The Pennsylvania State University

The Graduate School

Eberly College of Science

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

A Dissertation in
Statistics
by
Kuang-Yao Lee

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2012

The dissertation of Kuang-Yao Lee was reviewed and approved* by the following:

Bing Li
Professor of Statistics
Dissertation Adviser, Chair of Committee

Francesca Chiaromonte
Professor of Statistics and Public Health Sciences

Runze Li
Professor of Statistics and Public Health Sciences, Graduate Program Chair

Lee Giles
Professor of Information Sciences and Technology

Bruce G. Lindsay
Willaman Professor of Statistics, Department Head

*Signatures are on file in the Graduate School.

Abstract

Classical sufficient dimension reduction aims at searching directions in a Euclidean space that can preserve the information about the relation between a predictor and a response, both of which can be vectors. We reformulate sufficient dimension reduction in a nonlinear setting where the effective predictors are allowed to be arbitrary functions. This formulation 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 population-level properties of existing methods and develop new ones. We begin at the completely general level of σ-fields, and proceed to that of measurable and generating classes of functions. This leads to the notions of sufficient, complete and sufficient, 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 slice inverse regression estimator (GSIR). When completeness does not hold, this estimator captures only part of the central class (i.e. remains unbiased but is no longer exhaustive). However, 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.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to show my gratitude to my advisor, Dr. Bing Li, who has constantly supported me in the completion of this dissertation. With extreme patience and inspiration, he provided an excellent environment for conducting research; above all, he has demonstrated to me how rigorous a scientific attitude can and must be.

In addition to my advisor, I would like to thank my committee members, Dr. Francesca Chiaromonte, Dr. Lee Giles, and Dr. Runze Li, for their guidance, insightful criticism, and precious time.

I would like to thank the SGI members in central Pennsylvannia area, for chanting with me, encouraging and accompanying me during my five years' adventure at Penn State.

Last but not the least, I am indebted to to my wife, Yi-Hsia, for her always being there boosting me up through the good and bad times.

Chapter 1

Introduction

1.1 Overview

The topics of dimension reduction are becoming increasingly important in statistics, machine learning and data mining today. Complex and high-dimensional data are massively produced in many contemporary scientific researches. Very often dimension reduction is employed as a preprocessing step to reduce the input dimension, prior to the more detailed and refined analysis. The goal of this dissertation is to provide a novel framework for nonlinear dimension reduction (SDR); meanwhile, we develop new estimators and study the properties of the estimators under this new framework.

Sufficiet Dimension Reduction (SDR) provides a solution when it comes to the formulating the purpose of dimension reduction. SDR begins with the search a small-dimensional $\mathcal{S}$ of $\mathbb{R}^p$ that can explain all the dependence between response Y and predictors X; that is, when $\mathcal{S}$ is given, no information of Y from X will be lost. $\mathcal{S}$ can be a linear or nonlinear mapping of X. In the linear case we can represent SDR as

$$Y \perp\!\!\!\perp X | P_{\mathcal{S}} X, \tag{1.1}$$

where $P_{\mathcal{S}}$ is the orthogonal projection matrix of $\mathcal{S}$. This means Y and X are conditionally independent when $P_{\mathcal{S}} X$ is given.

Let $B = (\beta_1, \cdots, \beta_m) : \mathbb{R}^p \to \mathbb{R}^m$, and $\text{Span}(B) = \text{Span}(P_{\mathcal{S}})$, where $P_{\mathcal{S}}$ is defined in (1.1). An equivalent statement to (1.1) is

$$Y|X \overset{\mathcal{D}}{=} Y|B^{\mathsf{T}}X. \tag{1.2}$$

$\text{Span}(B)$, carrying all the information of Y through X, is called a *dimension reduction space* and is denoted by $\mathcal{S}_B$. The *central subspace* (Li, 1991, 1992, Li and Duan 1989, Cook and Weisberg 1991, Cook, 1994, 1998), written as $\mathcal{S}_{Y|X}$, is defined via the interaction of all possible $\mathcal{S}_B$s,

$$\mathcal{S}_{Y|X} = \cap_B \mathcal{S}_B. \tag{1.3}$$

The existence of $\mathcal{S}_{Y|X}$ is guaranteed under some mild conditions - see Chiaromonte and Cook (2002), and Yin, Li and Cook (2008).

SDR can be viewed as an extension of principle component analysis (PCA), or its supervised version due to the fact that SDR takes into account for the influence from the response on the regression model. PCA looks for the linear subspace of X that captures the most variation from X itself, while SDR seeks the subspace of X that explains the most variation from Y (Chen and Li, 1998).

Built on the structure of reproducing kernel Hilbert space (RKHS), Schölkopf, Smola, and Müller (1997, 1998) introduce a nonlinear extension of PCA called *Kernel PCA*. Kernel PCA is designed to capture more complicated (possibly nonlinear) structure in X that explains the most variation of X. For example, when the original data space is in elliptical shape, PCA can be useful since the direction that explains the most variation of the data cloud is linear; when data is in donut shape, or is distributed in a nonlinear manner, Kernel PCA handles better in explaining the intrinsic structure.

Via kernel mapping, the original space is projected into a high-dimensional space (possibly infinite) where the data are more easily to be isolated and explained by simple structure. This is known as the *Kernel trick* that is the driving force of Kernel PCA. Another pro of Kernel PCA is, the dimension of the matrices operated in the estimating procedure stays with the sample size. This is beneficial in the computation when the dimension of X is

much larger than the sample size.

There are some recent work that combines SDR with RKHS, e.g. *Kernel Sliced Inverse Regression* (KSIR) (Wu, Liang and Mukherjee, 2008, Hsin and Ren, 2009, Yeh, Huang and Lee, 2009) - see also Zhu and Li (2011) and Li, Artemiou and Li (2011). These work provide flexible and accurate methods for estimating nonlinear features of predictor; in addition, they create the possibility of extending the notion of sufficiency in classical SDR, to high-dimensional or functional spaces.

This dissertation aims at investigating this possibility even further, and more importantly in a much broader sense. One of our goals is to provide a deeper connection between SDR and classical statistical inference. We start with the introductory to a general formation of SDR that succeeds from the pioneer work by Li (1991,1992) and Cook (1994,1996,1998), but not limited to linear SDR. Under this new formulation we are able to study linear and nonlinear SDR in a comparative manner; furthermore, because of the richness supplying by this formulation, we are able to relax some of the restrictions made in linear SDR, e.g. linearity condition (Li, 1991, Cook, 1998). We also introduce the notions of *unbiasedness*, *exhaustiveness*, *completeness* and *minimal sufficiency* for nonlinear SDR, which are developed in parallel to their classical counterparts - see Fisher (1992), Neyman (1935), Halmos and Savage (1949), Lehmann and Scheffé (1950, 1955) and Bahadur (1954).

1.2 Summaries of Remaining Chapters

We summarize our development and each subsequent chapter in the following. In the rest of this chapter, we build upon the idea of Cook (2007) and Li, Artemiou, and Li (2011), which allows us to introduce *SDR σ-fields* and *SDR classes*. To extend the scope of SDR so that the sufficient predictors are not restricted to the linear functions of X, we re-formulate the conditional independence in (1.1), and define any sub σ-field $\mathcal{G}$ of $\sigma(X)$ as an SDR σ-field if it satisfies

$$Y \perp\!\!\!\perp X|\mathcal{G}. \tag{1.4}$$

To achieve the maximal dimension reduction, we define the intersection of all such $\mathcal{G}$'s as the *central sufficient dimension reduction σ-field*. We show existence of central SDR σ-field under a very general condition - much more general than the condition for central subspace in the linear setting.

We then turn to the notion of the SDR class. In linear SDR, a sufficient dimension reduction subspace $\mathcal{S}$ is determined by linear functions of the form $\beta_1^\mathsf{T} X, \cdots, \beta_d^\mathsf{T} X$ where $\beta_1, \cdots, \beta_d$ are the vectors to span $\mathcal{S}$. Following this logic, in the nonlinear case we characterize the SDR σ-field using functions; in particular, we consider functions in L^2 classes. There are some nice properties in L^2 class, such as inner product and orthogonal projection; and these make concrete the structure of σ-fields, which allows us to explore the similarities between linear and nonlinear dimension reduction in a more comparable way. We define the SDR class a subspace in $L_2(P_X)$ that is the collection of all measurable functions with respect to $\mathcal{G}$; furthermore, when $\mathcal{G}$ is the central SDR σ-field we call the corresponding class *central class*. Clearly, the goal of nonlinear sufficient dimension reduction is to make influence about the central class.

In chapter 2, we first of all review some linear SDR methods and explore their population-level properties such as unbiasedness and exhaustiveness - see Cook (2008) and Li, Zha, Chiaromonte (2005). The formulation of the central class allows us to develop these concepts for nonlinear SDR. Therefore, we provide the definitions of unbiasedness and exhaustiveness that generalize the ideas in Cook (1998), Li, Zha, and Chiaromonte (2005) and Li, Artemiou and Li (2011). In the nonlinear setting, a function $f \in L_2(P_X)$ is said to be an *unbiased* estimator of central class if it is measurable with respect to the yielding central dimension reduction σ-field; in addition, if a set of functions generates the same central SDR σ-field, then it is *exhaustive*.

Along with the development of these important properties, we come to a critical concept of the *complete dimension reduction class*; the completeness in nonlinear seeting is in the sense similar to the conventional statistics but is more general and applicable to wider statistical models. We say a subspace in $L_2(P_X)$ is *complete* if for each of its member g

$$E[g(X)|Y] = 0 \ \ a.s. \ \text{ implies } \ g(X) = 0 \ \ a.s.$$

We also demonstrate that, under completeness condition the central SDR class can be represented as a subspace in $L_2(P_X)$, $L_2(P_X) \ominus [L_2(P_X) \ominus L_2(P_Y)]$, where $A \ominus B$ indicates the direct difference $A \cap B^{\perp}$. This class plays an important role in nonlinear SDR. We specifically call this class *regression class*. The following summarizes our core discoveries about regression class:

- **[Unbiasedness]**: regression class is contained in central class;
- **[Exhaustiveness]**: if the central class is complete, then regression class and the central class are identical;
- **[Minimal Sufficiency]**: if there exists a complete SDR class, then it is the central class.

In chapter 3, we establish a crucial relationship between the regression class and the extended covariance operator in the RKHSs of X and Y. We begin with the construction of the RKHS and study its fundamental properties. When linear methods are applicable in RKHSs that are in high-dimensional spaces (possibly infinite), as the richness and complexity in these functional spaces, the original linear methods can have "nonlinear" interpretations. In other words, a simple structure in RKHS, e.g. linear, is adequate to offer quite complicated pattern in the original space. This is the *Kernel Trick* that is the key of several existing nonlinear dimension reduction methods. We review some of these methods and give some insight of how flexible and versatile in the way dimension reduction and RKHS can be coorporated.

In chapter 4, we introduce a new estimator for regression class using the covariance operator in RKHS. The formulation in (1.4) accommodates both linear and nonlinear sufficient dimension reduction; in addition, it enables us to use the tools from linear sufficient dimension reduction, to develop estimators for nonlinear sufficient dimension reduction. We then introduce *Generalized Sliced Inverse Regression* (GSIR). GSIR expands the idea of Slice Inverse Regression (SIR, Li, 1991) to the nonlinear setting. We prove that GSIR is an unbiased estimator of the central class; furthermore, we show when the central class is complete, GSIR is exhaustive.

Kernel Sliced Inverse Regression (KSIR) (Wu, 2008 and Yeh, Huang, and Lee, 2009) is proposed to hurdle similar task as GSIR; we show KSIR is unbiased under our framework. It is worth noting that, in the nonlinear sufficient dimension reduction the unbiasedness of both GSIR and KSIR does not require a linearity assumption on the conditional mean - this is often needed in the linear SDR. We also provide a sample version of GSIR, along with a new parameter selection procedure.

In chapter 5, we first discuss an *incomplete class* and show its existence; that is, completeness no longer hold for the central class. Given that an incomplete central class is presented, GSIR is no longer guaranteed exhaustive; this implies certain portion of the central class cannot be captured by GSIR. It's known that, SIR fails at estimating the central subspace when a symmetric structure is present - see Cook and Weisberg, 1991. When this occurs *Sliced Average Variance Estimation* (SAVE) provides a more general estimation than SIR. This motivates us the *generalized SAVE* (GSAVE). We introduce a new operator called *heteroscedastic conditional variance operator*, and use it to construct GSAVE. We show GSAVE is an unbiased estimator for the central class; and when the central class is incomplete it is capable of estimating functions outside the regression class .

Here, we also point out that the relation between the regression class and the central class is somewhat akin to that between the central mean subspace (Cook and Li, 2002) and the central subspace (Cook, 1994) in linear SDR. Cook and Li (2002) consider the following conditional independence structure

$$Y \perp\!\!\!\perp E(Y|X)|B^{\mathsf{T}}X.$$

The above focuses dimension reduction on the regression function $E(Y|X)$. The smallest subspace span(B) that satisfies this relation is called the *central mean subspace*, denoted by $\mathcal{S}_{E(Y|X)}$. One can easily see that this space is a subspace of the central subspace.

We compare our methods with some existing methods by simulation datasets in chapter 6. Varied settings of covariates, including Gaussian, non-Gaussian, and correlated covariates, are examined. In chapter 7 we evaluate our methods on two actual datasets.

1.3 Sufficient Dimesion Reduction σ-fields

Let (X, Y) be a pair of random vectors of dimensions p and q, respectively, defined on a probability space $(\Omega, \mathcal{F}, P)$. Let P_X, P_Y, and P_{XY} be the distributions of X, Y, and (X, Y), which have densities f_X, f_Y, and f_{XY} with respect to some σ-finite measures on $\mathbb{R}^{p+q}$, $\mathbb{R}^p$, and $\mathbb{R}^q$. Let Ω_{XY}, Ω_X, and Ω_Y be the supports of these densities. Let $\mathcal{F}_X$, $\mathcal{F}_Y$, and $\mathcal{F}_{XY}$ be the σ-fields of Borel sets in Ω_X, Ω_Y, and Ω_{XY}. Let $\sigma(X) = X^{-1}(\mathcal{F}_X)$ and $\sigma(Y) = Y^{-1}(\mathcal{F}_Y)$ be the sub-σ-fields of $\mathcal{F}$ generated by X and Y. Let $P_{Y|X}(\cdot|\cdot) : \mathcal{F}_Y \times \Omega_X \to \mathbb{R}$ be the conditional distribution of Y given X. Define $P_{X|Y}$ similarly.

Definition 1.1 *A sub σ-field $\mathcal{G}$ of $\sigma(X)$ is a sufficient dimension reduction (SDR) σ-field for Y versus X if it satisfies (1.4) – that is, if Y and X are independent given $\mathcal{G}$.*

Clearly, the conditional independence (1.4) is a generalization of (1.1) for linear SDR: if we take $\mathcal{G} = \sigma(B^\intercal X)$ then (1.4) reduces to (1.1). However, $\mathcal{G}$ in (1.4) can be generated by any set of measurable functions, not restricted to the linear form $B^\intercal X$. The notion of sufficiency in Definition 1.1 is different from the classical notion of sufficiency (Fisher, 1922), because $\mathcal{G}$ is allowed to depend on any parameter in the joint distribution of P_{XY}. For example $\mathcal{G} = \sigma(\beta^\intercal X)$ depends on the 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 first through a predictor and the second through a statistic. Indeed, it is partly through exploring this similarity that we develop the theory of nonlinear SDR.

Obviously there are many sub σ-fields of X that satisfy (1.4), for example $\sigma(X)$ itself satisfies this relation. 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 1.1 *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* $\mathcal{G}^* \subseteq \mathcal{G}$.

Before proving the theorem, let us introduce some notation. Let $\mathcal{P}$ and $\mathcal{Q}$ be two classes of probability measures on $(\Omega, \mathcal{F})$. We say that $\mathcal{Q}$ is dominated by $\mathcal{P}$ if, for each $A \in \mathcal{F}$,

$$P(A) = 0 \text{ for all } P \in \mathcal{P} \quad \Rightarrow \quad Q(A) = 0 \text{ for all } Q \in \mathcal{Q}.$$

We denote this by $\mathcal{Q} \ll \mathcal{P}$. If $\mathcal{Q} \ll \mathcal{P}$ and $\mathcal{P} \ll \mathcal{Q}$ then we say the two families are equivalent, and write $\mathcal{Q} \equiv \mathcal{P}$. We use $\mathbb{N}$ to denote the natural numbers $\{1, 2, \ldots\}$. We also need to use the following well known fact. Let U and V be two random variables defined on $(\Omega, \mathcal{F})$, and $\mathcal{G}$ be a sub σ-field of $\mathcal{F}$. Then

$$E[E(U|\mathcal{G})V] = E[UE(V|\mathcal{G})]. \tag{1.5}$$

Proof of Theorem 1.1. Let Π_y denote the measure $P_{X|Y}(\cdot|y)$. Since Π_y is dominated by a σ-finite measure, by Lemma 7 of Halmos and Savage (1949), there is a countable subset of $\{Q_k : k \in \mathbb{N}\} \subseteq \{\Pi_y : y \in \Omega_Y\}$ such that $\{Q_k : k \in \mathbb{N}\} \equiv \{\Pi_y : y \in \Omega_Y\}$. Let $\{c_k : k \in \mathbb{N}\}$ be a sequence of positive numbers that sum to 1, and let $Q_0 = \sum_{k\in\mathbb{N}} c_k Q_k$. Then Q_0 itself is a probability measure on Ω_X and $\{Q_0\} \equiv \{Q_k : k \in \mathbb{N}\} \equiv \{\Pi_y : y \in \Omega_Y\}$. 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 $\Rightarrow$ *2.* Let $B \in \mathcal{F}_X$. Then

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

By 1, $\Pi_y(X \in B|\mathcal{G})$ is the same for all $y \in \Omega_Y$. Hence $\Pi_y(B|\mathcal{G}) = Q_k(B|\mathcal{G})$ for all $k \in \mathbb{N}$, which implies $\Pi_y(B|\mathcal{G}) = Q_0(B|\mathcal{G})$. By this relation and the hermitian property (1.5), we

can rewrite the right hand side of the above equalities as

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

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

Proof of 2 $\Rightarrow$ 1. We will show that, for any $B \in \mathcal{F}_X$,

$$\Pi_y(B|\mathcal{G}) = Q_0(B|\mathcal{G}). \tag{1.6}$$

If this holds then $\Pi_y(B|\mathcal{G})$ does not depend on y, implying 1. Let $A \in \mathcal{G}$. By (1.5) again,

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

Since $A \in \mathcal{G}$, we have $E_{Q_0}(I_A(X)\pi_y(X)|\mathcal{G}) = I_A(X)E_{Q_0}(\pi_y(X)|\mathcal{G})$. By 2, $E_{Q_0}(\pi_y(X)|\mathcal{G}) = \pi_y(X)$. Hence the right hand side of the above display becomes

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

Now (1.6) follows from the definition of conditional probability. The claim is proved.

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 measurable with respect to $\mathcal{G}^*$ for all $y \in \Omega_Y$. Consequently, $\mathcal{G}^*$ is itself an SDR σ-field.

If $\mathcal{G}$ is any SDR σ-field, then it satisfies 2. Hence it contains $\mathcal{G}^*$. Thus $\mathcal{G}^*$ is 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 1.2 *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 1.1 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 allow us to bring out some structures that are very consequential for further development. In particular, these structures allow us to work with powerful ideas such as Hilbert spaces and completeness.

1.4 Sufficient Dimension Reduction Classes

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 $\mathcal{M}_\mathcal{G}$ 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 $\mathcal{M}_X$ to abbreviate $\mathcal{M}_{\sigma(X)}$. Note that, for any $\mathcal{G}$, $\mathcal{M}_\mathcal{G}$ is a linear subspace of $L_2(P_{XY})$.

Definition 1.3 *Let* $\mathcal{G}$ *be an* SDR σ*-field and* $\mathcal{G}_{Y|X}$ *be the central* σ*-field. Then the subspaces* $\mathcal{M}_\mathcal{G}$ *and* $\mathcal{M}_{\mathcal{G}_{Y|X}}$ *of* $L_2(P_X)$ *are called an* SDR *class, and the central class, respectively. 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. Therefore, the goal of nonlinear SDR is to make inference about $\mathfrak{S}_{Y|X}$. Note that the notion of central dimension reduction class is not equivalent to central

dimension reduction σ-field. The latter is a more general concept that does not rely on square-integrability.

Chapter 2

Unbiasedness, Exhaustiveness and Completeness

The purpose of SDR is to seek important features in the space of predictors X - these features are either linear or nonlinear mappings that explains the most variation of response Y; in other words, the feature extracted by SDR is sufficient to provide the full understanding of the regression model. In the first half of this chapter, we review some of the most representative methods in linear SDR; in addition, these methods are studied and evaluated using the two criteria - unbiasedness and exhaustiveness, developed by Cook (1998), Li, Zha and Chiaromonte (2005), and Li, Artemiou, and Li (2011). Specifically, we explore the theoretical properties of these methods in the population level and examine accordingly which condition is required for the unbiasedness or exhaustiveness to be satisfied.

Definition 2.1 *Let F_{XY} be the distribution of (X, Y), and let $\mathcal{S}_{Y|X}$ be the central subspace. Suppose $T(F_{XY})$ is a statistic that is matrix-valued. Then we say $T(F_{XY})$ is unbiased if*

$$\text{Span}(T(F_{XY})) \subseteq \mathcal{S}_{Y|X}; \tag{2.1}$$

in particular, when the equality holds, i.e. $T(F_{XY})$ spans the entire central subspace, we say it's exhaustive.

2.1 Review of Linear Sufficient Dimension Reduction Methods

1 Inverse Regression Methods

We give a brief introductory to some linear SDR methods that are based on inverse regression, or inverse moments. Methods of this type often depend on the calculation of two inverse moments - first order $E(X|Y)$ and second order $E(XX^{\mathsf{T}})$. Some methods consider only the first moment such as *Ordinary Least Squares* (Li, 1989) and *Sliced Inverse Regression* (Li, 1991); others use both first and second inverse moments - see *Sliced Average Variance Estimator* (Cook and Weisburg, 1991), *Contour Regression* (Li, Zha, and Chiaromonte, 2005), and *Directional Regression* (Li and Wang, 2007).

There are some advantages shared among inverse regression methods; for example, the validity of the methods is established without any further assumptions on the link function of the regression model; in addition, their estimators are easy to compute, mostly involved only the solutions of eigenvalue problems.

Sliced Inverse Regression

One of the most commonly used SDR methods is the Sliced Inverse Regression (SIR, Li, 1991). SIR approaches the problem of dimension reduction by computing the *inverse regression* $E(X|Y)$. Li (1991) proves that, $E(X|Y)$ provides an unbiased estimator for $\mathcal{S}$, with some regularities on the predictors. Therefore, the estimation of the $\mathcal{S}_{Y|X}$ depends on the estimation of the space spanned by $E(X|Y)$. We first of all offer the theoretical foundation of SIR.

Assumption 2.1 *Let $\Gamma \in \mathbb{R}^{p\times d}$ whose column space is $\mathcal{S}_{Y|X}$, i.e.* $\mathrm{Span}(\Gamma) = \mathcal{S}_{Y|X}$. *Suppose the following relation holds,*

$$E(X|\Gamma^{\mathsf{T}}X) \text{ is linear in } \Gamma^{\mathsf{T}}X.$$

That is, for each $a \in \mathbb{R}^p$, there exists $b \in \mathbb{R}^d$ such that $E(a^{\mathsf{T}}X|\Gamma^{\mathsf{T}}X) = b^{\mathsf{T}}\Gamma^{\mathsf{T}}X$.

Theorem 2.1 *Under Assumption (2.1), we have for all* $y \in \Omega_Y$,

$$E(X|y) - E(X) \in \Sigma_{XX}\mathcal{S}_{Y|X}, \tag{2.2}$$

where $\Sigma_{XX} = \mathrm{var}(X)$.

PROOF. Suppose $Z = \Sigma_{XX}^{-1/2}(X - E(X))$ is the standardized predictor. Then (2.2) is equivalent to

$$E(Z|y) = \Sigma_{XX}^{-1/2}E[X - E(X)|y] \in \Sigma_{XX}^{1/2}\mathcal{S}_{Y|X} = \mathcal{S}_{Y|Z}, \tag{2.3}$$

where the last equality is the invariance property of central subspace; that is, we can work on Z-scale directly without loss of generosity. Note that by Assumption (2.1),

$$E(Z|\Gamma^{\mathsf{T}}X) = E(Z|\Gamma^{\mathsf{T}}\Sigma_{XX}^{1/2}Z) = P_{\Sigma_{XX}^{1/2}\Gamma}Z,$$

where in $P_{\Sigma_{XX}^{1/2}\Gamma}$ is the projection matrix on $\mathrm{Span}(\Sigma_{XX}^{1/2}\Gamma)$. Therefore, we have

$$E(Z|y) = E[E(Z|\Gamma^{\mathsf{T}}\Sigma_{XX}^{1/2}Z)|y] = E(P_{\Sigma_{XX}^{1/2}\Gamma}Z|y) = P_{\Sigma_{XX}^{1/2}\Gamma}E(Z|y),$$

for each $y \in \Omega_Y$. This implies $E(Z|y) \in \Sigma_{XX}^{1/2}\mathcal{S}_{Y|X}$; therefore, $E(X|y) - E(X) \in \Sigma_{XX}\mathcal{S}_{Y|X}$ by transforming back to X-scale. □

According to (2.1) and the fact that $\mathrm{Span}(E(X|Y)) = \mathrm{Span}(\mathrm{var}[E(X|Y)]) \subseteq \mathcal{S}_{Y|X}$, $\mathcal{S}_{Y|X}$ can be estimated by the principle component of $\mathrm{var}(E(X|Y))$ (after rescaling by Σ_{XX}).

Algorithm of SIR The steps of SIR are listed below. Given data $\{(X_i, Y_i)\}_{i=1}^n$,

1. Slice the covariate $X = \begin{bmatrix} X_1^{\mathsf{T}} \\ \vdots \\ X_n^{\mathsf{T}} \end{bmatrix}$ into h slices by their associated Y.
2. Within each slice, $s = 1, \ldots, h$, calculate the mean, written as $\bar{X}_s$, as well as the grand mean, written as $\bar{X}$.

3. Replace each point X_i by its slice mean $\bar{X}_s$ and then build a new data matrix X_h.

4. Compute the between-slice covariance matrix $\Lambda = \mathrm{Cov}(X_h)$ from:

$$\Lambda = \frac{1}{n}\sum_{s=1}^{h} n_s(\bar{X}_s - \bar{X})(\bar{X}_s - \bar{X})^{\mathsf{T}},$$

where n_s is the number of data points in the sth slice. And also calculate the sample covariance matrix of X: $\hat{\Sigma} = \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^{\mathsf{T}}/n$.

5. Solve the generalized eigenvalue problem of Λ with respect to $\hat{\Sigma}$; i.e. find the leading m ($m \geq H - 1$) eigenvalues λ and corresponding eigenvectors β:

$$\Lambda\beta_j = \lambda_j\hat{\Sigma}\beta_j.$$

6. Determine the leading m SIR directions and collect them in a matrix $V_{p\times m} = [v_1, \ldots, v_m]$. V is orthogonal in terms of $\hat{\Sigma}$ inner product, i.e., $V^{\mathsf{T}}\hat{\Sigma}V = I_m$. Then project X onto V,

$$V^{\mathsf{T}}X = \begin{bmatrix} v_1^{\mathsf{T}}X \\ \vdots \\ v_m^{\mathsf{T}}X \end{bmatrix},$$

$v_1^{\mathsf{T}}X$ is called the first SIR variate, $v_2^{\mathsf{T}}X$ the second SIR variate, and so on.

SIR searches directions in $\mathcal{S}_{Y|X}$ by switching the roles of Y and X. The information of conditional distribution of Y given X can be captured via the inverse regression $E(X|Y)$, One of its superiority is there is no need to know the form of $E(Y|X)$ which is often required in classical regression methods; moreover, its algorithm is easy to implement. Under linearity condition SIR is shown to be unbiased for $\mathcal{S}_{Y|X}$. Nonetheless, there are some downsides of SIR; for example, if there exists symmetric dependence in the model, e.g. U-shape, then SIR cannot detect the directions in the central subspace. This can be seen by the following example.

$$Y = Z_1^2 + \epsilon, \tag{2.4}$$

where Z_1 and ϵ are both standard normals. In (2.4), the inverse regression $E(Z|Y)$ is zero, in which SIR can no longer extract any feature in the predictor space. If there exists a symmetric structure between Y and X, SIR may not be necessary to cover the entire central subspace. When this occurs, the space spanned by $E(X|Y)$ is generally smaller than $\mathcal{S}_{Y|X}$

$$\text{Span}(E(X|Y)) \subseteq \mathcal{S}_{Y|X}.$$

Sliced Average Variance Estimation

Sliced Average Variance Estimation (SAVE) (Cook and Weisberg, 1991) is another SDR method. It serves as a complementary tool of SIR which in general is able to estimate larger space than SIR. For example, SAVE can detect the directions in $\mathcal{S}_{Y|X}$ even when the symmetric dependence between Y and X exists in the model. The development of SAVE is based on the second moment of X given Y, or the conditional variance $\text{cov}(X|Y)$. Let's first of all look at the following relation,

$$\text{cov}[E(X|Y)] = E[\text{cov}(X) - \text{cov}(X|Y)]. \tag{2.5}$$

On the left hand side of (2.5) it's the covariance matrix of random vector $E(X|Y)$, where SIR is to estimate. The right hand side is based on the second moment $\text{cov}(X|Y)$. Therefore, SAVE uses the column space of $\text{cov}(X|Y)$ to estimate the directions in central subspace.

Assumption 2.2 (Constant Variance Condition) *Let $\Gamma \in \mathbb{R}^{p\times d}$ be a matrix whose column space is $\mathcal{S}_{Y|X}$. Suppose the following relation holds,*

$$\text{var}(X|\Gamma^{\mathsf{T}}X) \quad \textit{is a nonrandom matrix}$$

Theorem 2.2 *Under the Assumptions (2.1) and (2.2), we have*

$$\text{Span}\left(\text{cov}(X) - \text{cov}(X|y)\right) \subseteq \Sigma_{XX}\mathcal{S}_{Y|X}, \tag{2.6}$$

for each $y \in \Omega_Y$.

