Functional sufficient dimension reduction through distance covariance

Xing Yang\textsuperscript{a}, Jianjun Xu\textsuperscript{b,*}

\textsuperscript{a}Department of Mathematics, Northeastern University, Boston, MA 02115, USA; \textsuperscript{b}International Institute of Finance, School of Management, University of Science and Technology of China, Hefei, 230026, Anhui, China.

**ARTICLE HISTORY**
Compiled September 26, 2023

**ABSTRACT**
Our research proposes a novel method for reducing the dimensionality of functional data, specifically for the case where the response is a scalar and the predictor is a random function. Our method utilizes distance covariance, and has several advantages over existing methods. Unlike current techniques which require restrictive assumptions such as linear conditional mean and constant covariance, our method has mild requirements on the predictor. Additionally, our method does not involve the use of the unbounded inverse of the covariance operator. The link function between the response and predictor can be arbitrary, and our proposed method maintains the advantage of being model-free, without the need to estimate the link function. Furthermore, our method is naturally suited for sparse longitudinal data. We utilize functional principal component analysis with truncation as a regularization mechanism in the development of our method. We provide justification for the validity of our proposed method, and establish statistical consistency of the estimator under certain regularization conditions. To demonstrate the effectiveness of our proposed method, we conduct simulation studies and real data analysis. The results show improved performance compared to existing methods.

**KEYWORDS**
Sufficient dimension reduction; functional data; distance covariance.

1. Introduction

In contemporary data analysis, functional data are prevalent in many applications such as speech recognition, magnetic resonance imaging (MRI), online handwriting recognition and longitudinal data analysis [1]. Under a functional data analysis (FDA) framework, each sample element is considered to be a function. A hot issue is to study how a response variable varies with a random function $X(t)$, where $t$ is an index variable defined on an interval.

Take Tecator data as an example to introduce the problem of functional regression. Each sample of this dataset contains finely chopped pure meat with different moisture, fat and protein contents. Using analytical chemistry to measure fat content is expensive, while infrared analysis is substantially cheaper. The aim of the analysis is predicting the fat content of pieces of meat from a near infrared absorbance spectrum.

\footnote{Corresponding authors: Jianjun Xu. E-mail: xjj1994@ustc.edu.cn}
which is a curve, see Figure 1. This is a typical functional regression problem which has been investigated from both parametric and nonparametric point of views [2, 3]. However, parametric modeling can be restrictive in some applications, while nonparametric modeling can be unworkable due to the infinite-dimensionality of the functional data. For instance, most functional regression and correlation measure problems involve the inverse of compact operators which are unbounded. This is triggered by the infinite-dimensionality of the functional data, therefore, dimension reduction is key for functional data modeling and analysis.

Despite the infinite-dimensional nature of functional data, interestingly, the data set tends to have a certain pattern which might be represented by finite indexes. With the rapid development of functional data analysis, functional sufficient dimension reduction (FSDR) problems have received increasing attention in the literature. [4] first extended sliced inverse regression (SIR, 5) to the functional case and assumed

\[ Y = g(\langle \beta_1, X \rangle, \ldots, \langle \beta_K, X \rangle, \varepsilon), \]

(1)

where the response \( Y \) is a random variable, the predictor \( X \) takes values in a functional space \( \mathcal{H} \) with inner product denoted by \( \langle \cdot, \cdot \rangle \). \( \beta_1, \ldots, \beta_K \) are \( K \) linearly independent functions in \( \mathcal{H} \), \( \varepsilon \) is a scalar random noise, and \( g \) is an unknown function from \( \mathbb{R}^{K+1} \) to \( \mathbb{R} \). In the FSDR literature, the subspace spanned by \( \beta_1, \ldots, \beta_K \) is called the functional sufficient dimension reduction subspace. After [4] proposed the functional SIR (FSIR), quite a few different methods have been developed for estimating the SDR space. For instance, functional inverse regression [6], functional contour regression [7], functional K-means inverse regression [8], functional sliced average variance estimation (FSAVE, 9), the hybrid method of FSIR and FSAVE [10], functional cumulative slicing [11], robust FSIR [12] and functional directional regression [13]. [14] consider FSIR and FSAVE via a Tikhonov regularization approach and show that their convergence rates are the same as the minimax rates for functional linear regression. [15] proposed functional generalized SIR and functional generalized SAVE for nonlinear sufficient dimension reduction where both the predictor and the response may be random functions. [16] and [17] developed sufficient dimension reduction methods for function-on-function regression through weak conditional moments and average Fréchet derivatives respectively. [18] proposed a method under reproducing kernel Hilbert space (RKHS)
framework which can be applied to finite or infinite dimensional predictor space in a unified framework. Other researches include but are not limited to localized and regularized versions of FSIR and FSAVE [19] and functional surrogate assisted slicing [20] which are aimed at SDR for binary classification. More information can be referred to the review article [21] and the monograph [22]. However all of the above methods need the linear conditional mean assumption or constant covariance assumption or both and these assumptions are not easy to verify in practice. Most of the above methods need to estimate the inverse of a covariance operator which is unbounded since the covariance operator is defined on an infinite-dimensional space.

Existing literature on functional sufficient dimension reduction is mainly based on the sliced inverse moment methods. However, take FSIR [4] as an example, applying FSIR to sparsely observed longitudinal data is practically infeasible, since choosing a sufficiently large number of slices would result in too few observations in each slice with which to estimate a conditional covariance operator. Therefore, there is very little literature on sufficient dimension reduction for sparse longitudinal data. [23] extended the method of [6] for sparse longitudinal data. Functional cumulative slicing [11] and some other methods are also suitable for longitudinal data. We will show that the proposed method is applicable to both dense functional data and sparse longitudinal data.

In multivariate setting, [24] proposed a method for SDR via distance covariance [25–27]. Inspired by this method, we extend it to the functional context which has not been considered before. The goal of this paper is to develop a class of sufficient dimension reduction techniques for functional data that require no inversion of the covariance operator, using the idea of distance covariance. To the best of our knowledge, this is the first time that distance covariance methodology is extended beyond the usual multivariate regression setting to functional data analysis. An important contribution of this paper is to bridge the gap between the nascent area of dependence measure, functional data analysis, and sufficient dimension reduction.

In this article, following the work of [24], we first use distance covariance for functional sufficient dimension reduction. This method does not require linear conditional mean assumption and constant covariance assumption. It also does not involve the inverse of the covariance operator which is not bounded. Under mild conditions, we prove the validity of the proposed method as a sufficient functional dimension reduction method and we use functional principal components method as a form of regularization to make it feasible to estimate a infinite-dimension function in a finite-dimensional subspace. We also construct the consistency of the proposed estimator. Simulation and real data analysis are conducted to exhibit the superiority of the proposed method.

The rest of the paper is organized as follows. In Section 2, we introduce distance covariance, functional sufficient dimension reduction and propose our method for functional sufficient dimension reduction via distance covariance at the population level. Finite-sample estimation and its statistical consistency are presented in Section 3. Simulations and real data analysis are carried out in Section 4. Section 5 concludes the paper, and all the proofs are deferred to the Appendix.
2. Methodology

2.1. Distance covariance

Distance covariance (DCOV) proposed by [25] is a new measure of dependence between random vectors. The appealing property of distance covariance is that it is zero if and only if the random variables are independent. In this subsection random vectors. The characteristic functions of distance covariance are random in $\mathbb{R}^p$ and $V$ in $\mathbb{R}^q$ are random vectors, where $p$ and $q$ are positive integers. The Euclidean norm of $x$ in $\mathbb{R}^p$ is $|x|_p$. The characteristic functions of $U$, $V$ and $(U, V)$ are denoted by $f_U$, $f_V$ and $f_{UV}$ respectively. Then the DCOV defined in [25] is the nonnegative number $\mathcal{V}(U, V)$ with

$$\mathcal{V}^2(U, V) = \int_{\mathbb{R}^2} |f_{UV}(s, t) - f_U(s)f_V(t)|^2 w(s, t) ds dt,$$

where $|f|^2 = f \cdot \overline{f}$ and $w(s, t)$ is a weight function. If we choose $w(s, t) = (\pi^2 s^2 t^2)^{-1}$, [26] gave an equivalent form of DCOV as

$$\mathcal{V}^2(U, V) = E|U - U'|_p|V - V'|_q + E|U - U'|_pE|V - V'|_q - E|U - U'|_pE|V - V'|_q,$$

where $(U, V)$, $(U', V')$ and $(U'', V'')$ are i.i.d. Here we list several useful properties of DCOV. For random vectors $U \in \mathbb{R}^p$ and $V \in \mathbb{R}^q$ such that $E(|U|_p + |V|_q) < \infty$, the following properties hold:

(i) $\mathcal{V}(U, V) = 0$ if and only if $U$ and $V$ are independent.
(ii) $\mathcal{V}(a_1 + b_1 C_1 U, a_2 + b_2 C_2 V) = \sqrt{b_1b_2}\mathcal{V}(U, V)$ for all constant vectors $a_1 \in \mathbb{R}^p$, $a_2 \in \mathbb{R}^q$, scalars $b_1$, $b_2$ and orthonormal matrix $C_1$, $C_2$ in $\mathbb{R}^{p \times p}$ and $\mathbb{R}^{q \times q}$, respectively.
(iii) If the random vector $(U_1, V_1)$ is independent of the random vector $(U_2, V_2)$, then

$$\mathcal{V}(U_1 + U_2, V_1 + V_2) \leq \mathcal{V}(U_1, V_1) + \mathcal{V}(U_2, V_2).$$

Equality holds if and only if $U_1$ and $V_1$ are both constants, or $U_2$ and $V_2$ are both constants, or $U_1$, $U_2$, $V_1$, $V_2$ are mutually independent.

