DIMENSION REDUCTION FOR FUNCTIONAL DATA 
BASED ON WEAK CONDITIONAL MOMENTS

BY BING LI* AND JUN SONG

The Pennsylvania State University and The University of North Carolina 
at Charlotte

We develop a general theory and estimation methods for functional linear sufficient dimension reduction, where both the predictor and the response can be random functions, or even vectors of functions. Unlike the existing dimension reduction methods, our approach does not rely on the estimation of conditional mean and conditional variance. Instead, it is based on a new statistical construction — the weak conditional expectation, which is based on Carleman operators and their inducing functions. Weak conditional expectation is a generalization of conditional expectation. Its key advantage is to replace the projection on to an $L^2$-space — which defines conditional expectation — by projection on to an arbitrary Hilbert space, while still maintaining the unbiasedness of the related dimension reduction methods. This flexibility is particularly important for functional data, because attempting to estimate a full-fledged conditional mean or conditional variance by slicing or smoothing over the space of vector-valued functions may be inefficient due to the curse of dimensionality. We evaluated the performances of the our new methods by simulation and in several applied settings.

1. Introduction. We introduce a general theory and estimation procedures for sufficient dimension reduction (SDR) for functional data, where both the predictor and the response can be functions or even vector-valued functions, and the dimension reduction is linear in the sense that the predictors are compressed by linear functions. Both the formulation and the methods are quite different from the existing functional SDR methods. These new features — especially the weak conditional moments — are designed specifically to adapt to the special characteristics of the functional data.

In its original form, SDR is concerned with reducing the dimension of a random vector $X$ without losing its predictive power on another random

*Supported in part by the NSF Grant DMS-1713078.

MSC 2010 subject classifications: Primary 62G08; secondary 62H99, 62B05

Keywords and phrases: Carleman operator, Weak conditional moment, Weak Inverse Regression Estimate, Weak Average Variance Estimate, Weak Directional Regression, Function-on-function regression.
variable $Y$. The theory is based on the assumption
\begin{equation}
Y \perp \! \! \! \perp X|\beta^T X,
\end{equation}
where $\beta$ is a $p \times d$ matrix with $d \ll p$, and $A \perp \! \! \! \perp B|C$ stands for “$A$ is independent of $B$ given $C$”. The above relation depends on $\beta$ only through the space spanned by its columns, denoted by span$(\beta)$. The goal of SDR is to estimate the smallest subspace span$(\beta)$ that satisfies (1.1) without making strong assumptions about the relation between $Y$ and $X$. This smallest subspace is called the central subspace, and is denoted by $\mathcal{S}_Y|X$. The low-dimension predictor obtained by projecting $X$ on to the estimated central subspace can then be fed into any regression, classification, or machine learning procedures for further analysis. See Li [21], Cook and Weisberg [8], Cook [6], and Cook and Li [7].

This theory has been extended to function-valued $X$ — that is, $X$ is a random function defined on some interval $J$ — via two main routes. The first is proposed by Ferré and Yao [11, 12], which assumes that $X$ is a random element in a Hilbert space $\mathcal{H}$ such that
\begin{equation}
Y \perp \! \! \! \perp X|\langle f_1, X \rangle_{\mathcal{H}}, \ldots, \langle f_d, X \rangle_{\mathcal{H}}
\end{equation}
where $f_1, \ldots, f_d$ are fixed but unknown members of $\mathcal{H}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$. By making analogy with classical SDR, the authors extended the Sliced Inverse Regression (SIR) of Li [21] to functional Sliced Inverse Regression (FSIR). The advantage of this approach is that the theory has a parallel structure to the classical SDR, which provides us a path of generalization. Under the same framework, Wang, Lin and Zhang [34] extended the contour regression of Li, Zha and Chiaromonte [29] to functional contour regression (FCR), which overcomes some shortcomings of the FSIR. Lian and Li [30] extended Sliced Average Variance Estimator (SAVE, Cook and Weisberg [8]) to the functional setting (FSAVE) and investigated convergence rate of FSIR and FSAVE. Wang et al. [35] proposed a hybrid method combining the advantages of FSIR and FSAVE.

The second route, taken by Hsing and Ren [16], assumes there exist random variables $\xi_1, \ldots, \xi_d$ in the closed linear span of the set $\{X(t) : t \in J\}$ such that $Y \perp \! \! \! \perp X|\xi_1, \ldots, \xi_d$. In order to have a nonrandom target to estimate, the authors define a reproducing kernel Hilbert space (RKHS) using the covariance function $(s, t) \mapsto \text{cov}[X(s), X(t)]$ as its kernel. This space is isomorphic to the closed linear span of $\{X(t) : t \in J\}$, and contains the nonrandom mirror images of $\xi_1, \ldots, \xi_d$, which are taken to be the target of estimation. The advantage of this approach is that one need not assume $X$ to live in a Hilbert space, which can be too strong for some applications.
In this paper we further develop functional SDR using the first framework where \( X \) is assumed to be a random element in a Hilbert space. Unlike the previous extensions, we do not attempt to estimate the conditional mean \( E(X|Y) \) and conditional variance var(\( X|Y \)). Instead, we rely on a new statistical construct — the weak conditional expectation — to carry out the extension. The weak conditional expectation is defined as the inducing function of a Carleman operator (Weidmann [36]). The key advantage of this approach is that it replaces the \( L^2 \)-space for the projection that characterizes the conditional expectation by an arbitrary Hilbert space, while still maintaining the unbiasedness of the dimension reduction estimate. This flexibility is particularly important for functional data, because attempting to estimate the conditional mean function \( E(X|Y) \) and the conditional variance operator var(\( X|Y \)) when \( Y \) is a multivariate random function by taking slices or smoothing over a space of vector-valued functions may be inefficient due to the curse of dimensionality.

Intuitively, our method and its main property can be described as follows. We create a feature space based on the random function \( Y \) using a reproducing kernel, and perform projection of the random function \( X \), or the tensor product \( X \otimes X \), or some other related objects, on to the feature space. Under certain symmetric conditions on the distribution of \( X \), these projections can be used to recover the space spanned by \( f_1, \ldots, f_d \) in (1.2). The weak conditional moment is introduced as a way to define these projections in precise mathematical terms, and to suggest means to implement them. The projections reduce to conditional expectations if the feature space is sufficiently large, but if not, they still provide valid estimates of the dimension reduction space.

In addition to the novelty of weak conditional expectation, this paper breaks new grounds in several other ways. First, we give a new formulation of the SDR problem (1.2) using linear operators, which makes its similarity with the classical SDR more transparent and the structure of the generalized theory more coherent. Second, we generalize the Directional Regression (Li and Wang [28]), a commonly used SDR method in the classical setting, to the functional setting, which has never been done previously. Third, our new approach accommodates multivariate functional predictor and response conveniently and in a systematic manner. Finally, we expect the weak conditional moment to have wide applications in Statistics and Machine Learning beyond the current context, because conditional expectation plays the central role in many methods, and is hard to estimate in the high-dimensional setting. Replacing it by a more flexible object opens up a new avenue for solving difficult estimation problems in these areas.
Li and Song [27] extended nonlinear sufficient dimension reduction (Li, Artemiou and Li [25], Lee, Li and Chiaromonte [19]) to functional data. Although nonlinear SDR is a more general theoretical framework than linear SDR, as argued in Li, Artemiou and Li [25], the latter plays a special role that cannot be replaced by the former. For example, linear SDR is more interpretable and more parsimonious than nonlinear SDR.

Functional data analysis have received increasing attention in the recent years due to the demands from modern applications. Many statistical methods have been extended to the functional setting. In a broader sense, the development here is also a natural extension of the recent works on functional generalized linear models (Müller and Stadtmüller [31], Yao, Müller and Wang [37, 38]), functional single index models (Jiang and Wang [18]), and function-on-function regression (Ferraty, Van Keilegom and Vieu [10], Ivanescu et al. [17]).

The rest of the paper is organized as follows. In Section 2, we give an overview of random elements in Hilbert spaces, their moments and conditional moments. In Section 3, we introduce the theoretical framework of functional SDR. In Section 4, we introduce the concept of weak conditional expectation and develop its theoretical properties. In Sections 5, 6, and 7, we introduce three functional SDR estimators at the population level: the Weak Inverse Regression Estimator, the Weak Average Variance Estimator, and the Weak Directional Regression. In Sections 8 and 9, we establish the consistency and develop the convergence rate the WIRE, as well as the convergence rate of the optimal tuning parameter. In Section 10, we conduct simulation experiments to investigate the performance of the three methods and compare them with existing methods. In Section 11, we apply our methods to a bike sharing data set, where both the predictor and the response are functions.

Due to limited space some technical details, further extensions, and additional simulations and applications are presented in the online Supplementary Material. Specifically, it includes:

1. some additional theoretical results;
2. all the proofs of the theorems in the paper, as well as the lemmas that are needed in the proofs;
3. the implementing algorithms for WIRE, WAVE, and WDR;
4. the extension to the cases where both $X$ and $Y$ are multivariate functional data;
5. a cross-validation procedure to determine the tuning parameter in the Tychonoff regularization;
6. some additional simulation studies to investigate the sensitivity of the
2. Random functions and operators and their moments. Let \((\Omega, \mathcal{F}, P)\) be a probability space. Let \(\mathcal{H}\) be a Hilbert space with inner product \((\cdot, \cdot)_{\mathcal{H}}\). Let \(\mathcal{B}\) be the Borel \(\sigma\)-field generated by open sets in \(\mathcal{H}\), and \(U : \Omega \to \mathcal{H}\) a random element measurable with respect to \(\mathcal{F}/\mathcal{B}\). In our context, \(U\) is a random function representing functional data. If \(E\|U\|_{\mathcal{H}} < \infty\), then the linear functional \(\mathcal{H} \to \mathbb{R}, f \mapsto E\langle f, U \rangle_{\mathcal{H}}\) is bounded. By Riesz’s representation theorem, there is a unique \(f_0 \in \mathcal{H}\) such that

\[
\langle f_0, f \rangle_{\mathcal{H}} = E\langle f, U \rangle_{\mathcal{H}} \quad \text{for all } f \in \mathcal{H}.
\]

The function \(f_0\) is defined as the mean of \(U\), written as \(E(U)\) (see, for example, Bosq [3], Bonaccorsi and Priola [2]). We will also denote \(E(U)\) by \(M_U\).