PROOF. Similar to the proof in Theorem 2.1, we work on the Z-scale rather than the original predictor. Therefore, we need to show that

$$\text{Span}\,(I_p - \text{cov}(Z|y)) \subseteq \mathcal{S}_{Y|Z}. \tag{2.7}$$

Note that, by the well-known relation the conditional variance of Z given y, $\text{cov}(Z|y)$ can be re-written as

$$\text{cov}(Z|y) = E[\text{cov}(Z|P_{\Sigma_{XX}^{1/2}\Gamma})|y] + \text{cov}[E(Z|P_{\Sigma_{XX}^{1/2}\Gamma})|y].$$

It can be shown the first part on the right of the equality is $I_p - P_{\Sigma_{XX}^{1/2}\Gamma}$ by Assumption (2.2), and the second part is $P_{\Sigma_{XX}^{1/2}\Gamma}\text{cov}(Z|y)P_{\Sigma_{XX}^{1/2}\Gamma}$ by Assumption (2.1). Therefore, we have

$$I_p - \text{cov}(Z|y) = P_{\Sigma_{XX}^{1/2}\Gamma}[I_p - \text{cov}(Z|y)]P_{\Sigma_{XX}^{1/2}\Gamma},$$

which implies (2.7). □

Based on the above result and

$$\text{Span}\,(\Sigma_{XX} - \Sigma_{X|Y}) = \text{Span}\,(\text{cov}[\Sigma_{XX} - \Sigma_{X|Y}])\,,$$

$\mathcal{S}_{Y|X}$ can be estimated by $\text{Span}\,(\text{cov}[\Sigma_{XX} - \Sigma_{X|Y}])$.

Algorithm of SAVE:

1. Slice the covariate $X = \begin{bmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{bmatrix}$ into h slices by their associated Y.
2. Within each slice, $s = 1, \ldots, h$, calculate the mean, written as $\bar{X}_s$; and also the grand mean, written as $\bar{X}$.

3. For each slice compute the within-slice covariance matrix, denoted as $\widehat{\Sigma}_i$ for $1 \leq s \leq h$:

$$\widehat{\Sigma}_s = \frac{1}{n_s} \sum_{\{X_i \in \ s^{th} \text{ slice}\}} (X_i - \bar{X}_s)(X_i - \bar{X}_s)^\mathsf{T},$$

where n_s is the number of data points in the s^{th} slice. And also calculate the sample covariance matrix of X: $\widehat{\Sigma} = \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\mathsf{T}/n$.

4. Estimate $\text{cov}[\Sigma_{XX} - \Sigma_{X|Y}]$, denoted by $\widehat{\Gamma}$, from

$$\widehat{\Gamma} = \sum_{s=1}^{h} \frac{n_s}{n}[\widehat{\Sigma} - \widehat{\Sigma}_s]^2.$$

5. Solve the generalized eigenvalue problem of $\widehat{\Gamma}$ with respect to $\widehat{\Sigma}$; i.e. find the leading m ($m \geq H - 1$) eigenvalues λ and corresponding eigenvectors β:

$$\widehat{\Gamma}\beta_j = \lambda_j \widehat{\Sigma}\beta_j.$$

6. Determine the leading m SAVE directions and collect them as a matrix $V_{p\times m} = [v_1, \ldots, v_m]$. V is orthogonal in terms of Σ inner product, i.e., $V^\mathsf{T}\Sigma V = I_m$. Then project X onto V,

$$V^\mathsf{T}X = \begin{bmatrix} v_1^\mathsf{T}X \\ \vdots \\ v_m^\mathsf{T}X \end{bmatrix},$$

$v_1^\mathsf{T}X$ is called the first SAVE variate, $v_2^\mathsf{T}X$ the second SAVE variate, and so on.

Identically, one can instead solve an ordinary eigen problem:

$$\widehat{\Sigma}^{-1/2}\widehat{\Gamma}\widehat{\Sigma}^{-1/2}\alpha_j = \lambda_j \alpha_j.$$

SAVE in general covers a larger portion of the central subspace than SIR. It fixes the problem when a symmetric dependence is presented in the model; in addition, the rank of GSAVE is not limited to the number of slices. When we have categorical response, e.g. the

domain of Y takes only m distinct values $\Omega_Y = \{y_1, \cdots, y_m\}$, we can only divide the data into m slices. As a result, the space spanned by SIR estimator is at most rank $m-1$, while the rank of SAVE estimator can go beyond that and possibly recover a larger subspace in $\mathcal{S}_{Y|X}$. Let $\mathcal{S}_{\text{SIR}}$ and $\mathcal{S}_{\text{SAVE}}$ denote the spaces spanned by SIR and SAVE, respectively; then the following relation in general holds,

$$\mathcal{S}_{\text{SIR}} \subseteq \mathcal{S}_{\text{SAVE}} = \mathcal{S}_{Y|X}.$$

Nonetheless we should mention there are some downsides of SAVE; for instance, the computation of SAVE involves the estimation of conditional variance, and this may lead to certain level of loss of efficiency; moreover, in order to show the unbiasedness of SAVE, we need additional assumption on the predictor, i.e. the constant variance assumption. Ye and Weiss (2003) and Cook and Forzani (2009) have more details about the comparisons between SIR and SAVE.

Contour Regression (Simple Contour Regression)

Contour Regression (CR) (Li, Zha, and Chiaromonte, 2005) is another SDR method. The criterion behind CR is to estimate $\mathcal{S}_{Y|X}$ based on the "empirical directions". Given $(X_1, X_2, \cdots, X_n)$, the empirical directions is the collection of all possible pair differences $\{(X_i - X_j) : 1 \leq i < j \leq n\}$. The key is, the empirical direction somehow portraits the pattern of the predictor and show us how the directions in X are distributed - this comes from the intuition of how empirical distribution works. In addition, each pair of difference $X_i - X_j$ provides same amount of importance of describing the predictor, just like each observation provides equal probability mass in the empirical distribution. Therefore, within the range of empirical directions, CR searches the directions in the central subspace.

CR starts with characterizing the contour directions that is the orthogonal complement of the central subspace. Contour directions are the empirical directions whose responses rarely or do not change within a tiny contour band. Given a small number c, the contour directions can be captured by the principle components of the matrix $H(c)$, which is defined

as

$$H(c) := E[(X - \widetilde{X})(X - \widetilde{X})^{\mathsf{T}} \mid |Y - \widetilde{Y}| \leq c], \tag{2.8}$$

where $(\widetilde{X}, \widetilde{Y})$ is an independent copy of (X, Y). The given condition $|Y - \widetilde{Y}| \leq c$ describes the size of the contour, and $H(c)$ is the expected square of the empirical directions. After the contour direction is captured, we take its orthogonal complement as the estimate of $\mathcal{S}_{Y|X}$. Under some mild conditions, Li, Zha, and Chiaromonte (2005) show that, the eigenvectors corresponding to the smallest eigenvalues of $H(c)$ can fully span the central subspace.

Algorithm of CR (Simple Contour Regression)

1. Compute the sample mean and variance matrix:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^{\mathsf{T}}.$$

2. For a given c, estimate $H(c)$ defined in (2.8) by

$$\hat{H}(c) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} (X_j - X_i)(X_j - X_i)^{\mathsf{T}} I(|Y_j - Y_i| \leq c). \tag{2.9}$$

3. Solve the eigenvalue problem:

$$\hat{\Sigma}^{-\frac{1}{2}} H(c) \hat{\Sigma}^{-\frac{1}{2}} \gamma = \lambda \gamma,$$

   and $\hat{\gamma}_{p-q+1}, \cdots, \hat{\gamma}_p$ are the eigenvectors associated with the smallest q eigenvalues.

4. Compute $(\hat{\Sigma}^{-\frac{1}{2}}\hat{\gamma}_{p-q+1}, \cdots, \hat{\Sigma}^{-\frac{1}{2}}\hat{\gamma}_p)$.

Directional Regression

Directional Regression (DR) Li and Wang (2007) introduce an alternative SDR approach. DR makes better use of the empirical directions than CR; specifically, DR utilizes empirical directions in a more efficient way. The algorithm of CR involves the selection of contour directions from all $\binom{n}{2}$ empirical directions, and this is apparently not favorable from the

computational perspective. DR regresses the empirical directions directly on the responses; it calculates the variation of empirical directions, and then projects it (in L_2 sense) onto the space of response Y. In Dr, we first of all consider the following expectation

$$A(Y,\widetilde{Y}) = E[(Z-\widetilde{Z})(Z-\widetilde{Z})^{\mathsf{T}}|Y,\widetilde{Y}], \tag{2.10}$$

where $(\widetilde{Y},\widetilde{Z})$ is an independent copy of (Y,Z). It can be shown (Li and Wang, 2007), under the same conditions as SAVE, the column space of the matrix $2I_p - A(Y,\widetilde{Y})$ is unbiased, i.e. $\text{Span}(2I_p - A(Y,\widetilde{Y})) \subseteq \mathcal{S}_{Y|X}$. In the following we demonstrate this result on Z-scale.

Theorem 2.3 *Let $A(Y,\widetilde{Y})$ as defined in (2.10). Suppose*

a $E(\alpha^{\mathsf{T}}Z|\Gamma^{\mathsf{T}}Z)$ is linear in $\Gamma^{\mathsf{T}}Z$, and

b $\text{cov}(Z|\Gamma^{\mathsf{T}}Z)$ is a nonrandom matrix,

then the following property holds,

$$2I_p - A(Y,\widetilde{Y}) \in \mathcal{S}_{Y|Z}. \tag{2.11}$$

PROOF. (2.11) is equivalent to $\text{Ker}(\mathcal{S}_{Y|Z}) \subseteq \text{Ker}(2I_p - A(Y,\tilde{Y}))$. First note that $2I_p - A(Y,\tilde{Y})$ can be expressed as

$$\begin{aligned} 2I_p - A(Y,\tilde{Y}) &= 2I_p - E(ZZ^{\mathsf{T}}|Y) - E(\tilde{Z}\tilde{Z}^{\mathsf{T}}|\tilde{Y}) \\ &\quad + E(Z|Y)E(\tilde{Z}^{\mathsf{T}}|\tilde{Y}) + E(\tilde{Z}|\tilde{Y})E(Z^{\mathsf{T}}|Y) \\ &= [(I_p - \text{cov}(Z|Y)] + [I_p - \text{cov}(\tilde{Z}|\tilde{Y})] \\ &\quad + E(Z|Y)E(\tilde{Z}^{\mathsf{T}}|\tilde{Y}) + E(\tilde{Z}|\tilde{Y})E(Z^{\mathsf{T}}|Y) \\ &\quad - E(Z|Y)E(Z|Y) - E(\tilde{Z}|\tilde{Y})E(\tilde{Z}^{\mathsf{T}}|\tilde{Y}). \end{aligned} \tag{2.12}$$

Suppose $a \in \text{Ker}(\mathcal{S}_{Y|Z})$. Then it follows from Theorems 2.1 and 2.2 that

$$E(a^{\mathsf{T}}Z|Y) = 0, \text{ and } a^{\mathsf{T}}[I_p - \text{cov}(Z|Y)]a = 0,$$

which also imply $E(a^{\mathsf{T}}\tilde{Z}|\tilde{Y}) = 0$ and $a^{\mathsf{T}}[I_p - \mathrm{cov}(\tilde{Z}|\tilde{Y})]a = 0$ because (Z, Y) and $(\tilde{Z}, \tilde{Y})$ have the same distribution. Therefore, $a^{\mathsf{T}}[2I_p - A(Y, \tilde{Y})]a = 0$. □

Theorem 2.3 suggests the central subspace $\mathcal{S}_{Y|Z}$ can be estimated by the column space of $[2I_p - A(Y, \tilde{Y})]$, for any $(Y, \tilde{Y})$. Let G be calculated by the variance of $[2I_p - A(Y, \tilde{Y})]$, i.e.

$$G = E[2I_p - A(Y, \widetilde{Y})]^2. \tag{2.13}$$

Then G provides an estimate of $\mathcal{S}_{Y|Z}$. Actually, the column space of G coincides with the central subspace, i.e. G is exhaustiveness, when one additional condition is given.

Assumption 2.3 *For any nonzero $a \in \mathcal{S}_{Y|Z}$, it is true that*

$$a^{\mathsf{T}} A(Y, \tilde{Y})a \quad \textit{is nondegenerate.}$$

Theorem 2.4 *Let G be the matrix in (2.13). Under the assumptions in Theorem 2.3, and Assumption 2.3, G is exhaustiveness; that is,*

$$\mathrm{Span}\,(G) = \mathcal{S}_{Y|Z}. \tag{2.14}$$

PROOF. The unbiasedness of G is by Theorem 2.3, i.e. $\mathrm{Span}(G) \subset \mathcal{S}_{Y|Z}$. Suppose $\mathcal{S}_{Y|Z}$ cannot be full recovered by the column space of G. That is, there exists nonzero vector $a \in \mathcal{S}_{Y|Z}$ such that a is in the orthogonal complement of $\mathrm{Span}(G)$, or $a^{\mathsf{T}}Ga = 0$. However, this is not true because

$$\begin{aligned} a^{\mathsf{T}}Ga = a^{\mathsf{T}}E[2I_p - A(Y, \tilde{Y})]^2 a &\geq a^{\mathsf{T}}E\left([2I_p - A(Y, \tilde{Y})]aa^{\mathsf{T}}[2I_p - A(Y, \tilde{Y})]\right)a \\ &> \left(a^{\mathsf{T}}E[2I_p - A(Y, \tilde{Y})]a\right)^2 \\ &= 0, \end{aligned}$$

where the second inequality follows by applying Jensen's inequality on $a^{\mathsf{T}}[2I_p - A(Y, \tilde{Y})]a$, which is nondegenerate by assumption. □

Algorithm of DR:

1. First Compute $Z_i = \hat{\Sigma}^{-1/2}(X_i - \hat{\mu})$, where $\hat{\mu}$ & $\hat{\Sigma}$ are the sample mean and sample covariance matrix,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^{\mathsf{T}}.$$

2. Let $\{J_1, ..., J_m\}$ be a partition of Ω_Y. Then estimate $E[2I_p - A(Y, \widetilde{Y}]^2$ by

$$\widehat{E}[2I_p - A(Y, \widetilde{Y}]^2 = \binom{m}{2}^{-1} \sum_{k<l} [2I - \hat{A}(J_k, J_l)]^2, \text{where}$$
$$\hat{A}(J_k, J_l) = \frac{\sum_{i<j}(Z_i - Z_j)(Z_i - Z_j)^{\mathsf{T}} I(Y_i \in J_k, Y_j \in J_l)}{\sum_{i<j} I(Y_i \in J_k, Y_j \in J_l)}.$$

   Or, since $E[2I_p - A(Y, \widetilde{Y}]^2$ can also be represented in the following form

$$\begin{aligned} E[2I_p - A(Y, \widetilde{Y}]^2 &= 2E[E^2(ZZ^{\mathsf{T}} - I|Y)] + 2E^2[E(Z|Y)E(Z^{\mathsf{T}}|Y] \\ &+ 2E[E(Z^{\mathsf{T}}|Y)E(Z|Y)]E[E(Z|Y)E(Z^{\mathsf{T}}|Y)]. \end{aligned}$$

   Li and Wang (2007) provides another estimator:

$$\begin{aligned} \widehat{E}[2I_p - A(Y, \widetilde{Y}]^2 &= 2\sum E_n^2(ZZ^{\mathsf{T}} - I|Y \in J_k)\hat{p}_k \\ &+ 2\left[\sum E_n(Z|Y \in J_k)E_n(Z^{\mathsf{T}}|Y \in J_k)\hat{p}_k\right]^2 \\ &+ \sum E_n(Z^{\mathsf{T}}|Y \in J_k)E_n(Z|Y \in J_k)\hat{p}_k \\ &\times \sum E_n(Z|Y \in J_k)E_n(Z^{\mathsf{T}}|Y \in J_k)\hat{p}_k, \end{aligned}$$

   where $\hat{p}_k = E_n(Y \in J_k)$. $E_n(\cdot)$ stands for the moment estimate of $E(\cdot)$.

3. Find the largest q eigenvalues and their associated eigenvectors of $\widehat{E}[2I_p - A(Y, \widetilde{Y}]^2$. Let $(\hat{\gamma}_1, \cdots, \hat{\gamma}_q)$ be the eigenvectors.

4. Multiply the eigenvectors by $\hat{\Sigma}^{-1/2}$: $(\hat{\Sigma}^{-1/2}\hat{\gamma}_1, \cdots, \hat{\Sigma}^{-1/2}\hat{\gamma}_q)$.

Many other methods provide the estimation of the central subspace. Cook and Ni (2005) introduce a class of inverse regression estimators called IRE. Some others are variants of SIR that tackles the problem when multivariate response is observed - see *Nearest Sliced Inverse Regression* (Hsing, 1999), *K-means Sliced Inverse Regression* (Setodgi and Cook, 2004) and Li et al (2003). Other methods such as Chen and Li (1998), Fung et al (2002), and Zhou and He (2008) connect the idea of canonical correlation of Y and X with SDR.

2 Central Mean Subspace

Central subspace contains all the information about conditional distribution of $Y|X$; however, in most of statistical problems the regression function, i.e. $E(Y|X)$ is truly what we only concern. Let's first consider some examples.

- *Multiple regression model:* $Y = \Gamma^{\mathsf{T}} X + \epsilon$.
- *Single-index model:* $Y = f(\Gamma^{\mathsf{T}} X) + \epsilon$.

It's obvious in these models the dependence between Y and X are fully characterized by the conditional means; in other words, $E(Y|X)$ is sufficient to supply all the information about the conditional distribution $Y|X$. This suggests we can consider a more parsimonious modeling in which the conditional distribution is not entirely involved but the conditional mean. Cook and Li (2002) first introduce the notions of *mean dimension reduction space* and *central mean subspace*.

Definition 2.2 *Let α be a set of vectors in $\Re^p$. Then the space spanned by α, written as $\mathcal{S}_\alpha$, is called a mean dimension reduction space if*

$$Y \perp\!\!\!\perp E(Y|X) \mid \alpha^{\mathsf{T}} X. \tag{2.15}$$

An equivalent statement is, the conditional mean is a function of α, i.e given $\alpha^{\mathsf{T}} X$, $E(Y|X)$ is constant.

Definition 2.3 *Let $\mathcal{S}_{E(Y|X)}$ be the intersection of all mean dimension reduction space,*

$$\mathcal{S}_{E(Y|X)} = \cap_\alpha \mathcal{S}_\alpha. \tag{2.16}$$

Then $\mathcal{S}_{E(Y|X)}$ is called the central mean subspace.

It is obvious from the definition that the central mean subspace is a subspace of the central subspace, i.e.

$$\mathcal{S}_{E(Y|X)} \subseteq \mathcal{S}_{Y|X}.$$

Cook and Li (2002) show that ordinary least squares (OLS; Li and Duan, 1989), principal Hessian directions (PHD; Li, 1992; Cook, 1998) and iterative Hessian transformations (IHT; Cook and Li, 2002, 2004) are bound to recover directions inside the central mean subspace.

Ordinary Least Square

Li and Duan (1989) introduce OLS, which targets at the estimation of the central mean subspace. Let Δ be a basis of $\mathcal{S}_{E(Y|X)}$. Suppose $E(\gamma^\mathsf{T} X|\Delta^\mathsf{T} X)$ is linear in $\Delta^\mathsf{T} X$. Then one can show the minimizer of an objective function based on some convex functions, is within $\mathcal{S}_{E(Y|X)}$. That is, if $Z = \Sigma_{xx}^{-1/2} X$ is the standardized covariate, then $E(YZ) \in \mathcal{S}_{E(Y|X)}$. The following is the Fisher consistency of OLS.

Theorem 2.5 *Consider the objective function $L(a, \mathbf{b})$ defined as*

$$L(a, \mathbf{b}) = -Y(a + \mathbf{b}^\mathsf{T} Z) + \phi(a + \mathbf{b}^\mathsf{T} Z), \tag{2.17}$$

where ϕ is convex. Let (β_0, β) be the minimizer of the expectation of (2.17):

$$(\beta_0, \beta) = \text{argmin}_{a,\mathbf{b}} E[L(a, \mathbf{b})].$$

Suppose Assumption 2.1 holds, i.e.

$$E[\gamma^\mathsf{T} Z|\Delta^\mathsf{T} Z] \quad \textit{is linear in } \Delta^\mathsf{T} Z.$$

We then have

$$\beta \in \mathcal{S}_{E(Y|Z)}. \tag{2.18}$$

Note in (2.17), when $\phi(t) = t^2/2$, one can show that $\beta = E(YZ)$.

OLS is easy to compute; in addition it correctly identifies the direction in the central mean subspace - without any specification on the link function on $E(Y|X)$. However, OLS fails when there exists symmetric dependency between Y and X, e.g, $E(YZ) = 0$, or when the U-shape structure is presented as well. Furthermore, when $\mathcal{S}_{E(Y|X)}$ is multiple-dimensional, OLS is certainly inadequate as it's not able to detect more than one directions.

Principle Hessian Directions

Principle Hessian Directions (PHD) is introduced by Li (Li, 1992, Cook 1998, Lue 2010). It improves the estimation of OLS, in the way that PHD can detect multiple directions in $\mathcal{S}_{E(Y|X)}$. The idea of PHD is to use the Hessian matrix of the conditional means function. Let $\Gamma = (\gamma_1, \cdots, \gamma_m)$ be a basis of $\mathcal{S}_{E(Y|Z)}$; thus, by definition, the regression function $E(Y|Z)$ is identical to $E(Y|\Gamma^\intercal Z)$. Let f be the link function. We have

$$E(Y|Z) := f(Z) = f(\Gamma^\intercal Z). \tag{2.19}$$

Then the $p \times p$ Hessian matrix of $f(Z)$, denoted by $H(Z)$, can be written as

$$H(Z) = \frac{\partial^2 f(Z)}{\partial Z_i \partial Z_j} = \Gamma \frac{\partial^2 f(Z)}{\partial(\Gamma^\intercal Z_i)\partial(\Gamma^\intercal Z_j)} \Gamma^\intercal. \tag{2.20}$$

Based on (2.20), we can use the eigenvectors of $E[H(Z)]$ to estimate Γ.

Proposition 2.1 *The rank of* $E[H(Z)]$ *is at most* m*; moreover, its eigenvectors that associate with nonzero eigenvalues are aligned with* $\mathrm{Span}(\Gamma)$*. In other words, the column space of* $E[H(Z)]$ *is unbiased for* $\mathcal{S}_{E(Y|Z)}$*.*

This suggests that the central mean subspace can be estimated using the mean Hessian. By Stein's Lemma (Stein, 1981), $E[H(Z)]$ can be computed from the *weighted covariance*

matrix, which we denote as Σ_{yxx},

$$\Sigma_{yzz} = E[(Y - \mu_y)ZZ^{\mathsf{T}}], \tag{2.21}$$

where $\mu_y = E(Y)$.

Proposition 2.2 *(Stein's Lemma) Suppose Z follows a standard normal distribution. Then the mean Hessian matrix $E[H(Z)]$ is identical to Σ_{yzz}, i.e.*

$$E[H(Z)] = \Sigma_{yzz}. \tag{2.22}$$

Therefore, PHD solves the following eigenvalue problem,

$$\Sigma_{yzz}\beta = \lambda\beta. \tag{2.23}$$

Suppose $(\beta_1, \cdots, \beta_m)$ be the eigenvectors from (2.23) with nonzero eigenvalues. Then we have

$$\text{Span}(\beta_1, \cdots, \beta_m) \subseteq \mathcal{S}_{E(Y|X)}. \tag{2.24}$$

Li (1992) proves that the result in (2.24) given that the predictor has an elliptical distribution. Cook and Li (2002) show similar result with even milder conditions, in which we only require a linearity mean condition (Assumption 2.1), and a condition on the conditional covariance.

Theorem 2.6 *Assume*

a $E(\alpha^{\mathsf{T}}Z|\Gamma^{\mathsf{T}}Z)$ *is linear in* $\Gamma^{\mathsf{T}}Z$*, and*

b $\text{cov}(Z|\Gamma^{\mathsf{T}}Z)$ *is uncorrelated with* Y*,*

then (2.24) is satisfied.

Algorithm of pHd:

1. Compute the sample means of X and Y respectively, and covariance matrix of X:

$$\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\Sigma}_{xx} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^{\mathsf{T}}.$$

2. Standardize the covariate by $Z_i = \hat{\Sigma}_{xx}^{-1/2}(X_i - \hat{\mu}_x)$. Then compute the weighted covariance matrix Σ_{yzz}:

$$\hat{\Sigma}_{yzz} = \frac{1}{n}\sum_{i=1}^{n}[(Y_i - \hat{\mu}_y)(Z_i - \hat{\mu})(Z_i - \hat{\mu})^{\mathsf{T}}].$$

3. Solve the eigenvalue problem:

$$\hat{\Sigma}_{yzz}\beta = \lambda\beta,$$

and let $\hat{\beta}_1, \cdots, \hat{\beta}_m$ be the eigenvectors corresponding to eigenvalues $\lambda_1 \geq \cdots \lambda_m > 0$.

The linearity condition is critical, to those we have discussed thus far, and it can be verified provided an elliptical distribution on the predictor - see Eaton (1986). Some SDR methods may need additional assumptions, such as PHD, SAVE, and other second-order moments methods that generally assume the constant variance condition (Assumption 2.2).

Iterative Hessian Transformation

Cook and Li (2002, 2004) introduce another SDR method for the central mean subspace called *Iterative Hessian Transformation* (IHT). IHT combines the ideas of OLS and pHd and provides estimation of the central mean space.

It's shown by Cook (1998) that, when a linear structure is presented in the model, PHD may fail to detect it. Let's first of all consider the following example.

Example 2.1 *Consider the multiple regression model*

$$Y = \beta_0 + \beta^{\mathsf{T}} Z + \epsilon, \tag{2.25}$$

where $Z \sim \text{MVN}$ *and* $\epsilon \perp\!\!\!\perp Z$ *with* $E(\epsilon) = 0$.

PHD searches the directions of the central subspace based on the mean Hessian matrix Σ_{yzz}. However, it's easy to see Σ_{yzz} is zero in this example; this implies that the central mean subspace may not be necessary found from PHD. Therefore, by incorporating OLS in PHD, IHT provides alternative estimation of central mean space such that it is possible to capture the linear features in the model.

Theorem 2.7 *Let Γ be a basis of $\mathcal{S}_{E(Y|Z)}$. Suppose $U : \mathbb{R} \mapsto \mathbb{R}$ and $V : \mathbb{R} \mapsto \mathbb{R}$ are both measurable functions of $\Gamma^{\mathsf{T}} Z$. If Assumption 2.1 is satisfied, we have*

$$E[(UY + V)Z] \in \mathcal{S}_{E(Y|Z)}, \tag{2.26}$$

provided that $(UY + V)Z$ is integrable.

This theorem suggests that one can estimate $\mathcal{S}_{E(Y|Z)}$ by an iterative procedure which we describe below.

Iterative process of estimating $\mathcal{S}_{E(Y|Z)}$:

1. Find a vector $\delta_0 \in \mathcal{S}_{E(Y|Z)}$; that is, the OLS. Denoted it by β_{ols}.

2. Select appropriate functions U and V, and then defined a new response via

$$Y^{(1)} = U(\delta_0^{\mathsf{T}} Z)Y + V(\delta_0^{\mathsf{T}} Z).$$

3. Compute the covariance between $Y^{(1)}$ and Z and denote it by δ_i,

$$\delta_1 = E(Y^{(1)}Z) = E\{[U(\delta_0^{\mathsf{T}} Z)Y + V(\delta_0^{\mathsf{T}} Z)]Z\}.$$

   For example, let $\delta_0 = \beta_{\text{ols}}$ and if $U(t) = t$ and $V(t) = -tE(Y)$, then $\delta_1 = \Sigma_{yzz}\beta_{\text{ols}}$.

4. Iteratively compute the new response $Y^{(j)}$ from

$$Y^{(j)} = U(\delta_{j-1}^{\mathsf{T}} Z)Y + V(\delta_{j-1}^{\mathsf{T}} Z), \quad j = 2, 3, \cdots,$$

and then the vector δ_j, the covariance between the updated response $Y^{(j)}$ and Z:

$$\delta_j = E(Y^{(j)}Z) = E\{[U(\delta_{j-1}^{\mathsf{T}}Z)Y + V(\delta_{j-1}^{\mathsf{T}}Z)]Z\}.$$

For example, $\delta_j = \Sigma_{yzz}^{j}\beta_{\text{ols}}$.

Corollary 2.1 *Suppose Assumption 2.1 holds. Then*

$$\text{Span}\{\Sigma_{yzz}^{j}\beta_{\text{ols}} : j = 0, 1, \cdots\} \subseteq \mathcal{S}_{E(Y|Z)}. \tag{2.27}$$

This corollary shows only a special case of IHT estimators. Considering all measurable functions U and V in the above process, one can produce a class of estimators. In any direction, the purpose here is to stress that IHT can be handled as an assembly work of OLS and PHD.

Algorithm of IHT:

1. Compute the OLS, $\hat{\beta}_{\text{ols}} = \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i$.
2. Compute the weighted covariance matrix (see algorithm of pHd), $\hat{\Sigma}_{yzz}$.
3. Compute $p \times p$ matrix $B = \left(\hat{\beta}_{\text{ols}}, \hat{\Sigma}_{yzz}\hat{\beta}_{\text{ols}}, \cdots, \hat{\Sigma}_{yzz}^{p-1}\hat{\beta}_{\text{ols}}\right)$.
4. Solve the following eigenvalue problem:

$$BB^{\mathsf{T}}\gamma = \lambda\gamma,$$

and let $\hat{\gamma}_1, \cdots, \hat{\gamma}_m$ be the eigenvectors corresponding to eigenvalues $\lambda_1 \geq \cdots \lambda_m > 0$.

Unlike PHD, IHT only assumes the linearity condition. Moreover, it is capable of detecting multiple directions on $\mathcal{S}_{E(Y|X)}$, which is generally not possible in OLS.

3 Nonparametric Methods

We have seen some SDR methods; some seeks for the directions in the central subspace $\mathcal{S}_{Y|X}$, while others estimate the central mean subspace $\mathcal{S}_{E(Y|X)}$. In general these methods are easy

to compute - one only needs to calculate the empirical estimate of the inverse conditional moments, and then proceeds to a singular value decomposition problem. However, these methods also make assumptions on the distribution of predictors. For example, SIR requires a linearity condition to be satisfied, while some second-order methods also need $\text{var}(X|\Gamma^{\mathsf{T}}X)$ to be constant.

Another limitation of inverse moment methods is from the slicing step. That is, it is rather impractical to slice the response that is with high dimensions; furthermore, in SIR, the number of slices may falsely set up a bond of the rank of the space being estimated.

Some nonparametric methods have been recently proposed to overcome these challenges, e.g. *Minimum Average Variance Estimation* (MAVE) (Xia et al., 2002, 2007, Li et al 2010). MAVE assembles the idea of nonparametric estimation and directional estimation in SDR.

