The Annals of Statistics
2017, Vol. 45, No. 3, 1059–1095
DOI: 10.1214/16-AOS1475

NONLINEAR SUFFICIENT DIMENSION REDUCTION FOR FUNCTIONAL DATA

BY BING LI[1] AND JUN SONG

Pennsylvania State University

We propose a general theory and the estimation procedures for nonlinear sufficient dimension reduction where both the predictor and the response may be random functions. The relation between the response and predictor can be arbitrary and the sets of observed time points can vary from subject to subject. The functional and nonlinear nature of the problem leads to construction of two functional spaces: the first representing the functional data, assumed to be a Hilbert space, and the second characterizing nonlinearity, assumed to be a reproducing kernel Hilbert space. A particularly attractive feature of our construction is that the two spaces are nested, in the sense that the kernel for the second space is determined by the inner product of the first. We propose two estimators for this general dimension reduction problem, and establish the consistency and convergence rate for one of them. These asymptotic results are flexible enough to accommodate both fully and partially observed functional data. We investigate the performances of our estimators by simulations, and applied them to data sets about speech recognition and handwritten symbols.

1. Introduction. Functional data are prevalent in contemporary statistical applications such as Chemometrics, speech recognition, meteorology and longitudinal data analysis. As a result, estimation and inference methods to study the interrelations between functions are becoming increasingly important in data analysis. See Ramsay and Silverman (2005), Yao, Müller and Wang (2005a, 2005b), Ferraty and Vieu (2006), Horváth and Kokoszka (2012) and Hsing and Eubank (2015). In this paper, we develop a general theory along with estimation procedures for nonlinear sufficient dimension reduction of functional data, where both the predictor and response are allow to be random functions, and their relations can be arbitrary. In particular, the function-valued predictor and response do not have to be related through linear indices, as assumed by the recently-developed sufficient dimension reduction methods for functional data [Müller and Stadtmüller (2005), Ferré and Yao (2003, 2005), Hsing and Ren (2009)].

Received December 2015.

[1] Supported in part by NSF Grant DMS-14-07537.

MSC2010 subject classifications. 62B05, 62G08, 62G20, 62H99.

Key words and phrases. Convergence rate, linear operator, handwriting data, reproducing kernel Hilbert space, speech recognition, Sliced Average Variance Estimator, Sliced Inverse Regression.

Classical sufficient dimension reduction is characterized by conditional independence

$$Y \perp\!\!\!\perp X | \beta^{\mathsf{T}} X, \tag{1}$$

where X is a p-dimensional random vector, Y is a random variable, and β is a $p \times d$ matrix ($d \ll p$). The goal is to estimate the space spanned by the columns of β. That is, we seek a few linear combinations of X that are sufficient to describe the conditional distribution of Y given X [see Cook and Weisberg (1991), Li (1991), Cook (1998), Cook and Li (2002)]. This problem is linear in the sense that the reduced predictor takes the linear form $\beta^{\mathsf{T}} X$. For this reason, we refer to it as linear sufficient dimension reduction (linear SDR). The situation where Y is also a vector was considered in Cook and Setodji (2003), Yin and Bura (2006) and Li, Wen and Zhu (2008).

The theory of linear SDR was extended to functional data by Ferré and Yao (2003, 2005) and Hsing and Ren (2009), where the random element X takes values in a functional space, say $\mathcal{H}$, whose members are functions defined on an interval, representing time. The goal of functional linear SDR is to find members $f_1, \dots, f_d$ of $\mathcal{H}$ such that

$$Y \perp\!\!\!\perp X | \langle f_1, X \rangle_{\mathcal{H}}, \dots, \langle f_d, X \rangle_{\mathcal{H}},$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ represents the inner product in $\mathcal{H}$. On a different front, linear SDR was generalized to the nonlinear case by Li, Artemiou and Li (2011) and Lee, Li and Chiaromonte (2013), which seek a set of nonlinear functions $f_1(X), \dots, f_d(X)$ such that

$$Y \perp\!\!\!\perp X | f_1(X), \dots, f_d(X). \tag{2}$$

This was accomplished by enlarging the Euclidean space of linear coefficient vectors for linear SDR to a Hilbert space of functions of X. Lee, Li and Chiaromonte (2013) showed that the nonlinear functions $f_1, \dots, f_d$ in (2) can be obtained from the eigenfunctions of certain linear operators, and developed two methods to estimate them. The precursors of this theory include Bach and Jordan (2003), Cook (2007), Wu (2008) and Yeh, Huang and Lee (2009), which introduced a variety of practical nonlinear sufficient dimension reduction methods without articulating a unifying framework. The generalization frees us from the linear constraint in (1), so that we can handle relations between X and Y that cannot be described by linear indices $\beta^{\mathsf{T}} X$, thus achieving further reduction of dimension.

In this paper, we go one step further to propose the theory and methods for nonlinear sufficient dimension reduction for functional data, which we abbreviate as functional nonlinear SDR. Specifically, let X and Y be random functions defined an interval $T \subseteq \mathbb{R}$ representing time. Our goal is to find a set of nonlinear functions $f_1, \dots, f_d$ of X such that the random functions Y and X are independent conditioning on $f_1(X), \dots, f_d(X)$. The functional and nonlinear nature of this problem demands that we consider two nested functional spaces. First, X and Y are themselves functions on T, and they reside in functional spaces whose domains are T. Second, $f_1(X), \dots, f_d(X)$ reside in a space of functions whose domains are the first space. To employ the nested Hilbert spaces to construct estimators of $f_1, \dots, f_d$ and to develop asymptotic theories special to functional nonlinear SDR are the core of this paper.

This generalization is motivated and justified by many recent applications. For example, in a speech recognition problem [Ferraty and Vieu (2006), Section 2.2], the predictor is a voice signal as a function of time, and the response is the name of the vowel sounds pronounced, which can be regarded as a discrete random variable. And, regarding a Canadian weather data set [Ramsay and Silverman (2005), page 17] asked the following question:

> Can the temperature record Temp be used as a predictor of the entire precipitation profile, not merely the total precipitation?

This is a case where one is interested in predicting one set of functions by another set of functions. As a more visual example, consider the problem of training the computer to learn to associate two sets of handwritten symbols, one numerical and one alphabetical, as illustrated in Figure 1. In this problem, both the predictor and the response are in the form of two-dimensional functions $t \mapsto (f_1(t), f_2(t))^{\mathsf{T}}$, which describe the curves in a two-dimensional space. Moreover, the relation is too complicated to be described through linear index as in classical sufficient dimension reduction. As we show in Section 9, the methods proposed in this paper are sufficiently flexible to handle all these situations.

An alternative approach to functional nonlinear SDR is to represent each function as coordinates in a basis expansion and then perform nonlinear SDR on the coefficients as if they were multivariate data. However, the chief advantage of treating observed units as functions rather than vectors is that we can smooth the data across time; that is, to borrow information from adjacent time points. This special feature of functional data analysis is reflected in the convergence rate we develop

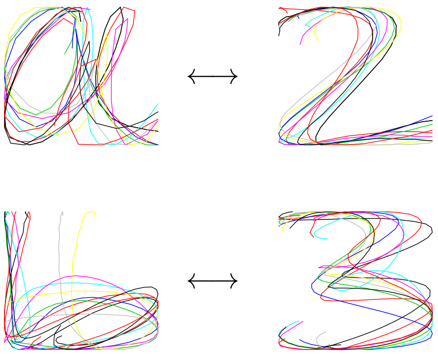


FIG. 1. *Associating two sets of symbols.*

in Section 7, where the convergence rate improves as the observe frequency of the functional data increases. Such an asymptotic behavior cannot be expected in the multivariate setting, where increase of dimension tends to hamper, rather than enhance, accuracy.

The rest of the paper is organized as follows. In Section 2, we give the formal definition of functional nonlinear SDR. In Section 3, we construct the two nested Hilbert spaces. In Sections 4 and 5, we extend the Generalized Sliced Inverse Regression (GSIR) and the Generalized Average Variance Estimator (GSAVE) introduced by Lee, Li and Chiaromonte (2013) to the functional case, resulting in function GSIR (f-GSIR) and functional GSAVE (f-GSAVE). In Section 6, we develop the algorithms for the two estimators, as well as accompanying procedures for selecting the tuning parameters involved in these estimators. In Section 7, we establish the consistency and convergence of the proposed f-GSIR method, and the consistency of a dimension estimation procedure. In Section 8, we investigate the performances of f-GSIR and f-GSAVE by simulation studies. In Section 9, we apply the new methods to two data sets involving speech recognition and handwritten symbols. Some concluding remarks are made in Section 10. Due to limited space all proofs are presented in an Online Supplement [Li and Song (2016)].

2. Functional nonlinear sufficient dimension reduction.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and T_X, T_Y be subsets in $\mathbb{R}^{k_1}$ and $\mathbb{R}^{k_2}$. Let $\mathcal{H}_X$ be a Hilbert space of functions from T_X to $\mathbb{R}^p$, and $\mathcal{H}_Y$ be a Hilbert space of functions from $T_Y \to \mathbb{R}^q$. Let $\mathcal{F}_X$ and $\mathcal{F}_Y$ be the Borel σ-fields generated by the topologies induced by the inner products in $\mathcal{H}_X$ and $\mathcal{H}_Y$. Let $X : \Omega \to \mathcal{H}_X$ and $Y : \Omega \to \mathcal{H}_Y$ be random elements measurable with respect to $\mathcal{F}/\mathcal{F}_X$ and $\mathcal{F}/\mathcal{F}_Y$, respectively. This general framework accommodates both random processes and random fields: if T_X and T_Y are intervals in $\mathbb{R}$, then X and Y are random processes; if T_X and T_Y are subsets of $\mathbb{R}^2$, then X and Y are random fields.

Let P_X and P_Y be the distributions of X and Y; that is, they are induced measures $P \circ X^{-1}$ and $P \circ Y^{-1}$ on $(\mathcal{H}_X, \mathcal{F}_Y)$ and $(\mathcal{H}_Y, \mathcal{F}_Y)$. For simplicity, we assume the pair of random elements takes values in the product space $(\mathcal{H}_X \times \mathcal{H}_Y, \mathcal{F}_X \times \mathcal{F}_Y)$. Let $\sigma(X)$ be the sub σ-field in $\mathcal{F}$ generated by X; that is, $\sigma(X) = X^{-1}(\mathcal{F}_X)$. Let $P_{X|Y} : \mathcal{H}_Y \times \mathcal{F}_X \to \mathbb{R}$ be the conditional distribution of X given Y.

DEFINITION 1. Suppose the family of probability measures $\{P_{X|Y}(\cdot|y) : y \in \mathcal{H}_Y\}$ is dominated by a σ-finite measure. A sub σ-field $\mathcal{G}$ of $\sigma(X)$ is a sufficient dimension reduction (SDR) σ-field for Y verses X iff Y and X are independent given $\mathcal{G}$, or in symbols,

$$Y \perp\!\!\!\perp X | \mathcal{G}. \tag{3}$$

The intersection of all sufficient sub σ-fields is called the central σ-field.

We denote the central σ-field as $\mathcal{G}_{Y|X}$. Following the proof of Theorem 1 of Lee, Li and Chiaromonte (2013) it can be shown that, if $\{P_{X|Y}(\cdot|y) : y \in \mathcal{H}_Y\}$ is dominated by a σ-finite measure, then the central σ-field also satisfies (3). This is the target of functional nonlinear SDR.

Definition 1 generalizes the previously developed functional linear SDR. Dauxois, Ferré and Yao (2001), Ferré and Yao (2003, 2005), and Amato, Antoniadis and De Feis (2006) developed the following framework for functional sufficient dimension reduction

$$Y \perp\!\!\!\perp X | \langle \beta_1, X \rangle_{\mathcal{H}_X}, \dots, \langle \beta_d, X \rangle_{\mathcal{H}_X},$$

where $\beta_1, \dots, \beta_d$ are members of $\mathcal{H}_X$. This is a special case of (3) if we take $\mathcal{G}$ to be the σ-field generated by $\langle \beta_1, X \rangle_{\mathcal{H}_X}, \dots, \langle \beta_d, X \rangle_{\mathcal{H}_X}$. Hsing and Ren (2009) proposed a more general framework of functional linear SDR where $\langle \beta_1, X \rangle_{\mathcal{H}_X}, \dots, \langle \beta_d, X \rangle_{\mathcal{H}_X}$ are replaced by a finite set of random variables in the closure of the span of $\{X(\cdot, t) : t \in T_X\}$, where T_X is a subset of $\mathbb{R}$. Wang, Lin and Zhang (2013) extended Contour Regression of Li, Zha and Chiaromonte (2005) to functional linear SDR. In all of the above extensions, the response Y is a random variable rather than a random function.

The following example illustrates the idea of functional nonlinear SDR in Definition 1.

EXAMPLE 1. Suppose that $\mathcal{H}_X$ is a separable Hilbert space and $\Sigma : \mathcal{H}_X \to \mathcal{H}_X$ is a trace-class operator. Let X be a Gaussian random element in $\mathcal{H}_X$ with Σ as its variance operator. That is, for any $h \in \mathcal{H}_X$ we have $E(e^{i\langle X,h \rangle}) = e^{i\langle \mu,h\rangle - \langle \Sigma h,h\rangle/2}$, where $\langle \cdot, \cdot \rangle$ stands for the inner product in $\mathcal{H}_X$ and $i = \sqrt{-1}$. Let $\beta_1(t) = \sin((3/2)\pi t)$, $\beta_2(t) = \sin((5/3)\pi t)$. Let

$$Y = e^{\langle \beta_1, X \rangle} + \langle \beta_2, X^2 \rangle + 0.3\varepsilon, \tag{4}$$

where $\varepsilon \perp\!\!\!\perp X$ and $\varepsilon \sim N(0,1)$. Then the central σ-field $\mathcal{G}_{Y|X}$ is the σ-field generated by $e^{\langle \beta_1, X \rangle} + \langle \beta_2, X^2 \rangle$, and the central class is spanned by the class of all strictly monotone functions of this random variable in $\mathcal{H}_X$.

3. Nested Hilbert spaces.

For convenience, assume that $T_X = T_Y = T$. As laid out in the last section, the Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ define the ranges of the random functions X and Y. To characterize the distributions of X and Y, and in particular their conditional independence, we need the second-level spaces $\mathfrak{M}_X$ and $\mathfrak{M}_Y$. In this section, we propose a convenient method for constructing these spaces.

A function $f \in \mathcal{H}_X$ has p components: $f = (f_1, \dots, f_p)^{\mathsf{T}}$, where each f_i is a function on T to $\mathbb{R}$. We assume $\mathcal{H}_X = \mathcal{H}_X^0 \times \cdots \times \mathcal{H}_X^0$. Furthermore, we assume that $\mathcal{H}_X^0$ is a Hilbert with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_X^0}$. We define $\mathcal{H}_X$ to be a Hilbert space with inner product

$$\langle f, g\rangle_{\mathcal{H}_X} = \sum_{i=1}^{p} \langle f_i, g_i\rangle_{\mathcal{H}_X^0}.$$

We define $\mathcal{H}_Y^0$ and $\mathcal{H}_Y$ similarly.

Without distinguishing between X and Y, let $\mathcal{H} = \mathcal{H}^0 \times \cdots \times \mathcal{H}^0$ and $\langle \cdot, \cdot\rangle_{\mathcal{H}}$ be as defined above. To construct the second-level space $\mathfrak{M}$, we need a kernel that maps $\mathcal{H} \times \mathcal{H}$ to $\mathbb{R}$. The next definition suggests a convenient way to define such a kernel.