Let \(\mathcal{G}\) be another Hilbert space, and \(C(\mathcal{G}, \mathcal{H})\) a Banach space of linear operators from \(\mathcal{G}\) to \(\mathcal{H}\). A random operator \(A\) in \(C(\mathcal{G}, \mathcal{H})\) is a mapping from \(\Omega\) to \(C(\mathcal{G}, \mathcal{H})\) measurable with respect to the Borel \(\sigma\)-field generated by the open sets in \(C(\mathcal{G}, \mathcal{H})\). If \(E(\|A\|_{\mathcal{C}(\mathcal{G}, \mathcal{H})}) < \infty\), then the mapping \(\mathcal{G} \times \mathcal{H} \to \mathbb{R}, (f_1, f_2) \mapsto E(\langle Af_1, f_2 \rangle_{\mathcal{H}})\) is a bounded bilinear form. By Theorem 2.2 of Conway [5], there is a unique operator \(B \in C(\mathcal{G}, \mathcal{H})\) such that \(\langle B f_1, f_2 \rangle_{\mathcal{H}} = E(\langle Af_1, f_2 \rangle_{\mathcal{H}})\). The operator \(B\) is defined to be the mean of \(A\), written as \(E(A)\). See, for example, Li and Song [27].

This paper will involve several types of \(C(\mathcal{G}, \mathcal{H})\): the set of bounded operators \(\mathcal{B}(\mathcal{G}, \mathcal{H})\), compact operators \(\mathcal{B}_\infty(\mathcal{G}, \mathcal{H})\), Hilbert-Schmidt operators \(\mathcal{B}_2(\mathcal{G}, \mathcal{H})\), trace-class operators \(\mathcal{B}_1(\mathcal{G}, \mathcal{H})\). It is well known (Weidmann [36]) that

\[
\mathcal{B}_1(\mathcal{G}, \mathcal{H}) \subseteq \mathcal{B}_2(\mathcal{G}, \mathcal{H}) \subseteq \mathcal{B}_\infty(\mathcal{G}, \mathcal{H}) \subseteq \mathcal{B}(\mathcal{G}, \mathcal{H}).
\]

When \(\mathcal{G} = \mathcal{H}\), we write \(C(\mathcal{G}, \mathcal{H}) = C(\mathcal{H})\).

For \(f \in \mathcal{H}\), \(g \in \mathcal{G}\), the tensor product \(f \otimes g\) is the operator \(\mathcal{G} \to \mathcal{H}\), \(h \mapsto f(g, h)_{\mathcal{G}}\). It is easy to check that \(f \otimes g \in \mathcal{B}_1(\mathcal{G}, \mathcal{H})\) with \(\|f \otimes g\|_{\mathcal{B}_1(\mathcal{G}, \mathcal{H})} \leq \|f\|_{\mathcal{H}}\|g\|_{\mathcal{G}}\). Let \(V : \Omega \to \mathcal{H}\), \(U : \Omega \to \mathcal{G}\) be a random elements in \(\mathcal{H}\) and \(\mathcal{G}\), respectively. If \(E\|U\|^2_{\mathcal{G}}\) and \(E\|V\|^2_{\mathcal{H}}\) are finite, then the covariance operator between \(V\) and \(U\) is defined as the linear operator from \(\mathcal{G}\) to \(\mathcal{H}\):

\[
(2.2) \quad \text{cov}(V, U) = E[(V - E(V)) \otimes (U - E(U))] = \Sigma_{VU}.
\]

It can be shown that \(\Sigma_{VU} \in \mathcal{B}_1(\mathcal{G}, \mathcal{H})\). Moreover,

\[
(2.3) \quad \Sigma_{VU} = E(V \otimes U) - (EV) \otimes (EU).
\]
When $U = V$, we denote $\text{cov}(U, V)$ by $\text{var}(U)$ and call it the variance operator. We denote the uncentered moment $E(V \otimes U)$ by $M_{VU}$. An alternative way of defining the covariance operator is via Bochner’s integral: see, for example, Hsing and Eubank [15].

Conditional moments are defined in a similar fashion. Let $\mathcal{I}$ be a third Hilbert space and $W : \Omega \to \mathcal{I}$ be a random element. For a fixed $w \in \mathcal{I}$, consider the linear functional $\mathcal{H} \to \mathbb{R}$, $f \mapsto E(\langle f, U \rangle_{\mathcal{H}} | W = w)$. 

Under the assumption $E(\|U\|_{\mathcal{H}} | W = w) < \infty$ for each $w$, the above linear functional is bounded. By Riesz’s representation theorem, there is $f_w \in \mathcal{H}$ such that $\langle f, E(\langle f, U \rangle_{\mathcal{H}} | W = w) \rangle_{\mathcal{H}} = E(\langle f, U \rangle_{\mathcal{H}} | W = w)$. We define $f_w$ as the conditional mean $E(U | W = w)$, and denote it by $M_{U|W}(w)$.

Similarly, we define the conditional second-order moment operator $M_{VU|W}(w) = E(V \otimes U | W = w)$ through the relation

$$\langle f, E(V \otimes U | W = w) \rangle_{\mathcal{H}} = E(\langle f, U \rangle_{\mathcal{H}} \langle U, g \rangle_{\mathcal{G}} | W = w).$$

(2.4)

We define conditional covariance as

$$\text{cov}(V, U | W = w) = M_{VU|W}(w) - M_{V|W}(w) \otimes M_{U|W}(w) \equiv \Sigma_{VU|W}(w).$$

We denote $\text{cov}(U, U | W)$ by $\text{var}(U | W)$.

The following well known relation about conditional mean and conditional variance still holds for random elements in Hilbert spaces:

$$\text{var}(U) = E[\text{var}(U | V)] + \text{var}[E(U | V)].$$

(2.5)

Note that this also implies that, if $E(\|U\|_{\mathcal{H}}^2 | W) < \infty$, then both $E[\text{var}(U | V)]$ and $\text{var}[E(U | V)]$ are trace-class operators.

3. Functional dimension reduction subspaces. For a linear operator $A \in \mathcal{B}(\mathcal{K}, \mathcal{H})$, let $\ker(A) = \{ h : A(h) = 0 \}$ be the kernel of $A$, let $\text{ran}(A) = \{ A(h) : h \in \mathcal{K} \}$ be the range of $A$, let $\overline{\text{ran}}(A)$ be the closure of $\text{ran}(A)$, and let $\text{dom}(A)$ be the domain of $A$. Let $J$ be an interval in $\mathbb{R}$ representing time. Let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be separable Hilbert spaces of functions defined on $J$. These functions can be real- or vector-valued functions. Let $X : \Omega \to \mathcal{H}_X$ and $Y : \Omega \to \mathcal{H}_Y$ be Borel random functions in $\mathcal{H}_X$ and $\mathcal{H}_Y$. Here, we allow $\mathcal{H}_Y$ to be a Euclidean space, so as to cover the cases where $Y$ is a random variable or vector, as assumed by the current literature. In this case $\mathcal{H}_Y$ is unrelated to the interval $J$. 

**Definition 1.** If there is a $T \in \mathcal{B}(\mathcal{H}_X, \mathbb{R}^d)$ such that

\[(3.1) \quad Y \mathbb{1} X | T(X),\]

then $\text{ran}(T^*)$ is a functional sufficient dimension reduction subspace (or functional SDR subspace), where $(\cdot)^*$ indicates the adjoint operator.

This formulation, though appearing different, is equivalent to that of Ferré and Yao [11]. Indeed, it can be shown, by Riesz’s representation theorem, that there exist $f_1, \ldots, f_d \in \mathcal{H}_X$ such that

$$T(X) = (T_1(X), \ldots, T_d(X)) = (\langle f_1, X \rangle_{\mathcal{H}_X}, \ldots, \langle f_d, X \rangle_{\mathcal{H}_X}).$$

Since, for any $v \in \mathbb{R}^d$, 

$$v^T T(X) = \sum v_i \langle f_i, X \rangle_{\mathcal{H}_X} = \sum v_i f_i,$$

we have $T^*(v) = \sum v_i f_i$, and hence $\text{ran}(T^*) = \text{span}(f_1, \ldots, f_d)$, which is the formulation of Ferré and Yao [11].

An advantage of this alternative formulation is that it focusses our attention on the linear operator $T$, whose estimation is more direct. A second advantage is that it is completely parallel to the classical SDR problem $Y \mathbb{1} X | \beta^T X$, making the analogies of the two settings transparent.

The linear operator $T$ in the relation (3.1) is not identifiable: this relation is equivalent to $Y \mathbb{1} X | (AT)(X)$ for any invertible matrix $A \in \mathbb{R}^{d \times d}$. However, the subspace

$$\text{ran}(T^*) = \text{ran}((AT)^*) = \text{ran}(T^* A^T)$$

is uniquely defined. That is why $\text{ran}(T^*)$, instead of $T^*$, is taken as the object of estimation. Let $\mathcal{S}$ be the class of all sufficient dimension reduction subspaces:

$$\mathcal{S} = \{ \mathcal{I} \subseteq \mathcal{H}_X : \mathcal{I} \text{ is a functional SDR subspace} \}.$$

We make the following assumption.

**Assumption 1.** $\cap_{\mathcal{I} \in \mathcal{S}} \mathcal{I} \in \mathcal{S}.$

That is, the intersection of all functional SDR subspaces is itself a functional SDR subspace. This condition is very mild in the classical SDR setting (see Yin, Li and Cook [39]), and we do not expect it to exert any real restriction here, either. In this paper, we take this condition for granted without developing its further sufficient conditions.
Definition 2. Under Assumption 1, we call \( \cap_{\mathcal{F} \in \mathcal{F}} S \) the functional central dimension reduction subspace — or functional central subspace — of \( Y \) versus \( X \), and denote it by \( \mathcal{S}_{Y|X} \).

Throughout the rest of the paper \( T \in \mathcal{B}(\mathcal{H}_X, \mathbb{R}^d) \) always denotes an operator such that \( \text{ran}(T^*) = \mathcal{S}_{Y|X} \). To provide intuition about the functional central subspace we consider the following example. Suppose \( \mathcal{H}_X \) and \( \mathcal{H}_Y \) are Hilbert spaces of functions defined on \([0, 1]\) furnished with the \( L^2([0, 1])\)-inner product \( \int_0^1 f(t)g(t)dt \). Let \( X \) be a random element in \( \mathcal{H}_X \), and let \( Y, \epsilon \) be random elements in \( \mathcal{H}_Y \) satisfying

\[
Y(s) = f_1 \left( \int_0^1 \beta_1(t)X(t)dt, s \right) + f_2 \left( \int_0^1 \beta_2(t)X(t)dt, s \right) \epsilon(s),
\]

(3.2)

where \( f_1, f_2 \) are arbitrary functions from \( \mathbb{R} \times [0, 1] \) to \( \mathbb{R} \). In this case the functional central subspace is again spanned by \( \{\beta_1, \beta_2\} \).