Minimum Average Variance Estimation (MAVE)

MAVE searches for directions in the central mean subspace $\mathcal{S}_{E(Y|X)}$. It first of all approximates the regression function $E(Y|X)$ based on *Local linear Regression* (Fan and Gijbels, 1996). Let $E(Y|X) = g(X)$, then given x_0, we can estimate $G(X)$ using Taylor expansion

$$\begin{aligned} g(X) &\approx g(x_0) + g'(x_0)(X - x_0), &(2.28)\\ &:= \beta_0 + \beta^{\mathsf{T}}(X - x_0), &(2.29)\end{aligned}$$

where g' is the derivative of g. We then calculate the coefficients by solving the *weighted least squares*. Let $(g(x_0), g'(x_0)) = (\beta_0, \beta)$ and $(\hat{\beta}_0, \hat{\beta})$ be its sample counterpart. One can show that

$$(\hat{\beta}_0, \hat{\beta}) = \text{argmin}_{(\beta_0,\beta)} \sum_{i=1}^{n} \{Y_i - [\beta_0 + \beta^{\mathsf{T}}(X_i - x_0)]\}^2 w_{i0}, \tag{2.30}$$

where $w_{i0} \geq 0$ are some weights satisfying $\sum_i w_{io} = 1$. w's are determined via kernel functions; for example, given a kernel $K_h(\cdots)$, (h is the width parameter), w_{i0} is computed by

$$w_{i0} = \frac{K_h(x_i - x_0)}{\sum_i K_h(x_i - x_0)}. \tag{2.31}$$

Suppose Γ is a basis of $\mathcal{S}_{E(Y|X)}$, By definition, $E(Y|X) = E(Y|\Gamma^{\mathsf{T}}X)$ and we can approximate $E(Y|\Gamma^{\mathsf{T}}X)$ by the following equivalence

$$E(Y|\Gamma^{\mathsf{T}}X) = g(\Gamma^{\mathsf{T}}X) \approx \beta_0 + \beta^{\mathsf{T}}\Gamma^{\mathsf{T}}(X - x_0),$$

which implies that the weighted least squares can be written as

$$\sum_{i=1}^{n}\{Y_i - [\beta_0 + \beta^{\mathsf{T}}\Gamma^{\mathsf{T}}(X_i - x_0)]\}^2 w_{i0}^{\Gamma}, \tag{2.32}$$

where w_{i0}^{Γ} can be computed from

$$w_{i0}^{\Gamma} = \frac{K_h(\Gamma^{\mathsf{T}}(x_i - x_0))}{\sum_i K_h(\Gamma^{\mathsf{T}}(x_i - x_0))}.$$

Note (2.32) is an approximation to the conditional variance $\mathrm{cov}(Y|\Gamma^{\mathsf{T}}x_0)$; we denote it as $\sigma^2(\Gamma^{\mathsf{T}}x_0)$. To estimate Γ, one can minimize the expectation of $\sigma^2(\Gamma^{\mathsf{T}}x_0)$

$$\hat{E}[\sigma(\Gamma^{\mathsf{T}}x_0)] = \sum_{i=1}^{n}\sigma(\Gamma^{\mathsf{T}}X_i). \tag{2.33}$$

By substituting (2.33) with (2.32),

$$(\hat{\Gamma}, \hat{\beta}_{0i}, \hat{\beta}_i) = \mathrm{argmin}_{(\Gamma,\beta_{0i},\beta_i)} \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\{Y_j - [\beta_{0i} + \beta_i^{\mathsf{T}}\Gamma^{\mathsf{T}}(X_j - X_i)]\}^2 w_{ji}^{\Gamma}\right), \tag{2.34}$$

where Γ can be estimated by an iterative weighted least square algorithm.

Unlike inverse regression methods in SDR (See Section 2.1), MAVE considers the forward regression $E(Y|X)$. By solving an optimization problem in (2.34), MAVE provides the estimation of directions in $\mathcal{S}_{E(Y|X)}$. In general, MAVE does not impose any conditions on the distribution of predictors.

Algorithm of MAVE:

1. Compute $\Gamma^{(0)}$ from moment methods, e.g. pHd or IHT. Then set $t = 0$.

2. Let $\Gamma = \Gamma^{(t)}$, then compute $(\beta_{0i}^{(t)}, \beta_i^{(t)})$ in the optimization problem in (2.34) by:

$$\begin{pmatrix} \beta_{0i}^{(t)} \\ \beta_i^{(t)} \end{pmatrix} = \left\{ \sum_{j=1}^{n} w_{ji}^{\Gamma} \begin{pmatrix} 1 \\ \Gamma^{\mathsf{T}}(X_j - X_i) \end{pmatrix} \begin{pmatrix} 1 \\ \Gamma^{\mathsf{T}}(X_j - X_i) \end{pmatrix}^{\mathsf{T}} \right\}^{-1} \times \left\{ \sum_{j=1}^{n} w_{ji}^{\Gamma} \begin{pmatrix} 1 \\ \Gamma^{\mathsf{T}}(X_j - X_i) \end{pmatrix} Y_j \right\}.$$

3. Let $(\beta_{0i}, \beta_i) = (\beta_{0i}^{(t)}, \beta_i^{(t)})$, then update $\Gamma^{(t)} = (\gamma_1^{(t)}, \cdots, \gamma_d^{(t)})$ by (which is another weighted least squares):

$$\begin{pmatrix} (\gamma_1^{(t+1)})^{\mathsf{T}} \\ \vdots \\ (\gamma_d^{(t+1)})^{\mathsf{T}} \end{pmatrix} = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ji}^{\Gamma} X_{ij} X_{ij}^{\mathsf{T}} \right\}^{-1} \times \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ji}^{\Gamma} X_{ij} (Y_j - \beta_{0i}) \right\},$$

where $X_{ij} = \beta_i \otimes (X_j - X_i)$.

4. Repeat 2 & 3 until the parameters converge.

Average Derivative Estimation and Outer product of Gradient

Average Derivative Estimation (ADE) (Härdle and Stoker 1989) is another SDR approach based on nonparametric regression. ADE estimates the directions in central mean subspace based on a single-index model:

$$Y = m(X) + \epsilon, \tag{2.35}$$

where the conditional mean $E(Y|X) = m(X) = g(\delta^{\mathsf{T}} X)$. In the above model the *average derivative* of the regression function can be calculated by

$$E(\partial m(X)/\partial X),$$

and it depends on the parameter δ because of the following relation

$$E(\frac{\partial m(X)}{\partial X}) = E[\frac{\partial m(X)}{\partial(\delta^{\mathsf{T}} X)}]\delta. \tag{2.36}$$

Let $\beta = E(\frac{\partial m(X)}{\partial X})$. Then δ and β are identical up to a scalar. Therefore, the estimation of $\mathcal{S}_{E(Y|X)}$ relies on the estimation of the average gradient of $E(Y|X)$. Again, by local linear regression we can estimate the derivative $\partial m(X)/\partial X$. ADE provides a simple procedure for estimating central mean subspace $\mathcal{S}_{E(Y|X)}$; however, it can not extract more than one direction.

Outer product of Gradient (OPG) (Härdle and Tsybakov, 1991, Xia et al 2002, 2007, Hristache, 2001) generalizes ADE and is capable of estimating multiple directions, i.e. when $\dim(\mathcal{S}_{E(Y|X)}) \geq 1$. Suppose $\Delta = (\delta_1, \cdots, \delta_d)$ is a basis of $\mathcal{S}_{E(Y|X)}$; by definition the conditional mean $E(Y|X)$ is identical to $E(Y|\Delta^{\mathsf{T}} X)$. Let $g(\Delta^{\mathsf{T}} X) = E(Y|\Delta^{\mathsf{T}} X)$. OPG computes the *expected outer product* of $\partial m(X)/\partial X$, the derivative of $m(X)$ with respect to X,

$$E\left\{\left(\frac{\partial m(X)}{\partial X}\right)\left(\frac{\partial m(X)}{\partial X}\right)^{\mathsf{T}}\right\} = \Delta E\left\{\left(\frac{\partial m(X)}{\partial(\Delta^{\mathsf{T}} X)}\right)\left(\frac{\partial m(X)}{\partial(\Delta^{\mathsf{T}} X)}\right)^{\mathsf{T}}\right\}\Delta^{\mathsf{T}}. \tag{2.37}$$

Suppose the central subspace is the column space of Δ. Then one can estimate Δ (up to scalars) by computing the eigenvectors of the expected outer product of gradient. The following result verifies the theoretical foundation of OPG.

Proposition 2.3 *Suppose* $\mathrm{Span}(\Delta) = \mathcal{S}_{E(Y|X)}$ *and* Δ *is of rank* d. *Suppose the expected outer product of* $\partial m(X)/\partial X$ *can be decomposed as*

$$E\left\{\left(\frac{\partial m(X)}{\partial X}\right)\left(\frac{\partial m(X)}{\partial X}\right)^{\mathsf{T}}\right\} = B\Lambda B^{\mathsf{T}},$$

where $B = (\beta_1, \cdots, \beta_p)$ *and* $\Lambda = diag(\lambda_1, \cdots, \lambda_p)$ *(in decreasing order) are the eigenvectors and eigenvalues respectively. Then we have*

1. *There are at most* d *nonzero eigenvalues, i.e.* $\exists d'$ *with* $(d' \leq d)$ *s.t.* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda'_d > 0 = \lambda_{d'+1} = \cdots = \lambda_p$.

2. *Suppose there are d' nonzero eigenvalues and their corresponding eigenvectors are $(\beta_1, \cdots, \beta_{d'})$. These eigenvectors locate within $\mathcal{S}_{E(Y|X)}$*

$$\text{Span}(\beta_1, \cdots, \beta_{d'}) \subseteq \mathcal{S}_{E(Y|X)}.$$

Algorithm of OPG:

1. Compute $[\widehat{g(X_i)}, \widehat{g'(X_i)}]$ by:

$$\begin{bmatrix} \widehat{m(X_i)} \\ \widehat{m'(X_i)} \end{bmatrix} = \left\{ \sum_{j=1}^{n} w_{ji} \begin{pmatrix} 1 \\ (X_j - X_i) \end{pmatrix} \begin{pmatrix} 1 \\ (X_j - X_i) \end{pmatrix}^{\intercal} \right\}^{-1} \times \left\{ \sum_{j=1}^{n} w_{ji} \begin{pmatrix} 1 \\ (X_j - X_i) \end{pmatrix} Y_j \right\},$$

   where w_{ji} is calculated from

$$w_{ji} = \frac{K_h(X_j - X_i)}{\sum_j K_h(X_j - X_i)}.$$

2. Take the average of $[\widehat{m'(X_i)}][\widehat{m'(X_i)}]^{\intercal}$:

$$\Lambda = \frac{1}{n} \sum_{i=1}^{n} [\widehat{m'(X_i)}][\widehat{m'(X_i)}]^{\intercal}.$$

3. Find the eigen decomposition of Λ. Suppose $\hat{B} = (\hat{\beta}_1, \cdots, \hat{\beta}_d)$ is the collection of eigenvectors, which have nonzero eigenvalues.

4. Estimate $\mathcal{S}_{E(Y|X)}$ by $\text{Span}(\hat{B})$.

Similar to PHD, OPG computes the expected product of the gradient, and then use its principle components to estimate the directions in $\mathcal{S}_{E(Y|X)}$; however, OPG considers the outer product of the first-order derivative, while PHD takes advantages of the second-order derivative (see (2.20). Besides, PHD extracts the eigenvectors from the a weighted covari-

ance matrix, where OPG finds the eigenvectors from the gradient. From the theoretical perspective, OPG literally requires no conditions on the predictors, unlike PHD.

2.2 Regression Class

In linear SDR, the goal is to find a set of vectors that span $\mathcal{S}_{Y|X}$. From Definition 2.1, a matrix Γ is called unbiased if satisfying $\text{span}(\Gamma) \subseteq \mathcal{S}_{Y|X}$; if $\text{span}(\Gamma) = \mathcal{S}_{Y|X}$, then Γ is exhaustive. Assume there exists a matrix B such that the central SDR σ-field $\mathcal{G}_{Y|X}$ coincides with the σ-field generated by $B^{\mathsf{T}}X$; that is,

$$\mathcal{G}_{Y|X} = \sigma(B^{\mathsf{T}}X). \tag{2.38}$$

Therefore, $\Gamma^{\mathsf{T}}X$ is a linear function of $B^{\mathsf{T}}X$ when Γ is unbiased; $\Gamma^{\mathsf{T}}X$ is a one-to-one transformation if Γ is exhaustive.

We provide two additional definitions that generalize Definition 2.1. In the nonlinear setting, we follow the same logic as in the linear setting but remove the linear requirement in (2.38), i.e. the central SDR σ-field doesn't have to be generated by the linear function of X. Part of the following definition is also given in Li, Artemiou, and Li (2011).

Definition 2.4 *A class of function in $L_2(P_X)$ is unbiased for $\mathfrak{S}_{Y|X}$ if each of its members is measurable with respect to $\mathcal{G}_{Y|X}$, 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. We first of all consider a sub-class of functions in $L_2(P_X)$,

$$L_2(P_X) \ominus (L_2(P_X) \ominus L_2(P_Y)), \tag{2.39}$$

where $\ominus$ is the direct difference. 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 $\mathcal{S}_1 \cap \mathcal{S}_2^{\perp}$. We call this class *regression class*, and its orthogonal complement $L_2(P_X) \ominus L_2(P_Y)$ *residual class*. It turns out that regression class is unbiased.

Before demonstrating the main theorem we deploy the following lemma, which provides a characterization of the orthogonal complement of $\mathcal{M}_{\mathcal{G}}$ and will be used many times in the subsequent development.

Lemma 2.1 *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 $\mathcal{M}_{\mathcal{G}}$ (in symbols, $f \perp \mathcal{M}_{\mathcal{G}}$) if and only if $E[f(U)|\mathcal{G}] = 0$.*

PROOF.

If $f \perp \mathcal{M}_{\mathcal{G}}$, then $E[f(X)g(X)] = 0$ for all $g \in \mathcal{M}_{\mathcal{G}}$. In particular

$$\int f(X) I_B(X) dP = \int 0 I_B(X) dP$$

for all $B \in \mathcal{G}$, which means 0 is a version of $E[f(X)|\mathcal{G}]$. Conversely, if $E[f(X)|\mathcal{G}] = 0$, then for any $g \in \mathcal{M}_{\mathcal{G}}$,

$$\langle f, g \rangle_{L_2(P_X)} = E[f(X)g(X)] = E[E(f(X)|\mathcal{G})g(X)] = 0.$$

Hence $f \perp \mathcal{M}_{\mathcal{G}}$. □

There is a geometric interpretation of the above lemma. Any conditional expectation can be seen as an L_2-projection; therefore, for any function f such that f is orthogonal to $\mathcal{M}_{\mathcal{G}}$, the conditional expectation of f on $\mathcal{G}$ is zero. The following theorem says that regression class is unbiased for $\mathfrak{S}_{Y|X}$.

Theorem 2.8 *If the family $\{\Pi_y : y \in \Omega_Y\}$ is dominated by a σ-finite measure, then*

$$L_2(P_X) \ominus [L_2(P_X) \ominus L_2(P_Y)] \subseteq \mathfrak{S}_{Y|X}. \tag{2.40}$$

PROOF. The statement in (2.40) is equivalent to

$$L_2(P_X) \ominus \mathfrak{S}_{Y|X} \subseteq L_2(P_X) \ominus L_2(P_Y).$$

If $f \in L_2(P_X) \ominus \mathfrak{S}_{Y|X}$, then, by Lemma 2.1, $E[f(X)|\mathcal{G}_{Y|X}] = 0$. Since $\mathcal{G}_{Y|X}$ is a sufficient σ-field,

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

By Lemma 2.1 again, $f \perp \mathcal{M}_Y$. Because $\mathcal{M}_Y = L_2(P_Y)$, 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 classical 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."

2.3 A Sufficient and Complete Dimension Reduction Class

After showing that the regression class (2.39) is unbiased, we turn to the task of investigating 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 and sufficient dimension reduction (CSDR) class.

Definition 2.5 *A sufficient* SDR *class $\mathcal{M}_\mathcal{G}$ within $L_2(P_X)$ is complete if, for any $g \in \mathcal{M}_\mathcal{G}$,*

$$E[g(X)|Y] = 0 \ \ a.s.\ P \ \Rightarrow \ g(X) = 0 \ a.s.\ P.$$

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

$$Y = h(X) + \varepsilon, \tag{2.41}$$

where $\varepsilon \perp\!\!\!\perp X$ and $E(\varepsilon) = 0$. Then $\mathcal{M}_{h(X)}$ 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. Suppose $m \in \mathcal{M}_{h(X)}$ 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 2.1, for any $f \in L_2(P_Y)$, we have

$$E[g(U)f(Y)] = 0. \tag{2.42}$$

In particular, $E[g(U)e^{it^\top Y}] = 0$, where $i = \sqrt{-1}$. Because $U \perp\!\!\!\perp \varepsilon$, this implies

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

Hence $E[g(U)e^{it^\top U}] = 0$. 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 (2.41) 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, Cook, and Li, 2008), are special cases of (2.41). Thus, complete and sufficient dimension reduction classes 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 2.5 *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 $\mathcal{M}_{g(X)}$ 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, $\mathcal{M}_{g(X)}$ is a sufficient dimension reduction class. Let $u \in \mathcal{M}_{g(X)}$. 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. Then

$$E[v(Y + \varepsilon)|Y] = 0$$

a.s. P. Because $Y \perp\!\!\!\perp \varepsilon$, the above implies $P_Y(\{y : Ev(y + \varepsilon) = 0\}) = 1$. In other words

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

a.s. P_Y. This implies

$$\int v(t) e^{-t^\intercal \Sigma^{-1} t/2} e^{y^\intercal \Sigma^{-1} t} dt = 0 \Longrightarrow \int v(\Sigma s) e^{-s^\intercal \Sigma s/2} e^{y^\intercal s} ds = 0$$

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

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

where λ is the Lebesgue measure on $\mathbb{R}^q$. But, because $e^{-s^\tau \Sigma s/2} > 0$, 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 nonparametric forward and inverse regressions.

2.4 Minmal Sufficiency in Nonlinear SDR

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

Proof. Since $\mathcal{M}_\mathcal{G}$ is a sufficient dimension reduction class, $\mathfrak{S}_{Y|X} \subseteq \mathcal{M}_\mathcal{G}$. To show $\mathcal{M}_\mathcal{G} \subseteq \mathfrak{S}_{Y|X}$, let $\mathcal{G}_{Y|X}$ be the central σ-field. Let $h \in \mathcal{M}_\mathcal{G}$. Since $\mathcal{G}_{Y|X}$ is sufficient,

$$E[h(X)|Y] = E\{E[h(X)|\mathcal{G}_{Y|X}]|Y\}$$

Since $\mathcal{G}_{Y|X} \subseteq \mathcal{G}$, we have $E[h(X)|\mathcal{G}_{Y|X}] = E\{E[h(X)|\mathcal{G}_{Y|X}]|\mathcal{G}\}$. Hence

$$E[h(X)|Y] = E\{E[E(h(X)|\mathcal{G}_{Y|X})|\mathcal{G}]|Y\}.$$

Hence $E[h(X) - E(E(h(X)|\mathcal{G}_{Y|X})|\mathcal{G})|Y] = 0$, which, together with Since $h \in \mathcal{M}_{\mathcal{G}}$, implies

$$E\{E[h(X) - E(h(X)|\mathcal{G}_{Y|X})|\mathcal{G}]|Y\} = 0.$$

Since $E[h(X) - E(h(X)|\mathcal{G}_{Y|X})|\mathcal{G}] \in \mathcal{M}_{\mathcal{G}}$ and $\mathcal{M}_{\mathcal{G}}$ is complete, we have

$$E[h(X) - E(h(X)|\mathcal{G}_{Y|X})|\mathcal{G}] = 0 \text{ a.s. } P.$$

By Lemma 2.1, this implies $h(X) - E(h(X)|\mathcal{G}_{Y|X}) \perp \mathcal{M}_{\mathcal{G}}$. But we also know that $h(X) - E(h(X)|\mathcal{G}_{Y|X}) \in \mathcal{M}_{\mathcal{G}}$. Hence $h(X) - E(h(X)|\mathcal{G}_{Y|X}) = 0$ a.s. P, which implies $h \in \mathcal{G}_{Y|X}$. □

2.5 Exhaustiveness of Regression Class

Here we consider again the unbiased regression class (2.39) and show that it is exhaustive for $\mathfrak{S}_{Y|X}$ under completeness. As we will see in Chapter 4, this characterization is very consequential, because it allows us to formulate a generalized version of sliced inverse regression (Li, 1991) that, under completeness, recovers the whole central class.

Theorem 2.10 *Suppose that the family $\{\Pi_y : y \in \Omega_Y\}$ is dominated by a σ-finite measure. If a complete and sufficient dimension reduction class exists, then*

$$\mathfrak{S}_{Y|X} = L_2(P_X) \ominus [L_2(P_X) \ominus L_2(P_Y)].$$

Proof. In Theorem 2.8 we have already shown that the right hand side is contained in the left hand side. We now show that $\mathfrak{S}_{Y|X} \subseteq L_2(P_X) \ominus [L_2(P_X) \ominus L_2(P_Y)]$, or

$$L_2(P_X) \ominus \mathfrak{S}_{Y|X} \supseteq L_2(P_X) \ominus L_2(P_Y). \tag{2.43}$$

Suppose $f \in L_2(P_X)$ and $f \perp L_2(P_Y)$. Then, by Lemma 2.1, $E(f(X)|Y) = 0$ a.s. P. Because $E(f(X)|Y) = [E(E(f(X)|\mathcal{G}_{Y|X})|Y)]$, we have

$$E\{[E(f(X)|\mathcal{G}_{Y|X})]|Y\} = 0 \text{ almost surely } P.$$

Since $\mathfrak{S}_{Y|X}$ is complete, $E(f(X)|\mathcal{G}_{Y|X}) = 0$ a.s. P, which, by Lemma 2.1, implies $f \perp \mathfrak{S}_{Y|X}$, proving (2.43). □

In the following, we will indicate the regression class (2.39) as $\mathfrak{C}_{Y|X}$. The above developments thus show that, when a CSDR class exists, it coincides with our ultimate object of interest i.e. the central class $\mathfrak{S}_{Y|X}$, and is exhaustively captured by the regression class $\mathfrak{C}_{Y|X}$.

Chapter 3

Operators in Reproducing Kernel Hilbert Spaces

3.1 Basic Definition and Reproducing Property

A Hilbert space $\mathcal{H}$ is a complete and possibly infinitely-dimensional vector space with its norm $||\cdot||_{\mathcal{H}}$ induced by an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, i.e. $||\cdot||_{\mathcal{H}} = \sqrt{\langle\cdot,\cdot\rangle_{\mathcal{H}}}$. An inner product satisfies the following properties:

1. Symmetry: for all $s, t \in \mathcal{H}$, $\langle s,t\rangle_{\mathcal{H}} = \langle t,s\rangle_{\mathcal{H}}$.

2. Linearity: for all $s, t, u \in \mathcal{H}$ and for all $a, b \in \mathbb{R}$, $\langle as+bt, u\rangle_{\mathcal{H}} = a\langle s,u\rangle_{\mathcal{H}} + b\langle t,u\rangle_{\mathcal{H}}$.

3. Positive definiteness: for each $s \in \mathcal{H}$, $\langle s,s\rangle_{\mathcal{H}} \geq 0$ and $\langle s,s\rangle_{\mathcal{H}} = 0$ implies $s = 0$.

Examples of Hilbert spaces include $\mathbb{R}^n$, l_2 and L_2 classes, with respect to their inner products. A Reproducing Kernel Hilbert Spaces (RKHS) is essentially a Hilbert space that is more restricted and smoothed in the sense that all its evaluation functionals are continuous, i.e. let $\mathcal{H}$ be an RKHS of functions defined on Ω_X, then the mapping $l_x : \mathcal{H} \to \mathbb{R}$ is continuous for each $x \in \Omega_X$. Given that the continuity of an evaluation functional can be implied by its boundedness, we provide a formal definition of RKHS.

Definition 3.1 *Let $\mathcal{H}$ be a Hilbert space of functions on Ω_X. Given each $x \in \Omega_X$, a functional l_x can be defined on $\mathcal{H}$. Then $\mathcal{H}$ is also an RKHS if the following holds,*

$$\sup_{f\in\mathcal{H}} \frac{|l_x(f)|}{||f||_{\mathcal{H}}} = \sup_{f\in\mathcal{H}} \frac{|f(x)|}{||f||_{\mathcal{H}}} < \infty. \tag{3.1}$$

An immediate result from this definition is that norm convergence in RKHS implies pointwise convergence. That is, when a function $f \in \mathcal{H}$ is close to another function g in $\mathcal{H}$-norm, $f(x)$ is also close to $g(x)$ for all x in the domain. This is one of the nice features of RKHS where it's generally not true in L_2 spaces.

Definition 3.2 *Let κ_X be a function defined on $\Omega_X \times \Omega_X$:*

$$(x, y) \mapsto \kappa_X(x, y). \tag{3.2}$$

Then $\mathcal{H}$ is an RKHS if there exists κ_X such that

a. $\kappa_X(\cdot, x) \in \mathcal{H}$ for each $x \in \Omega_X$, and

b. $\langle f, \kappa_X(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ for each $f \in \mathcal{H}$ and $x \in \Omega_X$.

In particular, such κ_X is called reproducing kernel. And assertion b. is called reproducing property*; that is, the evaluation of function f at point x can be reproduced by the inner product between f and $\kappa_X(\cdot, x)$.*

Theorem 3.1 *Let $\mathcal{H}$ be a Hilbert space. Then the following two statements are equivalent,*

1. $\mathcal{H}$ contains a reproducing kernel.

2. All the evaluation functional l_x, for each $x \in \mathcal{H}$ are continuous.

Proof. Let us first show 1 implies 2. Suppose κ_X is the reproducing kernel of $\mathcal{H}$. Then for each x, the evaluation functional l_x can be represented as

$$l_x(f) = f(x) = \langle f, \kappa_X(\cdot, x)\rangle_{\mathcal{H}}.$$

Then the boundedness (or continuity) of l_x can be derived from Cauchy-Schwarz inequality.

$$\begin{aligned} ||l_x|| = \sup_{f\in\mathcal{H}} \frac{|l_x(f)|}{||f||_{\mathcal{H}}} &= \sup_{f\in\mathcal{H}} \frac{|\langle f, \kappa_X(\cdot,x)\rangle_{\mathcal{H}}|}{||f||_{\mathcal{H}}} \\ &\leq \sup_{f\in\mathcal{H}} \frac{||f||_{\mathcal{H}} \cdot \sqrt{|\kappa_X(x,x)|}}{||f||_{\mathcal{H}}} = \sqrt{|\kappa_X(x,x)|}. \end{aligned}$$

Conversely, since all the functionals l_x is continuous, by Riesz's representation theorem for each $f \in \mathcal{H}$ there exists a $\kappa_X^\star(\cdot) \in \mathcal{H}$ such that

$$f(x) = l_x(f) = \langle f, \kappa_X^\star(\cdot)\rangle_{\mathcal{H}}.$$

Let $\kappa_X^\star(\cdot)$ be the same as $\kappa_X(\cdot, x)$ and κ_X is the reproducing kernel. □

The second definition provides alternative way to construct RKHS as long as a proper kernel function is presented. On the other hand, verifying that a space of functions is an RKHS should become more apparent under Definition 3.2. Actually, the property of being a reproducing kernel is equivalent to the property of being a *positive-definite function*, which will be introduced in the next section. Let us now conclude this section by probing a few examples of both RKHS and non-RKHS.

Example 3.1 *Let $\mathcal{H}$ be a Hilbert space and $(e_1, \cdots, e_n)$ be an orthogonal system of $\mathcal{H}$. Consider the following function κ_X defined on $\Omega_X \times \Omega_X$. For each $x, y \in \Omega_X$,*

$$(x, y) \mapsto \sum_{i=1}^{n} e_i(x)e_i(y).$$

We first examine the assertions in Definition 3.2. Assertion a. is obvious since for each $x \in \Omega_X$, $\kappa_X(\cdot, x) = \sum_{i=1}^n e_i(x)e_i(\cdot)$ which belongs to $\mathcal{H}$. Note that all the elements in $\mathcal{H}$ can be written as

$$f(\cdot) = \sum_{i}^{n} \alpha_i e_i(\cdot),$$

where $\alpha_i \in \mathbb{R}$ for $1 \leq i \leq n$. Therefore, the inner product of f and $\kappa_X(\cdot, x)$ is

$$\begin{aligned} \langle f, \kappa_X(\cdot, x)\rangle_{\mathcal{H}} &= \langle \sum_{i=1}^{n} \alpha_i e_i(\cdot), \sum_{i=1}^{n} e_i(x) e_i(\cdot)\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{n} \alpha_i e_i(x) = f(x). \end{aligned}$$

The above example claims that any finite-dimensional Hilbert space posses a reproducing kernel and hence is an RKHS. Next example is based on a sub-class of functions in L_2. When some nice features, e.g. continuity, are considered in L_2 spaces, we can also build up an RKHS.

Example 3.2 *Let $\mathcal{C}^1 = \{h \in L_2(\mathcal{R}) : h$ is absolutely continuous; $h' \in L_2(\mathcal{R})\}$ where h' is the derivative of h. $\mathcal{C}^1$ is a Hilbert space with the the inner product, for $h_1, h_2 \in \mathcal{C}^1$,*

$$\langle h_1, h_2\rangle_{\mathcal{C}^1} = \int_{\mathbb{R}} (h_1(x)h_2(x) + h_1'(x)h_2'(x))dx.$$

Moreover, $\mathcal{C}^1$ is also an RKHS with its reproducing kernel function κ

$$\kappa(x, y) = \frac{1}{2}\exp(-|x - y|),$$

for all $x, y \in \mathcal{R}$.

For any fixed $y \in \mathbb{R}$, $\kappa(x, y)$ is a function of x. It's easy to see the first two order derivatives of $\kappa(x, y)$ with respect to x have the following forms

$$\begin{aligned} &\frac{\partial \kappa(x, y)}{\partial x} = \text{sgn}(x - y)\kappa(x, y), \text{ and} \\ &\frac{\partial \kappa^2(x, y)}{\partial x^2} = \kappa(x, y), \end{aligned}$$

except when $x = y$. To show the space $\mathcal{C}^1$ is an RKHS, we first of all consider the following integral

$$\begin{aligned}\int_{-\infty}^{\infty} h_1'(x)h_2'(x)dx &= \int_{-\infty}^{y} h_1'(x)h_2'(x)dx + \int_{y}^{\infty} h_1'(x)h_2'(x)dx \\ &= \left(h_1(x)h_2'(x)|_{-\infty}^{y} - \int_{-\infty}^{y} h_1(x)h_2''(x)dx\right) + \left(h_1(x)h_2'(x)|_{y}^{\infty} - \int_{y}^{\infty} h_1(x)h_2''(x)dx\right) \\ &= h_1(y) - \int_{-\infty}^{\infty} h_1(x)h_2(x)dx.\end{aligned}$$

The second equality comes from integration by parts. The last equality can be derived by replacing $h_2(x)$ by $\kappa(x, y)$. Therefore, we have shown the reproducing property

$$\langle h_1, \kappa(\cdot, y)\rangle_{\mathcal{C}^1} = \int_{-\infty}^{\infty} (h_1(x)h_(x) + h_1'(x)h_2'(x))dx = h_1(y),$$

for any $y \in \mathbb{R}$.