As mentioned in [24] and [28], property (i) makes it possible that DCOV can be used as a sufficient dimension reduction tool. The properties (ii) and (iii) will be applied in the subsequent text.

2.2. Functional sufficient dimension reduction

The functional sufficient dimension reduction (FSDR) is characterized by conditional independence

$$Y \perp \!\!\!\! X \mid (\langle \beta_1, X \rangle, \ldots, \langle \beta_K, X \rangle),$$

where the response $Y$ is a scalar random variable, the predictor $X$ takes values in a functional space $\mathcal{H}$, $\langle \cdot, \cdot \rangle$ represents the inner product in $\mathcal{H}$, $\beta_1, \ldots, \beta_K$ are $K$ linearly independent vectors in $\mathcal{H}$ and $\perp \!\!\!\!$ indicates independence. The multi-index model (1) is included in (4). Without loss of generality, we consider $\mathcal{H} = L_2([0, 1])$, the space spanned by all the square integrable functions on $[0, 1]$. For any $f, g \in L_2([0, 1])$, the
inner product is defined by \( \langle f, g \rangle = \int_0^1 f(t)g(t)dt \). The FSDR subspace is denoted by \( S = \text{span}\{\beta_1, \ldots, \beta_K\} \). Obviously, FSDR subspace is not unique, so we only consider the smallest FSDR subspace, which can be defined as the intersection of all FSDR subspaces. Following the convention of [20], we call it the functional central subspace, denoted as \( S_{g|x} \). Throughout the article, we assume \( S_{g|x} \) exists, which is unique. Let \( K \) denote the dimension of \( S_{g|x} \). Our primary goal is to identify \( S_{g|x} \) by estimating \( K \) basis functions that span \( S_{g|x} \).

2.3. DCOV as a FSDR method

We assume \( E(X) = \mu_X \) and denote the covariance operator of \( X \) by \( \Sigma_X = E[(X - \mu_X) \otimes (X - \mu_X)] \), where \( \otimes \) is defined as \( (f \otimes g)v = \langle f, v \rangle g \) for any \( f, g, v \in L_2([0, 1]) \). For the sequel, we define another inner product on \( L_2([0, 1]) \). For any positive definite self-adjoint and linear operator \( A \), the inner product \( \langle f, g \rangle_A \), for any \( f, g \in L_2([0, 1]) \), is defined as \( \langle f, g \rangle_A = \langle Af, g \rangle = \int \int f(s)A(s,t)g(t)dsdt \). A vector \( f = (f_1, \ldots, f_d) \) has \( d \) components, where each \( f_i \) is a function in \( L_2([0, 1]) \). We define \( \langle f, f \rangle_A = (\langle f_i, f_j \rangle_{\mathbb{I}_{1 \leq i, j \leq d}}) \) as a \( d \times d \) matrix. For a function \( g \in L_2([0, 1]) \), we define \( \langle f, g \rangle = ((f_1, g), \ldots, (f_d, g))^T \) as a \( d \)-dimensional vector.

Denote \( \beta = (\beta_1, \ldots, \beta_K) \) where each \( \beta_i \) is a function in \( L_2([0, 1]) \). We will show that under mild conditions, a basis of \( S_{g|x} \) can be obtained by solving the following optimization problem:

\[
\max_{\langle \beta, \beta \rangle_{\Sigma_X} = I_K} V^2((\beta, X), Y),
\]

where \( \langle \beta, \beta \rangle_{\Sigma_X} = (\langle \beta_i, \beta_j \rangle_{\Sigma_X})_{1 \leq i, j \leq K} \) is a \( K \times K \) matrix, \( I_K \) is the \( K \times K \) identity matrix and \( \langle \beta, X \rangle = ((\beta_1, X), \ldots, (\beta_K, X))^T \) is a column vector. Here we need a scale constraint \( \langle \beta, \beta \rangle_{\Sigma_X} = I_K \) to make the maximization procedure work. The reason is that \( V^2((c\beta, X), Y) = |c|^2 V^2((\beta, X), Y) \) for any constant \( c \), so we can always get a bigger value of \( V^2((\beta, X), Y) \) by multiplying \( \beta \) by a constant with bigger absolute value.

The following propositions guarantees the validity of DCOV as a tool of functional sufficient dimension reduction. The solution of the optimization problem (5) indeed spans the functional central subspace.

Proposition 2.1. Let \( \eta = (\eta_1, \ldots, \eta_K) \) be a basis of \( S_{g|x} \) with \( \langle \eta, \eta \rangle_{\Sigma_X} = I_K \), \( \beta = (\beta_1, \ldots, \beta_K) \) with \( K_1 \leq K \) and \( \langle \beta, \beta \rangle_{\Sigma_X} = I_{K_1} \). Assume \( \text{span}(\beta) \subseteq \text{span}(\eta) \), then \( V^2((\beta, X), Y) \leq V^2((\eta, X), Y) \). The equality holds if and only if \( \text{span}(\beta) = \text{span}(\eta) \).

Proposition 2.1 means that the DCOV between \( (\beta, X) \) and \( Y \) is always no more than the DCOV between \( (\eta, X) \) and \( Y \) when \( \text{span}(\beta) \) is a subspace of the functional central subspace \( \text{span}(\beta) \subseteq \text{span}(\eta) \). The equality holds if and only if \( \text{span}(\beta) = \text{span}(\eta) \). However, this result is not enough to guarantee DCOV as a tool of functional sufficient dimension reduction. We also need to consider the situation that \( \text{span}(\beta) \not\subseteq \text{span}(\eta) \).

The next proposition gives the result of this situation under a mild condition.

Condition 1. Let \( \eta = (\eta_1, \ldots, \eta_K) \) be a basis of \( S_{g|x} \). Denote \( \text{span}(\eta)^1 \) the orthogonal complement space of \( \text{span}(\eta) \) with respect to the inner product \( \langle \cdot, \cdot \rangle_{\Sigma_X} \). We assume that for any \( f = (f_1, \ldots, f_I) \), \( g = (g_1, \ldots, g_J) \) where \( f_i \in \text{span}(\eta) \) and \( g_j \in \text{span}(\eta)^1 \), \( i = 1, \ldots, I \), \( j = 1, \ldots, J \), we have \( \langle f, X \rangle \perp \langle g, X \rangle \).
This condition is not as strong as it seems to be. When $X$ is a Gaussian process, the independence condition $(f, X) \perp (g, X)$ will be satisfied, because $\text{Cov}((f, X), (g, X)) = 0$. However, Gaussianity is not necessary. Condition 1 asymptotically holds when the dimension of $X$ gets reasonably high, see [24, 29] for details.

**Proposition 2.2.** Let $\eta = (\eta_1, \ldots, \eta_K)$ be a basis of $S_{glx}$ with $\langle \eta, \eta \rangle_{\Sigma_X} = I_K$, $\beta = (\beta_1, \ldots, \beta_K)$ with $\langle \beta, \beta \rangle_{\Sigma_X} = I_{K_2}$. Here $K_2$ could be bigger, less, or equal to $K$. Assume Condition 1 holds and $\text{span}(\beta) \not\subseteq \text{span}(\eta)$, then $\mathcal{V}^2((\beta, X), Y) < \mathcal{V}^2((\eta, X), Y)$.

Proposition 2.2 indicates that if $\text{span}(\beta) \not\subseteq \text{span}(\eta)$, the DCOV between $(\beta, X)$ and $Y$ is always less than the DCOV between $(\eta, X)$ and $Y$. Propositions 2.1 and 2.2 in this article are extensions of propositions 1 and 2 in [24]. Propositions 2.1 and 2.2 guarantee the validity of DCOV as a functional sufficient dimension reduction method, so we can obtain a basis of $S_{glx}$ by solving the optimization problem (5). Note that the optimization problem (5) is on the population level and the infinite dimensional functions $X$ and $\beta$ make the problem more complicated. Take into account of these problems, some form of regularization is needed in the estimation procedure. We will elaborate it in detail in the next section.

3. Estimation

In the previous section, we have established the method of functional SDR via distance covariance at the population level. In this section, we will give algorithms for both completely observed functional data and sparse longitudinal data at the sample level. The structural dimension $K$ is assumed to be known in this section.

