Communications for Statistical Applications and Methods
2024, Vol. 31, No. 2, 247–262

https://doi.org/10.29220/CSAM.2024.31.2.247
Print ISSN 2287-7843 / Online ISSN 2383-4757

A selective review of nonlinear sufficient dimension reduction

Sehun Jang[a], Jun Song[1,a]

[a]Department of Statistics, Korea University, Korea

Abstract

In this paper, we explore nonlinear sufficient dimension reduction (SDR) methods, with a primary focus on establishing a foundational framework that integrates various nonlinear SDR methods. We illustrate the generalized sliced inverse regression (GSIR) and the generalized sliced average variance estimation (GSAVE) which are fitted by the framework. Further, we delve into nonlinear extensions of inverse moments through the kernel trick, specifically examining the kernel sliced inverse regression (KSIR) and kernel canonical correlation analysis (KCCA), and explore their relationships within the established framework. We also briefly explain the nonlinear SDR for functional data. In addition, we present practical aspects such as algorithmic implementations. This paper concludes with remarks on the dimensionality problem of the target function class.

Keywords: nonlinear sufficient dimension reduction, central class, GSIR, GSAVE, KSIR, KCCA

1. Introduction

In the rapidly expanding field of data science, we frequently encounter high-dimensional datasets, which present challenges in extracting meaningful information due to the curse of dimensionality. Central to addressing these challenges is the concept of sufficient dimension reduction (SDR), which is a collection of methodologies about supervised dimension reduction without significant loss of regression information. More precisely, the classic linear SDR seeks to find a lower-dimensional space of linear transformation of a predictor random variable X without losing regression information. For example, let X and Y be random elements in $\mathbb{R}^p$ and $\mathbb{R}$. The goal of the linear SDR is to find $\beta \in \mathbb{R}^{p\times d}$, $d < p$ such that

$$Y \perp\!\!\!\perp X \mid \beta^{\mathsf{T}} X. \tag{1.1}$$

It implies that the relationship between Y and X is fully described d-dimensional random variable $\beta^{\mathsf{T}}X$. Since $d < p$, if β is available, we do not need to use all the variables in X. Instead, we replace it with $\beta^{\mathsf{T}}X$, which means that dimension reduction is possible. The subspace spanned by columns of β satisfying (1.1) is called the sufficient dimension reduction subspace and the minimal subspace of sufficient dimension reduction subspace is called the central dimension reduction subspace or central subspace, which is denoted by $\mathcal{S}_{Y|X}$. The existence of the central subspace is described in Cook (1998).

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1C1C1003647, No. RS-2023-00219212, and No. 2022M3J6A1063595) and a Korea University Grant (K2312771).

[1]Corresponding author: Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea. E-mail: junsong@korea.ac.kr

Published 31 March 2024 / journal homepage: http://csam.or.kr

The representative methods for estimating $\mathcal{S}_{Y|X}$ are the sliced inverse regression (SIR) introduced in Li (1991) and the sliced average variance estimation (SAVE) introduced in Cook and Weisberg (1991). Both methods are based on inverse moments: $E(X|Y))$, or $\mathrm{var}(X|Y)$. The term "inverse" comes from the fact that we normally deal with $E(Y|X)$ in the regression setting. A more comprehensive review on sufficient dimension reduction is illustrated in Li (2018).

With increasing complexity, linear indices are not enough to capture core characteristics in the relation between response and predictor. Hence, there is a need to extend the linear SDR to the nonlinear SDR, which induces the following setup.

$$Y \perp\!\!\!\perp X \mid h_1(X), \ldots, h_d(X), \tag{1.2}$$

where h_i is a function from the space in which X reside to $\mathbb{R}$ for $i = 1, \ldots, d$. In this setting, the SDR subspace is not clearly defined since the subspace is a concept for linear mapping. Regardless, (1.2 implies that we only need d predictors, $h_1(X), \ldots, h_d(X)$, in describing the relation between Y and $X \in \mathbb{R}^p$. Wu (2008) and Yeh *et al.* (2008) extend the SIR to the kernel sliced inverse regression (KSIR), which uses the kernel function to represent $h_1, \ldots, h_d$. Specifically, they use the kernel function κ that generates a reproducing kernel Hilbert space (RKHS), $\mathcal{H}$, by which (1.2) can be reformulated as

$$Y \perp\!\!\!\perp X \mid \langle h_1, \kappa(\cdot, X)\rangle_{\mathcal{H}}, \ldots, \langle h_d, \kappa(\cdot, X)\rangle_{\mathcal{H}}. \tag{1.3}$$

Hence, we can conclude that the KSIR conducts linear SDR on $\mathcal{H}$. In other respects, Fukumizu *et al.* (2007) proposes the kernel canonical correlation analysis (KCCA). The goal is to find functions such that $\sup_{f\in\mathcal{H}_X, g\in\mathcal{H}_Y} \mathrm{cor}(f(X), g(Y))$, where $\mathcal{H}_X, \mathcal{H}_Y$ are RKHSs generated by κ_X, κ_Y, respectively. Like KSIR, the KCCA is a linear problem on $\mathcal{H}_X \times \mathcal{H}_Y$ in that it seeks to maximize the correlation.

Lee *et al.* (2013) illustrates a general theory of nonlinear sufficient dimension reduction by introducing central σ-field as below.

$$Y \perp\!\!\!\perp X \mid \mathcal{G},$$

where $\mathcal{G}$ is a sub σ-field of $\sigma(X)$. Under mild conditions, there exists a unique minimal sub σ-field $\mathcal{G}_{Y|X}$ called by the central σ-field. However, $\mathcal{G}_{Y|X}$ is an algebraic structure, which is difficult to estimate directly. To estimate from data directly, Lee *et al.* (2013) provide the set of all $\mathcal{G}_{Y|X}$-measurable, square-integrable functions and calls it the central class denoted by $\mathfrak{S}_{Y_X}$. To recover $\mathfrak{S}_{Y_X}$ from data, the generalized sliced inverse regression (GSIR) and the generalized sliced average variance estimation (GSAVE) are developed. In addition, Li and Song (2017) adapts the GSIR and the GSAVE for functional data and develops f-GSIR and f-GSAVE then provides asymptotic theories for nonlinear SDR methods. More recently, Song *et al.* (2023) proposed a nonlinear extension of the principal fitted component, which is based on the likelihood of the predictors.

In this paper, we incorporate various nonlinear SDR methods into the framework of Lee *et al.* (2013). We show that the target function spaces of KSIR and KCCA are included in the central class $\mathfrak{S}_{Y|X}$ and delineate the relationship between KCCA and GSIR. The result functions of KCCA and GSIR are both in the range of some bounded operators.

The section of our paper is as follows. In Section 2, we introduce the kernel trick used in RKSH-based methods. In Section 3, we deal with the framework of nonlinear SDR and theoretical results of GSIR and GSAVE. Then, in Section 4, we review the KSIR in the case the response variable is scalar. In Section 5, the KCCA for nonlinear SDR of multivariate response variable is considered. Next, Section 6 deal with the nonlinear SDR for functional data: f-GSIR. Implementations of GSIR, GSAVE, KSIR, and KCCA are provided in Section 7. Finally, the concluding remarks regarding the application and the dimensionality are described in Section 8.

2. Preliminary: Kernel trick

Kernel methods can be extended to nonlinear methods through the use of the so-called kernel trick. The underlying key idea of this technique involves converting the original dataset into a higher-order space composed of nonlinear functions of the original dataset. Linear methods are then applied in this feature space. The most commonly used feature space is the reproducing kernel Hilbert space (RKHS), which facilitates easier computation and theoretical justification through the representer theorem. Aronszajn (1950) originally developed the concept of RKHS, which has been widely applied in the probability space context. Schölkopf and Smola (2002) provide a concise introduction to the main ideas of statistical learning theory, support vector machines, and kernel feature spaces, and Berlinet and Thomas-Agnan (2011) provide a rigorous framework and related theories of the RKHS within the context of probability space.

The RKHS can be defined in various ways using equivalent statements. One illustration is that the RKHS is a Hilbert space with a reproducing kernel. It is a function that, given a space of functions, allows for the evaluation of any function in that space at any point. For example, suppose that we want to construct a RKHS over a set $\mathcal{S}$ in a metric space and $\kappa(\cdot,\cdot)$ is a positive definite kernel such that $\kappa : \mathcal{S}\times\mathcal{S} \to \mathbb{R}$. Then the RKHS over $\mathcal{S}$ with the reproducing kernel can be represented by the completion of the following linear transformation of kernels

$$\mathcal{H} = \left\{ \sum_{i=1}^{\infty} a_i \kappa(\cdot, s_i) : \sum_{i=1}^{\infty} a_i^2 \kappa(\cdot, s_i) < \infty,\ a_i \in \mathbb{R}, s_i \in \mathcal{S} \right\},$$

where the inner product is,

$$\left\langle \sum_{i=1}^{\infty} a_i \kappa(\cdot, s_i), \sum_{j=1}^{\infty} b_j \kappa(\cdot, t_j) \right\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_i b_j \kappa(s_i, t_j).$$

Then the term reproducing comes from a property that for any $f(\cdot) = \sum_{i=1}^{\infty} a_i \kappa(\cdot, s_i) \in \mathcal{H}$, we have

$$f(s) = \langle f, \kappa(\cdot, s)\rangle_{\mathcal{H}}.$$

Now, suppose that we have a random sample $\{(X_1, \ldots, X_n)\}$ in $\mathbb{R}^p$. Then we map the original data to the RKHS such that $\{(\kappa(\cdot, X_1), \ldots, \kappa(\cdot, X_n)\}$ in

$$\mathcal{H}_n = \left\{ \sum_{i=1}^{n} a_i \kappa(\cdot, X_i) \right\}.$$

By treating $\kappa(\cdot, X_i)$ is a random element in $\mathcal{H}_n$, a linear method applied to $\kappa(\cdot, X_i)$ has a form of $\sum_{i=1}^{n} a_i \kappa(\cdot, X_i)$ which is a nonlinear method in terms of X_i.

