The Annals of Statistics
2022, Vol. 50, No. 2, 904–929
https://doi.org/10.1214/21-AOS2131

FUNCTIONAL SUFFICIENT DIMENSION REDUCTION THROUGH AVERAGE FRÉCHET DERIVATIVES

BY KUANG-YAO LEE[1,a] AND LEXIN LI[2,b]

[1]*Department of Statistical Science, Temple University,* [a]*kuang-yao.lee@temple.edu*
[2]*Department of Biostatistics, University of California, Berkeley,* [b]*lexinli@berkeley.edu*

Sufficient dimension reduction (SDR) embodies a family of methods that aim for reduction of dimensionality without loss of information in a regression setting. In this article, we propose a new method for nonparametric function-on-function SDR, where both the response and the predictor are a function. We first develop the notions of functional central mean subspace and functional central subspace, which form the population targets of our functional SDR. We then introduce an average Fréchet derivative estimator, which extends the gradient of the regression function to the operator level and enables us to develop estimators for our functional dimension reduction spaces. We show the resulting functional SDR estimators are unbiased and exhaustive, and more importantly, without imposing any distributional assumptions such as the linearity or the constant variance conditions that are commonly imposed by all existing functional SDR methods. We establish the uniform convergence of the estimators for the functional dimension reduction spaces, while allowing both the number of Karhunen–Loève expansions and the intrinsic dimension to diverge with the sample size. We demonstrate the efficacy of the proposed methods through both simulations and two real data examples.

1. Introduction. Sufficient dimension reduction (SDR) embodies a family of methods in regressions that seek a low-dimensional representation of the high-dimensional data while minimizing the loss of regression information. Since the pioneering work of sliced inverse regression (Li (1991)), SDR has enjoyed a rapid development in the past three decades, and has been widely used in a variety of applications, including biology, finance, medical science, among others. In its full generality, SDR aims at the entire conditional distribution of $Y|X$, where Y is a response variable and X a p-dimensional predictor vector, and (X, Y) has a joint distribution. It seeks a lower-dimensional representation $R(X)$ such that $Y \perp\!\!\!\perp X \mid R(X)$. In plenty of applications, the regression interest focuses on the conditional mean $E(Y|X)$ only. In that case, Cook and Li (2002) developed the notion of sufficient mean dimension reduction, by seeking $R(X)$ such that $Y \perp\!\!\!\perp E(Y|X) \mid R(X)$, or equivalently, $E(Y|X) = E\{Y|R(X)\}$. The reduction $R(X)$ usually takes the form of linear combinations of X, that is, $R(X) = \beta^{\mathsf{T}} X$, where β is a $p \times q$ matrix with $q \leq p$. Sufficient reduction or sufficient mean reduction then pursues the minimum subspace spanned by β, which uniquely exists under very mild conditions (Yin, Li and Cook (2008)). Such a space is called the central subspace or the central mean subspace, and its dimension q is called the intrinsic dimension. There have been a large body of methods proposed for SDR, usually in a nonparametric fashion. Depending on their estimation strategies, broadly speaking, those methods can be grouped into two categories. One category utilizes the first and second moments of inverse regression $X|Y$, which under certain distributional assumptions, contain useful information about $R(X)$; see, for example, Cook and Weisberg (1991), Li (1991), Li, Zha and Chiaromonte (2005), Li and Wang

Received April 2021; revised August 2021.
MSC2020 subject classifications. 62B05, 62G08, 62G20, 62R10.
Key words and phrases. Functional central mean subspace, functional central subspace, function-on-function regression, unbiasedness, exhaustiveness, consistency, reproducing kernel Hilbert space.

(2007), among others. The other category utilizes the gradient of the forward regression mean $E(Y|X)$; see, for example, Hardle (1989), Xia (2007), Xia et al. (2002), Yin and Li (2011), Fukumizu and Leng (2014), among others. See Li (2018a) for a comprehensive review of SDR.

In this article, we target the problem of sufficient reduction and sufficient mean reduction when both the response and the predictor are a function. Function-on-function regression is receiving increasing attention in recent years, and is being widely used in applications such as environmental science, neuroimaging analysis, and e-commerce (Kim et al. (2018), Luo and Qi (2017), Müller and Yao (2008), Reimherr, Sriperumbudur and Taoufik (2018), Sun et al. (2018), Luo and Qi (2019)). We aim to relax the parametric or semiparametric model assumptions, and propose a new method of nonparametric function-on-function SDR. We first develop the notions of functional central mean subspace and functional central subspace, which are the population targets of our functional SDR inquiries. Motivated by the fact that the gradient of the regression mean function $E(Y|X)$ lies in the central mean or central subspace (Xia et al. (2002), Xia (2007)), we extend the idea to the functional setting, and show that the Riesz representation of the Fréchet derivative of the regression functional is located in the functional central mean or central subspace. We then propose the corresponding functional dimension reduction estimators based on the average Fréchet derivative, and show that the resulting estimators are both unbiased and exhaustive. Moreover, our proposal leads naturally to a procedure for predicting any functional of the response after dimension reduction. Theoretically, we establish the uniform convergence for the estimated bases of the dimension reduction spaces, while allowing both the number of Karhunen–Loève expansions and the intrinsic dimension to diverge with the sample size.

Our proposal is closely related to but also clearly different from a number of lines of research on sufficient dimension reduction and functional regression. We next review the relevant literature and discuss the connections and differences with our proposal.

First, our proposal extends the gradient-based dimension reduction methods such as Hardle (1989), Xia (2007), Xia et al. (2002), Yin and Li (2011), Fukumizu and Leng (2014) from the random variable setting to the random function setting. However, such an extension is far beyond routine, and our new solution is utterly different, in both computation and theory, from the existing ones. On the computation side, most existing gradient-based methods approximated the high-dimensional gradients using iterative least squares, and the resulting computation can be intensive (Xia (2007), Xia et al. (2002), Yin and Li (2011)). To address this issue, we establish the interchangeability between the Fréchet derivative and the reproducing kernel Hilbert space (RKHS), which in effect provides a closed form of the Fréchet derivative. Consequently, our algorithm requires only spectral decompositions of linear operators and no iterative optimization, and hence is computationally much simpler. On the theory side, in the random variable setting, the variable dimension p and the intrinsic dimension q after dimension reduction are usually treated as fixed. Only recently, Lin et al. (2017) and Lin, Zhao and Liu (2019) established the consistency of sliced inverse regression with a diverging intrinsic dimension q. By contrast, in the random function setting, both the response and the predictor are random elements in a potentially infinite-dimensional Hilbert space. We need to handle a diverging number of Karhunen–Loève (KL) expansions from the random functions, and this portion of asymptotic analysis requires a more complex treatment than in the classical setting. Besides, we allow the intrinsic dimension q to diverge with the sample size.

Second, our work is also related to a family of proposals that extended the inverse regression-based SDR methods to the functional setting. The first generalization was introduced by Ferré and Yao (2003, 2005), in which they extended the predictor space from an Euclidean space to a Hilbert space, then proposed a functional slice inverse regression

method. Hsing and Ren (2009), Jiang, Yu and Wang (2014), Wang, Lin and Zhang (2013), Wang et al. (2015), Yao, Lei and Wu (2015) further extended and developed a series of inverse regression-based estimators in the functional setting. All these methods involved a scalar response Y and a functional predictor X, targeted sufficient reduction of the full conditional distribution $Y|X$, and were based on some moments of the inverse regression $X|Y$. Our proposal on one hand extends SDR to the setting where not only the predictor but also the response is a function. More importantly, our method is built on the derivative of the forward regression, instead of the inverse regression. A critical implication of this difference is that, our functional SDR estimators achieve the unbiasedness and exhaustiveness *without* imposing distributional assumptions such as the linearity condition or the constant variance condition that are commonly imposed by all existing inverse moments-based functional SDR methods. As pointed out by Ma and Zhu (2012, 2013) under the random variable setting, those distributional assumptions can be restrictive, and our proposal is the first unbiased and exhaustive solution that relaxes those conditions under the random function setting.

More recently, Li and Song (2017) proposed a nonlinear SDR method for functional data. Our proposal is similar, in that we both target SDR for functional data, and both methods are built on some linear operators. However, the two also differ considerably in terms of methodology and theory. On the methodology side, Li and Song (2017) focused on nonlinear reduction, rather than linear reduction. Although nonlinear reduction is more flexible in the sense that its reduced predictors do not have to be linear, the linear framework of our proposal is still pivotally important. Linear SDR is usually easier to interpret, and is better connected with other modeling techniques, since it preserves the original coordinates of the predictors; see Li ((2018a), Chapter 14) for more comparison between linear and nonlinear SDR. Moreover, our estimator can not be directly induced by that of Li and Song (2017). On the theory side, Li and Song (2017) required the constant variance assumption to show their generalized sliced average variance estimator is unbiased, which can be restrictive. By contrast, our estimator does not impose such a condition. Furthermore, our asymptotic analysis requires a different set of tools from that of Li and Song (2017). Specifically, we employ the leading KL coefficients to approximate the sample covariance operators to accelerate the computation, which was not considered in Li and Song (2017). As a result, we need to take into account this extra layer of approximation when proving the consistency. More importantly, Li and Song (2017) treated the intrinsic dimension q as fixed, while we allow q to diverge, which leads to a completely different regime of asymptotics.

In summary, our proposal is profoundly different from the existing sufficient dimension reduction methods. The extension from random variables to random functions, and the change from inverse regression to forward regression-based estimation are both far from straightforward. It leads to a new set of methodological and theoretical results, and makes a useful addition to the toolbox of dimension reduction and functional data analysis.

The rest of the article is organized as follows. Section 2 introduces the functional central mean and central subspaces. Section 3 develops the average Fréchet derivative estimator for functional SDR, and establishes the population properties including the unbiasedness and exhaustiveness. Section 4 develops the estimation procedure, both at the operator level and under a coordinate system. Section 5 derives the asymptotic properties. Section 6 presents the simulations and two real data examples. Section 7 concludes the paper with a discussion. The Supplementary Material (Lee and Li (2022)) collects additional results and proofs.

2. Functional dimension reduction subspaces. In this section, we first formally define the functional central mean and central subspaces, which are the population targets of functional SDR. We then introduce a series of linear operators useful for SDR estimation.

2.1. *Functional central mean subspace and central subspace.* Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $X : \Omega \to \Omega_X$, $Y : \Omega \to \Omega_Y$ be the Ω_X-valued and Ω_Y-valued random elements, where Ω_X and Ω_Y are Hilbert spaces of functions on an interval $T \subseteq \mathbb{R}$, and their inner products are $\langle \cdot, \cdot \rangle_{\Omega_X}$ and $\langle \cdot, \cdot \rangle_{\Omega_Y}$. Let $P_X = P \circ X^{-1}$ and $P_Y = P \circ Y^{-1}$ be the distributions of X and Y. We first seek sufficient reduction for the mean of the conditional distribution $Y|X$.

DEFINITION 1. Suppose $\mathcal{S}$ is a linear, closed subspace of Ω_X, and $P_{\mathcal{S}} : \Omega_X \to \mathcal{S}$ is the projection onto $\mathcal{S}$. If, for all $\psi \in \Omega_Y$,

$$E\big(\langle \psi, Y \rangle_{\Omega_Y} \mid X\big) = E\big(\langle \psi, Y \rangle_{\Omega_Y} \mid P_{\mathcal{S}} X\big), \tag{1}$$

we call $\mathcal{S}$ a functional mean dimension reduction subspace. Let $\mathcal{T}$ denote the collection of all $\mathcal{S}$ satisfying (1). If $\mathcal{S}_{E(Y|X)} = \cap\{\mathcal{S} : \mathcal{S} \in \mathcal{T}\}$ is in $\mathcal{T}$, we call $\mathcal{S}_{E(Y|X)}$ the functional central mean subspace.

We next seek sufficient reduction for the entire conditional distribution $Y|X$. The idea follows that of Xia (2007), Yin and Li (2011), in that one can extend the estimation of central mean subspace to that of central subspace through a class of functions that characterize the conditional distribution $Y \mid X$. More specifically, note that the collection of conditional means $\{E\{g(Y)|X\} : g \in \mathcal{H}_Y\}$ characterize the full information of $Y \mid X$, as long as the class of functions $\mathcal{H}_Y$ is sufficiently rich. This implies, if there is a functional dimensional reduction space for the conditional mean $E\{g(Y)|X\}$ for all $g \in \mathcal{H}_Y$, then it is also a functional dimension reduction space for $Y \mid X$. We formalize this idea below.

DEFINITION 2. Let $\mathcal{H}_Y$ be a dense subset of $L_2(P_Y)$ modulo constants. Suppose $\mathcal{S}$ is a linear, closed subspace of Ω_X, and $P_{\mathcal{S}} : \Omega_X \to \mathcal{S}$ is the projection onto $\mathcal{S}$. If, for all $g \in \mathcal{H}_Y$,

$$E\big\{g(Y) \mid X\big\} = E\big\{g(Y) \mid P_{\mathcal{S}} X\big\}, \tag{2}$$

we call $\mathcal{S}$ a functional dimension reduction subspace. Let $\mathcal{T}$ denote the collection of all $\mathcal{S}$ satisfying (2). If $\mathcal{S}_{Y|X} = \cap\{\mathcal{S} : \mathcal{S} \in \mathcal{T}\}$ is in $\mathcal{T}$, we call $\mathcal{S}_{Y|X}$ the functional central subspace.

Here we say $\mathcal{H}_Y$ is dense in $L_2(P_Y)$ modulo constants if, for any $g \in L_2(P_Y)$, there exists a sequence $\{g^k\}_{k\in\mathbb{N}}$ of $\mathcal{H}_Y$, such that $E\{g^k(X) - g(X)\}^2 \to 0$, as $k \to \infty$, where $\mathbb{N} = \{0, 1, 2, \ldots\}$ denotes the collection of natural numbers. We also note that, if $\mathcal{H}_Y$ is a dense subset of $L_2(P_Y)$, $\mathcal{H}_Y$ is characteristic (Fukumizu, Bach and Jordan (2009)). Therefore, (2) is equivalent to $Y \perp\!\!\!\perp X \mid P_{\mathcal{S}} X$. This justifies why (2) can characterize the conditional distribution $Y \mid X$. It also shows that $\mathcal{S}_{Y|X}$ is independent of the selection of $\mathcal{H}_Y$.

In both definitions of $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{Y|X}$, we are seeking the smallest subspace in $\mathcal{T}$ that satisfies (1) or (2), so to achieve maximal dimension reduction. Toward that goal, we take the intersection of all subspaces in $\mathcal{T}$, and assume such an intersection still belongs to $\mathcal{T}$. This actually holds under some mild conditions, as we show next. We introduce the concept of M-set in an Hilbert space.

DEFINITION 3. Let Ω_1 and Ω_2 denote two generic Hilbert spaces, and M a subset of the product space $\Omega_1 \times \Omega_2$. If, for any two pairs of points $(\omega_1, \omega_2), (\omega_1^*, \omega_2^*) \in M$, there exist a sequence of pairs of points $\{(\omega_1^j, \omega_2^j)\}_{j=1}^J \in \Omega_1 \times \Omega_2$, with $(\omega_1^1, \omega_2^1) = (\omega_1, \omega_2)$ and $(\omega_1^J, \omega_2^J) = (\omega_1^*, \omega_2^*)$, such that, (a) $(\omega_1^j, \omega_2^j) \in M$, for $j = 2, \ldots, J-1$, and (b) $\omega_1^{j+1} = \omega_1^j$ or $\omega_2^{j+1} = \omega_2^j$, for $j = 1, \ldots, J-1$, then we call M an M-set in $\Omega_1 \times \Omega_2$.

Yin, Li and Cook (2008) introduced the notion of M-set for the random variable setting, and Definition 3 extends it to the random function setting. The conditions in Definition 3 require that any two points in an M-set can be connected via a "stairway", and that only the corner points need to belong to the M-set. These are fairly mild conditions; see, for example, any convex and open subset of $\Omega_1 \times \Omega_2$ is an M-set. Let $\text{supp}(X)$ denote the support of X. For linear and closed subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ of Ω_X, and g a member in $\{P_{\mathcal{S}_1 \cap \mathcal{S}_2} f : f \in \text{supp}(X)\}$, denote $\Omega(\mathcal{S}_1, \mathcal{S}_2) = \{(P_{\mathcal{S}_1} f, P_{\mathcal{S}_2} f, P_{\mathcal{S}_1 \cap \mathcal{S}_2} f) : f \in \text{supp}(X)\}$, and $\Omega(\mathcal{S}_1, \mathcal{S}_2, g) = \{(P_{\mathcal{S}_1} f, P_{\mathcal{S}_2} f) : (P_{\mathcal{S}_1} f, P_{\mathcal{S}_2} f, g) \in \Omega(\mathcal{S}_1, \mathcal{S}_2)\}$. Let $\mathcal{T}$ be the collection of all subspaces satisfying (1) or (2). The next theorem justifies the existence of $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{Y|X}$.

