A General Theory for Nonlinear Sufficient Dimension Reduction: Formulation and Estimation

K.Y. Lee
B. Li
F. Chiaromonte

May 2, 2014

Outline

1. σ-Field as the condition for DR;
2. Sufficiency and minimal sufficiency;
3. Unbiasedness and exhaustiveness;
4. Population criterion;

Set-up

$(\Omega, \mathcal{F}, P)$ be a probability space,

$$\Rightarrow \quad (\Omega_X, \mathcal{F}_X, P_X)$$
$$(\Omega_Y, \mathcal{F}_Y, P_Y)$$
$$(\Omega_{XY}, \mathcal{F}_{XY}, P_{XY}) \text{ where } \Omega_{XY} = \Omega_X \times \Omega_Y, \mathcal{F}_{XY} = \mathcal{F}_X \times \mathcal{F}_Y$$

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

SDR σ-Field

Let $\mathcal{G} \subseteq \sigma(X)$ be a sub σ-field, if

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

then $\mathcal{G}$ is called a **SDR σ-field** for Y vs X.

Remark: $\mathcal{G}$ can be induced by some rv, say U, with the measurable space $(U, \mathcal{F}_U)$, i.e.

$$\mathcal{G} = U^{-1}(\mathcal{F}_U).$$

Then, SDR is achieved by using U, which is a transformation of X. Since the transformation is not necessarily to be linear, non-linear SDR can be achieved.

σ-Field SDR as a General Framework

Example 1: Let

$$\Omega_X = \mathbb{R}^P, \Omega_Y = \mathbb{R}^q$$
$$\mathcal{F}_X, \mathcal{F}_Y, \mathcal{F}_{XY} \text{ are Borel } \sigma\text{-fields}$$

If $U = B^T X$ and $\mathcal{G} = \sigma(U)$, then we have the usual linear DR

$$Y \perp\!\!\!\perp X | B^T X.$$

Example 2: Let λ be the Lebesgue on $[a, b]$ and

$$\Omega_X = L^2_\lambda, \Omega_Y = \mathbb{R}.$$

If $\{h_1, \ldots, h_d\} \subset L^2_\lambda$ and $U = \left(\langle X, h_1\rangle_{L^2_\lambda}, \cdots, \langle X, h_d\rangle_{L^2_\lambda}\right)$, then we have the functional DR problem considered by [Ferre and Yao 2003]

$$Y \perp\!\!\!\perp X | \langle X, h_1\rangle_{L^2_\lambda}, \cdots, \langle X, h_d\rangle_{L^2_\lambda}.$$

Remarks: Generalize SDR to the infinite-dimensional case, but still linear in X.

Minimal Sufficiency

$\mathcal{G}$ is not unique, for example, $\mathcal{G} = \sigma(X)$ is valid but not reduction. We want to find the smallest $\mathcal{G}$.

The following Theorem shows existence and uniqueness of the minimal sufficient $\mathcal{G}$.

Theorem 1

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

(1) $Y \perp\!\!\!\perp X|\mathcal{G}^*$;

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

$\mathcal{G}^*$ $(= \mathcal{G}_{Y|X})$ is called **central $\boldsymbol{\sigma}$-field**.

Adding More Structures

Let $L^2_{P_X}$, $L^2_{P_Y}$, and $L^2_{P_{XY}}$ be the function spaces on Ω_X, Ω_Y and Ω_{XY}. They are all 0 mean functions.

$$\mathcal{M}_{\mathcal{G}} = \left\{ f \in L^2_{P_{XY}} : f \text{ is } \mathcal{G}\text{-measurable} \right\}.$$

Remark: $\mathcal{G}$ is a linear sub-space of $L^2_{P_{XY}}$.

Definitions:

1) If $\mathcal{G}$ is sufficient, then $\mathcal{M}_{\mathcal{G}}$ is called a **SDR class**. $\mathcal{G}^*$ is minimal sufficient (central σ-algebra), then $\mathcal{M}_{\mathcal{G}^*}$ is called the **central class**.
2) If $\mathcal{G} = \sigma(U)$, then we also use $\mathcal{M}_U = \mathcal{M}_{\mathcal{G}}$.

Remark: $\mathcal{M}_{\mathcal{G}^*}$ is the generalization of the usual central space $\mathcal{S}_{Y|X}$ in linear SDR.

Unbiasedness and Exhaustiveness

If $\mathcal{M} \subset L^2_{P_X}$ is a collection of $\mathcal{G}^*$ measurable function, then $\mathcal{M}$ is **unbiased** for $\mathcal{M}_{\mathcal{G}^*}$.

If the members of $\mathcal{M}$ generate $\mathcal{G}^*$, then it is **exhaustive**.

Example:

In linear DR, if B is a DR matrix and $\text{span}(B) \subset \mathcal{S}_{Y|X}$, then it is unbiased.

If $\text{span}(B) = \mathcal{S}_{Y|X}$, it is exhaustive.

Unbiasedness and Exhaustiveness

If $\mathcal{M} \subset L^2_{P_X}$ is a collection of $\mathcal{G}^*$ measurable function, then $\mathcal{M}$ is **unbiased** for $\mathcal{M}_{\mathcal{G}^*}$.

