychonoff regularized inverses $(G_X + \theta_X I_n)^{-1}$ and $(G_Y + \theta_Y I_n)^{-1}$, where $\theta_X > 0$ and $\theta_Y > 0$ are tuning constants. This results in

$$
\begin{aligned}
&[f]^\mathsf{T} G_X(G_X + \theta_X I_n)^{-1} G_Y\\
&\quad (G_Y + \theta_Y I_n)^{-1} G_X(G_X + \theta_X I_n)^{-1}[f].
\end{aligned} \tag{4}
$$

To maximize equation (4) over $\mathfrak{M}^0_X$, the inner products in Corollary 1 can be expressed as $\langle f, f\rangle_{\mathfrak{M}_X} = [f]^\mathsf{T} G_X [f]$ and $\langle f, f_l\rangle_{\mathfrak{M}_X} = [f]^\mathsf{T} G_X [f_l]$. By defining $v = G_X^{1/2}[f]$ and applying Tychonoff regularization, we obtain

$$
[f] = (G_X + \theta_X I_n)^{-1/2} v.
$$

Reformulating the problem in terms of v, Corollary 1 leads to the following eigenvalue problem

$$
\begin{aligned}
\text{maximize} \quad & v^\mathsf{T}(G_X + \theta_X I_n)^{-3/2} G_X G_Y\\
& (G_Y + \theta_Y I_n)^{-1} G_X (G_X + \theta_X I_n)^{-3/2} v,\\
\text{subject to} \quad & v^\mathsf{T} v = 1,\ v^\mathsf{T} v_1 = 0,\ \ldots,\ v^\mathsf{T} v_{k-1} = 0,\\
& k \in \{2, \ldots, d\}.
\end{aligned}
$$

Thus, the vectors $v_1, \ldots, v_d$ correspond to the first d eigenvectors of the matrix

$$
\begin{aligned}
&(G_X + \theta_X I_n)^{-3/2} G_X G_Y\\
&\quad (G_Y + \theta_Y I_n)^{-1} G_X (G_X + \theta_X I_n)^{-3/2}.
\end{aligned} \tag{5}
$$

From this, the coefficients $[f_l]_{\mathfrak{B}_X}$ can be recovered as

$$
[f_l]_{\mathfrak{B}_X} = (G_X + \theta_X I_n)^{-1/2} v_l,
$$

where $\mathfrak{B}_X = \{\kappa(\cdot, X_i) - E_n\,\kappa(\cdot, X) : i = 1, \ldots, n\}$
$= \{b_1^{(X)}, \ldots, b_n^{(X)}\}$.

The final set of functions is given by

$$
\hat{f}_l = v_l^\mathsf{T}(G_X + \theta_X I_n)^{-1/2} Q b^{(X)}, \quad l = 1, \ldots, d,
$$

which represent the nonlinear sufficient predictors that span the approximate central class.

3.3 Tuning parameters

We adopt the Gaussian RBF as the kernel, defined as

$$
\kappa(s_1, s_2) = \exp(-\gamma \|s_1 - s_2\|^2),
$$

where $\|\cdot\|$ denotes the norm. Given observations $S_1^X, \ldots, S_n^X$ of X and $S_1^Y, \ldots, S_n^Y$ of Y, the tuning parameters γ_X and γ_Y are selected according to

$$
\gamma_j = \frac{\binom{n}{2}^2}{\sum_{a<b} \|S_a^j - S_b^j\|^2}, \quad j \in \{X, Y\}.
$$

The optimal values of the tuning parameters θ_X and θ_Y are selected using generalized cross-validation (GCV), following the approach in Li (2018) with Tychonoff regularization. Let $\lambda_{\max}(A)$ denote the largest eigenvalue of a matrix A, $\|\cdot\|_F$

represent the Frobenius norm, and tr(·) denote the trace of a matrix.

The GCV criteria are defined as

$$\mathrm{GCV}_X(\theta_X) = \frac{\|G_Y - G_X(G_X + \theta_X \lambda_{\max}(G_X) I_n)^{-1} G_Y\|_F^2}{\left(\mathrm{tr}[I_n - G_X(G_X + \theta_X \lambda_{\max}(G_X) I_n)^{-1}]\right)^2},$$

$$\mathrm{GCV}_Y(\theta_Y) = \frac{\|G_X - G_Y(G_Y + \theta_Y \lambda_{\max}(G_Y) I_n)^{-1} G_X\|_F^2}{\left(\mathrm{tr}[I_n - G_Y(G_Y + \theta_Y \lambda_{\max}(G_Y) I_n)^{-1}]\right)^2}.$$

By minimizing $\mathrm{GCV}_X(\theta_X)$ and $\mathrm{GCV}_Y(\theta_Y)$, we obtain the optimal tuning parameters that achieve a balance between preserving fidelity to the data and confirming smoothness in the estimator.

3.4 Dimension selection

A critical step in SDR methods is determining the structural dimension d, the minimal number of nonlinear components $f_1, \ldots, f_d \in \mathfrak{M}_X$. Our approach to selecting d follows the methodology established in Li et al. (2011) and Li and Song (2017). For classification problems with a relatively small categorical responses $Y \in \{1, \ldots, k\}$ where $k = 8$, it is common practice to set $d = k - 1$. However, to address situations involving continuous responses or categorical responses with a large number of classes, we employ an approach inspired by the Bayesian Information Criterion that has been adapted to the nonlinear functional setting. Following the framework in Li and Song (2017), we define a selection criterion as

$$G_n(k) = \sum_{i=1}^{k} \hat{\lambda}_i - \hat{\lambda}_1 n^{-\frac{1}{4}} \log(n) \cdot k, \tag{6}$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n$ denote the eigenvalues of the estimated operator $\widehat{\Sigma}_{XX}^{\dagger} \widehat{\Sigma}_{XY} \widehat{\Sigma}_{YY}^{\dagger} \widehat{\Sigma}_{YX} \widehat{\Sigma}_{XX}^{\dagger}$ in descending order. The first term of equation (6), $\sum_{i=1}^{k} \hat{\lambda}_i$, quantifies the total explained variation by the first k nonlinear eigenfunctions. The second term serves as a penalty for model complexity, scaling with sample size n, the leading eigenvalue $\hat{\lambda}_1$, and structural dimension candidate k. The penalty thus grows proportionally with k. The estimated optimal structural dimension $\hat{d}$ is then determined by maximizing the criterion as

$$\hat{d} = \underset{1 \leq k \leq n}{\arg\max}\, G_n(k).$$

Theoretical results in Li and Song (2017) demonstrate that this criterion consistently recovers the true structural dimension as $n \to \infty$, given appropriate regularity conditions and spectral decay assumptions. In practical implementation, we employ a systematic grid search procedure across a suitable range, $k = 1, \ldots, 10$ or until eigenvalues become negligible, and select the dimension that maximizes $G_n(k)$.

4 Asymptotic theory

In this section, we establish the asymptotic properties of the sample-level estimator

$$\widehat{T} = \widehat{\Sigma}_{XX}^{\dagger} \widehat{\Sigma}_{XY} \widehat{\Sigma}_{YY}^{\dagger} \widehat{\Sigma}_{YX} \widehat{\Sigma}_{XX}^{\dagger}.$$

Here, $\Sigma_{XX}^{\dagger}$ and $\Sigma_{YY}^{\dagger}$ are the Tikhonov-regularized inverses of the covariance operators Σ_{XX} and Σ_{YY}, respectively, while $\widehat{\Sigma}_{\cdot\cdot}$ are their empirical estimates. Under appropriate smoothness conditions and a suitable choice of the regularization parameters, we derive the convergence rate of $\widehat{T}$ to T and establish its dependence on the spectral decay of the population-level covariance operators.

