The Annals of Statistics
2013, Vol. 41, No. 1, 221–249
DOI: 10.1214/12-AOS1071

A GENERAL THEORY FOR NONLINEAR SUFFICIENT DIMENSION REDUCTION: FORMULATION AND ESTIMATION

BY KUANG-YAO LEE, BING LI[1] AND FRANCESCA CHIAROMONTE

Yale University, Pennsylvania State University and Pennsylvania State University

In this paper we introduce a general theory for nonlinear sufficient dimension reduction, and explore its ramifications and scope. This theory subsumes recent work employing reproducing kernel Hilbert spaces, and reveals many parallels between linear and nonlinear sufficient dimension reduction. Using these parallels we analyze the properties of existing methods and develop new ones. We begin by characterizing dimension reduction at the general level of σ-fields and proceed to that of classes of functions, leading to the notions of sufficient, complete and central dimension reduction classes. We show that, when it exists, the complete and sufficient class coincides with the central class, and can be unbiasedly and exhaustively estimated by a generalized sliced inverse regression estimator (GSIR). When completeness does not hold, this estimator captures only part of the central class. However, in these cases we show that a generalized sliced average variance estimator (GSAVE) can capture a larger portion of the class. Both estimators require no numerical optimization because they can be computed by spectral decomposition of linear operators. Finally, we compare our estimators with existing methods by simulation and on actual data sets.

1. Introduction. In this paper we propose a general theory for nonlinear sufficient dimension reduction (SDR), develop novel estimators and investigate their properties under this theory. Along with these developments we also introduce a new conditional variance operator, which can potentially be used to generalize all second-order dimension reduction methods to the nonlinear case.

In its classical form, linear SDR seeks a low-dimensional linear predictor that captures in full a regression relationship. Imagining a regression setting that comprises multiple predictor variables and multiple responses, let X and Y be random vectors of dimension p and q. If there is a matrix $\beta \in \mathbb{R}^{p\times d}$ with $d < p$ such that

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

then the subspace spanned by the columns of β is called a *sufficient dimension reduction* (*SDR*) *subspace*. Under mild conditions, the intersection of all such subspaces still satisfies (1), and is called the *central subspace*, denoted by $\mathcal{S}_{Y|X}$; see

Received June 2012.

[1] Supported in part by NSF Grants DMS-08-06058 and DMS-11-06815.

MSC2010 subject classifications. 62B05, 62G08, 62H30.

Key words and phrases. Dimension reduction σ-field, exhaustivenes, generalized sliced average variance estimator, generalized sliced inverse regression estimator, heteroscedastic conditional covariance operator, sufficient and complete dimension reduction classes, unbiasedness.

Li (1991, 1992), Li and Duan (1989), Duan and Li (1991), Cook and Weisberg (1991), Cook (1994, 1998b). A general condition for the existence of the central subspace is given by Yin, Li and Cook (2008).

Several recent papers have combined sufficient dimension reduction and kernels; see Akaho (2001), Bach and Jordan (2002), Fukumizu, Bach and Gretton (2007), Wu (2008), Wu, Liang and Mukherjee (2008), Hsing and Ren (2009), Yeh, Huang and Lee (2009), Zhu and Li (2011) and Li, Artemiou and Li (2011). This proliferation of work, in addition to producing versatile methods for extracting nonlinear sufficient predictors, points toward a general synthesis between the notions of sufficiency at the core of SDR and the ability to encompass nonlinearity afforded by kernel mappings. To achieve this synthesis, explore its many ramifications and broad scope and develop new estimators based on it, are the goals of this paper.

Specifically, we articulate a general formulation that comprises both linear and nonlinear SDR, and parallels the basic theoretical developments pioneered by Li (1991, 1992) and Cook (1994, 1998a, 1998b). This formulation allows us to study linear and nonlinear SDR comparatively and, somewhat surprisingly, to *relax* some stringent conditions required by linear SDR. For example, a linear conditional mean [Li (1991), Cook (1998b)] is no longer needed for unbiasedness, and the sufficient conditions for existence and uniqueness of the central subspace are far more general and transparent. Finally, our general formulation links linear and nonlinear SDR to the classical notions sufficiency, completeness and minimal sufficiency, which brings insights and great clarity to the SDR theory.

Our developments and the sections of this paper, can be summarized as follows. In Section 2, we build upon the ideas of Cook (2007) and Li, Artemiou and Li (2011) to define an SDR σ-field as a sub σ-field $\mathcal{G}$ of $\sigma(X)$ (the σ-field generated by X) such that $Y \perp\!\!\!\perp X|\mathcal{G}$, and the corresponding SDR class as the set of all square-integrable, $\mathcal{G}$-measurable functions. Under very mild conditions—much milder than the corresponding conditions for linear SDR [Yin, Li and Cook (2008)]—we show that there exists a unique minimal σ-field $\mathcal{G}_{Y|X}$ that satisfies $Y \perp\!\!\!\perp X|\mathcal{G}_{Y|X}$, which we call the *central σ-field*. The set of all $\mathcal{G}_{Y|X}$-measurable, square-integrable functions is named the *central class*.

In Section 3, we provide two additional definitions that generalize concepts in Cook (1998b), Li, Zha and Chiaromonte (2005) and Li, Artemiou and Li (2011): a class of functions is *unbiased* if its members are $\mathcal{G}_{Y|X}$-measurable, and *exhaustive* if they generate $\mathcal{G}_{Y|X}$. Next, we show that the special class

$$L_2(P_X) \ominus \left[L_2(P_X) \ominus L_2(P_Y)\right] \tag{2}$$

is unbiased, where $L_2(P_X)$ and $L_2(P_Y)$ are the spaces of square-integrable functions of X and Y. For reasons detailed in Section 3, we call this class the *regression class*.

In Section 4, we introduce the *complete dimension reduction class*: if $\mathcal{G} \subseteq \sigma(X)$ is a σ-field and for each $\mathcal{G}$-measurable $f \in L_2(P_X)$ we have

$$E\big(f(X)|Y\big) = 0 \qquad \text{almost surely} \quad \Rightarrow \quad f(X) = 0 \qquad \text{almost surely},$$

then we say that the class of $\mathcal{G}$-measurable functions in $L_2(P_X)$ is complete. We prove that when a complete sufficient dimension reduction (CSDR) class exists it is unique and coincides with the central class. We further show that the CSDR class coincides with the regression class—which is therefore not just unbiased, but also exhaustive.

In Section 5 we establish a critical relationship between the regression class and a covariance operator linking X and Y and, based on this, we generalize sliced inverse regression [SIR; Li (1991)] to a method (GSIR) that can recover the regression class—and hence is unbiased and exhaustive under completeness. In Section 6, we consider the case where the central class is not complete, so that GSIR is unbiased but no longer exhaustive. By introducing a novel conditional variance operator, we generalize sliced average variance estimation [SAVE, Cook and Weisberg (1991)] to a method (GSAVE) that can recover a class larger than the regression class. Here, the situation is similar to that in the linear SDR setting, where it is well known that

$$\text{SIR subspace} \subseteq \text{SAVE subspace} \subseteq \mathcal{S}_{Y|X}; \tag{3}$$

see Cook and Critchley (2000), Ye and Weiss (2003), Li, Zha and Chiaromonte (2005) and Li and Wang (2007).

In Section 7 we develop algorithms for the sample versions of GSIR and GSAVE, and a cross-validation algorithm to determine regularizing parameters. In Section 8 we compare GSIR and GSAVE with some existing methods by simulation and on actual data sets. Section 9 contains some concluding remarks. Some highly technical developments are provided in the supplementary material [Lee, Li and Chiaromonte (2013)].

2. Sufficient dimension reduction σ-fields and classes.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\Omega_X, \mathcal{F}_X)$, $(\Omega_Y, \mathcal{F}_Y)$ and $(\Omega_{XY}, \mathcal{F}_{XY})$ be measurable spaces. For convenience, assume that $\Omega_{XY} = \Omega_X \times \Omega_Y$ and $\mathcal{F}_{XY} = \mathcal{F}_X \times \mathcal{F}_Y$. Let X, Y and (X, Y) be random elements that take values in Ω_X, Ω_Y and Ω_{XY}, with distributions P_X, P_Y, P_{XY}, which are dominated by σ-finite measures. Let

$$\sigma(X) = X^{-1}(\mathcal{F}_X), \qquad \sigma(Y) = Y^{-1}(\mathcal{F}_Y), \qquad \sigma(X, Y) = (X, Y)^{-1}(\mathcal{F}_{XY}),$$

and finally let $P_{Y|X}(\cdot|\cdot) : \mathcal{F}_Y \times \Omega_X \to \mathcal{R}$ be the conditional distribution of Y given X.

DEFINITION 1. A sub σ-field $\mathcal{G}$ of $\sigma(X)$ is an SDR σ-field for Y versus X if it satisfies

$$Y \perp\!\!\!\perp X|\mathcal{G}, \tag{4}$$

that is, if Y and X are independent given $\mathcal{G}$.

This definition is sufficiently general to accommodate the two cases of nonlinear sufficient dimension reduction that interest us the most. The first case is when $\Omega_X = \mathbb{R}^p$ and $\Omega_Y = \mathbb{R}^q$ for some positive integers p and q, and $\mathcal{F}_X$, $\mathcal{F}_Y$ and $\mathcal{F}_{XY}$ are Borel σ-fields generated by the open sets in $\mathbb{R}^p$, $\mathbb{R}^q$ and $\mathbb{R}^{p+q}$. Clearly, in this case, the conditional independence (4) is a generalization of (1) for linear SDR: if we take $\mathcal{G} = \sigma(\beta^{\mathsf{T}} X)$, then (4) reduces to (1).

The second case is when X or Y, or both of them, are random functions. In this case Definition 1 is a generalization of the linear SDR for functional data introduced by Ferré and Yao (2003), and Hsing and Ren (2009). Specifically, let $[a, b]$ be a closed interval, λ the Lebesgue measure and $L_2(\lambda)$ the class of functions on $[a, b]$ that are square integrable with respect to λ. Let $\Omega_X = L_2(\lambda)$ and $\Omega_Y = \mathbb{R}$. In this case, each $X(\omega)$ is a function in $L_2(\lambda)$, which, depending on applications, could be, say, a growth curve or the fluctuation of a stock price. Let $h_1, \ldots, h_d$ be functions in $L_2(\lambda)$. Ferré and Yao (2003) considered the following functional dimension reduction problem:

$$Y \perp\!\!\!\perp X | \langle X, h_1 \rangle_{L_2(\lambda)}, \ldots, \langle X, h_d \rangle_{L_2(\lambda)}. \tag{5}$$

This generalizes linear SDR to the infinite-dimensional case, but not to the nonlinear case, because $\langle X, h_1 \rangle_{L_2(\lambda)}, \ldots, \langle X, h_d \rangle_{L_2(\lambda)}$ are linear in X. Hsing and Ren (2009) considered a more general setting where the sample paths $\{X_t(\omega) : t \in J\}$ need not lie within $L_2(\lambda)$. Still, their generalization is inherently linear in the same sense that problem (5) is linear. In contrast, our formulation in (4) allows an arbitrary sub σ-field of $\sigma(X)$, which need not be generated by linear functionals. Interestingly, as we will see Section 5, it is the relaxation of linearity that allows us to remove a restrictive linear conditional mean assumption used both in Ferré and Yao (2003) (Theorem 2.1), and in Hsing and Ren (2009), assumption (IR2).

The notion of sufficiency underlying SDR, as defined by (1) and (4), is different from the classical notion of sufficiency because $\mathcal{G}$ is allowed to depend on any parameter in the joint distribution of P_{XY}. For example, $\mathcal{G} = \sigma(\beta^{\mathsf{T}} X)$ depends on the parameter β [or rather, the meta-parameter span(β)] which characterizes the conditional distribution of $Y|X$. Nevertheless, both notions imply a reduction, or simplification, in the representation of a stochastic mechanism—the SDR one through a newly constructed predictor, and the classical one through a statistic. Indeed, it is partly by exploring and exploiting this similarity that we developed our theory of nonlinear SDR.

Obviously there are many sub σ-fields of X that satisfy (4), starting with $\sigma(X)$ itself—which induces no reduction. For maximal dimension reduction we seek the smallest such σ-field. As in the case of classical sufficiency, the minimal SDR σ-field does not universally exist, but exists under very mild assumptions. The next theorem gives the sufficient condition for the minimal SDR σ-field to uniquely exist. The proof echoes Bahadur (1954), which established the existence of the minimal sufficient σ-field in the classical setting.

THEOREM 1. *Suppose that the family of probability measures* $\{P_{X|Y}(\cdot|y) : y \in \Omega_Y\}$ *is dominated by a* σ*-finite measure. Then there is a unique sub* σ*-field* $\mathcal{G}^*$ *of* $\sigma(X)$ *such that*:

(1) $Y \perp\!\!\!\perp X|\mathcal{G}^*$;
(2) *if* $\mathcal{G}$ *is a sub* σ*-field of* $\sigma(X)$ *such that* $Y \perp\!\!\!\perp X|\mathcal{G}$, *then* $\mathcal{G}^* \subseteq \mathcal{G}$.