DEFINITION 2. We say that a positive definite kernel $\kappa : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is induced by $\langle \cdot, \cdot\rangle_{\mathcal{H}}$ if there is a function $\rho : \mathbb{R}^3 \to \mathbb{R}^+$ such that, for any $\phi, \psi \in \mathcal{H}$,

$$\kappa(\phi, \psi) = \rho\bigl(\langle \phi, \phi\rangle_{\mathcal{H}}, \langle \phi, \psi\rangle_{\mathcal{H}}, \langle \psi, \psi\rangle_{\mathcal{H}}\bigr).$$

An example of nested kernel is $k(\phi, \psi) = \exp(-\gamma \|\phi - \psi\|_{\mathcal{H}}^2)$, which corresponds to the Gaussian radial basis function often used in machine learning literature, except that the Euclidean scalar product is now replaced by the inner product $\langle \cdot, \cdot\rangle_{\mathcal{H}}$ in $\mathcal{H}$. We define the second-level space $\mathfrak{M}$ to be the RKHS generated by κ. Since the inner product in $\mathfrak{M}$ is uniquely determined by κ, and κ is uniquely determined by $\langle \cdot, \cdot\rangle_{\mathcal{H}}$, the inner product in $\mathfrak{M}$ is uniquely determined by the inner product in $\mathcal{H}$; thus in this sense they are nested. There are many ways to choose $\mathcal{H}$ and ρ. The following examples give several choices of $\mathcal{H}$ and ρ, where $\mathcal{H}$ is itself chosen to be a reproducing kernel Hilbert space generated by a kernel function κ_T defined on $T \times T$.

EXAMPLE 2. Assume $s = 1$, so that $\mathcal{H} = \mathcal{H}^0$. Let $\mathcal{H}_0$ be the reproducing-kernel Hilbert space generated by the Gaussian radial basis kernel on $T \times T$: $\kappa_T(t_1, t_2) = e^{-\gamma_1(t_1-t_2)^2}$ where $\gamma_1 > 0$. It can be shown [see, e.g., Minh (2010)] any function in $\mathcal{H}$ is of the form $\phi(t) = e^{-\gamma_1 t^2} \sum_{k=0}^{\infty} w_k t^k$ such that $\sum_{k=0}^{\infty} k! w_k^2/(2\gamma_1)^k < \infty$. Thus, if we let

$$\mathcal{C} = \left\{ \{w_k\} : \sum_{k=0}^{\infty} \bigl[k! w_k^2/(2\gamma_1)^k\bigr] < \infty \right\},$$

then a function in $\mathcal{H}$ can be equivalently represented as a sequence in $\mathcal{C}$. We will write $\phi \sim \{w_k\}$ to represent this one-to-one correspondence. It can be shown that for any $\phi \sim \{w_k\}$, $\psi \sim \{v_k\}$ in $\mathcal{C}$, we have

$$\langle \phi, \psi\rangle_{\mathcal{H}} = \sum_{k=0}^{\infty} k! w_k v_k/(2\gamma_1)^k.$$

If we take Gauss radial basis function, then the second-level kernel κ for $\mathfrak{M}$ is

$$\kappa: \qquad \mathcal{H} \times \mathcal{H} \to \mathbb{R}, \qquad (\phi, \psi) \mapsto \exp\left(-\gamma_2 \sum_{k=0}^{\infty} [k!(w_k - v_k)^2/(2\gamma_1)^k]\right).$$

The space $\mathfrak{M}$ is the completion of the collection of all functionals that are linear combinations of $\kappa(\cdot, \psi_1), \dots, \kappa(\cdot, \psi_m)$ where $\psi_1, \dots, \psi_m$ are members of $\mathcal{H}$. That is, $\mathfrak{M}$ is the completion of the set of functionals of the form

$$\phi \mapsto \sum_{\ell=1}^{m} c_\ell \exp\left(-\gamma_2 \sum_{k=0}^{\infty} [k!(w_k - v_k^{(\ell)})^2/(2\gamma_1)^k]\right),$$

where $\phi \sim \{w_k\}$, and $\phi_\ell \sim \{v_k^{(\ell)}\}$, $c_\ell \in \mathcal{C}$ for all $\ell = 1, \dots, m$.

EXAMPLE 3. Let $\{U(t) : t \in T\}$ be a random process. Then it can be shown that the mapping $(t_1, t_2) \mapsto \text{cov}[U(t_1), U(t_2)]$ is positive definite [see Berlinet and Thomas-Agnan (2004), page 58]. Then $\kappa_T(t_1, t_2) = \text{cov}[U(t_1), U(t_2)]$ can be used as a kernel to generate RKHS. An important special case is the standard Brownian motion, where this kernel takes the form $\kappa_T(t_1, t_2) = \min(t_1, t_2)$. The space $\mathcal{H}$ generated by this kernel consists of $\phi : T \to \mathbb{R}$ that are absolutely continuous, $\phi(0) = 0$, and $\int [\dot{\phi}(t)]^2 \, dt < \infty$. The inner product in $\mathcal{H}$ is $\langle \phi, \psi \rangle_{\mathcal{H}} = \int \dot{\phi}(t) \dot{\psi}(t) \, dt$. The eigenvalues and eigenfunctions for this kernel in Mercer's theorem are

$$\lambda_j = [(j - 1/2)\pi]^{-2}, \qquad v_j(t) = \sqrt{2} \sin[(j - 1/2)\pi t], \tag{5}$$

respectively; see Amini and Wainwright (2012). The kernel κ for the second-level functional space is then $\kappa(\phi, \psi) = \rho(\int (\dot{\phi}(t) - \dot{\psi}(t))^2 \, dt)$. The space $\mathfrak{M}$ is the RKHS generated by this kernel.

For more choices of kernels κ_T, see, for example, Berlinet and Thomas-Agnan (2004), Appendix and Rasmussen and Williams (2006), Chapter 4. In order for the class of functions $\mathfrak{M}_X$ to characterize the central σ-field, we need to make the following assumption. Let $L_2(P_X)$ denote the class of all functions of X such that $Ef^2(X) < \infty$ under P_X.

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

We now use a subset of $\mathfrak{M}_X$ to characterize functional nonlinear SDR. Comparing to the central σ-field $\mathcal{G}_{Y|X}$, this alternative representation gives a concrete object to estimate.

DEFINITION 3. Under Assumption 1, the class of functions in $\mathfrak{M}_X$ that are $\mathcal{G}_{Y|X}$-measurable is the central dimension reduction class, or the central class.

The central class will be denoted by $\mathfrak{S}_{Y|X}$, and our goal is to recover $\mathfrak{S}_{Y|X}$ from a random sample of (X, Y). If a subspace $\mathfrak{S}$ of $\mathfrak{M}_X$ is contained in $\mathfrak{S}_{Y|X}$, then we say it is unbiased; if it is equal to $\mathfrak{S}_{Y|X}$, then we say it is exhaustive. Similar to Lee, Li and Chiaromonte (2013) we define a complete sub-σ field of $\mathcal{G}$ and complete class as follows.

DEFINITION 4. A sub σ-field $\mathcal{G}$ of $\sigma(X)$ is complete if, for each f measurable with respect to $\mathcal{G}$ such that $E[f(X)|Y] =$ constant almost surely P_Y, we have $f(X) = 0$ almost surely P_X. The class of functions in $\mathfrak{M}_X$ that is measurable with respect to a complete $\mathcal{G}$ is called a complete class.

4. Functional generalized sliced inverse regression. In this section, we extend the Generalized Sliced Inverse Regression [GSIR; Lee, Li and Chiaromonte (2013)] for nonlinear SDR to the functional case. We refer to this extension as functional GSIR, or f-GSIR. For two generic Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, let $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ denote the class of bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$; if $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$, we abbreviate $\mathcal{B}(\mathcal{H}, \mathcal{H})$ by $\mathcal{B}(\mathcal{H})$. For any operator $A \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$, let A^* denote the adjoint operator of A, $\ker(A)$ the kernel of A, $\operatorname{ran}(A)$ the range of A, and $\overline{\operatorname{ran}}(A)$ the closure of the range of A.

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

Let $L_2^{(c)}(P_X) = \{f - Ef(X) : f \in L_2(P_X)\}$ be the centered $L_2(P_X)$ space. Assumption 2 ensures that the mapping $\mathcal{H}_X \to L_2^{(c)}(P_X)$, $f \mapsto f$ is continuous. This assumption guarantees that the bilinear form $\mathfrak{M}_X \times \mathfrak{M}_X \to \mathbb{R}$, $(f, g) \mapsto \operatorname{cov}[f(X), g(X)]$ is bounded. Hence, there is an operator $\Sigma_{XX} \in \mathcal{B}(\mathfrak{M}_X)$ such that $\langle f, \Sigma_{XX} g\rangle_{\mathfrak{M}_X} = \operatorname{cov}[f(X), g(X)]$. We can define $\Sigma_{YY} \in \mathcal{B}(\mathfrak{M}_Y)$, and $\Sigma_{XY} \in \mathcal{B}(\mathfrak{M}_Y, \mathfrak{M}_X)$ in the same way. By definition, Σ_{XX} is self adjoint and $\Sigma^*_{XY} = \Sigma_{YX}$. Similar constructions were used in Fukumizu, Bach and Jordan (2009) and Lee, Li and Chiaromonte (2013). See also Baker (1973).

Although Σ_{XX} is defined on $\mathfrak{M}_X$, its effective domain is the space $\overline{\operatorname{ran}}(\Sigma_{XX})$. This is because members of $\ker(\Sigma_{XX})$ are constants almost surely, which are unimportant to our consideration. Understanding this fact is instructive for constructing estimators at the sample level. The subspace $\overline{\operatorname{ran}}(\Sigma_{XX})$ has an explicit expression, as given by the next lemma. Under Assumption 2, there exists $C > 0$ such that $E|f(X)| < C\|f\|_{\mathfrak{M}_X}$ for any $f \in \mathfrak{M}_X$. This implies that the linear functional $f \mapsto E[f(X)]$ from $\mathfrak{M}_X$ to $\mathbb{R}$ is bounded. Let μ_X be the Riesz representation of this linear functional. Let

$$\mathfrak{M}_X^0 = \overline{\operatorname{span}}\{\kappa_X(\cdot, x) - \mu_X : x \in \mathcal{H}_X\}, \tag{6}$$

where $\overline{\text{span}}$ denotes the closure of linear span. We can define $\mathfrak{M}_Y^0$ and μ_Y in the same way.

LEMMA 1. *Under Assumption 2, we have* $\overline{\text{ran}}(\Sigma_{XX}) = \mathfrak{M}_X^0$ *and* $\overline{\text{ran}}(\Sigma_{YY}) = \mathfrak{M}_Y^0$.

Interestingly, although this explicit expression has been used at the sample level [see, e.g., Fukumizu, Bach and Jordan (2009), Lee, Li and Chiaromonte (2013)], to our knowledge, it has never been stated at the population level.

Since any $f \in \text{ran}(\Sigma_{XX})$ can be uniquely written as $\Sigma_{XX}(g_1 + g_2)$ where $g_1 \in \ker(\Sigma_{XX})$ and $g_2 \in \overline{\text{ran}}(\Sigma_{XX}) = \mathfrak{M}_X^0$, the mapping $f \mapsto g_2$ from $\text{ran}(\Sigma_{XX})$ to $\mathfrak{M}_X^0$ is well defined. We call it the Moore–Penrose inverse of Σ_{XX}, and denote it by $\Sigma_{XX}^\dagger$. For more information about Moore–Penrose inverse in this setting, see Hsing and Eubank (2015), Section 3.5.

ASSUMPTION 3. $\text{ran}(\Sigma_{YX}) \subseteq \text{ran}(\Sigma_{YY})$ and $\Sigma_{YY}^\dagger \Sigma_{YX}$ is a bounded operator.

Note that, under $\text{ran}(\Sigma_{YX}) \subseteq \text{ran}(\Sigma_{XX})$, $\Sigma_{YY}^\dagger \Sigma_{YX}$ is well define. We do not assume the operators $\Sigma_{YY}^\dagger$ to be bounded, which would be unrealistic because we typically assume Σ_{XX} and Σ_{YY} are compact operators, whose eigenvalues decay to 0. However, it is not unreasonable to assume $\Sigma_{XX}^\dagger \Sigma_{XY}$ and $\Sigma_{YY}^\dagger \Sigma_{YX}$ to be bounded, which is determined by the interaction between two operators. In the following, we refer to the bounded operator $\Sigma_{XX}^\dagger \Sigma_{XY}$ as the regression operator, due to its similarity in appearance to the coefficient vector in multivariate linear regression, and denote it by R_{YX}.

PROPOSITION 1. *Under Assumptions 1 through 3 we have, for any* $f \in \mathfrak{M}_X$,

$$\Sigma_{YY}^\dagger \Sigma_{YX} f = E\big(f(X)|Y\big) - Ef(X) + E\big[\big(\Sigma_{YY}^\dagger \Sigma_{YX} f\big)(Y)\big].$$

The next theorem is the theoretical basis for unbiasedness and exhaustiveness of f-GSIR. The proof is similar to that given in Lee, Li and Chiaromonte (2013) for vector-valued X and Y, and is omitted.

THEOREM 1. *Under Assumptions 1 through 3, we have* $\overline{\text{ran}}(R_{YX}^*) \subseteq \text{cl}(\Sigma_{XX}\mathfrak{S}_{Y|X})$. *Furthermore, if* $\mathfrak{S}_{Y|X}$ *is complete, then* $\overline{\text{ran}}(R_{YX}^*) = \text{cl}(\Sigma_{XX}\mathfrak{S}_{Y|X})$, *where* $\text{cl}(\cdots)$ *indicates the closure of a set.*

Intuitively, R_{YX}^* can be calculated as $\Sigma_{XY}\Sigma_{YY}^\dagger$. However, recall that $\Sigma_{YY}^\dagger$ is defined on $\text{ran}(\Sigma_{YY})$ instead of $\overline{\text{ran}}(\Sigma_{YY}) = \mathfrak{M}_Y^0$. Nevertheless, because $\Sigma_{XY}\Sigma_{YY}^\dagger$ is bounded, its domain can be extended to $\overline{\text{ran}}(\Sigma_{YY})$. Thus, we take the domain $R_{YX}^* = \Sigma_{XY}\Sigma_{YY}^\dagger$ as $\overline{\text{ran}}(\Sigma_{YY})$.

Since $\overline{\mathrm{ran}}(R^*_{YX}) = \overline{\mathrm{ran}}(R^*_{YX} A R_{YX})$ for any invertible operator $A : \mathfrak{M}^0_Y \to \mathfrak{M}^0_Y$, at the population level we can use any operator of the form $R^*_{YX} A R_{YX}$ to recover the same portion of $\mathrm{cl}(\Sigma_{XX}\mathfrak{S}_{Y|X})$. Choosing $A = \Sigma_{YY}$ results in the following operator

$$R^*_{YX}\Sigma_{YY}R_{YX} = \Sigma_{XY}\Sigma^\dagger_{YY}\Sigma_{YY}\Sigma^\dagger_{YY}\Sigma_{YX} = \Sigma_{XY}\Sigma^\dagger_{YY}\Sigma_{YX} \tag{7}$$

which resembles SIR in the sense that, for any $f \in \mathfrak{M}_X$,

$$\langle f, R^*_{YX}\Sigma_{YY}R_{YX}f\rangle_{\mathfrak{M}_X} = \langle R_{YX}f, \Sigma_{YY}R_{YX}f\rangle_{\mathfrak{M}_X} = \mathrm{var}\big[E\big(f(X)|Y\big)\big]. \tag{8}$$

To make Theorem 1 into a form that can be mimicked at the sample level, we make the following assumption.