Theorem 2 *Let $\{(X_i, Y_i)\}_{i=1}^n$ be an i.i.d. sample of functional data taking values in suitably defined RKHS $\mathcal{H}_X$ and $\mathcal{H}_Y$. Suppose the following conditions hold.*

(i) There exists an exponent $\alpha > 1$ and a constant $C_\alpha > 0$ such that, for all $j \geq 1$,

$$\lambda_j(\Sigma_{XX}) \leq C_\alpha\, j^{-\alpha},$$
$$\lambda_j(\Sigma_{YY}) \leq C_\alpha\, j^{-\alpha},$$

where $\{\lambda_j(\Sigma_{XX})\}$ and $\{\lambda_j(\Sigma_{YY})\}$ are the eigenvalues of the covariance operators Σ_{XX} and Σ_{YY}, respectively.

(ii) The regularization parameters θ_X and θ_Y satisfy

$$\theta_X \sim n^{-\frac{2\alpha}{2\alpha+1}} \quad \text{and} \quad \theta_Y \sim n^{-\frac{2\alpha}{2\alpha+1}},$$

balancing the bias–variance trade-off in the Tikhonov-regularized inverses $\Sigma_{XX}^{\dagger}$ and $\Sigma_{YY}^{\dagger}$ where $\sim$ indicates asymptotically proportional.

(iii) Each of the empirical covariance operators $\widehat{\Sigma}_{XY}$, $\widehat{\Sigma}_{YX}$, $\widehat{\Sigma}_{XX}$, $\widehat{\Sigma}_{YY}$ consistently estimates its population counterpart in operator norm at order $O_p(n^{-1/2})$ (Sriperumbudur et al. 2010; Fukumizu et al. 2007).

Then, there exists a constant $C > 0$ (independent of n) such that, with high probability,

$$\left\|\widehat{T} - T\right\|_{\mathcal{B}} \leq C\, n^{-\frac{\alpha}{2\alpha+1}},$$

as $n \to \infty$. Equivalently,

$$\left\|\widehat{T} - T\right\|_{\mathcal{B}} = O_p\left(n^{-\frac{\alpha}{2\alpha+1}}\right).$$

Proof The proof follows standard operator perturbation arguments and is omitted.

The condition $\theta_X, \theta_Y \to 0$ as $n \to \infty$ in Theorem 2 is an asymptotic requirement ensuring that the regularization bias vanishes in the limit of infinite data. This assumption is a sufficient condition for convergence, not a statement that overfitting becomes irrelevant for large but finite n. Similar asymptotic behavior occurs in kernel smoothing, spline models, and other nonparametric settings, where bandwidths or smoothing parameters typically diminish with growing sample size (Wahba 1990). In practice, one still needs to select θ_X and θ_Y adaptively via GCV at each fixed n to balance bias and variance. Therefore, θ_n does not literally approach zero at finite n, but rather tends to zero gradually as $n \to \infty$.

In contrast to the usual finite dimensional setting where convergence rates depend explicitly on the dimension, our infinite-dimensional framework replaces 'dimension' with the decay rate of the eigenvalues of the covariance operators. The exponent α in condition (i) characterizes this decay. A larger α indicates faster eigenvalue decay, leading to a faster convergence rate. Conversely, smaller values of α arise when the data exhibit more complex or less smooth structure. In practice, α is not directly observable, and its value can be estimated or approximated by examining how quickly the empirical eigenvalues decay. □

5 Simulation studies

In this section, we assess the performance of our proposed method under different scenarios for dimension reduction of functional data. We consider two scenarios: (i) the response is a random variable and the predictor is a random function; (ii) both the response and predictor are random functions. We use the weak inverse regression estimator (WIRE) in Li and Song (2022) as one of the comparison methods.

WIRE generalizes classical sliced inverse regression (Li 1991; Cook and Weisberg 1991) to functional data as Ferré and Yao (2003). However, unlike the existing methods, WIRE avoids direct estimation of $E(X \mid Y)$. Instead, it relies on weak conditional expectation, which defined as the inducing function of a Carleman operator in Weidmann (1980). Li and Song (2022) define $\mathcal{H}_X$ and $\mathcal{H}_Y$ as the RKHS spanned by $\{\kappa_T(\cdot, \tau_i) : i = 1, \cdots, m\}$, where $\kappa_T : J \times J \to \mathbb{R}$ be a positive definite kernel and time points $J_i = \{t_{i1}, \cdots, t_{im_i}\} \subseteq J$. They refer to the pair of kernels (κ_T, κ_Y) as nested kernels, and K_T as $m \times m$ matrix $\{\kappa_T(s,t) : s,t \in J\}$. The main difference between WIRE method and our approach lies in the coordinate matrix, since goal in WIRE is to find $\text{ran}(\Sigma_{XX}^{\dagger 1/2}\Sigma_{XY}M_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{\dagger 1/2})$, where $M_{YY} = n^{-1}K_Y$ and K_Y is the Gram matrix $\{\kappa_Y(Y_i, Y_j) : i, j = 1, \cdots, n\}$. The eigenvalue problem is

$$\begin{aligned} \text{maximize} \quad & \langle f, \Sigma_{XX}^{\dagger 1/2}\Sigma_{XY}M_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{\dagger 1/2} f\rangle_{\mathcal{H}_X} \\ \text{subject to} \quad & \langle f, f\rangle_{\mathcal{H}_X} = 1, \\ & \langle f, f_1\rangle_{\mathcal{H}_X} = \cdots = \langle f, f_{k-1}\rangle_{\mathcal{H}_X} = 0, \\ & k \in 2, \cdots, d. \end{aligned}$$

This can be reformulated as to find the first d eigenvectors of $A_X K_T^{1/2} B_{XY} K_T^{1/2} A_X$. Here, $A_X = (n^{-1}K_T^{1/2}[X_{1:n}]Q_n[X_{1:n}]^{\mathsf{T}}K_T^{1/2} + \theta_X I_m)^{-1/2}$, and $B_{XY} = [X_{1:n}]Q_n(n^{-1}K_Y)(n^{-1}K_Y + \theta_Y I_n)^{-1}Q_n[X_{1:n}]^{\mathsf{T}}$. In this structure, WIRE is adept at capturing linear relationship between predictor and response. For more details, see Li and Song (2022).

To evaluate the performance in each model, we compare the estimated and true predictors using a multivariate version of Spearman's correlation called the Multiple Correlation of Multivariate rank (MCMR). Let $U_1, \cdots, U_n \in \mathbb{R}^r$ and $V_1, \cdots, V_n \in \mathbb{R}^s$ be two samples of random vectors representing the estimated and true predictors, respectively. The multivariate ranks are defined as $\tilde{U}_i = n^{-1}\sum_{l=1}^n (U_l - U_i)/\|U_l - U_i\|$ and $\tilde{V}_i = n^{-1}\sum_{l=1}^n (V_l - V_i)/\|V_l - V_i\|$. The MCMR is then defined as

$$\text{mcm} \quad \text{r}_n(U, V) = \Big(\text{tr}\Big\{[\text{var}_n(\tilde{V})]^{-1/2}\text{cov}_n(\tilde{V}, \tilde{U}) [\text{var}_n(\tilde{U})]^{-1}\text{cov}_n(\tilde{U}, \tilde{V})[\text{var}_n(\tilde{V})]^{-1/2}\Big\}\Big)^{1/2}.$$

Here, a higher MCMR value signifies better performance.

In each simulation setting, we generate a total of n independent datasets. For massive datasets, subsampling based GCV approach in which a smaller subset can be used to evaluate the cross validation criterion, rather than using all n observations. A moderately sized subsample can provide a reliable estimate of the regularization parameter at a fraction of the cost. In practice, one may draw multiple subsamples, compute the GCV selected tuning parameter on each, and then use a median of those estimates as the final choice. Furthermore, although one could compute cross validated tuning parameters individually for each dataset, we find empirically that these tuning parameters exhibit minimal variability across replicates. Therefore, to reduce computational costs, we conduct the GCV procedure only on the first ten simulation replicates. For each of these ten replicates, we determine the optimal tuning parameters by minimizing the GCV criterion over prespecified grids. We define a grid of 20 candidate values for θ_X and θ_Y, where the first point is $n^{-1/4}$, and the remaining 19 points form a log-spaced sequence from $\frac{n^{-1/4}}{50}$ to $50\,n^{-1/4}$. We then average these ten optimal values to yield a single tuning parameter used for the remaining datasets.