PROOF. Let $\Pi_y = P_{X|Y}(\cdot|y)$ and $\mathbb{P} = \{\Pi_y : y \in \Omega_Y\}$. Since $\mathbb{P}$ is dominated by a σ-finite measure, it contains a countable subset $\mathbb{Q} = \{Q_k : k = 1, 2, \ldots\}$ such that $\mathbb{Q} \equiv \mathbb{P}$, where $\equiv$ means two families of measures dominating each other. Let $\{c_k : k = 1, 2, \ldots\}$ be a sequence of positive numbers that sum to 1, and let $Q_0 = \sum_{k=1}^{\infty} c_k Q_k$. Then Q_0 is a probability measure on Ω_X such that $\{Q_0\} \equiv \mathbb{Q} \equiv \mathbb{P}$. Let $\pi_y = d\Pi_y/dQ_0$ and $\mathcal{G}$ be a sub σ-field of $\sigma(X)$. We claim that the following statements are equivalent:

(1) $Y \perp\!\!\!\perp X|\mathcal{G}$;
(2) π_y is essentially measurable with respect to $\mathcal{G}$ for all $y \in \Omega_Y$ modulo Q_0.

PROOF OF 1 ⇒ 2. Let $B \in \mathcal{F}_X$. Then

$$\begin{aligned} E_{Q_0}(\pi_y(X) I_B(X)) &= E_{\Pi_y}(I_B(X)) = E_{\Pi_y}[E_{\Pi_y}(I_B(X)|\mathcal{G})] \\ &= E_{Q_0}[E_{\Pi_y}(I_B(X)|\mathcal{G})\pi_y(X)]. \end{aligned}$$

By 1, $\Pi_y(B|\mathcal{G})$ is the same for all $y \in \Omega_Y$. Hence $\Pi_y(B|\mathcal{G}) = Q_k(B|\mathcal{G})$ for all k, which implies $\Pi_y(B|\mathcal{G}) = Q_0(B|\mathcal{G})$. Hence we can rewrite the right-hand side of the above equalities as

$$E_{Q_0}[E_{Q_0}(I_B(X)|\mathcal{G})\pi_y(X)] = E_{Q_0}[I_B(X) E_{Q_0}(\pi_y(X)|\mathcal{G})].$$

Thus the following equality holds for all $B \in \mathcal{F}_X$:

$$E_{Q_0}(\pi_y(X) I_B(X)) = E_{Q_0}[I_B(X) E_{Q_0}(\pi_y(X)|\mathcal{G})],$$

which implies $\pi_y(X) = E_{Q_0}(\pi_y(X)|\mathcal{G})$ a.s. Q_0.

PROOF OF 2 ⇒ 1. For any $A \in \mathcal{G}$,

$$\begin{aligned} E_{\Pi_y}[E_{Q_0}(I_B(X)|\mathcal{G}) I_A(X)] &= E_{Q_0}[E_{Q_0}(I_B(X)|\mathcal{G}) I_A(X)\pi_y(X)] \\ &= E_{Q_0}[I_B(X) I_A(X) E_{Q_0}(\pi_y(X)|\mathcal{G})]. \end{aligned}$$

By 2, $E_{Q_0}(\pi_y(X)|\mathcal{G}) = \pi_y(X)$. Hence the right-hand side becomes

$$E_{Q_0}[I_B(X) I_A(X)\pi_y(X)] = E_{\Pi_y}[I_B(X) I_A(X)] = \Pi_y(X \in A \cap B).$$

Thus $E_{Q_0}(I_B(X)|\mathcal{G}) = Q_0(B|\mathcal{G})$ is the conditional probability $\Pi_y(B|\mathcal{G})$, which means $\Pi_y(B|\mathcal{G})$ does not depend on y. That is, 1 holds.

Now let $\mathcal{G}^*$ be the intersection of all SDR σ-fields $\mathcal{G}$. Then $\mathcal{G}^*$ is itself a σ-field. Moreover, since π_y is essentially measurable with respect to all SDR σ-fields for all $y \in \Omega_Y$, it is also essentially measurable with respect to $\mathcal{G}^*$ for all $y \in \Omega_Y$.

Consequently, $\mathcal{G}^*$ is itself an SDR σ-field, which implies that it is also the smallest SDR σ-field. If $\mathcal{G}^{**}$ is another smallest SDR σ-field, then we know $\mathcal{G}^* \subseteq \mathcal{G}^{**}$ and $\mathcal{G}^{**} \subseteq \mathcal{G}^*$. Thus $\mathcal{G}^*$ is unique. □

We can now naturally introduce the following definition:

DEFINITION 2. Suppose that the class of probability measures $\{P_{X|Y}(\cdot|y) : y \in \Omega_Y\}$ on Ω_X is dominated by a σ-finite measure. Then we call the σ-field $\mathcal{G}^*$ in Theorem 1 the central σ-field for Y versus X, and denote it by $\mathcal{G}_{Y|X}$.

Notably, this set up characterizes dimension reduction solely in terms of conditional independence. However, explicitly turning to functions and introducing an additional mild assumption of square integrability are very consequential for further development because they allow us to work with structures such as orthogonality and projection.

Let $L_2(P_{XY})$, $L_2(P_X)$ and $L_2(P_Y)$ be the spaces of functions defined on Ω_{XY}, Ω_X and Ω_Y that are square-integrable with respect to P_{XY}, P_X and P_Y, respectively. Since constants are irrelevant for dimension reduction, we assume throughout that all functions in $L_2(P_X)$, $L_2(P_Y)$ and $L_2(P_{XY})$ have mean 0. Given a sub σ-field $\mathcal{G}$ of $\sigma(X, Y)$, we use $\mathfrak{M}_{\mathcal{G}}$ to denote the class of all functions f in $L_2(P_{XY})$ such that $f(X)$ is $\mathcal{G}$-measurable. If $\mathcal{G}$ is generated by a random vector, say X, then we use $\mathfrak{M}_X$ to abbreviate $\mathfrak{M}_{\sigma(X)}$. It can be easily shown that, for any $\mathcal{G}$, $\mathfrak{M}_{\mathcal{G}}$ is a linear subspace of $L_2(P_{XY})$.

DEFINITION 3. Let $\mathcal{G}$ be an SDR σ-field and $\mathcal{G}_{Y|X}$ be the central σ-field. Then $\mathfrak{M}_{\mathcal{G}}$ is called an SDR class, and $\mathfrak{M}_{\mathcal{G}_{Y|X}}$ is called the central class. The latter class is denoted by $\mathfrak{S}_{Y|X}$.

The central class, comprising square-integrable functions that are measurable with respect to the central σ-field $\mathcal{G}_{Y|X}$, represents our generalization of the central space $\mathcal{S}_{Y|X}$ defined in linear SDR; see the Introduction.

3. Unbiasedness and exhaustiveness. In linear SDR, the goal is to find a set of vectors that span $\mathcal{S}_{Y|X}$. If a matrix γ satisfies $\text{span}(\gamma) \subseteq \mathcal{S}_{Y|X}$, we say that γ is unbiased [Cook (1998b)]. If $\text{span}(\gamma) = \mathcal{S}_{Y|X}$, we say that γ is exhaustive [Li, Zha and Chiaromonte (2005)]. Note that when $\text{span}(\gamma) \subseteq \mathcal{S}_{Y|X}$, $\gamma^{\mathsf{T}}X$ is a linear function of $\beta^{\mathsf{T}}X$, where β is any matrix such that $\text{span}(\beta) = \mathcal{S}_{Y|X}$; if $\text{span}(\gamma) = \mathcal{S}_{Y|X}$, then $\gamma^{\mathsf{T}}X$ is an injective linear transformation of $\beta^{\mathsf{T}}X$. In the nonlinear setting, we follow the same logic but remove the linear requirement. Part of the following definition was given in Li, Artemiou and Li (2011).

DEFINITION 4. A class of functions in $L_2(P_X)$ is unbiased for $\mathfrak{S}_{Y|X}$ if its members are $\mathcal{G}_{Y|X}$-measurable, and exhaustive for $\mathfrak{S}_{Y|X}$ if its members generate $\mathcal{G}_{Y|X}$.

Next, we look into what type of functions are unbiased. The lemma below provides a characterization of the orthogonal complement of $\mathfrak{M}_{\mathcal{G}}$ that will be used many times in the subsequent development. Its proof is essentially the definition of the conditional expectation, and is omitted.

LEMMA 1. *Suppose U is a random element defined on $(\Omega, \mathcal{F})$, $\mathcal{G}$ is a sub σ-field of $\sigma(U)$ and $f \in L_2(P_U)$. Then f is orthogonal to $\mathfrak{M}_{\mathcal{G}}$ $(f \perp \mathfrak{M}_{\mathcal{G}})$ if and only if $E[f(U)|\mathcal{G}] = 0$.*

Note that $\perp\!\!\!\perp$ and $\perp$ have different meanings: the former means independence; the latter means orthogonality. For two subspaces, say $\mathcal{S}_1$ and $\mathcal{S}_2$, of a generic Hilbert space $\mathcal{H}$, we use $\mathcal{S}_1 \ominus \mathcal{S}_2$ to denote the subspace $\mathcal{S}_1 \cap \mathcal{S}_2^{\perp}$. The following theorem explicitly specifies a class of functions, which we call *regression class*, that is unbiased for $\mathfrak{S}_{Y|X}$.

THEOREM 2. *If the family $\{\Pi_y : y \in \Omega_Y\}$ is dominated by a σ-finite measure, then*

$$L_2(P_X) \ominus \big[L_2(P_X) \ominus L_2(P_Y)\big] \subseteq \mathfrak{S}_{Y|X}. \tag{6}$$

PROOF. It is equivalent to show that $L_2(P_X) \ominus \mathfrak{S}_{Y|X} \subseteq L_2(P_X) \ominus L_2(P_Y)$. If $f \in L_2(P_X) \ominus \mathfrak{S}_{Y|X}$, then, by Lemma 1, $E[f(X)|\mathcal{G}_{Y|X}] = 0$. Since $\mathcal{G}_{Y|X}$ is a sufficient σ-field,

$$E\big[f(X)|Y\big] = E\big[E\big(f(X)|Y, \mathcal{G}_{Y|X}\big)|Y\big] = E\big[E\big(f(X)|\mathcal{G}_{Y|X}\big)|Y\big] = 0.$$

By Lemma 1 again, $f \perp \mathfrak{M}_Y$. Because $\mathfrak{M}_Y = L_2(P_Y)$, we have $f \in L_2(P_X) \ominus L_2(P_Y)$. □

The intuition behind the term "regression class" is that $L_2(P_X) \ominus L_2(P_Y)$ resembles the residual in a regression problem; thus $L_2(P_X) \ominus [L_2(P_X) \ominus L_2(P_Y)]$ is simply the orthogonal complement of the "residual class." Henceforth we write the regression class as $\mathfrak{C}_{Y|X}$.

4. Complete and sufficient dimension reduction classes. After showing that the regression class (2) is unbiased, we investigate under what conditions it is also exhaustive for the central class $\mathfrak{S}_{Y|X}$. To this end we need to introduce the notion of complete classes of functions in $L_2(P_X)$.

DEFINITION 5. Let $\mathcal{G} \subseteq \sigma(X)$ be a sub σ-field. The class $\mathfrak{M}_{\mathcal{G}}$ is said to be complete if, for any $g \in \mathfrak{M}_{\mathcal{G}}$,

$$E\big[g(X)|Y\big] = 0 \quad \text{a.s. } P \quad \Rightarrow \quad g(X) = 0 \quad \text{a.s. } P.$$

Again there are similarities and differences between completeness as defined here and in the classical setting. A complete and sufficient statistic in the classical setting is a rather restrictive concept, often associated with exponential families, the uniform distribution, or the order statistics. In contrast, completeness here is a rather general concept. To demonstrate this point, in the next two propositions we give two examples of complete and sufficient dimension reduction classes. In particular, the first shows that if Y is related to X through *any* regression model with additive error, then the subspace of $L_2(P_X)$ determined by the regression function is a complete and sufficient dimension reduction class. In the following, $[L_2(P_X)]^q$ denotes the q-fold Cartesian product of $L_2(P_X)$.

PROPOSITION 1. *Suppose there exists a function $h \in [L_2(P_X)]^q$ such that*

$$Y = h(X) + \varepsilon, \tag{7}$$

where $\varepsilon \perp\!\!\!\perp X$ and $E(\varepsilon) = 0$. Then $\mathfrak{M}_{h(X)}$ is a complete and sufficient dimension reduction class for Y versus X.