4. Weak conditional expectation. The germ of the idea of the weak conditional expectation (or weak conditional moment) was already apparent in Lee, Li and Chiaromonte [19], where the authors defined the heteroscedastic conditional covariance operator through a conditional bilinear form, which itself is the evaluation of the conditional mean operator. This is, indeed, also the route taken here. Tracing the idea further back, the Parametric Inverse Regression (PIR) proposed by Bura and Cook [4], and the Canonical Correlation SDR estimator (CANCOR) proposed by Fung et al. [14] can also be viewed as the special applications of weak conditional moment. When \( X \) is a random vector and \( \mathcal{M}_Y \) is a finite-dimensional Hilbert space of real-valued functions of \( Y \), as in the cases of PIR and CANCOR, the weak conditional expectation is relatively easy to write down. However, in the current infinite-dimensional setting, a fully rigorous definition has to be made in a rather abstract form. This is perhaps to be expected, considering that even the rigorous definition of conditional expectation involves abstract constructions such as the Radon-Nikodym derivative (see, for example, Billingsley [1]).

To motivate weak conditional expectation, let us first outline the construction of the Parametric Inverse Regression in the classical setting where \( X = (X^1, \ldots, X^p)^T \) is a random vector and \( Y \) is a random variable. Without loss of generality, assume that \( E(X) = 0 \). Instead of estimating \( E(X|Y) \) via taking slices, Bura and Cook [4] proposed to regress \( X \) on a set of functions of \( Y \). Let \( \mathcal{M}_Y \) be the linear space spanned by a finite set of functions, say \( f_1, \ldots, f_m \), and \( f_{1:m} = (f_1, \ldots, f_m)^T \). The \( L_2(P) \)-projection of the \( i \)th
component of $X$ on to $\mathcal{M}_Y$ is

$$h_i(y) = E[X^t f_{1:m}^T(Y)](E[f_{1:m}(Y)f_{1:m}^T(Y)])^{-1} f_{1:m}(y).$$

It can be shown, under the assumption that $E(X|b^TX)$ is a linear function of $b^TX$ for any $b \in \mathbb{R}^p$, that the vector $h_{1:p}(y) = (h_1(y), \ldots, h_p(y))^T$ belongs to the central subspace $\mathcal{S}_{Y|X}$ for any $y$. For a proof, see, for example, Li [23].

Despite the above relatively simple construction, there is no elementary way of extending the object $h_{1:p}(y)$ to functional $X$, which need not have a finite or countable number of components. This is the chief reason for introducing the weak conditional moment.

The following alternative representation of $h_{1:p}(y)$ gives us a hint at how it might be extended to the infinite-dimensional case. Let $P_{\mathcal{M}_Y}$ be the $L_2(P)$-projection on to $\mathcal{M}_Y$. That is, for any $g(x) \in L_2(P)$, $P_{\mathcal{M}_Y} g$ is the member $E[g(X) f_{1:m}^T(Y)](E[f_{1:m}(Y)f_{1:m}^T(Y)])^{-1} f_{1:m}$ of $\mathcal{M}_Y$. Next, let $A : \mathbb{R}^p \to \mathcal{M}_Y$ be the linear operator defined by

$$A(b) = P_{\mathcal{M}_Y} g_b,$$

where $g_b(x)$ is the linear function $b^Tx$, $b \in \mathbb{R}^p$. Then

$$A(b) = P_{\mathcal{M}_Y} (b^T x) = b^T (P_{\mathcal{M}_Y} x^1, \ldots, P_{\mathcal{M}_Y} x^p)^T = b^T h_{1:p}.$$

Thus, if, for each $y \in \mathbb{R}$, we let $A_y$ be the linear functional on $\mathbb{R}^p$:

$$A_y(b) = (Ab)(y),$$

then $A_y(b) = b^T h_{1:p}(y)$. In other words, $h_{1:p}(y)$ is the Riesz representation of the linear functional $A_y$ on $\mathbb{R}^p$. The advantage of this alternative representation is that it does not rely on the fact that $X$ has a finite number of components, thus paving the way for the extension to the infinite-dimensional setting.

To sum up, we extract the following key ingredients from the above construction to form the basis for the construction of weak conditional moments:

1. $\mathcal{G} = \mathbb{R}$;
2. $\mathcal{H} = \mathbb{R}^p$;
3. $\mathcal{M} = \text{span}\{f_1, \ldots, f_m\}$, where $f_i$ are real-valued function on $\mathcal{G}$;
4. the linear operator $A : \mathcal{H} \to \mathcal{M}$ defined by $A(b) = P_{\mathcal{M}_Y} g_b$, where $g_b(x) = b^T x$ and $P_{\mathcal{M}_Y}$ is the $L_2(P)$-projection on to $\mathcal{M}$. 
Interestingly, the idea of PIR or CANCOR was never extended to second-order methods such as Sliced Average Variance Estimate (SAVE) and Directional Regression (DR) in the classical setting where \( X \) is a random vector and \( Y \) is a random variable, perhaps due to the fact that, without the aid of the weak conditional moment, the path to the generalization is not easy to discern. Thus, the two estimators to be developed later in this paper, the Weak Average Variance Estimator and the Weak Directional Regression, are novel even for classical SDR.

We begin our introduction of the weak conditional expectation with the notions of Carleman operator and its inducing function.

**Definition 3.** Let \( \mathcal{G} \) be a set, \( \mathcal{M} \) a Hilbert space of \( \mathbb{R} \)-valued functions on \( \mathcal{G} \), \( \mathcal{H} \) another Hilbert space, and \( A : \mathcal{H} \to \mathcal{M} \) a linear operator. If, for each \( y \in \mathcal{G} \), the linear functional \( A_y : \mathcal{H} \to \mathbb{R}, f \mapsto (Af)(y), \) is bounded, then we call \( A \) an extended Carleman operator. Let \( \lambda_A(y) \) be the Riesz representation of \( A_y \). The function \( \lambda_A : \mathcal{G} \to \mathcal{H}, y \mapsto \lambda_A(y) \) is called the inducing function of \( A \).

Thus, the defining relation of the inducing function \( \lambda_A(y) \) is, for each \( f \in \mathcal{H}, \langle \lambda_A(y), f \rangle_{\mathcal{H}} = (Af)(y) \). In the theory of linear operators (Weidmann [36]), when \( \mathcal{M} \) is a subset of the class of square-integrable functions on \( \mathcal{G} \), an operator \( A \) that satisfies the conditions in Definition 3 is called a Carleman operator, and the function \( \lambda_A \) is called the inducing function of \( A \). The above definition is slightly more general because \( \mathcal{M} \) is not required to be a subspace of \( L_2(P_Y) \), and, in particular, the inner product in \( \mathcal{M} \) is not required to be the same as that in \( L_2(P_Y) \). This extra generality is needed for our development: we will assume \( \mathcal{M} \) to be a reproducing kernel Hilbert space, whose inner product is different from that of \( L_2(P_Y) \).

We now define weak conditional expectation. Let \( \kappa_Y : \mathcal{H}_Y \times \mathcal{H}_Y \to \mathbb{R} \) be a positive definite kernel function. This kernel can be constructed by the inner product in \( \mathcal{H}_Y \). For example,

\[
(4.1) \quad \kappa_Y(y_1, y_2) = \exp(-\gamma \|y_1 - y_2\|^2_{\mathcal{H}_Y}), \quad \kappa_Y(y_1, y_2) = (1 + \alpha \langle y_1, y_2 \rangle_{\mathcal{H}_Y})^d,
\]

where \( \gamma > 0, \alpha \in \mathbb{R}, \) and \( d \in \{1, 2, \ldots\} \). In general, \( \kappa_Y(y_1, y_2) \) is a function of the form

\[
\kappa_Y(y_1, y_2) = \rho(\|y_1\|^2_{\mathcal{H}_Y}, \langle y_1, y_2 \rangle_{\mathcal{H}_Y}, \|y_2\|^2_{\mathcal{H}_Y}).
\]
Let \( \mathcal{M}_Y \) be the reproducing kernel Hilbert space generated by the kernel \( \kappa_Y \). We call the pair of Hilbert spaces \((\mathcal{H}_Y, \mathcal{M}_Y)\) nested, because the kernel of the latter is determined by the inner product of the former. This structure is also used in Li and Song [27].

Let \( \mathcal{H}_U \) be a Hilbert space and let \( U \) be a random element that takes values in \( \mathcal{H}_U \). In later sections, the random element \( U \) can be either \( X \) or \( X \otimes X \). Let \( M_{YU} = E[\kappa(\cdot, Y) \otimes U] \), let \( M_{YY} = E[\kappa(\cdot, Y) \otimes \kappa(\cdot, Y)] \). Note that, if we follow the notation in Section 2 (fourth paragraph) exactly, then we should have used \( M_{VU} \) to denote \( E[\kappa_Y(\cdot, Y) \otimes U] \) where \( V \) is the random element \( \kappa_Y(\cdot, Y) \) of \( \mathcal{M}_Y \). However, doing so would keep the response out of the notation. Thus we use the notation \( M_{YU} \) with this caveat in mind.

Throughout the rest of the paper, we make the following assumption.

**Assumption 2.** \( \ker(M_{YY}) = \{0\}, \quad \operatorname{ran}(M_{YU}) \subseteq \operatorname{ran}(M_{YY}) \).

The first condition is mild: since \( \langle f, M_{YY}f \rangle_{\mathcal{M}_Y} = E[f^2(Y)] \), \( \langle f, M_{YY}f \rangle_{\mathcal{M}_Y} = 0 \) implies \( f(Y) = 0 \) almost surely. Thus, if \( \kappa_Y \) is a continuous kernel, then \( \ker(M_{YY}) = 0 \). The second condition requires that, for any \( f \in \mathcal{H}_U \), \( M_{YU}f \) is sufficiently concentrated on the low-frequency components of \( M_{YY} \), which is a kind of collective smoothness (see Li [24] for further discussions on this point). Under the first condition, \( M_{YY} : \mathcal{M}_Y \to \operatorname{ran}(M_{YY}) \) is a one-to-one transformation. We define \( M_{YY}^* \) to be the linear operator from \( \operatorname{ran}(M_{YY}) \) to \( \mathcal{M}_Y \) such that, for any \( f \in \operatorname{ran}(M_{YY}) \), \( M_{YY}^*f \) is the unique \( g \in \mathcal{M}_Y \) satisfying \( f = M_{YY}g \). We call the linear operator \( M_{YY}^* \) the Moore-Penrose inverse of \( M_{YY} \). This definition is essentially (but not exactly) the same as that given in [15]. Some properties of the Moore-Penrose inverse are given in Section 1.2 on the Supplementary Material. Under the second condition, \( \operatorname{ran}(M_{YU}) \subseteq \operatorname{dom}(M_{YY}^*) \), thus the operator \( M_{YY}^*M_{YU} \equiv R_{YU} \) is well defined. Lee, Li and Zhao [20] refers to the operator \( R_{YU} \) as the regression operator. See also Li [22].