Table 1 Comparison of average MCMR values and their standard errors for various methods across Models I-1 to I-3.

Model	FSIR	WIRE	NFPFC	FGSIR	FPTU
Model I-1	0.153 (0.006)	0.271 (0.011)	0.943 (0.002)	0.953 (0.001)	0.745 (0.015)
Model I-2	0.162 (0.007)	0.331 (0.013)	0.792 (0.007)	0.706 (0.003)	0.341 (0.016)
Model I-3	0.157 (0.006)	0.309 (0.014)	0.727 (0.009)	0.734 (0.004)	0.447 (0.009)

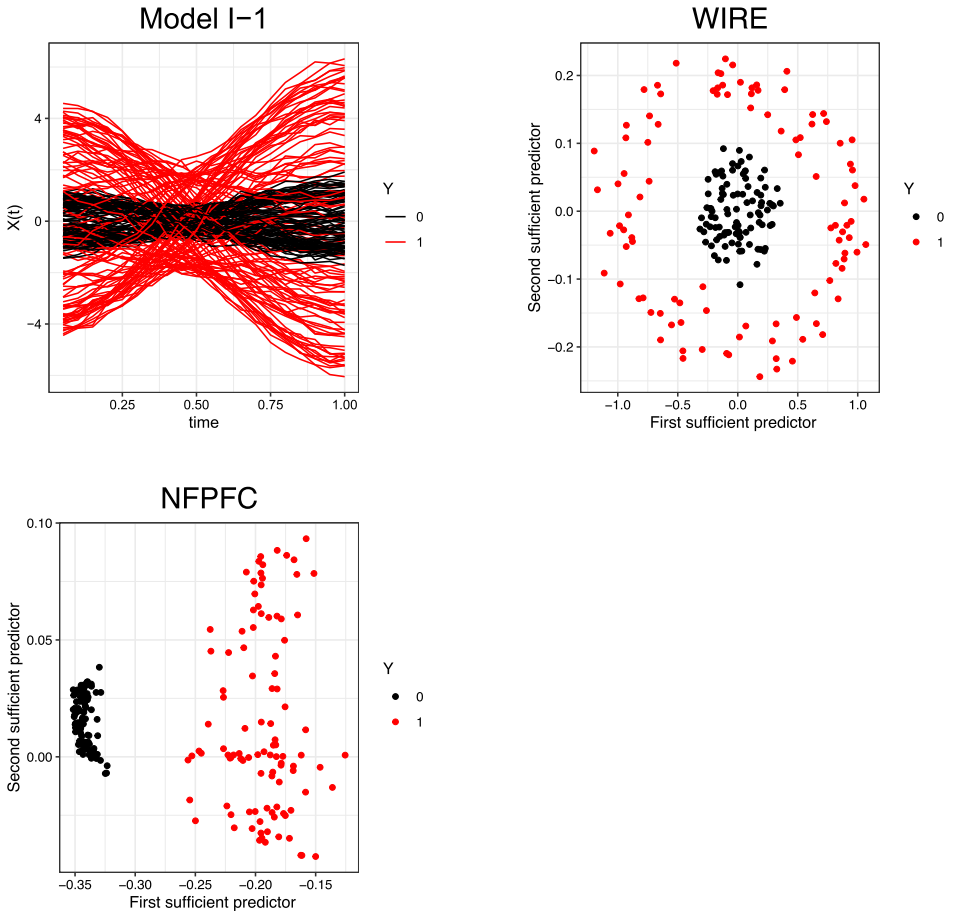

Fig. 2 Observed curves of $X_i(t)$ and the first two sufficient predictors or FPTU components from various methods for Model I-1. (Black: Class 0, Red: Class 1.)

5.1 Scenario I: scalar-on-function (Inverse Regression)

We initially present our methodology under the broad rubric of function to function regression. Our framework naturally encompasses scalar or vector responses as well, since any finite dimensional Euclidean space $\mathbb{R}^m$ can be endowed with a standard inner product and thus regarded as a Hilbert space. Let $\mathcal{H}_Y$ denote the usual Hilbert space structure on $\mathbb{R}^m$, where the inner product is given by $\langle u, v\rangle_{\mathcal{H}_Y} = \sum_{j=1}^m u_j\, v_j, \quad u, v \in \mathbb{R}^m$. As a result, if $m = 1$, $Y \in \mathbb{R}$, then Y is simply an element of a one dimensional Hilbert space. Thus, the special case of having a scalar or vector response $Y \in \mathbb{R}^m$ is fully covered by the general Hilbert space framework. We consider three models (Model I-1, Model I-2, and Model I-3) with regularly observed univariate functional data. Each model explores different relationships between the binary response variable Y (where $Y = 0$ or $Y = 1$) and the functional predictor $X(t)$ with inverse regression. We set the sample size $n = 100$, and each curve $X(t)$ is observed at 10 equally spaced time points in $[0, 1]$. We use Gaussian RBF kernel to construct the Gram matrix.

Model I-1 captures a highly nonlinear relationship between Y and $X(t)$. We generate a random sample $Y_i \sim$ Bernoulli(0.5) for $i = 1, \cdots, n$, and generate $X_i(t)$ based on

Model I-1 :

$$\begin{cases} X_i(t) \mid Y_i = 0 \sim Z_{1i}\cos(\theta_{1i})\cos(\pi t) \\ \qquad\qquad + Z_{1i}\sin(\theta_{1i})\sin(t) + \epsilon_i(t), \\ X_i(t) \mid Y_i = 1 \sim Z_{2i}\cos(\theta_{2i})\cos(\pi t) \\ \qquad\qquad + Z_{2i}\sin(\theta_{2i})\sin(t) + \epsilon_i(t), \end{cases}$$

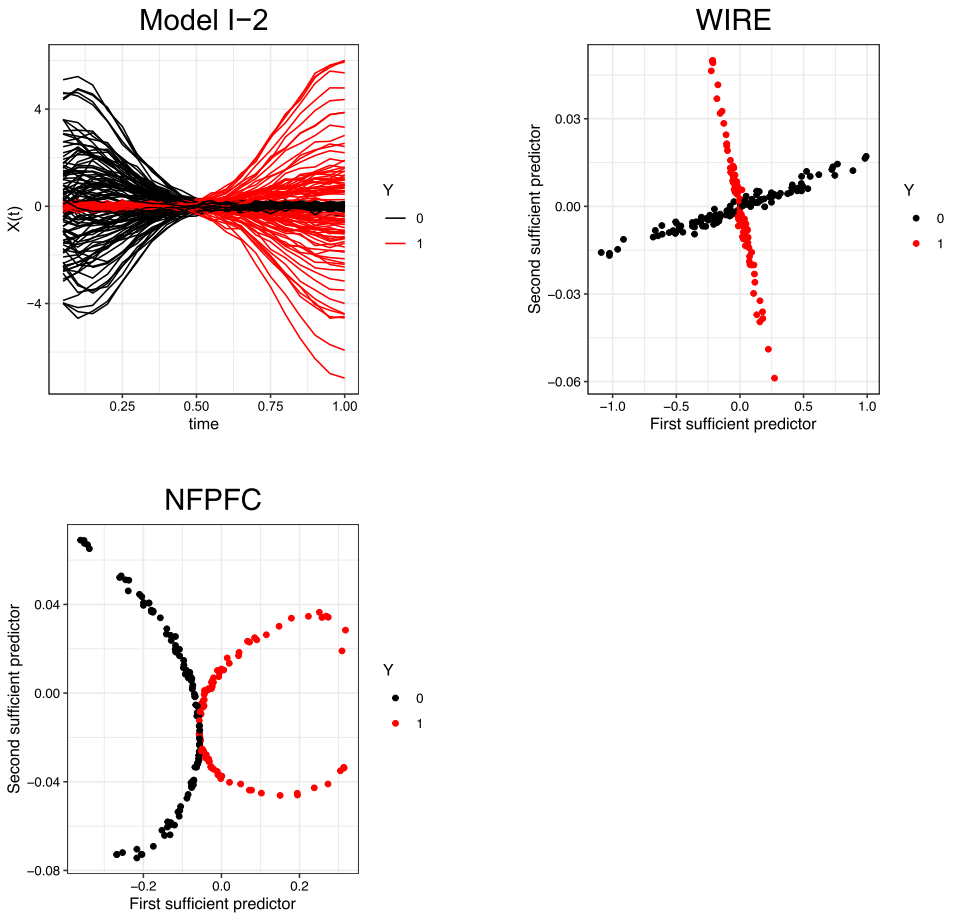

Fig. 3 Observed curves of $X_i(t)$ and the first two sufficient predictors or FPTU components from various methods for Model I-2. (Black: Class 0, Red: Class 1.)