THEOREM 1. *Suppose, for any $\mathcal{S}_1$ and $\mathcal{S}_2$ in $\mathcal{T}$, $\Omega(\mathcal{S}_1, \mathcal{S}_2, g)$ is an M-set in $\mathcal{S}_1 \times \mathcal{S}_2$, for all $g \in \{P_{\mathcal{S}_1 \cap \mathcal{S}_2} f : f \in \text{supp}(X)\}$.*

(a) *If the subspaces in $\mathcal{T}$ satisfy (1), then there exists a unique intersection, $\mathcal{S}_{E(Y|X)} = \cap\{\mathcal{S} : \mathcal{S} \in \mathcal{T}\} \subseteq \Omega_X$, satisfying that $\mathcal{S}_{E(Y|X)}$ is in $\mathcal{T}$, and $\mathcal{S}_{E(Y|X)} \subseteq \mathcal{S}$, for all $\mathcal{S} \in \mathcal{T}$;*
(b) *If the subspaces in $\mathcal{T}$ satisfy (2), then there exists a unique intersection, $\mathcal{S}_{Y|X} = \cap\{\mathcal{S} : \mathcal{S} \in \mathcal{T}\} \subseteq \Omega_X$, satisfying that $\mathcal{S}_{Y|X}$ is in $\mathcal{T}$, and $\mathcal{S}_{Y|X} \subseteq \mathcal{S}$, for all $\mathcal{S} \in \mathcal{T}$.*

We also note that, the sufficient predictors in Definitions 1 and 2 are linear mappings on Ω_X. Consider full reduction as an example, and let $\mathcal{S} = \text{Span}\{\phi_i \in \Omega_X : i = 1 \ldots, q\}$. Then

$$Y \perp\!\!\!\perp X \mid \langle X, \phi_1 \rangle_{\Omega_X}, \ldots \langle X, \phi_q \rangle_{\Omega_X}.$$

For this reason, our functional SDR is *linear* dimension reduction. By contrast, Li and Song (2017) studied *nonlinear* dimension reduction, by relaxing the linear constraint and allowing the sufficient predictors to be nonlinear mappings, in that,

$$Y \perp\!\!\!\perp X \mid f_1(X), \ldots, f_q(X),$$

where $f_1, \ldots, f_q$ are elements of the RKHS $\mathcal{H}_X$. This key difference leads to two sets of utterly different methodology and theory for our proposal and for Li and Song (2017), as detailed in Sections 1 and 7.

We next give some concrete examples of $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{Y|X}$. For a positive definite kernel function $\kappa_T(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, let Ω_T denote the RKHS induced by κ_T, which can be constructed by $\Omega_T = \overline{\text{Span}}\{\kappa_T(\cdot, t) : t \in \mathbb{R}\}$, where $\overline{\text{Span}}$ is the closure of the space spanned by the set of functions. Next, consider the Brownian motion kernel, $\kappa_T(t_1, t_2) = \min(t_1, t_2)$, and let (α_j, β_j) denote the jth pair of eigenvalue and eigenfunction of κ_T, with $\alpha_j = 1/\{(j - 0.5)\pi\}^2$, and $\beta_j(t) = \sin\{(j - 0.5)\pi t\}$, $j \in \mathbb{N}$ (see, e.g., Amini and Wainwright (2012)). We then construct the predictor function $X(t)$ and the error function $\epsilon(t)$ using the leading pairs of eigenvalues and eigenfunctions, where

$$X(t) = \sum_{j=1}^{J} a_j \sqrt{\alpha_j} \beta_j(t), \quad \epsilon(t) = \sum_{k=1}^{K} b_k \sqrt{\alpha_k} \beta_k(t), \quad \text{for some integer values } J, K,$$

$E(a_j) = E(b_k) = 0$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, and $(a_1, \ldots, a_J)^\mathsf{T}$ and $(b_1, \ldots, b_K)^\mathsf{T}$ are independent. By this construction, both $\epsilon(t)$ and $X(t)$ are elements in Ω_T. Besides, for any $\phi \in \Omega_T$, $E\langle \phi, \epsilon(t) \rangle_{\Omega_T} = \sum_{k=1}^{K} \sqrt{\alpha_k} E(b_k) \langle f, b_k(t) \rangle_{\Omega_T} = 0$. Moreover, because there are one-to-one correspondences between $X(t)$ and $(a_1, \ldots, a_J)^\mathsf{T}$, and between $\epsilon(t)$ and $(b_1, \ldots, b_K)^\mathsf{T}$, we have $\epsilon(t) \perp\!\!\!\perp X(t)$. Following this construction, we consider the following examples.

EXAMPLE 1. Suppose there exist $\phi_1, \phi_2 \in \Omega_T$, such that $X(t)$ and $Y(t)$ satisfy that

$$Y(t) = m_1\big(\langle \phi_1, X\rangle_{\Omega_T}\big)\sin(0.5\pi t) + m_2\big(\langle \phi_2, X\rangle_{\Omega_T}\big)\sin(1.5\pi t) + \sigma\varepsilon(t),$$

where $m_1(\cdot)$ and $m_2(\cdot)$ are mappings from $\mathbb{R}$ to $\mathbb{R}$, σ is a positive constant, and $\langle \phi, \psi\rangle_{\Omega_T} = \int\{d\phi(t)/dt\, d\psi(t)/dt\}dt$, for any $\phi, \psi \in \Omega_T$. Then Y is an Ω_T-valued random element, and $\mathcal{S}_{E(Y|X)} = \operatorname{Span}\{\phi_1, \phi_2\}$.

EXAMPLE 2. Suppose there exist $\phi_1, \phi_2, \phi_3 \in \Omega_T$, such that $X(t)$ and $Y(t)$ satisfy that

$$Y(t) = m_1\big(\langle \phi_1, X\rangle_{\Omega_T}\big)\sin(0.5\pi t) + m_2\big(\langle \phi_2, X\rangle_{\Omega_T}\big)\sin(1.5\pi t) + m_3\big(\langle \phi_3, X\rangle_{\Omega_T}\big)\varepsilon(t),$$

where $m_1(\cdot)$ to $m_3(\cdot)$ are mappings from $\mathbb{R}$ to $\mathbb{R}$, and $\langle \phi, \psi\rangle_{\Omega_T} = \int\{d\phi(t)/dt\, d\psi(t)/dt\}dt$, for any $\phi, \psi \in \Omega_T$. Then Y is an Ω_T-valued random element, and $\mathcal{S}_{Y|X} = \operatorname{Span}\{\phi_1, \phi_2, \phi_3\}$.

Finally, we note that it is straightforward to show $\mathcal{S}_{E(Y|X)} \subseteq \mathcal{S}_{Y|X}$. This echos the classical result in the random variable setting. In Example 2, $\mathcal{S}_{E(Y|X)}$ is a proper subset of $\mathcal{S}_{Y|X}$.

2.2. *Functional regression operator.* By Definitions 1 and 2, the subspaces of interest are defined via the conditional expectation $E(\langle \psi, Y\rangle_{\Omega_X}|X)$ or $E\{g(Y)|X\}$, that is, the regression functionals that link the predictor function X and the response function Y. In order to characterize such a regression relationship, we introduce the functional regression operator as an extension of regression coefficient to both functional and nonlinear settings.

We first define the nested kernel and the associated RKHS. A positive definite kernel $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$ can be constructed via the inner product $\langle\cdot,\cdot\rangle_{\Omega_X}$, if there exists a mapping $\rho : \mathbb{R}^3 \to \mathbb{R}$, such that

$$\kappa_X(\phi_1, \phi_2) = \rho\big(\langle \phi_1, \phi_1\rangle_{\Omega_X}, \langle \phi_1, \phi_2\rangle_{\Omega_X}, \langle \phi_2, \phi_2\rangle_{\Omega_X}\big), \quad \text{for any } \phi_1, \phi_2 \in \Omega_X. \tag{3}$$

Because κ_X is uniquely characterized by Ω_X, it is called a nested kernel. Examples of nested kernels include the radial basis function kernel $\kappa_X(\phi_1, \phi_2) = \exp(-\gamma\|\phi_1 - \phi_2\|^2_{\Omega_X})$, the polynomial kernel $\kappa_X(\phi_1, \phi_2) = (1 + \langle \phi_1, \phi_2\rangle_{\Omega_X})^\gamma$, among others. Then given κ_X, a nested RKHS $\mathcal{H}_X$ can be induced by this kernel. Similarly, a positive definite kernel $\kappa_Y : \Omega_Y \times \Omega_Y \to \mathbb{R}$, and a nested RKHS $\mathcal{H}_Y$ can be constructed via the inner product $\langle\cdot,\cdot\rangle_{\Omega_Y}$.

We next introduce some background and notation. Let $\mathcal{H}$, $\mathcal{H}'$ be two generic Hilbert spaces, and A a linear mapping from $\mathcal{H}$ to $\mathcal{H}'$. Define the norm of A as $\|A\| = \sup\{\|Af\|_{\mathcal{H}'} : f \in \mathcal{H}, \|f\|_{\mathcal{H}} = 1\}$, where $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}'}$ are the norms of $\mathcal{H}$ and $\mathcal{H}'$, respectively. An operator is said to be bounded, if its norm is finite. A linear operator $A : \mathcal{H} \to \mathcal{H}'$ is said to be Hilbert-Schmidt, if $\sum_{k\in\mathbb{N}} \|Af^k\|^2_{\mathcal{H}'} < \infty$ for an orthonormal basis $\{f^k\}_{k\in\mathbb{N}}$ of $\mathcal{H}$. Define the Hilbert-Schmidt norm of A as $\|A\|_{\mathrm{HS}} = (\sum_{k\in\mathbb{N}} \|Af^k\|^2_{\mathcal{H}'})^{1/2}$. Let $B(\mathcal{H}, \mathcal{H}')$ be the collection of all linear bounded operators from $\mathcal{H}$ to $\mathcal{H}'$, and $B_2(\mathcal{H}, \mathcal{H}')$ the collection of all Hilbert-Schmidt operators from $\mathcal{H}$ to $\mathcal{H}'$. Note that $B(\mathcal{H}, \mathcal{H}')$ is a Banach space endowed with the operator norm $\|\cdot\|$, and $B_2(\mathcal{H}, \mathcal{H}')$ is a Hilbert space with its inner product defined as $\langle A_1, A_2\rangle_{\mathrm{HS}} = \sum_{k\in\mathbb{N}}\langle A_1 f^k, A_2 f^k\rangle_{\mathcal{H}'}$, for any $A_1, A_2 \in B_2(\mathcal{H}, \mathcal{H}')$. Because $\|A\| \le \|A\|_{\mathrm{HS}}$, it holds that $B_2(\mathcal{H}, \mathcal{H}') \subseteq B(\mathcal{H}, \mathcal{H}')$ (Weidmann (1980)). We abbreviate $B(\mathcal{H}) = B(\mathcal{H}, \mathcal{H})$ and $B_2(\mathcal{H}) = B_2(\mathcal{H}, \mathcal{H})$ whenever appropriate. Furthermore, let $\ker(A)$, $\operatorname{ran}(A)$, $\overline{\operatorname{ran}}(A)$, A^* denote the null space, the range, the closure of the range, and the adjoint of an operator A, respectively.

We now develop a series of linear operators. We begin with a regularity condition. It ensures the square-integrability of the sample path of Y and that of every function in $\mathcal{H}_X$ and $\mathcal{H}_Y$. The first part is a standard moment condition; see, for example, Yao, Lei and Wu (2015). The second part holds for all bounded kernels. To avoid digression, we defer a detailed discussion of this condition, along with all other conditions in this article, to Section S1 of the Supplementary Material (Lee and Li (2022)).

ASSUMPTION 1. Suppose $E\|Y\|^2_{\Omega_Y} < \infty$, $E\{\kappa_X(X,X)\} < \infty$, and $E\{\kappa_Y(Y,Y)\} < \infty$. In addition, $\mathcal{H}_X$ and $\mathcal{H}_Y$ are dense in $L_2(P_X)$ and $L_Y(P_Y)$ modulo constants, respectively.

Let m_X denote the mean element of $\mathcal{H}_X$, such that $\langle f, m_X\rangle_{\mathcal{H}_X} = E\langle f, \kappa_X(\cdot, X)\rangle_{\mathcal{H}_X} = E\{f(X)\}$ for any $f \in \mathcal{H}_X$. Similarly, let μ_Y and m_Y denote the mean elements of Ω_Y and $\mathcal{H}_Y$, respectively. We next define a number of covariance operators:

$$\begin{aligned}
&\Lambda_{XY} : \Omega_Y \to \mathcal{H}_X, \quad \langle f, \Lambda_{XY}\psi\rangle_{\mathcal{H}_X} = E\big\{\big\langle f, \kappa_X(\cdot, X) - m_X\big\rangle_{\mathcal{H}_X}\langle \psi, Y - \mu_Y\rangle_{\Omega_Y}\big\},\\
&\Lambda_{YY} : \Omega_Y \to \Omega_Y, \quad \langle \psi', \Lambda_{YY}\psi\rangle_{\Omega_Y} = E\big\{\big\langle \psi', Y - \mu_Y\big\rangle_{\Omega_Y}\langle \psi, Y - \mu_Y\rangle_{\Omega_Y}\big\},\\
&\Sigma_{XX} : \mathcal{H}_X \to \mathcal{H}_X, \quad \langle f', \Sigma_{XX} f\rangle_{\mathcal{H}_X} = E\big\{\big\langle f', \kappa_X(\cdot, X) - m_X\big\rangle_{\mathcal{H}_X}\big\langle f, \kappa_X(\cdot, X) - m_X\big\rangle_{\mathcal{H}_X}\big\},\\
&\Sigma_{XY} : \mathcal{H}_Y \to \mathcal{H}_X, \quad \langle f, \Sigma_{XY} g\rangle_{\mathcal{H}_Y} = E\big\{\big\langle f, \kappa_X(\cdot, X) - m_X\big\rangle_{\mathcal{H}_X}\big\langle g, \kappa_Y(\cdot, Y) - m_Y\big\rangle_{\mathcal{H}_Y}\big\},\\
&\Sigma_{YY} : \mathcal{H}_Y \to \mathcal{H}_Y, \quad \langle g', \Sigma_{YY} g\rangle_{\mathcal{H}_Y} = E\big\{\big\langle g', \kappa_Y(\cdot, Y) - m_Y\big\rangle_{\mathcal{H}_Y}\big\langle g, \kappa_Y(\cdot, Y) - m_Y\big\rangle_{\mathcal{H}_Y}\big\},
\end{aligned}$$

for any $\psi, \psi' \in \Omega_Y$, $f, f' \in \mathcal{H}_X$, and $g, g' \in \mathcal{H}_Y$. Note that $\Sigma_{YX} = \Sigma^*_{XY}$. The next lemma justifies the existence of the mean elements and covariance operators we define above.

LEMMA 1. *Suppose Assumption 1 holds. Then there exist the mean elements m_X, μ_Y, and m_Y, and the covariance operators Λ_{XY}, Λ_{YY}, Σ_{XX}, Σ_{XY}, Σ_{YY}.*

Next, we define two functional regression operators,

$$m_{E(Y|X)} = \Sigma^\dagger_{XX}\Lambda_{XY}, \qquad m_{Y|X} = \Sigma^\dagger_{XX}\Sigma_{XY},$$

where $\Sigma^\dagger_{XX} : \mathrm{ran}(\Sigma_{XX}) \to \overline{\mathrm{ran}}(\Sigma_{XX})$ is the Moore–Penrose inverse of Σ_{XX}. That is, for any $f \in \mathrm{ran}(\Sigma_{XX})$, there exist $g \in \ker(\Sigma_{XX})$ and $h \in \overline{\mathrm{ran}}(\Sigma_{XX})$ such that $f = \Sigma_{XX}(g + h)$, and the Moore–Penrose inverse $\Sigma^\dagger_{XX}$ maps $f \in \Sigma_{XX}$ to $h \in \overline{\mathrm{ran}}(\Sigma_{XX})$ (Li (2018b)). Because they resemble the notion of classical regression coefficients, we refer $m_{E(Y|X)}$ and $m_{Y|X}$ as the functional regression operators. We next introduce another regularity condition to ensure such definitions. See Section S1 of the Supplementary Material (Lee and Li (2022)) for more discussion on this condition.