3.1. **Estimation of $S_{glx}$ for functional data**

In functional context, the estimation procedure is more intractable than that in multivariate context because of the infinite dimensionality of the predictor $X$ and the parameters $\beta_k$, $k = 1, \ldots, K$. Practically feasible approaches must include some form of dimensionality reduction. A standard method in functional data analysis is to embed the infinite dimensional curves into a finite dimensional space. Specifically, $X$ and $\beta_k$ are approximated using the series expansion method. In this article, we consider functional principal components (FPC) basis, which is a common choice in practice [1, 30–32].

In reality, the predictor trajectories are observed intermittently. For densely observed $X$, individual smoothing can be used as a pre-processing step to recover smooth trajectories, and the error introduced by individual smoothing has been shown to be asymptotically negligible under certain design conditions [31]. Thus, to simplify the notation, we only consider completely observed functional data.

Given i.i.d sample $(X, Y) = \{(X_i(t), Y_i), i = 1, 2, \ldots, n\}$ and let $\bar{X}(t) = 1/n\sum_{i=1}^{n} X_i(t)$, the sample covariance of $X$ can be estimated by $\bar{\Sigma}_X(s, t) = 1/n\sum_{i=1}^{n} (X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t))$. By Mercer’s theorem [33], $\Sigma_X$ and $\bar{\Sigma}_X$ admit the following eigen-decomposition:

$$\Sigma_X(s, t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(s)\phi_i(t), \quad \bar{\Sigma}_X(s, t) = \sum_{i=1}^{\infty} \bar{\lambda}_i \bar{\phi}_i(s)\bar{\phi}_i(t), \quad (6)$$
where \( \{ \lambda_j : j \geq 1 \} \) and \( \{ \tilde{\lambda}_j : j \geq 1 \} \) are the population and empirical eigenvalues, \( \{ \phi_j(t) : j \geq 1 \} \) and \( \{ \tilde{\phi}_j(t) : j \geq 1 \} \) are the corresponding eigenfunctions, or functional principal components, each forming an orthonormal basis of \( L^2([0,1]) \). Then, \( X_i \) and \( \tilde{\beta}_k \) can be expanded as

\[
X_i(t) = \mu_X(t) + \sum_{j=1}^{\infty} \theta_{ij} \tilde{\phi}_j(t), \quad \beta_k(t) = \sum_{j=1}^{\infty} b_{kj} \tilde{\phi}_j(t),
\]

where \( \theta_{ij} = \langle X_i - \mu_X, \tilde{\phi}_j \rangle \) and \( b_{kj} = \langle \beta_k, \tilde{\phi}_j \rangle \). We approximate \( X_i \) and \( \beta_k \) by

\[
X_i^D(t) = \mu_X(t) + \sum_{j=1}^{D} \theta_{ij} \tilde{\phi}_j(t) = \mu_X(t) + \theta_i^T \Phi_D(t),
\]

\[
\beta_k^D(t) = \sum_{j=1}^{D} b_{kj} \tilde{\phi}_j(t) = b_k^T \Phi_D(t),
\]

for some suitably large \( D \), where \( \theta_i = (\theta_{i1}, \ldots, \theta_{iD})^T \), \( b_k = (b_{k1}, \ldots, b_{kD})^T \), \( \Phi_D(t) = (\tilde{\phi}_1(t), \ldots, \tilde{\phi}_D(t))^T \).

According to [24], the sample version of \( \mathcal{V}_n^2(\langle \beta, X \rangle, Y) \) denoted by \( \mathcal{V}_n^2(\langle \beta, X \rangle, Y) \) has the following form:

\[
\mathcal{V}_n^2(\langle \beta, X \rangle, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}(\beta) B_{ij},
\]

where

\[
A_{ij}(\beta) = a_{ij}(\beta) - \bar{\pi}_i(\beta) - \bar{\pi}_j(\beta) + \bar{\pi}_.(\beta),
\]

\[
a_{ij}(\beta) = |\langle \beta, X_i \rangle - \langle \beta, X_j \rangle|_K, \quad \bar{\pi}_i(\beta) = \frac{1}{n} \sum_{j=1}^{n} a_{ij}(\beta),
\]

\[
\bar{\pi}_j(\beta) = \frac{1}{n} \sum_{i=1}^{n} a_{ij}(\beta), \quad \bar{\pi}_.(\beta) = \frac{1}{n^2} \sum_{i,j=1}^{n} a_{ij}(\beta),
\]

and \( B_{ij} = b_{ij} - \bar{b}_i - \bar{b}_j + \bar{b}_. \), \( b_{ij} = |Y_i - Y_j| \), and the definition of \( \bar{b}_i, \bar{b}_j, \bar{b}_. \) is similar to those of \( \bar{\pi}_i(\beta), \bar{\pi}_j(\beta), \bar{\pi}_.(\beta) \). Then, an estimate of a basis of \( S_{y|x} \), say \( \tilde{\eta}_{n,D} \) is obtained by solving the following optimization problem:

\[
\tilde{\eta}_{n,D} = \arg\max_{(\beta^D, \beta^D)} \mathcal{V}_n^2(\langle \beta^D, X^D \rangle, Y),
\]

where \( \beta^D = (\beta^D_1(t), \ldots, \beta^D_K(t)) \) and \( X^D = (X^D_1, \ldots, X^D_n)^T \). We just need to replace \( \langle \beta, X_i \rangle \) and \( \langle \beta, X_j \rangle \) in (10) with \( \langle \beta^D, X^D_i \rangle \) and \( \langle \beta^D, X^D_j \rangle \) to calculate \( \mathcal{V}_n^2(\langle \beta^D, X^D \rangle, Y) \). Particularly, let \( B = (b_1, \ldots, b_K) \), then \( \langle \beta^D, X^D_i \rangle = B^T \theta_i \) and

\[
\langle \beta^D, \beta^D \rangle_{\Sigma_{\beta}} = B^T \Sigma_{\theta} B \text{ where } \Sigma_{\theta} = 1/n \sum_{i=1}^{n} (\theta_i - \bar{\theta})(\theta_i - \bar{\theta})^T, \bar{\theta} = 1/n \sum_{i=1}^{n} \theta_i. \]

Thus, the optimization problem (11) is equivalent to

\[
\tilde{B}_{n,D} = \arg\max_{B^T \Sigma_{\theta} B = I_K} \mathcal{V}_n^2(\langle B^T \theta, Y \rangle, \bar{

7
where \( \theta = (\theta_1, \ldots, \theta_K) \) and \( \hat{B}_{n,D} \) is a \( D \times K \) matrix. Similarly, we can replace \( \langle \beta, X_i \rangle \) and \( \langle \beta, X_j \rangle \) in (10) with \( B^T \theta_i \) and \( B^T \theta_j \) to calculate \( V_{n}^2(B^T \theta, Y) \). Although the optimization problem (5) with respect to \( \beta_k \) is taken over an infinite dimensional space, the solution can actually be found in a finite dimensional subspace by regularization. It suffices to estimate the coefficients matrix \( B \) in (12).

Note that it is complicated to solve the optimization problem (12) over a \( D \times K \) matrix. A projection pursuit type of sufficient searching algorithm [34] is adopted to break down the problem (12) into successive single-index searching. The algorithm can be described as follows:

1. Solve the single-index searching problem \( \hat{\gamma}_1 = \arg \max_{B^T \Sigma_{a} B = 1} V_{n}^2(B^T \theta, Y) \), where \( B \) is a \( D \times 1 \) vector. \( \hat{\gamma}_1 \) is the first intermediate direction.
2. Construct \( D \times (D-1) \) matrix \( \Gamma_1 \) such that \( \Sigma_{\theta}^{1/2}(\hat{\gamma}_1, \Gamma_1) \) is an orthogonal matrix.
3. Let \( a \in \mathbb{R}^{D-1} \) and consider the predictor matrix \( (\hat{\gamma}_1, \Gamma_1 a) \), where \( \hat{\gamma}_1 \) is fixed. Solve the problem

\[
\hat{a}_1 = \arg \max_{(\hat{\gamma}_1, \Gamma_1 a) \in \Sigma_{a}(\hat{\gamma}_1, \Gamma_1 a) = I_2} \left\{ V_{n}^2((\hat{\gamma}_1, \Gamma_1 a)^T \theta, Y) : a \in \mathbb{R}^{D-1} \right\},
\]

then the second intermediate direction is \( \hat{\gamma}_2 = \Gamma_1 \hat{a}_1 \).
4. Let the \( D \times 1 \) vectors \( \hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_k \) be the first \( k \) intermediate directions, and let \( \Sigma_{\theta}^{1/2}(\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_k, \Gamma_k) \) form an orthogonal matrix. Then we search for a \( (D-k) \times 1 \) vector \( a_k \) based on the predictor matrix \( (\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_k, \Gamma_k a_k) \). Then the \((k+1)\)-th intermediate direction is \( \hat{\gamma}_{k+1} = \Gamma_k a_k \).
5. The estimate for the coefficients matrix \( B \) in (12) is \( \hat{B}_{n,D} = (\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_K) \).