**Definition 4.** If the regression operator \( R_{YU} \) is a Carleman operator, then the random element

\[
\Omega \to \mathcal{M}_Y, \quad \omega \mapsto \lambda_{R_{YU}}(Y(\omega)),
\]

where \( \lambda_{R_{YU}} \) is the inducing function of the Carleman operator \( R_{YU} \) as defined in Definition 3, is called the weak conditional moment of \( U \) given \( Y \), and is written \( E(U;Y) \).

As shown in Fukumizu, Bach and Gretton [13], the operator \( M_{YY} \), when defined, is a Hilbert Schmidt operator, which implies that \( M_{YY}^* \) is an unbounded operator. Since an unbounded operator is discontinuous, it cannot
be estimated accurately by statistical functionals. Fortunately, in our development $M^*_YY$ never appears alone; it always appear in the form $M^*_YYA$, where $A$ is an operator that maps into the domain of $M^*_YY$. By Weidmann [36], Theorem 6.12, $M^*_YYA$ is a Carleman operator if it is a Hilbert Schmidt operator. It is reasonable to assume $M^*_YYA$ to be a Hilbert Schmidt operator because, as argued in Li and Song [27], this amounts to imposing a type of smoothness in the relation between $U$ and $Y$.

Two explicit examples of the weak conditional expectation are given in Section 4.1 in the Supplementary Material, one of which differs from the conditional expectation, and the other agrees with the conditional expectation. These examples illustrate the construction of the weak conditional expectation and provide more insights into this concept.

The next theorem shows that the weak conditional moment is indeed a generalization of the conditional expectation. In the following, we use $E(U;Y = y)$ to denote the evaluation of $\lambda_{RYU}$ at $y$; that is, $E(U;Y = y) = \lambda_{RYU}(y)$. Thus, $E(U;Y = y)$ is a member of $\mathcal{H}_U$.

**Theorem 1.** If $R_{YU}$ is a Carleman operator and $\mathcal{M}_Y$ is dense in $L_2(P_Y)$, then $E(U;Y) = E(U|Y)$.

This theorem shows that the weak conditional expectation is a generalization of the conditional expectation. Thus, for example, the Weak Inverse Regression Estimate we introduce further on can be viewed as a generalization of the Sliced Inverse Regression, the latter of which is based on the conditional expectation $E(X|Y)$. Recall that conditional expectation $E(U|Y)$ satisfies the following condition

$$E[E(U|Y)g(Y)] = E[Ug(Y)] \quad \text{for all } g \in L_2(P_Y).$$

This identity holds in a weaker form for weak conditional expectation.

**Proposition 1.** Suppose $\mathcal{H}$ is a Hilbert space and $U$ is a random element in $\mathcal{H}$. Then, for any $f \in \mathcal{H}$ and $h \in \mathcal{M}_Y$,

$$E[(f, E(U;Y))_{\mathcal{H}}h(Y)] = E[(f, U)_{\mathcal{H}}h(Y)].$$

We should mention here that, unlike in the case conditional expectation, weak conditional expectation does not in general obey the identity

$$(4.2) \quad E[E(U;Y)] = E(U).$$

In fact, this holds only when $\mathcal{M}_Y$ contains the constant function $1 : \mathcal{H}_Y \rightarrow \mathbb{R}, y \mapsto 1$. The lack of this equality explains why some of the SDR estimators developed later are of different forms than their classical counterparts.
Note that in this paper we do not assume \( \mathcal{M}_Y \) to be dense in \( L_2(P_Y) \), so that the proposed methods cannot be described using conditional expectation. For example, if \( \mathcal{M}_Y \) is generated by the second kernel in (4.1), known as the polynomial kernel, then \( \mathcal{M}_Y \) is not dense in \( L_2(P_Y) \).

We now develop two special cases of weak conditional expectation that are of interest to us where \( U = X \) and \( U = X \otimes X \). We make the following assumptions.

**Assumption 3.**
(a) \( E(\|X\|^2_{\mathcal{H}_X}) < \infty \), (b) \( E(\|X\|^4_{\mathcal{H}_X}) < \infty \).

Note that condition (b) is stronger than (a), but they will be used in different places.

**Assumption 4.** \( E[\kappa_Y(Y,Y)] < \infty \).

This assumption implies \( \mathcal{M}_Y \subseteq L_2(P_Y) \), because, by the Cauchy-Schwarz inequality, for any \( f \in \mathcal{M}_Y \),

\[
E[f^2(X)] = E[(f, \kappa_Y(\cdot, Y))^2_{\mathcal{M}_Y}] \leq \|f\|^2E[\kappa_Y(Y,Y)] < \infty.
\]

Let \( \mathcal{H}_X \otimes \mathcal{H}_X \) be the tensor product space — that is, the Hilbert space spanned by the operators \( \{f \otimes g : f, g \in \mathcal{H}_X\} \), with its inner product determined by

\[
(f_1 \otimes g_1, f_2 \otimes g_2)_{\mathcal{H}_X \otimes \mathcal{H}_X} = \langle f_1, f_2 \rangle_{\mathcal{H}_X} \langle g_1, g_2 \rangle_{\mathcal{H}_X}.
\]

Let

\[
M_{YY} = E[\kappa(\cdot, Y) \otimes \kappa(\cdot, Y)], \quad M_{YX} = E[\kappa(\cdot, Y) \otimes (X - EX)], \quad M_{YXX} = E[\kappa(\cdot, Y) \otimes (X - EX) \otimes (X - EX)].
\]

It is shown in Theorem S.1 in the Supplementary Material that, under Assumptions 3 and 4, these are bounded operators in \( \mathcal{B}(\mathcal{M}_Y, \mathcal{M}_Y) \), \( \mathcal{B}(\mathcal{H}_X, \mathcal{M}_Y) \), and \( \mathcal{B}(\mathcal{H}_X \otimes \mathcal{H}_X, \mathcal{M}_Y) \), respectively.

We apply the general Definition 4 to the random function \( X \) and random operator \( X \otimes X \) to define \( E(X; Y) \) and \( E(X \otimes X; Y) \) as the first- and second-order weak conditional moments of the random function \( X \). These definitions require the following assumption.

**Assumption 5.** The regression operators

\[
R_{YX} = M_{YY}^t M_{YX} : \mathcal{H}_X \to \mathcal{M}_Y, \quad R_{YXX} = M_{YY}^t M_{YXX} : \mathcal{H}_X \otimes \mathcal{H}_X \to \mathcal{M}_Y
\]

are Carleman operators.
We now define the weak conditional variance operator.

**Definition 5.** Under Assumptions 3 through 5, the operator

\[
E(X \otimes X |_Y) - E(X |_Y) \otimes E(X |_Y)
\]

is called the weak conditional variance operator of \(X\) given \(Y\) with respect to \(\mathcal{M}_Y\), and is written as \(\text{var}(X |_Y)\).

The next corollary is a direct consequence of Theorem 1.

**Corollary 1.** If Assumptions 3 through 5 hold and \(\mathcal{M}_Y\) is dense in \(L^2(P_Y)\), then

\[
E(X |_Y) = E(X |_Y), \quad E(X \otimes X |_Y) = E(X \otimes X |_Y), \quad \text{var}(X |_Y) = \text{var}(X |_Y).
\]

5. **Weak Inverse Regression Estimator (WIRE).** Sliced Inverse Regression (Li [21], Cook [6]) is based on the conditional mean \(E(X |_Y)\) in \(\mathbb{R}^p\), or more specifically, the generalized eigenvalue problem

\[
\text{var}[E(X |_Y)]v = \lambda \text{var}(X)v.
\]

Here, the eigenvector \(v\) of a generalized eigenvalue problem \(Av = \lambda Bv\) means \(v = B^{-1/2}u\), where \(u\) is an eigenvector of \(B^{-1/2}AB^{-1/2}\). Previous generalizations of SIR to the functional setting, such as those proposed by Ferré and Yao [11] and Ferré and Yao [12], also use this formula, except that the meaning of \(E(X |_Y), \text{var}(X)\) and \(\text{var}[E(X |_Y)]\) have been extended to the infinite-dimensional case. We now use weak conditional expectation to further extend functional SIR through the generalized eigenvalue problem

\[
E[E(X - EX |_Y) \otimes (X - EX |_Y)]f = \lambda \text{var}(X)f.
\]

(5.1)

The significance of this generalization is that it steers us away from attempting to estimate \(E(X |_Y)\) through slicing or kernel smoothing, which can be cumbersome and inefficient when the response is also a random function. Furthermore, being able to limit the scope and dimension of \(\mathcal{M}_Y\) without causing bias offers us a regularizing mechanism on the dimension reduction estimator, which is particularly important when \(E(X |_Y)\) is a highly complex function in the infinite-dimensional \(L^2(P_Y)\) space. For example, we can choose \(\kappa_Y\) to be a linear kernel regardless of the complexity of \(E(X |_Y)\), dampening the effect of curse of dimensionality.

As in classical SDR (Li [21]), we assume \(E[X | T(X)]\) to be linear in \(T(X)\).
Assumption 6. There is bounded linear operator \( L : \text{ran}(T) \to \mathcal{H}_X \) such that \( E(X|T(X)) = LT(X) \).

This assumption leads to the following theorem which is parallel to a well known result in the classical setting (see, for example, Cook [6], page 57).