3. Nonlinear sufficient dimension reduction

In this section, we introduce the framework for nonlinear SDR developed by Lee *et al.* (2013) and Li and Song (2017). Let $(\Omega, \mathcal{F}, P)$ be probability space and let $(\Omega_X, \mathcal{F}_X)$, $(\Omega_Y, \mathcal{F}_Y)$, and $(\Omega_{XY}, \mathcal{F}_{XY})$ be measurable spaces, where $\Omega_{XY} = \Omega_X \times \Omega_Y$ and $\mathcal{F}_{XY} = \mathcal{F}_X \times \mathcal{F}_Y$. Let X, Y and (X, Y) be random elements that take in Ω_X, Ω_Y and Ω_{XY}, with distributions P_X, P_Y and P_{XY}. It is assumed that these distributions are dominated by σ-finite measures. Let σ-fields generated by random elements as $\sigma(X) = X^{-1}(\mathcal{F}_X)$,

$\sigma(Y) = Y^{-1}(\mathcal{F}_Y)$, and $\sigma(X,Y) = (X,Y)^{-1}(\mathcal{F}_{XY})$. Let $P_{X|Y}(\cdot|\cdot) : \mathcal{F}_X \times \Omega_Y \to \mathbb{R}$ be the conditional distribution of X given Y.

Definition 1. *A sub σ-field $\mathcal{G}$ of Ω_X is an SDR σ-field for Y versus X if it satisfies*

$$Y \perp\!\!\!\perp X \mid \mathcal{G}. \tag{3.1}$$

There are many SDR σ-field for Y versus X. Obviously, $\sigma(X)$ satisfies (3.1), which means that there is no reduction. For this reason, it is natural to seek the minimal SDR σ-field. In Lee *et al.* (2013), it is shown that if the family of probability measures $\{P_{X|Y}(\cdot|y) : y \in \Omega_Y\}$ is dominated by a finite σ-finite measure, then there is a unique sub σ-field $\mathcal{G}^*$ of $\sigma(X)$ such that:

1. $Y \perp\!\!\!\perp X \mid \mathcal{G}^*$;

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

Hence, it is natural to define the minimal SDR σ-field to $\mathcal{G}^*$, and we call $\mathcal{G}^*$ the central σ-field for Y versus X, and denoit it by $\mathcal{G}_{Y|X}$.

For further theoretical developments, a mild condition of square integrability is assumed. 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. In the context of SDR, a difference up to constants is irrelevant. Therefore, there is no problem to assume that all functions of $L_2(P_{XY})$, $L_2(P_X)$, and $L_2(P_Y)$ have zero mean. From now on, let $L_2'(P_X)$, $L_2'(P_Y)$, and $L_2'(P_{XY})$ denote centered space for $L_2(P_X)$, $L_2(P_Y)$, and $L_2(P_{XY})$, respectively. Since a σ-field is rather abstract concept, it is difficult to estimate it directly. Instead, we utilize a class of functions measurable to a target σ-field. Given a sub σ-filed $\mathcal{G}$ of $\sigma(X,Y)$, let $\mathfrak{M}_{\mathcal{G}}$ be the class of $\mathcal{G}$-measurable functions in $L_2'(P_{XY})$. For simplicity, we abbreviate $\mathfrak{M}_{\sigma(X)}$ to $\mathfrak{M}_X$. It is easily checked that $\mathfrak{M}_{\mathcal{G}}$ is a subspace of $L_2'(P_{XY})$. Now, we introduce a target class of functions in the nonlinear SDR.

Definition 2. *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. $\mathfrak{M}_{\mathcal{G}_{Y|X}}$ is denoted by $\mathfrak{S}_{Y|X}$.*

The central class $\mathfrak{S}_{Y|X}$ is the generalization of the central subspace in the classical linear SDR. In the linear SDR, we model

$$Y \perp\!\!\!\perp X \mid \beta^{\mathsf{T}} X, \tag{3.2}$$

where X and Y are random elements in $\mathbb{R}^p$ and $\mathbb{R}$ with square-integrability, and β is a matrix in $\mathbb{R}^{p\times d}$. For β satisfying the condition (3.2), the corresponding SDR σ-field $\mathcal{G}_\beta$ is $\sigma(\beta^{\mathsf{T}}X)$, and the central σ-field $\mathcal{G}_{Y|X}$ is $\cap\{\mathcal{G}_\beta : \beta \text{ satisfies the condition (3.2)}\}$. With this setup, we can represent the central class $\mathfrak{S}_{Y|X}$ as $\{f_\beta : \beta \in \mathcal{S}_{Y|X}\}$, where $f_\beta : \mathbb{R}^p \to \mathbb{R}$ by $x \mapsto \beta^{\mathsf{T}}x$ and $\mathcal{S}_{Y|X}$ is the central subspace of linear SDR.

As linear SDR, we can define the concept of unbiasedness and exhaustiveness for the nonlinear case.

Definition 3. *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 $\mathcal{G}_{Y|X}$.*

For the description of what classes of functions are unbiased, we introduce the operation $\ominus$. For two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ of a generic Hilbert space $\mathcal{H}$, $\mathcal{S}_1 \ominus \mathcal{S}_2$ denote the subspace $\mathcal{S}_1 \cap \mathcal{S}_2^{\perp}$. Now, we specify an unbiased class of functions, which is called regression class.

Proposition 1. *(Theorem 2 of Lee et al. (2013)) If the family* $\{P_{X|Y}(\cdot|y) : y \in \Omega_Y\}$ *is dominated by a* σ*-finite measure, then*

$$L_2'(P_X) \ominus [L_2'(P_X) \ominus L_2'(P_Y)] \subseteq \mathfrak{S}_{Y|X}. \tag{3.3}$$

If some function $f \in L_2'(P_X) \ominus L_2'(P_Y)$, then $f(X)$ is square-integrable, and $\mathrm{cov}(f(X), g(Y)) = 0$ for any $g \in L_2'(P_Y)$, which implies that $E(f(X) \mid Y) = 0$. It resembles the residual in a regression problem. Hence, we call $L_2'(P_X) \ominus [L_2'(P_X) \ominus L_2'(P_Y)]$ regression class and denote it by $\mathfrak{C}_{Y|X}$.

Naturally, the next topic is what condition the regression class is exhaustive for the central class $\mathfrak{S}_{Y|X}$. To this end, we introduce the notion of complete classes of functions in $L_2(P_X)$.

Definition 4. *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* $g \in \mathfrak{M}_{\mathcal{G}}$, $E(g(X) \mid Y) = 0$ *a.s. P implies that* $g(X) = 0$ *a.s. P.*

The completeness is satisfied in many situations that are common settings in SDR. For the details, see Lee *et al.* (2013). In classical statistics, a sufficient and complete statistic is the minimal sufficient statistic. A similar logic can be applied to the central class.

Proposition 2. *(Theorem 3 of Lee et al. (2013)) Suppose* $\{P_{X|Y}(\cdot|y) : y \in \Omega_Y\}$ *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*

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

Thus, we can conclude that, with completeness, the regression class is identical to the central class, and even without completeness, it is guaranteed that the regression class is at least included in the central class. Moreover, it turns out that the regression class can be specified as the range of some bounded linear operators. Through this representation, we can apply a simple spectral decomposition method for estimating the regression class; the generalized sliced inverse regression.

3.1. Generalized SIR - GSIR

Before dealing with GSIR, let us introduce necessary concepts; modulo constants. For A and B, we say that $A \subseteq B$ modulo constants if, for each $f \in A$, there is $c \in \mathbb{R}$ such that $f + c \in B$, and 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 on Ω_X and Ω_Y satisfying the condition:

1. $\mathcal{H}_X$ and $\mathcal{H}_Y$ are dense subsets of $L_2'(P_X)$ and $L_2'(P_Y)$ modulo constants, respectively.
2. There are constants $C_1 > 0$ and $C_2 > 0$ such that $\mathrm{var}(f(X)) \le C_1 \|f\|^2_{\mathcal{H}_X}$ and $\mathrm{var}(g(Y)) \le C_2 \|g\|^2_{\mathcal{H}_Y}$ for any $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$.