ASSUMPTION 2. Suppose $\ker(\Sigma_{XX}) = \{0\}$, $\mathrm{ran}(\Lambda_{XY}) \subseteq \mathrm{ran}(\Sigma_{XX})$, and $\mathrm{ran}(\Sigma_{XY}) \subseteq \mathrm{ran}(\Sigma_{XX})$. In addition, $m_{E(Y|X)} \in B_2(\Omega_Y, \mathcal{H}_X)$, and $m_{Y|X} \in B_2(\mathcal{H}_Y, \mathcal{H}_X)$.

Assumption 2 ensures that the mappings of Λ_{XY} and Σ_{XY} are both in that of Σ_{XX}, which in turns ensures that the two operators $m_{E(Y|X)}$ and $m_{Y|X}$ are well defined. Moreover, it also ensures that the Moore–Penrose inverse $\Sigma^\dagger_{XX}$ is well defined.

The next proposition shows that the regression functionals $E(\langle\psi, Y\rangle_{\Omega_Y} \mid X)$ and $E\{\langle g, \kappa_Y(\cdot, Y)\rangle_{\mathcal{H}_Y} \mid X\}$ can be induced by the regression operators $m_{E(Y|X)}$ and $m_{Y|X}$. Its proof is similar to Li and Song ((2017), Proposition 1) and is omitted.

PROPOSITION 1. *Suppose Assumptions 1 and 2 hold. Then for any $\psi \in \Omega_Y$ and $g \in \mathcal{H}_Y$, there exist constants c_ψ and c_g, such that*

$$\begin{aligned}
E\big(\langle\psi, Y - \mu_Y\rangle_{\Omega_Y} \mid X\big) &= m_{E(Y|X)}\psi + c_\psi, \quad \textit{and}\\
E\big\{\big\langle g, \kappa_Y(\cdot, Y) - m_Y\big\rangle_{\mathcal{H}_Y} \mid X\big\} &= m_{Y|X}g + c_g.
\end{aligned}$$

3. Average Fréchet derivative estimators. In this section, we first establish the connection between the Fréchet derivative of the regression functional and the functional central mean or central subspace. Built on this connection, we introduce the average Fréchet derivative estimators, and develop their population-level properties.

3.1. *Fréchet derivative of regression functionals.* We first derive a key property for our proposal: the Riesz representation of the Fréchet derivative of the regression functional is located within the dimension reduction spaces of interest. Let $M_\psi(\cdot) = E(\langle \psi, Y\rangle_{\Omega_Y} \mid X = \cdot)$ for any $\psi \in \Omega_Y$, and $M_g(\cdot) = E\{g(Y) \mid X = \cdot\}$ for any $g \in \mathcal{H}_Y$, both of which are mappings from $\Omega_X \to \mathbb{R}$. Let $\partial_x M_\psi(x, \cdot)$ and $\partial_x M_g(x, \cdot)$ denote the Fréchet derivative of M_ψ and M_g, respectively, at any $x \in \Omega_X$. If $A : \mathcal{H} \to \mathbb{R}$ is a bounded linear mapping, then, by Riesz representation theorem, there exists a unique $g \in \mathcal{H}$, such that $Af = \langle g, f\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. In other words, g is the Riesz representation of A in $\mathcal{H}$ (Conway (2010)).

THEOREM 2. *If M_ψ and M_g are Fréchet differentiable for any $\psi \in \Omega_Y$ and $g \in \mathcal{H}_Y$, then*:

(a) *for any $\psi \in \Omega_Y$, the Riesz representation of $\partial_x M_\psi(x, \cdot)$ is in $\mathcal{S}_{E(Y|X)}$;*
(b) *for any $g \in \mathcal{H}_Y$, the Riesz representation of $\partial_x M_g(x, \cdot)$ is in $\mathcal{S}_{Y|X}$.*

Its proof is given in Section S4 of the Supplementary Material (Lee and Li (2022)). Theorem 2 shares the same spirit as the gradient-based SDR methods such as Xia et al. (2002), Yin and Li (2011), but is stated in a more general setting. It shows that the Riesz representations of the Fréchet derivatives $\partial_x M_\psi(x, \cdot)$ and $\partial_x M_g(x, \cdot)$ are elements of the functional dimension reduction spaces. Next, we show that such representations have closed forms.

We introduce an additional assumption, which is mild and is satisfied for many standard kernels κ_X, with the ρ function in (3) properly chosen, for example, the radial basis kernel and the polynomial kernel. In Section S1 of the Supplementary Material (Lee and Li (2022)), we further give the sufficient condition under which this assumption holds, and the explicit form of the derivative of κ_X for two commonly used kernels, the radial basis kernel and the polynomial kernel.

ASSUMPTION 3. $\kappa_X(\cdot, \cdot)$ is continuously differentiable.

We next establish the interchangeability between the Fréchet derivative and the reproducing kernel Hilbert space for the general case. Its proof is given in Section S4 of the Supplementary Material (Lee and Li (2022)).

PROPOSITION 2. *Suppose Assumption 3 holds. Then, for any $f \in \mathcal{H}_X$ such that $f(x) = \langle f, \kappa_X(x, \cdot)\rangle_{\Omega_X}$, $x \in \Omega_X$, the Fréchet derivative of $f(x)$ at x, $\partial_x f(x, \cdot) : \Omega_X \to \mathbb{R}$, $u \mapsto \partial_x f(x, u)$, is equal to $\partial_x f(x, u) = \langle f, \partial_x \kappa_X(x, u)\rangle_{\mathcal{H}_X}$.*

Steinwart and Christmann (2008) established the differentiability of RKHS in the random variable setting. Proposition 2 extends their result to the functional setting. See also Fukumizu and Leng (2014) for a simpler version of this result in the random variable setting. Note that the regression functionals $M_\psi(\cdot)$ and $M_g(\cdot)$ are members in $\mathcal{H}_X$, and thus by Proposition 2, their Fréchet derivatives $\partial_x M_\psi(x, \cdot)$ and $\partial_x M_g(x, \cdot)$ can be calculated using the derivative of κ_X. By Proposition 1, this further implies that, via the mappings of the regression operators, we can explicitly compute $\partial_x M_\psi(x, \cdot)$ and $\partial_x M_g(x, \cdot)$, as well as their Riesz representations. The next theorem summarizes the above statement. Hereafter, we write $\partial_x \kappa(x, u)$ as $\{\partial_x \kappa(x)\} \circ u$, and $\partial_x \kappa(x) : \Omega_X \to \mathcal{H}_X$ is a member in $B(\Omega_X, \mathcal{H}_X)$.

THEOREM 3. *Suppose Assumptions 1 to 3 hold. Then, for any $x \in \Omega_X$,*

$$(a)\quad \overline{\operatorname{ran}}\big[\{\partial_x \kappa_X(x)\}^* m_{E(Y|X)}\big] \subseteq \mathcal{S}_{E(Y|X)}; \qquad (b)\quad \overline{\operatorname{ran}}\big[\{\partial_x \kappa_X(x)\}^* m_{Y|X}\big] \subseteq \mathcal{S}_{Y|X}.$$

Theorem 3 shows that $\overline{\text{ran}}[\{\partial_x \kappa_X(x)\}^* m_{E(Y|X)}]$ and $\overline{\text{ran}}[\{\partial_x \kappa_X(x)\}^* m_{Y|X}]$ are subspaces of the functional central mean and functional central subspaces, respectively, which provides the essential foundation for our functional SDR estimation. In particular, this result does *not* require any distributional or structural assumptions, such as the linearity condition or the constant variance condition, that are required in all existing inverse moment-based functional SDR methods (Ferré and Yao (2003), Hsing and Ren (2009), Jiang, Yu and Wang (2014), Wang, Lin and Zhang (2013), Yao, Lei and Wu (2015)). To gain further insight about Theorem 3, we observe that $\{\partial_x \kappa_X(x)\}^* m_{Y|X}$ has an intuitive explanation that can be linked to the classical linear regression coefficient. Consider a linear kernel $\kappa_X = \langle \phi, \phi' \rangle_{\Omega_X}$ for any $\phi, \phi' \in \Omega_X$. Suppose $\Omega_X = \mathbb{R}^p$, $\Omega_Y = \mathbb{R}$, Y and X are centered and satisfy that $E(Y \mid X = x) = \langle \beta, x \rangle = \beta^\mathsf{T} x$ with $\beta \in \mathbb{R}^p$. Then for any $x' \in \mathbb{R}^p$, we have $\langle \{\partial_x \kappa_X(x)\}^* m_{E(Y|X)}, x' \rangle = \langle \beta^\mathsf{T} x, \{\partial_x \kappa_X(x)\} \circ x' \rangle$, which is equal to $\langle \beta^\mathsf{T} x, (x')^\mathsf{T} x \rangle_{\mathcal{H}_X} = \beta^\mathsf{T} x'$. This implies that $\{\partial_x \kappa_X(x)\}^* m_{E(Y|X)} = \beta$.

3.2. *Average Fréchet derivative.* Built on Theorem 3, let $B = \{\partial_x \kappa_X(x)\}^* m_{E(Y|X)}$ or $\{\partial_x \kappa_X(x)\}^* m_{Y|X}$. Because $\overline{\text{ran}}(B) = \overline{\text{ran}}(BB^*)$, we can define two mappings that map X in Ω_X to the random operators $F_{E(Y|X)}(X)$ and $F_{Y|X}(X)$ in $B(\Omega_X)$, respectively, via

$$F_{E(Y|X)}(x) = \{\partial_x \kappa_X(x)\}^* m_{E(Y|X)} m_{E(Y|X)}^* \{\partial_x \kappa_X(x)\},$$
$$F_{Y|X}(x) = \{\partial_x \kappa_X(x)\}^* m_{Y|X} m_{Y|X}^* \{\partial_x \kappa_X(x)\}.$$

Note that the adjoints of $m_{E(Y|X)}$ and $m_{Y|X}$ are defined on $\text{ran}(\Sigma_{XX})$ instead of $\mathcal{H}_X$. Nonetheless, because $m_{E(Y|X)}$ and $m_{Y|X}$ are both bounded, their domains can be extended to $\overline{\text{ran}}(\Sigma_{XX})$, which, by Assumption 2, is equal to $\mathcal{H}_X$. Next, we establish the existence of the expectations of $F_{E(Y|X)}(X)$ and $F_{Y|X}(X)$. We first need a condition on the moments of the Fréchet derivative of κ_X. In Section S1 of the Supplementary Material (Lee and Li (2022)), we give further results based on which we can show this moment condition holds. We also derive the explicit bounds of $E\|\partial_x \kappa_X(X)\|^4$ for the radial basis kernel and the polynomial kernel.

ASSUMPTION 4. $E\|\partial_x \kappa_X(X)\|^4 < \infty$.

The next proposition establishes the existence of the expectations of $F_{E(Y|X)}(X)$ and $F_{Y|X}(X)$. Its proof immediately follows by the representation theorem, and is omitted.

PROPOSITION 3. *Suppose Assumptions 1 to 4 hold. Then the mean elements of* $F_{E(Y|X)}(X)$ *and* $F_{Y|X}(X)$ *in* $B(\Omega_X)$, *denoted by* $E\{F_{E(Y|X)}\}$ *and* $E(F_{Y|X})$, *respectively, uniquely exist via the following relations*: *for any* $\phi, \phi' \in \Omega_X$,

$$\langle \phi, E\{F_{E(Y|X)}\}\phi' \rangle_{\Omega_X} = E[\langle \phi, \{F_{E(Y|X)}(X)\}\phi' \rangle_{\Omega_X}],$$
$$\langle \phi, E(F_{Y|X})\phi' \rangle_{\Omega_X} = E[\langle \phi, \{F_{Y|X}(X)\}\phi' \rangle_{\Omega_X}].$$

The two operators, $E\{F_{E(Y|X)}\}$ and $E(F_{Y|X})$, share the same spirit as the average derivative estimator of Hardle (1989) in the classical random variable SDR setting, and are generalized to the functional setting. For this reason, we refer $E\{F_{E(Y|X)}\}$ and $E(F_{Y|X})$ as the average Fréchet derivatives (AFD).

In SDR for random variables, we say a basis matrix is an unbiased estimator for the central subspace if the spanning space of this basis matrix is located within the central subspace. In addition, we say the basis matrix is exhaustive if the spanning space recovers the entire central subspace (Li and Wang (2007)). The same definitions apply to the central mean subspace. We now extend these concepts to the functional setting.

DEFINITION 4. Let $\mathcal{S}$ be a subspace of Ω_X.

(a) We say $\mathcal{S}$ is unbiased for $\mathcal{S}_{E(Y|X)}$ if $\mathcal{S} \subseteq \mathcal{S}_{E(Y|X)}$, and exhaustive for $\mathcal{S}_{E(Y|X)}$ if $\mathcal{S} = \mathcal{S}_{E(Y|X)}$;
(b) We say $\mathcal{S}$ is unbiased for $\mathcal{S}_{Y|X}$ if $\mathcal{S} \subseteq \mathcal{S}_{Y|X}$, and exhaustive for $\mathcal{S}_{Y|X}$ if $\mathcal{S} = \mathcal{S}_{Y|X}$.

Next, we show that $E\{F_{E(Y|X)}\}$ and $E(F_{Y|X})$ are unbiased and exhaustive estimators for $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{Y|X}$, respectively. We add another condition, and discuss it in detail in Section S1 of the Supplementary Material (Lee and Li (2022)).

ASSUMPTION 5. The supports, $\operatorname{supp}(P_{\mathcal{S}_{E(Y|X)}} X)$ and $\operatorname{supp}(P_{\mathcal{S}_{Y|X}} X)$, are convex.

THEOREM 4. *Suppose Assumptions 1 to 4 hold. Then*:

(a) *Unbiasedness*: $\overline{\operatorname{ran}}[E\{F_{E(Y|X)}\}] \subseteq \mathcal{S}_{E(Y|X)}$, *and* $\overline{\operatorname{ran}}[E(F_{Y|X})] \subseteq \mathcal{S}_{Y|X}$.
(b) *Exhaustiveness*: $\overline{\operatorname{ran}}[E\{F_{E(Y|X)}\}] = \mathcal{S}_{E(Y|X)}$, *and* $\overline{\operatorname{ran}}[E\{F_{Y|X}\}] = \mathcal{S}_{Y|X}$, *if Assumption 5 further holds*.

Theorem 4 suggests that we can use the space spanned by $E\{F_{E(Y|X)}\}$ or $E(F_{Y|X})$ to estimate the functional central mean subspace or the functional central subspace. In other words, we can use the leading eigenfunctions of these operators to estimate the bases of $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{Y|X}$. The next corollary summarizes this observation, which provides the estimators for our functional SDR spaces at the population level. To ensure that we can compute the two functional spaces in Theorem 4, $\overline{\operatorname{ran}}[E(F_{E(Y|X)})]$ and $\overline{\operatorname{ran}}[E(F_{Y|X})]$, at the sample level, we let the rank of $E\{F_{E(Y|X)}\}$ and $E\{F_{Y|X}\}$ equal q, while we allow q to diverge.

COROLLARY 1. *Suppose the conditions in Theorem 4 hold.*

(a) *If* $\phi_1', \ldots, \phi_q'$ *are the solutions to the sequential optimization problem*

$$\max\big[\langle \phi, E\{F_{E(Y|X)}\}\phi\rangle_{\Omega_X} : \phi \in \Omega_X\big]$$
$$\text{subject to } \|\phi\|_{\Omega_X} = 1, \langle \phi, \phi_1'\rangle_{\Omega_X} = \cdots = \langle \phi, \phi_q'\rangle_{\Omega_X} = 0,$$

then $\mathcal{S}_{E(Y|X)} = \operatorname{Span}\{\phi_1', \ldots, \phi_q'\}$.
(b) *If* $\phi_1', \ldots, \phi_q'$ *are the solutions to the sequential optimization problem*

$$\max\big[\langle \phi, E(F_{Y|X})\phi\rangle_{\Omega_X} : \phi \in \Omega_X\big]$$
$$\text{subject to } \|\phi\|_{\Omega_X} = 1, \langle \phi, \phi_1'\rangle_{\Omega_X} = \cdots = \langle \phi, \phi_q'\rangle_{\Omega_X} = 0,$$

then $\mathcal{S}_{Y|X} = \operatorname{Span}\{\phi_1', \ldots, \phi_q'\}$.