Finally, the estimate of \( S_{y|x} \) is

\[
\hat{\eta}_{n,D} = \Phi_D^T(t) \hat{B}_{n,D}.
\] (13)

For the asymptotic analysis, it is typical to allow \( D \) to diverge with the increase of \( n \). We add a subscript \( n \) to \( D \) to emphasize this relationship. The following theorem states the consistency of the estimator \( \hat{\eta}_{n,D_n} \).

**Theorem 3.1.** Assume \( \eta = (\eta_1, \ldots, \eta_K) \) is a basis of \( S_{y|x} \) with \( \langle \eta, \eta \rangle \Sigma_X = I_K \). For \( \hat{\eta}_{n,D_n} \) defined in (11), as \( n \to \infty \), \( D_n \to \infty \) we have \( \hat{\eta}_{n,D_n} \xrightarrow{P} \eta \) provided that Condition 1 holds.

Theorem 3.1 establishes the consistency of the estimator \( \hat{\eta}_{n,D_n} \) of \( \eta \). Here, we denote \( \hat{\eta}_{n,D_n} = (\hat{\eta}_{n,D_n,1}, \ldots, \hat{\eta}_{n,D_n,K}) \), and the expression \( \hat{\eta}_{n,D_n} \xrightarrow{P} \eta \) means \( \| \hat{\eta}_{n,D_n} - \eta \| = \left( \sum_{k=1}^{K} \| \hat{\eta}_{n,D_n,k} - \eta_k \|^2 \right)^{1/2} \xrightarrow{P} 0 \).

---

3.2. Estimation for sparse longitudinal data

The focus of this subsection is to estimate \( S_{y|x} \) for intermittently and sparsely measured longitudinal covariates. When only a few observations are available for some or even all subjects, individual smoothing to recover \( X_i \) is infeasible and one must pool data across subjects for consistent estimation. For the i.i.d sample \( (X, Y) = \{(X_i(t), Y_i), i = 1, 2, \ldots, n \} \), the predictors \( X_i \) are observed intermittently, contaminated with noise,
and observed in the form of \( \{(T_{ij}, U_{ij}) : i = 1, \ldots, n; j = 1, \ldots, N_i\} \) where
\[
U_{ij} = X_i(T_{ij}) + \varepsilon_{ij}.
\]
The i.i.d. measurement error \( \varepsilon_{ij} \) satisfies \( E(\varepsilon_{ij}) = 0 \) and \( \text{var}(\varepsilon_{ij}) = \sigma^2 \). The numbers of observations \( \{N_i\}_{i=1}^n \) are assumed to be random, reflecting sparse and irregular designs. The observation time points \( \{T_{ij}\} \) are assumed to be i.i.d. realizations of a random variable and independent of all other random variables. Another assumption is that the pooled time points \( \{T_{ij}\} \) are sufficiently dense in the domain of \( X(t) \).

Firstly, we estimate the mean function \( \mu_X(t) \) based on the pooled data. Following [30], local linear smoothing is conducted for estimating \( \mu_X(t) \) by minimizing
\[
\sum_{i=1}^{n} \sum_{j=1}^{N_i} K_i \left( \frac{T_{ij} - t}{h_{\mu}} \right) \{U_{ij} - a_0 - a_1(t - T_{ij})\}^2,
\]
with respect to \( a_0 \) and \( a_1 \), where \( K_i \) is a univariate kernel function and \( h_{\mu} \) is the bandwidth. Then the estimate of \( \mu_X(t) \) is \( \hat{\mu}_X(t) = \hat{a}_0 \). For the covariance function \( \Sigma_X(s, t) \), [30] defined the observed raw covariance by \( G_i(T_{ij}, T_{il}) = (U_{ij} - \hat{\mu}_X(T_{ij}))(U_{il} - \hat{\mu}_X(T_{il})) \).

Solving the local linear surface smoothing problem
\[
\min_{(b_0,b_1,b_2)} \sum_{i=1}^{n} \sum_{j=1}^{N_i} K_2 \left( \frac{T_{ij} - s}{h_{\Sigma}}, \frac{T_{il} - t}{h_{\Sigma}} \right) \{G_i(T_{ij}, T_{il}) - b_0 - b_1(T_{ij} - s) - b_2(T_{il} - t)\}^2,
\]
yields \( \hat{\Sigma}_X(s, t) = \hat{\Sigma}_0 \), where \( K_2 \) is a bivariate kernel function with bandwidth \( h_{\Sigma} \). Then, similar to the previous subsection, \( \hat{\lambda}_j \) and \( \hat{\phi}_j \) can be obtained from the eigen decomposition of \( \hat{\Sigma}_X(s, t) \).

From the optimization problem (12), we know that the only quantity we need to acquire is the FPC scores \( \theta_{ij} = \langle X_i - \mu_X, \phi_j \rangle \). For sparse longitudinal data, individual smoothing to recover \( X_i \) is infeasible, thus numerical integration for calculating \( \theta_{ij} \) will not provide reasonable approximations to the real FPC scores. We adopt the efficient Principal Analysis by Conditional Expectation (PACE) [30] method specifically designed for sparse longitudinal data to estimate FPC scores. Denote \( \hat{X}_i = (X_i(T_{i1}), \ldots, X_i(T_{iN_i}))^T, \hat{U}_i = (U_{i1}, \ldots, U_{iN_i})^T, \mu_i = (\mu_X(T_{i1}), \ldots, \mu_X(T_{iN_i}))^T, \) and \( \phi_i = (\phi_1(T_{i1}), \ldots, \phi_j(T_{iN_i}))^T \). Then \( \theta_{ij} \) and \( \varepsilon_{ij} \) are jointly Gaussian, the best prediction of the FPC score \( \theta_{ij} \) given \( i \)-th subject is the conditional expectation
\[
\bar{\theta}_{ij} = E(\theta_{ij}|\hat{U}_i) = \lambda_j \phi_j^T \Sigma_{U_i}^{-1}(\hat{U}_i - \mu_i),
\]
where \( \Sigma_{U_i} = \text{cov}(\hat{U}_i, \hat{U}_i) = \text{cov}(\bar{X}_i, \bar{X}_i) + \sigma^2 I_{N_i} \) and the \( N_i \times N_i \) matrix \( \text{cov}(\bar{X}_i, \bar{X}_i) = (\Sigma_X(T_{ij}, T_{il}))_{1 \leq j, l \leq N_i} \). By substituting estimates of \( \mu_i, \lambda_j, \phi_{ij} \) and \( \Sigma_{U_i} \) obtained from the pooled data, we have an estimate of \( \theta_{ij} \),
\[
\hat{\theta}_{ij} = \hat{E}(\theta_{ij}|\hat{U}_i) = \hat{\lambda}_j \hat{\phi}_{ij}^T \hat{\Sigma}_{U_i}^{-1}(\hat{U}_i - \hat{\mu}_i),
\]
where \( \hat{\Sigma}_{U_i} = \text{cov}(\bar{X}_i, \bar{X}_i) + \hat{\sigma}^2 I_{N_i}, \text{cov}(\bar{X}_i, \bar{X}_i) = (\hat{\Sigma}_X(T_{ij}, T_{il}))_{1 \leq j, l \leq N_i} \), and \( \hat{\sigma}^2 \) is an estimate of \( \sigma^2 \) which can be found in [30]. In this paper, we do not introduce the specific estimate of \( \sigma^2 \) for the sake of brevity. [30] has shown that under some regularization
conditions, $\tilde{\theta}_{ij} \xrightarrow{p} \bar{\theta}_{ij}$, which means that $\tilde{\theta}_{ij}$ is a good estimate of $\theta_{ij}$. We substitute \{\tilde{\theta}_{ij}\} into the optimization problem (12) to get the coefficients matrix $\tilde{B}_{n,D}$. The optimization steps are exactly the same as those of completely observed case.

### 3.3. Selection of tuning parameters

Computation of the proposed method relies on the choice of two parameters: the truncation number $D$ and the structural dimension $K$. Determining $D$ and $K$ can be tackled in different ways and it depends on the goal of the analysis. If the purpose is prediction, $D$ and $K$ can be treated as parameters of the whole model and adjusted according to the performance of the prediction, such as cross-validation. This has been successfully experimented with in applications [6, 11, 18]. Consider model (1):

$$Y = g(\langle \beta_1, X \rangle, \ldots, \langle \beta_K, X \rangle, \varepsilon).$$

The data $(X, Y)$ were randomly divided into $k$ equal portions $\{(X^{(1)}, Y^{(1)}), \ldots, (X^{(k)}, Y^{(k)})\}$. For each feasible $D$, $K$ and $i = 1, \ldots, k$, we leave out $(X^{(i)}, Y^{(i)})$ and use the rest of the data $(X^{(-i)}, Y^{(-i)})$ to compute the $\tilde{\eta}_{n,D}$ in (13) and nonparametrically estimate $g$. Then, use the $\tilde{\eta}_{n,D}$, the estimate $g$, and $X^{(i)}$ to compute predicted values $\tilde{Y}^{(i)}_{D,K}$. Let $\text{CV}(D, K) = 1/k \cdot \sum_{l=1}^{k} \left\| Y^{(i)} - \tilde{Y}^{(i)}_{D,K} \right\|_2^2$, where $\| \cdot \|_2$ is the Euclidean norm, and pick $D$ and $K$ to minimize $\text{CV}(D, K)$. However, the cross-validation procedures are not ideal since the nonparametric fitting adds an extra layer of complication. A more satisfactory of the selection of $D$ and $K$ is currently not available. We will apply this cross-validation method in the real data analysis.