In fact, condition 2 has a relationship with the boundedness of bilinear operators. Hence, we assign symbols to the class of bounded operators. 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 $\mathrm{ran}(A)$, the kernel of A by $\ker(A)$, and the closure of $\mathrm{ran}(A)$ by $\overline{\mathrm{ran}}(A)$. Since the bilinear operator $c : \mathcal{H}_X \times \mathcal{H}_X \to \mathbb{R}$ defined by $f, g \mapsto \mathrm{cov}(f(X), g(X))$ is bounded, there is the unique operator $M_{XX} \in \mathcal{B}(\mathcal{H}_X)$ such that $\langle f, M_{XX} g \rangle_{\mathcal{H}_X} = \mathrm{cov}(f(X), g(X))$. Similarly, there is the unique operator $M_{YY} \in \mathcal{B}(\mathcal{H}_Y)$ such that $\langle f, M_{YY} g \rangle_{\mathcal{H}_Y} = \mathrm{cov}(f(Y), g(Y))$. From now

on, let $\mathcal{G}_X$ and $\mathcal{G}_Y$ be $\overline{\text{ran}}(M_{XX})$ and $\overline{\text{ran}}(M_{YY})$, respectively. In considering dependency, a difference up to constants can be negligible, and if $f \in \ker(M_{XX})$, then $\text{var}(f(X)) = 0$, which means that f is a constant function a.s. Hence, we restrict considered spaces to $\mathcal{G}_X$ and $\mathcal{G}_Y$. In fact, it is easily shown that $\mathcal{G}_X = L_2'(P_X)$ and $\mathcal{G}_Y = L_2'(P_Y)$. Specifically, $f \in L_2'(P_X)^\perp$ implies that $\text{cov}(f(X), g(X)) = 0$ for any $g \in L_2'(P_X)$, which is equivalent to $\text{var}(f(X)) = 0$. Moreover, $\text{var}(f(X)) = 0$ is equivalent to $f \in \ker(M_{XX}) = \mathcal{G}_X^\perp$. Thus, it follows that $\mathcal{G}_X = L_2'(P_X)$. Similarly, we can derive that $\mathcal{G}_Y = L_2'(P_Y)$. Since $L_2'(P_X) = \mathcal{G}_X \subseteq \mathcal{H}_X$ and $L_2'(P_Y) = \mathcal{G}_Y \subseteq \mathcal{H}_Y$, we can use inner products of $\mathcal{H}_X$ and $\mathcal{H}_Y$ for all functions in $L_2(P_X)$ and $L_2(P_Y)$. On the other hand, there is a unique cross-covariance operator $M_{YX} : \mathcal{H}_X \to \mathcal{H}_Y$ such that $\text{cov}(f(Y), g(X)) = \langle f, M_{YX} g \rangle_{\mathcal{H}_Y}$ for any $f \in \mathcal{H}_Y$ and $g \in \mathcal{H}_X$, and it is represented as $M_{YX} = M_{YY}^{1/2} R_{YX} M_{XX}^{1/2}$, where R_{YX} is called the correlation operator with properties of $\|R_{YX}\|_{op} \leq 1$ and $\text{ran}(R_{YX}) = \overline{\text{ran}}(M_{YY})$ (Baker, 1973). By Li (2018), we have $\text{ran}(M_{YX}) \subseteq \overline{\text{ran}}(M_{YY}) = L_2'(P_Y)$. Hence, we can define domain-restricted covariance operators.

Definition 5. *We define the restricted covariance operators $\Sigma_{XX} : L_2'(P_X) \to L_2'(P_X)$, $\Sigma_{YY} : L_2'(P_Y) \to L_2'(P_Y)$, and $\Sigma_{YX} : L_2'(P_Y) \to L_2'(P_X)$ to be identical to M_{XX}, M_{YY}, and M_{YX} on the restricted space.*

$$\langle f_1, \Sigma_{XX} f_2 \rangle_{\mathcal{H}_X} = \langle f_1, M_{XX} f_2 \rangle_{\mathcal{H}_X}, \quad \langle g_1, \Sigma_{YY} g_2 \rangle_{\mathcal{H}_Y} = \langle g_1, M_{YY} g_2 \rangle_{\mathcal{H}_Y}, \quad \langle g_1, \Sigma_{YX} f_1 \rangle_{\mathcal{H}_Y} = \langle g_1, M_{YX} f_1 \rangle_{\mathcal{H}_Y}.$$

By the definitions of Σ_{XX} and Σ_{YY}, it is easily shown that $\ker(\Sigma_{XX}) = \{0\}$ and $\ker(\Sigma_{YY}) = \{0\}$, which means that both operators are invertible. Using the inverse operators, we define the conditional expectation operator.

Definition 6. *We call the operator $\Sigma_{YY}^{-1}\Sigma_{YX} : L_2'(P_X) \to L_2'(P_Y)$ the conditional expectation operator, and denote it by $E_{X|Y}$.*

The reason why this operator is called the conditional expectation operator is as follows.

1. For any $f \in L_2'(P_X)$, $(E_{X|Y} f)(Y) = E(f(X)|Y)$.

2. For any $g \in L_2'(P_Y)$, $(E_{X|Y}^* g)(X) = E(g(Y)|X)$.

That is because, for any $g \in L_2'(P_Y)$,

$$\text{cov}\left((E_{X|Y} f)(Y), g(Y)\right) = \langle E_{X|Y} f, \Sigma_{YY} g \rangle_{\mathcal{H}_Y} = \langle \Sigma_{YX} f, g \rangle_{\mathcal{H}_Y} = \text{cov}\left(f(X), g(Y)\right),$$

which means that $(E_{X|Y} f)(Y) = E(f(X) \,|\, Y) + \text{constant}$. Since $E_{X|Y} f \in L_2'(P_Y)$, the constant should be equal to $-E(E(f(X) \,|\, Y)) = -E(f(X)) = 0$. Thus, it follows that $(E_{X|Y} f)(Y) = E(f(X) \,|\, Y)$. With this fact, we can easily induce that $(E_{X|Y}^* g)(X) = E(g(Y) \,|\, X)$.

Proposition 3. *(Corollary 1 of Lee et al. (2013)) For any $f, g \in L_2'(P_X)$,*

$$\langle g, E_{X|Y}^* E_{X|Y} f \rangle_{L_2(P_Y)} = cov\left[E(g(X) \,|\, Y), E(f(X) \,|\, Y)\right]. \tag{3.4}$$

Furthermore, $\|E_{X|Y}^* E_{X|Y}\|_{op} \leq 1$, thus $E_{X|Y}^* E_{X|Y} \in \mathcal{B}(L_2(P_Y))$.

Proof: (3.4) can be proved easily.

$$\langle g, E_{X|Y}^* E_{X|Y} f \rangle_{L_2(P_Y)} = \langle E_{X|Y} g, E_{X|Y} f \rangle_{L_2(P_Y)} = \text{cov}\left[E(g(X) \,|\, Y), E(f(X) \,|\, Y)\right].$$

The right-hand side of (3.4) is greater than equal to $\text{var}(E(g(X) \,|\, Y))^{1/2}\, \text{var}(E(f(X) \,|\, Y))^{1/2}$ by Cauchy-Schwarz inequality. In addition, it is greater than equal to $\text{var}(g(X))^{1/2} \text{var}(f(X))^{1/2} = \|g\|_{L_2'(P_Y)} \|f\|_{L_2'(P_Y)}$. Hence, it follows that $\langle g, E_{X|Y}^* E_{X|Y} f \rangle_{L_2'(P_Y)} \leq \|g\|_{L_2'(P_Y)} \|f\|_{L_2'(P_Y)}$, which means that $\|E_{X|Y}^* E_{X|Y}\|_{op} \leq 1$. □

In linear SDR, $\mathrm{cov}[E(X \mid Y)]$ has a crucial role. Likewise, the operator $E^*_{X|Y}E_{X|Y}$ has a key relationship with the central class $\mathfrak{S}_{Y|X}$.

Proposition 4. *(Theorem 5 of Lee et al. (2013)) If $\mathfrak{S}_{Y|X}$ is complete, then*

$$\overline{\mathrm{ran}}\left(E^*_{X|Y}E_{X|Y}\right) = \mathfrak{S}_{Y|X}.$$

Proof: $f \in \mathfrak{C}^{\perp}$ if and only if $f \in L'_2(P_X)$ and $E(f(X) \mid Y) = 0$, which is equivalent to $f \in \ker(E_{X|Y})$. It follows that $\ker(E_{X|Y})) = \mathfrak{C}^{\perp}_{Y|X}$. Since $\ker(E^*_{X|Y}E_{X|Y}) = \ker(E_{X|Y})$, we have $\overline{\mathrm{ran}}(E^*_{X|Y}E_{X|Y}) = C_{Y|X}$. Considering the completeness of $\mathfrak{S}_{Y|X}$, $\overline{\mathrm{ran}}(E^*_{X|Y}E_{X|Y}) = \mathfrak{S}_{Y|X}$. □

3.2. Generalized SAVE - GSAVE

Without completeness, the GSIR is guaranteed to be unbiased. Lee *et al.* (2013) proposed the GSAVE that it can recover functions beyond the regression class. In the GSAVE, the working space is not restricted to centered space. Thus, we work on $L_2(P_X), L_2(P_X)$. Define the noncentered conditional mean operator $E'_{X|Y} : L_2(P_X) \to L_2(P_Y)$ by, for any $f \in L_2(P_X), g \in L_2(P_Y)$,

$$\langle g, E'_{X|Y}f\rangle_{L'_2(P_X)} = E\left(g(Y)f(X)\right). \tag{3.5}$$

From this definition, it is easily shown that $(E'_{X|Y}f)(Y) = E(f(X)|Y)$. Additionally, we introduce a new type of conditional variance operator.