The above kernel function is an *exponential* function which is commonly used in the applications of RKHS approaches. This example demonstrates how the RKHS can be constructed when applying this type of kernel function; on the other hand, it also implicitly points out what type of functions are being considered as the searching domain in which the targeting function is located.

Example 3.3 $L_2([0, 1]) = \{f : f$ *is square integrable over* $[0, 1]\}$.

The purpose of this example is to show that L_2 space is not an RKHS. We first consider a sequence of functions $\{x^n,\ n \geq 1\}$. By direct integration it can be shown this sequence converges to null function in L_2-norm. However, the point-wise convergence doesn't hold in this example, e.g. let $x = 1$, which implies that $L_2[(0, 1)]$ is not an RKHS.

3.2 Postive Type Function

We have explored some basic properties of RKHS, including its definition and a few examples characterizing both RKHS and non-RKHS. A Hilbert space is said to be an RKHS when it posses a reproducing kernel. Next, we study the question that, for an arbitrary function

what condition is required to be a reproducing kernel. When such function exists, it can induce a mapping from the elements of its domain to a space of functions that is essentially an RKHS. This mapping is the so called *kernel map* and it provides an explicit way to construct RKHS.

Definition 3.3 (Positive Type Function) *A function $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$ is said to be a positive type function if the following conditions hold,*

a. κ_X is symmetric, i.e. for $x, y \in \Omega_X$ $\kappa_X(x, y) = \kappa_X(y, x)$.

b. κ_X is positive definite; that is, for any $x_i \in \Omega_X$, and $\alpha_i \in \mathbb{R}$, $1 \leq i \leq n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \kappa_X(x_i, x_j) \geq 0.$$

We now show that an RKHS can be constructed based on a positive type function. Let κ_X be positive type satisfying a. and b. in Definition 3.3. Suppose Φ is a kernel map, $\Phi : x \mapsto \kappa_X(\cdot, x)$; that is, to associate a function $\kappa_X(\cdot, x)$ to each $x \in \Omega_X$. Suppose $\mathcal{H}_0$ is a vector space contains all the linear combinations of $\kappa_X(\cdot, x)$, i.e.

$$\mathcal{H}_0 = \left\{ \sum_{i=1}^{n} \alpha_i \kappa_X(\cdot, x_i) : (\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n,\ n \geq 1 \right\}.$$

Theorem 3.2 *Let $\mathcal{H}$ be the closure of $\mathcal{H}_0$, where the closure is under the norm operation induced by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$,*

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \kappa_X(x_i, x_j),$$

for any $f = \sum_{i=1}^{n} \alpha_i \kappa_X(\cdot, x_i)$ and $g = \sum_{j=1}^{m} \beta_j \kappa_X(\cdot, x_j)$. Then $\mathcal{H}$ is an RKHS with K as its reproducing kernel.

PROOF. We first examine that $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is a valid inner product. Both symmetry and linearity are immediate. To check the positive definiteness of $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, we need to use the property of

reproducing kernel, which is true by definition

$$\langle f, \kappa_X(\cdot, x)\rangle_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_i \kappa_X(x, x_i) = f(x).$$

Therefore, by Cauchy-Schwartz inequality and the reproducing property of κ_X,

$$|f(x)| = |\langle f, \kappa_X(\cdot, x)| \leq \sqrt{\kappa_X(x, x)}\sqrt{\langle f, f\rangle_{\mathcal{H}}}.$$

This implies that $\langle f, f\rangle_{\mathcal{H}} = 0 \to f = 0$. □

The above theorem claims that an RKHS is associated with a positive type function that has the reproducing property. Actually, it's also true in the reversed direction, i.e. any reproducing kernel function is of positive type. This can be easily verified by observing that, for any $f = \sum_{i=1}^{n} \alpha_i \kappa_X(\cdot, x_i)$,

$$0 \leq \langle f, f\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \kappa_X(x_i, x_j).$$

3.3 Kerenl Trick and Mercer's Theorem

We have shown an RKHS can be constructed when a positive type function is provided - see Theorem 3.2; furthermore, this construction has the property that the inner product of the RKHS is identical to the value of the kernel function. This is a critical property to the development of the kernel methods. The inner product measures the similarity between data points and very often can serve the solution to some linear methods; that is, the process of optimization involves only the calculation of $X_i^{\mathsf{T}} X_j$. When a linear method fits in this situation, we can simply replace $X_i^{\mathsf{T}} X_j$ by the value of $\kappa(X_i, X_j)$; by doing so, the original method is given the nonlinear interpretation because it is carried out in an RKHS rather than the space of X. This technique is the *Kernel Trick* which we will study further in this section.

Support Vector Machines Recently the development of Support Vector Machines (SVM) (Cristianini and Shawe-Taylor 2000, Hofmann et al, 2008 and Hastie et al, 2009) has become rather popular in statistics community. SVM initialized with a classification problem. The criterion behind is the *optimal separating hyperplan*, or sometimes called *maximal margin*. The purpose of SVM is to seek a good hyperplane that best separates the points from different classes. One can do so by maximizing the margins of the hyperplanes that separate different classes.

Let's consider a binary classification case, e.g. $y_i = \pm 1$. Suppose two classes are linearly separable. Given a linear function $f(x) = x^{\mathsf{T}}\beta + \alpha$, a separating hyperplane can be written as

$$\{x : f(x) = x^{\mathsf{T}}\beta + \alpha = 0, \}. \tag{3.3}$$

where β is the slope and α is the intercept. Since this is a separable case such $f(x)$ can always be found with following property:

$$x_i^{\mathsf{T}}\beta + \alpha \geq +\frac{C}{2} \quad \forall\ y_i = +1 \tag{3.4}$$

$$x_i^{\mathsf{T}}\beta + \alpha \leq -\frac{C}{2} \quad \forall\ y_i = -1. \tag{3.5}$$

The distance between the two hyperplanes, aka *margin*, in (3.4) and (3.5) is $C/||\beta||$. Therefore, SVM is to find the optimal β, in the sense that one can get the largest margin. This can be accomplished by solving the problem

$$\max_{\beta,\alpha} \frac{C}{||\beta||} \quad \text{subject to} \quad \frac{C}{2} - y_i(x_i^{\mathsf{T}}\beta + \alpha) \leq 0, \quad \forall i. \tag{3.6}$$

After properly scaling, one can show that (3.6) is equivalent to

$$\min_{\beta,\alpha} ||\beta|| \quad \text{subject to} \quad 1 - y_i(x_i^{\mathsf{T}}\beta + \alpha) \leq 0. \tag{3.7}$$

Nonetheless, most of the real world datasets are almost not possible to be linearly separable. Let's now turn to the tasks with the non-separable data. That is, classes are apparently overlapped in the sense there are certain points that locate on the other side of

the hyperplanes in (3.4) and (3.5). The relations in (3.4) and (3.5) obviously don't hold in the non-separable case; to this end slack variables $\xi = (\xi_1, \xi_2, \cdots, \xi_n)$ are introduced to increase the flexibility.

$$x_i^{\mathsf{T}}\beta + \alpha \geq +\frac{C}{2} - \xi_i \quad \forall\ y_i = +1 \tag{3.8}$$

$$x_i^{\mathsf{T}}\beta + \alpha \leq -\frac{C}{2} + \xi_i \quad \forall\ y_i = -1, \tag{3.9}$$

where all $\xi_i \geq 0$. One additional constraint on the slack variables, $\sum_i \xi_i \leq D$ where D is constant, is imposed. Therefore, we now consider a new optimization problem

$$\min_{\beta,\alpha} ||\beta|| \quad \text{subject to} \quad \left[\begin{array}{l} 1 - \xi_i - y_i(x_i^{\mathsf{T}}\beta + \alpha) \leq 0. \\ \quad \xi \geq 0, \sum_i \xi_i \leq D, \end{array} \right. \tag{3.10}$$

It can be shown that (3.10) is computationally equivalent to

$$\min_{\beta,\alpha} \frac{||\beta||^2}{2} + C\sum_i \xi_i \quad \text{subject to} \quad \xi_i \geq 0, 1 - \xi_i - y_i(x_i^{\mathsf{T}}\beta + \alpha) \leq 0. \tag{3.11}$$

The Lagrange function of (3.11) is

$$\mathcal{L}(\beta, \alpha, \xi, \gamma_i, \delta_i) = \frac{||\beta||^2}{2} + C\sum_i \xi_i - \sum_i \gamma_i[y_i(x_i^{\mathsf{T}}\beta + \alpha) - (1 - \xi_i)] - \sum_i \delta_i\xi_i. \tag{3.12}$$

Taking $\partial\mathcal{L}/\partial\beta = 0$, $\partial\mathcal{L}/\partial\alpha = 0$ and $\partial\mathcal{L}/\partial\xi_i = 0$, we have three equalities,

$$\beta = \sum_i \gamma_i y_i x_i, \tag{3.13}$$

$$0 = \sum_i \gamma_i y_i, \tag{3.14}$$

$$\gamma_i = C - \delta_i,\ \forall i. \tag{3.15}$$

Plugging (3.13-3.15) into (3.12), we have

$$\max_{\gamma_i} \sum_i \gamma_i - \frac{1}{2} \sum_{ij} \gamma_i \gamma_j y_i y_j x_i^\top x_j \quad \text{subject to} \quad \sum_i \gamma_i y_i = 0 \quad \text{and} \quad 0 \leq \gamma_i \leq C. \tag{3.16}$$

It can be shown that under the *Karush-Kuhn-Tucker* conditions, the solution to (3.16) uniquely exists.

Kernel Trick

The goal of SVM is to find a hyperplane (or a linear boundary), which maximizes the margin between classes. What if in certain cases there is no appropriate linear boundaries, and instead a nonlinear structure is preferable? Can we have a nonlinear extension of SVM? The answer is positive and the technique we need is kernel trick. It is essentially to map the original data into the space with vey large dimension (possibly infinite), which is called the *feature space*. Then in this high dimensional space, the linear methods can still be adopted. Due to the richness and complexity in feature space, the linear methods can be interpreted in a "nonlinear" way.

Let's fist consider SVM in the nonlinear case, which we refer as nonlinear SVM. Recall the objective function in the linear case in (3.16). The solution is characterized only by the inner product $x_i^\top x_i$. The kernel trick is to replace the linear similarity measure $x_i^\top x_i$, by a nonlinear similarity measure $\kappa(x_i, x_j)$, the kernel measure. Since $\kappa(x_i, x_j)$ is defined as the inner product in the RKHS, we operate the same estimating procedure except it's implicitly handled in a functional space. The following theorem offers the validity of using kernel trick.

Theorem 3.3 (Mercer Theorem) (Mercer 1909) *Let κ_X be a positive type function κ_X : $\Omega_X \times \Omega_X \to \mathbb{R}$ satisfying*

$$\int \kappa_X^2(x, y) d[\mu(x) \times \mu(y)] < \infty, \tag{3.17}$$

where $\mu(\cdot)$ is a measure on Ω_X. Then for all $x, y \in \Omega_X$, $\kappa_X(x, y)$ has the following decom-

position:

$$\kappa_X(x,y) = \sum_{i=1}^{n} \lambda_i \psi_i(x)\psi(y), \tag{3.18}$$

where the eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ *satisfies* $\sum_{i=1}^{\infty} \lambda_i < \infty$ *and* $\{\psi_i(x)\}_{i=1}^{\infty}$ *are the associated eigenfunctions.*

Mercer Theorem claims that, the RKHS can be built by the closure of all linear combinations of eigenfunctions. First we define a feature map $\Psi(x)$ that projects each point in data space Ω_X into the feature space Ψ_{κ_X},

$$\Psi(x) : x \in \Omega_X \mapsto \{\sqrt{\lambda_i}\psi(x)\}_{i=1}^{\infty}. \tag{3.19}$$

It can be shown that the feature space, defined as

$$\Psi_{\kappa_X} := \overline{\text{Span}\{\Psi(x) : x \in \Omega_X\}}, \tag{3.20}$$

is a RKHS. In addition, the inner product in Ψ_{κ_X}, $\langle \Psi(x), \Psi(y) \rangle$, is identical to $\kappa_X(x,y)$

$$\langle \Psi(x), \Psi(y) \rangle = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(y) = \kappa_X(x,y).$$

It can be seen the above representation, and that in Theorem 3.2 are equivalent; that is, the $\Psi(X)$ and $\kappa_X(\cdot, X)$ are isometrically isomorphic.

The idea of nonlinear SVM is to project the data Ω_X onto a feature space Ψ_{κ_X}, and then conduct a linear SVM in Ψ_{κ_X}. Let $\Psi_i := \Psi(x_i)$, then the constraint in Ψ_{κ_X} can be represented as

$$\begin{aligned} \Psi_i^{\intercal}\beta + \alpha &\geq +\frac{C}{2} - \xi_i \quad \forall\ y_i = +1 \\ \Psi_i^{\intercal}\beta + \alpha &\leq -\frac{C}{2} + \xi_i \quad \forall\ y_i = -1. \end{aligned}$$

These are parallel to (3.8-3.9). The rest procedure can be similarly carried out as in the linear SVM. One can show that nonlinear SVM is to solve the following problem,

$$\max_{\gamma_i} \sum_i \gamma_i - \frac{1}{2} \sum_{ij} \gamma_i \gamma_j y_i y_j \Psi_i^{\mathsf{T}} \Psi_j \text{ subject to } \sum_i \gamma_i y_i = 0 \text{ and } 0 \leq \gamma_i \leq C. \quad (3.21)$$

Substituting the inner product $\Psi_i^{\mathsf{T}} \Psi_j$ with its kernel measure, $\kappa_X(x_i, x_j)$ in (3.16), we have

$$\max_{\gamma_i} \sum_i \gamma_i - \frac{1}{2} \sum_{ij} \gamma_i \gamma_j y_i y_j \kappa_X(x_i, x_j) \text{ subject to } \sum_i \gamma_i y_i = 0 \text{ and } 0 \leq \gamma_i \leq C. \quad (3.22)$$

Given a well-conditioned kernel function, e.g. positive type and square integrable, linear SVM can be generalized via the kernel trick. Actually, we do not try to solve the nonlinear problem directly; that is, we don't know explicit form of the feature mapping Ψ. The inner product defined via the kernel $\kappa_X(x_i, x_j)$ is all we need to proceed the whole analysis.

One of the most common kernel function is *Gaussian Kernel*, which is defined as

$$\kappa(x, y) = \exp(-\sigma ||x - y||^2), \quad (3.23)$$

where $||\cdot||$ measure the norm in $\mathbb{R}^p$ and σ is the width parameter of Gaussian Kernel. In the following we list some other kernel functions; in addition, two examples of Mercer Theorem are provided.

Examples of kernels

- `Linear kernel.` $\kappa(x, y) = x^{\mathsf{T}} y$.
- d^{th} `degree of polynomial kernel.` $\kappa(x, y) = (x^{\mathsf{T}} y + c)^d$.
- `Laplacian kernel.` $\kappa(x, y) = \exp(-\sigma ||x - u||)$.
- `Gaussian kernel.` $\kappa(x, y) = \exp(-\gamma ||x - y||^2)$.
- `Epanechnikov kernel.` $\kappa(x, y) = (1 - ||x - y||^2)_+$.

Examples of Mercer Theorem (Minh et al 2006)

We provides two examples in this section to have better insight of how the kernel function is related to the feature mapping $\Psi(x)$, i.e. what is the explicit forms of $\Psi(x)$ when using certain kernel functions. Noteworthily, in data analysis we do not require any knowledge of $\Psi(x)$.

In the first example, the dimension of original space increases from 2 to 4 due to the feature map, while in the other example the dimension increases from 2 to 3.

- Let $\Omega_X = \mathcal{S}$, a unit sphere on $\mathbb{R}^p$. The measure μ, defined in (3.17), is uniformly distributed on $\mathcal{S}$. If $p = 2$, e.g. $x = (x_1, x_2)$, then a second order of polynomial kernel has an eigenvalue decomposition and the feature map $\Psi(x)$ can be represented explicitly with:

$$\Psi(x) = \left(\sqrt{\frac{2}{3}}, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, \frac{x_1^2 - x_2^2}{\sqrt{2}} \right).$$

- The covariate is binary, e.g. $\Omega_X = \{1, -1\}$ with dimension 2. μ is uniformly distributed on Ω_X. Then with a second order of polynomial kernel there exists a feature map in the following form:

$$\Psi(x) = \left(\sqrt{3}, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2 \right).$$

3.4 Reviews of Nonlinear Dimension Reduction Methods

Triggered by SVM, the application of kernel trick can be extended to the topics of dimension reduction. Classical dimension reduction has two primary branches - unsupervised and supervised methods. Roughly speaking, unsupervised methods seek for linear directions from the covariates X that explains the most variation from X itself, e.g. *Principle Component Analysis* (PCA, Hotelling, 1993), while supervised methods look for directions that explains the variation from response Y, e.g. SIR. When combined with kernel machines, these methods are given the ability to capture the nonlinear features in the data. In this section we

provide a review of some nonlinear dimension reduction methods that successfully assemble the kernel trick and classical dimension reduction methodology.

Kernel Principle Component Analysis PCA searches only the linear combination of variables (or predictors). In real word data can be very complicated and linear structure may not be sufficient to explain the intrinsic structure. Instead, nonlinear models are considered. *Kernel Principle Component Analysis* (KPCA) (Schölkopf et al 1998) is another example of adopting kernel trick. We list a detailed procedure of how KPCA can be proceeded.

Recall in PCA, the following eigenvalue problem is considered,

$$\Sigma_{XX} u = \lambda u,$$

where $\Sigma_{XX} = \mathrm{cov}(X)$. By substitute Σ_{XX} withit sample version, $\widehat{\Sigma}_{XX} = X_n(I_n - \frac{1}{n}1_n 1_n)X_n^\top$,

$$X_n(I_n - \frac{1}{n}1_n 1_n)X_n^\top u = \lambda u. \tag{3.24}$$

This implies u is located in the column space of X_n, i.e. there exists $\alpha \in \mathbb{R}^n$ s.t. $X_n\alpha = u$. In (3.24), replace u by $X\alpha$ and meanwhile multiply $X^\top$ on both the equation,

$$X_n^\top X_n(I_n - \frac{1}{n}1_n 1_n)X_n^\top X_n\alpha = \lambda X_n^\top X_n\alpha. \tag{3.25}$$

Assume the inner product $X_n^\top X_n$ is non-singular, then (3.25) can be reduced to

$$(I_n - \frac{1}{n}1_n 1_n)X_n^\top X_n\alpha = \lambda\alpha. \tag{3.26}$$

This is an alternative way to derive PCA. Notice the solution in (3.26) only depends on $X_n^\top X_n$. By the kernel trick this means the inner product can be calculated by any given kernel functions. Let Ψ_n be the data matrix in feature space. Then (3.26) can be re-written as

$$(I_n - \frac{1}{n}1_n 1_n)\Psi_n^\top \Psi_n\alpha = \lambda\alpha. \tag{3.27}$$

Replacing $\Psi_n^{\mathsf{T}}\Psi_n$ by K, the gram kernel matrix

$$[K]_{ij} = \kappa(x_i, x_j), \quad \forall 1 \leq i, j \leq n, \tag{3.28}$$

then KPCA is to solve the following eigen problem

$$(I_n - \frac{1}{n}1_n 1_n)K\alpha = \lambda\alpha. \tag{3.29}$$

Algorithm of KPCA

1. Calculate the Gram Kernel matrix in (3.28).
2. Centralize K by $(I_n - \frac{1}{n}1_n 1_n)K$.
3. Find the solution of the eigenvalue problem in (3.29). Let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_m > 0$ be all the nonzero eigenvalues in Step 3, and with their associated eigenvectors $(\widehat{\alpha}_1, \widehat{\alpha}_2, \cdots, \widehat{\alpha}_m)$.
4. Project the feature space along the direction $\widehat{u}_i = \Psi_n\widehat{\alpha}_i$:

$$\Psi^{\mathsf{T}}(x)\left(\Psi_n\widehat{\alpha}_i\right) = \kappa(x, X_n)\widehat{\alpha}_i,$$

   where $\kappa(x, X_n) = (\kappa(x, x_1), \kappa(x, x_2), \cdots, \kappa(x, x_n))$ an $n \times 1$ vector.

5. The projection space is estimated by

$$\left(\kappa(x, X_n)\widehat{\alpha}_1, \kappa(x, X_n)\widehat{\alpha}_2, \cdots, \kappa(x, X_n)\widehat{\alpha}_m\right).$$

Kernel Canonical Correlation Analysis The kernelization of CCA, referred as *Kernel CCA*, (Bach and Joran, 2002, Fukumizu et al 2007 and Huang et al, 2009) is another application of kernel trick. Let Ψ_X and Ψ_Y be the feature spaces generated from X and Y. Then KCCA aims at maximizing the correlation between Ψ_X and Ψ_Y. The derivation is similar to KPCA which we omit here.

Algorithm of KCCA:

1. Calculate the gram kernel matrix for X and Y, denoted by K_X and K_Y.

2. Calculate the sample covariance matrices from K_X and K_Y,

$$\begin{aligned}
\widehat{\Sigma}_{XX} &= K_X(I_n - \frac{1}{n}1_n 1_n)K_X \\
\widehat{\Sigma}_{YY} &= K_Y(I_n - \frac{1}{n}1_n 1_n)K_Y \\
\widehat{\Sigma}_{XY} &= K_X(I_n - \frac{1}{n}1_n 1_n)K_Y \\
\widehat{\Sigma}_{YX} &= (\widehat{\Sigma}_{XY})^{\mathsf{T}}.
\end{aligned}$$

3. Solve the generalized eigenvalue problem:

$$\begin{pmatrix} 0 & \widehat{\Sigma}_{XY} \\ \widehat{\Sigma}_{YX} & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} \widehat{\Sigma}_{XX} & 0 \\ 0 & \widehat{\Sigma}_{YY} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

4. Let $\widehat{\rho}_1 \geq \widehat{\rho}_2 \geq \cdots \geq \widehat{\rho}_m > 0$ be all the nonzero eigenvalues in Step 3, and with their associated pairs of eigenvectors $\Big[(\widehat{\alpha}_1, \widehat{\beta}_1), (\widehat{\alpha}_2, \widehat{\beta}_2), \cdots, (\widehat{\alpha}_m, \widehat{\beta}_m)\Big]$, which are the KCCA coefficients.

KCCA is able to detect the nonlinear structures between X and Y. In addition, KCCA can serve as a tool to measure the associations between X and Y. The following theorem provides the theoretical result of KCCA; it claims that if no correlation can be detected in the process of KCCA, two random vectors are independent.

Theorem 3.4 *Let $\psi_x(\cdot)$ and $\psi_y(\cdot)$ be arbitrary elements in Ψ_X and Ψ_Y, respectively. Then the following two properties are equivalent:*

1. *X and Y are independent.*

2. $\sup_{\psi_x, \psi_x} \mathrm{cor}[\psi_x(X), \psi_y(Y)] = 0$.

Kernel Sliced Inverse Regression *Kernel Sliced Inverse regression* (KSIR) (Wu, 2008, Wu, 2008, Yeh et al, 2009, Hsing et al, 2009, Kim and Pavlovic, 2010) is a nonlinear extension of SIR. One of the results of SIR is the inverse regression $E(X|Y)$ is located in $\mathcal{S}_{Y|X}$ - see Theorem 2.1. Nonetheless, this result requires a linearity condition on X. Analogous to the framework of SIR, a linearity condition can be properly defined in RKHS; when this condition holds, KSIR is unbiased.

Definition 3.4 *(Linearity condition is RKHS Ψ_K) Let $\Phi(x) = (\phi_1(x), \phi_2(x), \cdots, \phi_m(x))$ be a collection of elements in RKHS Ψ_K. Then $\Phi(x)$ has the linearity condition if the following property holds, for all $\psi \in \Psi_K$*

$$E[\psi(x)|\Phi(x)] \quad \textit{is linear in } \Phi(x). \tag{3.30}$$

Theorem 3.5 *Let the inverse regression element in Ψ_κ be defined as $E[\kappa(\cdot, x)|y] - E[\kappa(\cdot, x)]$, and the covariance operator of Ψ_κ be Σ_{XX}. Then under the condition in (3.30), we have*

$$E[\kappa(\cdot, x)|y] - E[\kappa(\cdot, x)] \in \Sigma_{XX}\Psi(x) := (\Sigma_{XX}\psi_1(x), \Sigma_{XX}\psi_2(x), \cdots, \Sigma_{XX}\psi_m(x)). \tag{3.31}$$

The implementation of KSIR is similar to SIR. KSIR works on the *kernelized* data while SIR operates on the original data.

Algorithm of KSIR

1. Calculate the gram kernel matrix K.
2. Slice the covariate $K = \begin{bmatrix} k_1^\top \\ \vdots \\ k_n^\top \end{bmatrix}$ into H slices by their associated y.
3. Within each slice, $s = 1, \ldots, H$, calculate the mean, written as $\bar{k}_s$. And also calculate the grand mean, written by $\bar{k}$.
4. Replace each point k_i by its slice mean $\bar{k}_s$ and then build a new data matrix K_H.

5. Compute the between-slice covariance matrix $\Sigma_H = \mathrm{Cov}(K_H)$:

$$\Sigma_H = \frac{1}{n}\sum_{s=1}^{H} n_s(\bar{k}_s - \bar{k})(\bar{k}_s - \bar{k})^{\mathsf{T}},$$

where n_s is the number of data points in the sth slice. And also calculate the sample covariance matrix of K: $\Sigma_K = \sum_{i=1}^{n}(k_i - \bar{k})(k_i - \bar{k})^{\mathsf{T}}/n$.

6. Solve the generalized eigen problem of Σ_H with respect to Σ_K That is, to find the leading m ($m \geq H-1$) eigenvalues λ and corresponding eigenvectors β:

$$\Sigma_H \beta_j = \lambda_j \Sigma_K \beta_j.$$

7. Determine the leading m KSIR directions and collect them as a matrix $V_{n\times m} = [v_1, \ldots, v_m]$. V is orthogonal in terms of Σ_K inner product, i.e., $V^{\mathsf{T}}\Sigma_K V = I_m$. Then map K onto V,

$$V^{\mathsf{T}}K = \begin{bmatrix} v_1^{\mathsf{T}}K \\ \vdots \\ v_m^{\mathsf{T}}K \end{bmatrix},$$

$v_1^{\mathsf{T}}K$ is called the first KSIR variate, $v_2^{\mathsf{T}}K$ the second KSIR variate, and so on.

3.5 Covariance Operator and Conditional Covariance Operator

In the previous section we look into several nonlinear dimension reduction methods that are built on RKHS. Actually these methods can be represented by covariance operators, which can significantly simplify the theoretical development. In this section we provide the definition of covariance operator in Hilbert space (or RKHS), and also study some of its properties. We fist of all introduce the random element in Hilbert space. Suppose $(\Omega, \mathcal{F})$ is a measurable space, i.e. $\mathcal{F}$ is a Boreal σ-field of Ω. Let $\mathcal{H}$ be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, and $\mathcal{H}$ has a Borel σ-field. Then a random element is a mapping from Ω to

$\mathcal{H}$ that is measurable with respect to $\mathcal{F}$. If F is a random element in $\mathcal{H}$, then the mean element of F, denoted by $E(F)$, can be defined via the following equality, for each $h \in \mathcal{H}$,

$$\langle E(F), h\rangle_{\mathcal{H}} = E(\langle F, h\rangle_{\mathcal{H}}). \tag{3.32}$$

Theorem 3.6 (Existence of Mean Element) *Suppose* $E(\langle F, F\rangle_{\mathcal{H}}) < \infty$. *Then* $E(F)$ *uniquely exists.*

PROOF. This is an immediate result from Riesz's representation theorem. □

The above theorem verifies the uniqueness and existence of the mean element in the Hilbert space. In other words, we can exchange the order of integration and inner product, as justified in (3.32). Let us turn our attention to RKHS. Suppose κ is a measurable positive type function, and $\mathcal{H}$ is an RKHS with reproducing kernel κ. Then for a random variable $X : \Omega \to \Omega_X$, $\kappa(\cdot, X)$ is a random element in $\mathcal{H}$. By (3.32), we have an explicit form for the mean element $E(\kappa(\cdot, X))$, for any $u \in \Omega_X$,

$$\begin{aligned} E(\kappa(\cdot, X))(u) = \langle E(\kappa(\cdot, X), \kappa(\cdot, u)\rangle_{\mathcal{H}} &= E(\langle \kappa(\cdot, X), \kappa(\cdot, u)\rangle_{\mathcal{H}}) \\ &= E(\kappa(X, u)). \end{aligned}$$

Let $(X, Y) : \Omega \to \Omega_X \times \Omega_Y$ be a random vector. Suppose $\mathcal{H}_X$ and $\mathcal{H}_Y$ are two RKHSs such that κ_X and κ_Y are their associated reproducing kernels. Then we can define the *covariance operator* of (X, Y), written as Σ_{YX}, an operator from $\mathcal{H}_X$ to $\mathcal{H}_Y$ satisfying that, for each $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$

$$\langle g, \Sigma_{YX} f\rangle_Y = E[\langle f, \kappa_X(\cdot, X) - E(\kappa_X(\cdot, X))\rangle_{\mathcal{H}_X} \langle g, \kappa_Y(\cdot, Y) - E(\kappa_Y(\cdot, Y))\rangle_{\mathcal{H}_Y}], \tag{3.33}$$

where $E(\kappa_X(\cdot, X))$ and $E(\kappa_Y(\cdot, Y))$ are the mean elements of $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively.

Theorem 3.7 (Existence of Covariance Operator) *Suppose there exist* κ_X *and* κ_Y *such*

that

$$E(\kappa_X(X,X)) < \infty, \quad \textit{and} \quad E(\kappa_Y(Y,Y)) < \infty.$$

Then the covariance operator Σ_{YX} defined via (3.33) uniquely exists.