When FSDR is used in a descriptive way, we are mainly interested in recovering the directions per se. Much of the existing literature, such as [8, 14, 18], suggested that $D$ can be chosen subjectively. Specifically, the number $D$ of included eigenfunctions is chosen by fraction of variance explained criterion in practice,

$$D = \min \left\{ k : \sum_{l=1}^{k} \lambda_l / \sum_{l=1}^{n} \lambda_l \geq R \right\},$$

with a given threshold $R$ close to 1 and the eigenvalues $\lambda_l$, $1 \leq l \leq k$ are not “too small”. We can also draw a scree plot and choose the “elbow” point as the truncation number. In the simulations in this paper, we recommend $R = 95\%$, which includes 5 eigenfunctions, and this choice of $D$ yields satisfactory results.

We then consider the selection of $K$ when $D$ is known. Similar as in the multivariate case, the selection of $K$ relies on a criterion measuring the quality of the estimation of $S_{y|x}$. A bootstrap method suggested by [24] and [28] can be readily extended to our method and we apply it in the real data analysis. Specifically, the bootstrap method is based on the measure of distance of two functional spaces [11]:

$$\Delta_m(S_1, S_2) = \| P_{S_1} - P_{S_2} \|_H,$$

where $S_i$ is spanned by $\{\beta_1, \ldots, \beta_{d_i}\}$, $P_{S_i} = \sum_{j=1}^{d_i} \beta_j \otimes \beta_j$ is the projection operator onto $S_i$ for $i = 1, 2$, and $\| A \|^2_H = \int \int A^2(s, t) ds dt$ for a linear operator $A \in L_2([0, 1] \times [0, 1])$. Obviously, the smaller the distance is, the closer the two spaces are.
In order to use this measure, we will treat \((\theta, Y) = \{(\theta_i, Y_i), i = 1, \ldots, n\}\) as the sample. For each possible working dimensions \(1 \leq k \leq D - 1\), we solve the problem (12) to obtain an estimated coefficient matrix \(\hat{B}_k\) whose columns spans a subspace in \(\mathbb{R}^D\) and then we get the estimator of \(S_{\hat{g}_{1k}}\), \(\hat{B}_k = \Phi^T(t)\hat{B}_k\). Here we omit the subscript \(n\) and \(D\) to emphasize the status of \(k\). Then we randomly sample the data \((\theta, Y)\) with replacement \(B\) times, and obtain the estimated subspace based on the bootstrap samples, and denote them by \(\hat{B}^b_k\), \(b = 1, \ldots, B\). We calculate \(\Delta_m(\hat{B}_k, \hat{B}_k^b)\) for \(b = 1, \ldots, B\) and use the mean \(1/B \cdot \sum_{b=1}^B \Delta_m(\hat{B}_k, \hat{B}_k^b)\) as the measure of variability for each \(k\). We choose the \(k\) corresponding to the smallest variability as our estimated \(K\). The reason why this bootstrap method works well is mentioned in [24].

It is worth noting that directly using the bootstrap method is time consuming. However, the bootstrap method can easily be modified to a parallel version to significantly reduce the computation time. We use the \(R\) package parallel for parallel computing.

4. Numerical studies

4.1. Simulations

In this subsection, we conduct simulation studies to provide an insight in the empirical performance of the proposed method (FDCOV) and compare it with some existing methods. For the case of completely observed functional data, the competitors includes functional sliced inverse regression (FSIR, 4), functional sliced average variance estimation (FSAVE, 9), functional contour regression (FCR, 7) and functional directional slicing (FCS, 11).

In the simulation studies, the following five models are considered:

1. \(Y = \exp(\langle \eta_1, X \rangle) + \varepsilon\), \(\eta_1(t) = \sin(3\pi t/2)\),
2. \(Y = \exp(\langle \eta_1, X \rangle) + \exp(\langle \varepsilon, \eta_2, X \rangle) + \varepsilon\), \(\eta_1(t) = \sin(3\pi t/2)\), \(\eta_2(t) = \sin(5\pi t/2)\),
3. \(Y = \exp(\langle \eta_1, X \rangle) + \exp(\langle \varepsilon, \eta_2, X \rangle) + \varepsilon\), \(\eta_1(t) = (2t - 1)^3 + 1\), \(\eta_2(t) = \cos((2t - 1)\pi) + 1\),
4. \(Y = 5\langle \eta_1, X \rangle + 15\langle \eta_2, X \rangle^2 + \varepsilon\), \(\eta_1(t) = \sin(3\pi t/2)\), \(\eta_2(t) = \sin(5\pi t/2)\),
5. \(Y = 50\langle \eta_1, X \rangle \langle \eta_2, X \rangle^2 + \varepsilon\), \(\eta_1(t) = (2t - 1)^2 - 1\), \(\eta_2(t) = \sin(5\pi t/2)\),

Figure 2. Left panel: \(\eta_1\) (red smooth curve) and \(\hat{\eta}_1\) (black dots) versus \(t\). Middle panel: \(\langle \hat{\eta}_1, X \rangle\) versus \(\langle \eta_1, X \rangle\). Right panel: \(Y\) versus \(\langle \hat{\eta}_1, X \rangle\) and the true link function \(y = e^x\) (red smooth curve).
Table 1. Estimation error $\|\hat{P} - P\|_H$ for different estimators for Model 1. Numbers in parentheses are standard errors calculated from 100 generated data sets.

| n   | L  | r    | FDCOV | FSIR  | FSAVE | FDR  | FCR  |
|-----|----|------|-------|-------|-------|------|------|
| 100 | 5  | 0.05 | 0.121(0.047) | 0.255(0.117) | 0.287(0.128) | 0.253(0.129) | 0.174(0.076) |
| 10  | 0.10 | 0.121(0.047) | 0.190(0.077) | 0.379(0.243) | 0.296(0.127) | 0.168(0.073) |
| 15  | 0.15 | 0.121(0.047) | 0.169(0.072) | 1.127(0.400) | 0.359(0.162) | 0.160(0.069) |
| 20  | 0.20 | 0.121(0.047) | 0.177(0.077) | 1.299(0.248) | 0.415(0.162) | 0.163(0.073) |
| 200 | 5   | 0.05 | 0.085(0.036) | 0.165(0.060) | 0.166(0.066) | 0.164(0.071) | 0.120(0.050) |
| 10  | 0.10 | 0.085(0.036) | 0.130(0.062) | 0.154(0.064) | 0.194(0.094) | 0.115(0.053) |
| 15  | 0.15 | 0.085(0.036) | 0.114(0.054) | 0.195(0.149) | 0.201(0.084) | 0.113(0.050) |
| 20  | 0.20 | 0.085(0.036) | 0.123(0.050) | 0.239(0.190) | 0.232(0.114) | 0.116(0.048) |

Table 2. Estimation error $\|\hat{P} - P\|_H$ for different estimators for Model 2. Numbers in parentheses are standard errors calculated from 100 generated data sets.

| n   | L  | r    | FDCOV | FSIR  | FSAVE | FDR  | FCR  |
|-----|----|------|-------|-------|-------|------|------|
| 100 | 5  | 0.05 | 0.458(0.112) | 1.414(0.273) | 1.208(0.297) | 0.970(0.255) | 1.069(0.312) |
| 10  | 0.10 | 0.458(0.112) | 1.444(0.277) | 1.359(0.216) | 1.044(0.263) | 0.992(0.275) |
| 15  | 0.15 | 0.458(0.112) | 1.482(0.278) | 1.453(0.180) | 1.334(0.276) | 0.921(0.246) |
| 20  | 0.20 | 0.458(0.112) | 1.512(0.297) | 1.497(0.209) | 1.288(0.240) | 0.916(0.264) |
| 200 | 5   | 0.05 | 0.187(0.075) | 1.369(0.255) | 0.874(0.314) | 0.711(0.270) | 0.756(0.181) |
| 10  | 0.10 | 0.187(0.075) | 1.294(0.298) | 1.059(0.304) | 0.693(0.231) | 0.724(0.182) |
| 15  | 0.15 | 0.187(0.075) | 1.316(0.305) | 1.184(0.305) | 0.793(0.242) | 0.718(0.185) |
| 20  | 0.20 | 0.187(0.075) | 1.392(0.270) | 1.341(0.213) | 0.870(0.267) | 0.687(0.194) |

where $\varepsilon \sim N(0,0.1^2)$ and $X$ is the standard Brownian motion on $[0,1]$, independent of $\varepsilon$. These examples cover various situations. Model 1 is a single index model from [18]. Model 2 and Model 3 are taken from [4]. In Model 2, both $\eta_1$ and $\eta_2$ are eigenvectors of the Brownian motion, while both $\eta_1$ and $\eta_2$ in Model 3 are not eigenvectors of the Brownian motion. Model 4 was used in [14], considering heterogeneous errors. Model 5 was previously considered in [9] and [14], where $\eta_1$ is not an eigenvector of the Brownian motion and the link function is not additive.