Definition 7. *For each $y \in \Omega_Y$, the bilinear form*

$$L_2(P_X) \times L_2(P_X) \to \mathbb{R}, \quad (f, g) \mapsto \left[E'_{X|Y}(fg) - E'_{X|Y}fE'_{X|Y}g\right](y)$$

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

$$V_{X|Y} : \Omega_Y \to \mathcal{B}\left(L_2(P_Y)\right), \quad y \mapsto V_{X|Y}(y)$$

the heteroscedastic conditional variance operator given Y.

In summary, there exists uniquely as operator $V_{X|Y}(Y)$ such that $\mathrm{cov}(f(X), g(X) \mid Y) = \langle f, V_{X|Y}(Y)g\rangle_{L_2(P_X)}$. Let the operator S be $E(V - V_{X|Y})^2 : L_2(P_X) \to L_2(P_X)$, where $V : L_2(P_X) \to L_2(P_X)$ is the (unconditional) covariance operator as

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

Now, the reason the GSAVE works is provided.

Proposition 5. *(Theorem 7 of Lee et al. (2013)) Suppose* $\mathrm{var}\left(f(X) \mid \mathcal{G}_{Y|X}\right)$ *is nonrandom for any $f \in \mathfrak{S}^{\perp}_{Y|X}$. Then $\overline{\mathrm{ran}}S \subseteq \mathfrak{S}_{Y|X}$.*

The estimator derived from $\overline{\mathrm{ran}}S$ is called generalized SAVE or GSAVE. The notable result is that GSAVE can cover functions outside the regression class.

Proposition 6. *(Theorem 8 of Lee et al. (2013)) Suppose* $\mathrm{var}\left(f(X) \mid \mathcal{G}_{Y|X}\right)$ *is nonrandom for any $f \in \mathfrak{S}^{\perp}_{Y|X}$, then $\mathfrak{C}_{Y|X} \subseteq \overline{\mathrm{ran}}S$.*

Combining given propositions, the following relation is satisfied:

$$\mathfrak{C}_{Y|X} \subseteq \overline{\mathrm{ran}}S \subseteq \mathfrak{S}_{Y|X}.$$

4. Nonlinear SDR with a scalar response

In this section, we illustrate the extension of sliced inverse regression with the kernel trick introduced in Section 2. The kernel trick is applied to the predictor space, then use the relation between Y and $\kappa(\cdot, X) \in \mathfrak{M}_X$.

4.1. Kernel sliced inverse regression

Let $\Omega_Y \subseteq \mathbb{R}$ and $\Omega_X \subseteq \mathbb{R}^p$. The kernel sliced inverse regression (KSIR) applies the reproducing kernel Hilbert space (RKHS) to SDR method. Let $\mathcal{H}_X$ be a RKHS generated by a positive kernel κ_X. The KSIR assumes the population model:

$$Y \perp\!\!\!\perp X \mid h_1(X), \ldots, h_d(X),$$

where $\{h_1, \ldots, h_d\} \subset \mathcal{H}_X$. Thus, the goal of KSIR is to estimate $\{h_1, \ldots, h_d\}$, which is called e.d.r. directions.

Definition 8. *If $Y \perp\!\!\!\perp X \mid h_1(X), \ldots, h_d(X)$, then $\{h_1, \ldots, h_d\}$ is called the feature. e.d.r. directions, and the subspace $\mathcal{H} = \operatorname{span}(h_1, \ldots, h_d)$ is called the feature e.d.r. subspace.*

Note that, by the reproducing property of kernel function $h_i(X) = \langle h_i, \kappa(\cdot, X)\rangle_{\mathcal{H}_X}$, the model is represented as:

$$Y \perp\!\!\!\perp X \mid \langle h_1, \kappa(\cdot, X)\rangle_{\mathcal{H}_X}, \ldots, \langle h_d, \kappa(\cdot, X)\rangle_{\mathcal{H}_X},$$

which implies that KSIR is an extension of linear SDR in that the inner product in $\mathbb{R}^p$ is changed to that of $\mathcal{H}_X$. For further discussion, we assume that $E(\kappa_X(X, X)) < \infty$ to be followed:

1. There is a unique $E(\kappa_X(\cdot, X)) \in \mathcal{H}_X$ such that, for any $f \in \mathcal{H}_X$, $E(f(X)) = \langle f, E(\kappa_X(\cdot, X))\rangle_{\mathcal{H}}$.
2. There is a unique bounded linear operator $\Sigma_{XX} \in \mathcal{B}(\mathcal{H}_X)$ such that $\operatorname{cov}(f(X), g(X)) = \langle f, \Sigma_{XX} g\rangle_{\mathcal{H}_X}$ for any $f, g \in \mathcal{H}_X$.

For the success of KSIR, there is a sufficiency condition proposed by Yeh *et al.* (2008), and this condition is also just an extension of the linear conditional mean assumption in linear SDR.

Definition 9. *For $\{h_1, \ldots, h_p\} \subset \mathcal{H}_X$, it is said to satisfy the linear design condition, if, for any $f \in \mathcal{H}_X$,*

$$E\left(f(X) \mid h_1(X), \ldots, h_d(X)\right) = c_0 + c_1 h_1(X) + \cdots + c_d h_d(X),$$

for some constants $c_1, c_1, \ldots, c_p$.

With this condition, the unbiasedness that is different from ours can be proved in the population level.

Proposition 7. *(Theorem 1 of Yeh et al. (2008)) Assume the existence of a feature e.d.r. subspace $\operatorname{span}(h_1, \ldots, h_d)$, and the linear design condition holds. Then the central inverse regression vector falls into the subspace $\operatorname{span}(\Sigma_{XX} h_1, \ldots, \Sigma_{XX} h_d)$; $E(\kappa_X(\cdot, X) \mid Y) - E(\kappa_X(\cdot, X)) \in \operatorname{span}(\Sigma_{XX} h_1, \ldots, \Sigma_{XX} h_d)$.*

In a real application, it is difficult to deal with $E(\kappa_X(\cdot, X) \mid Y)$ directly. Instead, the discretized version of Y, $E(\kappa_X(\cdot, X) \mid Y \in J_j)$, is adopted, where $J_1, \ldots, J_h$ is a partition of Ω_Y with $P(Y \in J_j) > 0$ for all $j = 1, \ldots, h$. It can be shown that $E(\kappa_X(\cdot, X) \mid \tilde{Y}) - E(\kappa_X(\cdot, X)) \in \operatorname{span}(\Sigma_{XX} h_1, \ldots, \Sigma_{XX} h_d)$,

where $\tilde{Y} = \sum_{j=1}^{h} jI(Y \in J_j)$. It follows that the closure of the range of the covariance operator for $E(\kappa_X(\cdot, X) \mid \tilde{Y})$ is included in $\text{span}(\Sigma_{XX}h_1, \ldots, \Sigma_{XX}h_d)$. Now we can construct the covariance operator of $E(K_X(\cdot, X) \mid \tilde{Y})$ as

$$\Sigma_{E(K_X(\cdot,X)\mid\tilde{Y})} = \sum_{j=1}^{h} P(Y \in J_j)\left[E\left(K_X(\cdot, X) \mid Y \in J_j\right) - E\left(\kappa_X(\cdot, X)\right)\right] \otimes \left[E\left(K_X(\cdot, X) \mid Y \in J_j\right) - E\left(\kappa_X(\cdot, X)\right)\right].$$

Therefore, it is necessary to solve the eigenvalue problem, where k^{th} eigenvector is obtained by

$$\max_{a \in \mathcal{H}_X} \langle a\,,\, \Sigma_{E(K_X(\cdot,X)\mid\tilde{Y})}\, a\rangle_{\mathcal{H}_X}$$

subject to $\langle a, \Sigma_{XX}a\rangle_{\mathcal{H}_X} = 1$, $\langle a, \Sigma_{XX}a^j\rangle_{\mathcal{H}_X} = 0$ for $j = 1, \ldots, k-1$. Let us include KSiR into the framework proposed by Lee *et al.* (2013). Let $\text{span}(h_1, \ldots, h_d)$ be the central subspace $\mathcal{S}_{Y|X}$. Then, it follows that the central σ-field $\mathcal{G}_{Y|X}$ is $\sigma(h_1(X), \ldots, h_d(X))$. Hence the central subspace which considers only linear combinations is the subset of the central class $\mathfrak{S}_{Y|X}$. In other words, we have $S_{Y|X} \subseteq \mathfrak{S}_{Y|X}$. Since any measurable function g, $\mathcal{S}_{g(Y)|X} \subseteq \mathcal{S}_{Y|X}$, hence, it follows that $\mathcal{S}_{g(Y)|X} \subseteq \mathfrak{S}_{Y|X}$. Therefore, it can be concluded that the KSIR is the unbiased method at the population level, however, it cannot be exhaustive for $\mathfrak{S}_{Y|X}$.