ASSUMPTION 4. For a positive definite operator $A : \mathfrak{M}^0_X \to \mathfrak{M}^0_X$, the operator

$$\Sigma^\dagger_{XX}\Sigma_{XY}A\Sigma_{YX}\Sigma^\dagger_{XX} \tag{9}$$

has finite rank, say d.

Because of the resemblance of (8) to the defining property of SIR, Lee, Li and Chiaromonte (2013), refer to any sample estimator targeting

$$\overline{\mathrm{ran}}(\Sigma^\dagger_{XX}R^*_{YX}\Sigma_{YY}R_{YX}\Sigma^\dagger_{XX}) = \overline{\mathrm{ran}}(\Sigma^\dagger_{XX}R^*_{YX}) = \overline{\mathrm{ran}}(\Sigma^\dagger_{XX}\Sigma_{XY}A\Sigma_{YX}\Sigma^\dagger_{XX})$$

in the nonlinear SDR setting as GSIR. Thus, in the functional nonlinear SDR setting, we refer to any sample estimator targeting the above space as an f-GSIR.

A particularly convenient choice of A is Σ^2_{YY}, because it leads to $\Sigma^\dagger_{XX}\Sigma_{XY}\Sigma_{YX}\Sigma^\dagger_{XX}$, which avoids the inverse $\Sigma^\dagger_{YY}$. In our simulation studies, we do not find significant difference for choosing different A. Therefore, throughout the rest of the paper we take $A = \Sigma^2_{YY}$.

Combining Theorem 1 and Assumption 4, we arrive at the following population-level statement that suggests an algorithm. Henceforth, we call the subspace of functions in $\mathfrak{M}_X$ that are measurable with respect to $\sigma(f_1(X), \ldots, f_d(X))$ the subspace *generated by* $f_1, \ldots, f_d$.

COROLLARY 1. *Suppose Assumptions* 2 *through* 4 *are satisfied. Let* $f_1, \ldots, f_d$ *be solution to the following sequential maximization problem*: *for each* $k = 1, \ldots, d$,

$$\textit{maximize } \langle f, \Sigma^\dagger_{XX}\Sigma_{XY}\Sigma_{YX}\Sigma^\dagger_{XX}f\rangle_{\mathfrak{M}_X}$$

$$\textit{subject to} \qquad f \in \mathfrak{M}^0_X, \langle f, f\rangle_{\mathfrak{M}_X} = 1, \qquad \langle f, f_1\rangle_{\mathfrak{M}_X} = \cdots = \langle f, f_{k-1}\rangle_{\mathfrak{M}_X} = 0.$$

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

Note that we only need to carry out the maximization over $f \in \mathfrak{M}^0_X$ because the domain of $\Sigma_{YX}\Sigma^\dagger_{XX}$ is $\mathfrak{M}^0_X$. This fact will be important for constructing sample estimate.

5. Functional GSAVE. Another important dimension reduction estimator in the classical setting is the Sliced Average Variance Estimator (SAVE) introduced by Cook and Weisberg (1991), which was extended to the nonlinear case by Lee, Li and Chiaromonte (2013), GSAVE. We now further extend it to the functional case. We refer to our extension as the functional GSAVE, or f-GSAVE. Here, we adopt a somewhat different approach than Lee, Li and Chiaromonte (2013): we use RKHS $\mathfrak{M}_X$ and $\mathfrak{M}_Y$ as the basis for extension; whereas in Lee, Li and Chiaromonte (2013) used $L_2(P_X)$ and $L_2(P_Y)$ as the basis for extension. This alternative approach not only makes the extensions of SIR and SAVE more consistent, but also facilitates their asymptotic development, which will be carried out in Section 7.

Similar to the construction of Σ_{XX}, for each $y \in \mathcal{H}_Y$, the mapping

$$\mathfrak{M}_X \times \mathfrak{M}_X \to \mathbb{R}, \qquad (f_1, f_2) \mapsto \operatorname{cov}\big(f_1(X), f_2(X)|Y = y\big) \tag{10}$$

defines a bilinear form, which is bounded under the following assumption.

ASSUMPTION 5. There is a constant $C > 0$ such that, for each $f \in \mathfrak{M}_X$ and $y \in \mathcal{H}_Y$, $\operatorname{var}(f(X)|y) \le C\|f\|^2_{\mathfrak{M}_X}$.

Under Assumption 5, the bounded bilinear form (10) induces an operator $V_{XX|Y}(y) : \mathfrak{M}_X \to \mathfrak{M}_X$ such that

$$\langle f_1, V_{XX|Y}(y) f_2\rangle_{\mathcal{H}_X} = \operatorname{cov}\big(f_1(X), f_2(X)|Y = y\big).$$

The mapping $\Omega_Y \to \mathcal{B}(\mathfrak{M}_X)$, $y \mapsto V_{XX|Y}(y)$ then defines a random operator, which plays the role of $\operatorname{var}(X|Y)$ in SAVE. To proceed further, we need the notion of the expectation of a random operator such as $V_{XX|Y}(Y)$. Let A be any random operator taking values in $\mathcal{B}(\mathfrak{M}_X)$ such that $E\|A\| < \infty$ where $\|\cdot\|$ is the operator norm. Then A defines the bounded bilinear form

$$\mathfrak{M}_X \times \mathfrak{M}_X \to \mathbb{R}, \qquad (f_1, f_2) \mapsto E\langle f_1, A f_2\rangle_{\mathfrak{M}_X},$$

which induces a (nonrandom) operator $B \in \mathcal{B}(\mathfrak{M}_X)$ such that $\langle f_1, B f_2\rangle_{\mathfrak{M}_X} = E\langle f_1, A f_2\rangle_{\mathfrak{M}_X}$. The operator B is defined as the expectation of A, and is written as $E(A)$. That is, the expectation of a random operator is uniquely defined through the equation

$$\langle f_1, E(A) f_2\rangle_{\mathfrak{M}_X} = E\langle f_1, A f_2\rangle_{\mathfrak{M}_X}. \tag{11}$$

Our f-GSAVE is based on the following operator:

$$S \equiv E\big\{\big[\Sigma_{XX} - V_{XX|Y}(Y)\big]^2\big\}.$$

For this expectation to be defined, we need the following assumption.

ASSUMPTION 6. $E(\|\Sigma_{XX} - V_{XX|Y}(Y)\|^2) < \infty$.

Our goal is to show that the range of S is contained in $\Sigma_{XX}\mathfrak{S}_{Y|X}$. To establish this, we need an additional assumption.

ASSUMPTION 7. For any $f \in \mathfrak{M}_X$ such that $\operatorname{cov}[f(X), g(X)] = 0$ for all $g \in \mathfrak{S}_{Y|X}$, the conditional variance $\operatorname{var}[f(X)|\mathcal{G}_{Y|X}]$ is a constant almost surely P_X.

This is an extension of what is called *the constant conditional variance* assumption for SAVE in the classical setting. See, for example, Cook and Weisberg (1991) and Li, Zha and Chiaromonte (2005). Assumption 7 was also used in Lee, Li and Chiaromonte (2013). The next theorem is the theoretical basis of f-GSAVE. Since the assumptions here are different from Lee, Li and Chiaromonte (2013), the proof is also different.

THEOREM 2. *Under Assumptions* 2, 1, 6 *and* 7, *we have* $\overline{\operatorname{ran}}(S) \subseteq \operatorname{cl}(\Sigma_{XX}\mathfrak{S}_{Y|X})$.

According to this theorem, if $\Sigma_{XX}^{\dagger} S \Sigma_{XX}^{\dagger}$ is defined, then its range space is contained in $\mathfrak{M}_X^0 \cap \mathfrak{S}_{Y|X}$, which is equivalent to the central class $\mathfrak{S}_{Y|X}$ because any $f \in \mathfrak{S}_{Y|X}$ can be written as $c\mathbb{1} + f_0$ where $f_0 \in \mathfrak{M}_X^0 \cap \mathfrak{S}_{Y|X}$ and $\mathbb{1}$ is the constant function $\mathbb{1}(f) = 1$ for all $f \in \mathfrak{M}_X$. Similar to the construction of the f-GSIR operator, we make a strong enough assumption so that the range can be recovered via finite steps of optimizations.

ASSUMPTION 8. The operator $\Sigma_{XX}^{\dagger} S \Sigma_{XX}^{\dagger}$ has finite rank d.

We can now restate Theorem 2 in a form that suggests an algorithm.

COROLLARY 2. *Suppose Assumptions* 1, 2, 6, 7 *and* 8 *are satisfied. Let* $f_1, \dots, f_d$ *be solution to the following sequential maximization problem*: *for each* $k = 1, \dots, d$,

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

Then the functions $f_1(X), \dots, f_d(X)$ *generate a subspace of a subspace of* $\mathfrak{M}_X^0 \cap \mathfrak{S}_{Y|X}$.

We call the sample estimate that targets $\overline{\operatorname{ran}}(\Sigma_{XX}^{\dagger} S \Sigma_{XX}^{\dagger})$ an f-GSAVE. Recall that, by Corollary 1, the range of the operator $\Sigma_{XX}^{\dagger}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{\dagger}$ fully recovers the central class $\mathfrak{S}_{Y|X}$ when the latter is complete. When $\mathfrak{S}_{Y|X}$ is not complete, this range can be a proper subspace $\Sigma_{XX}\mathfrak{S}_{Y|X}$, but the operator S recovers a larger subspace of $\mathfrak{S}_{Y|X}$, as guaranteed by the next theorem. The proof is parallel that in Lee, Li and Chiaromonte (2013) and is omitted.

THEOREM 3. *Under Assumptions 3 and 6, we have*

$$\overline{\operatorname{ran}}(R^*_{YX}) \subseteq \overline{\operatorname{ran}}[\Sigma_{XX} - V_{XX|Y}(Y)]$$

almost surely P_Y.

6. Estimation.

Having defined f-GSIR and f-GSAVE at the population level, we now turn to its implementation at the sample level. To focus on the main ideas, we first discuss the case where $p = 1$ and $q = 1$, and then extend to the $p > 1, q > 1$ case. For simplicity, we only consider the case where T is an interval in $\mathbb{R}$; the more general case can be developed by analogy. The basic idea for constructing sample versions of f-GSIR and f-GSAVE is to replace the true distribution of (X, Y) by the empirical distribution based on an i.i.d. sample of (X, Y), and express the relevant operators as $n \times n$ matrices using a coordinate representation system [see Horn and Johnson (1985), Section 0.10].

6.1. *Coordinate representation.* Let $\mathcal{H}_1$ be a generic finite-dimensional vector space with basis $\mathcal{B} = \{b_1, \dots, b_n\}$. For each $f \in \mathcal{H}_1$, there is a vector $\alpha = (\alpha_1, \dots, \alpha_n)^{\mathsf{T}}$ such that $f = \sum_{i=1}^n \alpha_i b_i$. The vector α is called the coordinate of f with respect to $\mathcal{B}$, and is written as $[f]_{\mathcal{B}}$. Through the rest of this section we will reserve the square bracket $[\cdot]$ for coordinate representation by systematically avoiding using it anywhere else. Let $\mathcal{H}_2$ be another Hilbert spaces, spanned by $\mathcal{C} = \{c_1, \dots, c_m\}$ and $A : \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator. Then, for any $f \in \mathcal{H}_1$,

$$Af = A\left(\sum_{i=1}^n ([f]_B)_i b_i\right) = \sum_{i=1}^n ([f]_B)_i (Ab_i) = \sum_{i=1}^n ([f]_B)_i \sum_{j=1}^m ([Ab_i]_C)_j c_j.$$

The right-hand side can be rewritten as

$$\sum_{j=1}^m \sum_{i=1}^n ([f]_B)_i ([Ab_i]_C)_j c_j \equiv \sum_{j=1}^m \{({}_{\mathcal{C}}[A]_{\mathcal{B}})([f]_B)\}_j c_j,$$

where ${}_{\mathcal{C}}[A]_{\mathcal{B}}$ is the $m \times n$ matrix whose (j, i)th entry is $([Ab_i]_C)_j$. The above equation shows that $[Af]_{\mathcal{C}} = ({}_{\mathcal{C}}[A]_{\mathcal{B}})[f]_{\mathcal{B}}$. The matrix ${}_{\mathcal{C}}[A]_{\mathcal{B}}$ is called the coordinate of the operator A relative to bases $\mathcal{B}$ and $\mathcal{C}$. Let $\mathcal{H}_3$ be a third Hilbert space with basis $\mathcal{D} = \{d_1, \dots, d_\ell\}$ and $B : \mathcal{H}_2 \to \mathcal{H}_3$ be a linear operator. Then ${}_{\mathcal{D}}[BA]_{\mathcal{B}} = ({}_{\mathcal{D}}[B]_{\mathcal{C}})({}_{\mathcal{C}}[A]_{\mathcal{B}})$. In the following, if the bases involved are clear from the context, we will drop the subscripts and write ${}_{\mathcal{D}}[B]_{\mathcal{C}}$ and $[f]_{\mathcal{B}}$ as $[B]$ and $[f]$. We write the ith component of $[f]$ as $[f]_i$.

6.2. *Construction of $\mathcal{H}_X, \mathcal{H}_Y, \mathfrak{M}_X, \mathfrak{M}_Y$.* We first construct $\mathcal{H}_X$ and $\mathcal{H}_Y$, using $\mathcal{H}_X$ as an illustration. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be an i.i.d. sample of (X, Y). Suppose, rather than observing the function X_i in its entirety, we only observe it on a finite subset $\{t_{i1}, \dots, t_{im_i}\}$ of T, which is allowed to differ from subject to subject. Let $u = \sum_{i=1}^{n} m_i$ and

$$\tau = (t_{11}, \dots, t_{1m_1}, \dots, t_{n1}, \dots, t_{nm_n})^{\mathsf{T}} \equiv (\tau_1, \dots, \tau_u)^{\mathsf{T}}.$$

Let J_i be the set of indices of τ_k that belongs to the ith subject, and $T_i = \{\tau_k : k \in J_i\}$. For simplicity, assume that all time points are distinct: $\tau_k \neq \tau_\ell$ whenever $k \neq \ell$. When time points overlap, the following development is still valid with slight modification. Let K_T be the $u \times u$ Gram matrix $\{\kappa_T(\tau_k, \tau_\ell)\}$. Using the coordinate system in Section 6.1, we can express the inner product in $\mathcal{H}_X$ as

$$\langle \phi, \psi \rangle = \left\langle \sum_{k=1}^{u} [\phi]_k \kappa_T(\cdot, \tau_k), \sum_{\ell=1}^{u} [\psi]_\ell \kappa_T(\cdot, \tau_\ell) \right\rangle = [\phi]^{\mathsf{T}} K_T [\psi].$$

Since each X_i is observed at m_i time points, it only needs m_i functions in $\mathcal{H}_X$ to specify. Naturally, we choose these functions to be $\{\kappa_T(\cdot, \tau_k) : k \in J_i\}$. This means all except m_i entries of $[X_i]$ are 0. With this in mind, we have

$$X_i = \sum_{k=1}^{u} [X_i]_k \kappa_T(\cdot, \tau_k) = \sum_{k \in J_i} [X_i]_k \kappa_T(\cdot, \tau_k).$$

Furthermore, for each $\ell \in J_i$,

$$X_i(\tau_\ell) = \langle X_i, \kappa_T(\cdot, \tau_\ell) \rangle_{\mathcal{H}_X} = \sum_{k \in J_i} [X_i]_k \kappa_T(\tau_\ell, \tau_k).$$

Let $[X_i]^0$ denote the m_i-dimensional subvector of $[X_i]$ consisting of entries with indices in J_i, $K_T^{(i,j)}$ denote the $m_i \times m_j$ sub-matrix $\{(K_T)_{k\ell} : k \in J_i, \ell \in J_j\}$, and $X_i(T_i)$ denote the (column) vector $\{X_i(\tau) : \tau \in T_i\}$. Then $X_i(T_i) = K_T^{(i,i)}[X_i]^0$. Solve this equation with Tychonoff regularization to obtain

$$[X_i]^0 = \left(K_T^{(i,i)} + \varepsilon_T^{(X)} I_{m_i}\right)^{-1} X_i(T_i).$$

We define $\mathcal{H}_X$ as the space spanned by $\{\sum_{\ell=1}^{m_i} [X_i]_\ell^0 \kappa_T(\cdot, t_{i\ell}) : i = 1, \dots, n\}$ with inner product determined by $\langle \kappa_T(\cdot, s), \kappa_T(\cdot, t) \rangle_{\mathcal{H}_X} = \kappa_T(s, t)$. It follows that

$$\begin{aligned} \langle X_i, X_j \rangle_{\mathcal{H}_X} &= [X_i]^{0\mathsf{T}} K_T^{(i,j)} [X_j]^0 \\ &= X_i^{\mathsf{T}}(T_i)\left(K_T^{(i,i)} + \varepsilon_T^{(X)} I_{m_i}\right)^{-1} K_T^{(i,j)} \left(K_T^{(j,j)} + \varepsilon_T^{(X)} I_{m_j}\right)^{-1} X_j(T_j). \end{aligned} \tag{12}$$

We define $\mathcal{H}_Y$ similarly.