For the considered methods for completely observed data, we need to decide the following tuning parameters. For FSIR, FSAVE and FDR: the number of slices $L$; for FCR: the proportion $r$ of empirical directions. We consider $L = 5, 10, 15, 20, r = 0.05, 0.10, 0.15, 0.20$ and $n = 100, 200$. In each setting, we simulate 100 data sets and each random curve is sampled at $p = 100$ equally spaced points in $[0,1]$, $\{t_1, \ldots, t_{100}\}$ with $t_1 = 0$ and $t_{100} = 1$. Similar to the setting in [13], all these methods are implemented using functional principal component analysis with truncation $D$ chosen such that $95\%$ of variability in the predictor are retained. That is

$$D = \min \left\{ k : \left( \sum_{i=1}^{k} \hat{\lambda}_i \right) / \left( \sum_{i=1}^{n} \hat{\lambda}_i \right) \geq 95\% \right\}.$$ (19)

We assume that the structural dimension $K$ is known. Let $P = \sum_{k=1}^{K} \eta_k \otimes \eta_k$ and $\hat{P} = \sum_{k=1}^{K} \eta_{n,D,k} \otimes \eta_{n,D,k}$ be the projection operators onto the true $S_{y|x}$ and estimated $S_{y|x}$ respectively. We calculate $|\hat{P} - P|_H$ as the estimation error with smaller values.
indicating better estimation performance. All the simulation results for completely observed data are reported in Figure 2 and Tables 1-5.

For Model 1, some results of our proposed method are displayed in Figure 2. The left panel of Figure 2 gives the plots of $\eta_1$ and $\hat{\eta}_1$. The red smooth curve is the true direction $\eta_1 = \sin(3\pi t/2)$ and the black dots are corresponding estimator $\hat{\eta}_1$ at 100 equally spaced time points in $[0, 1]$. We see that $\hat{\eta}_1$ coincides almost perfectly with $\eta_1$. In prediction stage, what we care about is that the estimated projection $\langle \hat{\eta}_1, X \rangle$ is as close as possible to the true projection $\langle \eta_1, X \rangle$. We plot in the middle panel of Figure 2 the indexes $\langle \hat{\eta}_1, X \rangle$ versus $\langle \eta_1, X \rangle$. We find that these scatter plots reveal a strong correlation between $\langle \hat{\eta}_1, X \rangle$ and $\langle \eta_1, X \rangle$. We also present the plot for $Y$ versus $\langle \hat{\eta}_1, X \rangle$ along with the true link function $y = e^x$ in the right panel of Figure 2.

Note that the results in Figure 2 are based on one single simulation run. In order to get more representative results, we compare our proposed method with FSIR, FSAVE, FDR and FCR based on 100 Monte Carlo repetitions. Tables 1–5 report the mean and standard errors of $\| \hat{P} - P \|_H$ for Models 1–5. From Tables 1–5, FDCOV has the best performance in all five cases. For Model 1, FSAVE does not work well since it is known that SAVE is not efficient in estimating monotone trends for small to moderate data sets. For Model 2 and Model 3, it is not surprising that the absolute value of $\| \hat{P} - P \|_H$ of Model 2 is greater than that of Model 3 since Model 2 corresponding the ideal situation where the true direction is included into the a priori projection subspace. From Tables 2 and 4, we see that FDCOV can identify $S_{y|x}$ in heteroscedastic models, but it is not as efficient as in homoscedastic models. For all cases, the results become better as $n$ increases and the results are generally not very sensitive to the choice of

Table 3. Estimation error $\| \hat{P} - P \|_H$ for different estimators for Model 3. Numbers in parentheses are standard errors calculated from 100 generated data sets.

| $n$ | $L$ | $r$ | FDCOV  | FSIR  | FSAVE  | FDR   | FCR   |
|-----|-----|-----|--------|-------|--------|-------|-------|
| 100 | 5   | 0.05| 1.733(0.028) | 2.712(0.102) | 2.288(0.251) | 2.274(0.265) | 2.101(0.228) |
| 20  | 0.10| 1.733(0.028) | 2.698(0.128) | 2.444(0.255) | 2.396(0.269) | 2.122(0.229) |
| 15  | 0.15| 1.733(0.028) | 2.732(0.120) | 2.499(0.236) | 2.481(0.275) | 2.121(0.206) |
| 20  | 0.20| 1.733(0.028) | 2.713(0.123) | 2.568(0.272) | 2.497(0.228) | 2.180(0.225) |
| 200 | 5   | 0.05| 1.724(0.017) | 2.710(0.094) | 2.126(0.217) | 2.107(0.205) | 1.939(0.147) |
| 10  | 0.10| 1.724(0.017) | 2.678(0.131) | 2.193(0.235) | 2.166(0.227) | 1.974(0.156) |
| 15  | 0.15| 1.724(0.017) | 2.683(0.125) | 2.317(0.236) | 2.162(0.206) | 2.004(0.190) |
| 20  | 0.20| 1.724(0.017) | 2.692(0.123) | 2.268(0.255) | 2.219(0.228) | 2.020(0.164) |

Table 4. Estimation error $\| \hat{P} - P \|_H$ for different estimators for Model 4. Numbers in parentheses are standard errors calculated from 100 generated data sets.

| $n$ | $L$ | $r$ | FDCOV  | FSIR  | FSAVE  | FDR   | FCR   |
|-----|-----|-----|--------|-------|--------|-------|-------|
| 100 | 5   | 0.05| 0.606(0.124) | 1.464(0.294) | 1.214(0.282) | 1.049(0.268) | 0.884(0.239) |
| 10  | 0.10| 0.606(0.124) | 1.505(0.258) | 1.302(0.305) | 1.081(0.317) | 0.806(0.236) |
| 15  | 0.15| 0.606(0.124) | 1.493(0.316) | 1.454(0.247) | 1.460(0.320) | 0.830(0.219) |
| 20  | 0.20| 0.606(0.124) | 1.493(0.329) | 1.509(0.220) | 1.314(0.282) | 0.826(0.205) |
| 200 | 5   | 0.05| 0.356(0.108) | 1.415(0.210) | 0.865(0.279) | 0.769(0.220) | 0.695(0.215) |
| 10  | 0.10| 0.356(0.108) | 1.383(0.234) | 0.965(0.286) | 0.805(0.226) | 0.674(0.156) |
| 15  | 0.15| 0.356(0.108) | 1.371(0.245) | 1.090(0.280) | 0.803(0.200) | 0.638(0.157) |
| 20  | 0.20| 0.356(0.108) | 1.383(0.265) | 1.135(0.293) | 0.809(0.255) | 0.659(0.181) |
Table 5. Estimation error $\|\hat{P} - P\|_H$ for different estimators for Model 5. Numbers in parentheses are standard errors calculated from 100 generated data sets.

| $n$  | $L$  | $r$  | FDCOV | FSIR  | FSAVE | FDR   | FCR   |
|------|------|------|-------|-------|-------|-------|-------|
| 100  | 5    | 0.05 | 0.632(0.052) | 1.147(0.063) | 1.041(0.202) | 1.022(0.133) | 1.037(0.227) |
| 10   | 0.10 | 0.632(0.052) | 1.140(0.051) | 1.314(0.362) | 1.044(0.131) | 1.006(0.208) |
| 15   | 0.15 | 0.632(0.052) | 1.146(0.062) | 1.767(0.283) | 1.146(0.188) | 1.029(0.185) |
| 20   | 0.20 | 0.632(0.052) | 1.146(0.065) | 1.807(0.250) | 1.246(0.311) | 1.039(0.207) |
| 200  | 5    | 0.05 | 0.605(0.035) | 1.121(0.036) | 0.813(0.128) | 0.930(0.211) | 0.841(0.163) |
| 10   | 0.10 | 0.605(0.035) | 1.122(0.050) | 0.882(0.137) | 0.967(0.105) | 0.832(0.156) |
| 15   | 0.15 | 0.605(0.035) | 1.122(0.059) | 1.018(0.236) | 0.973(0.111) | 0.887(0.195) |
| 20   | 0.20 | 0.605(0.035) | 1.114(0.055) | 1.115(0.277) | 0.979(0.108) | 0.909(0.148) |

Table 6. Estimation error $\|\hat{P} - P\|_H$ for different estimators. Numbers in parentheses are standard errors calculated from 100 generated data sets.

| $n$  | methods | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|------|---------|---------|---------|---------|---------|---------|
| 100  | FDCOV   | 0.465(0.056) | 0.752(0.060) | 1.874(0.140) | 0.986(0.211) | 0.917(0.158) |
|      | FIR     | 1.007(0.060) | 2.016(0.087) | 2.532(0.697) | 1.985(0.326) | 1.282(0.265) |
|      | FCS     | 0.993(0.055) | 2.004(0.062) | 2.517(0.401) | 1.998(0.326) | 1.300(0.276) |
| 200  | FDCOV   | 0.416(0.055) | 0.718(0.054) | 1.688(0.119) | 0.851(0.177) | 0.834(0.102) |
|      | FIR     | 0.996(0.061) | 1.991(0.070) | 2.518(0.625) | 1.887(0.324) | 1.183(0.213) |
|      | FCS     | 0.984(0.049) | 1.996(0.059) | 2.507(0.400) | 1.897(0.245) | 1.296(0.190) |

$L$ and $r$. Note that FDCOV has no parameters to tune and is not related to $L$ and $r$.