5. Nonlinear SDR in Euclidean space

When the response variable is not a real random variable, slicing the response becomes inapplicable. Moreover, as the data structure grows in complexity, there is a need for generalized theories applicable to Hilbert space random elements. GSIR and GSAVE, introduced in Section 3, represent nonlinear SDR methods within a generic Hilbert space. In this section, we introduce another type of nonlinear method, drawing on the concept of canonical correlation analysis. We will briefly review Kernel Canonical Correlation Analysis (KCCA), as detailed in Fukumizu *et al.* (2007).

Let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be the reproducing kernel Hilbert spaces generated by positive definite kernels κ_X and κ_Y. For the existence of mean element and covariance operator, it is assumed that $E(\kappa_X(X, X)) < \infty$ and $E(\kappa_Y(Y, Y)) < \infty$. Then, we have, $E(f^2(X)) = E(\langle f, \kappa_X(\cdot, X)\rangle^2_{\mathcal{H}_X}) \leq \|f\|^2_{\mathcal{H}_X} E(\kappa_X(X, X)) < \infty$ for any $f \in \mathcal{H}_X$. Similarly, we can show that for any $g \in \mathcal{H}_Y$, $E(g^2(Y))$ is also finite. More explicitly, there exist uniquely covariance operators M_{XX}, M_{YY}, and M_{XY} such that $\text{cov}(f(X), g(X)) = \langle f, M_{XX}g\rangle_{\mathcal{H}_X}$, $\text{cov}(f(Y), g(Y)) = \langle f, M_{YY}g\rangle_{\mathcal{H}_Y}$, and $\text{cov}(f(X), g(Y)) = \langle f, M_{XY}g\rangle_{\mathcal{H}_X}$. For the connecting zero-mean and square-integrable space $L_2(P_X)$ and $L_2(P_Y)$, $\mathcal{H}_X$ and $\mathcal{H}_Y$ are reset to $\overline{\text{ran}}M_{XX}$ and $\overline{\text{ran}}M_{YY}$. Hence, it follows that $\mathcal{H}_X \subseteq L_2(P_X)$ and $\mathcal{H}_Y \subseteq L_2(P_Y)$, which implies that $\Sigma_{XX}, \Sigma_{YY}, \Sigma_{XY}$ on centered spaces are corresponding to M_{XX}, M_{YY}, M_{XY} on original spaces. With this setup, KCCA seeks to solve the following problem:

$$\max_{f \in \mathcal{H}_X, g \in \mathcal{H}_Y, f \neq 0, g \neq 0} \frac{\text{cov}(f(X), g(Y))}{\text{var}(f(x))^{1/2}\text{var}(g(Y))^{1/2}}. \tag{5.1}$$

In the population level, (5.1) is reformulated by RKHS form as

$$\max_{f \in \mathcal{H}_X, g \in \mathcal{H}_Y, f \neq 0, g \neq 0} \frac{\langle f, \Sigma_{XY}g\rangle_{\mathcal{H}_X}}{\langle f, \Sigma_{XX}f\rangle^{1/2}_{\mathcal{H}_X}\langle g, \Sigma_{YY}g\rangle^{1/2}_{\mathcal{H}_Y}}, \tag{5.2}$$

or equivalently,

$$\sup_{f\in\mathcal{H}_X, g\in\mathcal{H}_Y} \langle f, \Sigma_{XY} g\rangle_{\mathcal{H}_X} \text{ subject to } \langle f, \Sigma_{XX} f\rangle_{\mathcal{H}_X} = 1, \langle g, \Sigma_{YY} g\rangle_{\mathcal{H}_Y} = 1. \tag{5.3}$$

As with classical CCA, the solution of 5.3 is ven by the eigenfunctions corresponding to the largest eigenvalue of the following generalized eigenvalue problem:

$$\Sigma_{YX} f = \lambda_1 \Sigma_{YY} g, \quad \Sigma_{XY} g = \lambda_1 \Sigma_{XX} f. \tag{5.4}$$

As mentioned, Σ_{XY} can be represented as $\Sigma_{XY} = \Sigma_{XX}^{1/2} V_{XY} \Sigma_{YY}^{1/2}$, where V_{XY} is the correlation operator with its norm being less than equal to 1. Let ϕ, ψ be the unit eigenfunctions of V_{XY} corresponding to the largest singular value; that is,

$$\langle \phi, V_{XY}\psi\rangle = \max_{\|f\|_{\mathcal{H}_X}=1, \|g\|_{\mathcal{H}_Y}=1} \langle f, V_{XY} g\rangle_{\mathcal{H}_X}.$$

From the equation 5.4, it is easy to see that the solution of the KCCA is given the inverse images

$$f = \Sigma_{XX}^{-1/2}\phi, \quad g = \Sigma_{YY}^{-1/2}\psi.$$

At the sample level, we briefly introduce the estimating procedure. At first, let ϕ_n, ψ_n be the unit eigenfunctions corresponding to the largest singular value of the operator

$$V_{n,XY} = \left(\Sigma_{n,XX} + \epsilon_n I\right)^{-\frac{1}{2}} \Sigma_{n,XY} \left(\Sigma_{n,YY} + \epsilon_n I\right)^{-\frac{1}{2}}.$$

The empirical estimators f_n and g_n of KCCA are

$$f_n = \left(\Sigma_{n,XX} + \epsilon_n I\right)^{-\frac{1}{2}} \phi_n, \quad g_n = \left(\Sigma_{n,YY} + \epsilon_n I\right)^{-\frac{1}{2}} \psi_n.$$

In the practical analysis, when the inverse of an operator is treated, $\epsilon_n I$ is added to make the operator nonsingular. Moreover, this term acts like a regularize that a result function can be smooth. The authors of Fukumizu *et al.* (2007) shows convergence of f_n, g_n under the condition for the speed of ϵ_n. The KCCA can be interpreted in terms of conditional expectation operator $E_{X|Y} = \Sigma_{YY}^{-1}\Sigma_{YX}$. In fact, the solution of KCCA f, g has the following relationship.

$$\sup_{\|v\|=\|u\|=1} \langle u, V_{XY} u\rangle_{\mathcal{H}_X} = \langle \phi, V_{XY}\psi\rangle_{\mathcal{H}_X} = \langle \Sigma_{XX}^{1/2} f, V_{XY}\Sigma_{YY}^{1/2} g\rangle_{\mathcal{H}_X} = \langle f, \Sigma_{XX}^{-1}\Sigma_{XY} g\rangle_{\mathcal{H}_X} = \langle f, E_{X|Y}^* g\rangle_{\mathcal{H}_X}.$$

It follows that f, g is the solution of the following problem:

$$\max \langle E_{X|Y} u, v\rangle_{\mathcal{H}_Y} \text{ subject to } \langle u, \Sigma_{XX} u\rangle_{\mathcal{H}_X} = \langle v, \Sigma_{YY} v\rangle_{\mathcal{H}_Y} = 1.$$

Since $v = E_{X|Y} v, \langle u, \Sigma_{XX} u\rangle_{\mathcal{H}_X} = \langle u, u\rangle_{L_2'(P_X)}$, and $\langle v, \Sigma_{YY} v\rangle_{\mathcal{H}_Y} = \langle v, v\rangle_{L_2'(P_Y)}$, the above problem is reformulated as

$$\max \langle u, E_{X|Y}^* E_{X|Y} v\rangle_{\mathcal{H}_Y} \text{ subject to } \langle u, u\rangle_{L_2'(P_X)} = \langle v, v\rangle_{L_2'(P_Y)} = 1.$$

To achieve the maximization, u must be in $\text{ran}(E_{X|Y}^* E_{X|Y})$. Consequently, $f \in \text{ran}(E_{X|Y}^* E_{X|Y}) \subseteq \mathfrak{S}_{Y|X}$.