If the members of $\mathcal{M}$ generate $\mathcal{G}^*$, then it is **exhaustive**.

Example:

In linear DR, if B is a DR matrix and $\text{span}(B) \subset \mathcal{S}_{Y|X}$, then it is unbiased.

If $\text{span}(B) = \mathcal{S}_{Y|X}$, it is exhaustive.

For two spaces $\mathcal{S}_1$ and $\mathcal{S}_2$, denote $\mathcal{S}_1 \ominus \mathcal{S}_2 = \mathcal{S}_1 \cap \mathcal{S}_2$.

Theorem 2

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

$$L^2_{P_X} \ominus \left[L^2_{P_X} \ominus L^2_{P_Y}\right] \subseteq \mathcal{M}_{\mathcal{G}^*},$$

i.e. unbiased for $\mathcal{M}_{\mathcal{G}^*}$.

Proof: $\Leftrightarrow L^2_{P_X} \ominus \mathcal{M}_{\mathcal{G}^*} \subseteq L^2_{P_X} \ominus L^2_{P_Y}$

$$f \in L^2_{P_X} \ominus \mathcal{M}_{\mathcal{G}^*} \Rightarrow f \perp \mathcal{M}_{\mathcal{G}^*} \Rightarrow \mathbb{E}\left[f(X)|\mathcal{G}^*\right] = 0$$
$$\Rightarrow \mathbb{E}\left[f(X)|Y\right] = 0 \Rightarrow f \perp \mathcal{M}_Y \Rightarrow f \in L^2_{P_X} \ominus L^2_{P_Y}$$

Remarks:

1) $L^2_{P_X} \ominus L^2_{P_Y}$ resembles the **residual** in a regression.
2) $L^2_{P_X} \ominus \left[L^2_{P_X} \ominus L^2_{P_Y}\right]$ is the orthogonal complement of the residual class, called **regression class**.

For two spaces $\mathcal{S}_1$ and $\mathcal{S}_2$, denote $\mathcal{S}_1 \ominus \mathcal{S}_2 = \mathcal{S}_1 \cap \mathcal{S}_2$.

Theorem 2

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

$$L^2_{P_X} \ominus \left[L^2_{P_X} \ominus L^2_{P_Y}\right] \subseteq \mathcal{M}_{\mathcal{G}^*},$$

i.e. unbiased for $\mathcal{M}_{\mathcal{G}^*}$.

Proof: $\Leftrightarrow L^2_{P_X} \ominus \mathcal{M}_{\mathcal{G}^*} \subseteq L^2_{P_X} \ominus L^2_{P_Y}$

$$f \in L^2_{P_X} \ominus \mathcal{M}_{\mathcal{G}^*} \Rightarrow f \perp \mathcal{M}_{\mathcal{G}^*} \Rightarrow \mathbb{E}\left[f(X)|\mathcal{G}^*\right] = 0$$
$$\Rightarrow \mathbb{E}\left[f(X)|Y\right] = 0 \Rightarrow f \perp \mathcal{M}_Y \Rightarrow f \in L^2_{P_X} \ominus L^2_{P_Y}$$

Remarks:

1) $L^2_{P_X} \ominus L^2_{P_Y}$ resembles the **residual** in a regression.

2) $L^2_{P_X} \ominus \left[L^2_{P_X} \ominus L^2_{P_Y}\right]$ is the orthogonal complement of the residual class, called **regression class**.

For two spaces $\mathcal{S}_1$ and $\mathcal{S}_2$, denote $\mathcal{S}_1 \ominus \mathcal{S}_2 = \mathcal{S}_1 \cap \mathcal{S}_2$.

Theorem 2

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

$$L^2_{P_X} \ominus \left[L^2_{P_X} \ominus L^2_{P_Y}\right] \subseteq \mathcal{M}_{\mathcal{G}^*},$$

i.e. unbiased for $\mathcal{M}_{\mathcal{G}^*}$.

Proof: $\Leftrightarrow L^2_{P_X} \ominus \mathcal{M}_{\mathcal{G}^*} \subseteq L^2_{P_X} \ominus L^2_{P_Y}$

$$f \in L^2_{P_X} \ominus \mathcal{M}_{\mathcal{G}^*} \Rightarrow f \perp \mathcal{M}_{\mathcal{G}^*} \Rightarrow \mathbb{E}\left[f(X)|\mathcal{G}^*\right] = 0$$
$$\Rightarrow \mathbb{E}\left[f(X)|Y\right] = 0 \Rightarrow f \perp \mathcal{M}_Y \Rightarrow f \in L^2_{P_X} \ominus L^2_{P_Y}$$

Remarks:

1) $L^2_{P_X} \ominus L^2_{P_Y}$ resembles the **residual** in a regression.
2) $L^2_{P_X} \ominus \left[L^2_{P_X} \ominus L^2_{P_Y}\right]$ is the orthogonal complement of the residual class, called **regression class**.