To generate the sparse longitudinal data, we randomly selected 10 to 20 observations from $\{t_1, t_2, ..., t_{100}\}$ for each sample trajectory. The measurement error $\varepsilon_{ij}$ is independent and identically distributed as $N(0, 0.1^2)$. The simulation consists of 100 runs and Table 6 summarizes the results when $n$ is 100 and 200. For comparison, we also include the results of FIR [23] and FCS [11]. The estimation of functional principal components for spare longitudinal data is implemented through fdapace package in R system. The results suggest that our proposed method slightly outperforms the other methods we considered.

To examine the effectiveness of the bootstrap method for estimating $K$, we still consider the above five models. For each model, we consider $n = 200$ and $B = 100, 200$. As mentioned in Subsection 3.3, we use the average distance $1/B \cdot \sum_{b=1}^{B} \Delta_m(\hat{\eta}_k, \tilde{\eta}_b^k)$ as the measure of variability for each candidate $K$ and the results of these average distances under different settings are summarized in Table 7. The results show that the bootstrap method correctly chooses the dimension under different models.

4.2. Real data analysis

We consider the Tecator spectrometric data, available at http://lib.stat.cmu.edu/datasets/tecator and R package fda.usc. These data are recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850 - 1050 nm by the Near Infrared Transmission (NIT) principle. For each meat sample the data consists of a 100 channel spectrum of absorbance and the contents of moisture, fat and protein. The absorbance is $-\log_{10}$ of the transmittance measured by the spectrometer. The three contents, measured in percent, are determined by analytic chemistry. In this
example, the task is to predict the fat content $U$ of a meat sample on the basis of its near infrared absorbance spectrum $X$. The spectral data $X$ is the functional predictor and fat $U$ is the scalar variable. In accordance with the literature [6],[18], we use the transformed $Y = \log_{10}(U/(1 - U))$ as the response.

The sample size of the data is $n = 215$ and we use the first 150 for training and the remaining 65 for testing. We perform a spectral decomposition of $X$ and draw a scree plot of the eigenvalues in Figure 3. The scree plot shows that first 5 eigenvectors explain almost the total variation and the first 5 eigenvalues are not “too small”. Therefore, we select 5 as the truncation number. We apply all the methods in the simulation to this real data and we also consider $L = 5, 10, 15, 20$ and $r = 0.05, 0.10, 0.15, 0.20$ and use the bootstrap method mentioned in Subsection 3.3 to select the dimension $K$. The estimated dimension is $\hat{K} = 4$ in all cases. For every method, we obtain 4 estimated projections $(\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_3, \hat{\xi}_4) = (\langle \hat{\eta}_1, X \rangle, \langle \hat{\eta}_2, X \rangle, \langle \hat{\eta}_3, X \rangle, \langle \hat{\eta}_4, X \rangle)$ to estimate the unknown link function $g$. We use smoothing spline ANOVA method (ssanova function in the R package gss [35]). To measure the predictive performance of different methods, we use the root mean squared prediction error (RMSE) of the test sample which is

![Scree plot](Scree%20plot.png)

**Figure 3.** Eigenvalues of near infrared absorbance spectrum data.
Figure 4. $\hat{y}$ versus $y$ for different methods.

Table 8. Prediction errors for different estimators for real data.

| $L$ | $r$ | FDCOV | FSIR | FSAVE | FDR | FCR |
|-----|-----|-------|------|-------|-----|-----|
| 5   | 0.05| 0.094 | 0.135| 0.150 | 0.154| 0.147|
| 10  | 0.10| 0.094 | 0.132| 0.148 | 0.156| 0.159|
| 15  | 0.15| 0.094 | 0.127| 0.153 | 0.159| 0.146|
| 20  | 0.20| 0.094 | 0.133| 0.133 | 0.169| 0.137|

defined as

$$\text{RMSE} = \sqrt{n_{te}^{-1} \sum_{j=1}^{n_{te}} (y_j - \hat{y}_j)^2},$$

(20)

where $n_{te} = 65$ is the size of test sample, \( \hat{y}_j \) is the predictive value and \( y_j \) is the corresponding observed value. We also use 5-fold cross-validation discussed in Subsection 3.3 to select parameters $D$ and $K$. The parameters selected by the cross-validation method are $D = 5$, $K = 4$, the same as those of the previous method. The results of the prediction errors for different estimators are reported in Table 8. From Table 8, our proposed method outperforms FSIR, FSAVE, FDR and FCR under all settings. Figure 4 is the plot of $\hat{y} = \hat{g}(\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_3, \hat{\xi}_4)$ versus $y$ for our proposed FDCOV method and other competing methods on the test sample. The figure shows that the predictive responses for the proposed method are really closer to the test sample than some other methods, which indicates that our proposed method retains enough information for regression to predict the response variable.

5. Concluding remarks

In this work we propose a method of sufficient dimension reduction for functional data using distance covariance and establish its statistical consistency. In the estimation procedure, we adopt the commonly used functional PCA to project the infinite-
dimensional predictor onto a finite-dimensional subspace. We develop the FDCOV method to estimate $S_{y|x}$, along with procedures for determining the structural dimension. FDCOV requires very mild conditions on the predictor, unlike the existing methods require the restrictive linear conditional mean assumption and constant covariance assumption. It also does not involve the inverse of the covariance operator which is not bounded. In addition, the proposed method does not need to tune the parameters but other methods are needed. For example, the number of slices in FSIR, FS SAVE and FDR and the proportion of empirical directions in FCR.

In practice use, other basis such as wavelets and B-spline can also be considered for projection. The theoretical properties such as convergence rate and asymptotic normality need also be established. We consider the case that the response is a scalar in this article. However, the response can also be a random vector [36] or a random function [15] and the method will be adjusted accordingly in these cases. Nonlinear functional sufficient dimension reduction methods can also be developed by means of RKHS [15]. We leave these to future work.

6. Appendix

Proof of Proposition 2.1. Since $\text{span}(\beta) \subseteq \text{span}(\eta) = S_{y|x}$, $K_1 \leq K$ we can find a $K \times K_1$ matrix $A$, which satisfies $\beta = \eta A$. Therefore, $\mathcal{V}^2((\beta, X), Y) = \mathcal{V}^2(A^T(\eta, X), Y)$. Assume the single value decomposition of $A$ is $U \Sigma V^T$, where $U$ is a $K \times K$ orthogonal matrix, $V$ is a $K_1 \times K_1$ orthogonal matrix, and $\Lambda$ is a $K \times K_1$ diagonal matrix. Since $I_{K_1} = \langle \beta, \beta \rangle_{\Sigma_x} = \langle \eta A, \eta A \rangle_{\Sigma_x} = A^T(\eta, \eta)_{\Sigma_x}, A = A^T A$, we have that all nonzero numbers on the diagonal of $\Sigma$ are 1. According to the property (ii) in Subsection 2.1, $\mathcal{V}^2((\beta, X), Y) = \mathcal{V}^2(V \Lambda^T U^T(\eta, X), Y) = \mathcal{V}^2(\Lambda^T U^T(\eta, X), Y)$.

Denote $U^T(\eta, X) = (Z_1, \ldots, Z_K)^T$. Since all nonzero numbers on the diagonal of $\Lambda$ are 1, we have $\Lambda^T U^T(\eta, X) = (Z_1, \ldots, Z_K)^T$. Clearly, $\Lambda^T U^T(\eta, X)$ is a vector composed of the first $K_1$ components of $U^T(\eta, X)$. From this observation and Lemma A.1 in [24], we have $\mathcal{V}^2(\Lambda^T U^T(\eta, X), Y) \leq \mathcal{V}^2(U^T(\eta, X), Y)$ and the equality holds if and only if $K_1 = K$. By property (ii) in Subsection 2.1, $\mathcal{V}^2(U^T(\eta, X), Y) \leq \mathcal{V}^2((\eta, X), Y)$. Thus, we obtain $\mathcal{V}^2((\beta, X), Y) \leq \mathcal{V}^2((\eta, X), Y)$, and the equality holds if and only if $\text{span}(\beta) = \text{span}(\eta)$.