Note that, since $L_2(P_X)$ is centered, we have implicitly assumed that $E[h(X)] = 0$ [and hence $E(Y) = 0$]. However, this does not entail any real loss of generality because the proof below can be easily modified for the case where $L_2(P_X)$ is not centered.

PROOF OF PROPOSITION 1. Suppose $m \in \mathfrak{M}_{h(X)}$ and $E[m(X)|Y] = 0$ a.s. P. Then there is a measurable function $g : \mathbb{R}^q \to \mathbb{R}$ such that $m = g \circ h$. Let $U = h(X)$. Then $E(g(U)|Y) = 0$ a.s. P. By Lemma 1, for any $f \in L_2(P_Y)$, we have $E[g(U)f(Y)] = 0$. In particular, $E[g(U)e^{it^\mathsf{T}Y}] = 0$. Because $U \perp\!\!\!\perp \varepsilon$, this implies

$$E[g(U)e^{it^\mathsf{T}U}]E(e^{it^\mathsf{T}\varepsilon}) = E[g(U)e^{it^\mathsf{T}U}e^{it^\mathsf{T}\varepsilon}] = E[g(U)e^{it^\mathsf{T}Y}] = 0.$$

Hence $E[g(U)e^{it^\mathsf{T}U}] = 0$. By the uniqueness of inverse Fourier transformation we see that $g(U) = 0$ a.s. P, which implies $m(X) = (g \circ h)(X) = 0$ a.s. P. □

The expression in (7) covers many useful models in statistics and econometrics. For example, any homoscedastic parametric or nonparametric regression, such as the single index and the multiple index models [Ichimura and Lee (1991), Härdle, Hall and Ichimura (1993), Yin, Li and Cook (2008)], are special cases of (7). Thus, complete and sufficient dimension reduction classes exist for all those settings. The next proposition considers a type of inverse regression model, in which X is transformed into two components, one of which is related to Y by an inverse linear regression model, and the other independent of the rest of the data.

PROPOSITION 2. *Suppose $q < p$, Ω_Y has a nonempty interior, and P_Y is dominated by the Lebesgue measure on $\mathbb{R}^q$. Suppose there exist functions $g \in [L_2(P_X)]^q$ and $h \in [L_2(P_X)]^{p-q}$ such that*:

(1) *$g(X) = Y + \varepsilon$, where $Y \perp\!\!\!\perp \varepsilon$, and $\varepsilon \sim N(0, \Sigma)$;*

(2) *$\sigma(g(X), h(X)) = \sigma(X)$;*

(3) *$h(X) \perp\!\!\!\perp (Y, g(X))$;*

(4) *the induced measure $P_X \circ g^{-1}$ is dominated by the Lebesgue measure on $\mathbb{R}^q$.*

Then $\mathfrak{M}_{g(X)}$ is a complete sufficient dimension reduction class for Y versus X.

PROOF. Assumption 3 implies $Y \perp\!\!\!\perp h(X)|g(X)$, which, by assumption 2, implies $Y \perp\!\!\!\perp X|g(X)$. That is, $\mathfrak{M}_{g(X)}$ is an SDR class. Let $u \in \mathfrak{M}_{g(X)}$. Then $u = v \circ g$ for some measurable function $v : \mathbb{R}^q \to \mathbb{R}$. Let $U = g(X)$. Suppose that $E[v(U)|Y] = 0$ almost surely P. Because $Y \perp\!\!\!\perp \varepsilon$, this implies $P_Y(\{y : Ev(y+\varepsilon) = 0\}) = 1$. In other words,

$$\int_{\mathbb{R}^q} v(t) \frac{1}{(2\pi)^{q/2}|\Sigma|^{1/2}} e^{-(t-y)^\mathsf{T}\Sigma^{-1}(t-y)/2}\, dt = 0$$

a.s. P_Y. This implies

$$\int v(t) e^{-t^\mathsf{T}\Sigma^{-1}t/2} e^{y^\mathsf{T}\Sigma^{-1}t}\, dt = 0 \quad \Rightarrow \quad \int v(\Sigma s) e^{-s^\mathsf{T}\Sigma s/2} e^{y^\mathsf{T}s}\, ds = 0$$

a.s. P_Y, where $s = \Sigma^{-1}t$. Because Ω_Y contains an open set in $\mathbb{R}^q$ and the above function of y is analytic, by the analytic continuation theorem, the above function is 0 everywhere on $\mathbb{R}^q$. Hence, by the uniqueness of inverse Laplace transformation, we have

$$v(\Sigma s) e^{-s^\mathsf{T}\Sigma s/2} = 0 \qquad \text{almost surely } \lambda,$$

where λ is the Lebesgue measure on $\mathbb{R}^q$. But, because $e^{-s^\mathsf{T}\Sigma s/2} > 0$, we have $v(\Sigma s) = 0$ a.s. λ or equivalently $v(t) = 0$ a.s. λ. By the change of variable theorem,

$$\int_{v \circ g(x) \neq 0} dP_X = \int_{v(t) \neq 0} dP_X \circ g^{-1}.$$

By assumption 4, $P_X \circ g^{-1} \ll \lambda$. Hence the above integral is 0, implying $v \circ g(x) = 0$ a.s. P_X, or, equivalently, $v \circ g(X) = 0$ a.s. P. □

Inverse regressions of this type are considered in Cook (2007), Cook and Forzani (2009), and Cook, Li and Chiaromonte (2010) for linear SDR. The above two propositions show that a complete and sufficient dimension reduction class exists for a reasonably wide range of problems, including forward and inverse regressions of very general, nonparameterized form. The next theorem shows that when a complete and sufficient dimension reduction class exists, it is unique and coincides with the central class. Once again, the situation here echoes that in classical theory, where a complete and sufficient statistic, if it exists, coincides with the minimal sufficient statistic; see Lehmann (1981).

THEOREM 3. *Suppose $\{\Pi_y : y \in \Omega_Y\}$ is dominated by a σ-finite measure, and $\mathcal{G}$ is a sub σ-field of $\sigma(X)$. If $\mathfrak{M}_{\mathcal{G}}$ is a complete and sufficient dimension reduction class, then*

$$\mathfrak{M}_{\mathcal{G}} = \mathfrak{C}_{Y|X} = \mathfrak{S}_{Y|X}.$$

PROOF. If $f \perp \mathfrak{C}_{Y|X}$, then by Lemma 1, $E(f|Y) = 0$ which, because $\mathfrak{M}_{\mathcal{G}}$ is sufficient, implies

$$E[E(f|\mathcal{G})|Y] = 0.$$

Because $\mathfrak{M}_{\mathcal{G}}$ is complete and because $E(f|\mathcal{G}) \in \mathfrak{M}_{\mathcal{G}}$, we have $E(f|\mathcal{G}) = 0$. By Lemma 1, this implies $f \perp \mathfrak{M}_{\mathcal{G}}$. Thus we have proved $\mathfrak{M}_{\mathcal{G}} \subseteq \mathfrak{C}_{Y|X}$. However, by Theorem 2 we know that $\mathfrak{C}_{Y|X} \subseteq \mathfrak{S}_{Y|X} \subseteq \mathfrak{M}_{\mathcal{G}}$. This proves the desired equality. □

5. Generalizations of SIR and their population-level properties. From the previous developments we see that the subspace $L_2(P_X) \ominus L_2(P_Y)$ of $L_2(P_X)$ plays a critical role in nonlinear SDR. Its orthogonal complement in $L_2(P_X)$ coincides with the central class $\mathfrak{S}_{Y|X}$ under completeness, and even without completeness it is guaranteed to be inside $\mathfrak{S}_{Y|X}$. It turns out that this subspace can be expressed as the range of a certain bounded linear operators. This representation ensures that estimation procedures can rely on simple spectral decompositions, rather than complicated numerical optimizations. We first introduce some covariance operators which are the building block of this approach.

5.1. *Covariance operators.* Since constants are irrelevant here (e.g., f and $f + 3$ can be considered as the same function), we will speak of set relations modulo constants. If A and B are sets, then we say $A \subseteq B$ modulo constants if for each $f \in A$ there is $c \in \mathbb{R}$ such that $f + c \in B$. We say that A is a dense subset of B modulo constants if (i) $A \subseteq B$ modulo constants and (ii) for each $f \in B$, there is a sequence $\{f_n\} \subseteq A$ and a sequence of constants $\{c_n\} \subseteq \mathbb{R}$ such that $\{f_n + c_n\} \subseteq A$ and $f_n + c_n \to f$ in the topology for B. Let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be Hilbert spaces of functions of X and Y satisfying the conditions:

(A) $\mathcal{H}_X$ and $\mathcal{H}_Y$ are dense subsets of $L_2(P_X)$ and $L_2(P_Y)$ modulo constants;

(B) there are constants $C_1 > 0$ and $C_2 > 0$ such that $\mathrm{var}[f(X)] \le C_1\|f\|_{\mathcal{H}_X}$ and $\mathrm{var}[g(Y)] \le C_2\|g\|_{\mathcal{H}_Y}$.

Although we will later take $\mathcal{H}_X$ and $\mathcal{H}_Y$ to be reproducing kernel Hilbert spaces (RKHS), our theory is not restricted to such spaces. In particular, we do not require the evaluation functionals [such as $f \mapsto f(x)$ from $\mathcal{H}_X$ to $\mathbb{R}$] to be continuous.

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 all bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$, and let $\mathcal{B}(\mathcal{H}_1)$ abbreviate $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_1)$. We denote the range of a linear operator A by $\operatorname{ran} A$, the kernel of A by $\ker A$, and the closure of ran A by $\overline{\text{ran}}\,A$. Under assumption (B), the symmetric bilinear form $u:\mathcal{H}_X \times \mathcal{H}_X \to \mathbb{R}$ defined by $u(f,g) = \text{cov}[f(X), g(X)]$ is bounded and thus induces an operator $M_{XX} \in \mathcal{B}(\mathcal{H}_X)$ that satisfies $\langle f, M_{XX}g\rangle_{\mathcal{H}_X} = u(f,g)$. Similarly, the bounded bilinear form $(f,g) \mapsto \text{cov}[f(Y), g(Y)]$ from $\mathcal{H}_Y \times \mathcal{H}_Y$ to $\mathbb{R}$ defines an operator $M_{YY} \in \mathcal{B}(\mathcal{H}_Y)$. Let $\mathcal{G}_X$ and $\mathcal{G}_Y$ represent the subspaces $\overline{\text{ran}}\,M_{XX}$ and $\overline{\text{ran}}\,M_{YY}$.

DEFINITION 6. Suppose conditions (A) and (B) are satisfied. We define the covariance operators $\Sigma_{XX}:\mathcal{G}_X \to \mathcal{G}_X$, $\Sigma_{YY}:\mathcal{G}_Y \to \mathcal{G}_Y$ and $\Sigma_{YX}:\mathcal{G}_X \to \mathcal{G}_Y$ through the relations

$$\langle f, \Sigma_{XX}g\rangle_{\mathcal{G}_X} = \langle f, g\rangle_{L_2(P_X)}, \qquad \langle f, \Sigma_{YY}g\rangle_{\mathcal{G}_Y} = \langle f, g\rangle_{L_2(P_Y)},$$
$$\langle f, \Sigma_{YX}g\rangle_{\mathcal{G}_Y} = \langle f, g\rangle_{L_2(P_Y)}.$$

These operators are essentially the same as those introduced by Fukumizu, Bach and Jordan (2004, 2009), except that here we do not assume $\mathcal{H}_X$ and $\mathcal{H}_Y$ to be RKHS. By Baker [(1973), Theorem 1], there is a unique operator $R_{YX} \in \mathcal{B}(\mathcal{G}_X, \mathcal{G}_Y)$ such that $\Sigma_{YX} = \Sigma_{YY}^{1/2} R_{YX} \Sigma_{XX}^{1/2}$. We call R_{XY} the *correlation operator*. In order to connect these operators with the central class, which is an $L_2(P_X)$-object, we need to extend the domains of $\Sigma_{XX}^{1/2}$ and $\Sigma_{YY}^{1/2}$ from $\mathcal{G}_X$ and $\mathcal{G}_Y$ to $L_2(P_X)$ and $L_2(P_Y)$. The following extension theorem is important and nontrivial, but since the material presented here can be understood without its proof we relegate it to the supplementary material [Lee, Li and Chiaromonte (2013)].