where $Z_{1i} \sim N(1, 0.2^2)$, $\theta_{1i} \sim U(0, 2\pi)$, $Z_{2i} \sim N(4, 0.5^2)$, $\theta_{2i} \sim U(0, 2\pi)$, and $\epsilon_i(t_j) \sim N(0, 0.1^2)$. All random variables Z_{1i}, θ_{1i}, Z_{2i}, θ_{2i}, and $\epsilon_i(t)$ are independently sampled.

Next, we consider two models representing linear relationships, by using B-spline basis functions and monomial basis functions to construct dimension reduction directions. Our goal is to evaluate the performance of the NFPFC under linear conditions. Similar to Model I-1, we generate a random sample $Y_i \sim$ Bernoulli(0.5) for $i = 1, \cdots, n$, and generate $X_i(t)$ from following models as

Model I-2 :

$$\begin{cases} X_i(t) \mid Y_i = 0 \sim Z_{1i}(s_1(t) + s_2(t)) \\ \qquad\qquad + Z_{2i}s_2(t) + \epsilon_i(t), \\ X_i(t) \mid Y_i = 1 \sim (Z_{1i} + Z_{2i}) \\ \qquad\qquad (s_5(t) + s_6(t)) + \epsilon_i(t), \end{cases}$$

Model I-3 :

$$\begin{cases} X_i(t) \mid Y_i = 0 \sim (Z_{1i} + Z_{2i})m_2(t) + \epsilon_i(t), \\ X_i(t) \mid Y_i = 1 \sim (Z_{1i} + Z_{2i}) \\ \qquad\qquad (m_5(t) + m_6(t)) + \epsilon_i(t), \end{cases}$$

where $s_1(t), \cdots, s_6(t)$ are the 6 B-spline basis functions defined on $[0, 1]$, and $m_1(t), \cdots, m_6(t)$ are the 6 monomial basis functions with $m_j(t) = t^{j-1}, j = 1, \cdots, 6$ defined on $[0, 1]$. $Z_{1i} \sim N(0, 2^2)$, $Z_{2i} \sim N(0, 2^2)$, and $\epsilon_i(t) \sim N(0, 0.1^2)$ are independently sampled.

We first estimate two predictors using WIRE, NFPFC, functional sliced inverse regression (FSIR; Ferré and Yao 2003), functional generalized sliced inverse regression (FGSIR; Li and Song 2017), and functional parallel transport unfolding (FPTU; Tan et al. 2024). We then calculate the MCMR between the estimated and true predictors from the test set. This process is repeated 100 times, and we report the average MCMR values and their standard errors in Table 1. For Model I-1, NFPFC attains high MCMR value compared to WIRE and FPTU, which indicates that it performs better than the other competitors in this setting. In Model I-2, NFPFC achieves the highest MCMR, outperforming FGSIR as well as WIRE and FPTU. A similar trend is observed in Model I-3, where NFPFC and FGSIR again stand out, registering higher MCMR values than WIRE and FPTU. Therefore, we can observe that NFPFC exhibits strong performance in these scenarios. Table 1 further provides that FSIR (Ferré and Yao 2003), which is a linear functional

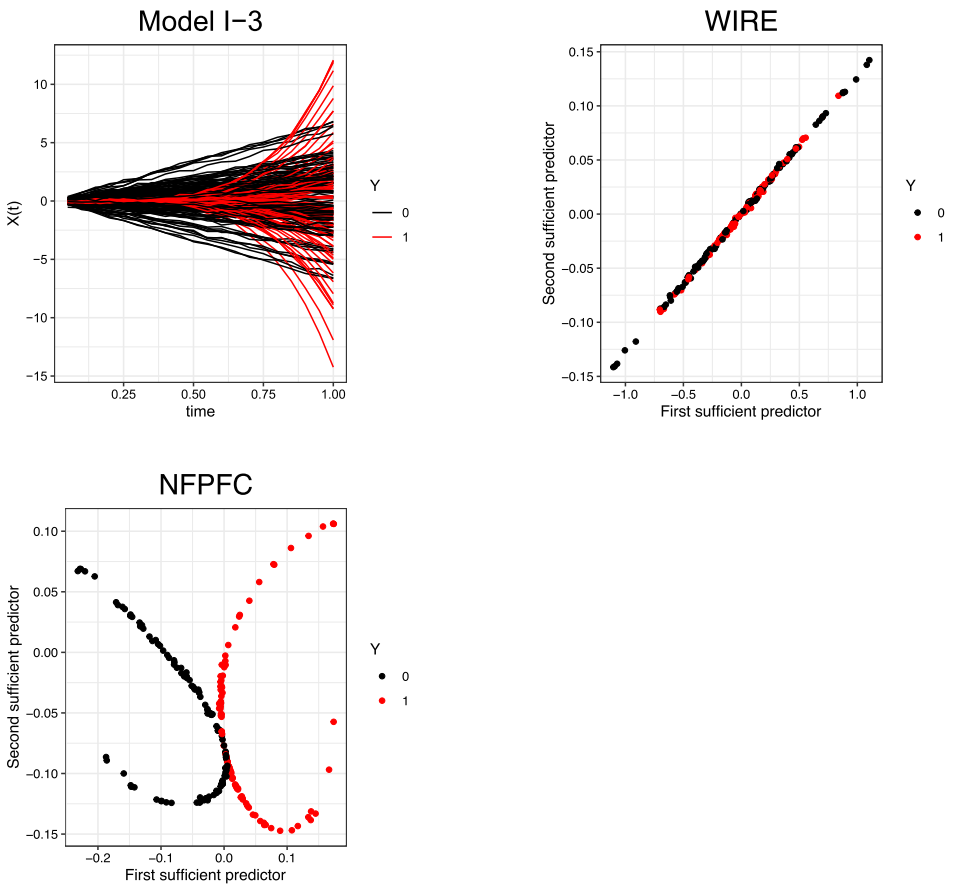

Fig. 4 Observed curves of $X_i(t)$ and the first two sufficient predictors or FPTU components from various methods for Model I-3. (Black: Class 0, Red: Class 1.)

SDR method, demonstrates consistently limited performance in all three models, with low average MCMR values ranging approximately between 0.15 and 0.16. This indicates that FSIR fails to effectively capture nonlinear relationships inherent in these simulation scenarios.

Figure 2-4 show the observed curves and the dimension reduction results for each model using WIRE, NFPFC, FSIR, FGSIR, and FPTU. In Figure 2, the observed curves display a highly nonlinear structure with noticeable overlap between two classes. WIRE, FSIR, and FPTU struggle to separate the classes due to this nonlinearity, while NFPFC effectively differentiates them by capturing the complex structure. In Figure 3, WIRE, FSIR, and FPTU partially separates two classes along a diagonal, while NFPFC presents successful classification results with nonlinearities. In Figure 4, WIRE captures the main linear pattern but leaves significant overlap, whereas NFPFC achieves clearer class separation.