PROOF. By Riesz's representation theorem, it suffices to show that the linear functional $l : g \mapsto \langle g, \Sigma_{YX} f\rangle_Y$ is bounded for any $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$. By definition,

$$\begin{aligned} |\langle g, \Sigma_{YX} f\rangle_Y| &= |E[\langle f, \kappa_X(\cdot, X) - E(\kappa_X(\cdot, X))\rangle_{\mathcal{H}_X} \langle g, \kappa_Y(\cdot, Y) - E(\kappa_Y(\cdot, Y))\rangle_{\mathcal{H}_Y}]| \\ &\leq E|\langle f, \kappa_X(\cdot, X)\rangle_{\mathcal{H}_X} \langle g, \kappa_Y(\cdot, Y)\rangle_{\mathcal{H}_Y}| + E|\langle f, \kappa_X(\cdot, X)\rangle_{\mathcal{H}_X}| \cdot E|\langle g, \kappa_Y(\cdot, Y)\rangle_{\mathcal{H}_Y}| \\ &\leq \left\{ E^{\frac{1}{2}} \kappa_X(X,X) E^{\frac{1}{2}} \kappa_Y(Y,Y) + E[\kappa_X^{\frac{1}{2}}(X,X)] E[\kappa_Y^{\frac{1}{2}}(Y,Y)] \right\} ||f||_{\mathcal{H}_X} ||g||_{\mathcal{H}_Y}. \end{aligned}$$

The last inequality follows from the Hölder's inequality, and the fact that

$$\langle f, \kappa_X(\cdot, X)\rangle_{\mathcal{H}_X} \leq ||\kappa_X(\cdot, X)||_{\mathcal{H}_X} ||f||_{\mathcal{H}_X}, \text{ and } \langle g, \kappa_Y(\cdot, Y)\rangle_{\mathcal{H}_Y} \leq ||\kappa_Y(\cdot, Y)||_{\mathcal{H}_Y} ||g||_{\mathcal{H}_Y}.$$

□

Let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be RKHSs of X and Y, and let Σ_{XY}, Σ_{YX}, Σ_{XX}, and Σ_{YY} be the covariance operators with their corresponding domains. Then the *conditional covariance operator* of Y given X, written as $\Sigma_{YY|X}$, can be defined via the following relation

$$\Sigma_{YY|X} := \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}. \tag{3.34}$$

Note that (3.34) is trivial when X, Y are both Gaussians. Fukumizu et al. (2004, 2009) establish the following result that links the conditional covariance operator, to the dependence structure between Y and X.

Theorem 3.8 *Assume $\mathcal{H}_X + \mathbb{R}$ is dense in $L_2(P_X)$. Then*

$$\langle g, \Sigma_{YY|X} g\rangle_{\mathcal{H}_Y} = E[\mathrm{var}(g(Y)|X)], \tag{3.35}$$

for all $g \in \mathcal{H}_Y$.

The expected value of the conditional variance of $g(Y)$ given X, can be calculated via the conditional covariance operator. This provides an insight that $\Sigma_{YY|X}$ is related to the variation of Y by X. Based on the conditional covariance operator $\Sigma_{YY|X}$, Fukumizu et al. (2004, 2009) introduce an alternative linear SDR method called *Kernel Dimension Reduction* (KDR).

Unlike moments methods that calculate the inverse regression, e.g. $E(X|Y)$ or $E(X^{\mathsf{T}}X)$, KDR considers the forward regression and computes the conditional variance of Y given X. KDR estimates the *conditional covariance operators* on the RKHS. On the other hand, KDR differs from the kernel methods we discuss previously - in the manner that KDR works on the operators, rather than the elements (functions) in the RKHS.

Let B be a matrix in $\mathbb{R}^{p\times d}$ and let $\Sigma_{YY|B^{\mathsf{T}}X}$ be the conditional covariance operator of Y given $B^{\mathsf{T}}X$. The following result is the theoretical foundation of KDR.

Theorem 3.9 *Assume* $\mathcal{H}_X + \mathbb{R}$ *and* $\mathcal{H}_Y + \mathbb{R}$ *are dense in* $L_2(P_X)$ *and* $L_2(P_Y)$*, respectively. Then we have*

$$\Sigma_{YY|X} \leq \Sigma_{YY|B^{\mathsf{T}}X}, \tag{3.36}$$

where the order " $\leq$ *" is defined via inner product in RKHS, which implies for all* $g \in \mathcal{H}_Y$

$$\langle g, \Sigma_{YY|X} g\rangle_{\mathcal{H}_Y} \leq \langle g, \Sigma_{YY|B^{\mathsf{T}}X} g\rangle_{\mathcal{H}_Y}.$$

Furthermore, in (3.36) the equality holds if and only if B *is exhaustive; i.e.*

$$\Sigma_{YY|X} = \Sigma_{YY|B^{\mathsf{T}}X} \quad \textit{if and only if} \quad Y|X \overset{\mathcal{D}}{=} Y|B^{\mathsf{T}}X. \tag{3.37}$$

The above suggests that, we can use the solution $B^\star$ to the following optimization to estimate the central subspace $\mathcal{S}_{Y|X}$,

$$B^\star = \text{argmin}_B \text{tr}(\Sigma_{YY|B^{\mathsf{T}}X}). \tag{3.38}$$

However, (3.38) is a non-convex optimization problem and may require complicated optimization procedure.

3.6 Extended Covariance Operator and Residual Class

Our primary goal is to make inference on the central class $\mathfrak{S}_{Y|X}$. On the other hand, we have demonstrated that the residual class

$$L_2(P_X) \ominus L_2(P_Y)$$

plays a critical role in the estimation of $\mathfrak{S}_{Y|X}$. Its orthogonal complement, i.e. the regression class fully estimates the central class $\mathfrak{S}_{Y|X}$ under completeness, and even without such a condition it is guaranteed to be within $\mathfrak{S}_{Y|X}$. It turns out this sub-class can be explicitly expressed as the kernel of a covariance operator. However, to derive it may not be obvious because the domain of a covariance operator is on an RKHS, not L_2.

In this section, we develop an extended operator that has domain on $L_2(P_X)$ and therefore eliminates the gap between RKHS and $L_2(P_X)$. We show that this extended operator has some nice properties such as isomorphism and invertibility, which are important to our theoretical development. In addition, when the extended operator is considered, the estimation procedures can rely on simple spectral decompositions, rather than complicated numerical optimizations.

Since constants are irrelevant here (for example, f and $f+3$ can be considered as the same function), we introduce notation that makes this explicit. If A and B are two sets, then we write $A \overset{\circ}{\subseteq} B$ if for each $f \in A$ there is a unique $c \in \mathbb{R}$ such that $f + c \in B$. We say that A is a dense$^\circ$ subset of B if, $A \overset{\circ}{\subseteq} B$ and, 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:

Assumption 3.1 *$\mathcal{H}_X$ and $\mathcal{H}_Y$ are dense$^\circ$ subsets of $L_2(P_X)$ and $L_2(P_Y)$, respectively.*

Assumption 3.2 *there are* $C_1 > 0$, $C_2 > 0$ *such that* $\mathrm{var}[f(X)] \leq C_1 \|f\|_{\mathcal{H}_X}$, $\mathrm{var}[g(Y)] \leq C_2 \|g\|_{\mathcal{H}_Y}$.

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

Under Assumption 3.2 the symmetric bilinear form $u : \mathcal{H}_X \times \mathcal{H}_X \to \mathbb{R}$ defined by $u(f, g) = \mathrm{cov}[f(X), g(X)]$ is bounded. Hence it induces a bounded and self-adjoint linear operator $M_{XX} : \mathcal{H}_X \to \mathcal{H}_X$ satisfying $\langle f, M_{XX}\, g\rangle_{\mathcal{H}_X} = u(f, g)$. In the following, we denote the range of a linear operator A by $\mathrm{ran}\, A$, the kernel of A by $\mathrm{Ker} A$, and the closure of $\mathrm{ran}\, A$ by $\overline{\mathrm{ran}}\, A$. Let Q_X be the projection on to $\overline{\mathrm{ran}}\, M_{XX}$.

Lemma 3.1 *For any constant* $c \in \mathcal{H}_X$, $Q_X c = 0$ *in* $\mathcal{H}_X$.

Proof. For any $g \in \overline{\mathrm{ran}}\, M_{XX}$, there is a sequence $\{g_n\}$ such that $g_n \to g$ in $\mathcal{H}_X$ and $g_n \in \mathrm{ran}\,(\Sigma_{XX})$. Then $g_n = M_{XX}\, h_n$ for some $h_n \in \mathcal{H}_X$. It follows that

$$\langle c - 0, g\rangle_{\mathcal{H}_X} = \lim_{n\to\infty} \langle c, g_n\rangle_{\mathcal{H}_X} = \lim_{n\to\infty} \langle c, M_{XX}\, h_n\rangle_{\mathcal{H}_X} = \lim_{n\to\infty} \mathrm{cov}(c, h_n(X)) = 0.$$

By definition, 0 is the projection $Q_X c$. □

Here, it is worth mentioning that in some cases the only constant in $\mathcal{H}_X$ is 0, in which case the lemma holds trivially. For an example of such cases see Steinwart, Hush, and Scovel (2004), and Fukumizu, Back, and Gretton (2007).

Lemma 3.2 *For each* $f \in \mathcal{H}_X$ *there is one and only one* $g \in \overline{\mathrm{ran}}\, M_{XX}$ *such that* g *differs from* f *by a constant.*

Proof. Since M_{XX} is self adjoint, we have $(\mathrm{ran}\, M_{XX})^{\perp} = (\overline{\mathrm{ran}}\, M_{XX})^{\perp} = \mathrm{Ker} M_{XX}$, which implies $M_{XX}\,(I - Q_X) f = M_{XX}\,(f - Q_X f) = 0$. Hence

$$\mathrm{var}(f(X) - Q_X f(X)) = \langle f - Q_X f, M_{XX}\,(f - Q_X f)\rangle_{\mathcal{H}_X} = 0.$$

That is, $f = Q_X f + c$ for some constant c. Since $Q_X f \in \overline{\text{ran}}\, M_{XX}$, existence is proved. Now suppose h is another member of $\overline{\text{ran}}\, M_{XX}$ that differs from f by a constant. Then, $h = Q_X h = Q_X f + Q_X c = Q_X f$, where the last equality follows from Lemma 3.1. □

Since we are not concerned with any constant, we need only consider functions inside $\overline{\text{ran}}\, M_{XX}$, and define covariance operators on $\overline{\text{ran}}\, M_{XX}$. In the following, when dealing with $\overline{\text{ran}}\, M_{XX}$ and $\overline{\text{ran}}\, M_{YY}$ we denote them by $\mathcal{L}_X$ and $\mathcal{L}_Y$, respectively.

Definition 3.5 *Suppose Assumption 3.1 and Assumption 3.2 are satisfied. We define the covariance operators $\Sigma_{XX} : \mathcal{L}_X \to \mathcal{L}_X$, $\Sigma_{YY} : \mathcal{L}_Y \to \mathcal{L}_Y$, and $\Sigma_{YX} : \mathcal{L}_X \to \mathcal{L}_Y$ through the relations:*

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

These operators are essentially the same as those introduced by Fukumizu, Back, and Jordan (2004, 2009), except that here we do not assume $\mathcal{H}_X$ and $\mathcal{H}_Y$ to be reproducing kernel Hilbert spaces.

By Baker (1972), Theorem 1, there is a unique operator $R_{YX} \in \mathcal{B}(\mathcal{L}_X, \mathcal{L}_Y)$ such that $\Sigma_{YX} = \Sigma_{YY}^{\frac{1}{2}} R_{YX} \Sigma_{XX}^{\frac{1}{2}}$. We call R_{XY} the *correlation operator*.

In order to connect the central class to covariance operators, we need to extend the operators $\Sigma_{XX}^{\frac{1}{2}} : \mathcal{L}_X \to \mathcal{L}_X$ and $R_{XY}\Sigma_{YY}^{\frac{1}{2}} : \mathcal{L}_Y \to \mathcal{L}_X$ to operators whose domains are $L_2(P_X)$ and $L_2(P_Y)$. Let $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ denote the collection of all bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$, and abbreviate $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_1)$ by $\mathcal{B}(\mathcal{H}_1)$.

Lemma 3.3 *Let $\mathcal{H}_1$, $\mathcal{H}_2$, and $\mathcal{H}_3$ be Hilbert spaces, $A \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_3)$, and $T \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$. Suppose*

1. *$\text{ran}\, T$ is dense in $\mathcal{H}_2$;*

2. *there is a constant $c > 0$ such that, for all $f \in \mathcal{H}_1$, $\|Af\|_{\mathcal{H}_3} \leq c\,\|Tf\|_{\mathcal{H}_2}$.*

Then there is a unique $\tilde{A} \in \mathcal{B}(\mathcal{H}_2, \mathcal{H}_3)$ such that $\tilde{A}T = A$ and $\|\tilde{A}\| \leq c$.

PROOF. Let $f \in \mathcal{H}_2$. Since $\operatorname{ran} T$ is dense in $\mathcal{H}_2$ there is a sequence $\{f_n\} \subseteq \mathcal{H}_1$ such that $Tf_n \to f$ in $\mathcal{H}_2$. Thus $\{Tf_n\}$ is Cauchy in $\mathcal{H}_2$. Because $\|A(f_n - f_m)\|_{\mathcal{H}_3} \le c\|T(f_n - f_m)\|_{\mathcal{H}_2}$, $\{Af_n\}$ is Cauchy in $\mathcal{H}_3$. Let $\tilde{f}$ be the $\mathcal{H}_3$-limit of Af_n. Let us show that $\tilde{f}$ is uniquely defined (that is, independently of the sequence $\{f_n\}$). Suppose $\{f'_n\} \subseteq \mathcal{H}_1$ is another sequence such that $Tf'_n \to f$ in $\mathcal{H}_2$, and let $\hat{f}$ be the $\mathcal{H}_3$-limit of Af'_n. Then

$$\|\tilde{f} - \hat{f}\|_{\mathcal{H}_3} \le \|\tilde{f} - Af_n\|_{\mathcal{H}_3} + \|\hat{f} - Af'_n\|_{\mathcal{H}_3} + \|A(f'_n - f_n)\|_{\mathcal{H}_3}.$$

By definition, the first two terms on the right tend to 0 as $n \to \infty$. The last term is no more than $c\|T(f'_n - f_n)\|_{\mathcal{H}_2}$, which tends to 0 as $n \to \infty$. Thus $\|\tilde{f} - \hat{f}\|_{\mathcal{H}_3} = 0$.

Since $\tilde{f}$ is independent of the sequence $\{f_n\}$, $f \mapsto \tilde{f}$ defines a mapping $\tilde{A} : \mathcal{H}_2 \to \mathcal{H}_3$. It is easy to see that $\tilde{A}$ is a linear operator. Furthermore, if $\{f_n\}$ is a sequence in $\mathcal{H}_1$ such that $Tf_n \to f$ in $\mathcal{H}_2$ and $Af_n \to \tilde{A}f$ in $\mathcal{H}_3$, then

$$\|\tilde{A}f\|_{\mathcal{H}_3} = \lim_{n\to\infty} \|Af_n\|_{\mathcal{H}_3} \le c \lim_{n\to\infty} \|Tf_n\|_{\mathcal{H}_2} = c\|f\|_{\mathcal{H}_2}.$$

Thus $\tilde{A} \in \mathcal{B}(\mathcal{H}_2, \mathcal{H}_3)$. For any $f \in \mathcal{H}_1$, we have $Tf \to Tf$ in $\mathcal{H}_2$ and $Af \to Af$ in $\mathcal{H}_3$. Hence $\tilde{A}Tf = Af$. That is, $\tilde{A}T = A$.

To show that $\tilde{A}$ is unique, suppose there is another $\hat{A} \in \mathcal{B}(\mathcal{H}_2, \mathcal{H}_3)$ such that $\hat{A}T = A$. Let $f \in \mathcal{H}_2$ and $\{f_n\}$ be a sequence in $\mathcal{H}_1$ such that $Tf_n \to f$ in $\mathcal{H}_2$. Then

$$\|(\tilde{A} - \hat{A})f\|_{\mathcal{H}_3} = \|(\tilde{A}T - \hat{A}T)f_n\|_{\mathcal{H}_3} + \|(\tilde{A} - \hat{A})(f - Tf_n)\|_{\mathcal{H}_3},$$

where the first term on the right $\|Af_n - Af_n\|_{\mathcal{H}_3} = 0$, and the second is no more than $(\|\tilde{A}\| + \|\hat{A}\|)\|f - Tf_n\|_{\mathcal{H}_2}$, which tends to 0 as $n \to \infty$. Hence $\tilde{A} = \hat{A}$. □

A similar extension can be found in Weidmann (1980, Theorem 4.5) which, however, assumes $\mathcal{H}_1$ and $\mathcal{H}_2$ to have the same inner product. For our purpose it is essential to make a distinction between the inner products for $\mathcal{H}_1$ and $\mathcal{H}_2$.

We will say that $A \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_3)$ is extendable to $\mathcal{B}(\mathcal{H}_2, \mathcal{H}_3)$ if there is a $T \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$

and $\tilde{A} \in \mathcal{B}(\mathcal{H}_2, \mathcal{H}_3)$ such that $\operatorname{ran} T$ is dense in $\mathcal{H}_2$ and $\tilde{A}Tf = Af$, and call $\tilde{A}$ the extension of A via T. Since all extensions in this paper are relative to the centering transformation $T(f) \mapsto f - Ef(X)$, we will omit "via T" when we mention an extension.

Theorem 3.10 *The operator $\Sigma_{XX}^{\frac{1}{2}} \in \mathcal{B}(\mathcal{L}_X, \mathcal{L}_X)$ is extendable to $\mathcal{B}(L_2(P_X), \mathcal{L}_X)$. Moreover, if $\tilde{\Sigma}_{XX}^{\frac{1}{2}}$ is the extension of $\Sigma_{XX}^{\frac{1}{2}}$, then $\operatorname{ran} \tilde{\Sigma}_{XX}^{\frac{1}{2}} = \overline{\operatorname{ran}}\, \Sigma_{XX}^{\frac{1}{2}}$. The same can be said of $\Sigma_{YY}^{\frac{1}{2}}$.*

Proof. Let $T : \mathcal{L}_X \to L_2(P_X)$ be the centering transformation. Let $f \in L_2(P_X)$. By Assumption 3.1, the family $\{f - Ef(X) : f \in \mathcal{H}_X\}$ is dense in $L_2(P_X)$. Let $\{f_n\} \subseteq \mathcal{H}_X$ be such that $f_n - Ef_n(X) \to f$ in $L_2(P_X)$. By Lemma 3.2, for each f_n, there is a unique member $g_n \in \mathcal{L}_X$ such that $T(g_n) = f_n - Ef_n(X)$. So $Tg_n \to f$. Hence condition 1 in Lemma 3.3 is satisfied for $\mathcal{H}_1 = \mathcal{L}_X$ and $\mathcal{H}_2 = L_2(P_X)$. In the meantime, for any $f \in \mathcal{L}_X$,

$$\|\Sigma_{XX}^{\frac{1}{2}} f\|_{\mathcal{L}_X}^2 = \langle \Sigma_{XX}^{\frac{1}{2}} f, \Sigma_{XX}^{\frac{1}{2}} f \rangle_{\mathcal{L}_X} = \operatorname{var}(f(X)) = \|Tf\|_{L_2(P_X)}^2.$$

Hence condition 2 in Lemma 3.3 is satisfied with $\mathcal{H}_2 = L_2(P_X)$ and $\mathcal{H}_3 = \mathcal{L}_X$. This proves the extendability of $\Sigma_{XX}^{\frac{1}{2}}$.

Let $\tilde{f} \in \operatorname{ran} \tilde{\Sigma}_{XX}^{\frac{1}{2}}$. Then $\tilde{f} = \tilde{\Sigma}_{XX}^{\frac{1}{2}} f$ for some $f \in L_2(P_X)$. Hence there is a sequence $\{f_n\} \subseteq \mathcal{L}_X$ such that $\Sigma_{XX}^{\frac{1}{2}} f_n \to \tilde{f}$ in $\mathcal{L}_X$. That is, there is a sequence $\{g_n\} \subseteq \operatorname{ran} \Sigma_{XX}^{\frac{1}{2}}$ such that $g_n \to \tilde{f}$ in $\mathcal{L}_X$. Hence $\operatorname{ran} \tilde{\Sigma}_{XX}^{\frac{1}{2}} \subseteq \overline{\operatorname{ran}}\, \Sigma_{XX}^{\frac{1}{2}}$. Let $\tilde{f} \in \overline{\operatorname{ran}}\, \Sigma_{XX}^{\frac{1}{2}}$. Then there is a sequence $\{f_n\} \subseteq \mathcal{L}_X$ such that $\Sigma_{XX}^{\frac{1}{2}} f_n \to \tilde{f}$. Since

$$\|\Sigma_{XX}^{\frac{1}{2}} f_n - \Sigma_{XX}^{\frac{1}{2}} f_m\|_{\mathcal{L}_X}^2 = \langle f_n - f_m, \Sigma_{XX}(f_n - f_m) \rangle_{\mathcal{L}_X} = \operatorname{var}(f_n - f_m) = \|Tf_n - Tf_m\|_{L_2(P_X)}^2,$$

$\{Tf_n\}$ is a Cauchy sequence in $L_2(P_X)$. Let f be the $L_2(P_X)$-limit $\{Tf_n\}$. Then $\tilde{\Sigma}_{XX}^{\frac{1}{2}} f = \tilde{f}$. Thus $\tilde{f} \in \operatorname{ran} \tilde{\Sigma}_{XX}^{\frac{1}{2}}$. □

The next lemma shows us how to construct an extension of the composition of two bounded operators.

Lemma 3.4 *Let $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$, $\mathcal{H}_4$ be Hilbert spaces, $A \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_3)$, $B \in \mathcal{B}(\mathcal{H}_3, \mathcal{H}_4)$. If A is extendable to $\mathcal{B}(\mathcal{H}_2, \mathcal{H}_3)$ then BA is extendable to $\mathcal{B}(\mathcal{H}_2, \mathcal{H}_4)$, and $\widetilde{BA} = B\tilde{A}$.*

PROOF. Let $\tilde{A}$ be the extension of A to $\mathcal{B}(\mathcal{H}_2, \mathcal{H}_3)$. Since A is extendable to $\mathcal{B}(\mathcal{H}_2, \mathcal{H}_3)$, there is a $T \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ such that $\operatorname{ran} T$ is dense in $\mathcal{H}_2$ and $\tilde{A}Tf = Af$ for all $f \in \mathcal{H}_1$. Since $BA \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_4)$ and, for any $f \in \mathcal{H}_1$,

$$\|BAf\|_{\mathcal{H}_4} = \|B\tilde{A}Tf\|_{\mathcal{H}_4} \leq \|B\tilde{A}\| \, \|Tf\|_{\mathcal{H}_2},$$

by Lemma 3.3, BA is extendable to $\mathcal{B}(\mathcal{H}_2, \mathcal{H}_4)$. □

We now consider extendability of the various operators in reproducing kernel Hilbert spaces, which will be used in subsequent developments.

Theorem 3.11 *Suppose Assumption 3.1 and Assumption 3.2 are satisfied. Then*

1. Σ_{XX} *is extendable to* $\mathcal{B}(L_2(P_X), \mathcal{L}_X)$, *and* $\tilde{\Sigma}_{XX} = \tilde{\Sigma}_{XX}^{\frac{1}{2}} \tilde{\Sigma}_{XX}^{\frac{1}{2}}$.
2. Σ_{YY} *is extendable to* $\mathcal{B}(L_2(P_Y), \mathcal{L}_Y)$, *and* $\tilde{\Sigma}_{YY} = \tilde{\Sigma}_{YY}^{\frac{1}{2}} \tilde{\Sigma}_{YY}^{\frac{1}{2}}$.
3. Σ_{YX} *is extendable to* $\mathcal{B}(L_2(P_Y), \mathcal{L}_X)$, *and* $\tilde{\Sigma}_{YX} = \Sigma_{YY}^{\frac{1}{2}} R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}}$.

PROOF. 1 & 2: for any $f \in \mathcal{H}_X$, $||\Sigma_{XX}^{\frac{1}{2}} f||_{\mathcal{H}_X} = ||f||_{L_2(P_X)}$; hence, $\Sigma_{XX}^{\frac{1}{2}}$ is extendable to $\mathcal{B}(L_2(P_X), \mathcal{L}_X)$. Since $\Sigma_{XX}^{\frac{1}{2}} \in \mathcal{B}(\mathcal{L}_X, \mathcal{L}_X)$, by Lemma 3.4, $\Sigma_{XX} = \Sigma_{XX}^{\frac{1}{2}} \Sigma_{XX}^{\frac{1}{2}}$ is also extendable to $\mathcal{B}(L_2(P_X), \mathcal{L}_X)$ and

$$\tilde{\Sigma}_{XX} = \Sigma_{XX}^{\frac{1}{2}} \tilde{\Sigma}_{XX}^{\frac{1}{2}} = \tilde{\Sigma}_{XX}^{\frac{1}{2}} \tilde{\Sigma}_{XX}^{\frac{1}{2}}.$$

Similarly we can prove part 2.

3: This follows from Lemma 3.4, and the fact that R_{XY} is bounded. □

Theorem 3.12 *Under conditions Assumption 3.1 and Assumption 3.2,* $\tilde{\Sigma}_{XX}^{\frac{1}{2}}$ *and* $\tilde{\Sigma}_{YY}^{\frac{1}{2}}$ *are isomorphisms.*

PROOF. Isomorphism of $\tilde{\Sigma}_{XX}^{\frac{1}{2}}$ means

$$\langle \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \tilde{\Sigma}_{XX}^{\frac{1}{2}} g \rangle_{\mathcal{L}_X} = \langle f, g \rangle_{L_2(P_X)}, \quad f, g \in L_2(P_X). \tag{3.39}$$

First, assume $f \in \operatorname{ran} T$, $g \in L_2(P_X)$. In this case there is $h \in \mathcal{L}_X$ such that $\tilde{\Sigma}_{XX}^{\frac{1}{2}} f = \tilde{\Sigma}_{XX}^{\frac{1}{2}} T h = \Sigma_{XX}^{\frac{1}{2}} h$. Let $\{g_n\}$ be a sequence in $\mathcal{L}_X$ such that $Tg_n \to g$ in $L_2(P_X)$ and $\Sigma_{XX}^{\frac{1}{2}} g_n \to \tilde{\Sigma}_{XX}^{\frac{1}{2}} g$ in $\mathcal{L}_X$. Because $Tg_n \to g$ in $L_2(P_X)$,

$$\langle \Sigma_{XX}^{\frac{1}{2}} T h, \Sigma_{XX}^{\frac{1}{2}} g_n \rangle_{\mathcal{L}_X} = \langle Th, \Sigma_{XX} T g_n \rangle_{\mathcal{L}_X} = \langle Th, Tg_n \rangle_{L_2(P_X)} \to \langle f, g \rangle_{L_2(P_X)}.$$

Because $\Sigma_{XX}^{\frac{1}{2}} g_n \to \tilde{\Sigma}_{XX}^{\frac{1}{2}} g$ in $\mathcal{L}_X$,

$$\langle \Sigma_{XX}^{\frac{1}{2}} f, \Sigma_{XX}^{\frac{1}{2}} g_n \rangle_{\mathcal{L}_X} \to \langle \Sigma_{XX}^{\frac{1}{2}} f, \tilde{\Sigma}_{XX}^{\frac{1}{2}} g \rangle_{\mathcal{L}_X}.$$

Therefore (3.39) holds. Now assume $f, g \in L_2(P_X)$. Let $\{f_n\}$ be a sequence in $\mathcal{L}_X$ such that $Tf_n \to f$ in $L_2(P_X)$ and $\Sigma_{XX}^{\frac{1}{2}} f_n \to \tilde{\Sigma}_{XX}^{\frac{1}{2}} f$ in $\mathcal{L}_X$. Then, by the previous case,

$$\langle \tilde{\Sigma}_{XX}^{\frac{1}{2}} T f_n, \tilde{\Sigma}_{XX}^{\frac{1}{2}} g \rangle_{\mathcal{L}_X} = \langle T f_n, g \rangle_{L_2(P_X)} \to \langle f, g \rangle_{L_2(P_X)}.$$

In the meantime, since $\Sigma_{XX}^{\frac{1}{2}} f_n \to \tilde{\Sigma}_{XX}^{\frac{1}{2}} f$ in $\mathcal{L}_X$, we have

$$\langle \Sigma_{XX}^{\frac{1}{2}} f_n, \tilde{\Sigma}_{XX}^{\frac{1}{2}} g \rangle_{\mathcal{L}_X} \to \langle \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \tilde{\Sigma}_{XX}^{\frac{1}{2}} g \rangle_{\mathcal{L}_X}.$$

Hence (3.39) holds. The result for $\tilde{\Sigma}_{YY}^{\frac{1}{2}}$ is proved similarly. □

Theorem 3.12 implies that $(\tilde{\Sigma}_{XX}^{\frac{1}{2}})^{-1}$ is a bounded operator (in fact, $\|(\tilde{\Sigma}_{XX}^{\frac{1}{2}})^{-1}\| = 1$). This property is of critical importance for asymptotic analysis. This is because, unless $\tilde{\Sigma}_{XX}^{-1/2}$ bounded (and hence continuous), it cannot be adequately approximated, no matter how large the sample size is. We note that the unextended operator $\Sigma_{XX}^{\frac{1}{2}}$ typically does not have a bounded inverse. For example, if $\mathcal{H}_X$ is a reproducing kernel Hilbert space then Σ_{XX} is compact. So, unless $\operatorname{ran} \Sigma_{XX}$ has finite dimension, the nonzero eigenvalues of Σ_{XX} goes to 0, and hence $(\Sigma_{XX}^{\frac{1}{2}})^{-1}$ cannot be bounded.

Neither does the invertibility of $\tilde{\Sigma}_{XX}^{\frac{1}{2}}$ imply the invertibility of $\tilde{\Sigma}_{XX}$. The reason is that $\langle f, \tilde{\Sigma}_{XX} f \rangle_{L_2(P_X)}$ is not a well defined object, because f is a member of $L_2(P_X)$ but $\tilde{\Sigma}_{XX} f$ is a member of $\mathcal{L}_X$, and they cannot be placed in the same inner product. So there is no parallel

argument to Theorem 3.12. Fortunately, as we will see $\tilde{\Sigma}_{XX}^{-1}$ is not an interesting object in our development. In summary, $\Sigma_{XX}^{\frac{1}{2}}$ and Σ_{XX} are both invertible but their inverses need not be bounded; $\tilde{\Sigma}_{XX}^{\frac{1}{2}}$ is invertible and its inverse is bounded; $\tilde{\Sigma}_{XX}$ need not be invertible. In the following we will write $(\tilde{\Sigma}_{XX}^{\frac{1}{2}})^{-1}$ as $\tilde{\Sigma}_{XX}^{-1/2}$.

The results of the last section allow us to characterize $L_2(P_X) \ominus L_2(P_Y)$ in terms of extended covariance operators, which is the key to deriving approaches to its estimation.