We now turn to the second-level spaces $\mathfrak{M}_X$ and $\mathfrak{M}_Y$, using $\mathfrak{M}_X$ as an illustration. For each $i, j = 1, \dots, n$, let

$$\begin{aligned} \kappa_X(X_i, X_j) &= \rho\left(\langle X_i - X_j, X_i - X_j \rangle_{\mathcal{H}_X}\right) \\ &= \rho\left(\langle X_i, X_i \rangle_{\mathcal{H}_X} - 2\langle X_i, X_j \rangle_{\mathcal{H}_X} + \langle X_j, X_j \rangle_{\mathcal{H}_X}\right), \end{aligned} \tag{13}$$

where ρ is the function introduced in Definition 2, and $\langle X_i, X_i\rangle_{\mathcal{H}_X}$ and so on are calculated by (12). Let $\mathfrak{M}_X$ be the space spanned by $\{\kappa_X(\cdot, X_i) : i = 1, \dots, n\}$ and K_X be $n \times n$ Gram matrix $\{\kappa_X(X_i, X_j)\}$. Then, for $f_1, f_2 \in \mathfrak{M}_X$, $\langle f_1, f_2\rangle_{\mathfrak{M}_X} = [f_1]^{\mathsf{T}} K_X [f_2]$.

6.3. *Implementation of f-GSIR.* To develop the f-GSIR estimator, we first derive the coordinates of Σ_{XX}, Σ_{YY}, and Σ_{YX}. With $\mathfrak{M}_X$ defined in the last subsection, let μ_X be the Riesz representation of the linear functional $f \mapsto E_n f(X)$ defined on $f \in \mathcal{H}_X$; let $\mathfrak{M}_X^0 = \operatorname{span}\{\kappa_X(\cdot, X_i) - \mu_X : i = 1, \dots, n\}$. Furthermore, let 1_n denote the n-dimension vector with its components identically 1, I_n denote the $n \times n$ identity matrix, and $Q = I_n - 1_n 1_n^{\mathsf{T}}/n$. Note that Q is the projection on to the orthogonal complement of $\operatorname{span}(1_n)$ in the Euclidean space $\mathbb{R}^n$. The following facts are easily checked.

PROPOSITION 2. *Let $\mathfrak{M}_X$, μ_X, and $\mathfrak{M}_X^0$ be as defined in this section. Then*:

1. $\mu_X = n^{-1} \sum_{i=1}^n \kappa_X(\cdot, X_i)$;
2. $f \in \mathfrak{M}_X^0$ *iff* $[f] = Q[f]$.

As we mentioned in Section 4, the relevant subspace of dimension reduction is $\mathfrak{M}_X^0$, because functions in its orthogonal complement are constant almost surely P_X. To reflect this at the sample level our coordinate representation for operators such as Σ_{XX}, Σ_{YX} and Σ_{YY} are with respect to $\mathfrak{M}_X^0$ rather than $\mathfrak{M}_X$. Using Proposition 2, we can easily prove the following results, which can be found in Fukumizu, Bach and Jordan (2009) and Lee, Li and Chiaromonte (2013). Let G_X and G_Y denote QK_XQ and QK_Y.

PROPOSITION 3. *We have the following coordinate expression at the sample level*:

$$[\Sigma_{XX}] = n^{-1} G_X, \qquad [\Sigma_{YY}] = n^{-1} G_Y, \qquad [\Sigma_{YX}] = n^{-1} G_X,$$
$$[\Sigma_{XY}] = n^{-1} G_Y, \qquad [\Sigma_{XX}^{\dagger}] = n G_X^{\dagger}, \qquad [\Sigma_{YY}^{\dagger}] = n G_Y^{\dagger}.$$

By Proposition 3, we can express the quantities in Corollary 1 in matrix form. The operator $\Sigma_{XX}^{\dagger} R_{YX}^{*} \Sigma_{YY}^{2} R_{YX} \Sigma_{XX}^{\dagger}$ can be expressed as

$$(n^{-1} G_X)^{\dagger} (n^{-1} G_Y)(n^{-1} G_X)(n^{-1} G_X)^{\dagger} = G_X^{\dagger} G_Y G_X G_X^{\dagger}.$$

Hence,

$$\begin{aligned}\langle f, (R_{YX}\Sigma_{XX}^{\dagger})^{*} \Sigma_{YY}^{2} (R_{YX}\Sigma_{XX}^{\dagger}) f\rangle_{\mathfrak{M}_X} &= [f]^{\mathsf{T}} K_X G_X^{\dagger} G_Y G_X G_X^{\dagger} [f] \\ &= [f]^{\mathsf{T}} G_X G_X^{\dagger} G_Y G_X G_X^{\dagger} [f],\end{aligned}$$

where the second equality follows from $[f] = Q[f]$ and $G_X^\dagger = QG_X^\dagger$. To prevent over-fitting, we use the Tychonoff regularized inverse $(G_X + \varepsilon_X I_n)^{-1}$ to replace the Moore–Penrose inverse $G_X^\dagger$, where $\varepsilon_X > 0$ is a tuning constant, which results in

$$[f]^\mathsf{T} G_X(G_X + \varepsilon_X I_n)^{-1} G_Y G_X (G_X + \varepsilon_X I_n)^{-1}[f]. \tag{14}$$

Since G_X and $(G_X + \varepsilon_X I_n)^{-1}$ commute, the matrix sandwiched between $[f]^\mathsf{T}$ and $[f]$ is symmetric.

Because it suffices to maximize (14) over $\mathfrak{M}_X^0$, the inner products in Corollary 1 are $\langle f, f\rangle_{\mathfrak{M}_X} = [f]^\mathsf{T} G_X[f]$ and $\langle f, f_\ell\rangle_{\mathfrak{M}_X} = [f]^\mathsf{T} G_X[f_\ell]$. Put $v = G_X^{1/2}[f]$ and solve for $[f]$ with Tychonoff regularization to obtain $[f] = (G_X + \varepsilon_X I_n)^{-1} v$. In terms of v, Corollary 1 is implemented as the following standard eigenvalue problem: for $k = 1, \dots, d$,

$$\begin{aligned} &\text{maximize: } v^\mathsf{T}(G_X + \varepsilon_X I)^{-3/2} G_X G_Y G_X (G_X + \varepsilon_X I)^{-3/2} v \\ &\text{subject to:} \qquad v^\mathsf{T} v = 1, v^\mathsf{T} v_1 = 0, \dots, v^\mathsf{T} v_{k-1} = 0. \end{aligned}$$

In other words, $v_1, \dots, v_d$ are the first d eigenvectors of

$$(G_X + \varepsilon_X I)^{-3/2} G_X G_Y G_X (G_X + \varepsilon_X I)^{-3/2}. \tag{15}$$

We then retrieve the coefficients $[f_\ell] = (G_X + \varepsilon_X I)^{-1/2} v_\ell$. Our final product is the set of functions

$$\hat{f}_\ell = v_\ell^\mathsf{T}(G_X + \varepsilon_X I_n)^{-1/2} Q b_X, \qquad \ell = 1, \dots, d,$$

which are the (nonlinear) sufficient predictors that span the approximate central class.

6.4. *Implementation of f-GSAVE.* We first need a procedure to estimate the conditional covariance

$$\operatorname{cov}\bigl(f_1(X), f_2(X)|y\bigr) = E\bigl(f_1(X) f_2(X)|y\bigr) - E\bigl(f_1(X)|y\bigr) E\bigl(f_2(X)|y\bigr) \tag{16}$$

for each function y and any $f_1, f_2 \in \mathfrak{M}_X$. For this purpose, we need to estimate $E(f(X)|y)$ for any $f \in \mathfrak{M}_X$. We introduce a few more operators

$$\begin{aligned} &M_{YX} : \text{induced by the bilinear form } \mathfrak{M}_X \times \mathfrak{M}_Y \to \mathbb{R}, \qquad (f, g) \mapsto E(fg), \\ &M_{YY} : \text{induced by the bilinear form } \mathfrak{M}_Y \times \mathfrak{M}_Y \to \mathbb{R}, \qquad (f, g) \mapsto E(fg), \qquad (17) \\ &E_{YX} = M_{YY}^{-1} M_{YX}. \end{aligned}$$

By the similar argument used in Section 4, it can be shown that, for any $f \in \mathfrak{M}_X$, $E_{YX} f = E(f(X)|Y)$. These operators are similar to and Σ_{YX}, Σ_{YY}, and R_{YX} except they are not centered. This is because the centering is automatically done by the expression (16) itself. Using these operators, we can express (16) as

$$\operatorname{cov}_n(f_1, f_2|y) = \bigl\{E_{YX}(f_1 f_2)\bigr\}(y) - \bigl\{(E_{YX} f_1)(y)\bigr\}\bigl\{(E_{YX} f_1)(y)\bigr\}. \tag{18}$$

By the argument similar to that used in Lee, Li and Chiaromonte (2013), it can be shown that the coordinate representation (in $\mathfrak{M}_X$) of these operators are

$$[M_{YY}] = n^{-1}K_Y, \qquad [M_{YX}] = n^{-1}K_X, \qquad [E_{YX}] = K_Y^{-1}K_X.$$

The next lemma gives the coordinate representation of $f_1 f_2$.

LEMMA 2. *For any* $f_1, f_2 \in \mathfrak{M}_X$, *we have* $[f_1 f_2] = K_X^{-1}(K_X[f_1] \odot K_X[f_2])$, *where* $\odot$ *is the Hadamard product.*

We are now ready to derive the explicit expression for $\mathrm{cov}_n(f_1, f_2|y)$. In the following, we will use the equality $(a \odot b)^{\mathsf{T}} c = a^{\mathsf{T}} \operatorname{diag}(c) b$, which can be easily verified.

LEMMA 3. *We have* $\mathrm{cov}_n(f_1, f_2|y) = [f_1]^{\mathsf{T}} A(y)[f_2]$ *where*

$$A(y) = K_X \operatorname{diag}\bigl(K_Y b_Y(y)\bigr)K_X - K_X K_Y^{-1} b_Y(y) b_Y^{\mathsf{T}}(y) K_Y^{-1} K_X. \tag{19}$$

The next theorem gives the coordinate representation of $V_{XX|Y}(y)$.

THEOREM 4. $[V_{XX|Y}(y)] = G_X^{\dagger} Q A(y) Q$.

We now describe an algorithm that implement the procedure in Corollary 2 at the sample level. For ease of exposition, we will keep using $G_X^{\dagger}$ in the following development. In actual estimation, like in the case of f-GSIR, we replace it by the Tychonoff regularized inverse $(G_X + \varepsilon_X I_n)^{-1}$.

By Corollary 2, we need to maximize $\langle f, \Sigma_{XX}^{\dagger} S \Sigma_{XX}^{\dagger} f\rangle_{\mathfrak{M}_X}$ successively over $f \in \mathfrak{M}_X^0$ under the constraint $\|f\|_{\mathfrak{M}_X} = 1$ and that f is orthogonal (in $\mathfrak{M}_X$) to the previously found maximizers. By (11),

$$\langle f, \Sigma_{XX}^{\dagger} S \Sigma_{XX}^{\dagger} f\rangle_{\mathfrak{M}_X} = E_n\bigl(\langle f, \Sigma_{XX}^{\dagger}(\Sigma_{XX} - V_{XX|Y}(Y))^2 \Sigma_{XX}^{\dagger} f\rangle_{\mathfrak{M}_X}\bigr).$$

By Theorem 4,

$$\begin{aligned}&\langle f, \Sigma_{XX}^{\dagger}(\Sigma_{XX} - V_{XX|Y}(Y))^2 \Sigma_{XX}^{\dagger}\rangle_{\mathfrak{M}_X}\\ &\quad = [f]^{\mathsf{T}} Q\bigl(G_X - G_X^{\dagger}A(Y)\bigr)^2 G_X^{\dagger} Q[f]\\ &\quad = [f]^{\mathsf{T}}\bigl(Q - G_X^{\dagger}A(Y)G_X^{\dagger}\bigr)G_X\bigl(Q - G_X^{\dagger}A(Y)G_X^{\dagger}\bigr)[f].\end{aligned}$$

Hence,

$$\begin{aligned}&\langle f, \Sigma_{XX}^{\dagger} S \Sigma_{XX}^{\dagger} f\rangle_{\mathfrak{M}_X}\\ &\quad = [f]^{\mathsf{T}} E_n\bigl\{\bigl(Q - G_X^{\dagger}A(Y)G_X^{\dagger}\bigr)G_X\bigl(Q - G_X^{\dagger}A(Y)G_X^{\dagger}\bigr)\bigr\}[f].\end{aligned} \tag{20}$$

To transform this into a standard eigenvalue problem, where the constraints are in terms of scalar product $a^{\mathsf{T}} b$ in $\mathbb{R}^n$ rather than $\langle f, g\rangle_{\mathfrak{M}_X}$, let $v = G_X^{1/2}[f]$. Then

$[f] = G_X^{\dagger 1/2} v$. The maximizer of (20) is of the form $G_X^{\dagger 1/2} v$ where v is the eigenvector of

$$E_n\{QG_X^{\dagger 1/2}(Q - G_X^{\dagger}A(Y)G_X^{\dagger})G_X(Q - G_X^{\dagger}A(Y)G_X^{\dagger})G_X^{\dagger 1/2}Q\}. \tag{21}$$

To sum up the f-GSAVE algorithm, we first replace the $G_X^{\dagger}$ in (21) by $(G_X + \varepsilon_X I_n)^{-1}$ (including the $G_X^{\dagger}$ in $G_X^{-1/2}$) and K_Y in $A(y)$ by $(K_Y + \varepsilon_Y I_n)$, where $\varepsilon_X > 0$ and $\varepsilon_Y > 0$ are constants to be determined by the tuning method in Section 6.6. With these replacements in place, we compute the first d eigenvectors $v_1, \dots, v_d$ of (21). Our sufficient predictors are $f_1, \dots, f_d$, where $[f_\ell] = (G_X + \varepsilon_X I_n)^{-1/2} v_\ell$, $\ell = 1, \dots, n$.