5.2 Scenario II: function-on-function (Forward Regression)

In scenario II, we consider three models (Model II-1, Model II-2, and Model II-3) to explore different relationships between the functional response $Y_i(t)$ and the functional predictor $X_i(t)$ with forward regression. In all models, we use the Brownian motion to generate $X_i(t)$ as

$$X_i(t) = \sum_{j=1}^{100} \sqrt{2} a_{ij} \frac{\sin((j-1/2)\pi t)}{(j-1/2)\pi},$$

where a_{ij} are independently sampled from $N(0,1)$ for $i = 1, \cdots, n$ and $j = 1, \cdots, 100$. We set the sample size $n = 200$, and each curve $X_i(t)$ and $Y_i(t)$ are observed at 20 equally spaced time points in $[0, 1]$. For the construction of Gram matrix, we also use Gaussian RBF kernel. We consider following models

Table 2 Comparison of average MCMR values and their standard errors for various methods across Models II-1 to II-3.

Model	WIRE	NFPFC	FGSIR
Model II-1	0.924 (0.003)	0.886 (0.007)	0.991 (0.000)
Model II-2	0.175 (0.009)	0.898 (0.004)	0.902 (0.003)
Model II-3	0.660 (0.011)	0.956 (0.002)	0.939 (0.002)

Model II-1 :

$$Y_i(t) = \{\langle X_i, b_1\rangle + \langle X_i, b_2\rangle\}\rho(t) + \sigma\epsilon_i(t),$$

Table 3 Average MCMR values with standard errors in parentheses for Model II-2 across different combinations of (θ_X, θ_Y) and (γ_X, γ_Y).

	$(\gamma_X, \gamma_Y) = (2, 2)$				$(\gamma_X, \gamma_Y) = (2, 4)$			
	$\theta_X = 0.001$	$\theta_X = 0.01$	$\theta_X = 0.05$	$\theta_X = 0.1$	$\theta_X = 0.001$	$\theta_X = 0.01$	$\theta_X = 0.05$	$\theta_X = 0.1$
$\theta_Y = 0.001$	0.766 (0.018)	0.912 (0.007)	0.916 (0.009)	0.914 (0.008)	0.767 (0.014)	0.865 (0.013)	0.912 (0.009)	0.910 (0.007)
$\theta_Y = 0.01$	0.752 (0.022)	0.915 (0.009)	0.914 (0.007)	0.910 (0.008)	0.721 (0.023)	0.908 (0.009)	0.913 (0.009)	0.910 (0.007)
$\theta_Y = 0.05$	0.819 (0.022)	0.906 (0.010)	0.906 (0.008)	0.907 (0.008)	0.725 (0.040)	0.894 (0.014)	0.907 (0.008)	0.903 (0.008)
$\theta_Y = 0.1$	0.841 (0.018)	0.905 (0.009)	0.906 (0.008)	0.906 (0.008)	0.744 (0.038)	0.893 (0.014)	0.905 (0.008)	0.903 (0.008)
	$(\gamma_X, \gamma_Y) = (4, 2)$				$(\gamma_X, \gamma_Y) = (4, 4)$			
	$\theta_X = 0.001$	$\theta_X = 0.01$	$\theta_X = 0.05$	$\theta_X = 0.1$	$\theta_X = 0.001$	$\theta_X = 0.01$	$\theta_X = 0.05$	$\theta_X = 0.1$
$\theta_Y = 0.001$	0.774 (0.019)	0.886 (0.011)	0.915 (0.008)	0.915 (0.008)	0.745 (0.017)	0.882 (0.013)	0.915 (0.008)	0.911 (0.007)
$\theta_Y = 0.01$	0.810 (0.026)	0.894 (0.012)	0.901 (0.008)	0.907 (0.007)	0.720 (0.024)	0.861 (0.017)	0.907 (0.007)	0.905 (0.008)
$\theta_Y = 0.05$	0.825 (0.026)	0.899 (0.010)	0.900 (0.008)	0.902 (0.007)	0.776 (0.023)	0.881 (0.014)	0.902 (0.009)	0.902 (0.008)
$\theta_Y = 0.1$	0.849 (0.021)	0.899 (0.010)	0.899 (0.008)	0.901 (0.008)	0.786 (0.030)	0.884 (0.014)	0.900 (0.009)	0.900 (0.008)

Model II-2 :

$$Y_i(t) = \left\{ \frac{\langle X_i, b_1\rangle \langle X_i, b_2\rangle}{1 + \exp(\langle X_i, b_3\rangle)} \right\} \rho(t) + \sigma \epsilon_i(t),$$

Model II-3 :

$$Y_i(t) = \cos\{\langle X_i, b_1\rangle + \langle X_i, b_2\rangle\} \sin\{\langle X_i, b_1\rangle + \langle X_i, b_2\rangle\} \rho(t) + \sigma \epsilon_i(t),$$

where $\langle X_i, b_j\rangle = \int_0^1 X_i(t) b_j(t) dt$, $b_j(t)$ are taken to the eigenfunctions $v_j(t) = \sqrt{2}\sin((j - 1/2)\pi t)$, $\rho(t) = \sum_{j=1}^{5} v_j(t)$, $\sigma = 0.1$, and $\epsilon_i(t)$ is generated from the standard Brownian motion. In Model II-1, we assume a linear relationship between $X_i(t)$ and $Y_i(t)$. Model II-2 and Model II-3 introduce nonlinearity through a variation of logistic function and trigonometric product, respectively. Again, we apply WIRE, NFPFC, and FGSIR to estimate sufficient predictors, and the MCMR results are in Table 2.

Table 2 compares the performance of WIRE, NFPFC, and FGSIR based on average MCMR values across three scenarios (Models II-1 to II-3). Since NFPFC is a nonlinear SDR method, it is natural that in linear structures, such as Model II-1, it does not perform as well as WIRE. However, in Models II-2 and II-3, NFPFC significantly outperforms WIRE due to its ability to capture nonlinearity in complex models. Overall, FGSIR and NFPFC demonstrate competitive performance, while WIRE exhibits variability and generally weaker performance in nonlinear settings.

We conducted a sensitivity analysis using Model II-2 to investigate how variations in the Gaussian RBF kernel parameters (γ_X, γ_Y) and the Tychonoff regularization parameters (θ_X, θ_Y) affect the performance of our method. Table 3 summarizes the MCMR results across a selected grid of these parameters. The results indicate limited fluctuations in performance across various parameter combinations. The best results from this sensitivity analysis closely align with those obtained using GCV, suggesting that GCV provides a practical approach for selecting optimal tuning parameters in practice.

5.3 Scenario III: Additional models

We consider the setup inspired by Liang et al. (2022) and Wang et al. (2015). We generate functional predictors $\{X_i(t)\}_{i=1}^n$ as realizations of independent standard Brownian motions observed on a uniform grid of 100 points in the interval [0, 1]. We then construct two variations of the response:

Model III-1 : $Y_i = \left|\langle X_i, \beta_1\rangle\right| + \langle X_i, \beta_2\rangle + \epsilon_i,$

Model III-2 : $Y_i = \left|\langle X_i, \beta_1\rangle\right| + \left|\langle X_i, \beta_2\rangle\right| + \epsilon_i,$

where $\epsilon_i \sim N(0,\ 0.1^2)$. The coefficient functions are defined as $\beta_1(t) = (2t - 1)^3 + 1$ and $\beta_2(t) = \cos\left[\pi(2t - 1)\right] +$ 1. By taking absolute values of inner products, we emulate nonlinear features while maintaining a fundamentally linear dimension reduction structure.

Table 4 Comparison of average MCMR values and their standard errors for WIRE and NFPFC in Models III-1 and III-2.

Model	WIRE	NFPFC
Model III-1	0.832 (0.004)	0.712 (0.006)
Model III-2	0.179 (0.007)	0.722 (0.005)

As we can observe from Table 4, under Model III-1, the MCMR measure for WIRE is higher than that of NFPFC, indicating that WIRE outperforms NFPFC in linear setting. By contrast, for Model III-2, WIRE's MCMR becomes substantially lower, demonstrating that NFPFC outperforms WIRE in this structure.