6. Nonlinear SDR for functional data

This section covers the nonlinear SDR in the case that both Y and X are random functions. We introduce the functional GSIR (f-GSIR) developed by Li and Song (2017) briefly. Let T_X and T_Y be subsets in $\mathbb{R}^{k_1}$ and $\mathbb{R}^{k_2}$. Let Ω_X be a Hilbert space of functions from T_X to $\mathbb{R}^p$ and Ω_Y be a Hilbert space of functions from T_Y to $\mathbb{R}^q$. Note that Ω_X and Ω_Y are both Hilbert spaces of functions. Since a similar procedure is applied to the Y case, we describe only the X case. For a $f \in \Omega_X$, we can represent f as a vector of functions: $f = (f_1, \ldots, f_p)$, where f_i is a function from T_X to $\mathbb{R}$. For simplicity, we assume that f_i is a random function in $\mathcal{H}^0$ which is a Hilbert space of function from T_X to $\mathbb{R}$ for all $i = 1, \ldots, p$. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}^0}$ is a inner product of $\mathcal{H}^0$. We define $\mathcal{H}$ as p Cartesian product space of $\mathcal{H}^0$, that is, $\mathcal{H} = \mathcal{H}^0 \times \cdots \times \mathcal{H}^0$. For $f, g \in \mathcal{H}$, the inner product of $\mathcal{H}$ is defined as $\langle f, g \rangle_{\mathcal{H}} = \sum_{j=1}^{p} \langle f_j, g_j \rangle_{\mathcal{H}^0}$. Note that $\mathcal{H}^0$ can be constructed by RKHS with kernel function $\kappa_{T_X} : T_X \times T_X \to \mathbb{R}$, hence, we call Ω_X the first level RKHS. Naturally, we call $\mathcal{H}_X$ the second level RKHS. Li and Song (2017) constrains the kernel function κ_X to be connected with the inner product of $\mathcal{H}$.

Definition 10. *We say that a positive definite kernel $\kappa : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is induced by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ if there is a function $\rho : \mathbb{R}^3 \to \mathbb{R}^+$ such that, for any $\phi, \psi \in \mathcal{H}$,*

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

We assume that κ_X is induced by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and assume that $\mathcal{H}_X$. In the same manner, κ_Y and $\mathcal{H}_Y$ can be constructed. We additionally assume that

1. $\mathcal{H}_X$ and $\mathcal{H}_Y$ are dense subsets of $L_2(P_X)$ and $L_2(P_Y)$ modulo constants, respectively.
2. There are constants $C_1 > 0$ and $C_2 > 0$ such that $\mathrm{var}(f(X)) \leq C_1 \|f\|^2_{\mathcal{H}_X}$ and $\mathrm{var}(g(Y)) \leq C_2 \|g\|^2_{\mathcal{H}_Y}$ for any $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$.

Then, there exist a unique element in $\mathcal{H}_X$, $E(\kappa_X(\cdot, X))$, such that $E(f(X)) = \langle f, E(\kappa_X(\cdot, X)) \rangle_{\mathcal{H}_X}$ for any $f \in \mathcal{H}_X$ and a unique element $\mathcal{H}_Y$, $E(\kappa_Y(\cdot, Y))$, such that $E(g(Y)) = \langle g, E(\kappa_Y(\cdot, Y)) \rangle_{\mathcal{H}_Y}$ for any $g \in \mathcal{H}_Y$. Furthermore, there exist uniquely covariance operators M_{XX}, M_{YY}, and M_{XY} such that $\mathrm{cov}(f(X), g(X)) = \langle f, M_{XX} g \rangle_{\mathcal{H}_X}$, $\mathrm{cov}(f(Y), g(Y)) = \langle f, M_{YY} g \rangle_{\mathcal{H}_Y}$, and $\mathrm{cov}(f(X), g(Y)) = \langle f, M_{XY} g \rangle_{\mathcal{H}_X}$. Let $\mathcal{H}^0_X = \overline{\mathrm{span}}\{\kappa_X(\cdot, X) - E(\kappa_X(\cdot, X)) : X \in \mathcal{H}_X\}$, and $\mathcal{H}^0_Y = \overline{\mathrm{span}}\{\kappa_Y(\cdot, Y) - E(\kappa_Y(\cdot, Y)) : Y \in \mathcal{H}_T\}$. Then, it can be shown that $\overline{\mathrm{ran}}(M_{XX}) = \mathcal{H}^0_X$, and $\overline{\mathrm{ran}}(M_{YY}) = \mathcal{H}^0_Y$. We denote the Moore-Penrose inverse of some operator A by $A^\dagger$. We call $M^\dagger_{XX} M_{XY}$ the regression operator and denote it by R_{YX}. The next proposition is a theoretical base of the functional GSIR (f-GSIR).

Proposition 8. *(Theorem 1 of Li and Song (2017)) Under some mild conditions, we have $\overline{\mathrm{ran}}(R^*_{YX}) \subseteq \mathrm{cl}(M_{XX} \Im_{Y|X})$, where cl(·) means the closure. Furthermore, if $\Im_{Y|X}$ is complete, then $\overline{\mathrm{ran}}(R^*_{YX}) = \mathrm{cl}(M_{XX} \Im_{Y|X})$.*

Since $\overline{\mathrm{ran}}(R^*_{YX}) = \overline{\mathrm{ran}}(R^*_{YX} A R_{XY})$ for any invertible operator $A : \mathcal{H}^0_Y \to \mathcal{H}^0_Y$, Li and Song (2017) recommends choosing A to M_{YY} to avoid inversion so that $R^*_{YX} A R_{XY} = M_{XY} M_{YX}$. We call the following entire procedure f-GSIR:

Let $f_1, \ldots, f_d$ be solution to the following sequential eigenvalue problem; for each $k = 1, \ldots, d$,

$$\max_{f \in \mathcal{H}^0_X} \langle f, M^\dagger_{XX} M_{XY} M_{YX} M^\dagger_{XX} f \rangle_{\mathcal{H}_X}$$

subject to $\langle f, f\rangle_{\mathcal{H}_X} = 1$, $\langle f, f_i\rangle_{\mathcal{H}_X} = 0$ for $i = 1, \ldots, k-1$. Corollary 1 of Li and Song (2017) shows that $f_(X), \ldots, f_d(X)$ generate a subspace of $\mathfrak{S}_{Y|X}$, if it is complete, then generate $\mathfrak{S}_{Y|X}$ itself. Note that f-GSIR is an extension of GSIR in that it does not restrict space to $L_2'(P_X)$, or $\mathcal{H}_X^0$ and it applies the RKHS to the construction of Ω_X to overcome the non-dense observation problem for functional data.

7. Implementation

7.1. Coordinate representation

Let $\mathcal{H}_i$ be a finite Hilbert space, $B_i = \{b_1^i, \ldots, b_{m_i}^i\}$ be $\mathcal{H}_i$'s basis system, and G_{B_i} be a gram matrix with $(G_{B_i})_{jk} = \langle b_j^i, b_k^i\rangle_{\mathcal{H}_i}$ for $i = 1, 2, 3$. We first of the coordinate representation of an element. If $f = \sum_j f_j b_j^i \in \mathcal{H}_i$, then the coordinate representation of f, $[f]_{B_i} = [f_1, \ldots, f_{m_i}]^\intercal$. For $A : \mathcal{H}_1 \to \mathcal{H}_2$, we define the coordinate representation of A, ${}_{B_2}[A]_{B_1} = [[Ab_1^1]_{B_2}, \ldots, [Ab_{m_1}^1]_{B_2}]$. It follows that $[Af]_{B_2} = {}_{B_2}[A]_{B_1}[f]_{B_1}$. Let C be an operator $\mathcal{H}_2 \to \mathcal{H}_3$. It can be shown that the coordinate representation AC is ${}_{B_3}[CA]_{B_1} = {}_{B_3}[C]_{B_2}{}_{B_2}[A]_{B_1}$.

Furthermore, we can calculate the inner product as $\langle f, g\rangle_{\mathcal{H}_1} = \sum_i \sum_j f_i g_j \langle b_i^1, b_j^1\rangle_{\mathcal{H}_1} = [f]_{B_1}^\intercal G_{B_1}[g]_{B_1}$. The tensor product can also be represented by the coordinated representation. If $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$, then,

$$\left[(g \otimes f) b_i^1\right]_{B_2} = \langle f, b_i^1\rangle_{\mathcal{H}_1}[g]_{B_2} = [g]_{B_2}[f]_{B_1}^\intercal G_{B_1} e_i,$$

where e_i is the i^{th} standard basis vector of $\mathbb{R}^{m_1}$. From this, we have

$$\begin{aligned} {}_{B_2}[g \otimes f]_{B_1} &= \left[\left[(g \otimes f) b_1^1\right]_{B_2} \ldots \left[(g \otimes f) b_{m_1}^1\right]_{B_2}\right] \\ &= [g]_{B_2}[f]_{B_1}^\intercal G_{B_1}(e_1 \ldots e_{m_1}) \\ &= [g]_{B_2}[f]_{B_1}^\intercal G_{B_1}. \end{aligned}$$

7.2. RKHS by coordinate representation

In this subsection, we introduce the implementation of RKHS, using the coordinate representation. Let $X_1, \ldots, X_n$ be an i.i.d. sample of X. Let κ_X be a positive definite kernel and let $\mathcal{H}_X$ be $\text{span}\{\kappa_X(\cdot, X_i) : i = 1, \ldots, n\}$. We estimate covariance operators at the sample level, by replacing E by the empirical E_n.

$$\hat{\Sigma}_{XX} = E_n\left[\kappa_X(\cdot, X) - E_n\kappa(\cdot, X)\right] \otimes \left[\kappa_X(\cdot, X) - E_n\kappa(\cdot, X)\right].$$