THEOREM 4. *Under assumptions* (A) *and* (B), *there exist unique isomorphisms*

$$\tilde{\Sigma}_{XX}^{1/2}: L_2(P_X) \to \mathcal{G}_X, \qquad \tilde{\Sigma}_{YY}^{1/2}: L_2(P_Y) \to \mathcal{G}_Y$$

that agree with $\Sigma_{XX}^{1/2}$ *and* $\Sigma_{YY}^{1/2}$ *on* $\mathcal{G}_X$ *and* $\mathcal{G}_Y$ *in the sense that, for all* $f \in \mathcal{G}_X$ *and* $g \in \mathcal{G}_Y$,

$$\tilde{\Sigma}_{XX}^{1/2}(f - Ef) = \Sigma_{XX}^{1/2} f, \qquad \tilde{\Sigma}_{YY}^{1/2}(g - Eg) = \Sigma_{YY}^{1/2} g.$$

Furthermore, for any $f \in L_2(P_X)$, $g \in L_2(P_Y)$ *we have*

$$\langle \tilde{\Sigma}_{YY}^{1/2} g, R_{YX} \tilde{\Sigma}_{XX}^{1/2} f\rangle_{\mathcal{G}_Y} = \text{cov}[g(Y), f(X)]. \tag{8}$$

The easiest way to understand equality (8) is through the special case where $f = f' - E(f')$, $g = g' - E(g')$ where $f' \in \mathcal{G}_X$, $g' \in \mathcal{G}_Y$. In this case,

$$\begin{aligned}\langle \tilde{\Sigma}_{YY}^{1/2} g, R_{YX} \tilde{\Sigma}_{XX}^{1/2} f\rangle_{\mathcal{G}_Y} &= \langle \Sigma_{YY}^{1/2} g', R_{YX} \Sigma_{XX}^{1/2} h'\rangle_{\mathcal{G}_Y} = \langle g', \Sigma_{YX} f'\rangle_{\mathcal{G}_Y} \\ &= \text{cov}[f(X), g(Y)].\end{aligned}$$

The theorem also implies that, for all $f, g \in L_2(P_X)$ and $s, t \in L_2(P_Y)$,

$$\langle \tilde{\Sigma}_{XX}^{1/2} g, \tilde{\Sigma}_{XX}^{1/2} f \rangle_{\mathcal{G}_X} = \langle g, f \rangle_{L_2(P_X)} = \operatorname{cov}[g(X), f(X)],$$

$$\langle \tilde{\Sigma}_{YY}^{1/2} s, \tilde{\Sigma}_{YY}^{1/2} t \rangle_{\mathcal{G}_Y} = \langle s, t \rangle_{L_2(P_Y)} = \operatorname{cov}[s(Y), t(Y)].$$

5.2. *Generalized SIR.* The results of the last subsection allow us to characterize $L_2(P_X) \ominus L_2(P_Y)$ in terms of extended covariance operators, which is the key to developing its estimator. Recall that classical SIR [Li (1991)] for linear SDR is based on the matrix

$$[\operatorname{var}(X)]^{-1} \operatorname{var}[E(X|Y)]. \tag{9}$$

Under the linear conditional mean assumption requiring that $E(X|\beta^{\mathsf{T}} X)$ be linear in X for any matrix β spanning $\mathcal{S}_{Y|X}$, the re-scaled "inverse" conditional mean $[\operatorname{var}(X)]^{-1} E(X|Y)$ is contained in this space. To generalize this to the nonlinear setting, we first introduce a conditional mean operator.

DEFINITION 7. We call the operator $\tilde{\Sigma}_{YY}^{-1/2} R_{YX} \tilde{\Sigma}_{XX}^{1/2} : L_2(P_X) \to L_2(P_Y)$ the conditional expectation operator, and denote it by $E_{X|Y}$.

The relation between the conditional expectation operator and conditional expectations is elucidated by the next proposition, which is followed by an important corollary.

PROPOSITION 3. *Under conditions* (A) *and* (B), *we have*:

(1) *for any* $f \in L_2(P_X)$, $E_{X|Y} f = E(f(X)|Y)$;
(2) *for any* $g \in L_2(P_Y)$, $E_{X|Y}^* g = E(g(Y)|X)$.

PROOF. For any $g \in L_2(P_Y)$,

$$\begin{aligned}\langle E_{X|Y} f, g \rangle_{L_2(P_Y)} &= \langle \tilde{\Sigma}_{YY}^{-1/2} R_{YX} \tilde{\Sigma}_{XX}^{1/2} f, g \rangle_{L_2(P_Y)} = \langle R_{YX} \tilde{\Sigma}_{XX}^{1/2} f, \tilde{\Sigma}_{YY}^{1/2} g \rangle_{\mathcal{H}_Y} \\ &= \operatorname{cov}(f(X), g(Y)),\end{aligned}$$

where the last equality follows from (8). Hence $\operatorname{cov}(f(X) - (E_{X|Y} f)(Y), g(Y)) = 0$. By the definition of conditional expectation, $E_{X|Y} f = E(f(X)|Y)$, which proves 1. Assertion 2 follows from the fact that $\tilde{\Sigma}_{YY}^{-1/2}$ and $\tilde{\Sigma}_{XX}^{1/2}$ are isomorphisms, and $R_{YX}^* = R_{XY}$. □

COROLLARY 1. *Under conditions* (A) *and* (B), *for any* $f, g \in L_2(P_X)$,

$$\langle g, E_{X|Y}^* E_{X|Y} f \rangle_{L_2(P_X)} = \operatorname{cov}[E(g(X)|Y), E(f(X)|Y)]. \tag{10}$$

Moreover, $E_{X|Y}^* E_{X|Y} \in \mathcal{B}(L_2(P_X))$, *and its norm is no greater than* 1.

PROOF. We have

$$\begin{aligned}\langle g, E_{X|Y}^* E_{X|Y} f\rangle_{L_2(P_X)} &= \langle E_{X|Y} g, E_{X|Y} f\rangle_{L_2(P_X)}\\ &= \langle E(g(X)|Y), E(f(X)|Y)\rangle_{L_2(P_X)},\end{aligned}$$

which is the right-hand side of (10). Moreover, since $\tilde{\Sigma}_{XX}^{1/2}$ is isomorphic, we have

$$E_{X|Y}^* E_{X|Y} = (\tilde{\Sigma}_{YY}^{-1/2} R_{YX} \tilde{\Sigma}_{XX}^{1/2})^* (\tilde{\Sigma}_{YY}^{-1/2} R_{YX} \tilde{\Sigma}_{XX}^{1/2}) = \tilde{\Sigma}_{XX}^{-1/2} R_{XY} R_{YX} \tilde{\Sigma}_{XX}^{1/2}.$$

Hence $\|E_{X|Y}^* E_{X|Y}\| \le \|\tilde{\Sigma}_{XX}^{-1/2}\| \|R_{XY}\| \|R_{YX}\| \|\tilde{\Sigma}_{XX}^{1/2}\|$. Because $\tilde{\Sigma}_{XX}^{1/2}$ and $\tilde{\Sigma}_{XX}^{-1/2}$ are isomorphisms, their norms are both 1. By Baker [(1973), Theorem 1], $\|R_{YX}\| \le 1$. Hence $\|E_{X|Y}^* E_{X|Y}\| \le 1$. □

From this corollary we see that the quadratic form

$$f \mapsto \langle f, E_{X|Y}^* E_{X|Y} f\rangle_{L_2(P_X)}, \qquad L_2(P_X) \times L_2(P_X) \to \mathbb{R}$$

generalizes the matrix $\text{var}[E(X|Y)]$ of the linear case, which is the essential ingredient of SIR for linear SDR. It is then not surprising that the operator $E_{X|Y}^* E_{X|Y}$ is closely connected to the central class for nonlinear SDR, as shown in the following theorem.

THEOREM 5. *If conditions* (A) *and* (B) *are satisfied and* $\mathfrak{S}_{Y|X}$ *is complete, then*

$$\overline{\text{ran}}(E_{X|Y}^* E_{X|Y}) = \mathfrak{S}_{Y|X}.$$

PROOF. By Lemma 1, $f \in \mathfrak{C}_{Y|X}$ if and only if $f \in L_2(P_X)$ and $E(f|Y) = 0$. By Proposition 3, this happens if and only if $f \in \ker E_{X|Y}$. This shows $\ker E_{X|Y} = \mathfrak{C}_{Y|X}^{\perp}$. However, because $\ker(E_{X|Y}) = \ker(E_{X|Y}^* E_{X|Y})$, we have

$$\overline{\text{ran}}(E_{X|Y}^* E_{X|Y}) = [\ker(E_{X|Y}^* E_{X|Y})]^{\perp} = (\ker E_{X|Y})^{\perp} = (\mathfrak{C}_{Y|X}^{\perp})^{\perp} = \mathfrak{C}_{Y|X}.$$

Since $\mathfrak{S}_{Y|X}$ is complete, we have $\mathfrak{C}_{Y|X} = \mathfrak{S}_{Y|X}$, as desired. □

Note that, unlike in classical SIR for linear SDR, here we do not have to consider an analogue to the rescaling $[\text{var}(X)]^{-1}$ in (9). This is because the $L_2(P_X)$-inner product absorbs the marginal variance in the predictor vector. We refer to the sample estimator based on $\overline{\text{ran}}(E_{X|Y}^* E_{X|Y})$ (see Section 7.2) as *generalized SIR* or GSIR. The GSIR estimator is related to kernel canonical component analysis (KCCA) introduced by Bach and Jordan (2002); see also Fukumizu, Bach and Gretton (2007). In Section 7.2 we will explore similarities and differences between these two methods.

5.3. *Kernel SIR*. We now turn to another nonlinear SDR estimator, which was proposed by Wu (2008) and further studied by Yeh, Huang and Lee (2009), called *kernel sliced inverse regression* (KSIR). In our setting, the population-level description of this estimator is as follows. Let $\mathcal{H}_X$ be a Hilbert space satisfying (A) and (B) (in this case an RKHS, but this assumption is unnecessary). Let $T:\mathcal{H}_X \to L_2(P_X)$ be the centering transformation $T(f) = f - E(f)$. Let $J_1, \ldots, J_h$ be a partition of Ω_Y, and let $\mu_1, \ldots, \mu_h \in \overline{\operatorname{ran}}\, T$ be the Riesz representations of the linear functionals

$$T_j : \overline{\operatorname{ran}}\, T \to \mathbb{R}, \qquad g \mapsto E\big(g(X)|Y \in J_i\big), \qquad i = 1, \ldots, h.$$

In our language, KSIR uses (the sample version of) the subspace $\operatorname{span}(\Sigma_{XX}^{-1}\mu_1, \ldots, \Sigma_{XX}^{-1}\mu_h)$ to estimate $\mathfrak{S}_{Y|X}$. The next theorem shows that any such representation must be a member of $\mathfrak{C}_{Y|X}$, and thus of $\mathfrak{S}_{Y|X}$ (since $\mathfrak{C}_{Y|X} \subseteq \mathfrak{S}_{Y|X}$)—which implies that KSIR is unbiased.

THEOREM 6. *If* (A) *and* (B) *hold, then* $\mu_j \in \mathfrak{C}_{Y|X}$.

PROOF. By condition (A), $\overline{\operatorname{ran}}\, T = L_2(P_X)$. If $f \in L_2(P_X) \ominus L_2(P_Y) \subseteq \overline{\operatorname{ran}}\, T$, then, by Lemma 1, $E(f|Y) = 0$. Hence $\langle f, \mu_i \rangle_{L_2(P_X)} = E[f(X)|Y \in J_i] = 0$. □

Yeh, Huang and Lee (2009) give another unbiasedness proof for KSIR, but they assume that the spanning functions of $\mathcal{H}_X$, say $f_1, \ldots, f_m$, satisfy the linear conditional mean assumption. That is, for any $f \in \mathcal{H}_X$, $E(f|f_1, \ldots, f_m)$ has the form $c_0 + c_1 f_1 + \cdots + c_m f_m$ for some $c_0, \ldots, c_m \in \mathbb{R}$. This condition is an analogue of the linear conditional mean assumption for linear SDR; see, for example, Li (1991) and Cook and Li (2002). Interestingly, our result no longer relies on this assumption. The reason Yeh, Huang and Lee need the assumption in the first place is that they define the central class [Definition 1 of Yeh, Huang and Lee (2009)] as the linear subspace spanned by $h_1, \ldots, h_d$ in $\operatorname{span}(f_1, \ldots, f_m)$ such that

$$Y \perp\!\!\!\perp X | h_1(X), \ldots, h_d(X), \tag{11}$$

whereas we define the central class as the class of all measurable functions of $h_1, \ldots, h_d$. Indeed, in the nonlinear setting there is no reason to restrict to this linear span formulation, since the conditional independence (11) only relies on the σ-field generated by $h_1, \ldots, h_d$.

6. Beyond completeness: Generalized SAVE.

We now turn to the more general problem of estimating the central class when it is not complete, in which case the regression class may be a proper subset of the central class. We will generalize SAVE [Cook and Weisberg (1991)] to the nonlinear case and show that it can recover functions beyond the regression class.