Theorem 3.13 *If Assumption 3.1 and Assumption 3.2 hold, then*

$$\mathrm{Ker}(\tilde{\Sigma}_{YX}) = L_2(P_X) \ominus L_2(P_Y).$$

PROOF. We claim that, for any $f \in L_2(P_X),\ g \in L_2(P_Y)$,

$$\mathrm{cov}[f(X), g(Y)] = \langle R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \tilde{\Sigma}_{YY}^{\frac{1}{2}} g \rangle_{\mathcal{L}_Y}. \tag{3.40}$$

We first prove this for $f \in L_2(P_X)$ and $g \in \mathrm{ran}\, T$. Then, $g = Th$ for some $h \in \mathcal{L}_Y$, and $\tilde{\Sigma}_{XX}^{\frac{1}{2}} Th = \Sigma_{XX}^{\frac{1}{2}} h$. Let $\{f_n\}$ be a sequence in $\mathcal{L}_X$ such that $Tf_n \to f$ in $L_2(P_X)$ and $\Sigma_{YX} f_n \to \tilde{\Sigma}_{YX} f$ in $\mathcal{L}_Y$. Then

$$\mathrm{cov}[Tf_n(X), g(Y)] \to \mathrm{cov}[f(X), g(Y)], \quad \langle \Sigma_{YX} f_n, h \rangle_{\mathcal{L}_Y} \to \langle \tilde{\Sigma}_{YX} f, h \rangle_{\mathcal{L}_Y}.$$

Since $f_n \in \mathcal{L}_X$ and $h \in \mathcal{L}_Y$, we have, by the definition of Σ_{YX},

$$\mathrm{cov}[Tf_n(X), g(Y)] = \mathrm{cov}[f_n(X), g(Y)] = \mathrm{cov}[f_n(X), h(X)] = \langle \Sigma_{YX} f_n, h \rangle_{\mathcal{L}_Y}.$$

Hence $\langle \tilde{\Sigma}_{YX} f, h \rangle_{\mathcal{L}_Y} = \mathrm{cov}[f(X), g(Y)]$. But since $\tilde{\Sigma}_{YX} = \Sigma_{YY}^{\frac{1}{2}} R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}}$ and $\Sigma_{YY}^{\frac{1}{2}}$ are self adjoint, we have

$$\langle \Sigma_{YX} f, h \rangle_{\mathcal{L}_Y} = \langle R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \Sigma_{YY}^{\frac{1}{2}} h \rangle_{\mathcal{L}_Y} = \langle R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \tilde{\Sigma}_{YY}^{\frac{1}{2}} Th \rangle_{\mathcal{L}_Y} = \langle R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \tilde{\Sigma}_{YY}^{\frac{1}{2}} g \rangle_{\mathcal{L}_Y}.$$

Thus (3.40) holds. Now let $f \in L_2(P_X)$ and $g \in L_2(P_Y)$, and let g_n be a sequence in $\mathcal{L}_Y$

such that $Tg_n \to g$ in $L_2(P_Y)$ and $\Sigma_{YX} g_n \to \tilde{\Sigma}_{YX} g$ in $\mathcal{L}_Y$. Then

$$\operatorname{cov}[f(X), g(Y)] = \lim_{n\to\infty} \operatorname{cov}[f(X), Tg_n(Y)].$$

But since $Tg_n \in \operatorname{ran} T$, we have

$$\begin{aligned}\operatorname{cov}[f(X), Tg_n(X)] = \langle R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \tilde{\Sigma}_{YY}^{\frac{1}{2}} Tg_n \rangle_{\mathcal{L}_X} &= \langle R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \Sigma_{YY}^{\frac{1}{2}} g_n \rangle_{\mathcal{L}_X} \\ &\to \langle R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \tilde{\Sigma}_{YY}^{\frac{1}{2}} g \rangle_{\mathcal{L}_X}.\end{aligned}$$

This proves (3.40).

Now if $f \in L_2(P_X) \ominus L_2(P_Y)$ then, by (3.40),

$$\langle R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \tilde{\Sigma}_{YY}^{\frac{1}{2}} g \rangle_{\mathcal{H}_Y} = 0 \tag{3.41}$$

for all $g \in L_2(P_Y)$. Take $g = T\Sigma_{YY}^{\frac{1}{2}} R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f$. Then

$$\begin{aligned}\langle R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \tilde{\Sigma}_{YY}^{\frac{1}{2}} T \Sigma_{YY}^{\frac{1}{2}} R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f \rangle_{\mathcal{L}_X} = \quad & \langle R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \Sigma_{YY}^{\frac{1}{2}} \Sigma_{YY}^{\frac{1}{2}} R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f \rangle_{\mathcal{L}_X} \\ = \langle \Sigma_{YY}^{\frac{1}{2}} R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f, \Sigma_{YY}^{\frac{1}{2}} R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f \rangle_{\mathcal{L}_X} &= \| \Sigma_{YY}^{\frac{1}{2}} R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f \|_{\mathcal{L}_X} = 0,\end{aligned}$$

which implies $f \in \operatorname{Ker}(\tilde{\Sigma}_{YX})$. Conversely, if $f \in \operatorname{Ker}(\tilde{\Sigma}_{YX})$, then

$$R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}} f = \tilde{\Sigma}_{YY}^{-1/2} (\tilde{\Sigma}_{YY}^{\frac{1}{2}} R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}}) f = \tilde{\Sigma}_{YY}^{-1/2} \tilde{\Sigma}_{YX} f = 0.$$

Hence (3.41) holds, which implies $\operatorname{cov}[f(X), g(Y)] = 0$, and hence $f \in L_2(P_X) \ominus L_2(P_Y)$. $\square$

Chapter 4

Generalized Sliced Inverse Regression (GSIR) - First Order Method

From the development in previous chapter, we are now ready to describe the population-level set up required to estimate the regression class $\mathfrak{C}_{Y|X}$, which under completeness is exhaustive for the central class $\mathfrak{S}_{Y|X}$. Fukumizu, Bach and Jordan (2004) introduce a *conditional mean operator*, and it characterizes the conditional expectation of $g(Y)$ given X via

$$\Sigma_{XX} E[g(Y)|X] = \Sigma_{XY} g,$$

for any $g \in \mathcal{H}_Y$ where $\mathcal{H}_Y$ is the RKHS of Y. We consider the conditional expectation in a reverse manner; given any $f \in \mathcal{H}_X$ where $\mathcal{H}_X$ is the RKHS of X, we compute the inverse regression $E(f(X)|Y)$ via *inverse conditional mean operator*, and show that this operator can be explicitly represented as products of extended operators. Based on inverse conditional mean operator, we develop a novel estimator of $\mathfrak{C}_{Y|X}$ called *Generalized Sliced Inverse Regression* (GSIR); furthermore, we show that GSIR is unbiased. This part of work is constructed in parallel with the classical SIR and again, this demonstrates the simplicity

of notions under our formulation, and ability to adopt classical concepts.

Secondly, we revisit the kernel SIR introduced by Wu (2008) and Yeh, Huang, and Lee (2009), and establish its unbiasedness in our formulation, whose result is much more general than the form of unbiasedness developed in Yeh, Huang, and Lee (2009). Last, we introduce a new notation system for the coordinates of elements in RKHS. Given this coordinate system and finite sample, we derive the matrix representations of operators in RKHS. Since the estimation of GSIR depends on the operators, we then propose a sample algorithm for GSIR.

4.1 Inverse Conditional Mean Operator

Recall that classical SIR (Li, 1991) for linear SDR is based on the matrix

$$[\mathrm{var}(X)]^{-1}\mathrm{var}[E(X|Y)]. \tag{4.1}$$

The reason for using this matrix is that, under Assumption 2.1, the rescaled "inverse" conditional mean $[\mathrm{var}(X)]^{-1}E(X|Y)$ is contained in this space. To generalize this to the nonlinear setting, we first generalize the inverse conditional mean $E(X|Y)$ to an inverse conditional mean operator.

Definition 4.1 *We call the operator* $\tilde{\Sigma}_{YY}^{-1/2}R_{YX}\tilde{\Sigma}_{XX}^{\frac{1}{2}} : L_2(P_X) \to L_2(P_Y)$ *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 theorem.

Theorem 4.1 *Under Assumption 3.1 and Assumption 3.2, 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 $g \in L_2(P_Y)$,

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

where the last equality follows from (3.40). Hence $\mathrm{cov}(f(X) - (E_{X|Y} f)(Y), g(Y)) = 0$. 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 $\tilde{\Sigma}_{XX}^{\frac{1}{2}}$ are isomorphisms, and $R_{YX}^* = R_{XY}$. □

Corollary 4.1 *Under Assumption 3.1 and Assumption 3.2, for any $f, g \in L_2(P_X)$,*

$$\langle g, E_{X|Y}^* E_{X|Y} f\rangle_{L_2(P_X)} = \mathrm{cov}[E(g(X)|Y), E(f(X)|Y)]. \tag{4.2}$$

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

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

which is the right hand side of (4.2). Moreover, since $\tilde{\Sigma}_{XX}^{\frac{1}{2}}$ is isomorphic, we have

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

Hence $\|E_{X|Y}^* E_{X|Y}\| \le \|\tilde{\Sigma}_{XX}^{-1/2}\| \, \|R_{XY}\| \, \|R_{YX}\| \, \|\tilde{\Sigma}_{XX}^{\frac{1}{2}}\|$. Because $\tilde{\Sigma}_{XX}^{\frac{1}{2}}$ and $\tilde{\Sigma}_{XX}^{-1/2}$ are isomorphisms, their norms are both 1. By Baker (1972, Theorem 1), $\|R_{YX}\| \le 1$. Hence $\|E_{X|Y}^* E_{X|Y}\| \le 1$. □

From this corollary we see that the quadratic form

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

generalizes the matrix $\mathrm{var}[E(X|Y)]$ 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.

4.2 Population-Level Estimation of GSIR

We are ready to establish the connection between the operator $E^*_{X|Y}E_{X|Y}$, and the regression class $\mathfrak{C}_{Y|X}$. Let's first look at two lemmas that are useful in our later derivation. Then our main result is presented, which claims that the closure of $E^*_{X|Y}E_{X|Y}$ is identical to $\mathfrak{C}_{Y|X}$.

Lemma 4.1 *If A^* is adjoint of A and A^*A is a compact operator then* $\mathrm{Ker}(A) = \mathrm{Ker}(A^*A)$*; or equivalently,* $\overline{\mathrm{ran}}\,(AA^*) = \overline{\mathrm{ran}}\,(A)$.

Proof. If $f \in \mathrm{Ker}(A)$, then $Af = 0$, and hence

$$\langle f, A^*Af\rangle = \langle Af, Af\rangle = 0$$

Since A^*A is self adjoint and compact we have that

$$\langle f, A^*Af\rangle = 0 \Leftrightarrow f \in \mathrm{Ker}(A^*A).$$

Thus we have proved $\mathrm{Ker}(A) \subseteq \mathrm{Ker}(A^*A)$.

If $f \in \mathrm{Ker}(A^*A)$ then

$$\langle Af, Af\rangle = \langle f, A^*Af\rangle = 0$$

So $f \in \mathrm{Ker}(A)$. □

Let us do a lemma.

Lemma 4.2 *$A : \mathcal{H}_1 \to \mathcal{H}_2$, $B : \mathcal{H}_2 \to \mathcal{H}_3$. Then*

1. $\mathrm{ran}\,(BA) = B\mathrm{ran}\,(A)$.
2. *Suppose $U : \mathcal{H}_1 \to \mathcal{H}_2$ is a bounded linear operator. $P : \mathcal{H}_2 \to \mathcal{H}_2$ is a projection and* $\mathrm{ran}\,(P) \subseteq \mathrm{ran}\,(U)$. *Then* $\mathrm{ran}\,(PU) = \mathrm{ran}\,(P)$.

PROOF. 1. By definition, $\mathrm{ran}\,(BA) = BA\mathcal{H}_1 = B(A\mathcal{H}_1) = B\mathrm{ran}\,A$.

2. We have

$$\mathrm{ran}\,(PU) = P\mathrm{ran}\,(U) \subseteq P\mathcal{H}_2 = \mathrm{ran}\,(P),$$

where the first equality follows from assertion 2, the last equality follows from assertion 1. In the meantime

$$\mathrm{ran}\,(PU) = P\mathrm{ran}\,(U) \supseteq P\mathrm{ran}\,(P) = \mathrm{ran}\,(P^2) = \mathrm{ran}\,(P).$$

This completes the proof. □

Theorem 4.2 *If Assumption 3.1 and Assumption 3.2 are satisfied, then*

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

PROOF. By Theorem 2.10 and Theorem 3.13, $\mathrm{Ker}(\tilde{\Sigma}_{YX})^\perp = \mathfrak{C}_{Y|X}$. Thus it suffices to show that $\overline{\mathrm{ran}}\,(E^*_{X|Y}E_{X|Y}) = \mathrm{Ker}(\tilde{\Sigma}_{YX})^\perp = \overline{\mathrm{ran}}\,(\tilde{\Sigma}^*_{YX})$. Since $\tilde{\Sigma}^{\frac{1}{2}}_{XX}$ is isomorphic and $\Sigma^{\frac{1}{2}}_{YY}$ is self adjoint, we have

$$\overline{\mathrm{ran}}\,\tilde{\Sigma}^*_{YX} = \overline{\mathrm{ran}}\,((\tilde{\Sigma}^{\frac{1}{2}}_{XX})^* R^*_{YX}(\Sigma^{\frac{1}{2}}_{YY})^*) = \overline{\mathrm{ran}}\,(\tilde{\Sigma}^{-1/2}_{XX}R_{XY}\Sigma^{\frac{1}{2}}_{YY}) = \mathrm{cl}[\tilde{\Sigma}^{-1/2}_{XX}R_{XY}\mathrm{ran}\,(\Sigma^{\frac{1}{2}}_{YY})],$$

where the last equality comes from part 2 of Lemma 4.2. Note that, for two operators A and B, $\mathrm{cl}[B\mathrm{ran}\,(A)] = \mathrm{cl}[B\overline{\mathrm{ran}}\,(A)]$. Hence the right hand side above is

$$\mathrm{cl}[\tilde{\Sigma}^{-1/2}_{XX}R_{XY}\mathrm{ran}\,(\Sigma^{\frac{1}{2}}_{YY})] = \mathrm{cl}[\tilde{\Sigma}^{-1/2}_{XX}R_{XY}\overline{\mathrm{ran}}\,(\Sigma^{\frac{1}{2}}_{YY})] = \mathrm{cl}[\tilde{\Sigma}^{-1/2}_{XX}R_{XY}\mathrm{ran}\,(\tilde{\Sigma}^{\frac{1}{2}}_{YY})],$$

where the second equality follows from Theorem 3.10. Let Q_Y be the projection on to $\text{ran}\,(\tilde{\Sigma}_{YY}^{\frac{1}{2}})$. Then $\text{ran}\,(\tilde{\Sigma}_{YY}^{\frac{1}{2}}) = \text{ran}\,Q_Y$, and the right hand side above is equal to

$$\begin{aligned}\text{cl}[\tilde{\Sigma}_{XX}^{-1/2} R_{XY} \text{ran}\, Q_Y] = \text{cl}(\tilde{\Sigma}_{XX}^{-1/2} \text{ran}\, R_{XY}) = \text{cl}(\tilde{\Sigma}_{XX}^{-1/2} \overline{\text{ran}}\, R_{XY}) \\ = \text{cl}[\tilde{\Sigma}_{XX}^{-1/2} \overline{\text{ran}}\,(R_{XY} R_{YX})] = \text{cl}[\tilde{\Sigma}_{XX}^{-1/2} \text{ran}\,(R_{XY} R_{YX})].\end{aligned} \tag{4.3}$$

where the first equality follows from the fact that $R_{XY} = R_{XY} Q_Y$ and part 1 of Lemma 4.2, the third equality follows from Lemma 4.1 and $R_{YX} = R_{XY}^*$. Let Q_X be the projection on to $\text{ran}\,\tilde{\Sigma}_{XX}^{\frac{1}{2}}$. Since $R_{YX} = R_{YX} Q_X$ and $\text{ran}\, Q_X = \overline{\text{ran}}\,\Sigma_{XX}^{\frac{1}{2}} = \text{ran}\,\tilde{\Sigma}_{XX}^{\frac{1}{2}}$, the right hand side of (4.3) is

$$\begin{aligned}\text{cl}[\tilde{\Sigma}_{XX}^{-1/2} \text{ran}\,(R_{XY} R_{YX} Q_X)] &= \text{cl}[\tilde{\Sigma}_{XX}^{-1/2} R_{XY} R_{YX} \text{ran}\,(Q_X)] \\ &= \text{cl}[\tilde{\Sigma}_{XX}^{-1/2} R_{XY} R_{YX} \text{ran}\, \tilde{\Sigma}_{XX}^{\frac{1}{2}}] = \overline{\text{ran}}\,(\tilde{\Sigma}_{XX}^{-1/2} R_{XY} R_{YX} \tilde{\Sigma}_{XX}^{\frac{1}{2}}),\end{aligned}$$

as to be demonstrated. □

Note that, unlike in classical SIR for linear SDR, here we do not have to consider $[\text{var}(X)]^{-1}$ in (4.1). This is because the operator $E_{X|Y}$ is defined on $L_2(P_X)$ rather than $\mathcal{L}_X$; that is, the $L_2(P_X)$-inner product absorbs any unconditional (i.e. marginal) variance in the predictor vector.

When $E_{X|Y}^* E_{X|Y}$ is a compact operator, its spectrum consists of a countable set of eigenvalues, and its range is spanned by the eigenvectors corresponding to these eigenvalues. Let Λ be the (countable) collection of nonzero eigenvalues of $E_{X|Y}^* E_{X|Y}$, which are real numbers because $E_{X|Y}^* E_{X|Y}$ is self-adjoint. Then

$$\text{ran}\,(E_{X|Y}^* E_{X|Y}) = \{f \in L_2(P_X) :\ E_{X|Y}^* E_{X|Y} f = \lambda f,\ \lambda \in \Lambda\}.$$

The GSIR estimator is related to kernel canonical component analysis (KCCA) (Bach and Jordan, 2002 and Fukumizu, Bach, and Gretton, 2007) - see Section 3.4. Later we will explore similarities and differences between these two methods.

4.3 Revisit KSIR

Let's us now tune to another nonlinear SDR method, kernel sliced inverse regression - see Section 3.4. In particular, we consider unbiasedness of this estimator under our formulation. In our setting, the population-level description of this estimator is as follows. Let $\mathcal{H}_X$ be a Hilbert space satisfying Assumption 3.1 and Assumption 3.2 (in this case a RKHS, but this assumption is unnecessary). Let $T : \mathcal{H}_X \to L_2(P_X)$ be the centering transformation. Let $J_1, \ldots, J_h$ be a partition of Ω_Y, and let $\mu_1, \ldots, \mu_h \in \overline{\text{ran}}\, T$ be the Riesz representations of the linear functionals

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

In our language, KSIR uses (the sample version of) the subspace $\text{span}(\Sigma_{XX}^{-1}\mu_1, \ldots, \Sigma_{XX}^{-1}\mu_h)$ to estimate $\mathfrak{S}_{Y|X}$. The next theorem shows that any such Riesz 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 4.3 *If Assumption 3.1 and Assumption 3.2 hold, then* $\mu_j \in \mathfrak{C}_{Y|X}$.

PROOF. By condition Assumption 3.1, $\overline{\text{ran}}\, T = L_2(P_X)$. If $f \in L_2(P_X) \ominus L_2(P_Y) \subseteq \overline{\text{ran}}\, T$, then, by Lemma 2.1, $E(f|Y) = 0$. 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 form of unbiasedness proof of KSIR, but they assume that the spanning functions of $\mathcal{H}_X$ satisfy the linear conditional mean assumption - see Definition 3.4. 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 that they need the assumption in the first place, is that they define the central class (e.d.r. subspace in Definition 1, Yeh, Huang, and Lee (2009)) as the linear subspace spanned by $h_1, \ldots, h_d$ in $\text{span}(f_1, \ldots, f_m)$ such that

$$Y \perp\!\!\!\perp X | h_1(X), \ldots, h_d(X); \tag{4.4}$$

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 (4.4) only depends on the σ-field generated by $h_1, \ldots, h_d$.

4.4 Sample-Level Estimation of GSIR

In the population level we have shown $\mathfrak{S}_{Y|X}$ can be estimated via the range of the $E^*_{X|Y}E_{X|Y}$; furthermore, this estimation is guaranteed to be exhaustive when the central class is complete. In this section we demonstrate how to construct the central class using finite sample. We first derive the matrix representation of $E^*_{X|Y}E_{X|Y}$; then through a spectral analysis of this matrix, $\mathfrak{S}_{Y|X}$ can be identified. Throughout this section, $A^\dagger$ represents the Moore-Penrose inverse of a matrix A, and $A^{\dagger\,\alpha}$ 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.

Matrix representations of operators

Our targeting class $\mathfrak{S}_{Y|X}$ is a subspace in $L_2(P_X)$ that is generally infinite-dimensional. Therefore, we require a finite sample approximation (or representation) in order to make the calculation possible; in other words, data of a functional form can be represented using finite sample and stored in a matrix or array.

To facilitate our derivation, we use the same notional system adopted by Li, Chun, and Zhao (2011). Let $\mathcal{H}$ be a generic finite-dimensional Hilbert space with spanning system $\mathcal{B} = \{b_1, \ldots, b_n\}$. For an $f \in \mathcal{H}$, let $[f]_\mathcal{B}$ denote the coordinates of f relative to $\mathcal{B}$; that is, $f = \sum_{i=1}^n ([f]_\mathcal{B})_i b_i$. Let $B(\cdot) : \Omega_X \to \mathbb{R}^n$ denote the $\mathbb{R}^n$-valued function $(b_1(\cdot), \ldots, b_n(\cdot))^\mathsf{T}$. Then we can write f as $B^\mathsf{T}(\cdot)[f]_\mathcal{B}$. In this section we reserve the square brackets $[\cdot]$ exclusively for coordinates. Let $A : \mathcal{H} \to \mathcal{H}'$, where $\mathcal{H}'$ is another finite-dimensional Hilbert spaces with spanning system $\mathcal{C} = \{c_1, \ldots, c_m\}$. Let $C(\cdot) = (c_1(\cdot), \ldots, c_m(\cdot))^\mathsf{T}$. Then, for $f \in \mathcal{H}$,

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

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

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

the coordinate vector of Af is a matrix times the coordinate vector $[f]_{\mathcal{B}}$. We denote this matrix by ${}_{\mathcal{C}}[A]_{\mathcal{B}}$. We define

$${}_{\mathcal{C}}[A]_{\mathcal{B}} = \begin{pmatrix} ([Ab_1]_{\mathcal{C}})_1 & \cdots & ([Ab_n]_{\mathcal{C}})_1 \\ \vdots & \ddots & \vdots \\ ([Ab_1]_{\mathcal{C}})_m & \cdots & ([Ab_n]_{\mathcal{C}})_m \end{pmatrix}.$$

Let $C(\cdot)$ denote the row vector $(c_1(\cdot), \ldots, c_n(\cdot))$. Then, by definition

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

In other words, $[Af]_{\mathcal{B}} = ({}_{\mathcal{C}}[A]_{\mathcal{B}})[f]_{\mathcal{B}}$. 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,

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

In the following, the spanning systems in the domain and range of an operator are self evident. So we simply use $[A]$ and $[f]$ to denote the coordinates of operators and functions, without explicit reference to the spanning systems.

Suppose $A \in \mathcal{B}(\mathcal{H})$ is self-adjoint, where $\mathcal{H}$ is an n-dimensional Hilbert space. Let $I : \mathcal{H} \to \mathcal{H}$ be the identity mapping. Let $\lambda_1, \ldots, \lambda_m$ be distinct nonzero eigenvalues of A, $P_1, \ldots, P_m$ be the projections on to the corresponding eigenspaces $\mathrm{Ker}(A - \lambda_i I)$, $i = 1, \ldots, m$, and P_0 be the projection onto $\mathrm{Ker}(A)$. It can be shown that $\lambda_1, \ldots, \lambda_m$ are the nonzero distinct eigenvalues of $[A]$, $[P_i]$ are the projection matrices onto the corresponding eigenspaces $\mathrm{Ker}([A] - \lambda_i I_n)$, and $[P_0]$ is the projection matrix onto $\mathrm{Ker}([A])$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a function, and define $\phi(A) = \sum_{i=1}^m \phi(\lambda_i) P_i + \phi(0) P_0$ and $\phi([A]) = \sum_{i=1}^m \phi(\lambda_i)[P_i] +$

$\phi(0)[P_0]$. Then it can be shown that $[\phi(A)] = \phi([A])$. Thus, for any $f : \mathcal{H} \to \mathbb{R}$ we have

$$\phi(A)f = \phi(A)B(\cdot)[f] = B(\cdot)[\phi(A)][f] = B(\cdot)\phi([A])[f].$$

For example, $[A^{-1/2}] = [A]^{-1/2}$, and $[A^{\frac{1}{2}}] = [A]^{\frac{1}{2}}$. Equipped with these notations we can represent the various operators introduced earlier at the sample level. Every finite-dimensional Hilbert space is isomorphic to a Euclidean space; therefore, properties in linear algebra can be applied to the functions or operators that are in forms of matrices. See Horn and Johnson (1985, P. 31) for more detail.

Sample version for GSIR

We develop here the algorithm for the GSIR and discuss its difference from, and relation with, the kernel canonical correlation analysis (KCCA) introduced by Bach and Jordan (2002), and Fukumizu, Bach, and Gretton (2007).

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample of (X, Y). At the sample level, $\Omega_X = \{X_1, \ldots, X_n\}$. Let $P_{n,X}$ represent the empirical distribution based on $X_1, \ldots, X_n$. Let κ_X be as defined before, and $\mathcal{H}_X$ be the RKHS spanned by the functions

$$\mathcal{B}_X = \{\kappa_X(\cdot, X_1), \ldots, \kappa_X(\cdot, X_n)\}.$$

Let $M_{XX} : \mathcal{H}_X \to \mathcal{H}_X$ be defined by the relation

$$\langle f, M_{XX}\, g\rangle_{\mathcal{H}_X} = \mathrm{cov}_n[f(X), g(X)], \tag{4.5}$$

where $\mathrm{cov}_n(\cdot, \cdot)$ denotes the sample covariance. Let Q be the projection matrix $I_n - 1_n 1_n^\intercal/n$. Let $\kappa_X(\cdot)$ denote the $\mathbb{R}^n$-valued function $(\kappa_X(\cdot, X_1), \ldots, \kappa_X(\cdot, X_n))^\intercal$, and K_X denote the $n \times n$ matrix $\{\kappa_X(X_i, X_j)\}$.

Lemma 4.3 *If $g_1, g_2 \in \mathcal{H}_Y$, then*

$$\langle g_1, g_2\rangle_{\mathcal{H}_Y} = [g_1]_{\mathcal{B}_Y}^\intercal K_Y [g_2]_{\mathcal{B}_Y}.$$

PROOF. We have

$$\begin{aligned}\langle g_1, g_2\rangle_{\mathcal{H}_Y} &= \langle \sum_{i=1}^{n} ([g_1]_{\mathcal{B}_Y})_i \kappa_Y(\cdot, Y_i), \sum_{j=1}^{n} ([g_2]_{\mathcal{B}_Y})_j \kappa_Y(\cdot, Y_j)\rangle_{\mathcal{H}_Y} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} [g_1]_{\mathcal{B}_Y})_i ([g_2]_{\mathcal{B}_Y})_j \langle (\kappa_Y(\cdot, Y_i), \kappa_Y(\cdot, Y_j)\rangle_{\mathcal{H}_Y} \\ &= [g_1]_{\mathcal{B}_Y}^{\mathsf{T}} K_Y [g_2]_{\mathcal{B}_Y}\end{aligned}$$

as desired. □

Proposition 4.1 *Suppose K_X is nonsingular. Then*

1. $[M_{XX}\,] = QK_X/n$;
2. *for any* $f \in \operatorname{ran} M_{XX}$, $[f] = Q[f]$.

PROOF. 1. By direct computation, it is easy to verify that, for any $f, g \in \mathcal{H}_X$,

$$\begin{aligned}\operatorname{cov}_n[f(X), g(X)] &= [f]^{\mathsf{T}}(K_X Q K_X/n)[g] \\ \langle f, M_{XX}\, g\rangle_{\mathcal{H}_X} &= [f]^{\mathsf{T}} K_X [M_{XX}\, g] = [f]^{\mathsf{T}} K_X [M_{XX}\,][g].\end{aligned}$$

Hence $[f]^{\mathsf{T}} K_X [M_{XX}\,][g] = [f]^{\mathsf{T}} K_X Q K_X [g]$. Since this holds for all $[g], [f] \in \mathbb{R}^p$, we have $K_X[M_{XX}\,] = K_X Q K_X/n$, which implies $[M_{XX}\,] = QK_X/n$.

2. For any $f \in \mathcal{L}_X$,

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

So $[f] = n^{-1} Q K_X [g] = Q(n^{-1} Q K_X [g]) = Q[f]$. □

Let us now find the representations of Σ_{XX}, Σ_{YY}, and Σ_{YX}. For any $f, g \in \mathcal{L}_X$,

$$\langle f, g\rangle_{\mathcal{H}_X} = [f]^{\mathsf{T}} K_X [g] = [f]^{\mathsf{T}} Q K_X Q [g] \equiv [f]^{\mathsf{T}} G_X [g],$$

where $G_X = QK_XQ$. In the following, $L_2(P_{n,X})$ and $L_2(P_{n,Y})$ denote the centered L_2-space based on the empirical measures $P_{n,X}$ and $P_{n,Y}$. For example, $L_2(P_{n,X})$ is the space spanned by $\{\kappa_X(\cdot, X_i) - E_n\kappa_X(X, X_i) : i = 1, \ldots, n\}$ with inner product $\langle f, g\rangle_{L_2(P_{n,X})} = n^{-1}[f]^\intercal K_X Q K_X [g]$.