4. Sample estimation. In this section, we develop the functional SDR estimation procedure, first at the operator level, then under a coordinate system. We also develop a postreduction prediction method based on the proposed dimension reduction framework.

4.1. *Estimation at the operator level.* Let $\{(X^k, Y^k)^\top\}_{k=1}^n$ denote i.i.d. samples of $(X, Y)^\top$. We estimate the mean elements μ_Y, m_X and m_Y of Ω_Y, $\mathcal{H}_X$ and $\mathcal{H}_Y$ by $\hat{\mu}_Y = E_n(Y)$, $\hat{m}_X = E_n\{\kappa_X(\cdot, X)\}$, and $\hat{m}_Y = E_n\{\kappa_Y(\cdot, Y)\}$, respectively, where $E_n(\cdot)$ is the mean operator such that $E_n(\omega) = n^{-1}\sum_{k=1}^n \omega^k$ for the samples $(\omega^1, \ldots, \omega^n)$. We estimate the three key covariance operators by

$$\hat{\Lambda}_{XY} = E_n\big[\{\kappa_X(\cdot, X) - \hat{m}_X\} \otimes (Y - \hat{\mu}_Y)\big],$$
$$\hat{\Sigma}_{XY} = E_n\big[\{\kappa_X(\cdot, X) - \hat{m}_X\} \otimes \{\kappa_Y(\cdot, Y) - \hat{m}_Y\}\big],$$
$$\hat{\Sigma}_{XX} = E_n\big[\{\kappa_X(\cdot, X) - \hat{m}_X\} \otimes \{\kappa_X(\cdot, X) - \hat{m}_X\}\big],$$

where $\otimes$ represents the tensor product. Besides, we estimate Λ_{YY} and Σ_{YY} by $\hat{\Lambda}_{YY} = E_n\{(Y - \hat{\mu}_Y) \otimes (Y - \hat{\mu}_Y)\}$, and $\hat{\Sigma}_{YY} = E_n[\{\kappa_Y(\cdot, Y) - \hat{m}_Y\} \otimes \{\kappa_Y(\cdot, Y) - \hat{m}_Y\}]$, respectively. For any $\psi, \psi' \in \Omega_Y$, we have $\langle \psi, \hat{\Lambda}_{YY}\psi' \rangle_{\Omega_Y} = \mathrm{cov}_n\{\langle \psi, Y\rangle_{\Omega_Y}, \langle \psi', Y\rangle_{\Omega_Y}\}$, and the rest of the sample covariance operators have similar properties, where $\mathrm{cov}_n(\cdot, \cdot)$ denotes the sample covariance between the designated quantities.

In the next section, we show that these sample covariance operators can be represented as functions of Gram kernel matrices, which often have fast decaying eigenvalues and can be well approximated by the spectral decomposition associated with only the leading eigenvalues (Lee and Huang (2007)). This suggests that using a reduced set of bases of RKHS can improve the estimation efficiency both statistically and computationally. We adopt this idea in our functional SDR estimation. Let $\{(\lambda^k, \psi^k)\}_{k\in\mathbb{N}}$, $\{(a^k, f^k)\}_{k\in\mathbb{N}}$, and $\{(b^k, g^k)\}_{k\in\mathbb{N}}$ denote the pairs of eigenvalues and eigenfunctions of Λ_{YY}, Σ_{XX}, and Σ_{YY}, such that

$$\Lambda_{YY} = \sum_{k\in\mathbb{N}} \lambda^k(\psi^k \otimes \psi^k), \qquad \Sigma_{XX} = \sum_{k\in\mathbb{N}} a^k(f^k \otimes f^k), \qquad \Sigma_{YY} = \sum_{k\in\mathbb{N}} b^k(g^k \otimes g^k).$$

Let $\{\theta^k, c^k, d^k\}_{k\in\mathbb{N}}$, denote the Karhunen–Loève coefficients satisfying that,

$$Y - \mu_Y = \sum_{k\in\mathbb{N}} \theta^k\psi^k, \qquad \kappa_X(\cdot, X) - m_X = \sum_{k\in\mathbb{N}} c^k f^k, \qquad \kappa_Y(\cdot, Y) - m_Y = \sum_{k\in\mathbb{N}} d^k g^k.$$

That is, $\theta^k = \langle Y - \mu_Y, \psi^k\rangle_{\Omega_Y}$, $c^k = \langle \kappa_X(\cdot, X) - m_X, f^k\rangle_{\mathcal{H}_X}$ and $d^k = \langle \kappa_Y(\cdot, Y) - m_Y, g^k\rangle_{\mathcal{H}_Y}$. By definition, the covariance operators Λ_{XY}, Σ_{XX}, and Σ_{XY} can be written as

$$\begin{aligned} \Lambda_{XY} &= \sum_{k,\ell\in\mathbb{N}} \mathrm{cov}(c^k, \theta^\ell)(f^k \otimes \psi^k), \qquad \Sigma_{XY} = \sum_{k,\ell\in\mathbb{N}} \mathrm{cov}(c^k, d^\ell)(f^k \otimes g^\ell), \\ \Sigma_{XX} &= \sum_{k} \mathrm{var}(c^k)(f^k \otimes f^k). \end{aligned} \tag{4}$$

Similarly, at the sample level, let $\{(\hat{\lambda}^k, \hat{\psi}^k)\}_{k\in\mathbb{N}}$, $\{(\hat{a}^k, \hat{f}^k)\}_{k\in\mathbb{N}}$ and $\{(\hat{b}^k, \hat{g}^k)\}_{k\in\mathbb{N}}$ denote the pairs of eigenvalues and eigenfunctions of $\hat{\Lambda}_{YY}$, $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{YY}$. Let $\{\hat{\theta}^k, \hat{c}^k, \hat{d}^k\}_{k\in\mathbb{N}}$ denote the estimated KL coefficients, $\hat{\theta}^k = \langle Y - \hat{\mu}_Y, \hat{\psi}^k\rangle_{\Omega_Y}$, $\hat{c}^k = \langle \kappa_X(\cdot, X) - \hat{m}_X, \hat{f}^k\rangle_{\mathcal{H}_X}$ and $\hat{d}^k = \langle \kappa_Y(\cdot, Y) - \hat{m}_Y, \hat{g}^k\rangle_{\mathcal{H}_Y}$. To estimate the covariance operators in (4), we then truncate and focus on the leading d terms, and obtain the estimators via

$$\begin{aligned} \hat{\Lambda}^d_{XY} &= \sum_{k,\ell=1}^{d} \mathrm{cov}_n(\hat{c}^k, \hat{\theta}^\ell)(\hat{f}^k \otimes \hat{\psi}^\ell), \qquad \hat{\Sigma}^d_{XY} = \sum_{k,\ell=1}^{d} \mathrm{cov}_n(\hat{c}^k, \hat{d}^\ell)(\hat{f}^k \otimes \hat{g}^\ell), \\ \hat{\Sigma}^d_{XX} &= \sum_{k=1}^{d} \mathrm{var}_n(\hat{c}^k)(\hat{f}^k \otimes \hat{f}^k), \end{aligned} \tag{5}$$

where $\mathrm{var}_n(\omega) = \mathrm{cov}_n(\omega, \omega)$.

Correspondingly, we estimate the functional regression operators $m_{E(Y|X)}$ and $m_{Y|X}$ by

$$\hat{m}^d_{E(Y|X)} = \big(\hat{\Sigma}^d_{XX} + \epsilon I\big)^{-1}\hat{\Lambda}^d_{XY}, \qquad \hat{m}^d_{Y|X} = \big(\hat{\Sigma}^d_{XX} + \epsilon I\big)^{-1}\hat{\Sigma}^d_{XY}, \tag{6}$$

respectively, where ϵ is a ridge parameter, and $I : \mathcal{H}_X \to \mathcal{H}_X$ is the identify mapping. Then, by Corollary 1, we solve the optimization problem

$$\max\big[\langle \phi, E_n\{\hat{F}^d(X)\}\phi\rangle_{\Omega_X} : \phi \in \Omega_X\big], \tag{7}$$

where $\hat{F}^d(x) = \hat{F}^d_{E(Y|X)}(x) = \{\partial_x\kappa_X(x)\}^*\hat{m}^d_{E(Y|X)}(\hat{m}^d_{E(Y|X)})^*\{\partial_x\kappa_X(x)\}$ when estimating $\mathcal{S}_{E(Y|X)}$, and $\hat{F}^d(x) = \hat{F}^d_{Y|X}(x) = \{\partial_x\kappa_X(x)\}^*\hat{m}^d_{Y|X}(\hat{m}^d_{Y|X})^*\{\partial_x\kappa_X(x)\}$ when estimating $\mathcal{S}_{Y|X}$. That is, we estimate $\mathcal{S}_{E(Y|X)}$ or $\mathcal{S}_{Y|X}$ by $\mathrm{Span}\{\hat{\phi}_k : k = 1, \ldots, q\}$, where $\hat{\phi}_k$ is the eigenfunction corresponding to the kth largest eigenvalue of $E_n\{\hat{F}^d(X)\}$ in (7). Lastly, we estimate the sufficient predictors $\langle \phi_k, X\rangle_{\Omega_X}$ by $\langle \hat{\phi}_k, X\rangle_{\Omega_X}$, for $k = 1, \ldots, q$.

4.2. *Estimation under a coordinate system.* In practice, a function can only be observed at a finite set of points. For simplicity, we focus on the balanced setting where all the subjects are observed under a common set of time points, and the observed data are $\{[X^k(t), Y^k(t)]^\mathsf{T} : t = t_1, \ldots, t_m, k = 1, \ldots, n\}$. Meanwhile, our method can be straightforwardly extended to the unbalanced setting with only minor modifications. We next develop an estimation procedure under a chosen coordinate system given the observed data.

We first introduce the notion of coordinate representation here, and give more details in Section S2.1 of the Supplementary Material (Lee and Li (2022)). Let $\lfloor f \rfloor_{\mathcal{B}} \in \mathbb{R}^m$ denote the coordinate of f in a generic Hilbert space $\mathcal{H}$ with respect to its bases $\mathcal{B}$. Moreover, let ${}_{\mathcal{B}'}\lfloor A \rfloor_{\mathcal{B}}$ denote the coordinate of a linear operator $A : \mathcal{H} \to \mathcal{H}'$ with respect to $\mathcal{B}$ and $\mathcal{B}'$, where $\mathcal{H}'$ is another Hilbert space spanned by $\mathcal{B}'$. For convenience, we write $\lfloor f \rfloor_{\mathcal{B}}$ as $\lfloor f \rfloor$, and ${}_{\mathcal{B}'}\lfloor A \rfloor_{\mathcal{B}}$ as $\lfloor A \rfloor$ when there is no confusion.

We divide our coordinate-level estimation procedure into four major steps.

In Step 1, we first construct the RKHS, Ω_X and Ω_Y, which are Hilbert spaces of functions on an interval $T \subseteq \mathbb{R}$. We begin with a positive definite kernel function, $\kappa_T : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. There are many choices, for example, the Brownian motion kernel, or the radial basis kernel. Suppose $\Omega_X = \Omega_Y$ is the linear span of a sequence of functional basis $\{\kappa_T(\cdot, t_s)\}_{s=1}^m$.

Next, we derive the coordinate representation of the observed data $X^k(t)$ and $Y^k(t)$. Let $X^k(\tau) = \{X^k(t_1), \ldots, X^k(t_m)\}^\mathsf{T}$, and $Y^k(\tau) = \{Y^k(t_1), \ldots, Y^k(t_m)\}^\mathsf{T}$, for the kth subject, $k = 1, \ldots, n$, and $(K_T)_{s,u=1}^m = \kappa_T(t_s, t_u)$ be the $m \times m$ Gram matrix of T with the spectral decomposition, $K_T = U_T D_T U_T^\mathsf{T}$. It is then straightforward to obtain that

$$\mathcal{B}_T(\cdot) = D_T^{-1/2} U_T^\mathsf{T} [\kappa_T(\cdot, t_1), \ldots, \kappa_T(\cdot, t_m)]^\mathsf{T}$$

is an orthonormal basis of Ω_X and Ω_Y. Therefore, we can write $X^k(\tau)$ and $Y^k(\tau)$ as $X^k(\tau) = U_T D_T^{1/2} \lfloor X^k \rfloor_{\mathcal{B}_T}$, and $Y^k(\tau) = U_T D_T^{1/2} \lfloor Y^k \rfloor_{\mathcal{B}_T}$, which further leads to the their coordinate representations,

$$\lfloor X^k \rfloor_{\mathcal{B}_T} = D_T^{-1/2} U_T^\mathsf{T} X^k(\tau), \qquad \lfloor Y^k \rfloor_{\mathcal{B}_T} = D_T^{-1/2} U_T^\mathsf{T} Y^k(\tau). \tag{8}$$

Note that if K_T is not of a full rank, we simply use its leading nonzero eigenvalues and corresponding eigenfunctions to construct $\mathcal{B}_T$.

In Step 2, we first construct the nested RKHS, $\mathcal{H}_X$ and $\mathcal{H}_Y$, which are Hilbert spaces of functions on X and Y, respectively. We choose positive definite nested kernels κ_X and κ_Y, using, for example, the radial basis function kernel, or the Laplacian kernel. We then follow the coordinates derived in (8), and estimate the inner products by $\langle X^k, X^\ell \rangle_{\Omega_X} = \lfloor X^k \rfloor_{\mathcal{B}_T}^\mathsf{T} \lfloor X^\ell \rfloor_{\mathcal{B}_T}$, and $\langle Y^k, Y^\ell \rangle_{\Omega_X} = \lfloor Y^k \rfloor_{\mathcal{B}_T}^\mathsf{T} \lfloor Y^\ell \rfloor_{\mathcal{B}_T}$. Let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be the linear span of $\{\kappa_X(\cdot, X^k) - E_n \kappa_X(\cdot, X)\}_{k=1}^n$ and $\{\kappa_Y(\cdot, Y^k) - E_n \kappa_Y(\cdot, Y)\}_{k=1}^n$, respectively.

Next, we construct the orthonormal bases of $\mathcal{H}_X$ and $\mathcal{H}_Y$, following a similar way as in Step 1. That is, let $(K_X)_{k,\ell=1}^n = \kappa_X(X^k, X^\ell)$ and $(K_Y)_{k,\ell=1}^n = \kappa_Y(Y^k, Y^\ell)$ be the $n \times n$ Gram matrix of X and Y, respectively, and $G_X = Q_n K_X Q_n$ and $G_Y = Q_n K_Y Q_n$ be their centered version, with $Q_n = I_n - n^{-1} 1_n^\mathsf{T} 1_n$ and 1_n being the n-dimensional vector of ones. We then perform the spectral decomposition, $G_X = U_X D_X U_X^\mathsf{T} + U_X^0 D_X^0 (U_X^0)^\mathsf{T}$, and $G_Y = U_Y D_Y U_Y^\mathsf{T} + U_Y^0 D_Y^0 (U_Y^0)^\mathsf{T}$, where $U_X D_X U_X^\mathsf{T}$ and $U_Y D_Y U_Y^\mathsf{T}$ correspond to the first d eigenvalues, and $U_X^0 D_X^0 (U_X^0)^\mathsf{T}$ and $U_Y^0 D_Y^0 (U_Y^0)^\mathsf{T}$ correspond to the last $n - d$ eigenvalues. Therefore, we construct the orthonormal basis of $\mathcal{H}_X$ and $\mathcal{H}_Y$, respectively, via

$$\begin{aligned} \mathcal{C}_X(\cdot) &= D_X^{-1/2} U_X^\mathsf{T} [\kappa_X(\cdot, X^1) - \hat{m}_X, \ldots, \kappa_X(\cdot, X^n) - \hat{m}_X]^\mathsf{T}, \\ \mathcal{C}_Y(\cdot) &= D_Y^{-1/2} U_Y^\mathsf{T} [\kappa_Y(\cdot, Y^1) - \hat{m}_Y, \ldots, \kappa_Y(\cdot, Y^n) - \hat{m}_Y]^\mathsf{T}. \end{aligned} \tag{9}$$

In Step 3, we derive the coordinate representation of the truncated sample covariance operators in (5), given the constructed Ω_X, Ω_Y, $\mathcal{H}_X$ and $\mathcal{H}_Y$, and their orthonormal bases.