Theorem 2. Under Assumptions 3(a), 6, and \( E(X) = 0 \), we have

\[
E[X|T(X)] = \Sigma_{XX} T^* (T \Sigma_{XX} T^*)^\dagger T(X).
\]

Because \( T \) belongs to \( B(\mathcal{H}_X, \mathbb{R}^d) \), \( T \Sigma_{XX} T^* \) is a \( d \times d \) symmetric matrix. Furthermore, we can deduce that \( T^* (T \Sigma_{XX} T^*)^\dagger T \Sigma_{XX} \) is the projection on to \( \text{ran}(T^*) \) relative to the \( L_2(P_X) \)-inner product \( \langle f, g \rangle \mapsto \langle f, \Sigma_{XX} g \rangle_{\mathcal{H}_X} \). For this reason we denote this operator by \( P_{T^*}(\Sigma_{XX}) \), bearing in mind, though, that \( P_{T^*}(\Sigma_{XX}) \) is not a projection in the \( \mathcal{H}_X \)-inner product.

To proceed further we make two more assumptions. First, note that the important part of the Hilbert space \( \mathcal{H}_X \) is the subspace \( \text{ran}(\Sigma_{XX}) \) because, for any \( f \perp \text{ran}(\Sigma_{XX}) \), \( \langle f, \Sigma_{XX} f \rangle_{\mathcal{H}_X} = \text{var}([\langle f, X \rangle]_{\mathcal{H}_X}) = 0 \), which implies \( f \) is orthogonal to the support of \( X - E(X) \). It is then reasonable to make the following assumption, which is not restrictive at all.

Assumption 7. \( \mathcal{S}_{Y|X} \subseteq \text{ran}(\Sigma_{XX}) \).

The next assumption requires that the kernel \( \kappa_Y \) is a continuous function.

Assumption 8. The kernel function \( \kappa_Y : \mathcal{H}_Y \times \mathcal{H}_Y \to \mathbb{R} \) is continuous.

The next theorem shows that, under the linear conditional mean assumption, \( E(X - E(X)|Y) \) belongs to \( \Sigma_{XX} \mathcal{S}_{Y|X} \). This implies the estimator based on \( E(X - E(X)|Y) \) is an unbiased estimate of the functional central subspace. A salient point of the theorem is that, even though the weak conditional expectation \( E(X - E(X)|Y) \) may be far away from the conditional expectation \( E(X - E(X)|Y) \), this does not disrupt the unbiasedness of the estimator derived from \( E(X - E(X)|Y) \). This is, indeed, the main point of introducing the weak conditional moment.

Theorem 3. Under Assumptions 3(a), 4, 5, 6, 7, and 8 we have

\[
\text{ran}[E(X - E(X)|Y)] \in \Sigma_{XX} \mathcal{S}_{Y|X}.
\]

This theorem implies that the collection of eigenfunctions with nonzero eigenvalues in the generalized eigenvalue problem (5.1) span a subspace of
\( \mathcal{S}_{Y|X} \). That is,

\[
\text{span}\{ f : E[E(X - E(X)|Y) \otimes E(X - E(X)|Y)]f = \lambda \Sigma_{XX} f, \lambda > 0\} \subseteq \mathcal{S}_{Y|X}.
\]

For the rest of this paper we assume that \( X \) is centered without loss of generality.

**Assumption 9.** \( E(X) = 0 \)

Under this assumption, (5.2) and (5.3) reduce to

\[
\text{ran}[E(X|Y)] \in \Sigma_{XX} \mathcal{S}_{Y|X},
\]

\[
\text{span}\{ f : E[E(X|Y) \otimes E(X|Y)]f = \lambda \Sigma_{XX} f, \lambda > 0\} \subseteq \mathcal{S}_{Y|X}.
\]

The previous generalizations of SIR to the functional case based on \( E(X|Y) \) are special cases of (5.3) when \( \mathcal{M}_Y \) is dense in \( L_2(\mathcal{P}_Y) \). We call the estimator based on (5.3) the Weak Inverse Regression Estimator, or WIRE. The next proposition gives an alternative expression of \( E[E(X|Y) \otimes E(X|Y)] \).

**Proposition 2.** Under Assumptions 3(a), 4, 5, and 9, we have

\[
E[E(X|Y) \otimes E(X|Y)] = \Sigma_{XY} M_{Y}^{\dagger} \Sigma_{YX}.
\]

When \( \kappa_Y \) is taken to be the polynomial kernel (the second kernel in (4.1)), the WIRE is a generalization of the parametric inverse regression of Bura and Cook [4] to the functional setting. In particular, when \( d = 1 \), it is a generalization of the Ordinary Least Squares for dimension reduction (Li and Duan [26]) to the functional setting. In these cases, the family \( \mathcal{M}_Y \) is rather small but still produces unbiased functional SDR estimates.

**6. Weak Averaged Variance Estimator (WAVE).** In the classical setting, the Sliced Average Variance Estimator (SAVE, Cook and Weisberg [8]) is based on the matrix \( M_{\text{SAVE}} = E[\text{var}(X) - \text{var}(X|Y)]^2 \). We propose

\[
E[\lambda H_{YXX}(Y) - \text{var}(X:Y)]^2, \text{ where } H_{YXX} = M_{YY} E[\kappa_Y(\cdot, Y)] \otimes E(X \otimes X)
\]

as the generalization \( M_{\text{SAVE}} \) to the functional setting. At first glance, this does not seem to be a generalized form of \( M_{\text{SAVE}} \), but the next lemma shows that it is. The reason underlying this apparent discrepancy is that (4.2) no longer holds in general.
**Proposition 3.** Suppose Assumptions 3(a), 3(b), 4, 5, 8, and 9 hold and $\mathcal{M}_Y$ contains the function $1 : \mathcal{H}_Y \to \mathbb{R}$, $f \mapsto 1$. Then $\lambda_{HYXX}(y) = \Sigma_{XX}$.

As in the classical setting (Cook and Weisberg, 1991), we make the following constant variance assumption, which is needed for establishing the unbiasedness of our extension of SAVE. This condition is satisfied if, for example, $X$ is a Gaussian random element in $\mathcal{H}_X$.

**Assumption 10.** $\operatorname{var}[X|T(X)]$ is a non-random operator.

We are now ready to establish that the range of the random operator $\lambda_{HYXX}(Y) - \operatorname{var}(X;Y)$ is contained in $\Sigma_{XX} \mathcal{F}_{Y|X}$, which implies the unbiasedness of our extension of SAVE. Once again, $\operatorname{var}(X;Y)$ can be very different from $\operatorname{var}(X|Y)$ but this does not affect unbiasedness.

**Theorem 4.** Under Assumptions 3(b), and Assumptions 4 through 10 we have

$$\operatorname{ran}[\lambda_{HYXX}(Y) - \operatorname{var}(X;Y)] \subseteq \Sigma_{XX} \mathcal{F}_{Y|X}.$$  

For any $\alpha > 0$, let $\Sigma^{\alpha}$ denote $(\Sigma^\alpha)^\dagger$. Theorem 4 implies that

$$\operatorname{ran}\{\Sigma^{\alpha}_{XX}(\lambda_{HYXX}(Y) - \operatorname{var}(X;Y))\Sigma^{\alpha}_{XX}\} \subseteq \Sigma^{\alpha}_{XX} \mathcal{F}_{Y|X}.$$  

Let $S(Y) = \Sigma^{1/2}_{XX}(\lambda_{HYXX}(Y) - \operatorname{var}(X;Y))\Sigma^{1/2}_{XX}$. The above relation implies that the range of the operator $E[S^2(Y)]$ is contained in $\Sigma^{1/2}_{XX} \mathcal{F}_{Y|X}$. Let $f_1, \ldots, f_r$ be the eigenfunctions of $E[S^2(Y)]$ corresponding to its nonzero eigenvalues (note that $r$ is no greater than the dimension of the central subspace), then $\Sigma^{1/2}_{XX}f_1, \ldots, \Sigma^{1/2}_{XX}f_r$ are members of the central subspace $\mathcal{F}_{Y|X}$. We call the estimation procedure based on this eigenvalue problem the Weak Average Variance Estimator, or WAVE for short. Under this definition, Theorem 4 implies that WAVE is an unbiased estimate of $\mathcal{F}_{Y|X}$.

**7. Weak Directional Regression.** Directional Regression (DR) was introduced by Li and Wang [28] to combine the first two conditional moments of $X$ and $Y$ so as to overcome some of the difficulties in SIR and SAVE. In many situations it enjoys better performance than SIR and SAVE. Let $(\tilde{X}, \tilde{Y})$ be an independent copy of $(X,Y)$. A direct generalization of DR to our setting would be based on the fact

$$2\Sigma_{XX} - E[(X - \tilde{X}) \odot (X - \tilde{X})|Y, \tilde{Y}] \in \Sigma_{XX} \mathcal{F}_{Y|X},$$

(7.1)
which is indeed true. We further generalize this to functional data using weak conditional moments. Applying the similar argument of Li and Wang [28] we can show that the above equals

\[ 2\Sigma_{XX} - E(X \otimes X | Y) - E(\tilde{X} \otimes \tilde{X} | \tilde{Y}) \]
\[ + E(X | Y) \otimes E(\tilde{X} | \tilde{Y}) + E(\tilde{X} | \tilde{Y}) \otimes E(X | Y). \]

Replacing \( \Sigma_{XX} \) by \( \lambda_{HY_{XX}}(Y) \) and the strong conditioning | by the weak conditioning \( \langle \), we propose the following operator as the generalization of (7.2):

\[ C(Y, \tilde{Y}) = \lambda_{HY_{XX}}(Y) - E(X \otimes X | Y) + E(X | Y) \otimes E(\tilde{X} | \tilde{Y}) + \lambda_{HY_{XX}}(Y) - E(\tilde{X} \otimes \tilde{X} | \tilde{Y}) + E(\tilde{X} | \tilde{Y}) \otimes E(X | Y). \]

The next theorem implies the unbiasedness of our extension of DR.

**Theorem 5.** Under Assumptions 3(b) and Assumptions 4 through 10,

\[ C(Y, \tilde{Y}) = P_{T^*}(\Sigma_{XX})C(Y, \tilde{Y})P_{T^*}(\Sigma_{XX}). \]

Letting \( D(Y, \tilde{Y}) = \Sigma_{XX}^{-1/2}C(Y, \tilde{Y})\Sigma_{XX}^{-1/2} \), by Theorem 5, if \( f_1, \ldots, f_r \) are the eigenfunctions of \( E[D^2(Y, \tilde{Y})] \), then \( \Sigma_{XX}^{1/2}f_1, \ldots, \Sigma_{XX}^{1/2} \) are members of \( \mathcal{H}_{Y|X} \).