Proposition 4.2 *If K_X is nonsingular, then*

$$
\begin{aligned}
[\Sigma_{XX}] &= [\tilde{\Sigma}_{XX}] = n^{-1}G_X, \\
[\Sigma_{YY}] &= [\tilde{\Sigma}_{YY}] = n^{-1}G_Y, \\
[\Sigma_{YX}] &= [\tilde{\Sigma}_{YX}] = n^{-1}G_X.
\end{aligned}
$$

PROOF. For any $f \in \mathcal{L}_X$, we have

$$
\Sigma_{XX} f = M_{XX}\, f = K_X^\intercal(\cdot)[M_{XX}][f] = n^{-1}K_X^\intercal(\cdot)QK_XQ[f] = n^{-1}K_X^\intercal(\cdot)G_X[f].
$$

Hence $[\Sigma_{XX} f] = [\Sigma_{XX}][f] = n^{-1}G_X[f]$. Since this is true for all $[f] \in \text{span}(Q)$, we have $[\Sigma_{XX}] = n^{-1}G_X$. The rest of the equalities can be proved similarly. □

Since $E_{X|Y} f = K_Y^\intercal(\cdot)Q[E_{X|Y}][f]$, we have

$$
\begin{aligned}
\langle f, E_{X|Y}^* E_{X|Y} f\rangle_{L_2(P_{n,X})} &= \langle K_Y^\intercal(\cdot)Q[E_{X|Y}][f], K_Y^\intercal(\cdot)Q[E_{X|Y}][f]\rangle_{L_2(P_{n,Y})} \\
&= [f]^\intercal [E_{X|Y}]^\intercal G_Y^2 [E_{X|Y}][f].
\end{aligned}
$$

The sample version of $\tilde{\Sigma}_{XX}^{\frac{1}{2}}$ is $G_X^{\frac{1}{2}}$. The sample version of R_{YX} is $(G_Y + \epsilon I_n)^{-1/2} G_X (G_X + \epsilon I_n)^{-1/2}$. The sample version of $\tilde{\Sigma}_{YY}^{-1/2}$ is $(G_Y + \epsilon I_n)^{-1/2}$. Thus,

$$
\begin{aligned}
[E_{X|Y}] &= [\tilde{\Sigma}_{YY}^{-1/2}][R_{YX}][\tilde{\Sigma}_{XX}^{\frac{1}{2}}] \\
&= (G_Y + \epsilon I_n)^{-1/2}(G_Y + \epsilon I_n)^{-1/2} G_X (G_X + \epsilon I_n)^{-1/2} (G_X)^{\frac{1}{2}} \\
&= (G_Y + \epsilon I_n)^{-1} G_X^{3/2} (G_X + \epsilon I_n)^{-1/2}.
\end{aligned}
$$

Now let us look at the side condition

$$\langle f, f\rangle_{L_2(P_{n,X})} = \langle K_X^\intercal(\cdot)Q[f], K_X^\intercal(\cdot)Q[f]\rangle_{L_2(P_{n,X})} = n^{-1}[f]^\intercal G_X^2[f] = 1.$$

Let $\phi = G_X[f]$. Then $[f] = (G_X + \epsilon I_n)^{-1}\phi$. So we would like to maximize

$$\phi^\intercal (G_X + \epsilon I_n)^{-3/2} G_X^{3/2} (G_Y + \epsilon I_n)^{-1} (G_Y)^2 (G_Y + \epsilon I_n)^{-1} G_X^{3/2} (G_X + \epsilon I_n)^{-3/2} \phi$$

subject to $\phi^\intercal\phi = 1$. In other words, our solution is $[f] = (G_X + \epsilon I_n)^{-1}\phi$, where ϕ is the eigenvector of the matrix

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

corresponding to nonzero eigenvalues.

The GSIR estimator is similar to a method introduced by Bach and Jordan (2002), Fukumizu, Bach, and Gretton (2007), which they call the KCCA. They proposed to maximize

$$\langle g, \Sigma_{YX} f\rangle_{\mathcal{L}_Y} = [g]^\intercal G_Y G_X [f]$$

subject to $\langle g, \Sigma_{YY} g\rangle_{\mathcal{L}_Y} = [g]^\intercal G_Y^2 [f] = 1$ and $\langle f, \Sigma_{XX} f\rangle_{\mathcal{L}_X} = [g]^\intercal G_Y^2 [g] = 1$. Let $\phi = G_X[f]$ and $\psi = G_Y[g]$. It is equivalent to maximize

$$\psi^\intercal (G_Y + \epsilon I_n)^{-1} G_Y G_X (G_X + \epsilon I_n)^{-1} \phi$$

subject to $\phi^\intercal\phi = \psi^\intercal\psi = 1$. The solution for ϕ is the eigen problem

$$(G_X + \epsilon I_n)^{-1} G_X G_Y (G_Y + \epsilon I_n)^{-2} G_Y G_X (G_X + \epsilon I_n)^{-1} \phi = \lambda\phi$$

The coefficient of the estimated predictors are then taken to be $[f] = (G_X + \epsilon I_n)^{-1}\phi$. We will compare these two methods by simulations and real world datasets.

Parameters Selection

In the estimating procedure of GSIR, the values of two parameters need to be determined. Since we will use Gaussian kernel (Section 3.3), there is a width parameter γ in the kernel function; another parameter is ridge regression parameter ϵ which takes of the inverses of matrices. We now introduce a data-based process of turning the parameters.

Two parameters deploy similar smooth effect on the data; that is, smaller ϵ or larger γ are in favor of mode complicated models, while larger ϵ or smaller γ are better for smooth models. Therefore, it's reasonable to fix one of them and choice the other; this trick is also considered in Fukumizu, Bach and Jordan (2009). Throughout our numerical analyses, we fix the value of ϵ and make it proportional to the dimensionality of the associating random vector

$$\epsilon_X = .01 \times \dim(X), \quad \text{and} \quad \epsilon_Y = .01 \times \dim(Y). \tag{4.6}$$

We then have the values of γ the width parameter left to be selected. To this end, we propose a goodness-of-fit criterion based on the mean square error of a kernel-transforming variable. Specifically, let $(X_1, \cdots, X_n)$ be a finite sample, then any element in $\mathcal{H}_X$ has a vector-value representation in terms of basis $\mathcal{B}_X(x)$

$$\mathcal{B}_X(x) = \{\kappa_X(x, X_1), \cdot, \kappa_X(x, X_n)\}.$$

Suppose $E(\mathcal{B}_X(X)|Y)$ is the conditional mean of the random vector $\mathcal{B}_X(X)$ given Y, and let its estimate denoted as $\widehat{\mathcal{B}_X}(X)$. Then the MSE of $E(\mathcal{B}_X(X)|Y)$ can be computed by $\sum_{i=1}^n ||\mathcal{B}_X(X_i) - \widehat{\mathcal{B}_X}(X_i)||^2$. Combined with the leave-one-out procedure, we can calculate the predicted residual $\Delta_X(i)$ and the PRESS:

$$\begin{aligned} \Delta_X(i) &= \mathcal{B}_X(X_i) - \widehat{\mathcal{B}_X}(X_{i,-i}), \quad \text{and} \\ \Delta_X &= \sum_{i=1}^n ||\Delta_X(i)||^2, \end{aligned} \tag{4.7}$$

where $\widehat{\mathcal{B}_X}(X_{i,-i})$ is the predicted value of $\mathcal{B}_X(Xi)$ without entry of X_i.

Procedure of selecting γ_X and γ_Y:

1. Determine the searching domain of γ_X.

$$\gamma_x \in (\frac{1}{3}\rho_x, 3\rho_x) \text{ with } 1/\sqrt{\rho_x} = \binom{n}{2}^{-1} \sum_{i<j} ||X_i - X_j||.$$

2. Calculate K_X, the first row of $K_X(\gamma_X)$ is $1_n^\intercal$; the bottom $n \times n$ block is the matrix $\{\kappa_X(X_i, X_j) : i, j = 1, \ldots, n\}$. Note that this kernel matrix depends on γ_X.

$$\begin{aligned}\Delta_X(i) = &(K_Y)_{-(k+1),i} - (K_Y)_{-(i+1),-i}(K_X)^\intercal_{-(i+1),-i} \\ &[(K_X)_{-(i+1),-i}\,(K_X)^\intercal_{-(i+1),-i} + \epsilon_X\, I_n]^{-1}(K_X)^\intercal_{-(i+1),i}\end{aligned}$$

   Here, for a generic $r \times s$ matrix A, $(A)_{-i,-j}$ denote the $(r-1) \times (s-1)$ matrix with the i^{th} row and the j^{th} column of A deleted; $(A)_{-i,j}$ denotes the j^{th}column of the $(r-1) \times s$ matrix obtained by deleting the i^{th} row of the matrix A.

3. Compute PRESS Δ_X and minimize this function over a grid of domain in 1 to find the optimal γ_X.

4. Switch the positions of X and Y and repeat 1-3 to find the best value of γ_Y

Chapter 5

Generalized Sliced Average Variance Estimation (GSAVE) - Second Order Method

We develop GSIR in Chapter 4 as an estimator for regression class $\mathfrak{C}_{Y|X}$, which is identical to the central class $\mathfrak{S}_{Y|X}$ when $\mathfrak{S}_{Y|X}$ is a complete and sufficient dimension class. Starting this chapter, we consider the issue of how to estimate the central class when the completeness does not hold, so that GSIR still provides an unbiased estimator of $\mathfrak{S}_{Y|X}$, but not necessarily exhaustive.

First of all we show the existence of *incomplete sufficient dimension reduction class* and supply examples. Via the example of the inverse regression model (Section 2.5), here we establish a connection to the incompleteness in its classical perspective. Secondly, an innovative *heteroscedastic conditional variance operator* for L_2 class is introduced; based on this operator we generalize Sliced Average Variance Estimator (SAVE, Section 2.1), to a method (GSAVE) that is unbiased for $\mathfrak{S}_{Y|X}$, and is capable to estimate a larger class than $\mathfrak{C}_{Y|X}$. We observe that, the relationship between $\mathfrak{C}_{Y|X}$ and $\mathfrak{S}_{Y|X}$ is similar to relationship between the central subspace $\mathcal{S}_{Y|X}$ and central mean subspace $\mathcal{S}_{E(Y|X)}$ - see Sections 2.1 and 2.3. Last, the population- and sample-level estimator for GSAVE are derived, along with a step-by-step algorithm.

5.1 Incomplete Sufficient Dimension Reduction Class

Let $f, g \in L_2(P_X)$ satisfying $\sigma(f(X), g(X)) = \sigma(X)$. Suppose $P_X \circ f^{-1}$ is dominated by Lebesque measure and $f(X)$ and Y has the following relation:

$$f(X) = Y + \epsilon, \tag{5.1}$$

where $Y \perp\!\!\!\perp \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We have shown $\sigma(f(X))$ is the central SDR σ-field, provided that $g(X)$ is conditionally independent with Y given $f(X)$ - see Section 2.5. Furthermore, $\mathcal{M}_{\sigma(f(X))}$ is a complete SDR class. The inverse type of Models provides an alternative way to characterize the dependence between X and Y. The forward conditional distribution $Y|X$ often serves as prime interest in the context of sufficient dimension reduction, while the backward model takes into account the distribution of $X|Y$. Nonetheless, the ultimate goal for SDR is to seek important directions in the sapce of X such that the information of $Y|X$ can be preserved. When the function f in (5.1) is linear, i.e. $f(\cdot) = \Gamma^{\mathsf{T}}$ a matrix in $\mathbb{R}^{d\times p}$, it coincides with the *principle fitted components model* (PFC, Cook (2007), Cook and Forzani (2009), and Cook, Li, and Chiaromonte (2010)).

PFC assumes a probabilistic model for $X|Y$, then one can show that $\Gamma^{\mathsf{T}} X$ is a sufficient dimension reduction subspace of $Y|X$ (Cook (2007), Proposition 1). In their discussion $X|Y$ has to be normally distributed or at least belonged to the exponential family. Below we provide a more general result without assuming a specific distribution.

Proposition 5.1 *Suppose the family of probability measures $\mathcal{P}_{\Omega_Y} = \{P_{X|Y}(\cdot|y) : y \in \Omega_Y\}$ is dominated by a σ-finite measure. Let $f(X)$ be a statistic. Then the following two statements are equivalent:*

1. $P_{X|f(X),Y}(\cdot|f(x), y) = P_{X|f(X)}(\cdot|f(x))$, *and*
2. $P_{Y|f(X),X}(\cdot|f(x), x) = P_{Y|f(X)}(\cdot|f(x))$.

The above result is obvious because statements 1 and 2 are simply two identical representations for the conditional independence of Y and X given $f(X)$. It also explains why PFC model can be used for the purpose of sufficient dimension reduction. In statement 1,

when treating each $y \in \Omega_Y$ as a parameter of the conditional distribution $X|Y = y$, $f(X)$ becomes a sufficient statistic of $\mathcal{P}_{\Omega_Y}$. On the other hand, statement 2 claims that when such $f(X)$ is available it forms a sufficient dimension reduction space for the central subspace $\mathcal{S}_{Y|X}$; that is, $Y|X, \mathcal{G} \overset{\mathcal{D}}{=} Y|\mathcal{G}$ where $\mathcal{G} = \sigma(f(X))$.

Corollary 5.1 *Suppose the family of probability measures $\mathcal{P}_{\Omega_Y} = \{P_{X|Y}(\cdot|y) : y \in \Omega_Y\}$ is dominated by a σ-finite measure. Let $f(X)$ be a statistic and $\mathcal{M}_{\mathcal{G}}$ is the class of functions in $L_2(P_X)$ that are measurable with respect to $\mathcal{G} = \sigma(f(X))$. Then the following two statements are equivalent:*

1. *$f(X)$ is a (complete) minimal sufficient statistic of $\mathcal{P}_{\Omega_Y}$, and*

2. *$\mathcal{M}_{\mathcal{G}}$ is a (complete) central dimension reduction class.*

PFC models seeks a linear subspace in $\mathcal{R}^p$ that is sufficient in $Y|X$, but not necessarily complete or minimal. From the above corollary we can see that, when the property of minimality (or completeness) holds in the backward relation $X|Y$, the SDR class generated by the sufficient statistic is also a minimal (or complete) SDR class. This observation again resembles classical and nonlinear notions of sufficiency, minimality and completeness. In this point of view, one can search important features from the space of X that is sufficient to the forward regression $Y|X$ while the model is built on an inverse regression $X|Y$.

Models of incompleteness

The minimality and completeness of sufficient statistics are rather different concepts in classical setting - See Lehmann (1981). A complete and sufficient statistic ensures the existence of minimal sufficient statistics, while the other direction is not always true, i.e. a minimal sufficient statistic is not necessarily complete. Since our formulation of central SDR class and the notion of completeness are defined in parallel to the classical setting, this suggests that there can exist central classes that are not complete. The following example offers an example of an incomplete central class.

Example 5.1 (Incomplete central SDR class) *Consider the following model:*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} Y \\ Y^2 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right). \tag{5.2}$$

It can be shown that $\sigma(X_1, X_2)$ is the central SDR σ-field; however, its associated central class $\mathcal{M}_{X_1,X_2}$ is not complete. In order to show this, we consider a function $f(X_1, X_2) = X_1^2 - X_2$. It's not hard to see that the conditional expectation $E(f(X)|Y)$ is constant in $y \in \Omega_Y$, which implies $\mathcal{M}_{X_1,X_2}$ is not complete.

5.2 Heteroscedastic Conditional Variance Operator

We now consider a more general problem of estimating the central class $\mathfrak{S}_{Y|X}$ when completeness does not hold, in which case the regression class $\mathfrak{C}_{Y|X}$ may be a proper subset of $\mathfrak{S}_{Y|X}$. Here, again, there is an analogy with sufficient linear dimension reduction, where the central class $\mathfrak{S}_{Y|X}$ corresponds to the central subspace $\mathcal{S}_{Y|X}$ and the complete central class $\mathfrak{C}_{Y|X}$ corresponds to the central mean subspace $\mathcal{S}_{E(Y|X)}$. 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 and 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 $\mathcal{L}_Y$ and $\mathcal{L}_X$. 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, and $\mathcal{L}_X$ and $\mathcal{L}_Y$ still denote the centered L_2-spaces. Define the noncentered conditional mean operator $E'_{X|Y} : L_2'(P_X) \to L_2'(P_Y)$ through the relation

$$\langle g, E'_{X|Y} f \rangle_{\mathcal{L}_Y} = E(g(Y)f(X)), \;\; f \in L_2'(P_X), \; g \in L_2'(P_Y). \tag{5.3}$$

By the same argument of Proposition 4.1, $E'_{X|Y} f = E(f(X)|Y)$. To generalize SAVE, we introduce a new type of conditional variance operator.

Definition 5.1 *For each $y \in \Omega_Y$, the bilinear form*

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

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

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

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) - see Section 3.5. In a sense, $\Sigma_{X|Y}$ is a generalization of $E[\text{var}(X|Y)]$ rather than $\text{var}(X|Y)$, because $\langle f, \Sigma_{X|Y} f \rangle_{\mathcal{H}_X} = E[\text{var}(f(X)|Y)]$. Note that $E[\text{var}(f(X)|Y)]$ becomes $\text{var}(f(X)|Y)$ 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 $\text{var}[f(X)|Y]$, hence the term heteroscedastic conditional variance operator. Here, we should also stress that $E'_{X|Y}$ is defined between the non-centered classes $L'_2(P_X)$ and $L'_2(P_Y)$, whereas $V_{X|Y}(y)$ is defined between centered classes $\mathcal{L}_X$ and $\mathcal{L}_X$.

5.3 Population-Level Estimation of GSAVE

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 $x \mapsto \int_{\Omega_Y} (A(y)f)(x) P_Y(dy)$, 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

$$S = E(V - V_{X|Y})^2 : L_2(P_X) \to L_2(P_X), \tag{5.4}$$

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

$$\langle f, Vg \rangle_{\mathcal{L}_X} = \text{cov}(f(X), g(X)).$$

This operator is similar to $\tilde{\Sigma}_{XX}$ in Chapter 3 and 4, except that it is not defined through RKHS. The operator S is an extension of the SAVE matrix (Cook and Weisberg, 1991) - see also Section 2.1.

$$\Sigma^{-1} E[\text{var}(X) - \text{var}(X|Y)]^2 \Sigma^{-1}. \tag{5.5}$$

Let B be a basis matrix of the central subspace $\mathcal{S}_{Y|X}$ of linear SDR. Cook and Weisberg show that, if Assumption 2.1 and Assumption 2.2 hold, then the column space of (5.5) is contained in $\mathcal{S}_{Y|X}$. We provide a more general result, but without requiring an analogue of Assumption 2.1. Let's do a lemma before showing our main theorem.

Lemma 5.1 *Let $\mathcal{H}$ be a Hilbert space and $A : \mathcal{H} \to \mathcal{H}$ is a compact, self-adjoint, and positive semi-definite operator. Then $h \in \text{Ker}A$ if and only if $\langle h, Ah \rangle = 0$.*

PROOF. Necessity is obvious. To prove sufficiency, let h be a member of $\mathcal{H}$ such that $\langle h, Ah \rangle = 0$. By Theorem 5.1 of Conway (1990), any compact, self-adjoint, and positive semi-definite operator $A : \mathcal{H} \to \mathcal{H}$ has the spectral decomposition

$$A = \sum_{i=1}^{\infty} \lambda_i P_i,$$

where $\{\lambda_1, \lambda_2, \ldots\}$ are positive and distinct eigenvalues of A, and P_i is the projection on to $\text{Ker}(A - \lambda_i I)$, I being the identity mapping from $\mathcal{H}$ to $\mathcal{H}$. Moreover, $P_i P_j = 0$ whenever $i \neq j$. Note that

$$Ah = \sum_{i=1}^{\infty} \lambda_i P_i(h) = \left(\sum_{j=1}^{\infty} \lambda_j P_j \right) \left(\sum_{i=1}^{\infty} P_i(h) \right) = A \sum_{i=1}^{\infty} P_i(h). \tag{5.6}$$

Hence

$$0 = \langle h, Ah \rangle = \left\langle \sum_{i=1}^{\infty} P_i(h), A \sum_{i=1}^{\infty} P_i(h) \right\rangle = \sum_{i=1}^{\infty} \lambda_i \|P_i(h)\|^2.$$

Because $\lambda_i > 0$ for all i, the norms $\|P_i(h)\|$ must be 0 for all i, implying $P_i(h) = 0$ for all i. Then, by (5.6), $Ah = 0$, as desired. □

Theorem 5.1 *Suppose Assumption 3.1 and Assumption 3.2 are satisfied, and* $\text{var}[f(X)|\mathcal{G}_{Y|X}]$ *is nonrandom for any* $f \in \mathfrak{S}_{Y|X}^{\perp}$. *Then* $\overline{\text{ran}}\, S \subseteq \mathfrak{S}_{Y|X}$.

Proof. It suffices to show that ${\mathfrak{S}_{Y|X}}^{\perp} \subseteq \text{Ker}S$. Let $f \in {\mathfrak{S}_{Y|X}}^{\perp}$. We claim that for any $y \in \Omega_Y$,

$$\langle f, [V - V_{X|Y}(y)]f \rangle_{L_2(P_X)} = 0. \tag{5.7}$$

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

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

Because, by Lemma 2.1, $E(f(X)|\mathcal{G}_{Y|X})$ is constant, the first term is 0. Because $\text{var}(f(X)|\mathcal{G}_{Y|X})$ is non-random, the second term is $\text{var}(f(X)|\mathcal{G}_{Y|X})$. Hence

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

Similarly,

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

Therefore $\text{var}(f(X)|Y) = \text{var}(f(X))$, which implies (5.7). Since $V - V_{X|Y}(y)$ is self-adjoint, (5.7) implies $f \in \text{Ker}(V - V_{X|Y}(y))$. 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

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

By Lemma 5.1, $f \in \text{Ker} E(V - V_{X|Y})^2$, as desired. □

Similar to GSIR, here we do not need to consider $[\text{var}(X)]^{-1}$; that is, when generalizing SAVE we do not need to employ the rescaling by Σ^{-1} in (5.5). This is because the $L_2(P_X)$-inner product absorbs any marginal variance. We call the estimator derived from $\overline{\text{ran}}\, S$ *generalized SAVE*, or GSAVE. The next theorem shows that GSAVE can recover functions outside $\mathfrak{C}_{Y|X}$.

Theorem 5.2 *If Assumption 3.1 and Assumption 3.2 are satisfied, then* $\mathfrak{C}_{Y|X} \subseteq \overline{\text{ran}}\, S$.

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

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

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

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

By Definition 5.1, the left hand side is $\text{var}[f(X)] - E[\text{var}(f(X)|Y)] = \text{var}[E(f(X)|Y)]$. Hence $\text{var}[E(f(X)|Y)] = 0$, which implies $E[f(X)|Y] = E[f(X)] = 0$. By Lemma 2.1, we have $f \in \mathcal{L}_X \ominus \mathcal{L}_Y = \mathfrak{C}_{Y|X}^{\perp}$, as desired. □

Combining Theorems 5.1 and 5.2 we see that

$$\mathfrak{C}_{Y|X} \subseteq \overline{\operatorname{ran}}\, S \subseteq \mathfrak{S}_{Y|X}. \tag{5.8}$$

In linear SDR, Cook and Li (2002) showed that the

$$\mathcal{S}_{E(Y|X)} \subseteq \mathcal{S}_{\text{SAVE}} \subseteq \mathcal{S}_{Y|X}, \tag{5.9}$$

where $\mathcal{S}_{\text{SAVE}}$ is the subspace of $\mathbb{R}^p$ spanned by the columns of the matrix in (5.5). 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 central mean subspace.

5.4 Sample-Level Estimation of GSAVE

In this section we derive the estimate of of $\operatorname{ran} S_n$, where $S_n : L_2(P_{n,X}) \to L_2(P_{n,X})$ is the operator $E_n(V - V_{X|Y})^2$. This corresponds to solving the generalized eigenvalue problem

$$\text{maximize } \langle f, S_n f\rangle_{L_2(P_{n,X})} \quad \text{subject to } \langle f, f\rangle_{L_2(P_{n,X})}. \tag{5.10}$$

The sample version of the noncentered L_2-classes $L_2'(P_{n,X})$ and $L_2'(P_{n,Y})$ are spanned by

$$\{1, K_X(\cdot, X_1), \ldots, K_X(\cdot, X_n)\}, \quad \{1, \kappa_y(\cdot, Y_1), \ldots, \kappa_y(\cdot, Y_n)\}, \tag{5.11}$$

respectively. Let K_X be the $(n+1) \times n$ matrix whose first row is 1_n^{T} and the lower $n \times n$ block is $\{K_X(X_i, X_j)\}_{i,j=1}^n$, and let K_Y be defined similarly. Let $[\cdot]$ represent the coordinates relative to spanning systems (5.11).

We present some lemma in the following that are essential to our later derivation.

Lemma 5.2 *Let K_X be the matrix mentioned above. Then the following two properties*

hold.

$$1.\ E_n[f(X)g(X)] = [f]^\intercal (K_X K_X^\intercal/n)[g], \quad \textit{and}$$
$$2.\ \mathrm{cov}_n[f(X)g(X)] = [f]^\intercal (K_X Q K_X/n)[g],$$

for $f, g \in L_2'(P_{n,X})$; $Q = I_n - 1_n 1_n^\intercal/n$.

PROOF. The result is immediate and omitter. □

We now define the inverse conditional mean operator $E'_{X|Y}$ via the relation, for $f \in L_2'(P_{n,X})$ and $g \in L_2'(P_{n,Y})$,

$$\langle g, E'_{X|Y} f \rangle_{L_2'(P_{n,Y})} = E_n[g(Y)f(X)]. \tag{5.12}$$

Then the following lemma verifies the matrix representation of $E'_{X|Y}$, which can be written as

$$[E'_{X|Y}] = (K_Y K_Y^\intercal)^\dagger (K_Y K_X^\intercal), \tag{5.13}$$

where † denotes the Moore-Penrose inverse.

Lemma 5.3 *Let K_X and K_Y be defined above. Then the equality in (5.13) holds.*

PROOF. Note that

$$\begin{aligned} \langle g, E'_{X|Y} f \rangle_{L_2(P_{n,Y})} &= [g]^\intercal (K_Y K_Y^\intercal/n)[E'_{X|Y}][f], \quad \text{and} \\ E_n[f(X)g(Y)] &= [g]^\intercal (K_Y K_X^\intercal/n)[f]. \end{aligned}$$

The result is shown by equating above two equations. □

Let $K_Y(\cdot)$ denote the function $y \mapsto (1, K_Y(y, Y_1), \ldots, K_Y(y, Y_n))^\intercal$, and let $K_X(\cdot)$ denote the similar function of x. Then we can express the product of functions, i.e. $f \times g$, in terms of matrix representation.

Lemma 5.4 *Let $[fg]$ be the matrix representation of $f \times g$. Then we have*

$$[fg] = (K_X K_X^\intercal)^\dagger K_X (K_X[f] \odot K_X^\intercal[g]),$$

where $\odot$ is the Hadamard product.

Proof. It it true that, for $fg(\cdot)$, its evaluation at $x \in \Omega_X$ is

$$f(x)g(x) = K_X(x)^\intercal [fg].$$

When evaluated at $x = X_1, \cdots, X_n$, we have

$$K_X[f] \odot K_X[g] = K_X^\intercal [fg].$$

Multiply both sides by K_X to close the proof. □

Lemma 5.5 *The covariance of $f(X)$ and $g(X)$ has the quadratic form*

$$\mathrm{cov}[f(X), g(X)|Y = y] = [f]^\intercal K_X \Lambda(y) K_X^\intercal [g], \quad \forall f, g \in L_2(P_{n,X}),$$

where $\Lambda(y) = \mathrm{diag}(C_Y(y)) - C_Y(y)C_Y(y)^\intercal$ with $C_Y(y) = K_Y(K_Y K_Y^\intercal)^\dagger K_Y(y)$.

Proof. By Definition 5.1, we have

$$\mathrm{cov}[f(X), g(X)|Y = y] = K_Y(y)^\intercal [E'_{X|Y}][fg] - [f]^\intercal [E'_{X|Y}]^\intercal K_Y(y) K_Y^\intercal(y) [E'_{X|Y}][g]. \quad (5.14)$$

By Lemma 5.3 and 5.4, the first term on the right is

$$\begin{aligned} K_Y^\intercal(y)(K_Y K_Y^\intercal)^\dagger (K_Y K_X^\intercal)(K_X K_X^\intercal)^\dagger K_X (K_X^\intercal[f] \odot K_X^\intercal[g]) &= K_Y^\intercal(y)(K_Y K_Y^\intercal)^\dagger K_Y (K_X^\intercal[f] \odot K_X^\intercal[g]) \\ &= [f]^\intercal K_X \mathrm{diag}[K_Y^\intercal (K_Y K_Y^\intercal)^\dagger K_Y(y)] K_X^\intercal[g] = [f]^\intercal K_X \mathrm{diag}(C_Y(y)) K_X^\intercal[g]. \end{aligned} \quad (5.15)$$

The second term is

$$[f]^\intercal K_X K_Y^\intercal (K_Y K_Y^\intercal)^\dagger K_Y(y) K_Y^\intercal(y)(K_Y K_Y^\intercal)^\dagger K_Y K_X^\intercal [g] = [f]^\intercal K_X K_Y(y) K_Y^\intercal(y) K_X^\intercal [g]. \quad (5.16)$$

Substitute (5.15) and (5.16) into (5.14) to complete the proof. □

We are interested in heteroscedastic variance operator, which is defined on centered $L_2(P_{n,X})$. Let $QK_X(x)$ be the bases of $L_2(P_{n,X})$, and let $[V_{X|Y}(y)]$ be the matrix representation of the heteroscedastic variance operator for each $y \in \Omega_Y$. Then the lemma below gives the the expression of $[V_{X|Y}(y)]$.

Lemma 5.6 *Let $\Lambda(y)$ and Q as defined before, for each $y \in \Omega_Y$. Then we have*

$$[V_{X|Y}(y)] = (K_X Q K_X^\intercal)^\dagger K_X Q \Lambda(y) Q K_X^\intercal.$$

Proof. For $f, g \in L_2(P_{n,X})$,

$$\begin{aligned} \langle g, V_{X|Y}(y) f \rangle_{L_2(P_{n,X})} &= \operatorname{cov}[f(X), g(X)|Y = y] \\ &= [f]^\intercal K_X Q \Lambda(y) Q K_X^\intercal [g], \end{aligned}$$

which implies $[f]^\intercal (K_X Q K_X^\intercal / n)[V_{X|Y}(y)][g] = [f]^\intercal K_X Q \Lambda(y) Q K_X^\intercal [g]$. Since this is true for all $[f], [g] \in \mathbb{R}^{n+1}$, the following relation also holds,

$$(K_X Q K_X / n)[V_{X|Y}(y)] = K_X Q \Lambda(y) Q K_X^\intercal.$$

□

Mimicking the procedure in deriving $[V_{X|Y}(y)]$, we can show the unconditional variance operator V has the representation as

$$[V] = (K_X Q K_X^\intercal / n)^\dagger (K_X Q K_X^\intercal / n). \quad (5.17)$$

Lemma 5.7 *Given finite sample, the unconditional variance operator has a matrix expression in (5.17).*

PROOF. This is an immediate result by the following relation

$$\langle f, Vg\rangle_{L_2(P_{n,X})} = [f]^\intercal (K_X Q K_X^\intercal/n)[V][g] = [f]^\intercal (K_X Q K_X^\intercal/n)[g] = \mathrm{cov}_n[f(X), g(X)].$$

□

We are now ready to write down the sample version of GSAVE operator S_n. By definition

$$\langle f, S_n f\rangle_{L_2(P_{n,X})} = E_n\left(\langle f, [V - V_{X|Y}(Y)]^2 f\rangle_{L_2(P_{n,X})}\right)$$

It follows that

$$\langle f, S_n f\rangle_{L_2(P_{n,X})} = E_n\left([f]^\intercal K_X Q(Q/n - \Lambda(y))Q(Q/n - \Lambda(y))QK_X^\intercal[f]\right).$$

Maximizing this subject to $[f]^\intercal(K_X Q K_X^\intercal/n)[f] = 1$ yields the following standard eigenvalue problem. First, compute the first d eigenvectors, say $\phi_1, \ldots, \phi_d$, of the matrix

$$(K_X Q K_X^\intercal)^{\dagger\frac{1}{2}} E_n\left(K_X Q(Q/n - \Lambda(y))Q(Q/n - \Lambda(y))QK_X^\intercal\right)(K_X Q K_X^\intercal)^{\dagger\frac{1}{2}}$$

Then use the functions

$$K_X^\intercal(x)(K_X Q K_X^\intercal)^{\dagger\frac{1}{2}}\phi_i, \quad i = 1, \ldots, d. \tag{5.18}$$

as the GSAVE sufficient predictors. To enhance performance, we use the regularized version $(K_X Q K_X^\intercal + \epsilon_X I_n)^{-1/2}$ in place of the $(K_X Q K_X^\intercal)^\dagger$ in (5.18) and $(K_Y Q K_Y^\intercal + \epsilon_Y I_n)^{-1}$ in place of the $(K_Y K_Y^\intercal)^\dagger$ in (5.13), where ϵ_X and ϵ_Y are some positive numbers.