The setting here is different from that for GSIR in two respects. First, since we now deal with the location-invariant quantity $f(X) - E[f(X)|Y]$, we no longer need to define the conditional mean operator through the centered L_2-spaces $L_2(P_Y)$ and $L_2(P_X)$. Second, we now define relevant operators through L_2-spaces instead of RKHSs, which is more convenient in this context. Let $L_2'(P_X)$ and $L_2'(P_Y)$ denote the noncentered L_2-spaces. Define the noncentered conditional mean operator $E'_{X|Y} : L_2'(P_X) \to L_2'(P_Y)$ through

$$\langle g, E'_{X|Y} f\rangle_{L_2'(P_X)} = E\big(g(Y)f(X)\big), \qquad f \in L_2'(P_X), g \in L_2'(P_Y). \tag{12}$$

By the same argument of Proposition 3, $E'_{X|Y} f = E(f(X)|Y)$. To generalize SAVE, we introduce a new type of conditional variance operator.

DEFINITION 8. For each $y \in \Omega_Y$, the bilinear form

$$L_2(P_X) \times L_2(P_X) \to \mathbb{R}, \qquad (f, g) \mapsto \big(E'_{X|Y}(fg) - E'_{X|Y} f E'_{X|Y} g\big)(y)$$

uniquely defines an operator $V_{X|Y}(y) \in \mathcal{B}(L_2(P_X))$ via the Riesz representation. We call the random operator

$$V_{X|Y} : \Omega_Y \to \mathcal{B}\big(L_2(P_X)\big), \qquad y \mapsto V_{X|Y}(y)$$

the heteroscedastic conditional variance operator given Y.

The operator $V_{X|Y}$ is different from the conditional variance operator $\Sigma_{X|Y}$ introduced by Fukumizu, Bach and Jordan (2004, 2009). In a sense, $\Sigma_{X|Y}$ is a generalization of $E[\mathrm{var}(X|Y)]$ rather than $\mathrm{var}(X|Y)$, because $\langle f, \Sigma_{X|Y} f\rangle_{\mathcal{H}_X} = E[\mathrm{var}(f(X)|Y)]$. Note that $E[\mathrm{var}(f(X)|Y)]$ becomes $\mathrm{var}(f(X)|Y)$ only when the latter is nonrandom. So $\Sigma_{X|Y}$ might be called a *homoscedastic* conditional variance operator. In contrast, $\langle f, V_{X|Y} f\rangle_{L_2(P_X)}$ gives directly the conditional variance $\mathrm{var}[f(X)|Y]$, hence the term heteroscedastic conditional variance operator. Here, we should also stress that $E'_{X|Y}$ is defined between noncentered $L_2'(P_X)$ and $L_2'(P_Y)$, whereas $V_{X|Y}(y)$ is defined between centered $L_2(P_X)$ and $L_2(P_X)$.

We now define the expectation of a generic random operator $A : \Omega_Y \to \mathcal{B}(L_2(P_X))$. For each $f \in L_2(P_X)$ and $x \in \Omega_X$, the mapping $y \mapsto (A(y)f)(x)$ defines a random variable. Its expectation defines a function $x \mapsto \int_{\Omega_Y} (A(y)f)(x) \times P_Y(dy)$, which is a member of $L_2(P_X)$. Denoting this member as $\tilde{f}$, we define the nonrandom operator $L_2(P_X) \to L_2(P_X)$, $f \mapsto \tilde{f}$ as the expectation $E(A)$. We now consider the operator

$$S = E(V - V_{X|Y})^2 : L_2(P_X) \to L_2(P_X), \tag{13}$$

where $V : L_2(P_X) \to L_2(P_X)$ is the (unconditional) covariance operator defined by

$$\langle f, Vg\rangle_{L_2(P_X)} = \mathrm{cov}\big(f(X), g(X)\big).$$

This operator is similar to $\tilde{\Sigma}_{XX}$ in Section 5 except that it is not defined through RKHS. The operator S is an extension of the SAVE matrix [Cook and Weisberg (1991)]

$$\Sigma^{-1} E\big[\mathrm{var}(X) - \mathrm{var}(X|Y)\big]^2 \Sigma^{-1}. \tag{14}$$

Let β be a basis matrix of the central subspace $\mathcal{S}_{Y|X}$ of linear SDR. Cook and Weisberg show that if $E(X|\beta^{\mathsf{T}} X)$ is linear in $\beta^{\mathsf{T}} X$ and $\mathrm{var}(X|\beta^{\mathsf{T}} X)$ is nonrandom, then the column space of (14) is contained in $\mathcal{S}_{Y|X}$. The next theorem generalizes this result, but without requiring an analogue of the linear conditional mean assumption.

THEOREM 7. *Suppose that conditions* (A) *and* (B) *are satisfied*, *and* $\mathrm{var}[f(X)|\mathcal{G}_{Y|X}]$ *is nonrandom for any* $f \in \mathfrak{S}_{Y|X}^{\perp}$. *Then* $\overline{\mathrm{ran}}\, S \subseteq \mathfrak{S}_{Y|X}$.

PROOF. Let $f \perp \mathfrak{S}_{Y|X}$. We claim that for any $y \in \Omega_Y$,

$$\langle f, [V - V_{X|Y}(y)] f\rangle_{L_2(P_X)} = 0. \tag{15}$$

Because $Y \perp\!\!\!\perp X|\mathcal{G}_{Y|X}$, we have

$$\mathrm{var}\big(f(X)|Y\big) = \mathrm{var}\big(E\big(f(X)|\mathcal{G}_{Y|X}\big)|Y\big) + E\big(\mathrm{var}\big(f(X)|\mathcal{G}_{Y|X}\big)|Y\big).$$

Because, by Lemma 1, $E(f(X)|\mathcal{G}_{Y|X})$ is constant, the first term is 0. Because $\mathrm{var}(f(X)|\mathcal{G}_{Y|X})$ is nonrandom, the second term is $\mathrm{var}(f(X)|\mathcal{G}_{Y|X})$. Hence

$$\mathrm{var}\big(f(X)|Y\big) = \mathrm{var}\big(f(X)|\mathcal{G}_{Y|X}\big).$$

Similarly,

$$\mathrm{var}\big(f(X)\big) = \mathrm{var}\big(E\big(f(X)|\mathcal{G}_{Y|X}\big)\big) + E\big(\mathrm{var}\big(f(X)|\mathcal{G}_{Y|X}\big)\big) = \mathrm{var}\big(f(X)|\mathcal{G}_{Y|X}\big).$$

Therefore $\mathrm{var}(f(X)|Y) = \mathrm{var}(f(X))$, which implies (15). Since $V - V_{X|Y}(y)$ is self-adjoint, (15) implies $f \in \ker V_{X|Y}(y)$. Hence

$$\langle f, [V - V_{X|Y}(y)]^2 f\rangle_{L_2(P_X)} = 0.$$

Now integrate both sides of this equation to obtain

$$\begin{aligned}
&\int_{\Omega_Y} \langle f, (V - V_{X|Y}(y))^2 f\rangle_{L_2(P_X)} P_Y(dy) \\
&\quad = \left\langle f, \int_{\Omega_Y} (V - V_{X|Y}(y))^2 f\, P_Y(dy)\right\rangle_{L_2(P_X)} \\
&\quad = \langle f, (E(V - V_{X|Y})^2) f\rangle_{L_2(P_X)} = 0.
\end{aligned}$$

Hence $f \in \ker E(V - V_{X|Y})^2$, as desired. □

Similar to the case of GSIR, we do not need to employ the rescaling by Σ^{-1} in (14) when generalizing SAVE, because the $L_2(P_X)$-inner product absorbs any marginal variance. We call the estimator derived from $\overline{\operatorname{ran}} S$ (see Section 7.3) *generalized SAVE* or GSAVE. The next theorem shows that GSAVE can recover functions outside $\mathfrak{C}_{Y|X}$.

THEOREM 8. *If conditions* (A) *and* (B) *are satisfied*, *then* $\mathfrak{C}_{Y|X} \subseteq \overline{\operatorname{ran}} S$.

PROOF. Since S is self-adjoint, it suffices to show that $\ker S \subseteq \mathfrak{C}_{Y|X}^{\perp}$. For any $f \in \ker S$,

$$\int_{\Omega_Y} \langle f, (V - V_{X|Y})^2(y) f \rangle P_Y(dy) = 0.$$

Hence $\langle f, (V - V_{X|Y}(y))^2 f \rangle_{L_2(P_X)} = 0$ a.s. P_Y, which implies $(V - V_{X|Y}(y)) f = 0$ a.s. P_Y. Then

$$\int_{\Omega_Y} \langle f, \bigl(V - V_{X|Y}(y)\bigr) f \rangle_{L_2(P_X)} P_Y(dy) = 0.$$

By Definition 8, the left-hand side is $\operatorname{var}[f(X)] - E[\operatorname{var}(f(X)|Y)] = \operatorname{var}[E(f(X)| Y)]$. Hence $\operatorname{var}[E(f(X)|Y)] = 0$, which implies $E[f(X)|Y] = E[f(X)] = 0$. By Lemma 1, we have $f \in L_2(P_X) \ominus L_2(P_Y) = \mathfrak{C}_{Y|X}^{\perp}$, as desired. □

Combining Theorems 7 and 8 we see that

$$\mathfrak{C}_{Y|X} \subseteq \overline{\operatorname{ran}} S \subseteq \mathfrak{S}_{Y|X}, \tag{16}$$

which is analogous to the relation (3) in the classical setting. Thus we can expect GSAVE to discover functions outside the class $\mathfrak{C}_{Y|X}$, just as we can expect SAVE to discover vectors outside the space spanned by SIR.

7. Algorithms.

We now develop algorithms for the sample versions of GSIR and GSAVE, together with a cross-validation scheme to select parameters in the GSIR and GSAVE algorithms. These sample versions involve representing the operators in Theorems 5 and 7 as matrices. To formulate the algorithms we need to introduce coordinate representations of functions and operators, which we adopt with modifications from Horn and Johnson [(1985), page 31]; see also Li, Chun and Zhao (2012).

Throughout this section, $A^{\dagger}$ represents the Moore–Penrose inverse of a matrix A, $A^{\dagger\alpha}$ represents $(A^{\dagger})^{\alpha}$, I_n denotes the $n \times n$ identity matrix, 1_n denotes the vector in $\mathbb{R}^n$ whose entries are all 1 and $Q = I_n - 1_n 1_n^{\mathsf{T}}/n$. Let $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$ be a positive definite function. Also, let K_X be the $n \times n$ the Gram matrix $\{\kappa_X(X_i, X_j) : i, j = 1, \ldots, n\}$, G_X its centered versions QK_XQ and L_X the Gram matrices with intercept; that is, $L_X = (1_n, K_X)^{\mathsf{T}}$. Finally, define κ_Y, K_Y, G_Y, L_Y in the same manner for Y.

7.1. *Coordinate representation.* Let $\mathcal{H}$ be a finite-dimensional Hilbert space with spanning system $\mathcal{B} = \{b_1, \ldots, b_n\}$. For an $f \in \mathcal{H}$, let $[f]_{\mathcal{B}} \in \mathbb{R}^n$ denote the coordinates of f relative to $\mathcal{B}$; that is, $f = \sum_{i=1}^{n}([f]_{\mathcal{B}})_i b_i$. Let $b : \Omega_X \to \mathbb{R}^n$ denote the $\mathbb{R}^n$-valued function $(b_1, \ldots, b_n)^{\mathsf{T}}$. Then we can write $f = [f]_{\mathcal{B}}^{\mathsf{T}} b$. Let $A : \mathcal{H} \to \mathcal{H}'$, where $\mathcal{H}'$ is another finite-dimensional Hilbert spaces with spanning system $\mathcal{C} = \{c_1, \ldots, c_m\}$ and let $c = (c_1, \ldots, c_m)^{\mathsf{T}}$. Then, for $f \in \mathcal{H}$,

$$Af = A\big(b^{\mathsf{T}}[f]_{\mathcal{B}}\big) = (Ab_1, \ldots, Ab_n)[f]_{\mathcal{B}} = \big(c^{\mathsf{T}}[Ab_1]_{\mathcal{C}}, \ldots, c^{\mathsf{T}}[Ab_n]_{\mathcal{C}}\big)[f]_{\mathcal{B}}.$$

Thus, if we let ${}_{\mathcal{C}}[A]_{\mathcal{B}} = ([Ab_1]_{\mathcal{C}}, \ldots, [Ab_n]_{\mathcal{C}})$, then $Af = c^{\mathsf{T}}({}_{\mathcal{C}}[A]_{\mathcal{B}})[f]_{\mathcal{B}}$. In other words,

$$[Af]_{\mathcal{C}} = \big({}_{\mathcal{C}}[A]_{\mathcal{B}}\big)[f]_{\mathcal{B}}.$$

Furthermore, if $A_1 : \mathcal{H}' \to \mathcal{H}''$ is another linear operator, where $\mathcal{H}''$ is a third finite-dimensional Hilbert space with spanning system $\mathcal{D}$, then, by a similar argument,

$${}_{\mathcal{D}}[A_1 A]_{\mathcal{B}} = \big({}_{\mathcal{D}}[A_1]_{\mathcal{C}}\big)\big({}_{\mathcal{C}}[A]_{\mathcal{B}}\big).$$

Since the spanning systems in the domain and range of an operator are self-evident in the following discussion, we will write ${}_{\mathcal{C}}[A]_{\mathcal{B}}$ and $[f]_{\mathcal{B}}$ simply as $[A]$ and $[f]$.