We call the estimation procedure derived from this fact the *Weak Directional Regression*, or WDR. With this definition, Theorem 5 implies WDR is an unbiased estimate of the functional central subspace.

Although the definition of \( D(Y, \tilde{Y}) \) involves \( (X, Y) \) and \( (\tilde{X}, \tilde{Y}) \), the operator \( ED^2(Y, \tilde{Y}) \) can be expressed in such a way that it only depends on the distribution of \( (X, Y) \). This expression greatly simplifies the implementation of WDR, because instead of dealing U-statistics (which is needed for estimating quantities such as \( ED^2(Y, \tilde{Y}) \)), we simply require sample moments based on i.i.d. samples. Let

\[ L(Y) = \Sigma_{XX}^{1/2}E(X | Y), \]
\[ Q(Y) = \Sigma_{XX}^{1/2}\lambda_{HY_{XX}}(Y)\Sigma_{XX}^{1/2} - \Sigma_{XX}^{1/2}E(X \otimes X | Y)\Sigma_{XX}^{1/2}. \]

Note that \( Q(Y) \) is a random operator in \( \mathcal{B}(\mathcal{H}_X) \), and \( L(Y) \) is a random function in \( \mathcal{H}_X \). The letters \( L \) and \( Q \) indicate “linear” and “quadratic”. The next theorem gives an alternative expression of \( E[D^2(Y, \tilde{Y})] \), which will be the basis for our estimation procedure.

**Theorem 6.** Under Assumptions 3(b), 6, and 10, we have

\[ E[D^2(Y, \tilde{Y})] = 2E[Q(Y)^2] + 2E[Q(Y)]^2 + 2\{E[L(Y) \otimes L(Y)]\}^2 \]
\[ + 2E((L(Y), L(Y))_{\mathcal{H}_X}) E[L(Y) \otimes L(Y)]. \]
Again, the difference between (7.4) and its counterpart in the original Directional Regression (Li and Wang [28]), is due to the fact that (4.2) no longer holds in general.

8. Consistency and Convergence rates of WIRE. In this section we prove the consistency and develop the convergence rates of WIRE and some related quantities. We focus on WIRE because of limited space, but the argument employed here can be extended to WAVE and WDR in a straightforward (albeit quite tedious) manner.

Since, by Assumption 7, \( \mathcal{Y}_{\mathcal{X}} \subseteq \text{ran}(\Sigma_{\mathcal{X}\mathcal{X}}) \), we can, without loss of generality, reset \( \mathcal{H}_{\mathcal{X}} \) to be \( \text{ran}(\Sigma_{\mathcal{X}\mathcal{X}}) \), which amounts to assuming \( \Sigma_{\mathcal{X}\mathcal{X}} : \text{ran}(\Sigma_{\mathcal{X}\mathcal{X}}) \to \text{ran}(\Sigma_{\mathcal{X}\mathcal{X}}) \) is injective. Similarly, by Corollary S.3, \( M_{\mathcal{Y}\mathcal{Y}} \) is injective under Assumptions 4 and 8. As before, let \( \Sigma_{\mathcal{X}\mathcal{X}}^{1/2} \) and \( M_{\mathcal{Y}\mathcal{Y}}^{1/2} \) be the Moore-Penrose inverses of \( \Sigma_{\mathcal{X}\mathcal{X}} \) and \( M_{\mathcal{Y}\mathcal{Y}} \). By Proposition 2, finding vectors in \( \mathcal{Y}_{\mathcal{X}} \) via WIRE at the population level amounts to solving the eigenvalue problem

\[
(\Sigma_{\mathcal{X}\mathcal{X}}^{1/2} \Sigma_{\mathcal{X}\mathcal{Y}} M_{\mathcal{Y}\mathcal{Y}}^{1/2} \Sigma_{\mathcal{Y}\mathcal{X}}^{1/2})f = \lambda f, \quad h = \Sigma_{\mathcal{X}\mathcal{X}}^{1/2} f,
\]

where \( \Sigma_{\mathcal{X}\mathcal{X}}^{1/2} \) denotes \( (\Sigma_{\mathcal{X}\mathcal{X}}^{1/2})^{-1} \), the Moore-Penrose inverse of \( \Sigma_{\mathcal{X}\mathcal{X}}^{1/2} \).

At the sample level, we estimate \( \Sigma_{\mathcal{X}\mathcal{X}}, \Sigma_{\mathcal{X}\mathcal{Y}}, \text{ and } M_{\mathcal{Y}\mathcal{Y}} \) by

\[
\hat{\Sigma}_{\mathcal{X}\mathcal{X}} = E_n[(X - E_n X) \otimes (X - E_n X)],
\]
\[
\hat{\Sigma}_{\mathcal{X}\mathcal{Y}} = E_n[(X - E_n X) \otimes \kappa_Y (\cdot, Y)],
\]
\[
\hat{M}_{\mathcal{Y}\mathcal{Y}} = E_n[\kappa_Y (\cdot, Y) \otimes \kappa_Y (\cdot, Y)],
\]

and estimate \( \Sigma_{\mathcal{X}\mathcal{X}}^{1/2} \) and \( M_{\mathcal{Y}\mathcal{Y}}^{1/2} \) by the regularized inverses

\[
(\hat{\Sigma}_{\mathcal{X}\mathcal{X}} + \epsilon_n^{(X)} I)^{-1/2}, \quad (\hat{M}_{\mathcal{Y}\mathcal{Y}} + \epsilon_n^{(Y)} I)^{-1}.
\]

We then compute the first \( d \) eigenfunctions \( \hat{f}_1, \ldots, \hat{f}_d \) of

\[
\hat{\Lambda} = (\hat{\Sigma}_{\mathcal{X}\mathcal{X}} + \epsilon_n^{(X)} I)^{-1/2} \hat{\Sigma}_{\mathcal{X}\mathcal{Y}} (\hat{M}_{\mathcal{Y}\mathcal{Y}} + \epsilon_n^{(Y)} I)^{-1} \hat{\Sigma}_{\mathcal{Y}\mathcal{X}} (\hat{\Sigma}_{\mathcal{X}\mathcal{X}} + \epsilon_n^{(X)} I)^{-1/2}
\]

and use the inner products

\[
\langle (\hat{\Sigma}_{\mathcal{X}\mathcal{X}} + \epsilon_n^{(X)} I)^{-1/2} \hat{f}_i, X \rangle_{\mathcal{H}_{\mathcal{X}}}, \quad i = 1, \ldots, d,
\]

as the estimate of a basis of the central subspace. In the online Supplementary Material, we will give more details of the numerical procedures using coordinate representation. But for asymptotic development it is sufficient –
and indeed beneficial – to stay at the operator level. Let \( \Lambda \) and \( f_1, \ldots, f_d \) be the population-level counterparts of \( \hat{\Lambda}, \hat{f}_1, \ldots, \hat{f}_d \). That is,
\[
\Lambda = \sum_{\Sigma_{XX}} \Sigma_{XY} M_{YY} \Sigma_{YX} \Sigma_{XX}^\dagger,
\]
and \( f_1, \ldots, f_d \) are the first \( d \) eigenvectors of \( \Lambda \).

To streamline the presentation, we collect some assumptions into the next assumption.

**Assumption 11.**

1. Assumptions 3(a), 4, 7, and 8 are satisfied;
2. there is \( \beta > 0 \) and a bounded linear operator \( D_{XY} : \mathcal{M}_Y \to \mathcal{H}_X \) such that
   \[
   \Sigma_{XY} = \Sigma_{XX}^{1+\beta} D_{XY} M_{YY}^{1+\beta};
   \]
3. each of the \( d \) nonzero eigenvalues of \( \Lambda \) has multiplicity 1;
4. \( \epsilon_n^{(X)} = c \epsilon_n^{(Y)} \) for some constant \( c > 0 \); thus we can assume, without loss of generality, \( \epsilon_n^{(X)} = \epsilon_n^{(Y)} = \epsilon_n \) for notation simplicity.

Among these assumptions, Assumptions 3(a) and 4 guarantee existence of certain moments of the random function \( X \) or the kernel \( \kappa_Y(Y,Y) \). Assumptions 7 and 8 are made so that \( \Sigma_{XX} \) and \( M_{YY} \) are injective. As discussed in Li [24], condition (8.1) is to guarantee a degree of collective smoothness of the classes of functions sent out by the operators \( \Sigma_{XY} \) and \( \Sigma_{YX} \). For example, one implication of this condition is \( \text{ran}(\Sigma_{XY}) \subseteq \text{ran}(\Sigma_{XX}^{1+\beta}) \). It is easy to see that
\[
\text{ran}(\Sigma_{XX}^{1+\beta}) = \left\{ f \in \mathcal{H}_X : \sum_{i=1}^{\infty} \frac{(f, f_i)^2_{\mathcal{H}_X}}{\lambda_i^{2(1+\beta)}} < \infty \right\},
\]
where \( (\lambda_1, f_1), (\lambda_2, f_2), \ldots \) are the eigenvalue-eigenfunction pairs of \( \Sigma_{XX} \). Because \( \Sigma_{XX} \) is a Hilbert-Schmidt operator (Fukumizu, Bach and Gretton [13]), \( \sum_{i=1}^{\infty} \lambda_i^2 < \infty \), which implies \( \lambda_i \to 0 \). The inequality in (8.2) means that, as \( \beta \) increases, a member \( f \) of \( \text{ran}(\Sigma_{XX}^{1+\beta}) \) has its spectrum increasingly concentrated on the lower-frequency region. Thus, the assumption \( \text{ran}(\Sigma_{XY}) \subseteq \text{ran}(\Sigma_{XX}^{1+\beta}) \) amounts to requiring the output of the operator \( \Sigma_{XY} \) are smooth functions in \( \mathcal{H}_X \). In full generality, we can require \( \Sigma_{XY} = \Sigma_{XX}^{1+\alpha} D_{XY} M_{YY}^{1+\beta} \) for different positive constants \( \alpha \) and \( \beta \), which means the outputs of \( \Sigma_{XY} \) and \( \Sigma_{YX} \) have different degrees of smoothness, but we take \( \alpha = \beta \) to keep things simple. We will refer to \( \beta \) as the smoothness index.
The assumption that nonzero eigenvalues of $\Lambda$ have multiplicity 1 is a technical condition to make the proofs simple, because under this assumption the eigenvectors corresponding to nonzero eigenvalues are uniquely defined (up to a sign). This can be relaxed if we work with eigen spaces. Finally, the fourth item of Assumption 11 is made correspondingly to item 2 which assumes the outputs of $\Sigma_{XY}$ and $\Sigma_{YX}$ have the same degree of smoothness, and can be relaxed if we allow different degrees of smoothness in item 2. Moreover, with a more complicated proof, we can also relax $\epsilon_n^{(X)} = c\epsilon_n^{(Y)}$ to

$$c_1 \leq \liminf_n (\epsilon_n^{(X)}/\epsilon_n^{(Y)}) \leq \limsup_n (\epsilon_n^{(X)}/\epsilon_n^{(Y)}) \leq c_2$$

for some $0 < c_1 \leq c_2 < \infty$.