6 Real data applications

We explore the phoneme classification dataset, available in the `fda.usc` R package. The dataset, introduced by Hastie et al. (1995), consists of 1, 000 log-periodograms of length 256, with five phoneme classes such as "sh", "iy", "dcl", "aa", and "ao". The phonemes are transcribed as follows: "sh" as in "she", "iy" as the vowel in "she", "dcl" as in "dark", "aa" as the vowel in "dark", and "ao" as the first vowel in "water". For speech recognition, a log-periodogram is computed from each speech frame and Figure 5 represent a sample of 10 log-periodograms per phoneme class.

Next, we utilize WIRE and NFPFC to reduce the dimensionality of these functional data representations, followed by classification based on the reduced predictors. Figure 6 displays the first two sufficient predictors from WIRE and NFPFC, evaluated for the phoneme training data. The left plot in Figure 6 shows the WIRE results, where the phoneme classes exhibit some clustering but overlap significantly, especially between "sh" and "dcl" as well as "iy" and "aa". This overlap indicates that WIRE struggles to capture the nuanced differences between these phonemes, as it primarily relies on linear structures. In contrast, the right plot in Figure 6 illustrates the NFPFC results, where the five phoneme classes are more distinctly separated. NFPFC's ability to capture nonlinear patterns in the data leads to better clustering of each class.

To quantitatively evaluate the classification performance, we conducted experiments with different numbers of predictors obtained from WIRE and NFPFC. These reduced predictors served as inputs to three classification methods: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and support vector machines (SVM). Table 5 summarizes the classification accuracies with 2, 4, and

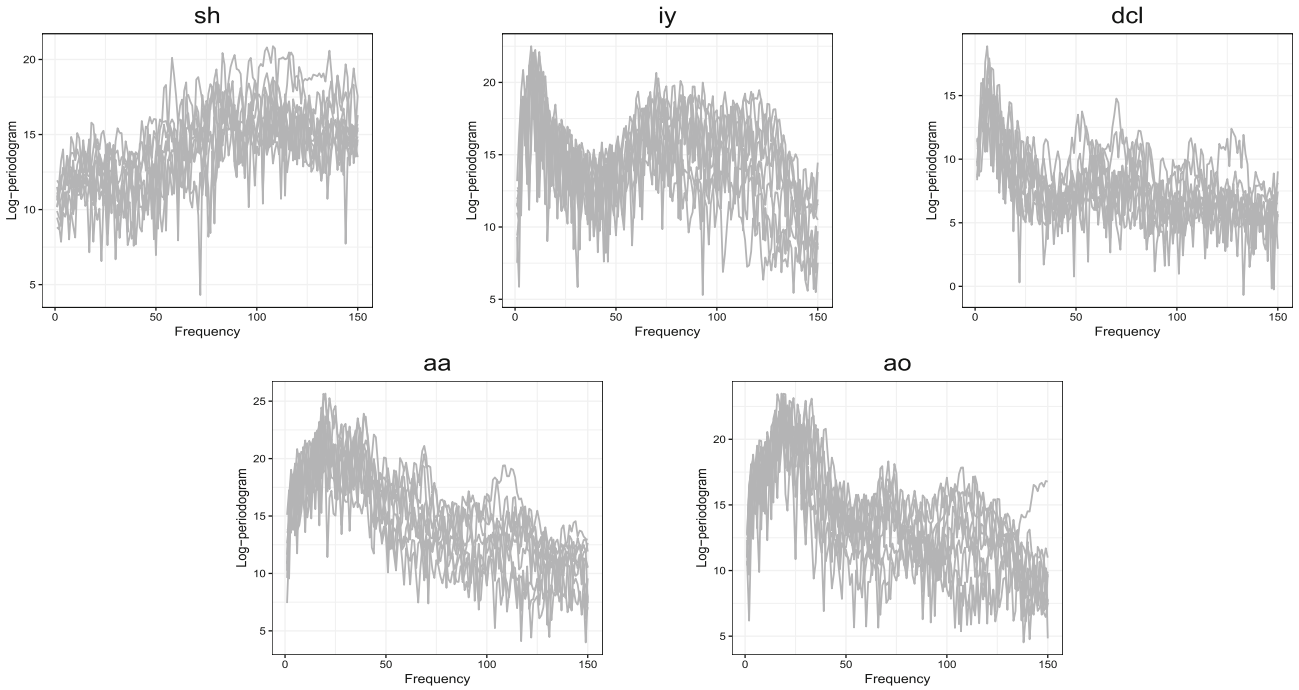

Fig. 5 Sample of 10 log-periodograms for each phoneme class

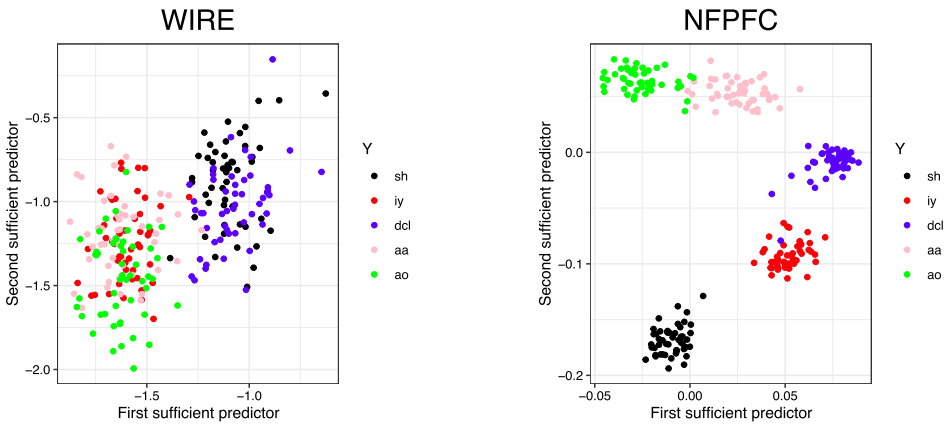

Fig. 6 The first two sufficient predictors from WIRE and NFPFC evaluated on the phoneme training data

Table 5 Percentages of correct classifications for the phoneme dataset using different numbers of predictors.

d	Sample Size	Method	WIRE			NFPFC		
			LDA	QDA	SVM	LDA	QDA	SVM
2	250	Accuracy(%)	53.6	52.4	50.4	90.4	91.6	90.8
4			93.6	89.2	92.0	93.6	94.4	95.2
5			94.4	91.2	91.2	94.4	94.4	95.2

Table 6 Percentages of correct classifications for the phoneme dataset with FPCA.

# of Principal Components	Method	FPCA		
		LDA	QDA	SVM
2	Accuracy(%)	84.8	82.4	82.4
4		92.8	90.0	91.2

5 predictors. The choice of 2 predictors follows from the dimension selection method described in Section 3.4, assuming a relatively large number (five categories) for the response variable. In addition, we considered 4 predictors, corresponding to scenarios with fewer categories, as discussed in Section 3.4. We also explored the performance with 5 predictors to further assess classification capabilities.

From Table 5, we observe that when using 2 predictors, NFPFC significantly outperforms WIRE across all three classifiers. It also provides better class separation, as shown in Figure 6. Similarly, when using 4 predictors, NFPFC continues to demonstrate superior performance compared to WIRE, particularly noticeable with QDA and SVM classifiers. Furthermore, increasing the number of predictors to 5 maintains classification accuracy for both methods, with NFPFC consistently outperforming or matching WIRE.

Table 6 compares classification accuracies after dimension reduction using an unsupervised functional principal component analysis (FPCA; Ramsay and Silverman 2005) approach. NFPFC produces consistently higher accuracy rates in Table 5, outperforming FPCA based results. These demonstrate the advantage of supervised dimension reduction by NFPFC, which appears more effective than the unsupervised FPCA in capturing discriminative features for classification.

7 Discussion

In this paper, we introduce a novel approach to SDR for functional data by extending the PFC model. We adapt the FPFC model and generalize it to the NFPFC model within the framework of RKHS.

Our work simultaneously addresses the challenges posed by the infinite dimensional nature of functional data and the complexities introduced by nonlinear relationships between functional predictors and responses. By leveraging the nested Hilbert space theory and the properties of RKHS, we establish a population-level formulation of the NFPFC model and provide a practical sample-level implementation using coordinate representations and regularization methods. Furthermore, we develop an asymptotic theory that characterizes the convergence properties of the proposed estimator. Our theoretical results establish the rate of convergence under mild regularity conditions.