Like GSIR, the estimation of GSAVE depends on the values of parameters. We follow the tuning procedure introduced in Section 4.4 to select the optimal values of $\gamma_X, \gamma_Y, \epsilon_X$, and ϵ_Y

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.

Chapter 6

Simulation Study

In this chapter we conduct numerical studies and compare the performance of GSIR and GSAVE with two nonlinear SDR methods, KSIR and KCCA. We design a variety of experiments to demonstrate the robustness of our methods and algorithms; in particular, we test all the methods and models under three predictor scenarios:

1. Gaussian with independent structure,
2. Mixture of Gaussians, and
3. Gaussian with correlated structure

In the first portion of the simulations, we compare GSIR with KSIR and KCCA in settings where the sufficient predictor appears in the conditional mean. Secondly, we compare GSAVE with GSIR, KSIR, and KCCA in settings where the sufficient predictor appears in the conditional variance.

6.1 When Central Class Depends on Conditional Mean

We use the following three models to compare GSIR, KSIR, and KCCA:

$$\begin{aligned}
\text{Model I}: &\quad Y = (X_1^2 + X_2^2)^{1/2} \log(X_1^2 + X_2^2)^{1/2} + \varepsilon, \\
\text{Model II}: &\quad Y = X_1/(1 + e^{X_2}) + \varepsilon, \\
\text{Model III}: &\quad Y = \sin(\pi(X_1 + X_2)/10) + \varepsilon.
\end{aligned}$$

where $X \perp\!\!\!\perp \varepsilon$, $\varepsilon \sim N(0, 0.25)$. The dimension p of X is taken to be 10. X is generated under three scenarios:

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

The central σ-fields for the three models are generated by their conditional means $E(Y|X)$. Since we are only interested in a monotone function of the predictor, we use Spearman's correlation with the true predictor to measure the performance of each estimator. For each model, $n = 400$ observations on (X, Y) are generated. We then divide the whole sample into two parts - 50% for the training and 50% for validation. The first predictor is computed using all three methods based on the training data. Spearman's correlations between estimated and true predictors are then computed on the testing data. This process is repeated $N = 200$ times. In Table 6.1 we list means and standard deviations of Spearman's correlations computed using the $B = 200$ simulated samples.

Note the kernels used in all these comparisons are the Gaussian radial basis kernel. Also, there are four parameters to be determined in each experiment, γ_X, γ_Y, ϵ_X and ϵ_Y. For each model, we tune the values of parameters at one time using our proposed selection procedure (Section 4.4); we then use the same values in all three methods. In other words, the parameter selection is rather model- than method-oriented.

We provide a step-by-step algorithm of how the simulation is carried out.

pseudo code

0. The following model is being tested

$$Y = f(X) + \epsilon,$$

where X is generated based on difference scenarios.

1. simulate $(X_1, \cdots, X_n) \sim \mathcal{N}(0, I_p)$

2. simulate $(Y_1, \cdots, Y_n)$ using the above model

3. given $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, determine the values of $(\gamma_X, \gamma_Y, \epsilon_X, \epsilon_Y)$.

4. divide $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ into two sets: $\{(X_1, Y_1), \cdot, (X_m, Y_m)\}$ for training; $\{(X_{m+1}, Y_{m+1}), \cdot, (X_n, Y_n)\}$ for validation; $m < n$.

5. given training sample $\{(X_1, Y_1), \cdot, (X_m, Y_m)\}$ and $(\gamma_X, \gamma_Y, \epsilon_X, \epsilon_Y)$ in 3, use GSIR, KSIR and KCCA to compute the first eigen function f_1.

6. calculate $(f_1(X_{m+1}), \cdots, f_1(X_n))$.

7. compute Spearman's correlation between $(f_1(X_{m+1}), \cdots, f_1(X_n))$ and the true predictor $(f(X_{m+1}), \cdots, f(X_n))$

8. repeat 1,2,4-7 B times.

Table 6.1: Comparison of KSIR, KCCA, and GSIR for models I–III, where sufficient predictors appear in the conditional means

Scenario	Model	Method		
		KSIR	KCCA	GSIR
A	I	0.777 (0.052)	0.813 (0.044)	0.795 (0.047)
	II	0.810 (0.049)	0.895 (0.032)	0.909 (0.026)
	III	0.765 (0.065)	0.891 (0.035)	0.908 (0.030)
B	I	0.882 (0.025)	0.880 (0.022)	0.868 (0.024)
	II	0.887 (0.030)	0.929 (0.020)	0.929 (0.019)
	III	0.904 (0.025)	0.973 (0.010)	0.974 (0.009)
C	I	0.792 (0.041)	0.816 (0.036)	0.806 (0.037)
	II	0.829 (0.049)	0.860 (0.058)	0.879 (0.045)
	III	0.831 (0.055)	0.958 (0.017)	0.960 (0.017)

From the table we see that the performances of KCCA and GSIR are similar, and both are slightly better than KSIR. In addition, GSIR is robust to varied predictor settings, with the absolute Spearman's correlations around .8 or higher. GSIR differs from KSIR in the way how the response is smoothed; KSIR slices the response, similar to the effect when adopting a Uniform kernel, while GSIR considers a Gaussian kernel. For this reason we expect the discrepancy between two methods to be more apparent when the data has multi-dimensional response.

6.2 When Central Class Depends on Conditional Variance

Next, we compare GSAVE, KSIR, KCCA, and GSIR using the following models:

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

where $X \perp\!\!\!\perp \varepsilon$, $\varepsilon \sim N(0, 0.25)$. Again, the predictor X varies among three different settings, Independent gaussians, mixture of gaussians and correlated gaussians. The specifications of

of n, m, N, p are the same as for Models I – III. Note that here sufficient predictors appear in the conditional variance var$(Y|X)$, rather than the conditional mean as in Models I – III. Means and standard deviations of Spearman's correlations are listed in Table 6.2.

pseudo code

0. The following model is being tested

$$Y = f(X)\epsilon,$$

where X is generated based on difference scenarios.

1. simulate $(X_1, \cdots, X_n) \sim \mathcal{N}(0, I_p)$

2. simulate $(Y_1, \cdots, Y_n)$ using the above model

3. given $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, determine the values of $(\gamma_X, \gamma_Y, \epsilon_X, \epsilon_Y)$.

4. divide $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ into two sets: $\{(X_1, Y_1), \cdot, (X_m, Y_m)\}$ for training; $\{(X_{m+1}, Y_{m+1}), \cdot, (X_n, Y_n)\}$ for validation; $m < n$.

5. given training sample $\{(X_1, Y_1), \cdot, (X_m, Y_m)\}$ and $(\gamma_X, \gamma_Y, \epsilon_X, \epsilon_Y)$ in 3, use GSAVE, GSIR, KSIR and KCCA to compute the first eigen function f_1.

6. calculate $(f_1(X_{m+1}), \cdots, f_1(X_n))$.

7. compute Spearman's correlation between $(f_1(X_{m+1}), \cdots, f_1(X_n))$ and the true predictor $(f(X_{m+1}), \cdots, f(X_n))$

8. repeat 1,2,4-7 B times.

Table 6.2: Comparison of KSIR, KCCA, GSIR, and GSAVE for Models IV – VI, where sufficient predictors appear in conditional variances

Scenario	Model	Method			
		GSAVE	KSIR	KCCA	GSIR
A	IV	0.886 (0.079)	0.101 (0.070)	0.357 (0.221)	0.414 (0.233)
	V	0.728 (0.190)	0.091(0.071)	0.168 (0.126)	0.196 (0.144)
	VI	0.841 (0.087)	0.099 (0.080)	0.250 (0.166)	0.266 (0.174)
B	IV	0.869 (0.077)	0.103 (0.073)	0.434 (0.246)	0.534 (0.251)
	V	0.878 (0.064)	0.088 (0.066)	0.106 (0.078)	0.114 (0.083)
	VI	0.756 (0.149)	0.268 (0.110)	0.614 (0.129)	0.636 (0.129)
C	IV	0.763 (0.205)	0.110 (0.071)	0.230 (0.163)	0.263 (0.183)
	V	0.824 (0.145)	0.095(0.071)	0.112 (0.088)	0.118 (0.092)
	VI	0.728 (0.147)	0.151 (0.103)	0.414 (0.167)	0.442 (0.171)

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 $E[\text{var}(f(X)|Y)]$, whereas GSAVE extracts more information from $\text{var}(f(X)|Y)$.

Note we consider a data-based tuning process for parameters $\gamma_X, \gamma_Y, \epsilon_X$, and ϵ_Y, where γ_X and γ_Y are the width parameters for the kernel matrices K_X and K_Y, respectively; ϵ_X and ϵ_Y are ridge parameters. For each model and each predictor setting, the optimal values of the parameters are listed in Table 6.3.

Table 6.3: Comparison of KSIR, KCCA, GSIR, and GSAVE for Models IV – VI, where sufficient predictors appear in conditional variances

Scenario	Model	Parameters			
		γ_X	γ_Y	ϵ_X	ϵ_Y
A	I	0.036	0.134	0.1	0.01
	II	0.033	0.354	0.1	0.01
	III	0.024	0.549	0.1	0.01
	IV	0.008	0.783	0.1	0.01
	V	0.020	130.67	0.1	0.01
	VI	0.012	2.65	0.1	0.01
B	I	0.040	0.123	0.1	0.01
	II	0.042	0.225	0.1	0.01
	III	0.012	0.335	0.1	0.01
	IV	0.004	0.339	0.1	0.01
	V	0.041	12.96	0.1	0.01
	VI	0.007	1.305	0.1	0.01
C	I	0.085	0.270	0.1	0.01
	II	0.083	0.467	0.1	0.01
	III	0.021	0.434	0.1	0.01
	IV	0.017	0.725	0.1	0.01
	V	0.033	109.29	0.1	0.01
	VI	0.017	2.085	0.1	0.01

Chapter 7

Applicaitons

7.1 Face Data

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. For each data point, 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 pick up 10 images and line up in Fig 7.1, which provides a better look of the data.

We use this data to demonstrate that the first 3 sufficient predictors extracted from KCCA and GSIR can effectively capture the 3-variate response. We use $n = 558$ of the images (roughly 80%) as training data, and the remaining $m = 140$ images as testing data. For each method, we compute 3 predictor functions from the training data, and

Figure 7.1: Face data

evaluate them on the testing data. The upper panel of Figure 7.2 is the perspective plot of the first 3 KCCA predictors evaluated on the 140 testing images, and the lower 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 close by regions in the 3-D cubes correspond to similar patterns of left-right rotation, up-down rotation, lighting direction, and different regions in the cubes correspond to different patterns, which reflects a good correspondence between the 3 sufficient predictors and the 3 responses.

7.2 USPS Handwritten Digits Data

Next, we apply KSIR, KCCA, and GSIR to the *handwritten digits data*, available at `http://www.cs.nyu.edu/~roweis/data.html` – the aim is again to show that all these methods provide good performance. This data set contains 2000 images of $p = 16 \times 16$ 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 the testing data. As for the faces data, for each of the methods we compute the first 3 sufficient predictors on the training data, and evaluate them on the testing data. Results are presented in the three perspective plots in Figure 7.3 – 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.

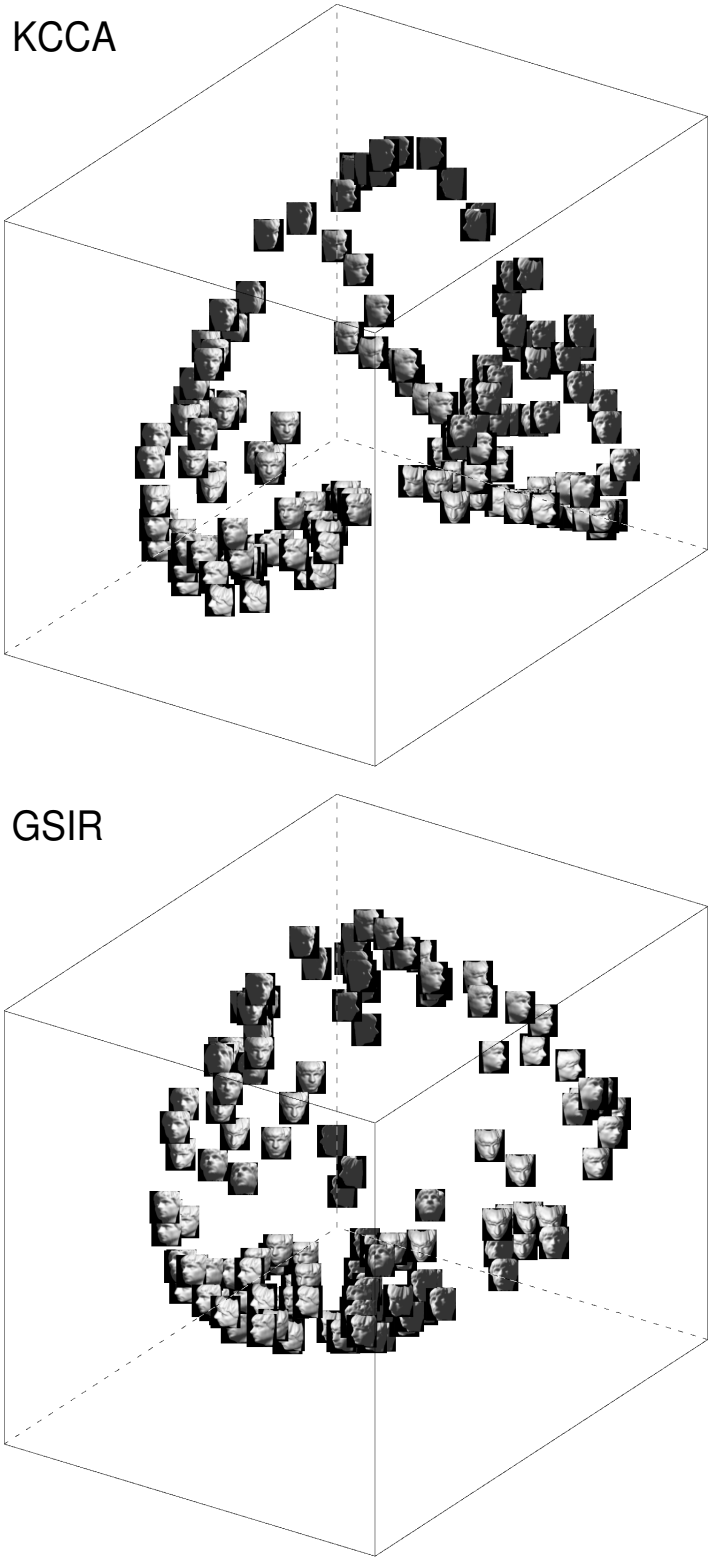

Figure 7.2: First 3 sufficient predictors by KCCA (upper panel) and GSIR (lower panel), computed from 558 training images, and evaluated on 140 testing images faces data.

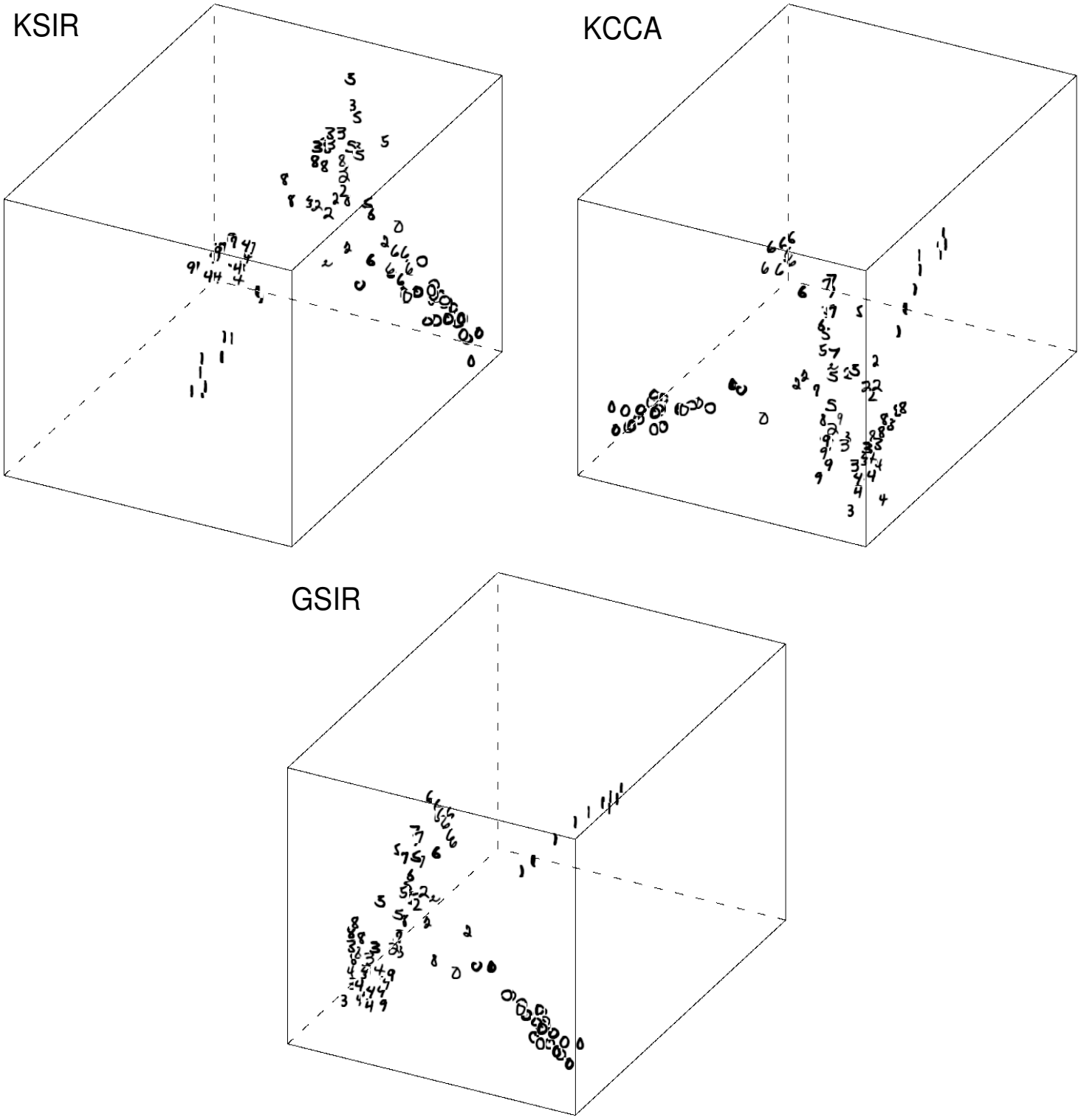

Figure 7.3: First 3 sufficient predictors by KSIR (upper-left panel), KCCA (upper-right panel) and GSIR (lower panel), computed on 1000 training images, and evaluated on 1000 testing images – handwritten digits data.

Chapter 8

Conclusion and Future Work

Both data analyses and simulation have suggested that GSIR, and GSAVE have comparable performance than other kernel-based nonlinear SDR methods recently proposed in the literature - in particular, GSAVE are obviously superior than other methods when the the dependence between predictor and the response is through the conditional variance. Furthermore, we demonstrate that both GSIR and GSAVE are applicable to an extensive predictor settings - predictors with moderately high dimensionality as well as large p small n problems. We also apply GSIR on dataset with categorical response.

In addition, the general framework and theoretical formulation we propose assembles much of the work in SDR in the literature; in addition under this framework we relax some of the restriction in existing methods and provide new insight on the context of nonlinear SDR. Most importantly, this novel framework subsumes the notion of sufficient dimension reduction, reproducing kernel Hilbert space, and classical statistical inference such as sufficiency, completeness and minimal sufficiency. Meanwhile, it also shows that how powerful these fundamental concepts can be useful to contemporary data analysis.

In the following we provide some potential research topics in the future.

Other Second-Order Methods The constructions of GSIR and GSAVE illuminated us that many existing estimators for linear sufficient dimension reduction can be generalized in the similar way. Our next step is to extend other linear methods, such as Directional

Regression (Li and Wang, 2007) or Inverse Regression Estimation (Cook and Ni, 2005) to improve the performance of GSIR and GSAVE in nonlinear dimension reduction.

DR considers the following inverse regression problem:

$$A(Y, \widetilde{Y}) = E[(Z - \widetilde{Z})(Z - \widetilde{Z})^{\top} | Y, \widetilde{Y}], \tag{8.1}$$

where $(\widetilde{Y}, \widetilde{Z})$ is an independent copy of (Y, Z). $A(Y, \tilde{Y})$ is conditional expectation of the square of empirical directions $Z - \tilde{Z}$ given Y. An extension of (8.1) to the nonlinear setting is to consider the empirical directions in Hilbert space; that is, for any $f \in \mathcal{H}$ where $\mathcal{H}$ is a Hilbert space, the empirical direction is $f(X) - f(\tilde{X})$. We then define an linear operator from $\mathcal{H}$ to $\mathcal{H}$, which is a generalization of $A(Y, \tilde{Y})$ in (8.1):

$$(f, g) \mapsto E\left[(f(X) - f(\tilde{X})))(g(X) - g(\tilde{X}) | Y, \tilde{Y}\right], \tag{8.2}$$

for any $f, g \in \mathcal{H}$

DR serves as a remedy tool to SAVE, as it provides higher efficiency than SAVE; analogously, the operator defined in (8.2) should potentially provide better estimation for the central class than GSAVE. Much of the work is under construction.

Consistency of GSIR and GSAVE The covariance operator plays a critical role in our development; not only does it much simplify the theoretical derivations, but provide efficient calculation when using it in the estimating procedure. Fukumizu, Bach and Gretton (2007) have shown its consistency, which is based on the *Hilbert-Schmidt* norm of operators.

Proposition 8.1 *Let Σ_{YX} be the covariance operator from $\mathcal{H}_X$ to $\mathcal{H}_Y$ where $\mathcal{H}_X$ and $\mathcal{H}_Y$ are both RKHSsU, and let $\widehat{\Sigma_{YX}}$ be its empirical estimate. Then with some regularities, we have*

$$||\widehat{\Sigma_{YX}} - \Sigma_{YX}||_{\text{HS}} = O_p(n^{-\frac{1}{2}}) \quad (n \to \infty) \tag{8.3}$$

where $||\cdot||_{\text{HS}}$ is the Hilbert-Schmidt norm.

The sample estimate $\widehat{\Sigma_{YX}}$ of finite rank converges to its population counterpart defined in RKHS, which can be infinite-dimensional. Our proposed forms of GSIR and GSAVE

are based on inverse conditional mean operators, which can be represented as covariance operators. Therefore, using the result in the above we can provide the convergence of GSIR and GSAVE. Much is left for future work.

Bibliography

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

2. Bach, F. R. and Jordan. M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, **3**, 1-48.

3. Bahadur, R. R. (1954). Sufficiency and Statistical Decision Functions. *The Annals of Mathematical Statistics*, **25**, 3, 423-462.

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

5. Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimension reduction and data representation. *Neural Computation*, **15**, 1373-1396.

6. Belkin, M. and Niyogi, P. (2008). Towards a theoretical foundation for Laplacian-based manifold methods. *Journal of Computer and System Sciences*, **74**, 1289-1308.

7. Berlinet, A. and Thomas-Agnan C. (2004). Reproducing Kernel Hilbert Spaces in Probability and Statistics. *Springer*.

8. Chen, C.-H. and Li, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica*, **8**, 289316

9. Chiaromonte, F. and Cook, R. D. (2002) Sufficient Dimension Reduction and Graphics in Regression. *Ann. Inst. Statist. Math* **54** 4, 768-795

10. Conway, J. (1990). *A Course in Functional Analysis*. Second edition. Springer.

11. 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*, Alexandria, VA: American Statistical Association, 18–25.

12. Cook, R. D. (1998). Principle Hessian Directions Revisited. *Journal of American Statistical Association*, **93**, 441, 84-94.

13. Cook, R. D. (1998). Regression Graphics: Ideas for Studying Regressions through Graphics. New York: Wiley.

14. Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, **22**, 1–40.

15. Cook, R. D. and Forzani, L. (2009). Likelihood-Based Sufficient Dimension Reduction. *Journal of the American Statistical Association*, **104**, 485, 197-208.

16. Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *The Annals of Statistics*, 30, 455–474.

17. Cook, R. D. and Li, B. (2004). Determining the Dimension of Iterative Hessian Transformation. *The Annuals of Statistics*, **32**, 6, 2501-2531.

18. Cook, R.D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, **100**, 410–428.

19. Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association.* **86**, 316–342.

20. Cristianini, N. and Shawe-Taylor, John (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

21. Eaton, M. L. (1986). A characterization of spherical distributions. *Journal of Multivariate Analysis*, **34**, 439–446.

22. Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Applications. Chapman and Hall, London.

23. Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Toy. Soc. London Ser. A*, **222**, 109-368.

24. Fukumizu, K., Bach, F. R. and Jordan M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *The Journal of Machine Learning Research*, **5**, 73–99.

25. Fukumizu, K., Bach, F. R. and Gretton, A. (2007). Statistical Consistency of Kernel Canonical Correlation Analysis. *The Journal of Machine Learning Research*, **8**, 361-383.

26. Fukumizu, K., Bach, F. R. and Jordan M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, **4** 1871 – 1905.

27. Fung, K. F., He, X., Liu, L. and Shi, P. (2002). Dimension reduction based on canonical correlation. *Statistica Sinica*, **12**, 1093–1113.

28. Halmos, P. R. and Savage, L. J. (1949). Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics. *The Annals of Mathematical Statistics*, **20**, 2, 225-241.

29. Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal Smoothing in Single-Index Models. *Annals of Statistics*, **21(1)**,157-178.

30. Härdle, W. and Stoker, T. M. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of the American Statistical Association*, **84**, 408, 986- 995.

31. Härdle, W. and Tsybakiv, A. B. (1991). Sliced Inverse Regression for Dimension Reduction: Comment. *Journal of the American Statistical Association*, **86**, 414, 333-335.

32. Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning. second edition, Springer.

33. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417-441.

34. Hristache, M., Juditsky, A., Polzehi, J. and Spokoiny, V. (2001). Structure Adaptive Approach for Dimension Reduction. *The Annals of Statistics*, **29**, 6, 1537-1566

35. Huang, S.-Y., Lee, M.-H. and Hsiao, C. K. (2009). Nonlinear measures of association with kernel canonical correlation analysis and applications. *Journal of Statistical Planning and Inference*, **139**, 2162 2174.

36. Hsing, T. (1999) Nearest neighbor inverse regression. *Ann. Statist.*, **27**, 2, 697-731.

37. Hsing, T. and Ren, H. (2009). An RKHS formulation of the inverse regression dimension-reduction problem. *The Annals of Statistics*, **37** 726–755.

38. Hofmann, T., Schölkopf, B. and Smola, A. J. (2008). Kernel Methods in Machine Learning. *The Annuals of Statistics*, **36**, 3, 1171-1220.

39. Ichimura, H. and L. Lee, (1991). Semiparametric least squares estimation of multiple index models: single equation estimation. W.A. Barnett, J.L. Powell, and G. Tauchen, eds., *Nonparametric and semiparametric methods in econometrics and statistics*, Cambridge: Cambridge University Press.

40. Kim, M. and Pavlovic, V. (2010). Central subspace dimension reduction using covariance operators *IEEE Transactions on Pattern Analysis and Machine Intelligence* **99**.

41. Lehmann, E. L. (1981). An interpretation of completeness and Basus theorem. *Journal of the American Statistical Association*, **76**, 335-340.

42. Li, B., Artemiou, A. and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Annals of Statistics*. To appear.

43. Li, B., Chun, H. and Zhao, H. (2011). Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*. To appear.

44. Li, B. and Dong, Y. (2009). Dimension reduction for non-elliptically distributed predictors. *The Annals of Statistics*, **37**, 1272–1298.

45. Li, L., Li, B. and Zhu, L.-X. (2010). Groupwise Dimension Reduction. *Journal of the American Statistical Association*, **105**, 491, 1188-1201.

46. Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, **102**, 997–1008.

47. Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, **33**, 1580-1616.

48. Li, K. -C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316 – 342.

49. Li, K. -C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association*, **86**, 316 – 342.

50. Li, K.-C., Aragon, Y., Shedden, K. and Agnan C. T. (2003) Dimension Reduction for Multivariate Response Data. *Journal of the American Statistical Association*, **98**, 461, 99- 109.

51. Li, K. -C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, **17**, 1009–1052.

52. Lue, H. H. (2010). On principal Hessian directions for multivariate response regressions. *Computational Statistics*, **25**, pp.619-632.

53. Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London*, **209**.

54. Minh, H. Q., Niyogi, P. and Yao, Y. (2006), Mercer's Theorem, Feature Maps, and Smoothing. *Proc. of Computational Learning Theory*

55. Setodji, C. M. and Cook, R. D. (2004) K-Means Inverse Regression. *Technometrics*, **46**, 4, 421-429.

56. Scholkopf, B., Smola, A. J. and Muller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 12991319.

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

58. Wu, Q., Liang, F., and Mukherjee, S. (2008). Regularized sliced inverse regression for kernel models. Technical report, Duke University.

59. Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of optimal regression subspace. *Journal of the Royal Statistical Society, Series B.*, **64**, 363–410.

60. Ye, Z., and Weiss, R. E. (2003). Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods. *Journal of the American Statistical Association*, **98**, 968979.

61. 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.

62. Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.*, **35**, 6, 2654-2690.

63. Xia, Y., Li, B. and Cook, R. D. (2008) Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, **99**, 8, 1733-175.

64. Zhou, J. and He, X. (2008) Dimension Reduction Based on Constrained Canonical Correlation and Variable Filtering. *The Annuals of Statistics*, **36**, 4, 1649-1668.

Vita

Kuang-Yao Lee

Kuang-Yao Lee was born in December 1979 in Taipei, Taiwan. He received the BS and MS degrees from the Department of Mathematics at Taiwan University, in 2002 and 2005. He is currently pursuing his Ph.D. in statistics at Pennsylvania State University with an expected graduation in 2012. After graduation, he will work as a Postdoctoral Associate in the Biostatistics Division at Yale School of Public Health School.