The next theorem gives the consistency and convergence rates of $\hat{\Lambda}$.

**Theorem 7.** If Assumption 11 holds, $\beta > 1$, and $n^{-\frac{1}{2}} < \epsilon_n < 1$, then $\hat{\Lambda}$, is a consistent estimate of $\Lambda$, and

$$\|\hat{\Lambda} - \Lambda\| = O_P(n^{-1}\epsilon_n^{-2} + n^{-\frac{1}{2}}\epsilon_n^{-\frac{1}{2}} + \epsilon_n).$$

An immediate consequence of this theorem is that the first $d$ eigenvalues and eigenfunctions of $\hat{\Lambda}$ converge to those of $\Lambda$ at the same rates, as stated in the next corollary.

**Corollary 2.** Under the assumptions of Theorem 7,

$$\bar{\lambda}_i - \lambda_i = O_P(n^{-1}\epsilon_n^{-2} + n^{-\frac{1}{2}}\epsilon_n^{-\frac{1}{2}} + \epsilon_n),$$

$$\|\hat{f}_i - f_i\|_{\mathcal{H}_X} = O_P(n^{-1}\epsilon_n^{-2} + n^{-\frac{1}{2}}\epsilon_n^{-\frac{1}{2}} + \epsilon_n).$$

After obtaining the eigenfunction estimates $\hat{f}_1, \ldots, \hat{f}_d$, we transform them by $(\hat{\Sigma}_{XX} + \epsilon_n I)^{-\frac{1}{2}}$ to estimate the central subspace. We next derive the consistency and convergence rates of transformed eigenfunctions $\hat{h}_i = (\hat{\Sigma}_{XX} + \epsilon_n I)^{-\frac{1}{2}}\hat{f}_i$, $i = 1, \ldots, d$, as well as the sufficient predictors $\hat{\eta}_i(X) = \langle \hat{h}_i, X \rangle_{\mathcal{H}_X}$. Let $h_i = \Sigma_{XX}^{1/2}f_i$ and $\eta_i(X) = \langle h_i, X \rangle_{\mathcal{H}_X}$. Note that, even though $\Sigma_{XX}^{1/2}$ is an unbounded linear operator, because $f_i = \lambda_i^{-1}\Sigma_{XX}f_i$, $h_i = \Sigma_{XX}^{1/2}f_i = \lambda_i^{-1}\Sigma_{XX}^{1/2}f_i$ is a well defined member of $\mathcal{H}_X$.

**Theorem 8.** If Assumption 11 holds, $\beta > 1$, and $n^{-2/5} < \epsilon_n < 1$, then, for $i = 1, \ldots, d$,

$$\|\hat{h}_i - h_i\|_{\mathcal{H}_X} = O_P(n^{-1}\epsilon_n^{-\frac{1}{2}} + n^{-\frac{1}{2}}\epsilon_n^{-\frac{1}{2}} + \epsilon_{n_{\min(1,\beta-1)}}),$$

$$\hat{\eta}_i(X) - \eta_i(X) = O_P(n^{-1}\epsilon_n^{-\frac{1}{2}} + n^{-\frac{1}{2}}\epsilon_n^{-\frac{1}{2}} + \epsilon_{n_{\min(1,\beta-1)}}).$$

Consequently, for such $\beta$ and $\epsilon_n$, $\hat{h}_i$ and $\hat{\eta}_i(X)$ are consistent estimates of $h_i$ and $\eta_i(X)$.
9. Optimal rate for tuning constant. Using the asymptotic results of the last section we can easily derive the optimal rate at which tuning constant $\epsilon_n$ converges to 0. Assuming that $\epsilon_n$ is of the form $cn^{-t}$ for some constants $c > 0$, $t > 0$. Our goal is to find the $t$ that makes the rate in
Theorem 8,
\[
\rho(cn^t, \beta) = n^{-\frac{5}{2} + \frac{2}{\beta - 1}} + n^{-\frac{1}{2} + t} + n^{-\min(1, \beta - 1)t},
\]
converges to 0 the fastest. We will also derive at what degree of smoothness $\beta$ the quantity $\min_{t > 0} \rho(cn^t, \beta)$ reaches its fastest possible level when the optimal tuning constant is used. In the following, if $a_n$ and $b_n$ are positive sequences such that $a_n/b_n$ is a bounded sequence, then we write $a_n \preceq b_n$.

**Theorem 9.** Suppose Assumption 11 holds and $\beta > 1$. Let
\[
\epsilon_n^*(\beta) = \begin{cases} 
  n^{-2\min(1, \beta - 1)}/(\beta - 2) & \beta < 3/2 \\
  n^{-2(\beta - 1)/(\beta - 1) - 1} & \beta \geq 3/2
\end{cases}
\]

Then $\rho(\epsilon_n^*(\beta), \beta) \preceq \rho(cn^t, \beta)$ for any $t > , c > 0$, and $\beta > 1$. Furthermore the convergence rate corresponding to the optimal tuning constant $\epsilon_n^*(\beta)$ is

\[
\rho(\epsilon_n^*(\beta), \beta) = \begin{cases} 
  n^{-\frac{2\min(1, \beta - 1)}{\beta - 2}} & \beta < 3/2 \\
  n^{-\frac{\min(1, \beta - 1)}{2(\beta - 1)}} & \beta \geq 3/2
\end{cases}
\]

It can be shown that, for each $n$, $\rho(\epsilon_n^*(\beta), \beta)$ is a monotone nonincreasing function of $\beta$. There are two “phase transition” values of the smoothness index $\beta$: before it reaches $3/2$, $\log[\rho(\epsilon_n^*(\beta), \beta)]$ is $-\frac{2\min(1, \beta - 1)}{\beta - 2}$ (ignoring the proportionality constant $\log(n)$); after it reaches $3/2$, $\log[\rho(\epsilon_n^*(\beta), \beta)]$ is $-\frac{\min(1, \beta - 1)}{2(\beta - 1)}$. Furthermore, because $\min(1, \beta - 1) = 1$ if $\beta \geq 2$, $\rho(\epsilon_n^*(\beta), \beta)$ reaches is minimum $n^{-\frac{2}{3}}$ after $\beta$ reaches 2. Figure 1 shows the shape of the function $\log[\rho(\epsilon_n^*(\beta), \beta)]$. 
The next corollary records the fastest possible rate of \( \rho(\epsilon_n^*(\beta), \beta) \) as well as the optimal tuning constant to achieve it. Of course, it is possible that the rate \( \rho(\epsilon_n^*(\beta), \beta) \) itself may be improved, which may lead to improvement of the \( n^{-\frac{2}{7}} \) bound.

**Corollary 3.** Under Assumption 11, if \( \beta \geq 2 \), then the convergence rate \( \rho(\epsilon_n^*(\beta), \beta) \) reaches its fastest level \( O_p(n^{-\frac{2}{7}}) \), and the optimal tuning constant to achieve this rate is \( \epsilon_n^* \propto n^{-\frac{2}{7}} \).

10. **Simulation studies.** In this section we evaluate the performances of WIRE, WAVE, and WDR under four simulated models, which are the special cases of the functional SDR model \( Y \mid X \mid T(X) \), where \( X \) is a random function or a vector of random functions in a Hilbert space \( \mathcal{H}_X \), and \( T(X) \) is defined by the linear mapping

\[
X \mapsto (\langle f_1, X \rangle_{\mathcal{H}_X}, \ldots, \langle f_d, X \rangle_{\mathcal{H}_X})
\]

with \( f_1, \ldots, f_d \) being nonrandom members of \( \mathcal{H}_X \). The \( d \)-dimensional sufficient predictor \( (\langle f_1, X \rangle_{\mathcal{H}_X}, \ldots, \langle f_d, X \rangle_{\mathcal{H}_X}) \) appears in the conditional distribution of \( Y \) given \( X \) either through the conditional mean \( E(Y|X) \) or the conditional variance \( E(Y|X) \). Due to limited space we only include the simulations for function-on-function regressions in this manuscript. Additional simulation studies for scalar-on-function regressions and for multivariate functional regression are presented in sections 6.1 and 6.3 of Supplementary Material.

To provide a comprehensive study we include the following four models with monotone trend, symmetric trend, both monotone and symmetric...
trends, and heteroscedasticity:

Model 1: \[ Y(t) = 0.2 \exp(1 + \langle \beta_1, X \rangle) v_1(t) + \sigma \epsilon(t), \]
Model 2: \[ Y(t) = (3 \langle \beta_1, X \rangle)^2 v_1(t) + \sigma \epsilon(t), \]
Model 3: \[ Y(t) = 3 \langle \beta_1, X \rangle v_1(t) + (9 \langle \beta_2, X \rangle)^2 v_2(t) + \sigma \epsilon(t), \]
Model 4: \[ Y(t) = 5 \langle \beta_1, X \rangle v_1(t) + 8 \langle \beta_2, X \rangle \sigma \epsilon(t), \]

where \( X \) is the standard Brownian motion on \([0, 1]\), \( \beta_1(t) = \sin(3/2\pi t) \), \( \beta_2(t) = \sin(5/2\pi t) \), \( \sigma = 0.1 \), \( v_1(t) = \sin(3/2\pi t) \), \( v_2(t) = \sin(5/2\pi t) \), and \( \epsilon(t) \) is generated from the standard Brownian motion independent of \( X(t) \). Model 4 is a variation of the prototype (3.2). Each \( X_i \) is observed at 11 time points equally spaced in \([0, 1]\). For \( n = 100 \), we generate 100 samples of sizes 200 from each model, which are then divided equally as training and test sets. Similarly, for \( n = 200 \), we generate 100 samples of size 400, which are then divided equally into training and test sets.

We apply our three functional SDR methods to these models. Since we are aware of no previous linear SDR methods for function-on-function regression, we compare our methods with the nonlinear function-on-function SDR methods, generalized sliced inverse regression (f-GSIR), proposed by Li and Song [27].