Proof of Proposition 2.2. For the $\beta$ and $\eta$ defined in Proposition 2.2, we can find a rotation matrix $R$ such that $\beta R = (\eta_a, \eta_b)$ and $\text{span}(\eta_a) \subseteq \text{span}(\eta)$. $\text{span}(\eta_b) \subseteq \text{span}(\eta)^\perp$ where $\text{span}(\eta)^\perp$ is the orthogonal complement space of $\text{span}(\eta)$ with respect to the inner product $(\cdot, \cdot)_{\Sigma_x}$.

The definition of $\eta$ indicates $Y \perp X|\langle \eta, X \rangle$, thus $Y \perp \langle \eta_b, X \rangle|\langle \eta, X \rangle$. By Condition 1, we have $\langle \eta_b, X \rangle \perp \langle \eta, X \rangle$. Therefore $\langle Y \rangle_{\langle \eta, X \rangle}$ and we can get $\langle Y \rangle_{\langle \eta_b, X \rangle}$ and $V_2 = 0$, then $(U_1, V_1) \perp (U_2, V_2)$. By property (iii) in Subsection 2.1, $\mathcal{V}^2(U_1 + U_2, V_1 + V_2) \leq \mathcal{V}^2(U_1 + V_1) + \mathcal{V}^2(U_2 + V_2)$, this means $\mathcal{V}^2(R^T(\beta, X), Y) = \mathcal{V}^2((\beta, X), Y) < \mathcal{V}^2((\eta, X), Y)$. $\mathcal{V}^2((\eta_b, X), Y) \leq \mathcal{V}^2((\eta, X), Y)$. $\mathcal{V}^2((\eta, X), Y)$.

Proof of Theorem 3.1. Suppose $\eta_{n,D_n}$ is not a consistent estimator of $S_{y|x}$, there exists a subsequence $\eta_{n^*, D_{n^*}}$ of $\eta_{n,D_n}$ such that $\eta_{n^*, D_{n^*}} \overset{P}{\longrightarrow} \eta^*$, where $\langle \eta^*, \eta^* \rangle_{\Sigma_x} = I_K$.
but \( \text{span}(\eta^*) \neq \text{span}(\eta) \). By Lemma A in [24], we have

\[
\mathcal{V}^2_n(\langle \eta_n^*, D_{n^*}, X \rangle, Y) - \mathcal{V}^2_n(\langle \eta^*, X \rangle, Y) \xrightarrow{P} 0.
\]

According to Theorem 2 in [25], \( \mathcal{V}^2_n(\langle \eta^*, X \rangle, Y) \xrightarrow{a.s.} \mathcal{V}^2(\langle \eta^*, X \rangle, Y) \), therefore \( \mathcal{V}^2_n(\langle \eta_n^*, D_{n^*}, X \rangle, Y) \xrightarrow{P} \mathcal{V}^2(\langle \eta^*, X \rangle, Y) \).

Besides, since \( \eta_{n,D_n} = \arg \max_{(\beta^n_{n,D_n})} \langle \beta^n_{D_n}, X \rangle \) \( \mathcal{V}^2_n(\langle \eta_{n,D_n}, X \rangle, Y) \geq \mathcal{V}^2_n(\langle \eta_{D_n}, X \rangle, Y) \), we have

\[
\mathcal{V}^2_n(\langle \eta_{n,D_n}, X \rangle, Y) \geq \mathcal{V}^2_n(\langle \eta_{D_n}, X \rangle, Y),
\]

where \( \eta_{D_n} \) is the representation of the function \( \eta \) in the \( D_n \)-truncated basis. Let \( n \to \infty \), we get \( \mathcal{V}^2(\langle \eta^*, X \rangle, Y) \geq \mathcal{V}^2(\langle \eta, X \rangle, Y) \), which contradicts to the definition of \( \eta \). Then we can conclude that \( \eta_{n,D_n} \) is a consistent estimator of \( \eta \). □

References

[1] Ramsay JO, Dalzell C. Some tools for functional data analysis. Journal of the Royal Statistical Society: Series B (Methodological). 1991;53(3):539–561.
[2] Cardot H, Ferraty F, Sarda P. Functional linear model. Statistics & Probability Letters. 1999;45(1):11–22.
[3] Ferraty F, Vieu P. The functional nonparametric model and application to spectrometric data. Computational Statistics. 2002;17(4):545–564.
[4] Ferré L, Yao AF. Functional sliced inverse regression analysis. Statistics. 2003;37(6):475–488.
[5] Li KC. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association. 1991;86(414):316–327.
[6] Ferré L, Yao AF. Smoothed functional inverse regression. Statistica Sinica. 2005;665–683.
[7] Wang G, Lin N, Zhang B. Functional contour regression. Journal of Multivariate Analysis. 2013;116:1–13.
[8] Wang G, Lin N, Zhang B. Functional k-means inverse regression. Computational Statistics & Data Analysis. 2014;70:172–182.
[9] Lian H, Li G. Series expansion for functional sufficient dimension reduction. Journal of Multivariate Analysis. 2014;124:150–165.
[10] Wang G, Zhou Y, Feng XN, et al. The hybrid method of fsir and fsave for functional effective dimension reduction. Computational Statistics & Data Analysis. 2015;91:64–77.
[11] Yao F, Lei E, Wu Y. Effective dimension reduction for sparse functional data. Biometrika. 2015;102(2):421–437.
[12] Wang G, Zhou J, Wu W, et al. Robust functional sliced inverse regression. Statistical papers. 2017;58(1):227–245.
[13] Wang G, Zhang F, Lian H. Directional regression for functional data. Journal of Statistical Planning and Inference. 2020;204:1–17.
[14] Lian H. Functional sufficient dimension reduction: Convergence rates and multiple functional case. Journal of Statistical Planning and Inference. 2015;167:58–68.
[15] Li B, Song J, et al. Nonlinear sufficient dimension reduction for functional data. The Annals of Statistics. 2017;45(3):1059–1095.
[16] Li B, Song J. Dimension reduction for functional data based on weak conditional moments. The Annals of Statistics. 2022;50(1):107–128.

[17] Lee KY, Li L. Functional sufficient dimension reduction through average Fréchet derivatives. The Annals of Statistics. 2022;50(2):904–929.

[18] Hsing T, Ren H, et al. An RKHS formulation of the inverse regression dimension-reduction problem. The Annals of Statistics. 2009;37(2):726–755.

[19] Wang G, Song X. Functional sufficient dimension reduction for functional data classification. Journal of Classification. 2018;35(2):250–272.

[20] Wang G, Liang B, Wang H, et al. Dimension reduction for functional regression with a binary response. Statistical Papers. 2019;1–16.

[21] Song J. On sufficient dimension reduction for functional data: Inverse moment-based methods. Wiley Interdisciplinary Reviews: Computational Statistics. 2019;11(4):e1459.

[22] Li B. Sufficient dimension reduction: Methods and applications with R. CRC Press; 2018.

[23] Jiang CR, Yu W, Wang JL. Inverse regression for longitudinal data. The Annals of Statistics. 2014;42(2):563–591.

[24] Sheng W, Yin X. Sufficient dimension reduction via distance covariance. Journal of Computational and Graphical Statistics. 2016;25(1):91–104.

[25] Székely GJ, Rizzo ML, Bakirov NK, et al. Measuring and testing dependence by correlation of distances. The Annals of Statistics. 2007;35(6):2769–2794.

[26] Székely GJ, Rizzo ML, et al. Brownian distance covariance. The Annals of Applied Statistics. 2009;3(4):1236–1265.

[27] Lyons R. Distance covariance in metric spaces. The Annals of Probability. 2013;41(5):3284–3305.

[28] Zhang J, Chen X. Robust sufficient dimension reduction via ball covariance. Computational Statistics & Data Analysis. 2019;140:144–154.

[29] Sheng W, Yin X. Direction estimation in single-index models via distance covariance. Journal of Multivariate Analysis. 2013;122:148–161.

[30] Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association. 2005;100(470):577–590.

[31] Hall P, Müller HG, Wang JL. Properties of principal component methods for functional and longitudinal data analysis. The Annals of Statistics. 2006;1493–1517.

[32] Wang JL, Chiou JM, Müller HG. Functional data analysis. Annual Review of Statistics and Its Application. 2016;3:257–295.

[33] Riesz F, Nagy S. Functional analysis. Dover Publications, Inc, New York First published in. 1955(3):35.

[34] Xue Y, Zhang N, Yin X, et al. Sufficient dimension reduction using hilbert–schmidt independence criterion. Computational Statistics & Data Analysis. 2017;115:67–78.

[35] Gu C. Smoothing spline anova models. Vol. 297. Springer Science & Business Media; 2013.

[36] Chen X, Yuan Q, Yin X. Sufficient dimension reduction via distance covariance with multivariate responses. Journal of Nonparametric Statistics. 2019;31(2):268–288.

[37] Cook RD. Regression graphics: Ideas for studying regressions through graphics. Vol. 482. John Wiley & Sons; 2009.