Hence, the subspace $\overline{\text{ran}}(\hat{\Sigma}_{XX})$ is spanned by

$$B_X = \{\kappa_X(\cdot, X_i) - E_n\kappa_X(\cdot, X) : i = 1, \ldots, n\} = \left\{b_1^{(X)}, \ldots, b_n^{(X)}\right\}.$$

First of all, the gram matrix G_{B_X} such that $(G_{B_X})_{ij} = \langle b_i^{(X)}, b_j^{(X)}\rangle_{\mathcal{H}_X}$ is considered.

$$\begin{aligned} \langle b_i^{(X)}, b_j^{(X)}\rangle_{\mathcal{H}_X} &= \langle \kappa_X(\cdot, X_i) - n^{-1}\sum_{k=1}^n \kappa(\cdot, X_k), \kappa_X(\cdot, X_j) - n^{-1}\sum_{\ell=1}^n \kappa_X(\cdot, X_\ell)\rangle_{\mathcal{H}_X} \\ &= \kappa_X\left(X_i, X_j\right) - n^{-1}\sum_{\ell=1}^n \kappa_X\left(X_i, X_\ell\right) - n^{-1}\sum_{k=1}^n \kappa_X\left(X_j, X_k\right) + n^{-2}\sum_{k=1}^n\sum_{\ell=1}^n \kappa_X\left(X_k, X_\ell\right). \end{aligned}$$

Let K_X denote the $n \times n$ matrix whose $(i, j)^{th}$ entry is $\kappa_X(X_i, X_j)$, and let Q denote the projection matrix $I_n - n^{-1}1_n 1_n^\top$. Then it is easily shown that $G_{B_X} = QK_XQ$. Now it is ready to represent $\hat{\Sigma}_{XX}$ by the coordinate representation.

$$\begin{aligned}\left[\hat{\Sigma}_{XX} b_i^{(X)}\right]_{B_X} &= n^{-1}\left[\left(\sum_{i=1}^n b_k^{(X)} \otimes b_k^{(X)}\right) b_i^{(X)}\right] \\ &= n^{-1}\left(\sum_{k=1}^n \left[b_k^{(X)}\right]_{B_X}\left[b_k^{(X)}\right]_{B_X}^\top G_{B_X}\right)\left[b_i^{(X)}\right]_{B_X} \\ &= n^{-1}\left(\sum_{k=1}^n e_k e_k^\top\right) QK_XQe_i = n^{-1}QK_XQe_i.\end{aligned}$$

Hence, it follows that

$$\begin{aligned}{}_{B_X}\left[\hat{\Sigma}_{XX}\right]_{B_X} &= \left[\left[\hat{\Sigma}_{XX} b_1^{(X)}\right]_{B_X} \ldots \left[\hat{\Sigma}_{XX} b_n^{(X)}\right]_{B_X}\right] \\ &= n^{-1}QK_XQ\,(e_1 \ldots e_n) = n^{-1}QK_XQ = n^{-1}G_{B_X}.\end{aligned}$$

The similar development can be carried out on Y. That is, $\mathcal{H}_Y$ is the space spanned by

$$B_Y = \{\kappa_Y(\cdot, Y) - E_n\kappa_Y(\cdot, Y) : i = 1, \ldots n\} = \left\{b_i^{(Y)} : i = 1, \ldots, n\right\}.$$

The cododinate representation of $\hat{\Sigma}_{YY}$ is

$${}_{B_Y}\left[\hat{\Sigma}_{YY}\right]_{B_Y} = n^{-1}QK_YQ = n^{-1}G_{B_Y}.$$

In similar way, we derive the coordinate representation of $\hat{\Sigma}_{XY}$ and $\hat{\Sigma}_{YX}$ as

$${}_{B_X}\left[\hat{\Sigma}_{XY}\right]_{B_Y} = n^{-1}G_{B_Y}, \quad {}_{B_Y}\left[\hat{\Sigma}_{YX}\right]_{B_X} = n^{-1}G_{B_X}.$$

From now on, if the domain and range of an operator are explicit, we abbreviate ${}_{B_1}[A]_{B_2}$ to just $[A]$.

7.3. GSIR

In the context of GSIR, the inner product in $L_2'(P_X)$ is used, which should also be represented by coordinate form. Since $\langle f, g\rangle_{L_2'(P_X)} = \langle f, \Sigma_{XX} g\rangle_{\mathcal{H}_X}$, its coordinate representation is $[f]^\top G_{B_X}[\Sigma_{XX}][g] = n^{-1}[f]^\top G_{B_X}^2[g]$. The conditional expectation operator $E_{X|Y}$ is $\Sigma_{YY}^{-1}\Sigma_{YX}$. The problem is that, although Σ_{YY} is invertible analytically, its coordinate representation would be singular. To avoid singular issue, it is often use $([A] + \epsilon_n I_n)^{-1}$ rather than $[A]^{-1}$ directly. Hence, let the coordinate form of $E_{X|Y}$ be

$$[E_{X|Y}] = \left[\hat{\Sigma}_{YY}\right]^{-1}\left[\hat{\Sigma}_{YX}\right] = \left(G_{B_Y} + \epsilon_n I_n\right)^{-1} G_{B_X}.$$

Consequently, it follows that

$$\begin{aligned}\langle f, \hat{E}_{X|Y}^* \hat{E}_{X|Y} f\rangle_{L_2'(P_X)} &= \langle \hat{E}_{X|Y} f, \hat{E}_{X|Y} f\rangle_{L_2'(P_Y)} \\ &= n^{-1}[f]^\top\left[\hat{E}_{X|Y}\right]^\top G_{B_Y}^2\left[\hat{E}_{X|Y}\right][f] \\ &= n^{-1}[f]^\top G_{B_X}\left(G_{B_Y} + \epsilon_n I_n\right)^{-1} G_{B_Y}^2\left(G_{B_Y} + \epsilon_n I_n\right)^{-1} G_{B_X}[f].\end{aligned}$$

The eigenvalue problem that GSIR solve is as follows

$$\max_{v\in\mathbb{R}^n} n^{-1}v^{\intercal}G_{B_X}\left(G_{B_Y}+\epsilon_n I_n\right)^{-1}G_{B_Y}^2\left(G_{B_Y}+\epsilon_n I_n\right)^{-1}G_{B_X}v, \text{ subject to } \langle v,v\rangle_{L_2'(P_X)} = n^{-1}v^{\intercal}G_{B_X}^2 v = 1.$$

Transforming v to $u = n^{-1/2}G_{B_X}v$, the above problem is changed to

$$\max_{u\in\mathbb{R}^n} u^{\intercal}\left(G_{B_X}+\epsilon_n I_n\right)^{-1}G_{B_X}\left(G_{B_Y}+\epsilon_n I_n\right)^{-1}G_{B_Y}^2\left(G_{B_Y}+\epsilon_n I_n\right)^{-1}G_{B_X}\left(G_{B_X}+\epsilon_n I_n\right)^{-1}u,$$

subject to $\|u\|_{\mathbb{R}^n} = 1$. Let the first d eigenvectors of this problem be $\{u_1,\ldots,u_d\}$, which means that the i^{th} solution of original problem is $v_i = n^{1/2}(G_{B_X}+\epsilon_n I_n)^{-1}u_i$ for $i = 1,\ldots,d$. From this result, i^{th} sufficient predictor is $h^{(i)}(x) = \sum_{j=1}^n v_{ij}\, b_i^{(X)}(x)$ for $i = 1,\ldots,d$.

7.4. KSIR

In the implementation of KSIR, $\Sigma_{E(\kappa_X(\cdot,X)\,|\,\tilde{Y})}$ is used and it has a explicit analytical form.

$$\begin{aligned}\Sigma_{E(K_X(\cdot,X)\,|\,\tilde{Y})} &= \sum_{j=1}^h P\left(Y\in J_j\right)\Sigma_{E(K_X(\cdot,X)\,|\,Y\in J_j)}\\ &= \sum_{j=1}^h P\left(Y\in J_j\right)\left[E\left(K_X(\cdot,X)-E(\kappa_X(\cdot,X))\,|\,Y\in J_j\right)\right]\otimes\left[E\left(K_X(\cdot,X)-E\left(\kappa_X(\cdot,X)\right)\,|\,Y\in J_j\right)\right].\end{aligned}$$

To implement the KSIR, $P(Y\in J_j)$, $E(\kappa_X(\cdot,X))$, and $E(K_X(\cdot,X)-E(\kappa_X(\cdot,X))\,|\,Y\in J_j)$ should be estimated, and their estimates are as follows:

- $\hat{P}(Y\in J_j) = E_n(I(Y\in J_j)) = n^{-1}\sum_{i=1}^n I(Y_i\in J_j) = n^{-1}n_j = \pi_j$
- $\hat{E}(\kappa_X(\cdot,X)) = E_n(\kappa_X(\cdot,X)) = n^{-1}\sum_{i=1}^n \kappa_X(\cdot,X_i)$.
- $\hat{E}(K_X(\cdot,X)-E(\kappa_X(\cdot,X))\,|\,Y\in J_j) = (n_j)^{-1}\sum_{i:y_i\in J_j}(\kappa_X(\cdot,X_i)-E(\kappa_X(\cdot,X)))$