Suppose $A \in \mathcal{B}(\mathcal{H})$ is self-adjoint. It can be shown that, for any $\alpha > 0$, $[A^{\alpha}] = [A]^{\alpha}$. Depending on the choice of the spanning system of $\mathcal{H}$, it is possible that A is invertible and yet $[A]$ is singular, but it is generally true that $A^{-\alpha} = [A]^{\dagger\alpha}$. Throughout this section the square brackets $[\cdot]$ will be used exclusively for denoting coordinate representations.

7.2. *Algorithm for GSIR.* At the sample level, P_X is replaced by the empirical measure $P_{n,X}$; $\mathcal{H}_X$ is the RKHS spanned by $\mathcal{B}_X = \{\kappa_X(\cdot, X_1), \ldots, \kappa_X(\cdot, X_n)\}$ with inner product $\langle f, g\rangle_{\mathcal{H}_X} = [f]^{\mathsf{T}} K_X [g]$, where $[\cdot]$ is coordinate with respect to $\mathcal{B}_X$. The space $L_2(P_{n,X})$ is spanned by $\kappa_X(\cdot, X_i) - E_n\kappa_X(X, X_i)$, $i = 1, \ldots, n$, with inner product $\langle f, g\rangle_{L_2(P_{n,X})} = \mathrm{cov}_n[f(X), g(X)] = n^{-1}[f]K_X Q K_X[g]$. The operator M_{XX} is defined through the relation $\langle f, M_{XX} g\rangle_{\mathcal{H}_X} = \mathrm{cov}_n(f(X), g(X))$; that is,

$$[f]^{\mathsf{T}} K_X [M_{XX}][g] = n^{-1}[f]^{\mathsf{T}} K_X Q K_X [g].$$

Since $[f]$ and $[g]$ are arbitrary members of $\mathbb{R}^n$, the above implies $[M_{XX}] = n^{-1} Q K_X$. Then any $f \in \operatorname{ran} M_{XX} \equiv \mathcal{G}_X$ can be written as $M_{XX} g$ for some $g \in \mathcal{H}_X$, which implies $[f] = Q K_X [g] = Q[f]$. Consequently, for any $f, g \in \mathcal{G}_X$, $\langle f, g\rangle_{\mathcal{H}_X} = [f]^{\mathsf{T}} G_X [g]$.

Let us now find the matrix representations of Σ_{XX}, Σ_{YY} and Σ_{YX}. In the following, h_X represents the function $x \mapsto (\kappa_X(x, X_1), \ldots, \kappa_X(x, X_n))^{\mathsf{T}}$. For any $f \in \mathcal{G}_X$, we have

$$\Sigma_{XX} f = M_{XX} f = h_X^{\mathsf{T}}[M_{XX}][f] = n^{-1} h_X^{\mathsf{T}} Q K_X Q [f] = n^{-1} h_X^{\mathsf{T}} G_X [f].$$

Hence $[\Sigma_{XX} f] = [\Sigma_{XX}][f] = n^{-1} G_X [f]$. Since this is true for all $[f] \in \mathrm{span}(Q)$, we have $[\Sigma_{XX}] = n^{-1} G_X$. By the same argument we can show that

$$\begin{aligned} &[\tilde{\Sigma}_{XX}] = n^{-1} G_X, \qquad [\Sigma_{YY}] = [\tilde{\Sigma}_{YY}] = n^{-1} G_Y, \\ &[\Sigma_{YX}] = [\tilde{\Sigma}_{YX}] = n^{-1} G_X, \qquad [E_{X|Y}] = G_Y^{\dagger} G_X G_X^{\dagger 1/2} G_X^{1/2}. \end{aligned} \tag{17}$$

Theorem 5 suggests that we use $\overline{\mathrm{ran}}(E^*_{X|Y} E_{X|Y})$ to estimate $\mathfrak{S}_{Y|X}$. Since $E^*_{X|Y} E_{X|Y}$ is an operator on $L_2(P_{n,X})$ to $L_2(P_{n,X})$, the vectors in $\overline{\mathrm{ran}}(E^*_{X|Y} E_{X|Y})$ can be found by

$$\text{maximizing} \quad \langle f, E^*_{X|Y} E_{X|Y} f \rangle_{L_2(P_{n,X})} = \| E_{X|Y} f \|^2_{L_2(P_{n,Y})}$$

subject to

$$\langle f, f \rangle_{L_2(P_{n,X})} = 1.$$

The coordinate representation of this problem is

$$\text{maximizing} \quad [f]^{\mathsf{T}} [E_{X|Y}]^{\mathsf{T}} G_Y^2 [E_{X|Y}][f] \quad \text{subject to} \quad [f]^{\mathsf{T}} G_X^2 [f] = 1.$$

The optimal solution is $[f] = G_X^{\dagger} \phi$, where ϕ are the leading eigenvectors of the matrix

$$\begin{aligned} &G_X^{\dagger} [E_{X|Y}]^{\mathsf{T}} G_Y^2 [E_{X|Y}] G_X^{\dagger} \\ &\quad = G_X^{\dagger} G_X^{1/2} G_X^{\dagger 1/2} G_X G_Y^{\dagger} G_Y^2 G_Y^{\dagger} G_X G_X^{\dagger 1/2} G_X^{1/2} G_X^{\dagger}. \end{aligned} \tag{18}$$

To enhance accuracy we replace the Moore–Penrose inverses $G_X^{\dagger}$ and $G_Y^{\dagger}$ by the ridge-regression-type regularized inverses $(G_X + \epsilon_X I_n)^{-1}$ and $(G_Y + \epsilon_Y I_n)^{-1}$. We summarize the algorithm as follows:

(1) Select the parameters $\gamma_X, \gamma_Y, \epsilon_X, \epsilon_Y$ using the algorithm in Section 7.4.
(2) Compute the matrix

$$(G_X + \epsilon_X I_n)^{-3/2} G_X^{3/2} (G_Y + \epsilon_Y I_n)^{-1} G_Y^2 (G_Y + \epsilon_Y I_n)^{-1} G_X^{3/2} (G_X + \epsilon_X I_n)^{-3/2}$$

and its first d eigenvectors $\phi_1, \ldots, \phi_d$ of this matrix.
(3) Form the sufficient predictors at x $\phi_i^{\mathsf{T}} (G_X + \epsilon_X I_n)^{-1} h_X(x)$, $i = 1, \ldots, d$.

GSIR estimation is similar to the kernel canonical correlation analysis (KCCA) developed by Akaho (2001), Bach and Jordan (2002) and Fukumizu, Bach and Gretton (2007). In our notation, KCCA maximizes

$$\langle g, \Sigma_{YX} f \rangle_{L_2(P_Y)} = [g]^{\mathsf{T}} G_Y G_X [f]$$

subject to $\langle g, \Sigma_{YY} g \rangle_{L_2(P_Y)} = [g]^{\mathsf{T}} G_Y^2 [f] = 1$ and $\langle f, \Sigma_{XX} f \rangle_{L_2(P_X)} = [g]^{\mathsf{T}} \times G_Y^2 [g] = 1$. The optimal solution for $[f]$ is $[f] = (G_X + \epsilon I_n)^{-1} \phi$, where ϕ is one of the first d eigenvectors of

$$(G_X + \epsilon I_n)^{-1} G_X G_Y (G_Y + \epsilon I_n)^{-2} G_Y G_X (G_X + \epsilon I_n)^{-1}.$$

We will compare GSIR and KCCA in Section 8.

7.3. *Algorithm for GSAVE.* We first derive the sample-level representation of the operator $V_{X|Y}(y)$. The sample version of the noncentered L_2-classes $L_2'(P_{n,X})$ and $L_2'(P_{n,Y})$ are spanned by

$$\mathcal{C}_X = \{1, \kappa_X(\cdot, X_1), \dots, \kappa_X(\cdot, X_n)\}, \qquad \mathcal{C}_Y = \{1, \kappa_Y(\cdot, Y_1), \dots, \kappa_Y(\cdot, Y_n)\}, \tag{19}$$

respectively. Let $[\cdot]$ represent the coordinates relative to these spanning systems. Then, for any $f \in L_2'(P_{n,X})$, $(f(X_1), \dots, f(X_n))^{\mathsf{T}} = L_X^{\mathsf{T}}[f]$. The operator $E'_{X|Y}$ is defined through the relation $\langle g, E'_{X|Y} f\rangle_{L_2'(P_{n,Y})} = E_n(g(Y)f(X))$, which yields the representation

$$[E'_{X|Y}] = (L_Y L_Y^{\mathsf{T}})^{\dagger}(L_Y L_X^{\mathsf{T}}). \tag{20}$$

Let ℓ_Y denote the function $y \mapsto (1, \kappa_Y(y, Y_1), \dots, \kappa_Y(y, Y_n))^{\mathsf{T}}$, and let ℓ_X denote the same function of x. For any $f, g \in L_2'(P_{n,X})$,

$$\begin{aligned} &\{E'_{X|Y}(fg) - (E'_{X|Y} f)(E'_{X|Y} g)\}(y) \\ &\quad = \ell_Y^{\mathsf{T}}(y)[E'_{X|Y}][fg] - [f]^{\mathsf{T}}[E'_{X|Y}]^{\mathsf{T}} \ell_Y(y)\ell_Y^{\mathsf{T}}(y)[E'_{X|Y}][g]. \end{aligned} \tag{21}$$

For any X_i, $f(X_i)g(X_i)$ can be expressed as the ith entry of the vector $L_X^{\mathsf{T}}[f] \odot L_X^{\mathsf{T}}[g]$, which is the same as $L_X^{\mathsf{T}}(L_X L_X^{\mathsf{T}})^{\dagger} L_X(L_X^{\mathsf{T}}[f] \odot L_X^{\mathsf{T}}[g])$, where $\odot$ is the Hadamard product. Thus we have the coordinate representation

$$[fg] = (L_X L_X^{\mathsf{T}})^{\dagger} L_X(L_X^{\mathsf{T}}[f] \odot L_X^{\mathsf{T}}[g]). \tag{22}$$

Substituting (20) and (22) into (21) we see that, for any $f, g \in L_2'(P_{n,X})$,

$$\begin{aligned} \langle f, V_{X|Y}(y) g\rangle_{L_2'(P_{n,X})} &= [f]^{\mathsf{T}} L_X(\operatorname{diag} C_Y(y) - C_Y(y)C_Y^{\mathsf{T}}(y))L_X^{\mathsf{T}}[g] \\ &\equiv [f]^{\mathsf{T}} L_X \Lambda(y) L_X^{\mathsf{T}}[g], \end{aligned} \tag{23}$$

where $C_Y(y) = L_Y^{\mathsf{T}}(L_Y L_Y^{\mathsf{T}})^{\dagger}\ell_Y(y)$.

Let $S_n : L_2(P_{n,X}) \to L_2(P_{n,X})$ be the operator $E_n(V - V_{X|Y}(Y))^2$. By Theorem 7, GSAVE is the class of functions $\overline{\operatorname{ran}}(S)$. At the sample level, this corresponds to

$$\text{maximizing} \quad \langle f, S_n f\rangle_{L_2(P_{n,X})} \quad \text{subject to} \quad \langle f, f\rangle_{L_2(P_{n,X})} = 1. \tag{24}$$

By (23), for each $y \in \Omega_Y$, and $f, g \in L_2(P_{n,X})$, we have

$$\langle g, V_{X|Y}(y) f\rangle_{L_2(P_{n,X})} = [f]^{\mathsf{T}} L_X Q \Lambda(y) Q L_X^{\mathsf{T}}[g].$$

From this we deduce that $[V_{X|Y}(y)] = (L_X Q L_X^{\mathsf{T}}/n)^{\dagger} L_X Q \Lambda(y) Q L_X^{\mathsf{T}}$. By a similar derivation we find $[V] = (L_X Q L_X^{\mathsf{T}}/n)^{\dagger}(L_X Q L_X^{\mathsf{T}}/n)$. Hence

$$[V - V_{X|Y}(y)] = (L_X Q L_X^{\mathsf{T}}/n)^{\dagger} L_X Q(Q/n - \Lambda(y)) Q L_X^{\mathsf{T}}.$$

It follows that

$$\langle f, S_n f\rangle_{L_2(P_{n,X})} = E_n\{[f]^{\mathsf{T}} L_X Q(Q/n - \Lambda(Y))Q(Q/n - \Lambda(Y))Q L_X^{\mathsf{T}}[f]\}.$$

To find $\operatorname{ran}(S_n)$ we maximize the above subject to $[f]^{\mathsf T}(L_X Q L_X^{\mathsf T}/n)[f] = 1$. Again we use the regularized inverses instead of the Moore–Penrose inverses to enhance performance. The algorithm is summarized as follows:

(1) Determine $\gamma_X, \gamma_Y, \epsilon_X, \epsilon_Y$ using the algorithm is Section 7.4.