We use the multiple correlation between the true and estimated predictors to evaluate the performance of each method. Specifically, let \( U \) and \( V \) be random vectors of the same dimension \( d \). Let \( C_{UV}, C_{UV}, \) and \( C_{VV} \) represent the sample covariance matrices. The multiple correlation between \( U \) and \( V \) is

\[
mcorr(U, V) = \text{tr}(C_{VV}^{-1/2} C_{UV} C_{UV}^{-1} C_{VV}^{-1/2}).
\]

This number ranges between 0 and \( d \), with larger number indicating stronger linear dependence. We applied WIRE, WAVE, and WDR to the four models. The kernels \( \kappa_T \) and \( \kappa_X \) are chosen to be the Gaussian radial basis function. The tuning constants \( \epsilon_T, \epsilon_X, \) and \( \epsilon_Y \) are chosen by the GCV criteria in Section 5. The constants \( \gamma_T, \gamma_Y \) are chosen by (S.65) in the Supplementary Material.

In Table 1 we present the averages and standard errors (in parentheses) of the multiple correlations between the true and the estimated predictors based on the testing sets. We see that, for these models, WAVE and WDR perform overall better than WIRE. It is also interesting to note that WAVE and WDR performs very similarly to WIRE under the monotone model. It is not surprising that f-GSIR, a nonlinear SDR method, does not perform as well as WIRE, WAVE, and WDR, because it does not take advantage of the linear index structure.
Table 1. Comparison of the functional SDR methods

<table>
<thead>
<tr>
<th>model</th>
<th>n</th>
<th>method</th>
<th>WIRE</th>
<th>WAVE</th>
<th>WDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>f-GSIR</td>
<td>0.953 (0.017)</td>
<td>0.982 (0.006)</td>
<td>0.966 (0.046)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td>0.963 (0.009)</td>
<td>0.986 (0.003)</td>
<td>0.979 (0.006)</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>f-GSIR</td>
<td>0.436 (0.263)</td>
<td>0.641 (0.283)</td>
<td>0.963 (0.041)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td>0.397 (0.245)</td>
<td>0.659 (0.271)</td>
<td>0.974 (0.011)</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>f-GSIR</td>
<td>1.098 (0.173)</td>
<td>1.538 (0.252)</td>
<td>1.772 (0.110)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td>1.091 (0.160)</td>
<td>1.448 (0.277)</td>
<td>1.807 (0.071)</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>f-GSIR</td>
<td>1.032 (0.137)</td>
<td>1.681 (0.241)</td>
<td>1.595 (0.254)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td>1.018 (0.113)</td>
<td>1.747 (0.175)</td>
<td>1.739 (0.130)</td>
</tr>
</tbody>
</table>

Note that the numbers in the parentheses are the standard errors of the mcorr(U,V) based on 100 simulated samples. Thus, the standard error of the averaged mcorr should be estimated by the number in the parentheses divided by 10. With this in mind the advantages of the proposed methods over the existing methods are quite significant in most cases.

To further investigate the sensitivity of our methods to the choice of the tuning parameters $\epsilon_X$ and $\epsilon_Y$, we run WIRE, WAVE, and WDR on the simulated data generated from Models 1 through 4, but with the alternative choice of $\epsilon_X = \epsilon_Y = n^{-1/4}$ without tuning by GCV. The results are given in Table 2. Overall, the procedures are quite stable under the change of tuning parameters.

Table 2. Performance of the methods without tuning for function-on-function model.

<table>
<thead>
<tr>
<th>model</th>
<th>n</th>
<th>method</th>
<th>WIRE</th>
<th>WAVE</th>
<th>WDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>WIRE</td>
<td>0.933 (0.083)</td>
<td>0.802 (0.247)</td>
<td>0.841 (0.211)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td>0.962 (0.017)</td>
<td>0.945 (0.052)</td>
<td>0.952 (0.052)</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>WIRE</td>
<td>0.515 (0.287)</td>
<td>0.917 (0.136)</td>
<td>0.919 (0.13)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td>0.574 (0.287)</td>
<td>0.955 (0.049)</td>
<td>0.955 (0.045)</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>WIRE</td>
<td>1.483 (0.303)</td>
<td>1.136 (0.288)</td>
<td>1.147 (0.289)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td>1.552 (0.283)</td>
<td>1.242 (0.327)</td>
<td>1.253 (0.334)</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>WIRE</td>
<td>1.521 (0.3)</td>
<td>1.442 (0.309)</td>
<td>1.487 (0.307)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td>1.57 (0.276)</td>
<td>1.646 (0.232)</td>
<td>1.686 (0.218)</td>
</tr>
</tbody>
</table>
11. Applications. In this section we apply WIRE, WAVE, and WDR to a bike sharing data set, where both the response and the predictor are functional data. Three additional applications, where the responses are categorical data and the predictors are one- or multi-dimensional functional data, are presented in section 7 in the Supplementary Material. Throughout this section and section 7 in the Supplementary Material, we choose the kernel $\kappa_T$ to be the Brownian-motion kernel, and the kernels $\kappa_X$ and $\kappa_Y$ to be the Gaussian radial basis kernel.

The bike sharing data were collected from the Capital Bike Share system in Washington D.C., which records, among other things, the hourly bike rental counts from January 1, 2011 to December 31, 2012. Fanaee-T and Gama [9] aggregated the data with the temperature data over the same period. Figure 2 shows the curves of temperature (left panel) and the curves of bike rental count per day (right panel), colored by the average temperature of the day, where the bike rental counts are transformed to log(count+1). Clearly, the rental curves are dependent on the temperature curves—a relation that could be useful for the optimal arrangement of the bikes. This is a function-on-function dimension reduction problem with the rental curves as the response and the temperature curves as the predictor. We first apply WIRE, WAVE, WDR, and functional PCA (FPCA) to the data, and then apply regression to the sufficient predictors to predict rental curves.

Since the bike rental behavior is highly affected by whether the day is a working day or a nonworking day, we only consider the nonworking days (weekends or holidays). The data for 2011 contains 100 nonworking days, which we use as the training set, and that for 2012 contains 112 nonworking, which we use as the test set. For day $i$, let $Y_i(t)$ represent the (transformed) rental count, and $X_i(t)$ the temperature, at time $t$. These functions are observed at equally-spaced time points $t_1, \ldots, t_{24}$, representing the hours of a day. We estimate the whole functions $X_i$ and $Y_i$ based on $\{(X_i(t_j), Y_i(t_j)) :
FUNCTIONAL DIMENSION REDUCTION

\[ j = 1, \ldots, 24 \} \) using the B-spline. To avoid complication, we still use the symbols \( X_i \) and \( Y_i \) to denote the estimated curves.

For the dimension reduction step, we use SDR to obtain the sufficient predictors \( \{(U_{1i}, \ldots, U_{di}) : i = 1, \ldots, n \} \). In the case of FPCA, \( \{(U_{1i}, \ldots, U_{di}) : i = 1, \ldots, n \} \) are the first \( d \) functional principal components. For the regression step, we use the kernel regression method

\[
\hat{y}(u)(t) = \frac{\sum_{i=1}^{n} K(\|u_i - u\|) y_i(t)}{\sum_{i=1}^{n} K(\|u_i - u\|)},
\]

(11.1)

where, on the left-hand side, \( u \) is a vector in \( \mathbb{R}^d \), \( \hat{y}(u) \) is the predicted response curve at \( U = u \), and \( \hat{y}(u)(t) \) is \( \hat{y}(u) \) evaluated at time \( t \); on the right-hand side, \( K(\cdot) \) is the Gaussian kernel, \( u_i = (u_{1i}, \ldots, u_{di})^T \) is the sufficient predictor, and \( \| \cdot \| \) is the \( \mathbb{R}^d \)-norm.

Table 3. Training MSE comparisons for the bike-share dataset.

<table>
<thead>
<tr>
<th>data</th>
<th>dim. red.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR</td>
<td>FPCA</td>
<td>0.2</td>
<td>0.19</td>
<td>0.18</td>
<td>0.18</td>
<td>0.18</td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>WIRE</td>
<td>0.27</td>
<td>0.26</td>
<td>0.18</td>
<td>0.18</td>
<td>0.12</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>WAVE</td>
<td>0.25</td>
<td>0.19</td>
<td>0.13</td>
<td>0.11</td>
<td>0.1</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>WDR</td>
<td>0.25</td>
<td>0.25</td>
<td>0.13</td>
<td>0.11</td>
<td>0.1</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>TE</td>
<td>FPCA</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
<td>0.48</td>
<td>0.48</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>WIRE</td>
<td>0.37</td>
<td>0.37</td>
<td>0.34</td>
<td>0.36</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>WAVE</td>
<td>0.35</td>
<td>0.35</td>
<td>0.37</td>
<td>0.34</td>
<td>0.32</td>
<td>0.32</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>WDR</td>
<td>0.35</td>
<td>0.36</td>
<td>0.35</td>
<td>0.34</td>
<td>0.33</td>
<td>0.32</td>
<td>0.31</td>
</tr>
</tbody>
</table>

After fitting regression models, we compute the mean squared error of the predictions

\[
\text{MSE} = (24n)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{24}[y_i(t_j) - \hat{y}(u_i)(t_j)]^2
\]

based on the test set. Table 3 shows the results, where the data column indicates the training (TR) and test (TE) data, and the dim. red. column indicates different dimension reduction methods. Overall, WDR or WAVE works the best among different dimension reduction methods. We also applied two existing function-on-function regression methods: one linear, as described in [33], and one nonlinear, which was proposed in [32], as well as the kernel regression without dimension reduction. The training set MSE for the three methods are 0.221, 0.198, and 0.220; the testing set MSE for the three methods are: 0.354, 0.326, and 0.310, respectively.
Acknowledgements. The authors would like to thank three referees and an Associate Editor for their many useful and constructive comments and suggestions, which helped us greatly in improving this work. The research of Bing Li is support in part by the U.S. National Science Foundation grant DMS-1713078.

References.


**Address of Bing Li**
Department of Statistics
The Pennsylvania State University
326 Thomas Building
University Park, PA 16802, USA
E-mail: bxl9@psu.edu

**Address of Jun Song**
Department of Mathematics and Statistics
The University of North Carolina at Charlotte
Fretwell 376
9201 University City Blvd.
Charlotte, NC 28223, USA
E-mail: jsong19@uncc.edu