The next three subsections are devoted to tuning parameters, of which there are three classes. The first class consists of tuning parameters for $\mathcal{H}_X$ and $\mathcal{H}_Y$, which include the Tychonoff regularization constants $\varepsilon_T^{(X)}$ and $\varepsilon_T^{(Y)}$, and, if the Gauss radial basis function is used, the constants $\gamma_T^{(X)}$ and $\gamma_T^{(Y)}$. This is discussed in Section 6.5. The second class of tuning parameters are for $\mathfrak{M}_X$ and $\mathfrak{M}_Y$, which are $\varepsilon_X, \varepsilon_Y, \gamma_X$ and γ_Y (if the Gauss radial basis function is used). This is discussed in Section 6.6. The third-class consists of one tuning parameter: the dimension d of the central class. This is discussed in Section 6.7. In the following, we assume that the GRB kernel is used, but the basic ideas apply to other kernels as well.

6.5. *Tuning first-level functions.* Since the procedures for tuning $(\gamma_T^{(X)}, \varepsilon_T^{(X)})$ and $(\gamma_T^{(Y)}, \varepsilon_T^{(Y)})$ are the same, we focus on the X-version and omit the superscripts of γ_T and ε_T. First, we recommend the default value for ε_T to be $\varepsilon_T^0 = 0.05\hat{\lambda}_1(K_T)$, where $\hat{\lambda}_1(K_T)$ the largest eigenvalue of K_T, and the default value for γ_T to be $\gamma_T^0 = 1/(2\rho)$ where $\rho = \binom{u}{2}^{-1}\sum_{i<j}(\tau_i - \tau_j)^2$.

We can either use the default values as the tuning parameters or use generalized cross validation (GCV) to fine tune the parameters around the default values. Recall from Section 6.2 that we fit the function X_i by its observed values $\{X_i(\tau_k) : k \in J_i\}$ according to

$$\hat{X}_i(t, \gamma_T, \varepsilon_T) = X_i^{\mathsf{T}}(T_i)\big(K_T^{(i,i)}(\gamma_T) + \varepsilon_T I_{m_i}\big)^{-1}\kappa_T(t, T_i; \gamma_T), \qquad t \in T,$$

where we have put γ_T in $K_T^{(i,i)}$ and κ_T to emphasize their dependence on it. Let

$$\begin{aligned}\operatorname{gcv}(\gamma_T, \varepsilon_T) &= \sum_{i=1}^{n} \frac{m_i^{-1}\sum_{j=1}^{m_i}[X_i(t_{ij}) - \hat{X}_i(t_{ij}, \gamma_T, \varepsilon_T)]^2}{\{m_i^{-1}\operatorname{tr}[I_{m_i} - \hat{H}_i(\gamma_T, \varepsilon_T)]\}^2} \\ &= \sum_{i=1}^{n} \frac{m_i^{-1}\|[I_{m_i} - \hat{H}_i(\gamma_T, \varepsilon_T)]X_i(T_i)\|^2}{\{m_i^{-1}\operatorname{tr}[I_{m_i} - \hat{H}_i(\gamma_T, \varepsilon_T)]\}^2},\end{aligned}$$

where $\hat{H}_i(\gamma_T, \varepsilon_T)$ is the matrix $K_T^{(i,i)}(\gamma_T)(K_T^{(i,i)}(\gamma_T) + \varepsilon_T I_{m_i})^{-1}$, a smoothed projector that projects the random function X_i on to the important eigenfunctions of the kernel $K_T^{(i,i)}$. This criterion pools observations from all subjects to determine ε_T, and allows J_i to overlap for different subjects. In particular, in the balanced case the criterion reduces to

$$\operatorname{gcv}(\varepsilon_T, \gamma_T) = \frac{m}{\operatorname{tr}[I_m - \hat{H}(\gamma_T, \varepsilon_T)]^2} \sum_{i=1}^{n} \| [I_m - \hat{H}(\gamma_T, \varepsilon_T)] X_i(T_i) \|^2.$$

We minimize $\operatorname{gcv}(\gamma_T, \varepsilon_T)$ over a grid of γ_T in $[\gamma_T^0/20, 20\gamma_T^0]$ and a grid of ε_T in $[\varepsilon_T^0/50, 50\varepsilon_T^0]$ using the Gauss–Seidel method. Each grid consists of 20 points, equally spaced in log scale.

6.6. *Tuning second-level functions.* Next, we tune the parameters γ_X, ε_X, γ_Y, ε_Y in the second-level functions, again focusing on the X-version. These procedures are further developed from the tuning methods in Li, Chun and Zhao (2012, 2014) and Lee, Li and Chiaromonte (2013).

We recommend the default values γ_X^0 and ε_X^0 to be the same as γ_T^0 and ε_T^0 but with $K_T^{(i,i)}$ replaced by K_X, u by n, and $|\tau_i - \tau_j|$ by $\|X_i - X_j\|$. We then use the leave-one-out cross validation (LOOCV) to fine tune γ_X and ε_X around their default values. In nonparametric regression, the LOOCV criterion is defined by $\sum_{i=1}^{n} \{Y_i - \hat{E}^{(-i)}(Y_i|X_i)\}^2$, where $\hat{E}^{(-i)}(Y|X)$ is some nonparametric estimate of the conditional expectation $E(Y|X)$ based on the sample with the ith subject left out. In our setting, however, the response is a space of functions of y spanned by $\{\kappa_Y(\cdot, Y_i), i = 1, \dots, n\}$ rather than a single y. This means we need to fit not one but a set of functions. For each $i = 1, \dots, n$, we predict $g_j = \kappa(\cdot, Y_j)$ for $j \neq i$ by $E_{XY}^{(-i)} g_j$, where $E_{XY}^{(-i)}$ is the third operator in (17) applied to the sample with the ith subject removed. That is, $[E_{XY}^{(-i)}] = (K_X^{(-i)} + \varepsilon_X I_{n-1})^{-1} K_Y^{(-i)}$, where $K_X^{(-i)}$ and $K_Y^{(-i)}$ are kernel matrices based on the sample with ith subject left out. The sum of squared errors of these predictions is

$$\sum_{i=1}^{n} \sum_{j \neq i} \{\kappa_Y(Y_i, Y_j) - (E_{XY}^{(-i)} \kappa_Y(\cdot, Y_j))(X_i)\}^2. \tag{22}$$

As a member of $\mathfrak{M}_X$, the function $E_{XY}^{(-i)} \kappa_Y(\cdot, Y_j)$ has coordinate

$$[E_{XY}^{(-i)} \kappa_Y(\cdot, Y_j)] = [E_{XY}^{(-i)}][\kappa_Y(\cdot, Y_j)],$$

where the coordinate of $\kappa_Y(\cdot, Y_j)$ with respect to the set $\{\kappa(\cdot, Y_j) : j \neq i\}$ is $e_j^{(n-1)}$, the jth column of I_{n-1}. Thus, $E_{XY}^{(-i)} \kappa_Y(\cdot, Y_j)$ is the mapping

$$x \mapsto (e_j^{(n-1)})^{\mathsf{T}} [E_{XY}^{(-i)}]^{\mathsf{T}} \kappa_X(X^{(-i)}, x),$$

where $\kappa_X(X^{(-i)}, x)$ denotes the vector $(\kappa_X(X_1, x), \dots, \kappa_X(X_n, x))^\mathsf{T}$ with its ith component removed. Criterion (22) can now be written as

$$\begin{aligned}\operatorname{loocv}(\gamma_X, \varepsilon_X) &= \sum_{i=1}^{n} \sum_{j \neq i} \{\kappa_Y(Y_i, Y_j) - (e_j^{(n-1)})^\mathsf{T} [E_{XY}^{(-i)}]^\mathsf{T} \kappa_X(X^{(-i)}, X_i)\}^2 \\ &= \sum_{i=1}^{n} \|\kappa_Y(Y^{(-i)}, Y_i) - [E_{XY}^{(-i)}]^\mathsf{T} \kappa_X(X^{(-i)}, X_i)\|^2,\end{aligned}$$

where $\|\cdot\|$ is the norm in the Euclidean space $\mathbb{R}^{n-1}$. We then use the Gauss–Seidel algorithm to minimize $\operatorname{loocv}(\gamma_X, \varepsilon_X)$ over the grids defined in the same way as before, except that $\gamma_T^0, \varepsilon_T^0$ are replaced by $\gamma_X^0, \varepsilon_X^0$.

The tuning parameters $(\gamma_Y, \varepsilon_Y)$ are determined in the same way with the roles of X and Y reversed.

6.7. *Dimension determination.* We propose two strategies to determine the dimension d. When the response Y is a categorical variable with k categories and k is relatively small (say $k \le 8$), we propose to choose d to be $k - 1$. The rational for this choice is that k points can occupy at most a $k - 1$ dimensional Euclidean space. Thus, if the clusters were points then $k - 1$ would be an upper bound of d. More generally, if all clusters are of the same shape, then d is upper bounded by $k - 1$. We choose d to be $k - 1$ to allow extra dimensions that may be caused by the different shapes of the clusters. Our experiences indicate that the shapes of clusters in the sufficient-predictor plot derived from nonlinear SDR are more regular than those derived from linear SDR, because the kernel mapping in nonlinear SDR "absorbed" nonlinearity from the clusters. For this reason, $d = k - 1$ is large enough in many practical problems.

For continuous response or categorial response with a large number of categories, we propose a criterion similar to CVBIC (cross-validated BIC) introduced by Li, Artemiou and Li (2011). Let

$$G_n(k) = \sum_{i=1}^{k} \hat{\lambda}_i - a\hat{\lambda}_1 n^{\alpha\beta/(1+\beta)} \log(n) k, \tag{23}$$

where $\hat{\lambda}_i$ are the eigenvalues of the matrix representations of f-GSIR and f-GSAVE, as given in (15) and described at the end of Section 6.4. The numbers $\alpha \in (0, 1/2]$ and $\beta \in (0, 1]$ are the constants in the optimal convergence rate to be derived in Section 7: α represents the frequency of the measurement schedule for the functional data, which is closer to 1/2 is the schedule is more frequent, and close to 0 if it is more sparse; β represents the smoothness of the relation between X and Y, with β closer to 1 representing a smoother relation. In all the examples in this paper, we take $\alpha = 1/2$ and $\beta = 1$, so that $n^{-\alpha\beta/(1+\beta)} = n^{-1/4}$. We propose to estimate d by

$$\hat{d} = \operatorname{argmax}\{G_n(k) : k = 0, 1, \dots, n\}. \tag{24}$$

In Section 7, we prove the consistency of this criterion for the case of f-GSIR.

The constant a in (23) is determined by LOOCV, as follows. For each fixed a in a grid, we find $\hat{d}(a)$ according to (24), and compute the first $\hat{d}(a)$ eigenfunctions $U = (\hat{f}_1(X), \dots, \hat{f}_{\hat{d}(a)}(X))^{\mathsf{T}}$ of the f-GSIR or f-GSAVE operator. We then use U to replace X and perform LOOCV as described in Section 6.6. Since the dimension of U is relatively low, we set the Tychonoff regularization tuning constant $\varepsilon_U = 0$ for computing $E_{UY}^{(-1)}$ in (22). We use the default value γ_U^0 for tuning parameter in the kernel κ_U. This round of cross validation is to be performed after all the other tuning parameters have been determined and fixed. We take the grid to be 20 points placed in $[0.1, 1]$ with equal distance in $\log_e$ scale.

In Section 6.6, we prove the consistency of the criterion (23). The BIC aspect of this criterion is related to the BIC-type criteria in Zhu, Miao and Peng (2006), Wang and Yin (2008) and Li, Li and Zhu (2010) in the classical sufficient dimension reduction setting.

6.8. *Vector-valued functions of t.* When $p > 1$ and $q > 1$, the estimation procedures remain largely the same except that $\mathcal{H}_X$ and $\mathcal{H}_Y$ and their inner products need to be redefined. Let

$$X_i(t) = \left(X_i^1(t), \dots, X_i^p(t)\right)^{\mathsf{T}}, \qquad Y_i(t) = \left(Y_i^1(t), \dots, Y_i^p(t)\right)^{\mathsf{T}}$$

be vector-valued functions of t, and let $(X_1, Y_1), \dots, (X_n, Y_n)$ be an i.i.d. sample of (X, Y). We construct $\mathcal{H}_{X^i}$ for each component X^i of X as described in Section 6.2 using a kernel κ_T, which, for convenience, is taken to be the same for all components. We then define $\mathcal{H}_X$ for the vector-valued function $t \mapsto X(t)$ as the direct sum $\mathcal{H}_{X^1} \oplus \cdots \oplus \mathcal{H}_{X^p}$; that is:

1. a function $\phi \in \mathcal{H}_X$ is a vector-valued function $(\phi_1, \dots, \phi_p)^{\mathsf{T}}$ where $\phi_i \in \mathcal{H}_{X^i}$;
2. the inner product between two functions $\phi, \psi \in \mathcal{H}_X$ is $\sum_{i=1}^p \langle \phi_i, \psi_i \rangle_{\mathcal{H}_{X^i}}$.

We define H_Y similarly. The rest of the algorithm is the same as the $p = q = 1$ case.

7. Asymptotic analysis.

In this section, we establish the consistency and convergence rate of f-GSIR. We first derive the convergence rate assuming the random functions X_i and Y_i are fully observed, and then extend the result to accommodate the situations where they are not fully observed. We also give the consistency of the dimension determination method. Although we focus on f-GSIR due to limited space, the techniques employed here are fully applicable to f-GSAVE.

7.1. *Convergence rate for fully observed functional data.* The next lemma is proved similarly to Lemma 5 of Fukumizu, Bach and Gretton (2007).

LEMMA 4. *If $E[\kappa_X(X,X)] < \infty$, $E[\kappa_Y(Y,Y)] < \infty$, then Σ_{XX}, Σ_{YY} and Σ_{YX} are Hilbert–Schmidt operators and*

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

Let M and $\hat{M}$ denote the population- and sample-level f-GSIR operators in Sections 4 and 6.3 with A taken to be Σ_{YY}^2; that is,

$$\begin{aligned} M &= \Sigma_{XX}^{\dagger}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{\dagger}, \\ \hat{M} &= (\hat{\Sigma}_{XX} + \varepsilon_n I)^{-1}\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}(\hat{\Sigma}_{XX} + \varepsilon_n I)^{-1}. \end{aligned} \tag{25}$$

Here, we have used ε_n to represent ε_X to emphasize the dependence on n (and we can do so because ε_Y is not involved in this version of f-GSIR). For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \prec b_n$ iff $a_n/b_n \to 0$, write $b_n \succ a_n$ iff $a_n \prec b_n$, and write $a_n \preceq b_n$ iff a_n/b_n either tends to 0 or is a bounded sequence.