(2) Compute $C = L_Y^{\mathsf T}(L_Y L_Y^{\mathsf T} + \epsilon_Y I_{n+1})^{-1/2} L_Y$. Let C_i be the ith column of C. Compute $\Lambda_i = \operatorname{diag}(C_i) - C_i C_i^{\mathsf T}$ and then compute $\Gamma_i = Q/n - \Lambda_i$ for $i = 1, \ldots, n$.

(3) Compute

$$n^{-1}\sum_{i=1}^{n}(L_X Q L_X^{\mathsf T} + \epsilon_X I_{n+1})^{-1/2} L_X Q \Gamma_i Q \Gamma_i Q L_X^{\mathsf T}(L_X Q L_X^{\mathsf T} + \epsilon_X I_{n+1})^{-1/2}$$

and the first d eigenvectors of this matrix, say $\phi_1, \ldots, \phi_d$.

(4) The sufficient predictors' values at any given $x \in \Omega_X$ are the set of d numbers

$$\ell_X^{\mathsf T}(x)(L_X Q L_X^{\mathsf T} + \epsilon_X I_{n+1})^{-1/2} Q \phi_i, \qquad i = 1, \ldots, d.$$

Here we should mention that, similar to SAVE for linear SDR, GSAVE works best for extracting predictors affecting the conditional variance of the response, but often not so well for extracting predictors affecting the conditional mean. However, we expect that other second-order methods for linear SDR, such as directional regression [Li and Wang (2007)] and the minimum discrepancy approach [Cook and Ni (2005)], will be amenable to similar generalizations to nonlinear SDR. These will be left for future research.

7.4. *Cross-validation algorithm.* We now develop a cross-validation scheme to determine the parameters γ_X, γ_Y, ϵ_X, ϵ_Y, which are used in the algorithms for both the GSIR and the GSAVE. We will only describe the algorithm for determining (γ_X, ϵ_X); that of (γ_Y, ϵ_Y) is completely analogous.

In the following, for a matrix A, $A_{-i,-j}$ represents the submatrix of A with its ith row and jth column removed, and $A_{-i,j}$ represents the jth column of A with the ith entry removed. Let $\mathcal{C}_Y^{-i} = \mathcal{C}_Y \setminus \{\kappa_Y(\cdot, Y_i)\}$, and define $\mathcal{C}_X^{-i}$ similarly. Our cross-validation strategy is to predict $f(Y_i)$ for each $f \in \mathcal{C}_Y^{-i}$, using the conditional mean operator developed from $(\mathcal{C}_X^{-i}, \mathcal{C}_Y^{-i})$. The regularized matrix representation of $E'_{Y|X}$ based on $(\mathcal{C}_X^{-i}, \mathcal{C}_Y^{-i})$ is

$$\begin{aligned}&[E'_{Y|X}]_{-(i+1),-(i+1)}\\ &\quad = \{(L_X)_{-(i+1),-i}(L_X)_{-(i+1),-i}^{\mathsf T} + \epsilon_X I_n\}^{-1}(L_X)_{-(i+1),-i}(L_Y)_{-(i+1),-i}^{\mathsf T}.\end{aligned}$$

The kth member f_k of $\mathcal{C}_Y^{-i}$ is the function $e_k^{\mathsf T}(\ell_Y)_{-(i+1)}(\cdot)$ where e_k is the vector in $\mathbb{R}^n$ whose kth entry is 1 and the remaining entries are 0. Therefore, the estimate of $E(f_k(Y)|X = x)$ based on on $\mathcal{C}_X^{-i}$ is

$$(\ell_X)_{-(i+1)}^{\mathsf T}(x)[E'_{Y|X} f_k]_{-(i+1)} = e_k^{\mathsf T}[E'_{Y|X}]_{-(i+1),-(i+1)}^{\mathsf T}(\ell_X)_{-(i+1)}(x),$$

and the prediction of $(f_1(Y_i), \dots, f_n(Y_i))^{\mathsf{T}}$ is $[E'_{Y|X}]^{\mathsf{T}}_{-(i+1),-(i+1)}(\ell_X)_{-(i+1)}(X_i)$. However, because $(\ell_X)_{-(i+1)}(X_i)$ is the vector $(L_X)_{-(i+1),i}$, and $(f_1(Y_i), \dots, f_n(Y_i))^{\mathsf{T}}$ is the vector $(L_Y)_{-(i+1),i}$, the difference between $(f_1(Y_i), \dots, f_n(Y_i))^{\mathsf{T}}$ and its prediction is

$$(L_Y)_{-(i+1),i} - [E'_{Y|X}]^{\mathsf{T}}_{-(i+1),-(i+1)}(L_X)_{-(i+1),i}.$$

To stress that this difference depends on $\gamma_X, \epsilon_X, \gamma_Y$, we denote it by $\Delta_i(\epsilon_X, \gamma_X, \gamma_Y)$. Our cross-validation criterion is defined as $\mathrm{CV}(\gamma_X, \epsilon_X, \gamma_Y) = \sum_{i=1}^n \|\Delta_i(\gamma_X, \epsilon_X, \gamma_Y)\|^2$. Since the role of γ_Y is only to determine the set of functions to be predicted, we exclude it from the optimization process (for the determination of ϵ_X, γ_X). Moreover, as argued in Fukumizu, Bach and Jordan (2009), the parameters ϵ_X and γ_X have similar smoothing effects and only one of them needs to be optimized. For these reasons we fix γ_Y and ϵ_X at

$$1/\gamma_{Y0} = \binom{n}{2}^{-1} \sum_{i<j} |Y_i - Y_j|^2, \qquad \epsilon_{X0} = 0.01 \tag{25}$$

and minimize $\mathrm{CV}(\gamma_X, \epsilon_{X0}, \gamma_{Y0})$ over a grid for γ_X. The grid consists of 20 subintervals in $[\gamma_{X0}/3, 3\gamma_{X0}]$, equally spaced in the log scale, where γ_{X0} is calculated using the first formula in (25) with $|Y_i - Y_j|$ replaced by $\|X_i - X_j\|$. The rationale for this formula can be found in Li, Artemiou and Li (2011).

The pair (γ_Y, ϵ_Y) is selected in the same way, except that ϵ_{Y0} is set to 0.001. This is because Y has dimension 1, so a weaker penalty is needed.

8. Simulations and data analysis.

In this section we present simulation comparisons among GSIR, GSAVE, KSIR and KCCA. For the reasons explained in the previous section, we compare GSIR with KSIR and KCCA in settings where the sufficient predictor appears in the conditional mean, and we compare GSAVE with GSIR, KSIR and KCCA in settings where the sufficient predictor appears in the conditional variance. We also apply GSIR, KSIR and KCCA to two real data sets.

8.1. *Simulation comparisons.*

To make a comprehensive comparison of GSIR, KSIR and KCCA we consider three regression models, namely:

$$\begin{cases} \text{I:} & Y = (X_1^2 + X_2^2)^{1/2} \log(X_1^2 + X_2^2)^{1/2} + \varepsilon; \\ \text{II:} & Y = X_1/(1 + e^{X_2}) + \varepsilon, \\ \text{III:} & Y = \sin(\pi(X_1 + X_2)/10) + \varepsilon, \end{cases}$$
$$\varepsilon \perp\!\!\!\perp X, \varepsilon \sim N(0, 0.25), p = 10;$$

as well as three distributional scenarios for the predictor vector X, namely: (A) independent Gaussian predictors, (B) independent non-Gaussian predictors and (C) correlated Gaussian predictors. In symbols:

$$\begin{cases} \text{A:} & X \sim N(0, I_p); \\ \text{B:} & X \sim (1/2)N(-1_p, I_p) + (1/2)N(1_p, I_p); \\ \text{C:} & X \sim N(0, 0.6 I_p + 0.4 1_p 1_p^{\mathsf{T}}). \end{cases}$$

TABLE 1

Comparison of KSIR, KCCA and GSIR when sufficient predictors appear in the conditional means

Models		**Spearman cor. with true predictor**			**Spearman cor. with response**		
X	$Y\|X$	**KSIR**	**KCCA**	**GSIR**	**KSIR**	**KCCA**	**GSIR**
A	I	0.78 (0.05)	0.81 (0.04)	0.80 (0.05)	0.63 (0.06)	0.66 (0.05)	0.64 (0.05)
	II	0.81 (0.05)	0.90 (0.03)	0.91 (0.03)	0.56 (0.06)	0.61 (0.05)	0.62 (0.05)
	III	0.76 (0.06)	0.89 (0.04)	0.91 (0.03)	0.47 (0.07)	0.56 (0.05)	0.56 (0.05)
B	I	0.88 (0.02)	0.88 (0.02)	0.87 (0.02)	0.82 (0.03)	0.81 (0.03)	0.80 (0.03)
	II	0.89 (0.03)	0.93 (0.02)	0.93 (0.02)	0.71 (0.04)	0.74 (0.04)	0.74 (0.04)
	III	0.90 (0.02)	0.97 (0.01)	0.97 (0.01)	0.72 (0.04)	0.77 (0.03)	0.77 (0.03)
C	I	0.79 (0.04)	0.82 (0.04)	0.81 (0.04)	0.64 (0.05)	0.66 (0.05)	0.65 (0.05)
	II	0.83 (0.05)	0.86 (0.06)	0.88 (0.04)	0.56 (0.06)	0.59 (0.06)	0.60 (0.06)
	III	0.83 (0.06)	0.96 (0.02)	0.96 (0.02)	0.56 (0.06)	0.65 (0.04)	0.65 (0.04)

Note that the central σ-fields for the three models I, II and III are generated by $X_1^2 + X_2^2$, $X_1/(1+e^{X_2})$ and $\sin(\pi(X_1+X_2)/10)$, respectively.

We assess the quality of an estimated sufficient predictor by its closeness to the true sufficient predictor and its closeness to the response. Since we are only interested in monotone functions of the predictor, we use Spearman's correlation as the measure of closeness. For each combination of the models and scenarios, we generate $n = 200$ observations on (X, Y) as the training data, and compute the first predicting function using the each of three methods. We then independently generate $m = 200$ observations on (X, Y) as the testing data, and evaluate the predicting functions at these points. Finally, we compute the mentioned Spearman's correlations from the testing data. This process is repeated $N = 200$ times. In Table 1 we list means and standard deviations of the Spearman's correlations computed using the $N = 200$ simulated samples. From the table we see that the performances of KCCA and GSIR are similar, and both are slightly better than KSIR.

Next, we compare GSAVE, KSIR, KCCA and GSIR when the predictors only affect the variance. We use the following models:

$$\begin{cases} \text{IV:} & Y = X_1\varepsilon; \\ \text{V:} & Y = (1/50)\big(X_1^3 + X_2^3\big)\varepsilon; \\ \text{VI:} & Y = \big(X_1/\big(1+e^{X_2}\big)\big)\varepsilon, \end{cases}$$

and again the scenarios (A), (B) and (C) for the distribution of X. The specifications of n, m, N, p are the same as in the previous comparison.

Because the sufficient predictors appear in the conditional variance $\operatorname{var}(Y|X)$ only, it is less meaningful to measure the closeness between the estimated sufficient predictor and the response. So in Table 2 we only report the means and standard deviations of Spearman's correlations between the estimated and true

TABLE 2
Comparison of KSIR, KCCA, GSIR and GSAVE when sufficient predictors appear in conditional variances

Models		**Spearman's correlation with true predictors**			
X	$Y\|X$	**GSAVE**	**KSIR**	**KCCA**	**GSIR**
A	IV	0.89 (0.08)	0.10 (0.07)	0.36 (0.22)	0.41 (0.23)
	V	0.73 (0.19)	0.09 (0.07)	0.17 (0.13)	0.20 (0.14)
	VI	0.84 (0.09)	0.10 (0.08)	0.25 (0.17)	0.27 (0.17)
B	IV	0.87 (0.08)	0.10 (0.07)	0.43 (0.25)	0.53 (0.25)
	V	0.88 (0.06)	0.09 (0.07)	0.11 (0.08)	0.11 (0.08)
	VI	0.76 (0.15)	0.27 (0.11)	0.61 (0.13)	0.64 (0.13)
C	IV	0.76 (0.20)	0.11 (0.07)	0.23 (0.16)	0.26 (0.18)
	V	0.82 (0.14)	0.10 (0.07)	0.11 (0.09)	0.12 (0.09)
	VI	0.73 (0.15)	0.15 (0.10)	0.41 (0.17)	0.44 (0.17)

sufficient predictors. We see that GSAVE performs substantially better than the other methods. The discrepancy can be explained by the fact that KSIR, KCCA and GSIR depend completely on $E[\mathrm{var}(f(X)|Y)]$, whereas GSAVE extracts more information from $\mathrm{var}(f(X)|Y)$.