Algorithm 1 Functional sufficient dimension reduction estimation

1: Choose the kernel κ_T, construct $\Omega_X = \Omega_Y = \text{Span}\{\kappa_T(\cdot, t_s)\}_{s=1}^m$ and their orthonormal basis $\mathcal{B}_T$, and compute the coordinates $\lfloor X^k \rfloor_{\mathcal{B}_T}$ and $\lfloor Y^k \rfloor_{\mathcal{B}_T}$ following (8).

2: Choose the nested kernels κ_X and κ_Y, construct $\mathcal{H}_X$, $\mathcal{H}_Y$, and their orthonormal bases following (9).

3: Compute the coordinates ${}_{\mathcal{C}_X}\lfloor \hat{\Lambda}_{XY}^d \rfloor_{\mathcal{B}_Y}$, ${}_{\mathcal{C}_X}\lfloor \hat{\Sigma}_{XY}^d \rfloor_{\mathcal{C}_Y}$, and ${}_{\mathcal{C}_X}\lfloor \hat{\Sigma}_{XX}^d \rfloor_{\mathcal{C}_X}$ following (10).

4: Compute the coordinates ${}_{\mathcal{B}_T}\lfloor \hat{F}^d(X) \rfloor_{\mathcal{B}_T}$ following (11), compute the sample average $E_n\{{}_{\mathcal{B}_T}\lfloor \hat{F}^d(X) \rfloor_{\mathcal{B}_T}\}$, obtain its eigenfunctions $\lfloor \hat{\phi}_1 \rfloor, \ldots, \lfloor \hat{\phi}_q \rfloor$, and estimate $\mathcal{S}_{E(Y|X)}$ or $\mathcal{S}_{Y|X}$ by $\text{Span}\{\hat{\phi}_1, \ldots, \hat{\phi}_q\}$, where $\hat{\phi}_i = \lfloor \hat{\phi}_i \rfloor^{\mathsf{T}}\{\mathcal{B}_T(\cdot)\}$, $i = 1, \ldots, q$.

Specifically, we perform the spectral decomposition on $E_n(\lfloor Y^k - \hat{\mu}_Y \rfloor_{\mathcal{B}_T} \lfloor Y^k - \hat{\mu}_Y \rfloor_{\mathcal{B}_T}^{\mathsf{T}})$, and denote its first d eigenvalues and eigenfunctions as $\{(\hat{\lambda}^k, \lfloor \hat{\psi}^k \rfloor_{\mathcal{B}_T})\}_{k=1}^d$. Let $\mathcal{B}_Y = \{\hat{\psi}^k : k = 1, \ldots, d\}$. We then compute the leading sample KL coefficients by, for $k = 1, \ldots, d$, $\ell = 1, \ldots, n$,

$$\hat{\theta}^{k,\ell} = \lfloor Y^\ell - \hat{\mu}_Y \rfloor_{\mathcal{B}_T}^{\mathsf{T}} \lfloor \hat{\psi}^k \rfloor_{\mathcal{B}_T}, \qquad \hat{c}^{k,\ell} = e_k^{\mathsf{T}}\{\mathcal{C}_X(X^\ell) - E_n\mathcal{C}_X(X)\},$$
$$\hat{d}^{k,\ell} = e_k^{\mathsf{T}}\{\mathcal{C}_Y(Y^\ell) - E_n\mathcal{C}_Y(Y)\},$$

where e_k is the d-dimensional vector with its kth position being one and the rest being zero. Stack the sample KL coefficients $\theta^{k,\ell}$, $\hat{c}^{k,\ell}$, and $\hat{d}^{k,\ell}$ vertically, and form the n-dimensional vectors $\hat{\Theta}^k = (\hat{\theta}^{k,1}, \ldots, \hat{\theta}^{k,n})^{\mathsf{T}}$, $\hat{C}^k = (\hat{c}^{k,1}, \ldots, \hat{c}^{k,n})^{\mathsf{T}}$, and $\hat{D}^k = (\hat{d}^{k,1}, \ldots, \hat{d}^{k,n})^{\mathsf{T}}$, for $k = 1, \ldots, d$. Next stack $\hat{\Theta}^k$, $\hat{C}^k$ and $\hat{D}^k$ horizontally, and form the $n \times d$ matrices $\hat{\Theta} = (\hat{\Theta}^1, \ldots, \hat{\Theta}^d)$, $\hat{C} = (\hat{C}^1, \ldots, \hat{C}^d)$ and $\hat{D} = (\hat{D}^1, \ldots, \hat{D}^d)$. Then we obtain the coordinate representation of the truncated sample covariance operators in (5) as

$$\begin{aligned} &{}_{\mathcal{C}_X}\lfloor \hat{\Lambda}_{XY}^d \rfloor_{\mathcal{B}_Y} = n^{-1}\hat{C}^{\mathsf{T}}\hat{\Theta}, \qquad {}_{\mathcal{C}_X}\lfloor \hat{\Sigma}_{XY}^d \rfloor_{\mathcal{C}_Y} = n^{-1}\hat{C}^{\mathsf{T}}D, \\ &{}_{\mathcal{C}_X}\lfloor \hat{\Sigma}_{XX}^d \rfloor_{\mathcal{C}_X} = n^{-1}\hat{C}^{\mathsf{T}}\hat{C}. \end{aligned} \tag{10}$$

In Step 4, we derive the coordinate representation of $\hat{F}^d(X)$ in (7) as

$$ {}_{\mathcal{B}_T}\lfloor \hat{F}^d(X) \rfloor_{\mathcal{B}_T} $$
$$= \begin{cases} \lfloor \partial_x \kappa_X(X) \rfloor^{\mathsf{T}}\{\lfloor \hat{\Sigma}_{XX}^d \rfloor + \epsilon I_d\}^{-1} \lfloor \hat{\Lambda}_{XY}^d \rfloor \lfloor \hat{\Lambda}_{XY}^d \rfloor^{\mathsf{T}}\{\lfloor \hat{\Sigma}_{XX}^d \rfloor + \epsilon I_d\}^{-1} \lfloor \partial_x \kappa_X(X) \rfloor \\ \quad \text{for } \mathcal{S}_{E(Y|X)}, \\ \lfloor \partial_x \kappa_X(X) \rfloor^{\mathsf{T}}\{\lfloor \hat{\Sigma}_{XX}^d \rfloor + \epsilon I_d\}^{-1} \lfloor \hat{\Sigma}_{XY}^d \rfloor \lfloor \hat{\Sigma}_{XY}^d \rfloor^{\mathsf{T}}\{\lfloor \hat{\Sigma}_{XX}^d \rfloor + \epsilon I_d\}^{-1} \lfloor \partial_x \kappa_X(X) \rfloor \\ \quad \text{for } \mathcal{S}_{Y|X}, \end{cases} \tag{11}$$

where $\lfloor \partial_x \kappa_X(X) \rfloor$ is the coordinate of the Fréchet derivative, which we derive explicitly for the radial basis kernel and the polynomial kernel in Section S2.2 of the Supplementary Material (Lee and Li (2022)). We take the value of ϵ to be $10^{-4} \times [D_X]_{1,1}$. Finally, we compute the sample average $E_n\{{}_{\mathcal{B}_T}\lfloor \hat{F}^d(X) \rfloor_{\mathcal{B}_T}\}$, perform its spectral decomposition, and obtain the eigenfunctions $\lfloor \hat{\phi}_1 \rfloor, \ldots, \lfloor \hat{\phi}_q \rfloor$ associated with the leading q eigenvalues. We estimate $\mathcal{S}_{E(Y|X)}$ or $\mathcal{S}_{Y|X}$ by $\text{Span}\{\hat{\phi}_1, \ldots, \hat{\phi}_q\}$, where $\hat{\phi}_i = \lfloor \hat{\phi}_i \rfloor^{\mathsf{T}}\{\mathcal{B}_T(\cdot)\}$, $i = 1, \ldots, q$.

We summarize the above estimation procedure in Algorithm 1.

Our estimation method requires specifying kernel functions for κ_T, κ_X, and κ_Y. We study the effect of kernel choices in Section S3 of the Supplementary Material (Lee and Li (2022)). In general, we have found that the estimated sufficient predictors are not overly sensitive to the choices of kernel functions.

In addition, we need to determine the number of leading KL coefficients d, and the intrinsic dimension q of the functional central subspace or central mean subspace. We adopt the commonly used strategy in principal component analysis (PCA) to select d and q, such that a certain proportion of total variation has been accounted for. Specifically, let $\mathrm{tr}(\cdot)$ and $\lambda_k(\cdot)$ denote the trace and the kth largest eigenvalue of a designated symmetric matrix, and p_d and p_q denote some prespecified proportion, say, 90% or 95%. We then determine d and q by

$$
\begin{aligned}
d &= \min\left\{d' \in \mathbb{N} : \frac{\sum_{k=1}^{d'} \lambda_k(\lfloor \hat{\Sigma}_{XX} \rfloor)}{\mathrm{tr}(\lfloor \hat{\Sigma}_{XX} \rfloor)} \geq p_d \ \text{ and } \ \frac{\sum_{k=1}^{d'} \lambda_k(\lfloor \hat{\Sigma}_{YY} \rfloor)}{\mathrm{tr}(\lfloor \hat{\Sigma}_{YY} \rfloor)} \geq p_d\right\}, \\
q &= \min\left\{q' \in \mathbb{N} : \frac{\sum_{k=1}^{q'} \lambda_k(E_n \lfloor \hat{F}^d(X) \rfloor)}{\mathrm{tr}(E_n \lfloor \hat{F}^d(X) \rfloor)} \geq p_q\right\}.
\end{aligned} \tag{12}
$$

4.3. *Postreduction prediction.* In function-on-function regression, in addition to finding the reduced-dimensional representation, it is of equal interest to predict the response function Y, or any of its functional $g(Y)$ in $\mathcal{H}_Y$, given the predictor function X. Our proposed dimension reduction framework leads naturally to a procedure for such prediction tasks.

Let $\hat{\phi}_i$, $i = 1, \ldots, q$, denote the estimates from Step 5 of Algorithm 1. Then, for any $x \in \Omega_X$, its projections onto $\mathrm{Span}\{\hat{\phi}_k : k = 1, \ldots, q\}$ is $x_q = \sum_{i=1}^{q} (\lfloor x \rfloor^{\mathsf{T}} \lfloor \hat{\phi}_i \rfloor \hat{\phi}_i)$, where $\lfloor x_q \rfloor = \{\lfloor x \rfloor^{\mathsf{T}} \lfloor \hat{\phi}_1 \rfloor, \ldots, \lfloor x \rfloor^{\mathsf{T}} \lfloor \hat{\phi}_q \rfloor\}^{\mathsf{T}}$. Therefore, for any $x, x' \in \Omega_X$, we can define a new kernel κ_X^q for the reduced predictors x_q via

$$
\kappa_X^q(x, x') = \rho\big(\langle x_q, x_q \rangle_{\Omega_X}, \langle x_q, x_q' \rangle_{\Omega_X}, \langle x_q', x_q' \rangle_{\Omega_X}\big) = \rho\big(\lfloor x_q \rfloor^{\mathsf{T}} \lfloor x_q \rfloor, \lfloor x_q \rfloor^{\mathsf{T}} \lfloor x_q' \rfloor, \lfloor x_q' \rfloor^{\mathsf{T}} \lfloor x_q' \rfloor\big),
$$

where ρ is from (3). Based on κ_X^q, we calculate the Gram matrix $K_X^q = [\kappa_X^q(X^k, X^\ell)]_{k,\ell=1}^n$. Suppose we aim to estimate the conditional mean of $g(Y)$ given $X = x$, that is, $E\{g(Y) \mid X = x\}$, for a given g in $\mathcal{H}_Y$. We first compute the coordinates of g with respect to $\mathcal{C}_Y$ by $\lfloor g \rfloor_{\mathcal{C}_Y} = D_Y^{-1/2} U_Y^{\mathsf{T}} \{g(Y^1), \ldots, g(Y^n)\}^{\mathsf{T}}$. We then estimate $E\{g(Y) \mid x\}$ by

$$
\hat{E}\{g(Y) \mid x\} = \big[\{(D_X^q)^{-1/2} (U_X^q)^{\mathsf{T}}\} U_Y D_Y^{1/2} \lfloor g \rfloor_{\mathcal{C}_Y}\big]^{\mathsf{T}} \mathcal{C}_X^q(x),
$$

where D_X^q and U_X^q are from the spectral decomposition $Q_n K_X^q Q_n = U_X^q D_X^q (U_X^q)^{\mathsf{T}}$ corresponding to the leading d eigenvalues, and $\mathcal{C}_X^q(\cdot)$ is the basis with respect to the new kernel κ_X^q, that is, $\mathcal{C}_X^q(\cdot) = D_X^{-1/2} (U_X^q)^{\mathsf{T}} [\kappa_X^q(\cdot, X^1) - E_n \kappa_X^q(\cdot, X), \ldots, \kappa_X^q(\cdot, X^n) - E_n \kappa_X^q(\cdot, X)]^{\mathsf{T}}$.

A special case that is of particular interest is to predict Y given X. Consider the mapping, $g^s : \Omega_Y \to \mathbb{R}$, $y \mapsto g^s(y) = \lfloor y \rfloor_{\mathcal{B}_T, s}$, that is, the sth coordinate of y, for each $s = 1, \ldots, m$. Therefore, we estimate $E\{\lfloor Y \rfloor_{\mathcal{B}_T, s} \mid x\}$ by $\hat{E}\{\lfloor Y \rfloor_{\mathcal{B}_T, s} \mid x\} = [\{(D_X^q)^{-1/2} (U_X^q)^{\mathsf{T}}\} (\lfloor Y^1 \rfloor_{\mathcal{B}_T, s}, \ldots, \lfloor Y^n \rfloor_{\mathcal{B}_T, s})^{\mathsf{T}}]^{\mathsf{T}} \mathcal{C}_X^q(x)$, and estimate $E\{Y(t) \mid x\}$ by

$$
\hat{E}\big\{Y(t) \mid x\big\} = \hat{E}\big[\lfloor Y \rfloor_{\mathcal{B}_T}^{\mathsf{T}} \big\{\mathcal{B}_T(t)\big\} \mid x\big] = \big\{\hat{E}\big(\lfloor Y \rfloor_{\mathcal{B}_T} \mid x\big)\big\}^{\mathsf{T}} \big\{\mathcal{B}_T(t)\big\},
$$

where $\hat{E}(\lfloor Y \rfloor_{\mathcal{B}_T} \mid x) = [\hat{E}\{\lfloor Y \rfloor_{\mathcal{B}_T, 1} \mid x\}, \ldots, \hat{E}\{\lfloor Y \rfloor_{\mathcal{B}_T, m} \mid x\}]^{\mathsf{T}}$.

5. Asymptotic properties. In this section, we establish the consistency and convergence rate of our functional SDR estimator for the functional central subspace $\mathcal{S}_{Y|X}$. The result for the functional central mean subspace $\mathcal{S}_{E(Y|X)}$ can be obtained in a similar fashion, and is thus omitted. We first show the convergence of the truncated sample covariance operators $\hat{\Sigma}_{XX}^d$ and $\hat{\Sigma}_{XY}^d$ in (5). We next show the convergence of the sample functional regression operator $\hat{m}_{Y|X}^d$ in (6). We then establish the convergence of the sample average Fréchet derivative of the regression function, $E_n\{\hat{F}_{Y|X}^d(X)\}$, which leads to the uniform convergence of our functional SDR estimator for $\mathcal{S}_{Y|X}$. We assume the trajectory of the random functions $\{X(t), Y(t)\}^{\mathsf{T}}$ is

fully observed on $t \in T$, and briefly comment on the setting when $\{X(t), Y(t)\}^\mathsf{T}$ is partially observed toward the end. We also derive the consistency and convergence rate of the estimator based on the un-truncated sample covariance operators in Section S2.3 of the Supplementary Material (Lee and Li (2022)). Unless otherwise stated, all proofs are relegated to Section S4 of the Supplementary Material (Lee and Li (2022)).