THEOREM 5. *Suppose $E\kappa_X(X,X) < \infty$, $E\kappa_Y(Y,Y) < \infty$, and $\Sigma_{XY} = \Sigma_{XX}^{1+\beta}S_{XY}$ for some linear operator $S_{XY} : \mathfrak{M}_X \to \mathfrak{M}_Y$ and $0 < \beta \leq 1$. Suppose $n^{-1/2} \prec \varepsilon_n \prec 0$:*

1. *If S_{XY} is bounded, then $\|\hat{M} - M\|_{\mathrm{OP}} = O_P(\varepsilon_n^{\beta} + \varepsilon_n^{-1}n^{-1/2})$.*
2. *If S_{XY} is Hilbert–Schmidt, then $\|\hat{M} - M\|_{\mathrm{HS}} = O_P(\varepsilon_n^{\beta} + \varepsilon_n^{-1}n^{-1/2})$.*

Because Σ_{XX} is a Hilbert–Schmidt operator, its eigenvalues go to 0, which means $\Sigma_{XX}^{\dagger}$ is an unbounded operator. In order for $\Sigma_{XX}^{\dagger}\Sigma_{XY}$ to be bounded, we need at a minimum that $\Sigma_{XY} = \Sigma_{XX}B$ for a bounded linear operator B. However, for the consistency of f-GSIR it is not enough for $\Sigma_{XX}^{\dagger}\Sigma_{XY}$ to be bounded: we need $(\Sigma_{XX}^{\dagger})^{1+\beta}\Sigma_{XY}$ to be bounded for some $\beta > 0$. This is actually a smoothness condition, because it implies the subspaces corresponding to the small singular values of Σ_{XY} be sufficiently aligned with the eigenspaces of small eigenvalues of Σ_{XX}—in other words, the range space of Σ_{XY} be sufficiently focussed on the eigenspaces of the large eigenvalues of Σ_{XX}. Since large eigenvalues are usually associated with low-frequency components, the range of Σ_{XY} is essentially spanned by the low-frequency components of Σ_{XX}. The larger β is, the stronger this tendency. Thus, β characterizes the degree of "smoothness" in the relation between X and Y. Because $\beta \leq 1$, the rate $\varepsilon_n^{\beta} + \varepsilon_n^{-1}n^{-1/2}$ is the fastest when $\beta = 1$, in which case it is $\varepsilon_n + \varepsilon_n^{-1}n^{-1/2}$. Hence, when $\beta = 1$, both the optimal convergence rate and the optimal regularity constant are of the order $n^{-1/4}$. We should mention the rate $\varepsilon_n^{\beta} + \varepsilon_n^{-1}n^{-1/2}$ is an upper bound of the error, rather that the error itself. The reason that we restrict β to be no greater than 1 is that further increasing β would not lead to faster rate than $\varepsilon_n^{\beta} + \varepsilon_n^{-1}n^{-1/2}$.

By a well-known result in perturbation theory [see Koltchinskii and Giné (2000), Lemma 5.2 and Zwald and Blanchard (2006), Theorem 2] the eigenspaces of $\hat{M}$ converge to those of M at the same rate.

COROLLARY 3. *Suppose that the assumptions in Theorem 5 hold, and the nonzero eigenvalues of M are distinct. Let $\hat{P}_k$ and P_k be the projection operators on to the subspaces spanned by the kth eigenfunctions of $\hat{M}$ and M, respectively, where $k = 1, \dots, d$. Then*

$$\|\hat{P}_k - P_k\| = O_P(\varepsilon_n^{\beta} + \varepsilon_n^{-1} n^{-1/2}),$$

where the norm is the operator norm if S_{XY} is bounded, and Hilbert–Schmidt norm if S_{XY} is Hilbert–Schmidt.

7.2. *Convergence rate for partially observed functional data.* The consistency and convergence rate in Section 7.1 are developed under the assumption that X_i and Y_i are observed in their entirety. In reality, they are observed on a finite set of time points called *measurement schedule* in functional data analysis [Wang, Chiou and Muller (2015)]. Following this convention, we refer the measurement schedules that are sufficiently frequent so that covariance operators Σ_{XX}, Σ_{XY}, Σ_{XY} and Σ_{YY} can be estimated at the $n^{-1/2}$ rate as *dense schedules*. Applications involving automated measurements by instruments, such as fMRI, EEG and smart wearable records, may be regarded as belonging to this category. Since Theorem 5 and Corollary 3 depend only on the $n^{-1/2}$-convergence of the estimators of covariance operators, they apply to dense schedules without change. At the other extreme, the measurement schedules where the number of time points does not go to infinity with n are referred to as *sparse schedules*, which suitably describe a typical longitudinal study. For sparse schedules consistency can be achieved by pooling time points from different subjects, provided that these time points are sufficiently varied from subject to subject to fill up the whole interval as $n \to \infty$. However, the convergence rates for sparse schedules are slower than $n^{-1/2}$, and depend on the type of time smoothers employed. There are also, of course, measurement schedules in between these two extreme cases, resulting in convergence rates between them. We refer to these schedules as *nondense schedules*.

In order to suit nondense measurement schedules, we extend Theorem 5 and Corollary 3 to the case where the covariance operators are estimated at an arbitrary rate $n^{-1/2} \preceq \delta_n \prec 1$. Due to the limited scope of this paper, we do not delve into specific convergence rates by various time smoothers under different measurement schedules, but instead make the convergence statements sufficiently flexible to accommodate any convergence rate that might be provided by future investigations of specific cases. The proof of the next theorem follows those of Theorem 5 and Corollary 3, and is omitted.

THEOREM 6. *Suppose the covariance operators* Σ_{XX}, Σ_{XY}, Σ_{YX} *and* Σ_{YY} *can be estimated by* $\hat{\Sigma}_{XX}$, $\hat{\Sigma}_{XY}$, $\hat{\Sigma}_{YX}$, *and* $\hat{\Sigma}_{YY}$ *at a rate* δ_n *in operator norm, where* $n^{-1/2} \preceq \delta_n \prec 0$; *that is,*

$$\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{\mathrm{OP}} = O_P(\delta_n), \qquad \|\hat{\Sigma}_{YY} - \Sigma_{YY}\|_{\mathrm{OP}} = O_P(\delta_n),$$

$$\|\hat{\Sigma}_{YX} - \Sigma_{YX}\|_{\mathrm{OP}} = O_P(\delta_n).$$

Furthermore, suppose the Tychonoff regularization sequence $\{\varepsilon_n\}$ *in* $\hat{M}$ *satisfies* $\delta_n \prec \varepsilon_n \prec 0$, *and the covariance operator* Σ_{XY} *satisfies* $\Sigma_{XY} = \Sigma_{XX}^{1+\beta} S_{XY}$ *for some bounded linear operator* $S_{XY} : \mathfrak{M}_X \to \mathfrak{M}_Y$ *and* $0 < \beta \le 1$. *Then*:

1. $\|\hat{M} - M\|_{\mathrm{OP}} = O_P(\varepsilon_n^{\beta} + \delta_n/\varepsilon_n)$,
2. $\|\hat{P}_k - P_k\|_{\mathrm{OP}} = O_P(\varepsilon_n^{\beta} + \delta_n/\varepsilon_n)$, $k = 1, \dots, d$.

Moreover, the above statements hold for the Hilbert–Schmidt norm if S_{XY} *is a Hilbert–Schmidt operator.*

From this theorem, it is easy to derive the optimal ε_n for a given β, δ_n. The proof of the next corollary is straightforward, and is omitted.

COROLLARY 4. *Under the assumptions of Theorem 6, the optimal rate of convergence and the regularization sequence that achieves the optimal convergence rate are, respectively,*

$$\rho_n(\delta_n, \beta) = \delta_n^{\beta/(1+\beta)}, \qquad \varepsilon_n(\delta_n, \beta) = \delta_n^{1/(1+\beta)}.$$

To provide more intuition about these optimal rates, suppose $\delta_n = n^{-\alpha}$ for some $0 < \alpha \le 1/2$. Then α reflects the frequency of the measurement schedule: the more frequent a measurement schedule is, the close α is to $1/2$, with $\alpha = 1/2$ representing dense schedules. Meanwhile, recall that β measures a degree of smoothness of the relation between X and Y. In terms of α and β, the above rates are

$$\rho_n(\alpha, \beta) = n^{-\alpha\beta/(1+\beta)}, \qquad \varepsilon_n(\alpha, \beta) = n^{-\alpha/(1+\beta)}.$$

These relations sum up nicely the how the optimal penalty and optimal convergence rate depend on the smoothness and the measurement frequency: the smoother the relation between X and Y, the faster the optimal convergence rate and the stronger the optimal penalty; the more frequently the functional data are measured, the faster the optimal convergence rate and the weaker the optimal penalty.

7.3. *Consistency of the dimension determination criterion.* We now state the consistency of the dimension determination criterion proposed in Section 6.7, which is similar to the proof of Theorem 3 in Li, Li and Zhu (2010).

THEOREM 7. *Suppose the conditions in Theorem 6 are satisfied, and the optimal regularization sequence* $\varepsilon_n = n^{-\alpha/(1+\beta)}$ *is used. Suppose all the nonzero eigenvalues of* M *in* (25) *are distinct*; *that is,* $\lambda_1 > \cdots > \lambda_d > \lambda_{d+1} = \cdots = 0$. *If* $\hat{d}$ *is the minimizer of* (23) *for any constant* a, *then* $P(\hat{d} = d) \to 1$.

8. Simulation studies.

In this section, we investigate the performances of f-GSIR and f-GSAVE by simulation. We consider three different scenarios of response versus predictor: random variable versus random function (scenario I), random function versus random vector (scenario II) and random function versus random function (scenario III). We take $T = [0, 1]$, and include both balanced cases (where $\{t_{i1}, \dots, t_{im_i}\}$ are the same for different i) and unbalanced cases (where they are not). For vector-valued predictor or response, we use the Euclidean space as $\mathcal{H}_X$ or $\mathcal{H}_Y$. We first describe how the random functions are generated.

8.1. *Simulation of random functions.* We choose the Gaussian radial basis function (GRB) or Brownian motion covariance (BMC) as the kernel for $\mathcal{H}_X$ and $\mathcal{H}_Y$. When the GRB kernel is used, we generate X by $\sum_{k=1}^{m} a_k \kappa_T(\cdot, t_k)$ where $a_1, \dots, a_m$ are independently sampled from $N(0, 1)$, $t_1, \dots, t_m$ are independently sampled from $U[0, 1]$, $m = 5$, and $\gamma_T = 7$. When the BMC kernel is used, we generate X by

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

with $a_1, \dots, a_{100}$ independently sampled from $N(0, 1)$. As we noted in Example 3, the functions in the summand are the eigenfunctions of the BMC kernel. The random function Y is generated in the same way. The two kernels are also used for estimation: in Section 8.2, we compare the results from the four combinations of two kernels and two estimators.

For the observed time points $\{t_{i1}, \dots, t_{im_i} : i = 1, \dots, n\}$, we simulated both the balanced and the unbalanced cases. For the balanced case, we chose equally-spaced 10 time points. For the unbalanced case, we fixed 100 equally-spaced time points in [0, 1], then randomly selected 10 points from them for each subject. To provide intuition, in Figure 2 we show 50 sample paths of the random function generated by each kernel.

8.2. *Scenario* I. We first consider the case where Y is a random variable and X is a random function:

$$\begin{aligned} &\text{Model I-1:} \quad Y = \langle b_1, X\rangle + \langle b_2, X\rangle + \varepsilon, \\ &\text{Model I-2:} \quad Y = \frac{\langle b_1, X\rangle}{1 + e^{\langle b_2, X\rangle}} + 0.2\langle X, X\rangle + \varepsilon, \qquad (26) \\ &\text{Model I-3:} \quad Y = \big(\langle b_1, X\rangle + \langle b_2, X\rangle\big)\varepsilon, \end{aligned}$$

where $\varepsilon \sim N(0, 0.1^2)$ and b_1, b_2, b_3 are nonrandom elements of $\mathcal{H}_X$. When the GRB kernel is used, we take $b_1(t) = \kappa(t - 0.1)$, $b_2(t) = \kappa(t - 0.5)$, $b_3(t) = \kappa(t - 0.9)$; when the BMC kernel is used, we take $b_j(t) = v_j(t)$, $j = 1, 2, 3$, as

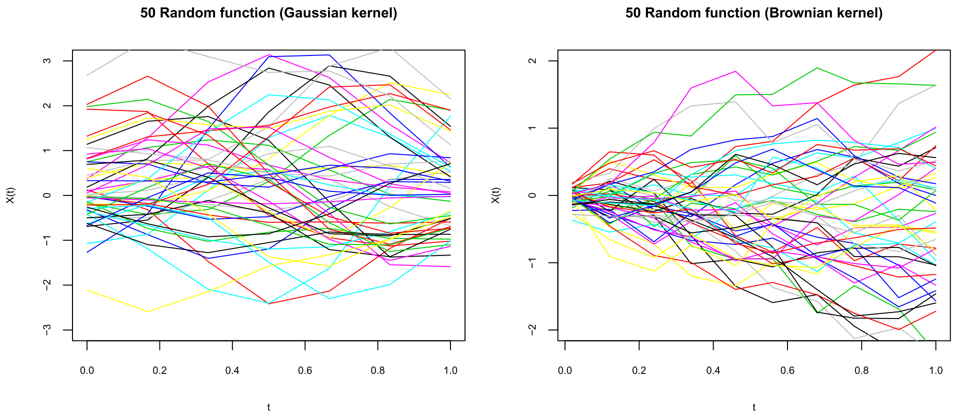


FIG. 2. *Sample paths generated by the GRB kernel* (*left panel*) *and BMC kernel* (*right panel*).

defined in (5). The sample size is $n = 100$. Note that Models I-1 and I-2 have complete central class, for which f-GSIR is exhaustive, but Model I-3 does not have a complete central class.

To compare the estimated and true predictors, which may be vectors of different dimensions, we propose a multivariate version of Spearman's correlation, which we call Multiple Correlation of Multivariate Rank (MCMR), as follows. Let $U_1, \dots, U_n \in \mathbb{R}^r$ and $V_1, \dots, V_n \in \mathbb{R}^s$ be two samples of random vectors representing the estimated and true predictors. Let $\tilde{U}_i$ and $\tilde{V}_i$ be their multivariate ranks

$$\tilde{U}_i = n^{-1} \sum_{\ell=1}^{n} (U_\ell - U_i)/\|U_\ell - U_i\|, \qquad \tilde{V}_i = n^{-1} \sum_{\ell=1}^{n} (V_\ell - V_i)/\|V_\ell - V_i\|.$$

See, for example, Oja (2010). We define MCMR between $\{U_1, \dots, U_n\}$ and $\{V_1, \dots, V_n\}$ to be the multiple correlation between the multivariate ranks of the two samples:

$$\begin{aligned} \text{mcmr}_n(U, V) = \big(\text{tr}\big\{\big[\text{var}_n(\tilde{V})\big]^{-1/2} \text{cov}_n(\tilde{V}, \tilde{U}) \big[\text{var}_n(\tilde{U})\big]^{-1} \\ \times \text{cov}_n(\tilde{U}, \tilde{V}) \big[\text{var}_n(\tilde{V})\big]^{-1/2}\big\}\big)^{1/2}. \end{aligned}$$

Lee, Li and Chiaromonte (2013) used Spearman's correlation to measure the (possibly nonlinear) dependence between the estimated and the true predictors that are both scalars. The Spearman's correlation is an ideal measurement of such dependence because it is invariant under monotone transformations the two random variables involved. Our use of MCMR is similarly motivated.