Simulation studies demonstrated the effectiveness of our proposed method. In scenarios where the relationship between the functional predictor and response is nonlinear, the NFPFC consistently outperforms other methods. Additionally, we apply our method to the phoneme classification dataset, where NFPFC exhibits superior performance in distinguishing different phoneme classes by effectively capturing nonlinear patterns in log-periodograms.

We provide a framework for SDR, particularly in cases involving nonlinear relationships. The ability to handle both functional predictors and responses broadens the applicability of dimension reduction across various disciplines, including bioinformatics, climatology, and speech recognition.

Future research could explore several extensions of our work. One potential direction is to apply these nonlinear functional settings to the extended PFC model proposed in Cook (2007) or the unstructured PFC model in Cook and Forzani (2008). Furthermore, extending the framework to accommodate functional data with more complex structures, such as sparsely observed or irregularly sampled functions, would further enhance the utility of the proposed methods. Linear FPFC will also be explored in future research to extend and complement our current methodological framework.

Appendix

The proof of Theorem 1

Proof The objective in equation (3) can be written as

$$R(\Gamma, \beta) \;=\; E_n \Big\| \Lambda'(X) \;-\; \sum_{k=1}^{d} (\Gamma_k \otimes \beta_k)\, \Theta(y) \Big\|_{\mathfrak{M}}^2,$$

where E_n denotes the empirical expectation. Expanding the squared norm,

$$\begin{aligned} R(\Gamma, \beta) \;=\; E_n \Big[& \|\Lambda'(X)\|_{\mathfrak{M}_X}^2 \\ & - 2 \sum_{k=1}^{d} \langle \Gamma_k, \Lambda'(X) \rangle_{\mathfrak{M}_X} \, \langle \beta_k, \Theta(y) \rangle_{\mathfrak{M}_Y} \\ & + \sum_{k=1}^{d} \langle \beta_k, \Theta(y) \rangle_{\mathfrak{M}_Y}^2 \Big]. \end{aligned} \tag{7}$$

Because $\langle \Gamma_k, \Gamma_j \rangle_{\mathfrak{M}_X} = \delta_{kj}$ (the Kronecker delta), cross-terms cancel out. Note that (7) is quadratic in each β_k with Γ fixed. Let $\mathrm{D}_\beta R : \mathfrak{M} \to \mathbb{R}$ denote the Fréchet derivative of R with respect to β. For any $h = (h_1, \ldots, h_k)^{\mathsf{T}}$ with each $h_i \in \mathcal{H}$, we obtain

$$\begin{aligned} D_\beta R(h, \Gamma, \beta) = \sum_{k=1}^{d} E_n \Big[& - 2 \, \langle \Gamma_k, \Lambda'(X) \rangle \, \langle h_k, \Theta(y) \rangle \\ & + 2 \, \langle \beta_k, \Theta(y) \rangle \, \langle h_k, \Theta(y) \rangle \Big] = 0. \end{aligned}$$

Solving for β_k gives

$$\hat{\beta}_k \;=\; E_n \big[\Theta(y) \otimes \Theta(y) \big]^{-1} \, E_n \big[\Theta(y) \otimes \Lambda'(X) \big] \, \Gamma_k.$$

Here, we use the property of $E_n(A)$ in $\mathfrak{M}$ with $\langle f, E_n(A) g \rangle_{\mathfrak{M}} = E_n\big(\langle f, A\, g \rangle_{\mathfrak{M}}\big)$ and $\langle f', f \rangle_{\mathfrak{M}} \langle g', g \rangle_{\mathfrak{M}} = \langle (g \otimes f)\, f', g' \rangle_{\mathfrak{M}}$. See Li (2018) for details. Substituting $\hat{\beta}_k$ back into the objec-

tive, one obtains

$$R(\Gamma,\beta) = E_n\|\Lambda'(X)\|^2_{\mathfrak{M}_X} - 2\,E_n\sum_{k=1}^{d}\langle\Gamma_k,\ \Lambda'(X)\rangle_{\mathfrak{M}_X}\langle E_n[\Theta(y)\otimes\Theta(y)]^{-1} E_n[\Theta(y)\otimes\Lambda'(X)]\,\Gamma_k,\,\Theta(y)\rangle_{\mathfrak{M}} + E_n\sum_{k=1}^{d}\Big\langle E_n[\Theta(y)\otimes\Theta(y)]^{-1} E_n[\Theta(y)\otimes\Lambda'(X)]\,\Gamma_k,\ \Theta(y)\Big\rangle^2_{\mathfrak{M}}. \tag{8}$$

The second term can be written as

$$-2\sum_{k=1}^{d}\langle\Gamma_k, E_n[\Lambda'(X)\otimes\Theta(y)]\,E_n[\Theta(y)\otimes\Theta(y)]^{-1} E_n[\Theta(y)\otimes\Lambda'(X)]\,\Gamma_k\rangle_{\mathfrak{M}},$$

and the third term as

$$\sum_{k=1}^{d}\Big\langle E_n[\Theta(y)\otimes\Theta(y)]^{-1}\,E_n[\Theta(y)\otimes\Lambda'(X)]\,\Gamma_k, E_n[\Theta(y)\otimes\Theta(y)]\,E_n[\Theta(y)\otimes\Theta(y)]^{-1} E_n[\Theta(y)\otimes\Lambda'(X)]\,\Gamma_k\Big\rangle_{\mathfrak{M}}.$$

Recognizing that $E_n[\Theta(y)\otimes\Theta(y)]^{-1}E_n[\Theta(y)\otimes\Lambda'(X)]$ acts as an adjoint operator, we use the identity $\langle A\Gamma_k, B\rangle = \langle\Gamma_k, A^*B\rangle$ to rewrite the third term equivalently as

$$\sum_{k=1}^{d}\langle\Gamma_k, E_n[\Lambda'(X)\otimes\Theta(y)]\,E_n[\Theta(y)\otimes\Theta(y)]^{-1} E_n[\Theta(y)\otimes\Lambda'(X)]\,\Gamma_k\rangle_{\mathfrak{M}}.$$

Hence, (8) simplifies to

$$R(\Gamma,\beta) = E_n\|\Lambda'(X)\|^2_{\mathfrak{M}_X} - \sum_{k=1}^{d}\langle\Gamma_k, E_n[\Lambda'(X)\otimes\Theta(y)]\,E_n[\Theta(y)\otimes\Theta(y)]^{-1} E_n[\Theta(y)\otimes\Lambda'(X)]\,\Gamma_k\rangle_{\mathfrak{M}}.$$

Therefore, the minimization problem in (3) reduces to identifying the first d eigenfunctions $f_1,\ldots,f_d$ of the operator $\hat{\Sigma}_{XY}\,\hat{\Sigma}^{\dagger}_{YY}\,\hat{\Sigma}_{YX}$, where d is specified in advance. □

Additional simulation results

Table 7 compares the performance of WIRE, NFPFC, and FGSIR based on average distance correlation (DCOR) values across Models I-1 to I-3. Distance correlation (Székely et al. 2007) measures both linear and nonlinear dependence, capturing a broader range of associations compared to classical correlation methods. DCOR values near 1 imply strong dependence, while values close to 0 indicate independence. In Models I-1 and I-2, NFPFC outperforms WIRE and FGSIR, demonstrating its competitiveness in detecting nonlinear associations. All methods show relatively low DCOR values in Model I-3, indicating weaker overall dependence, with WIRE and FGSIR slightly outperforming NFPFC. These results demonstrate that NFPFC generally offers stronger performance in capturing complex relationships. Table 8 reports the average DCOR results for WIRE, NFPFC, and FGSIR in Models II-1, II-2, and II-3. In all three models, NFPFC shows higher results compared to WIRE. The gap is largest in Model II-1 and Model II-3, while in Model II-2, WIRE's score is low relative to NFPFC. Although FGSIR shows slightly higher DCOR in Models II-1 and II-3, the difference is often small, illustrating that NFPFC remains highly competitive.