We begin with some supporting lemmas. The first lemma is about the perturbation of self-adjoint operators. Given $d \in \mathbb{N}$, let $\lambda_1 \geq \cdots \geq \lambda_{d+1}$ denote the top $d+1$ eigenvalues of a self-adjoint operator Σ, and $\nu_d(\Sigma)$ denote the minimum distance between these eigenvalues, $\nu_d(\Sigma) = \min\{\lambda^k - \lambda^{k+1} : k = 1, \ldots, d\}$.

LEMMA 2. *Let Σ and $\hat{\Sigma}$ be self-adjoint operators in $B(\mathcal{H})$, and $\lambda_1 > \lambda_2 > \cdots$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots$ be their sequences of eigenvalues, respectively. If $\nu_d(\Sigma) > 2\|\hat{\Sigma} - \Sigma\|$, then $\max_{k=1,\ldots,d} |\hat{\lambda}^k - \lambda^k| \leq 4\nu_d^{-1}(\Sigma)\, \|\hat{\Sigma} - \Sigma\|$.*

The next lemma provides the convergence rates of the relevant sample operators.

LEMMA 3. *Suppose Assumption 1 holds. Then the orders of magnitude of the terms,*

$$\|\hat{m}_X - m_X\|_{\mathcal{H}_X}, \quad \|\hat{m}_Y - m_Y\|_{\mathcal{H}_Y}, \quad \big\|E_n\{\kappa_X(\cdot, X) - m_X\} \otimes \{\kappa_X(\cdot, X) - m_X\} - \Sigma_{XX}\big\|_{\mathrm{HS}},$$
$$\big\|E_n\{\kappa_Y(\cdot, Y) - m_Y\} \otimes \{\kappa_Y(\cdot, Y) - m_Y\} - \Sigma_{YY}\big\|_{\mathrm{HS}},$$
$$\big\|E_n\{\kappa_X(\cdot, X) - m_X\} \otimes \{\kappa_Y(\cdot, Y) - m_Y\} - \Sigma_{XY}\big\|_{\mathrm{HS}},$$

are $O_P(n^{-1/2})$.

Next, we establish the convergence of the truncated sample covariance operators $\hat{\Sigma}_{XX}^d$ and $\hat{\Sigma}_{XY}^d$ in (5). Since our interest is on $\mathcal{S}_{Y|X}$, not $\mathcal{S}_{E(Y|X)}$, we do not discuss the property of $\hat{\Lambda}_{XY}^d$ here. We introduce two intermediate operators, Σ_{XX}^d and Σ_{XY}^d, which are defined as

$$\Sigma_{XX}^d = \sum_{k=1}^{d} \mathrm{var}(c^k)(f^k \otimes f^k), \qquad \Sigma_{XY}^d = \sum_{k,\ell=1}^{d} \mathrm{cov}(c^k, d^\ell)(f^k \otimes g^\ell).$$

Note that we allow d to grow with the sample size, but denote it as d instead of d_n to simplify the notation. Moreover, for three positive sequences $\{a_n\}$, $\{b_n\}$ and $\{c_n\}$, denote $a_n \prec b_n$ if $a_n = o(b_n)$, $a_n \preceq b_n$ if $a_n = O(b_n)$, and $a_n \wedge b_n = b_n$, or $a_n \vee b_n = a_n$, if $b_n \preceq a_n$. Denote $a_n \asymp b_n$ if both $a_n \preceq b_n$ and $b_n \preceq a_n$. Similarly, denote $a_n \wedge b_n \wedge c_n = (a_n \wedge b_n) \wedge c_n$, and $a_n \vee b_n \vee c_n = (a_n \vee b_n) \vee c_n$.

LEMMA 4. *Suppose Assumption 1 holds, and $\xi \equiv d^2 n^{-1/2}\{\nu_d(\Sigma_{XX}) \wedge \nu_d(\Sigma_{YY})\}^{-1} \preceq 1$. Then $\|\hat{\Sigma}_{XX}^d - \Sigma_{XX}^d\|_{\mathrm{HS}} = O_P(\xi)$, and $\|\hat{\Sigma}_{XY}^d - \Sigma_{XY}^d\|_{\mathrm{HS}} = O_P(\xi)$.*

Next, we establish the convergence of the sample functional regression operator $\hat{m}_{Y|X}^d$. We require two additional smoothness assumptions. We discuss the two assumptions in detail in Section S1 of the Supplementary Material (Lee and Li (2022)).

ASSUMPTION 6. There exist $\beta_1 > 0$ and $m^0 \in B_2(\mathcal{H}_Y, \mathcal{H}_X)$, such that $m_{Y|X} = \Sigma_{XX}^{\beta_1} m^0$.

ASSUMPTION 7. Let a^k and b^k denote the kth eigenvalues of Σ_{XX} and Σ_{YY}, respectively. There exists $\beta_2 > 0$, such that $\sum_{k \geq d+1} a^k \preceq d^{-\beta_2}$, and $\sum_{k \geq d+1} b^k \preceq d^{-\beta_2}$.

THEOREM 5. *Suppose Assumptions 1, 2, 6, 7 hold, and that* $\epsilon \prec 1$. *Then,*

$$\|\hat{m}^d_{Y|X} - m_{Y|X}\|_{\mathrm{HS}} = O_P\{\epsilon^{-1}(\xi \vee d^{-\beta_2}) + \epsilon^{\beta_1}\}.$$

Next, we present a lemma regarding the convergence of linear operators in the form of B^*B. We then establish the convergence of the sample average Fréchet derivative estimator.

LEMMA 5. *Let* $\{B_n\}_{n\geq 1}$ *and* B *be random elements in* $B_2(\mathcal{H}, \mathcal{K})$. *Suppose* $\|B_n - B\|_{\mathrm{HS}} = O_P(a_n)$. *Then* $\|B_n^* B_n - B^* B\|_{\mathrm{HS}} = O_P(a_n)$.

THEOREM 6. *Suppose Assumptions 1 to 7 hold, and that* $\epsilon \prec 1$. *Then,*

$$\|E_n\{\hat{F}^d_{Y|X}(X)\} - E(F_{Y|X})\|_{\mathrm{HS}} = O_P\{\epsilon^{-1}(\xi \vee d^{-\beta_2}) + \epsilon^{\beta_1}\}.$$

We are now ready to establish the uniform convergence of the estimated bases $\{\hat{\phi}_1, \ldots, \hat{\phi}_q\}$ for $\mathcal{S}_{Y|X}$. The proof follows Lemma 2 and Theorem 6, and is omitted.

THEOREM 7. *Let* $\{(\lambda_k^0, \phi_k^0)\}_{k=1}^q$ *denote the eigenvalues and eigenfunctions of* $E(F_{Y|X})$, *with* $\lambda_1^0 > \cdots > \lambda_q^0 > 0$. *Suppose the same conditions in Theorem 6 hold. Then,*

$$\max_{k=1,\ldots,q} \|\hat{\phi}_k - \phi_k^0\|_{\Omega_X} = O_P[\nu_q^{-1}\{E(F_{Y|X})\}\{\epsilon^{-1}(\xi \vee d^{-\beta_2}) + \epsilon^{\beta_1}\}].$$

We make some remarks about Theorems 5 to 7. First, we note that Li and Song (2017) also studied the convergence of the functional regression operator. Our Theorem 5 further extends their result to the setting where the sample covariance operators are based on the truncated bases. In addition, Li and Song (2017) obtained the consistency of their nonlinear basis functions, whereas our Theorem 7 studied the linear basis functions ϕ_k, which involves a different asymptotic analysis.

Second, and more importantly, in all our theorems, we allow the intrinsic dimension q to diverge, while Li and Song (2017) only considered the setting of a fixed q. This difference has profound implications, and makes our asymptotic analysis far from a straightforward extension of Li and Song (2017). Actually, Theorem 7 suggests that q affects the convergence rate of the basis estimates through $\nu_q\{E(F_{Y|X})\}$, that is, the minimal gap between the eigenvalues of $E(F_{Y|X})$. To gain further insight, we note that, by definition, the squared Hilbert-Schmidt norm of $E(F_{Y|X})$ is

$$\sum_{k,\ell=1}^{\infty} \langle \phi_k, E(F_{Y|X})\phi_\ell \rangle^2_{\Omega_X} = \sum_{k,\ell=1}^{\infty} (E[\langle \phi_k, \{F_{Y|X}(X)\}\phi_\ell \rangle_{\Omega_X}])^2,$$

where $\{\phi_k\}_{k=1}^{\infty}$ are a set of orthonormal bases in Ω_X. Following the proof of Theorem 6, the right-hand side of the above equation is bounded. This implies that $E(F_{Y|X})$ is Hilbert-Schmidt, which further indicates that its eigenvalues decay to zero. Therefore, Theorem 7 suggests that we need to restrict this decaying rate. We also comment that, Lin, Zhao and Liu (2019) has recently shown that the convergence rate of sliced inverse regression (SIR) relies on the smallest eigenvalue of the SIR matrix. Although we consider a very different problem than Lin, Zhao and Liu (2019), both works have suggested that the eigen-structure of the key quantity of interest, the SIR matrix in their work, and the average Fréchet derivative in ours, plays an important role in the consistency of the estimated basis.

Third, we examine more closely the convergence rate in Theorem 7. For a given d, we denote the rate of the regularization parameter ϵ in (6) as $\epsilon(d)$, and the resulting convergence rate in Theorem 7 as $\zeta(d)$. We study the rate of the $\epsilon(d)$ under which the best convergence rate $\zeta(d)$ is achieved. We have the following result.

COROLLARY 2. *Suppose the same conditions in Theorem 7 hold. Then*

$$\epsilon(d) = (\xi \vee d^{-\beta_2})^{1/(1+\beta_1)}, \qquad \zeta(d) = \nu_q^{-1}(EF_{Y|X})(\xi \vee d^{-\beta_2})^{\beta_1/(1+\beta_1)}.$$

Its proof is by direct calculation and is omitted. To better understand the rates in Corollary 2, suppose $d = n^a$, $\{\nu_d(\Sigma_{XX}) \wedge \nu_d(\Sigma_{YY})\} = n^{-b}$, and $\nu_q\{E(F_{Y|X})\} = n^{-c}$, for some constants $a, b, c > 0$. The value of a reflects the growing rate of the number of KL expansions, b restricts the decaying rate of the eigenvalues of Σ_{XX} and Σ_{YY}, and c controls the decaying rate of the eigenvalues of $E(F_{Y|X})$. Combined with Corollary 2, we have $\epsilon(a, b, \beta_1, \beta_2) = n^{-r(a,b,\beta_1,\beta_2)}$, and $\zeta(a, b, \beta_1, \beta_2) = n^{-\beta_1 r(a,b,\beta_1,\beta_2)+c}$, where $r(a, b, \beta_1, \beta_2) = \{(1/2 - 2a - b) \wedge (a\beta_2)\}/(1 + \beta_1)$. This further implies that, when $(2 + \beta_2)a + b = 1/2$, we obtain the best rates,

$$\epsilon(b, \beta_1, \beta_2) = n^{-\beta_2(1/2-b)/\{(1+\beta_1)(2+\beta_2)\}}, \qquad \zeta(b, \beta_1, \beta_2) = n^{-\beta_1\beta_2(1/2-b)/\{(1+\beta_1)(2+\beta_2)\}+c}.$$

There is a clear interpretation regarding how these best rates are affected by β_1, β_2 as defined in Assumptions 6 and 7, and also the value of b. Specifically, β_1 controls the smoothness of the functional regression operator, and a larger value of β_1 encourages a smoother relation between X and Y, which leads to a larger penalty from ϵ, and a faster convergence rate of ζ. Meanwhile, β_2 regulates the behavior of the tail eigenvalues of Σ_{XX} and Σ_{YY}, and a larger β_2 encourages a faster decaying of those tail eigenvalues, which leads to a smaller penalty from ϵ, and a faster rate of ζ. The quantity b restricts the decaying rate of the gaps from the leading eigenvalues of Σ_{XX} and Σ_{YY}, and a smaller b implies a slower decaying rate, which means a smaller ϵ is required to stabilize the inversion of Σ_{XX}^d, and a faster rate of ζ.

Finally, we have so far studied the asymptotics under the setting when the random functions are fully observed. We remark that all the results can be extended to the setting when the random function are only partially observed. In this case, the sample mean and covariance operators in Lemma 3 are to have a slower convergence rate than $n^{-1/2}$. Suppose the rate is $n^{-1/2+\delta}$, with δ between $[0, 1/2)$, and the denser the observed time points are for the functions, the closer δ is to 0. The specific value of δ depends on the actual sampling schedule of how functional data are collected; see Wang, Chiou and Muller (2016) for more details. Based on the rate of $n^{-1/2+\delta}$, we can extend the results in Theorems 5 to 7 to the setting of partially observed functions. We skip the details, since this extension is relatively straightforward.

6. Numerical studies. In this section, we first carry out simulations to examine the empirical performance of our proposed functional SDR estimators. We then illustrate our method with two real data examples, the classical Canadian weather dataset, and a more recent bike sharing dataset. We report the additional simulation results that compare our method with Li and Song (2017), and study the effect of diverging intrinsic dimensionality in Section S3 of the Supplementary Material (Lee and Li (2022)).

6.1. *Simulations*. We consider two sets of simulation models: in the first set, the response function is associated with the predictor function via the conditional mean only, and in the second set, with not only the conditional mean but also with the conditional variance. Let $\alpha_j = 1/\{(j - 0.5)\pi\}^2$ and $\beta_j(t) = \sin\{(j - 0.5)\pi t\}$ denote the jth eigenvalue and eigenfunction of the Brownian motion kernel, for $j = 1, \ldots, 100$. We independently sample the predictor function $X(t)$ and the error function $\varepsilon(t)$ from $\sum_{j=1}^{100} a_j\sqrt{\alpha_j}\beta_j(t)$, where a_j's are i.i.d. standard normal variables. Let $\phi_1(t) = \beta_4(t)$, and $\phi_2(t) = \beta_5(t)$. We then generate the response function $Y(t)$ as

I-1: $$Y(t) = \beta_1(t) \times \langle X(t), \phi_1(t)\rangle_{\Omega_X} + \beta_2(t) \times \langle X(t), \phi_2(t)\rangle_{\Omega_X} + \varepsilon(t);$$

$$\text{I-2:} \quad Y(t) = \beta_1(t) \times \sin\{\langle X(t), \phi_1(t)\rangle_{\Omega_X}\} + \beta_2(t) \times \sin\{\langle X(t), \phi_2(t)\rangle_{\Omega_X}\} + \varepsilon(t);$$

$$\text{II-1:} \quad Y(t) = \beta_1(t) \times \langle X(t), \phi_1(t)\rangle_{\Omega_X} + \langle X(t), \phi_2(t)\rangle_{\Omega_X} \times \varepsilon(t);$$

$$\text{II-2:} \quad Y(t) = \beta_1(t) \times \sin\{\langle X(t), \phi_1(t)\rangle_{\Omega_X}\} + \sin\{\langle X(t), \phi_2(t)\rangle_{\Omega_X}\} \times \varepsilon(t).$$

By construction, for models I-1 and I-2, $\mathcal{S}_{E(Y|X)} = \mathcal{S}_{Y|X} = \operatorname{Span}\{\phi_1(t), \phi_2(t)\}$, while for models II-1 and II-2, $\mathcal{S}_{E(Y|X)} = \operatorname{Span}\{\phi_1(t)\}$ and $\mathcal{S}_{Y|X} = \operatorname{Span}\{\phi_1(t), \phi_2(t)\}$. We then take the same 50 equally spaced points $\{t_{k1}, \ldots, t_{k50}\}$ between $[0, 1]$ as the observed time points, for all $k = 1, \ldots, n$. We consider different values of the sample size $n = \{100, 250, 500\}$. In our implementation of the functional SDR, we use a Brownian motion kernel for κ_T, and a radial basis kernel for κ_X and κ_Y.