Since this criterion is always evaluated at the test set, there is no over fitting. However, when the dimension of $\hat{U}$ or $\hat{V}$ is high relative to the sample size, there can be spurious correlation (i.e., two samples of high-dimensional vectors tend to be correlated even when they are uncorrelated at the population level). Hence, to

TABLE 1
Performances of f-GSIR and f-GSAVE under Models I-1 *through* I-3 *and the two kernels in the balanced case*

Models		Methods			
		BMC		GRB	
X	*Y*|*X*	f-GSIR	f-GSAVE	f-GSIR	f-GSAVE
BMC	I-1	0.933 (0.023)	0.814 (0.031)	0.916 (0.130)	0.849 (0.039)
	I-2	0.401 (0.360)	0.936 (0.021)	0.325 (0.321)	0.931 (0.023)
	I-3	0.214 (0.252)	0.844 (0.030)	0.226 (0.242)	0.813 (0.040)
GRB	I-1	0.973 (0.026)	0.845 (0.038)	0.980 (0.015)	0.928 (0.025)
	I-2	0.783 (0.253)	0.530 (0.094)	0.660 (0.339)	0.699 (0.083)
	I-3	0.445 (0.259)	0.864 (0.027)	0.193 (0.221)	0.867 (0.029)

objectively reflect the performance of the estimators, we only calculate MCMR for the first 10 estimated predictors when the dimension d is estimated to be greater than 10. This happened rather rarely: in only 3 out of the 42 scenarios considered $\hat{d}$ is larger than 10. These are Model II-1, 2, 3 and Model III-1 for f-GSAVE, where d is estimated to be 69, 67, 40, 50, respectively. In most cases, the estimate of d is no greater than 3.

We generate $2n = 200$ independent observations on (X, Y), of which 100 are used as the training set and the rest as the test set. We apply our methods to the training set to obtain the sufficient predictors, and then evaluate these predictors at X in the test set. We then evaluate MCMR between the estimated and true predictors from the test set. We repeat this process 100 times and report the averages and standard errors (in parentheses) of the MCMR in Table 1. The tuning parameters are determined by the methods in Sections 6.5 through 6.7. To save computing time, we estimate the tuning parameters based on 10 separately generated pilot samples, and then use their average (except $\hat{d}$) for estimation in the 100 training sample. For the dimension $\hat{d}$, we use the mode of the histogram of the 10 estimates instead of the average. Tuning parameters for the other two scenarios were determined the same way.

Table 1 shows the results for the balanced case from the three models in (26) and the two estimation methods (f-GSIR and f-GSAVE), as they are coupled with two kernels, resulting a combination of six models and four methods. We note that the performances of f-GSIR and f-GSAVE are comparable for Models I-1 and I-2, where the central class is complete; whereas f-GSAVE performs better than f-GSIR for Model I-3, where the central class is not complete. Overall, at sample size 100, with relatively sparsely positioned observation times and rather complicated nonlinear relations in (26), the two methods capture the true predictor quite well, with MCMR mostly ranging from $0.6 \sim 0.9$ with relatively low standard errors. The few cases with low MCMR, such as the 0.530 for the (GRB, I-2, BMC,

TABLE 2
Performances of f-GSIR and f-GSAVE under Models I-1–I-3 *in the unbalanced case*

Models	**f-GSIR**	**f-GSAVE**
I-1	0.830 (0.024)	0.871 (0.031)
I-2	0.388 (0.306)	0.920 (0.024)
I-3	0.237 (0.258)	0.866 (0.029)

f-GSAVE) combination, are due to underestimate of dimension d by the pilot sample, which is 1. The MCMR increases sharply to 0.622 when the second and third sufficient predictor is included.

We carry out rest of the simulations using the BMC kernel only. Table 2 shows the results for Models I-1, I-2, I-3 in the unbalanced case. Overall, the performances in the unbalanced case are slightly worse than in the balanced case.

8.3. *Scenarios* II *and* III. For these scenarios we only present the results for the balanced case, where the random functions are observed in equally spaced 10 time points in [0, 1]. As in scenario I, the training sample size, test sample size and simulation sample size are each taken to be 100. Let v_j be the eigenfunctions in (5). We consider following models for scenarios II:

$$\text{Model II-1:}\quad Y(t) = v_0(t) + (X_1 + X_2)\sum_{i=1}^{5} v_i(t) + \sigma\varepsilon(t),$$

$$\text{Model II-2:}\quad Y(t) = v_0(t) + \bigl(X_1/\bigl(1 + e^{X_2}\bigr) + X_3\bigr) + \sigma\varepsilon(t),$$

$$\text{Model II-3:}\quad Y(t) = X_1\|X\|^2\varepsilon(t),$$

where $X \sim N(0, I_{10})$ and $\varepsilon(t)$ is generated from the standard Brownian motion. The results are shown in the first 3 rows of Table 3.

TABLE 3
Performances of f-GSIR and f-GSAVE under Models II-1–II-3 *and* III-1–III-3

Models	**f-GSIR**	**f-GSAVE**
II-1	0.981 (0.005)	0.953 (0.013)
II-2	0.971 (0.007)	0.931 (0.013)
II-3	0.145 (0.151)	0.882 (0.019)
III-1	0.881 (0.023)	0.990 (0.003)
III-2	0.864 (0.085)	0.849 (0.036)
III-3	0.791 (0.173)	0.853 (0.028)

For scenario III, we consider the following models:

$$\text{Model III-1:}\quad Y(t) = \big(\langle X, b_1\rangle + \langle X, b_2\rangle\big)\rho(t) + \sigma\varepsilon(t),$$

$$\text{Model III-2:}\quad Y(t) = \left(\frac{\langle X, b_1\rangle}{1 + e^{\langle X, b_2\rangle}} + \langle X, b_3\rangle\right)\rho(t) + \sigma\varepsilon(t),$$

$$\text{Model III-3:}\quad Y(t) = \cos\big(\langle b_1, X\rangle\big) + \sin\big(\langle b_2, X\rangle\big)\rho(t) + \sigma\varepsilon(t),$$

where X is a random function on $T = [0, 1]$, $b_j(t)$ are taken to the eigenfunctions $v_j(t)$ defined in (5), $\rho(t) = \sum_{j=1}^{5} v_j(t)$, $\sigma = 0.5$, and $\varepsilon(t)$ is generated from the standard Brownian motion. The results are shown in the last 3 rows of Table 3.

From Table 3, we see that f-GSIR and f-GSAVE perform similarly in the complete cases, but f-GSAVE performs significantly better than f-GSIR in the incomplete case.

9. Applications.

9.1. *Speech recognition data.* We applied our functional nonlinear SDR methods to analyzed the speech recognition dataset from the TIMIT, available at http://statweb.stanford.edu/~tibs/ElemStatLearn/, which was also used in Hastie, Tibshirani and Friedman (2009), Rossi and Villa (2006) and Epifanio (2008). It consists of five phonemes transcribed as follows: “sh” as in “she,” “dcl” as in “dark,” “iy” as the vowel in “she,” “aa” as the vowel in “dark” and “ao” as the first vowel in “water.” A total of 4509 speech frames of 32 millisecond duration each are recorded. Figure 3 shows the phoneme curves for the first 10 speech frames corresponding to each of the five phonemes, computed by a log-periodogram with length 256. As in Epifanio (2008), we only used the first 150 frequencies.

For each phoneme type from the database, we randomly selected with replacement 50 samples of curves each containing 200 curves. We then randomly divide each sample of 200 curves into training and test dataset, each containing 100 curves. At the end of the process we obtained 50 samples, each sample containing a training and test dataset, each dataset containing 500 curves evenly distributed among 5 phoneme types.

We applied f-GSIR and f-GSAVE to each training set to derive the sufficient predictors. Figure 4 shows the scatter plots of first two predictors derived from f-GSIR and f-GSAVE based on one sample, where the left panels show the training set, and right panels show the test set. These plots show reasonably good separation of the phoneme types, but a much clear separation can be seen in a 3-d spin plot—which cannot be shown here—in which all except the blue and light blue groups are well separated. The blue and light blue groups represent the first vowel in “water” and the vowel in “dark,” respectively, which are quite close. In fact, as noted by Epifanio (2008) and Rossi and Villa (2006), these two sounds are the most difficult to classify in the database.

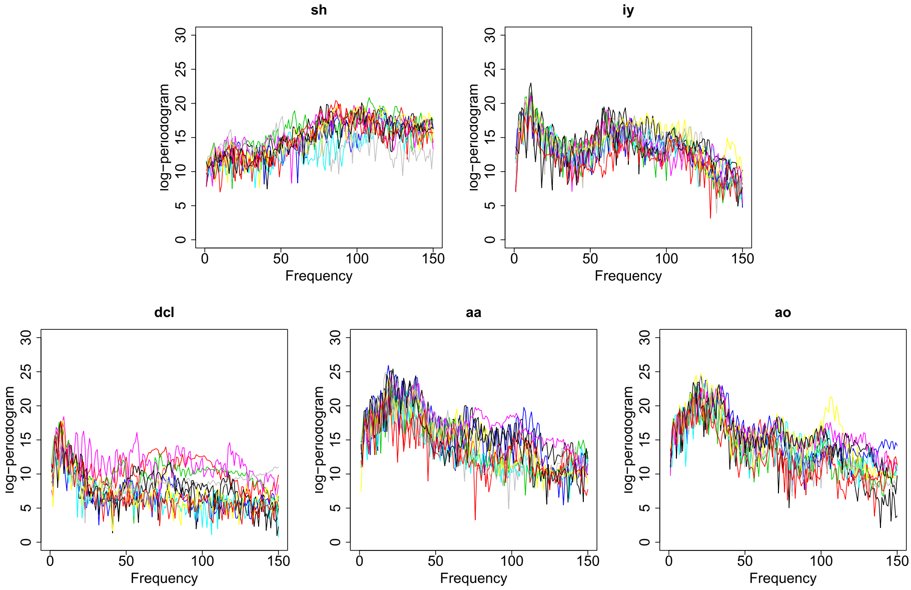


FIG. 3. *First* 10 *phoneme curves of each phoneme.*

We then applied three commonly used classifiers, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Support Vector Machine (SVM), to the first $\hat{d}$ sufficient predictors by the f-GSIR, f-GSAVE and f-SIR, and recorded the misclassification rate. As a baseline for comparison, we also computed the misclassification rate of the functional SIR (f-SIR) of Ferré and Villa (2006). We repeat the process for all 50 samples and report in Table 4 the mean and standard error of the misclassification rates from the combinations in {f-GSIR, f-GSAVE, f-SIR} × {LDA, QDA, SVM} × {training set, test set}.

The tuning parameters were chosen by the methods described in Sections 6.5–6.7. In particular, then dimension d is taken to be $5 - 1 = 4$ for both f-GSIR and f-GSAVE as proposed in Section 6.7. We used GBR for both the kernels of the first-level and second-level functions. For f-SIR, we used the regularization parameter to be 10 and dimension to be 4, which are taken from Ferré and Villa (2006), Table 2. The table shows that f-GSIR and f-GSAVE consistently perform better than f-SIR in the training and test sets according to all three classifiers; f-GSIR performs better than f-GSAVE in most cases, but worse than f-GSAVE for the test data set with QDA as classifier.

9.2. *Handwriting symbol association.* Our second application is the handwritten symbol association problem mentioned in the Introduction, where our goal is to train the computer to learn to associate two sets of handwritten symbols: a, b, c, . . . , and 1, 2, 3, ... Available handwritten data are usually

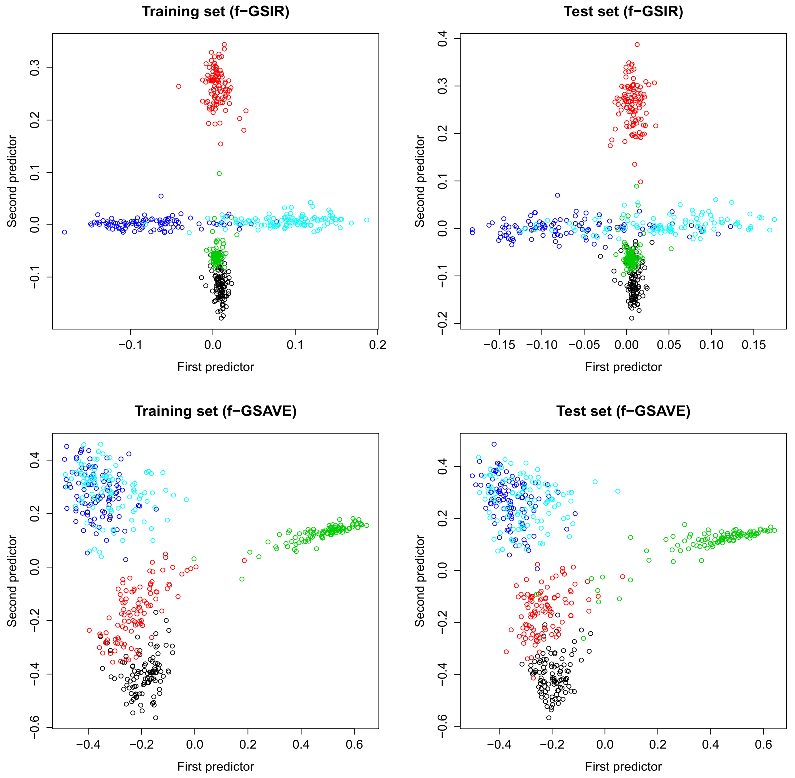


FIG. 4. *Scatter plots of first two f-GSIR predictors of the phoneme data* (*black* = "*sh*," *red* = "*iy*," *green* = "*dcl*," *dark blue* = "*aa*" *and light blue* = "*ao*"). *Left panels show the predictors evaluated at the training set*; *the right panels the test set*.

TABLE 4
Comparison of means and standard errors of misclassification rates based on 50 *samples*

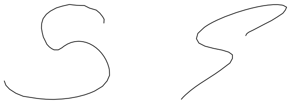
	Dimension reduction methods					
	Training set			**Test set**		
Classifier	**f-GSIR**	**f-GSAVE**	**f-SIR**	**f-GSIR**	**f-GSAVE**	**f-SIR**
LDA	0.0296	0.1091	0.1075	0.0883	0.1090	0.1084
	(0.0067)	(0.0137)	(0.0122)	(0.0128)	(0.0104)	(0.0119)
QDA	0.0315	0.0942	0.0981	0.0948	0.1037	0.1083
	(0.0066)	(0.0122)	(0.0103)	(0.0125)	(0.0112)	(0.0124)
SVM	0.0290	0.0947	0.0959	0.0878	0.1040	0.1126
	(0.0062)	(0.0116)	(0.0112)	(0.0120)	(0.0112)	(0.0123)

of two types: the on-line type, which is a parameterized curve in $\mathbb{R}^2$, and the off-line type, which is simply a 2-d image. The on-line type is a natural form of functional data, and we use it for our analysis. The data set can be found in the UJIpenchars2 database: http://archive.ics.uci.edu/ml/machine-learning-databases/uji-penchars/version2/. See also [Llorens et al. (2008)].

The database consists of 11,640 handwritten symbols from 60 writers × 97 characters × 2 repetitions. We used the first repetition as the training set, and the second as the test set. Mathematically, each predictor or response is a 2-dimensional random function. For example, $(X_1(t), Y_1(t))$ is subject 1's handwriting of "2" and "a," where $X_1(t) = (X_1^1(t), X_1^2(t))$ is the parameterized curve for "2" and $Y_1(t) = (Y_1^1(t), Y_1^2(t))$ is that for "a." Obviously, there is no reason to believe that the association between these symbols follows a linear index model; nevertheless, a relatively strong relation must exist because humans can easily associate them.

We normalize the data by rescaling the images to fit within a [0, 100] × [0, 180] rectangle and rescaling the sets of t to fit within the interval $[0, 0.1]$. After the rescaling, the data consist of $\mathbb{R}^2$-valued functions of t observed at different sets of times. We first used "2," "3" and "6" as the predictors, and "a," "b" and "c" as the responses. These symbols were selected for simplicity: it is slightly more complicated to parameterize symbols with loops, such as 4 and 8 (though a more careful treatment can solve this problem). Figure 5 shows the first two sufficient predictors from f-GSIR (upper panels) and f-GSAVE (lower panels). The left panels use the handwritten alphabets as the plotting symbols; the right panels use handwritten digits as the plotting symbols. We then repeated the analysis on with 2, 3, 6 replaced by 5, 7, 9, and obtained the similar degree of matching, as shown in Figure 6. The sufficient predictors and tuning parameters are based on the training set, and Figures 5 and 6 show the results for the test sets.