Thus, $\hat{\Sigma}_{E(K_X(\cdot,X)\,|\,Y\in J_j)} = (n_j)^{-2}(\sum_{k:y_k\in J_j} b_k^{(X)})\otimes(\sum_{\ell:y_\ell\in J_j} b_\ell^{(X)}) = (n_j)^{-2}\sum_{k:y_k\in J_j}\sum_{\ell:y_\ell\in J_j} b_k^{(X)}\otimes b_\ell^{(X)}$. Hence, the coordinate representation of it is

$$\left[\hat{\Sigma}_{E(K_X(\cdot,X)\,|\,Y\in J_j)}\right] = \left(n_j\right)^{-2}\sum_{k:y_k\in J_j}\sum_{\ell:y_\ell\in J_j} e_k e_\ell^{\intercal}G_{B_X} = (n_j)^{-2}\left(\sum_{k:y_k\in J_j}e_k\right)\left(\sum_{k:y_k\in J_j}e_k\right)^{\intercal}G_{B_X},$$

from which, it follows that

$$\left[\hat{\Sigma}_{E(K_X(\cdot,X)\,|\,\tilde{Y})}\right] = \left(\sum_{j=1}^h \pi_j\left(n_j\right)^{-2}\left(\sum_{k:y_k\in J_j}e_k\right)\left(\sum_{k:y_k\in J_j}e_k\right)^{\intercal}\right)G_{B_X}.$$

Let A denote $(\sum_{j=1}^h \pi_j (n_j)^{-2}(\sum_{k:y_k\in J_j}e_k)(\sum_{k:y_k\in J_j}e_k)^{\intercal})$. Note that A is symmetric.

The eigenvalue problem of the KSIR is $\max_{v\in\mathcal{H}_X}\langle v,\Sigma_{E(\kappa_X(\cdot,X)|\tilde{Y})}v\rangle_{\mathcal{H}_X}$ subject to $\langle v,\Sigma_{XX}v\rangle_{\mathcal{H}_X} = 1$. The coordinate form of this problem is

$$\max_{[v]\in\mathbb{R}^n}[v]^{\intercal}G_{B_X}AG_{B_X}[v], \text{ subjet to } n^{-1}[v]^{\intercal}G_{B_X}^2[v] = 1.$$

Let $u = n^{-1/2} G_{B_X}[v]$. Then, the above problem is transformed to the below problem:

$$\max_{u \in \mathbb{R}^n} u^\top \left(G_{B_X} + \epsilon_n I_n\right)^{-1} G_{B_X} A G_{B_X} \left(G_{B_X} + \epsilon_n I_n\right)^{-1} u, \text{ subjet to } \|u\|_{\mathbb{R}^n} = 1.$$

Let the first d eigenvector of this problem be $\{u_1, \ldots, u_d\}$, from which we can derive that the i^{th} solution of original problem is $v_i = n^{1/2}(G_{B_X} + \epsilon_n I_n)^{-1} u_i$. Therefore, i^{th} sufficient predictor is $h_i(x) = \sum_{j=1}^n v_{ij}\, b_j^{(X)}(x)$.

7.5. KCCA

(5.3) says that the KCCA solves the generalized singular value problem at the population level.

$$\sup_{f \in \mathcal{H}_X, g \in \mathcal{H}_Y} \langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_X} \text{ subject to } \langle f, \Sigma_{XX} f \rangle_{\mathcal{H}_X} = 1, \langle g, \Sigma_{YY} g \rangle_{\mathcal{H}_Y} = 1.$$

If we represent this problem in coordinate form, we have

$$\sup_{f,g \in \mathbb{R}^n} f^\top G_{B_X} [\Sigma_{XY}]\, g \text{ subject to } f^\top G_{B_X} [\Sigma_{XX}]\, f = g G_{B_Y} [\Sigma_{YY}]\, g = 1. \tag{7.1}$$

By representing covariance operators in form of gram matrix, we have

$$\sup_{f,g \in \mathbb{R}^n} n^{-1} f^\top G_{B_X} G_{B_Y} g \text{ subject to } n^{-1} f^\top G_{B_X}^2 f = n^{-1} g G_{B_Y}^2 g = 1.$$

Let $\phi = n^{-1/2} G_{B_X} f$ and $\psi = n^{-1/2} G_{B_Y} g$. Then, the above statement is reformulated as

$$\sup_{f,g \in \mathbb{R}^n} \phi^\top \left(G_{B_X} + \epsilon_n I_n\right)^{-1} G_{B_X} G_{B_Y} \left(G_{B_Y} + \epsilon_n I_n\right)^{-1} \psi \text{ subject to } \|\phi\|_{\mathbb{R}^n} = \|\psi\|_{\mathbb{R}^n} = 1. \tag{7.2}$$

Let $\phi, \psi \in \mathbb{R}^n$ be eigenvectors corresponding to the first singular value of (7.2). Then, the solution of (7.1), $f, g \in \mathbb{R}^n$, are $n^{1/2}(G_{B_X} + \epsilon_n I_n)^{-1}\phi$ and $n^{1/2}(G_{B_Y} + \epsilon_n I_n)^{-1}\psi$, respectively. It means that estimates of KCCA are represented by $\hat{f}(x) = \sum_{j=1}^n f_j\, b_j^{(X)}(x)$ and $\hat{g}(y) = \sum_{j=1}^n g_j\, b_j^{(Y)}(y)$.

8. Concluding remark

We endeavor to integrate a range of nonlinear SDR methods within the framework established by Lee *et al.* (2013) and Li and Song (2017). In the practical implementation of nonlinear SDR methods, we recommend using KCCA, GSIR, and GSAVE in general. This is because KSIR requires the response variable to be a scalar value, and other methods are known to have better performance than KSIR. Both KCCA and GSIR demonstrate similar performance, but GSAVE provides slightly different information. More precisely, KCCA and GSIR rely on $E[f(X)|Y]$ while GSAVE extracts information from $\mathrm{var}(f(X)|Y)$. Thus the performance depends on the structure of data. See Lee *et al.* (2013) for a numerical comparison of the methods. However, it is not easy to determine which method is superior based solely on the dataset. Therefore, we suggest trying different methods and choosing the one that works best under an appropriate criterion.

One aspect we have not addressed is the dimensionality of the reduced feature space. The target of nonlinear SDR is the central class $\mathfrak{S}_{Y|X}$. On the other hand, dimensionality is defined for linear subspaces as a geometric structure, leading to confusion in determining an appropriate dimension. For instance, the space spanned by the $\{f_1(x), f_2(x)\}$ where $f_1(x) = x$, $f_2(x) = x^3$ is a two-dimensional

space. But they provide the same sigma-algebra, $\sigma(f_1(X)) = \sigma(f_2(X)) = \sigma(f_1(X), f_2(X))$. This highlights the subtle differences encountered when dealing with nonlinear methods, making the selection of dimensionality a distinctly different problem from that in linear methods. Practically, we assume a finite rank for the target operator defined in the RKHS to extract the dimension. For example, Li and Song (2017) proposed a BIC-based criterion for selecting the dimension d. Nevertheless, more theoretical analysis is necessary to understand the dimensionality of the central class.

References

Aronszajn N (1950). Theory of reproducing kernels, *Transactions of the American Mathematical Society*, **68**, 337–404.

Baker CR (1973). Joint measures and cross-covariance operators, *Transactions of the American Mathematical Society*, **186**, 273–289.

Berlinet A and Thomas-Agnan C (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic Publishers, Boston, MA.

Cook RD (1998). *Regression Graphics*, Wiley, New York.

Cook RD and Weisberg S (1991). Sliced inverse regression for dimension reduction: Comment, *Journal of the American Statistical Association*, **86**, 328–332.

Fukumizu K, Bach FR, and Gretton A (2007). Statistical consistency of Kernel canonical correlation analysis, *Journal of Machine Learning Research*, **8**, 361–383.

Lee K-Y, Li B, and Chiaromonte F (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation, *The Annals of Statistics*, **41**, 221–249.

Li B (2018). *Sufficient Dimension Reduction: Methods and Applications with R*, CRC Press, Boca Raton, FL.

Li B and Song J (2017). Nonlinear sufficient dimension reduction for functional data, *The Annals of Statistics*, **45**, 1059–1095.

Li K-C (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 316–327.

Schölkopf B and Smola AJ (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT press, Cambridge, Mass.

Song J, Kim K, and Yoo JK (2023). On a nonlinear extension of the principal fitted component model, *Computational Statistics and Data Analysis*, **182**, 107707.

Wu H-M (2008). Kernel sliced inverse regression with applications to classification, *Journal of Computational and Graphical Statistics*, **17**, 590–610.

Yeh Y-R, Huang S-Y, and Lee Y-J (2008). Nonlinear dimension reduction with kernel sliced inverse regression, *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1590–1603.

Received January 21, 2024; Revised January 30, 2024; Accepted January 30, 2024