We first evaluate the accuracy of recovering $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{Y|X}$. Toward that end, we generate n i.i.d. pairs of random functions $\{X(t), Y(t)\}$ for each model, then divide them into a training set and a testing set, each with $n/2$ observations. We then apply our methods to the training data to estimate the basis functions for the functional dimension reduction subspaces. We next compute the top q sufficient predictors using the testing data. We first use the true value of q, then examine the estimation of q using (12) later. To evaluate how effective the estimated sufficient predictors approximate the true ones, we compute the trace of the average squared multivariate correlation coefficient matrix (MCC, Ferré and Yao (2005)),

$$\begin{aligned}\operatorname{MCC}(Z, \hat{Z}) = \big(q^{-1} \operatorname{tr}\big[\{\operatorname{cov}_n(Z, Z)\}^{-1/2}\{\operatorname{cov}_n(Z, \hat{Z})\} \\ \times \{\operatorname{cov}_n(\hat{Z}, \hat{Z})\}^{-1}\{\operatorname{cov}_n(\hat{Z}, Z)\}\{\operatorname{cov}_n(Z, Z)\}^{-1/2}\big]\big)^{1/2},\end{aligned}$$

where $Z = (\langle \phi_1, X\rangle_{\Omega_X}, \ldots, \langle \phi_1, X\rangle_{\Omega_X})^{\mathsf{T}}$, and $\hat{Z} = (\langle \hat{\phi}_1, X\rangle_{\Omega_X}, \ldots, \langle \hat{\phi}_1, X\rangle_{\Omega_X})^{\mathsf{T}}$ are the true and estimated sufficient predictors, respectively. The value of MCC is between 0 and 1, with 1 indicating a perfect recovery. Besides, it is calculated based on the testing samples, and hence avoids the overfitting issue. Figure 1 reports the box plots of this criterion for the four models based on 80 data replications, for both the functional central mean subspace (fCMS) and the functional central subspace (fCS) estimation. It is seen that, for models I-1 and I-2, both estimators perform well, since the conditional mean contains all the relevant regression information. For models II-1 and II-2, the functional central subspace estimator performs consistently better than the functional central mean subspace estimator, because the conditional variance contains additional relevant regression information that the functional central mean subspace misses. Moreover, the performance improves as the sample size n increases, which agrees with our asymptotic theory.

Next, we evaluate the performance of selecting the intrinsic dimension q using (12). We choose the top q eigenfunctions that account for 90% total variation in the estimated average Fréchet derivative. For models I-1 and I-2, both $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{Y|X}$ have the intrinsic dimension $q = 2$. For models II-1 and II-2, $\mathcal{S}_{E(Y|X)}$ has the intrinsic dimension $q = 1$, and $\mathcal{S}_{Y|X}$ has the intrinsic dimension $q = 2$. Table 1 reports the estimated q averaged over 80 data replications. It is seen that, the estimated q is slightly larger than the true intrinsic dimension, especially for models II-1 and II-2. This is acceptable from a dimension reduction perspective, as a slightly larger estimated q implies that the method recovers some additional redundant information, while a smaller estimated q means the method misses some important information. Moreover, the estimation of q becomes more accurate as the sample size n increases.

Finally, we study the performance of postreduction prediction. We compare our method with three state-of-the-art parametric function-on-function regression methods: the functional linear regression (LIN, Yao, Müller and Wang (2005)), the functional linear regression by signal compression (LQ, Luo and Qi (2017)), and the functional additive regression (ADD,

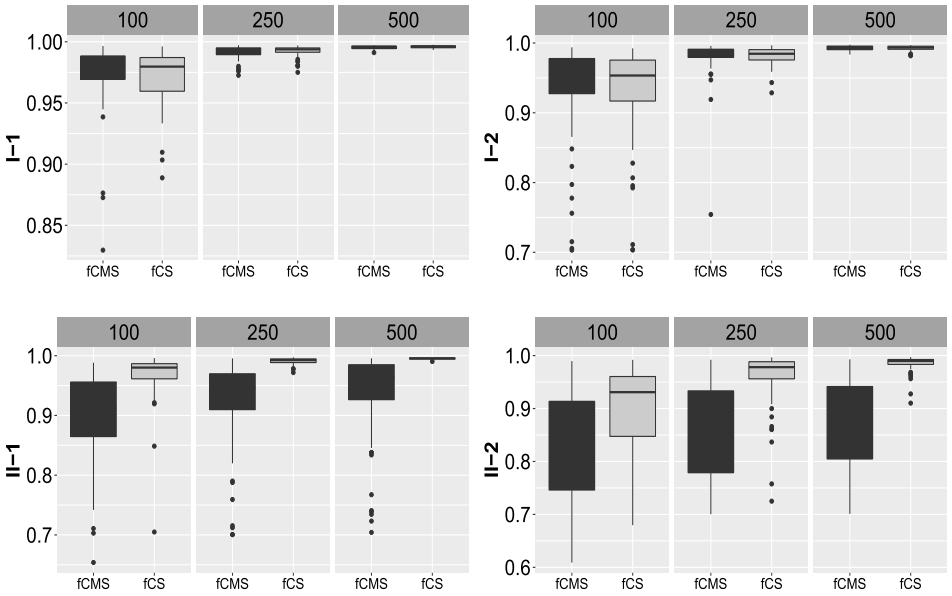


FIG. 1. *Box plots of the average squared multivariate correlation coefficients between the true and estimated sufficient predictors, for the functional central mean subspace estimator (fCMS), and the functional central subspace estimator (fCS), and the four models, respectively. Each panel consists of* 3 *boxes for three sample sizes* $n = 100, 250, 500$.

Müller and Yao (2008)). Both LIN and ADD are implemented in the `R` package `PACE`, and LQ is implemented in the `R` package `FRegSigCom`. Since the competing methods were designed to predict the mean function of the response function, it is more meaningful to focus our comparison on the conditional mean models I-1 and I-2. To evaluate the predictive performance, we fit the model using the training samples, then obtain the predicted response function $\hat{Y}(x,t) = \hat{E}\{Y(t) \mid x\}$ using the testing samples. We then compute the mean squared prediction error (MSPE),

$$\text{MSPE} = \sum_{k=1}^{n_t} \sum_{s=1}^{m} \{Y^k(t_s) - \hat{Y}(X^k, t_s)\}^2/(n_t m),$$

TABLE 1
The estimated intrinsic dimension q

Model		fCMS			fCS		
	n	100	250	500	100	250	500
I-1	mean	3.00	2.48	2.01	3.99	3.01	2.14
	s.d.	0.32	0.50	0.11	0.11	0.11	0.35
I-2	mean	3.58	3.00	2.58	4.03	3.92	3.02
	s.d.	0.50	0.23	0.50	0.16	0.27	0.16
II-1	mean	4.00	3.86	3.27	4.08	3.94	3.01
	s.d.	0.00	0.35	0.45	0.27	0.24	0.11
II-2	mean	4.00	3.95	3.39	4.41	4.01	3.88
	s.d.	0.00	0.22	0.49	0.50	0.11	0.33

TABLE 2
The mean squared prediction error for the prediction methods based on the functional central mean subspace (fCMS), the functional central subspace (fCS), the functional linear regression (LIN), the functional linear regression by signal compression (LQ), and the functional additive regression (ADD)

Model		fCMS	fCS	LIN	LQ	ADD
I-1	mean	0.715	0.695	2.584	**0.540**	2.618
	s.d.	0.159	0.121	0.151	0.042	0.147
I-2	mean	0.610	**0.603**	1.396	0.666	1.411
	s.d.	0.063	0.058	0.081	0.057	0.088

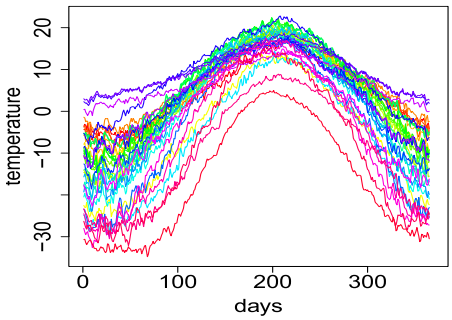
where $(X^1, Y^1)^\mathsf{T}, \ldots, (X^{n_t}, Y^{n_t})^\mathsf{T}$ are the testing samples. Table 2 reports the results based on 80 data replications. It is seen that, LQ has the best predictive performance for model I-1, while our prediction method based on the functional central subspace $\mathcal{S}_{Y|X}$ is the second best. This is not surprising though, since their method was specifically designed for a linear function-on-function regression model. Meanwhile, our method based on $\mathcal{S}_{Y|X}$ performs the best for model I-2, where there is a clear nonlinear association between the predictor and response functions.

6.2. *Canadian weather data*. We first illustrate our methods with the classical Canadian weather data, which is often used as a benchmark for function-on-function regression analysis (Ramsay and Silverman (2005)). The dataset consists of the average daily temperature and logarithm of the daily precipitation over 35 years, from 1960 to 1994, for 35 different locations in Canada. The data is available from the `R` package `fda`. Figure 2(a)–(b) show the observed time series for the daily temperature and precipitation in the logarithmic scale. The analysis goal is to study the association between the temperature and precipitation, and to use the temperature to predict the precipitation.

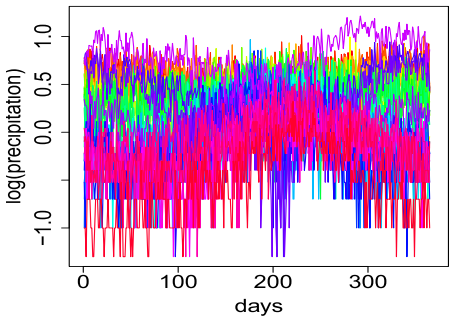
We apply the proposed functional SDR methods to this data. For $\mathcal{S}_{E(Y|X)}$, the top 3 sufficient predictors explain 45.1%, 33.6% 15.2% of total variation, respectively, and 93.9% accumulatively. For $\mathcal{S}_{Y|X}$, the top 3 sufficient predictors explain 48.5%, 37.3%, 9.2% of total variation, respectively, and 94.9% accumulatively. Figure 2(c)–(d) show these estimated bases, which look similar for $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{Y|X}$ in this example. From the plots, it is seen that the first basis function $\hat{\phi}_1(t)$ weighs more at the beginning and toward the end of the year, while the second basis function $\hat{\phi}_2(t)$ weighs more in the middle of the year.

Next, we apply the postreduction prediction method to estimate the conditional mean of the precipitation function in logarithm given the top 3 estimated bases from the temperature function. Figure 3(a) shows the scatterplot of the first sample KL coefficient of the logarithmic daily precipitation, as well as its estimated regression function, that is, the conditional mean, versus the first and second estimated sufficient predictors. We also interpolate the regression surface for a clearer visualization. From the plot, we see that the sample KL coefficient is well approximated by its regression function. We also observe an increasing trend of the conditional mean as the sufficient predictors increase.

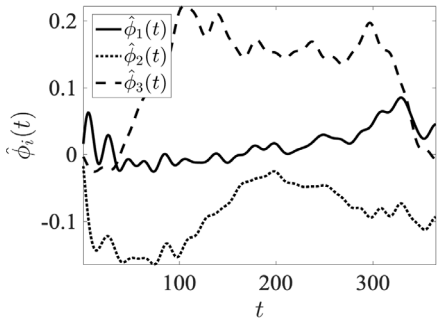
Finally, we perform an out-of-sample prediction analysis. That is, we randomly draw 30 out of 35 observations as the training samples, and use the rest as the testing samples. We record the mean squared prediction error based on the testing data, and compare the five methods studied in Section 6.1. We repeat this process 100 times. Table 3 reports the results averaged over such 100 replications. It is seen that our method based on the functional central subspace achieves the best prediction accuracy.



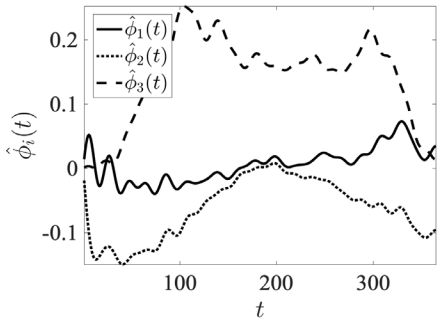
(a) Daily temperature



(b) Daily precipitation in logarithm



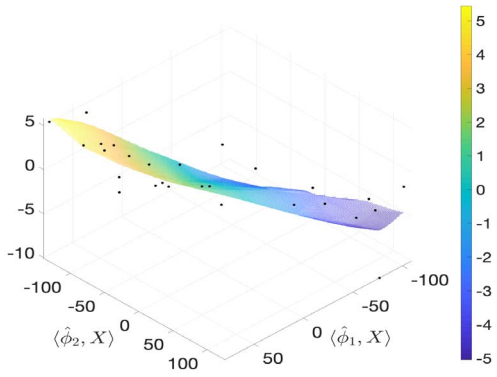
(c) Estimated top 3 bases for $\mathcal{S}_{E(Y|X)}$



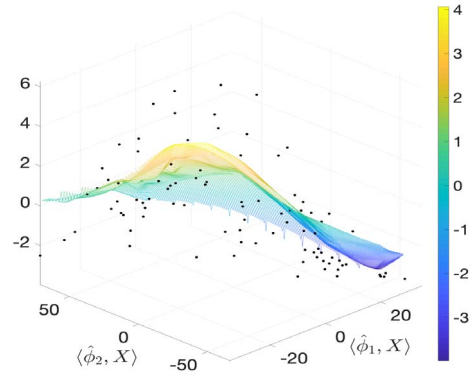
(d) Estimated top 3 bases for $\mathcal{S}_{Y|X}$

FIG. 2. *Canadian weather data.*

6.3. *Bike sharing data.* We next illustrate our methods with a more recent business application, the bike sharing data (Fanaee-T and Gama (2014)). The dataset consists of the hourly counts of total rental bikes for both casual and registered users, and weather related information, such as the hourly temperature, precipitation, wind speed, and humidity. The observations were recorded from the Capital Bike Share system in Washington, DC every day from January 1, 2011 to December 31, 2012. The data is available from https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset. One of the analysis goals is to understand how the bike rentals are affected by the temperature on Saturdays. Therefore, we



(a) Canadian weather data



(b) Bike sharing data

FIG. 3. *The first sample KL coefficient of the response function (black dots) and its regression function (colored mesh) versus the first two sufficient predictors.*

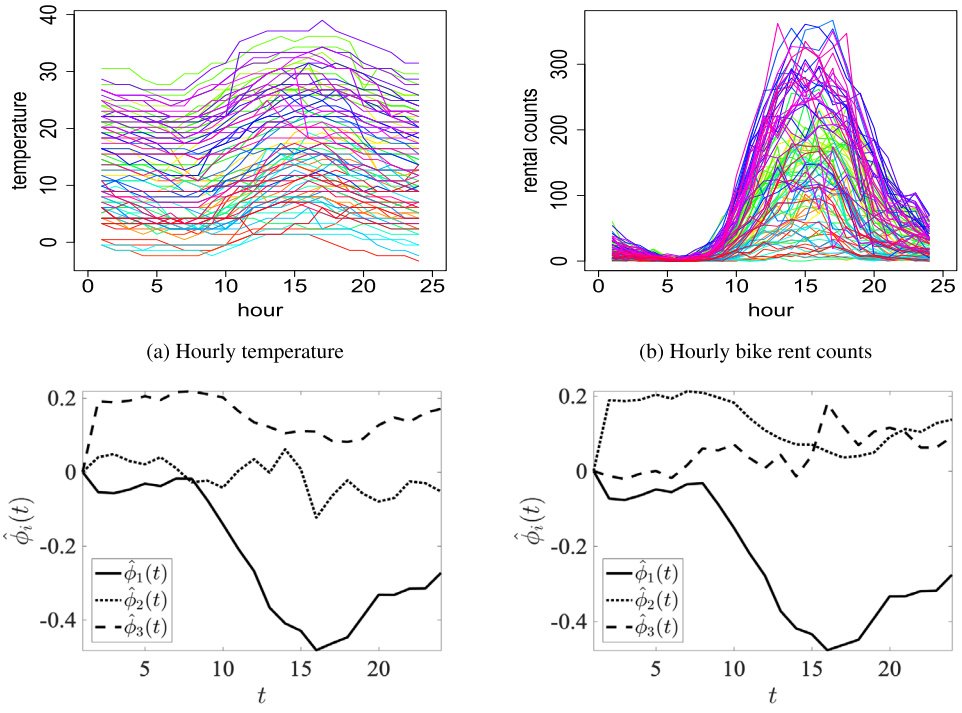
TABLE 3

The mean squared prediction error for the prediction methods based on the functional central mean subspace (fCMS), the functional central subspace (fCS), the functional linear regression (LIN), the functional linear regression by signal compression (LQ), and the functional additive regression (ADD)

Dataset		fCMS	fCS	LIN	LQ	ADD
Canadian weather	mean	0.0865	**0.0861**	0.1134	0.1230	0.1192
	s.d.	0.0301	0.0300	0.0401	0.0589	0.0492
Bike sharing	mean	**0.1357**	0.1361	0.1995	0.1775	0.1458
	s.d.	0.0421	0.0415	0.0732	0.0630	0.0545

treat the hourly bike rentals as the response function and the hourly temperature as the predictor function. After removing 3 weeks of observations with many missing values, we obtain 102 weeks of pairs of curves, which are shown in Figure 4(a)–(b).