As we see from Figures 5 and 6, f-GSAVE worked better for the first set of symbols, and f-GSIR worked better for the second. In each case, they gave nearly perfect separation. Note that this is not a classification problem: the responses (handwritten letters) are simply continuous curves whose meanings are not understood by a computer as symbols. Thus, it is more difficult than classification. It is also interesting to note that the cases that were not well separated were often due to poor handwriting. For example, the symbols



appeared in the middle of the three clusters in the upper left panel of Figure 6 (f-GSIR for the second set of symbols): the first symbol is 5 and the second is 9. However, due to the ambiguous handwriting the first symbol was placed near the 9 group, and the second symbol was placed near the 5 group.

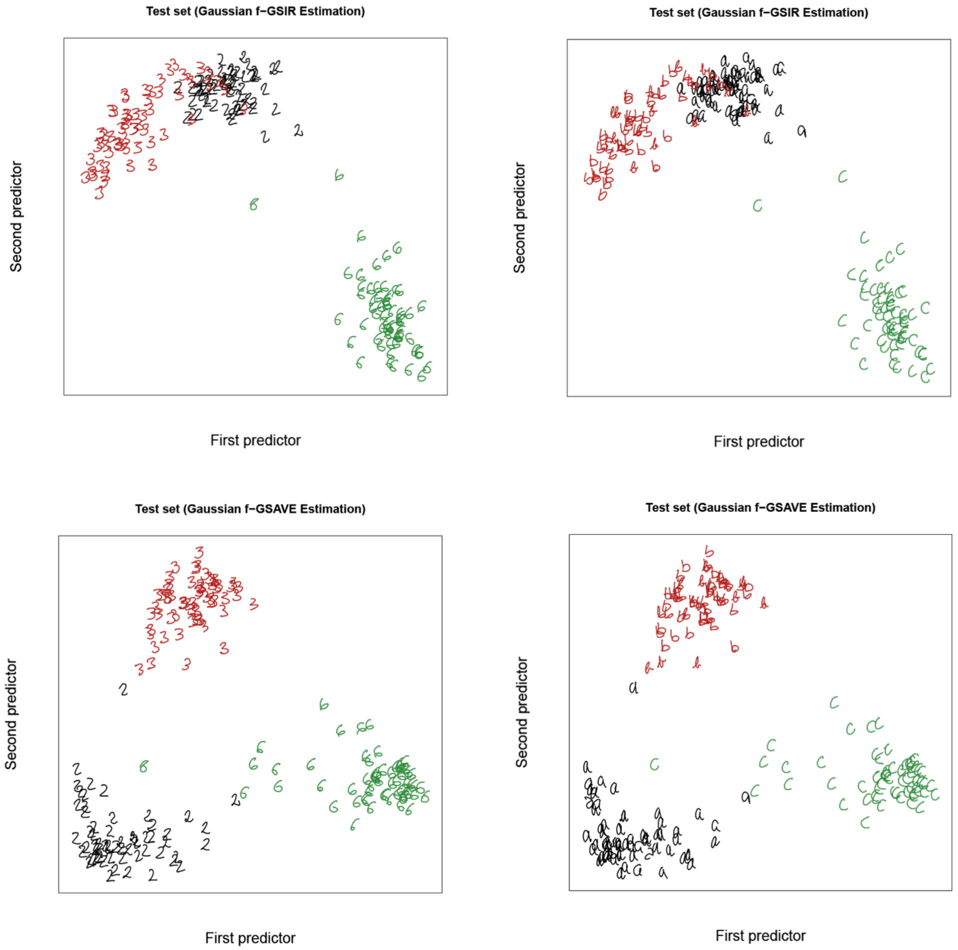


FIG. 5. *Scatter plots of first two predictors from f-GSIR* (*upper panels*) *and f-GSAVE* (*lower panels*). *The left panels use handwritten letters as the plotting symbol*; *the right panels use handwritten numerals as the plotting symbol.*

10. Conclusions.

We have extended the theory of nonlinear sufficient dimension reduction to situations where both the predictor and response can be functions. We developed two estimators, the f-GSIR and f-GSAVE, to estimate the central class for functional nonlinear SDR, along with procedures for choosing the tuning parameters and for determining the dimension of the central class.

While functional nonlinear SDR inevitably shares some common properties with nonlinear SDR for multivariate data in Lee, Li and Chiaromonte (2013), the development here goes far beyond that paper in several aspects. First, to account for the functional and nonlinear nature of this problem, we proposed the construction of two nested Hilbert spaces, where the inner product of the first determines

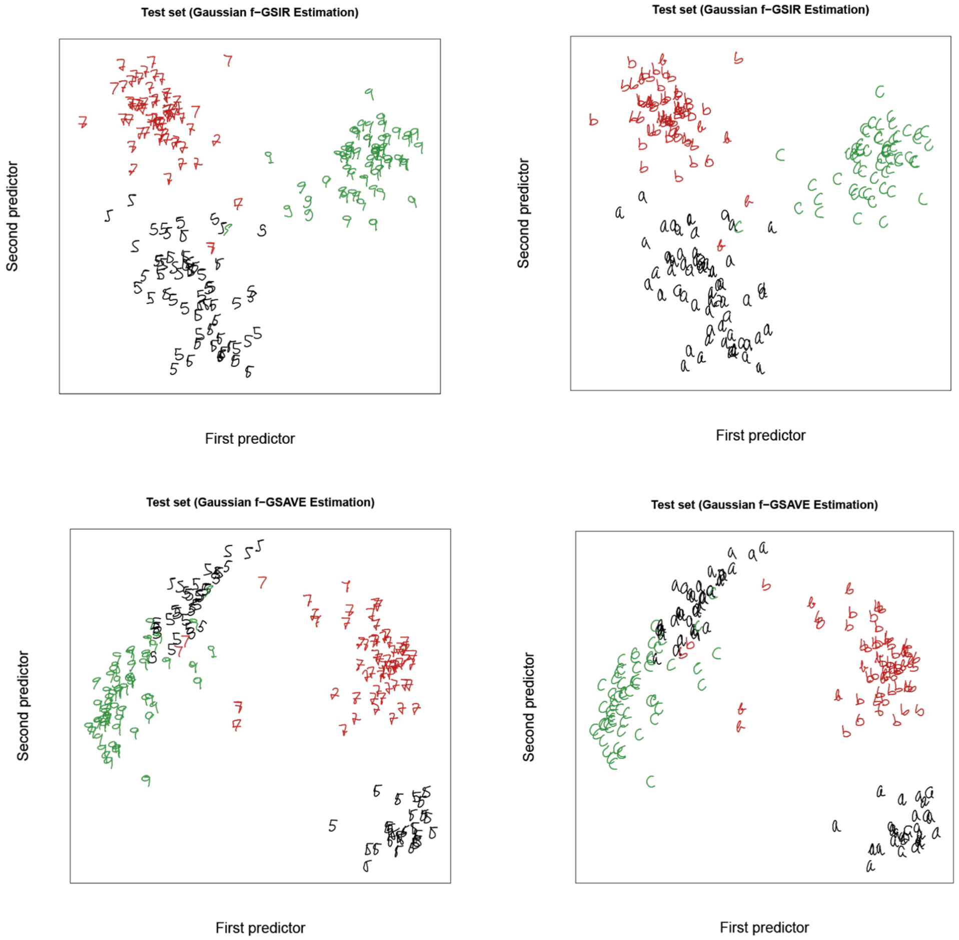


FIG. 6. *Scatter plots of first two predictors from f-GSIR (upper panels) f-GSAVE (lower panels) for associating handwritten* 5, 7, 9 *and handwritten* a, b, c.

the reproducing kernel of the second. Second, we derived the convergence rate of f-GSIR as a function of a smoothness index and an observation frequency index, which reflect the special features of functional data that set it apart from multivariate data. Third, we developed tuning and dimension estimation procedures specific to functional data, which cannot be deduced from Lee, Li and Chiaromonte (2013). Finally, the vast amount of functional data made available by recent techniques such as smart wearable devices [see, e.g., Bai et al. (2012)] create significant new demands for dimension reduction, which justifies fully developed theory and methodologies, in spite of their precursors in multivariate nonlinear SDR.

The theoretical framework laid out in this paper has many ramifications that cannot all be fully developed here due to limited space. For example, other dimension reduction methods in the classical setting, such as Directional Regression [Li and Wang (2007)], can be generalized in to functional nonlinear SDR in a similar manner. Other asymptotic properties such as asymptotic normality need also be established. We leave these to future research.

Acknowledgments. We thank two referees and an Associate Editor for their many thoughtful and constructive comments and suggestions, which helped us greatly in revising this paper.

SUPPLEMENTARY MATERIAL

External Appendix to "Nonlinear sufficient dimension reduction for functional data" (DOI: 10.1214/16-AOS1475SUPP; .pdf). The supplementary file provides the proofs of Lemmas 1, 2 and 3, Theorems 1, 4, 5 and 7, and Proposition 1.

REFERENCES

AMATO, U., ANTONIADIS, A. and DE FEIS, I. (2006). Dimension reduction in functional regression with applications. *Comput. Statist. Data Anal.* **50** 2422–2446. MR2225577

AMINI, A. A. and WAINWRIGHT, M. J. (2012). Sampled forms of functional PCA in reproducing kernel Hilbert spaces. *Ann. Statist.* **40** 2483–2510. MR3097610

BACH, F. R. and JORDAN, M. I. (2003). Kernel independent component analysis. *J. Mach. Learn. Res.* **3** 1–48. MR1966051

BAI, J., GOLDSMITH, J., CAFFO, B., GLASS, T. A. and CRAINICEANU, C. M. (2012). Movelets: A dictionary of movement. *Electron. J. Stat.* **6** 559–578. MR2988420

BAKER, C. R. (1973). Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.* **186** 273–289. MR0336795

BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, MA. MR2239907

COOK, R. D. (1998). *Regression Graphics*. Wiley, New York. MR1645673

COOK, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22** 1–26. MR2408655

COOK, R. D. and LI, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30** 455–474. MR1902895

COOK, R. D. and SETODJI, C. M. (2003). A model-free test for reduced rank in multivariate regression. *J. Amer. Statist. Assoc.* **98** 340–351. MR1995710

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

DAUXOIS, J., FERRÉ, L. and YAO, A.-F. (2001). Un modèle semi-paramétrique pour variables aléatoires hilbertiennes. *C. R. Acad. Sci. Paris Sér. I Math.* **333** 947–952. MR1873814

EPIFANIO, I. (2008). Shape descriptors for classification of functional data. *Technometrics* **50** 284–294. MR2528652

FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis*: *Theory and Practice*. Springer, New York. MR2229687

FERRÉ, L. and VILLA, N. (2006). Multilayer perceptron with functional inputs: An inverse regression approach. *Scand. J. Stat.* **33** 807–823. MR2300917

FERRÉ, L. and YAO, A. F. (2003). Functional sliced inverse regression analysis. *Statistics* **37** 475–488. MR2022235

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

FUKUMIZU, K., BACH, F. R. and GRETTON, A. (2007). Statistical consistency of kernel canonical correlation analysis. *J*. *Mach*. *Learn*. *Res*. **8** 361–383. MR2320675

FUKUMIZU, K., BACH, F. R. and JORDAN, M. I. (2009). Kernel dimension reduction in regression. *Ann*. *Statist*. **37** 1871–1905. MR2533474

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

HORN, R. A. and JOHNSON, C. R. (1985). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR0832183

HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. Springer, New York. MR2920735

HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis*, *with an Introduction to Linear Operators*. Wiley, Chichester. MR3379106

HSING, T. and REN, H. (2009). An RKHS formulation of the inverse regression dimension–reduction problem. *Ann*. *Statist*. **37** 726–755. MR2502649

KOLTCHINSKII, V. and GINÉ, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* **6** 113–167. MR1781185

LEE, K.-Y., LI, B. and CHIAROMONTE, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Ann*. *Statist*. **41** 221–249. MR3059416

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

LI, B., ARTEMIOU, A. and LI, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Ann*. *Statist*. **39** 3182–3210. MR3012405

LI, B., CHUN, H. and ZHAO, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *J*. *Amer*. *Statist*. *Assoc*. **107** 152–167. MR2949348

LI, B., CHUN, H. and ZHAO, H. (2014). On an additive semigraphoid model for statistical networks with application to pathway analysis. *J*. *Amer*. *Statist*. *Assoc*. **109** 1188–1204. MR3265690

LI, L., LI, B. and ZHU, L.-X. (2010). Groupwise dimension reduction. *J*. *Amer*. *Statist*. *Assoc*. **105** 1188–1201. MR2752614

LI, B. and SONG, J. (2016). Supplement to "Nonlinear sufficient dimension reduction for functional data." DOI:10.1214/16-AOS1475SUPP.

LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J*. *Amer*. *Statist*. *Assoc*. **102** 997–1008. MR2354409

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

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

LLORENS, D., PRAT, F., MARZAL, A., VILAR, J. M., CASTRO, M. J., AMENGUAL, J. C., BARRACHINA, S., CASTELLANOS, A., NA, S. E., GÓMEZ, J. A., GORBE, J., GORDO, A., PALAZÓN, V., PERIS, G., RAMOS-GARIJO, R. and ZAMORA, F. (2008). The ujipenchars database: A pen-based database of isolated handwritten characters. In *Proc*. 6*th Int*. *Conf*. *Language Resources Eval*, Marrakech, Morocco. 2647–2651.

MINH, H. Q. (2010). Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constr*. *Approx*. **32** 307–338. MR2677883

MÜLLER, H.-G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann*. *Statist*. **33** 774–805. MR2163159

OJA, H. (2010). *Multivariate Nonparametric Methods with R*: *An Approach Based on Spatial Signs and Ranks*. Springer, New York. MR2598854

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

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA. MR2514435

ROSSI, F. and VILLA, N. (2006). Support vector machine for functional data cliassification. *Neurocomputing* **79** 730–742.

WANG, J.-L., CHIOU, J.-M. and MULLER, H.-G. (2015). Review of functional data analysis. Preprint. Available at arXiv:1507.05135v1.

WANG, G., LIN, N. and ZHANG, B. (2013). Functional contour regression. *J. Multivariate Anal.* **116** 1–13. MR3049886

WANG, Q. and YIN, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Comput. Statist. Data Anal.* **52** 4512–4520. MR2432477

WU, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *J. Comput. Graph. Statist.* **17** 590–610. MR2528238

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106

YEH, Y.-R., HUANG, S.-Y. and LEE, Y.-J. (2009). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Trans. Knowl. Data Eng.* **11** 1590–1603.

YIN, X. and BURA, E. (2006). Moment-based dimension reduction for multivariate response regression. *J. Statist. Plann. Inference* **136** 3675–3688. MR2256281

ZHU, L., MIAO, B. and PENG, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Amer. Statist. Assoc.* **101** 630–643. MR2281245

ZWALD, L. and BLANCHARD, G. (2006). On the convergence of eigenspaces in kernel principal component analysis. In *Advances in Neural Information Processing Systems* **18**. MIT Press, Cambridge, MA.

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
326 THOMAS BUILDING
UNIVERSITY PARK, PENNSYLVANIA 16802
USA
E-MAIL: bing@stat.psu.edu