Table 7 Comparison of average DCOR values and their standard errors for WIRE, NFPFC, and FGSIR across Models I-1 to I-3.

Model	WIRE	NFPFC	FGSIR
Model I-1	0.325 (0.004)	0.813 (0.003)	0.686 (0.003)
Model I-2	0.277 (0.005)	0.538 (0.005)	0.462 (0.004)
Model I-3	0.226 (0.007)	0.186 (0.007)	0.246 (0.005)

Table 8 Comparison of average DCOR values and their standard errors for WIRE, NFPFC, and FGSIR across Models II-1 to II-3.

Model	WIRE	NFPFC	FGSIR
Model II-1	0.442 (0.006)	0.866 (0.002)	0.919 (0.002)
Model II-2	0.019 (0.014)	0.347 (0.006)	0.317 (0.006)
Model II-3	0.358 (0.006)	0.827 (0.004)	0.853 (0.004)

Illustration of the convergence rate and the role of α

The convergence rate presented in Theorem 2 depends on the spectral decay exponent α, which acts as a measure of smoothness or complexity in infinite dimensional functional data. To illustrate empirically how α influences the convergence behavior, we conducted a simulation study using Model II-1. In particular, we generated the functional predictor and response data from a truncated Karhunen–Loève expansion with eigenvalues decaying as $j^{-\alpha}$, where j indexes the eigenvalues, and α takes values in $\{1.2, 2, 3, 4\}$. To assess the performance of our estimator across different values of α, we computed the estimation error as $1-\text{MCMR}$. We repeated this procedure for increasing sample sizes n and plotted the resulting errors. As we can observe from Figure

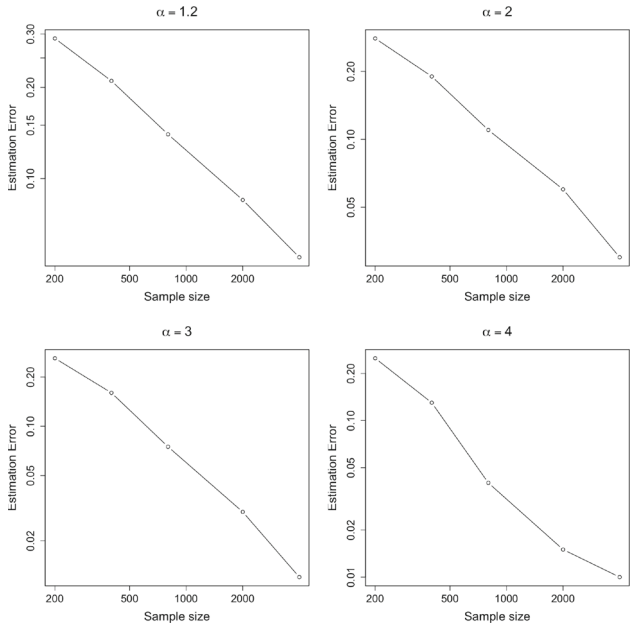

Fig. 7 Plots illustrating estimation error versus sample size for different values of the eigenvalue decay exponent α.

7, larger values of α led to a more rapid decrease in estimation error as the sample size increased, reflecting fewer effectively influential directions within smoother functional data. Conversely, smaller values of α corresponded to slower convergence rates.

Acknowledgements We sincerely thank the Editor, Associate Editor, and referees for their insightful comments and valuable suggestions, which have significantly improved the quality of this paper. For Jae Keun Yoo, this work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education (RS2023-00240564 and RS-2023-0021722). For Kyongwon Kim, this work is supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean government (MSIT) (RS-2023-00219212, RS-2025-00513476).

Author Contributions The manuscript conceptualization, writing, reviewing, and editing were equally shared by M.K, Y.P, K.K, and J.K.Y.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no potential conflict of interest.

References

Billingsley, P.: Probability and Measure. John Wiley & Sons, New Jersey (2012)

Cook, R.D., Forzani, L.: Principal fitted components for dimension reduction in regression. Statistical Science **23**(4), 485–501 (2008). https://doi.org/10.1214/08-STS275

Cook, R.D.: Fisher lecture: Dimension reduction in regression. Statistical Science **22**(1), 1–26 (2007). https://doi.org/10.1214/088342306000000682

Cook, R.D., Weisberg, S.: Sliced inverse regression for dimension reduction: Comment. Journal of the American Statistical Association **86**(414), 328–332 (1991). https://doi.org/10.2307/2290564

Delaigle, A., Hall, P.: Defining probability density for a distribution of random functions. The Annals of Statistics **38**(2), 1171–1193 (2010)

Dai, X., Müller, H.-G., Yao, F.: Optimal bayes classifiers for functional data and density ratios. Biometrika **104**(3), 545–560 (2017)

Fukumizu, K., Bach, F.R., Gretton, A.: Statistical consistency of kernel canonical correlation analysis. Journal of Machine Learning Research **8**(14), 361–383 (2007)

Ferré, L., Yao, A.-F.: Functional sliced inverse regression analysis. Statistics **37**(6), 475–488 (2003). https://doi.org/10.1080/0233188031000112845

Ferré, L., Yao, A.-F.: Smoothed functional inverse regression. Statistica Sinica **15**(3), 665–683 (2005)

Hastie, T., Buja, A., Tibshirani, R.: Penalized discriminant analysis. The Annals of Statistics **23**(1), 73–102 (1995). https://doi.org/10.1214/aos/1176324456

Li, B., Artemiou, A., Li, L.: Principal support vector machines for linear and nonlinear sufficient dimension reduction. Annals of Statistics **39**(6), 3182–3210 (2011)

Liang, B., Gao, T., Bai, D., Wang, G.: Functional dimension reduction based on fuzzy partition and transformation. Australian & New Zealand Journal of Statistics **64**(1), 45–66 (2022)

Li, K.-C.: Sliced inverse regression for dimension reduction. Journal of the American Statistical Association **86**(414), 316–327 (1991). https://doi.org/10.1080/01621459.1991.10475035

Li, B.: Sufficient Dimension Reduction: Methods and Applications with R. Chapman and Hall/CRC, New York (2018)

Lee, K.-Y., Li, B., Chiaromonte, F.: A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. The Annals of Statistics **41**(1), 221–249 (2013). https://doi.org/10.1214/12-AOS1071

Li, B., Song, J.: Nonlinear sufficient dimension reduction for functional data. The Annals of Statistics **45**(3), 1059–1095 (2017). https://doi.org/10.1214/16-AOS1475

Li, B., Song, J.: Dimension reduction for functional data based on weak conditional moments. The Annals of Statistics **50**(1), 107–128 (2022). https://doi.org/10.1214/21-AOS2091

Li, B., Wang, S.: On directional regression for dimension reduction. Journal of the American Statistical Association **102**(479), 997–1008 (2007). https://doi.org/10.1198/016214507000000536

Li, B., Zha, H., Chiaromonte, F.: Contour regression: A general approach to dimension reduction. The Annals of Statistics **33**(4), 1580–1616 (2005). https://doi.org/10.1214/009053605000000192

Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, New York (2005)

Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G.R.: Hilbert space embeddings and metrics on probability measures. The Journal of Machine Learning Research **11**, 1517–1561 (2010)

Song, J., Kim, K., Yoo, J.K.: On a nonlinear extension of the principal fitted component model. Computational Statistics & Data Analysis **182**, 107707 (2023). https://doi.org/10.1016/j.csda.2023.107707

Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. The Annals of Statistics **35**(6), 2769–2794 (2007)

Tan, R., Zang, Y., Yin, G.: Nonlinear dimension reduction for functional data with application to clustering. Statistica Sinica **34**, 1391–1412 (2024)

Wahba, G.: Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia (1990)
Weidmann, J.: Linear Operators in Hilbert Spaces, vol. 68. Springer, New York (1980)
Wang, G., Zhou, Y., Feng, X.-N., Zhang, B.: The hybrid method of fsir and fsave for functional effective dimension reduction. Computational Statistics & Data Analysis **91**, 64–77 (2015)