We apply the proposed functional SDR methods to this data. For the functional central mean subspace $\mathcal{S}_{E(Y|X)}$, the top 3 sufficient predictors explain 62.8%, 18.4% 14.6% of total variation, respectively, and 95.9% accumulatively. For the functional central subspace $\mathcal{S}_{Y|X}$, the top 3 sufficient predictors explain 69.3%, 16.2%, 10.9% of total variation, respectively, and 96.4% accumulatively. Figure 4(c)–(d) show these estimated bases. It is seen that, the first basis $\hat{\phi}_1(t)$ suggests that the temperature between 10 am and midnight has more impact on the bike rentals than the rest of the day, while the second and third bases suggest that early morning tends to affect more on the bike rentals. Moreover, we observe that, although the estimated bases of the $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{Y|X}$ are not numerically identical, they capture very similar patterns. This suggests that the leading estimated bases are most likely related to the change of the conditional mean of the hourly rental counts. Also, we see that the second and



(a) Hourly temperature

(b) Hourly bike rent counts

(c) Estimated top 3 bases for $\mathcal{S}_{E(Y|X)}$

(d) Estimated top 3 bases for $\mathcal{S}_{Y|X}$

FIG. 4. *Bike sharing data for* 102 *Saturdays.*

third estimated bases from the two estimations roughly span the same space, with the second estimated basis for $\mathcal{S}_{E(Y|X)}$ corresponding to the sign-flipped third estimated basis for $\mathcal{S}_{Y|X}$, and the third estimated basis for $\mathcal{S}_{E(Y|X)}$ corresponding to the second estimated basis for $\mathcal{S}_{Y|X}$. This order switching between the estimated bases is likely due to the closeness of the second and third largest eigenvalues.

Next, we apply the postreduction prediction method to estimate the conditional mean of the bike rental function given the top 3 estimated bases from the temperature function. Figure 3(b) shows the scatterplot of the first sample KL coefficient of the hourly rental, and its estimated regression function, versus the first and second estimated sufficient predictors. We again interpolate the regression surface. We see that the estimated regression function well approximates the estimated KL coefficient. Besides, as the first two sufficient predictors increase, the conditional mean of the first KL of the bike rental decreases.

Finally, we perform the out-of-sample prediction analysis, by randomly drawing 90 observations out of 102 as the training samples, and then using the rest as the testing samples. Table 3 reports the results averaged over 100 replications. For this example, our method based on the functional central mean subspace achieves the best prediction accuracy.

7. Discussion. In this section, we reiterate and further clarify some key differences between our proposal and the existing gradient-based SDR and functional SDR solutions.

We first compare to the gradient-based SDR methods for the random variable case.

(a) The gradient-based linear SDR methods such as Härdle and Stoker (1989), Xia (2007), Xia et al. (2002), Yin and Li (2011) targeted the central mean or central subspaces that are finite-dimensional subspaces of $\mathbb{R}^p$. In comparison, our functional central mean and central subspaces are defined as the projections of Hilbert space-valued random predictors, which themselves can be Hilbert space-valued and infinite-dimensional.

(b) A major step in the gradient-based methods is to estimate the gradient itself. Nevertheless, this step has imposed great challenges under the functional setting. Härdle and Stoker (1989) estimated the gradient of the density function; however, the density function does not always exist for random functions. Alternatively, Xia (2007), Xia et al. (2002), Yin and Li (2011) estimated the gradient of the regression functions, which requires to solve a high-dimensional local linear regression. Extending the local linear regression to the functional setting is feasible, but is analytically complicated. In contrast, we build our estimator on the Fréchet derivative of the regression functionals, which has a closed form. To achieve this, we rely on a key property of RKHS, in which we establish the interchangeability between the nested kernel function and the Fréchet derivative; see Proposition 2. This simplification in estimating the Fréchet derivative is useful in both estimation at the population and sample levels, as well as in deriving the consistency results.

(c) We have developed a number of useful properties regarding the Fréchet differentiability of the nested RKHS. For instance, we derive the conditions under which the nested kernel function is continuously Fréchet differentiable in Proposition S1. To show that the Fréchet derivative of the regression functionals have tractable forms, we illustrate with some popular kernel functions and derive their Fréchet derivatives in Proposition S2 at the population level, and in Proposition S5 at the coordinate level. Moreover, in Lemma S3 we derive a property on the interchangeability of partial Fréchet differentiability on the nested RKHS, which allows us to compute the moments of the Fréchet derivative; see Proposition S4. These results are sufficiently general to be applied to problems beyond the SDR framework, and are useful for functional data analysis involving the Fréchet derivative.

(d) We remark that, a possible alternative for our functional SDR is to discretize the response function as $Y(t_1), \ldots, Y(t_m)$ for the observed time points $t_1, \ldots, t_m$, then perform dimension reduction by treating Y as an m-dimensional random vector (Hsing (1999), Li, Wen and Zhu (2008)). This solution does not require to specify a working RKHS. However, a major advantage of our approach is that we are able to borrow information across the entire function to compute the KL coefficients, then to recover the entire function using only a few leading coefficients. This approach greatly alleviates the curse of dimensionality, as the working space is of the same dimension as the number of extracted KL coefficients d, rather than the number of observed time points m, and m is usually much larger than d. This is also reflected in our asymptotic convergence rates, which depend on d, not m.

We next compare to the functional SDR methods, particularly, Li and Song (2017).

(a) Although Li and Song (2017) has shown the existence of the nonlinear functional subspace, our Theorem 1 on the existence of the functional central mean and central subspaces can not be directly inferred by their result. To establish the existence, we introduce the concept of M-set in the functional space, a result that is not available in their work.

(b) A major difference between our work and Li and Song (2017) is that, their estimator was based on the inverse regression of $X \mid Y$, whereas our estimator is built on the forward regression of $Y \mid X$. The inverse regression approach has been taken by many other functional SDR estimators, for example, Hsing and Ren (2009), Jiang, Yu and Wang (2014), Wang, Lin and Zhang (2013), Wang et al. (2015), Yao, Lei and Wu (2015). Such a difference leads to completely different estimation approaches, theoretical analyses, as well as required conditions. For instance, the generalized sliced average variance estimator of Li and Song (2017) and all other inverse regression based functional SDR methods required some version of the linearity or constant variance conditions to establish the unbiasedness or exhaustiveness. In contrast, our method achieves the unbiasedness and exhaustiveness without imposing such distributional assumptions.

(c) Another major difference is that Li and Song (2017) targeted nonlinear SDR, whereas we aim at linear SDR, as we have clarified after Theorem 1. Each has its own strength and limitation; see Li ((2018a), Chapter 14) for more comparison. However, simply using a linear kernel in Li and Song (2017) would not produce the same result as our method, due to different approaches of inverse and forward regressions the two methods adopt.

(d) Compared to the estimator of Li and Song (2017) that was solely built on the functional regression operator, our estimator requires both the functional regression operator, and another key element, the Fréchet derivative. Consequently, we have developed a different set of tools regarding the Fréchet differentiability of the nested RKHS, None of those results are available in Li and Song (2017).

(e) A key novelty of our asymptotic analysis is that we allow the number of sufficient predictors to diverge with the sample size. Li and Song ((2017), Corollary 3) has shown the pointwise consistency, but assumed a fixed number of sufficient predictors. In comparison, in Theorem 7, we have established the uniform consistency with a diverging number of sufficient predictors, which is considerably more challenging than Li and Song (2017). We require additional technical tools such as Lemma 2 to deal with the perturbation of the sample linear operators. Moreover, our asymptotic framework needs to deal with the consistency of the Fréchet derivative estimation; for example, in the proof of Theorem 6, we derive the rate of convergence for the Fréchet derivative, based on which we are able to establish the consistency of our average Fréchet derivative estimator.

In summary, our proposal is far from an incremental extension of the existing SDR methods. We believe it is important and useful for both fields of sufficient dimension reduction and functional data analysis in general.

Acknowledgments. The authors would like to thank the three anonymous referees, the Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding. Lee's research was partially supported by the NSF Grant CIF-2102243, and the Seed Funding grant from Fox School of Business, Temple University.

Li's research was partially supported by the NSF Grant CIF-2102227, and the NIH Grants R01AG061303, R01AG062542 and R01AG034570.

SUPPLEMENTARY MATERIAL

Supplementary appendix for "Functional sufficient dimension reduction through average Fréchet derivatives" (DOI: 10.1214/21-AOS2131SUPP; .pdf). The supplementary appendix collects detailed discussions of the regularity assumptions, some additional results regarding the estimation, asymptotic theory and simulations, and all the technical proofs.

REFERENCES

AMINI, A. A. and WAINWRIGHT, M. J. (2012). Sampled forms of functional PCA in reproducing kernel Hilbert spaces. *Ann. Statist.* **40** 2483–2510. MR3097610 https://doi.org/10.1214/12-AOS1033

CONWAY, J. B. (2010). *A Course in Functional Analysis*, 2nd ed. Springer, New York.

COOK, R. D. and LI, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30** 455–474. MR1902895 https://doi.org/10.1214/aos/1021379861

COOK, R. D. and WEISBERG, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *J. Amer. Statist. Assoc.* **86** 328–332.

FANAEE-T, H. and GAMA, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Prog. Artif. Intell.* **2** 113–127. https://doi.org/10.1007/s13748-013-0040-3

FERRÉ, L. and YAO, A. F. (2003). Functional sliced inverse regression analysis. *Statistics* **37** 475–488. MR2022235 https://doi.org/10.1080/0233188031000112845

FERRÉ, L. and YAO, A.-F. (2005). Smoothed functional inverse regression. *Statist. Sinica* **15** 665–683. MR2233905

FUKUMIZU, K., BACH, F. R. and JORDAN, M. I. (2009). Kernel dimension reduction in regression. *Ann. Statist.* **37** 1871–1905. MR2533474 https://doi.org/10.1214/08-AOS637

FUKUMIZU, K. and LENG, C. (2014). Gradient-based kernel dimension reduction for regression. *J. Amer. Statist. Assoc.* **109** 359–370. MR3180569 https://doi.org/10.1080/01621459.2013.838167

HARDLE, W. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995.

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995. MR1134488

HSING, T. (1999). Nearest neighbor inverse regression. *Ann. Statist.* **27** 697–731. MR1714711 https://doi.org/10.1214/aos/1018031213

HSING, T. and REN, H. (2009). An RKHS formulation of the inverse regression dimension-reduction problem. *Ann. Statist.* **37** 726–755. MR2502649 https://doi.org/10.1214/07-AOS589

JIANG, C.-R., YU, W. and WANG, J.-L. (2014). Inverse regression for longitudinal data. *Ann. Statist.* **42** 563–591. MR3210979 https://doi.org/10.1214/13-AOS1193

KIM, J. S., STAICU, A.-M., MAITY, A., CARROLL, R. J. and RUPPERT, D. (2018). Additive function-on-function regression. *J. Comput. Graph. Statist.* **27** 234–244. MR3788315 https://doi.org/10.1080/10618600.2017.1356730

LEE, Y.-J. and HUANG, S.-Y. (2007). Reduced support vector machines: A statistical theory. *IEEE Trans. Neural Netw.* **18** 1–13. https://doi.org/10.1109/TNN.2006.883722

LEE, K.-Y. and LI, L. (2022). Supplement to "Functional sufficient dimension reduction through average Fréchet derivatives." https://doi.org/10.1214/21-AOS2131SUPP

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

LI, B. (2018a). *Sufficient Dimension Reduction*: *Methods and Applications with R. Monographs on Statistics and Applied Probability* **161**. CRC Press, Boca Raton, FL. MR3838449 https://doi.org/10.1201/9781315119427

LI, B. (2018b). Linear operator-based statistical analysis: A useful paradigm for big data. *Canad. J. Statist.* **46** 79–103. MR3767167 https://doi.org/10.1002/cjs.11329

LI, B. and SONG, J. (2017). Nonlinear sufficient dimension reduction for functional data. *Ann. Statist.* **45** 1059–1095. MR3662448 https://doi.org/10.1214/16-AOS1475

LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102** 997–1008. MR2354409 https://doi.org/10.1198/016214507000000536

LI, B., WEN, S. and ZHU, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *J. Amer. Statist. Assoc.* **103** 1177–1186. MR2462891 https://doi.org/10.1198/016214508000000445

LI, B., ZHA, H. and CHIAROMONTE, F. (2005). Contour regression: A general approach to dimension reduction. *Ann. Statist.* **33** 1580–1616. MR2166556 https://doi.org/10.1214/009053605000000192

LIN, Q., ZHAO, Z. and LIU, J. S. (2019). Sparse sliced inverse regression via Lasso. *J. Amer. Statist. Assoc.* **114** 1726–1739. MR4047295 https://doi.org/10.1080/01621459.2018.1520115

LIN, Q., LI, X., HUANG, D. and LIU, J. S. (2017). On the optimality of sliced inverse regression in high dimensions.

LUO, R. and QI, X. (2017). Function-on-function linear regression by signal compression. *J. Amer. Statist. Assoc.* **112** 690–705. MR3671763 https://doi.org/10.1080/01621459.2016.1164053

LUO, R. and QI, X. (2019). Interaction model and model selection for function-on-function regression. *J. Comput. Graph. Statist.* **28** 309–322. MR3974882 https://doi.org/10.1080/10618600.2018.1514310

MA, Y. and ZHU, L. (2012). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc.* **107** 168–179. MR2949349 https://doi.org/10.1080/01621459.2011.646925

MA, Y. and ZHU, L. (2013). Efficient estimation in sufficient dimension reduction. *Ann. Statist.* **41** 250–268. MR3059417 https://doi.org/10.1214/12-AOS1072

MÜLLER, H.-G. and YAO, F. (2008). Functional additive models. *J. Amer. Statist. Assoc.* **103** 1534–1544. MR2504202 https://doi.org/10.1198/016214508000000751

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2168993

REIMHERR, M., SRIPERUMBUDUR, B. and TAOUFIK, B. (2018). Optimal prediction for additive function-on-function regression. *Electron. J. Stat.* **12** 4571–4601. MR3893421 https://doi.org/10.1214/18-EJS1505

STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. Springer, New York.

SUN, X., DU, P., WANG, X. and MA, P. (2018). Optimal penalized function-on-function regression under a reproducing kernel Hilbert space framework. *J. Amer. Statist. Assoc.* **113** 1601–1611. MR3902232 https://doi.org/10.1080/01621459.2017.1356320

WANG, J. L., CHIOU, J. M. and MULLER, H. G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.

WANG, G., LIN, N. and ZHANG, B. (2013). Functional contour regression. *J. Multivariate Anal.* **116** 1–13. MR3049886 https://doi.org/10.1016/j.jmva.2012.11.005

WANG, G., ZHOU, Y., FENG, X.-N. and ZHANG, B. (2015). The hybrid method of FSIR and FSAVE for functional effective dimension reduction. *Comput. Statist. Data Anal.* **91** 64–77. MR3368006 https://doi.org/10.1016/j.csda.2015.05.011

WEIDMANN, J. (1980). *Linear Operators in Hilbert Spaces*. *Graduate Texts in Mathematics* **68**. Springer, New York–Berlin. MR0566954

XIA, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.* **35** 2654–2690. MR2382662 https://doi.org/10.1214/009053607000000352

XIA, Y., TONG, H., LI, W. K. and ZHU, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 363–410. MR1924297 https://doi.org/10.1111/1467-9868.03411

YAO, F., LEI, E. and WU, Y. (2015). Effective dimension reduction for sparse functional data. *Biometrika* **102** 421–437. MR3371014 https://doi.org/10.1093/biomet/asv006

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106 https://doi.org/10.1214/009053605000000660

YIN, X. and LI, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann. Statist.* **39** 3392–3416. MR3012413 https://doi.org/10.1214/11-AOS950

YIN, X., LI, B. and COOK, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *J. Multivariate Anal.* **99** 1733–1757. MR2444817 https://doi.org/10.1016/j.jmva.2008.01.006