8.2. *Data analysis*. We first consider the *faces data*, available at http://waldron.stanford.edu/isomap/datasets.html. This data set contains 698 images of the same sculpture of a face photographed at different angles and with different lighting directions. The predictor comprises 64 × 64 image pixels (thus $p = 4096$), and the response comprises horizontal rotation, vertical rotation and lighting direction measurements (thus $q = 3$). We use this data to demonstrate that the first three sufficient predictors estimated by KCCA and GSIR can effectively capture the 3-variate response. We use $n = 558$ of the images selected at random (roughly 80%) as training data, and the remaining $m = 140$ images as testing data. For each method, we estimate the first three predictor functions from the training data, and evaluate them on the testing data. The left panel of Figure 1 is the perspective plot of the first three KCCA predictors evaluated on the 140 testing images, and the right panel is the counterpart for GSIR. We did not include KSIR in this comparison because in its proposed form it cannot handle multivariate responses. The perspective plots indicate that nearby regions in the 3-D cubes have similar patterns of left–right rotation, up–down rotation and lighting direction, while distant regions have discernibly different patterns. This reflects the ability of the three sufficient predictors to capture the 3-variate responses.

Next, we apply KSIR, KCCA and GSIR to the *handwritten digits data*, available at http://www.cs.nyu.edu/~roweis/data.html. This data set contains 2000 images of

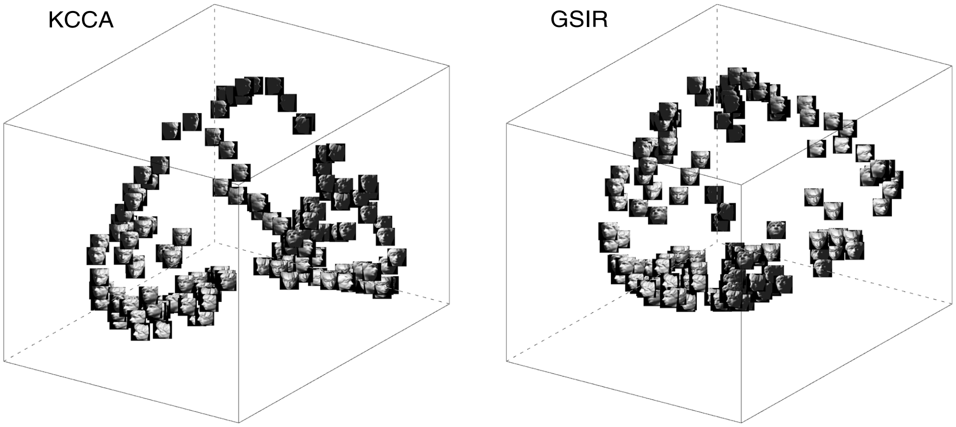

FIG. 1. *First* 3 *sufficient predictors by KCCA* (*left panel*) *and GSIR* (*right panel*), *computed from* 558 *training images*, *and evaluated on* 140 *testing images—faces data.*

$p = 16 \times 16$ pixels showing handwritten digits from 0 to 9—the response is thus categorical with 10 levels. We use 1000 images as training data and 1000 as testing data. Again, for each method we estimate the first three sufficient predictors on the training data, and evaluate them on the testing data. Results are presented in the three perspective plots in Figure 2—for visual clarity, these plots include only 100 randomly selected points from the 1000 in the testing data. The plots show that all three methods provide low-dimensional representations in which the digits are well separated.

9. Concluding remarks.

In this article we described a novel and very general theory of sufficient dimension reduction. This theory allowed us to combine linear and nonlinear SDR into a coherent system, to link them with classical statistical sufficiency, and to subsume several existing nonlinear SDR methods into a unique framework.

Our developments thus revealed important and previously unexplored properties of SDR methods. For example, unbiasedness of various nonlinear extensions of SIR proposed in recent literature was proved under the stringent linear conditional mean assumption. We were able to show that these methods are all unbiased under virtually no assumption, and that GSIR is exhaustive under the completeness assumption. We were also able to show that nonlinear extensions of SIR are in general *not* exhaustive when completeness is not satisfied, and that in these cases GSAVE can recover a larger portion of the central class. These insights could not have been obtained without paralleling linear and nonlinear SDR as allowed by our new theory.

In addition to achieving theoretical synthesis and important insights on SDR methods, we introduced a new *heteroscedastic* conditional variance operator—

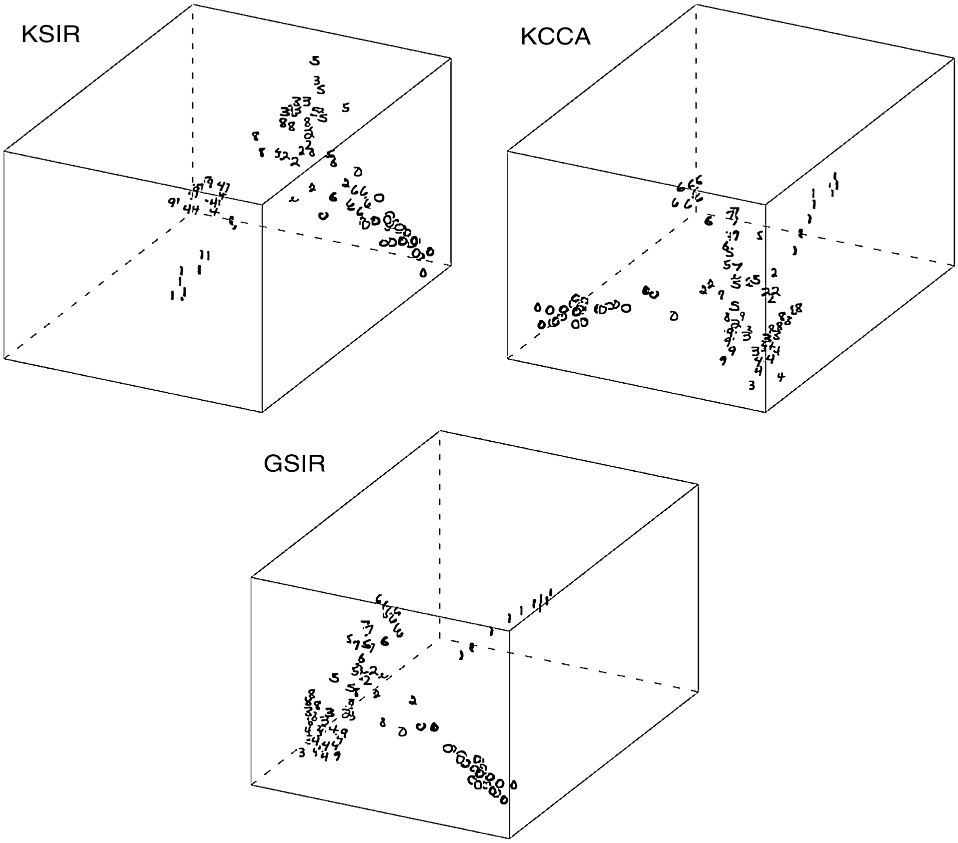

FIG. 2. *First* 3 *sufficient predictors by KSIR* (*upper-left panel*), *KCCA* (*upper-right panel*) *and GSIR* (*lower panel*), *computed on* 1000 *training images and evaluated on* 1000 *testing images—handwritten digits data.*

which is more general than the (homoscedastic) conditional variance operator in Fukumizu, Bach and Jordan (2004, 2009). This operator was crucial to generalizing SAVE to the nonlinear GSAVE, and thus to exploit dependence information in the conditional variance to improve upon the performance of the nonlinear extensions of SIR. We have no doubt that the heteroschedastic conditional variance operator can be used to generate nonlinear extensions of other second-order SRD methods such as contour regression [Li, Zha and Chiaromonte (2005)], directional regression [Li and Wang (2007)], SIR-II [Li (1991)] and other F2M methods [Cook and Forzani (2009)]. These extensions will be the topic of future work.

More generally, it is our hope that the clarity and simplicity that classical notions lend to the formulation of dimension reduction, as well as the transparent parallels we were able to draw between linear and nonlinear SDR, will provide fertile grounds for much research to come.

As we put forward a general theory that encompasses both linear and nonlinear SDR, it is also important to point out that linear SDR has its special values that cannot be replaced by nonlinear SDR via kernel mapping, one of which is its preservation of the original coordinates and as a result its strong interpretability. For example, when mapped to higher dimension spaces, kernel methods can sometimes interpret difference in variances in the original coordinates as location separation in the transformed coordinates, which can be undesirable depending on the goal and emphasis of particular applications. For further discussion and an example of this point, see Li, Artemiou and Li (2011).

Acknowledgments. We would like to thank two referees and an Associate Editor for their insightful comments and useful suggestions, which led to significant improvements to this paper. In particular, the consideration of nonlinear sufficient dimension reduction for functional data is inspired by the comments of two referees.

SUPPLEMENTARY MATERIAL

Supplement to "A general theory for nonlinear sufficient dimension reduction: Formulation and estimation" (DOI: 10.1214/12-AOS1071SUPP; .pdf). This is supplementary appendix that contains some techincal proofs of the results in the paper.

REFERENCES

AKAHO, S. (2001). A kernel method for canonical correlation analysis. In *Proceedings of the International Meeting of the Psychometric Society* (*IMPS*2001). Springer, Tokyo.

BACH, F. R. and JORDAN, M. I. (2002). Kernel independent component analysis. *J. Mach. Learn. Res.* **3** 1–48. MR1966051

BAHADUR, R. R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.* **25** 423–462. MR0063630

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

COOK, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In 1994 *Proceedings of the Section on Physical and Engineering Sciences* 18–25. Amer. Statist. Assoc., Alexandria, VA.

COOK, R. D. (1998a). *Regression Graphics*: *Ideas for Studying Regressions Through Graphics*. Wiley, New York. MR1645673

COOK, R. D. (1998b). Principal Hessian directions revisited. *J. Amer. Statist. Assoc.* **93** 84–94.

COOK, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22** 1–40. MR2408655

COOK, R. D. and CRITCHLEY, F. (2000). Identifying regression outliers and mixtures graphically. *J. Amer. Statist. Assoc.* **95** 781–794.

COOK, R. D. and FORZANI, L. (2009). Likelihood-based sufficient dimension reduction. *J. Amer. Statist. Assoc.* **104** 197–208. MR2504373

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

COOK, R. D., LI, B. and CHIAROMONTE, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statist. Sinica* **20** 927–1010.

COOK, R. D. and NI, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100** 410–428. MR2160547

COOK, R. D. and WEISBERG, S. (1991). Comment on "Sliced inverse regression for dimension reduction," by K.-C. Li. *J. Amer. Statist. Assoc.* **86** 328–332.

DUAN, N. and LI, K.-C. (1991). A bias bound for least squares linear regression. *Statist. Sinica* **1** 127–136. MR1101318

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

FUKUMIZU, K., BACH, F. R. and JORDAN, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.* **5** 73–99. MR2247974

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

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171

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

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

ICHIMURA, H. and LEE, L. F. (1991). Semiparametric least squares estimation of multiple index models: Single equation estimation. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics* (*Durham*, *NC*, 1988) (W. A. Barnett, J. L. Powell and G. Tauchen, eds.) 3–49. Cambridge Univ. Press, Cambridge. MR1174973

LEE, K. Y., LI, B. and CHIAROMONTE, F. (2013). Supplement to "A general theory for nonlinear sufficient dimension reduction: Formulation and estimation." DOI:10.1214/12-AOS1071SUPP.

LEHMANN, E. L. (1981). An interpretation of completeness and Basu's theorem. *J. Amer. Statist. Assoc.* **76** 335–340. MR0624335

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

LI, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **86** 316–342.

LI, B., ARTEMIOU, A. and LI, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Ann. Statist.* **9** 3182–3210.

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.

LI, K.-C. and DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052. MR1015136

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

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

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

WU, Q., LIANG, F. and MUKHERJEE, S. (2008). Regularized sliced inverse regression for kernel models. Technical report, Duke Univ., Durham, NC.

YE, Z. and WEISS, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.* **98** 968–979.

YEH, Y. R., HUANG, S. Y. and LEE, Y. Y. (2009). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Transactions on Knowledge and Data Engineering* **21** 1590–1603.

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

ZHU, H. and LI, L. (2011). Biological pathway selection through nonlinear dimension reduction. *Biostatistics* **12** 429–444.

K.-Y. LEE
DEPARTMENT OF BIOSTATISTICS
YALE SCHOOL OF PUBLIC HEALTH
60 COLLEGE STREET
P.O. BOX 208034
NEW HAVEN, CONNECTICUT
USA
E-MAIL: kuang-yao.lee@yale.edu

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