Completeness and Exhaustiveness

Definition 5

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

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

Examples:

1) **[Forward regression]** Suppose there exists a function $h \in \left[L^2_{P_X}\right]^q$ such that

$$Y = h(X) + \epsilon,$$

where $\epsilon \perp\!\!\!\perp X$ and $\mathbb{E}\left[\epsilon\right] = 0$. Then $\mathcal{M}_{h(X)}$ is a complete and sufficient dimension reduction class for Y versus X.

Definition 5

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

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

Examples:

1) **[Forward regression]** Suppose there exists a function $h \in \left[L^2_{P_X}\right]^q$ such that

$$Y = h(X) + \epsilon,$$

where $\epsilon \perp\!\!\!\perp X$ and $\mathbb{E}\left[\epsilon\right] = 0$. Then $\mathcal{M}_{h(X)}$ is a complete and sufficient dimension reduction class for Y versus X.

2) **[Inverse regression]** Suppose $q < p$, Ω_Y has a nonempty interior, and P_Y is dominated by the Lebesgue measure on $\mathbb{R}^q$. Suppose there exists functions $g \in \left[L^2_{P_X}\right]^q$ and $h \in \left[L^2_{P_X}\right]^{p-q}$ such that:

1. $g(X) = Y + \epsilon$, where $Y \perp\!\!\!\perp \epsilon$, and $\epsilon \sim N(0, \Sigma)$;
2. $\sigma\left(g(X), h(X)\right) = \sigma(X)$;
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.

When a complete and sufficient dimension reduction class exists, it is unique and coincides with the central class.

Theorem 3

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

$$\mathcal{M}_{\mathcal{G}} = \mathscr{C}_{Y|X} = \mathcal{M}_{\mathcal{G}^*},$$

i.e. it is exhaustive.

Summary of Sufficiency, Completeness, Unbiasedness and Exhaustiveness

With the fairly general assumption of the function spaces, we see that

$$L^2_{P_X} \ominus L^2_{P_Y} \quad \text{(residual class)}$$

plays a crucial role in nonlinear DR. If its **orthogonal complement** in $L^2_{P_X}$ is a complete and sufficient DR class for Y versus X, then it is the **central class**, i.e.

$$L^2_{P_X} \cap \left\{ L^2_{P_X} \ominus L^2_{P_Y} \right\}^{\perp} = \mathcal{M}_{\mathcal{G}^*}.$$

Without completeness, it is still **unbiased**, i.e.

$$L^2_{P_X} \cap \left\{ L^2_{P_X} \ominus L^2_{P_Y} \right\}^{\perp} \subseteq \mathcal{M}_{\mathcal{G}^*}.$$

Characterization of the Regression Class

Definition

For two sets A and B, we say $A \subseteq B$ **modulo constants** if for each $f \in A$ there is $c \in \mathbb{R}$ such that $f + c \in B$.

A is a dense subset of B modulo constants, if (i) $A \subseteq B$ modulo constants and (ii) any $f \in B$ can be approximated by a sequence $\{f_n + c_n\} \subseteq A$.

Remark: Recall the denseness assumption **(AS)** in [Fukumizu, Bach and Jordan 2009].

Examples: Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ with finite variances are dense in $L^2_{P_X}$ and $L^2_{P_Y}$ modulo constants.

Definition

For two sets A and B, we say $A \subseteq B$ **modulo constants** if for each $f \in A$ there is $c \in \mathbb{R}$ such that $f + c \in B$.

A is a dense subset of B modulo constants, if (i) $A \subseteq B$ modulo constants and (ii) any $f \in B$ can be approximated by a sequence $\{f_n + c_n\} \subseteq A$.

Remark: Recall the denseness assumption **(AS)** in [Fukumizu, Bach and Jordan 2009].

Examples: Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ with finite variances are dense in $L^2_{P_X}$ and $L^2_{P_Y}$ modulo constants.

Characterization of the Regression Class

Definition

For two sets A and B, we say $A \subseteq B$ **modulo constants** if for each $f \in A$ there is $c \in \mathbb{R}$ such that $f + c \in B$.

A is a dense subset of B modulo constants, if (i) $A \subseteq B$ modulo constants and (ii) any $f \in B$ can be approximated by a sequence $\{f_n + c_n\} \subseteq A$.

Remark: Recall the denseness assumption **(AS)** in [Fukumizu, Bach and Jordan 2009].

Examples: Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ with finite variances are dense in $L^2_{P_X}$ and $L^2_{P_Y}$ modulo constants.

(Recall) Due to Riesz Representation Theorem, we can define cross-covariance operators for RKHSs.

$$\langle f, \Sigma_{XX}^{RKHS} g \rangle_{\mathcal{H}_X} = \mathsf{cov}(f(X), g(X))_{P_X},$$
$$\langle f, \Sigma_{YY}^{RKHS} g \rangle_{\mathcal{H}_Y} = \mathsf{cov}(f(X), g(X))_{P_Y},$$
$$\langle f, \Sigma_{YX}^{RKHS} g \rangle_{\mathcal{H}_Y} = \mathsf{cov}(f(X), g(X))_{P_{XY}}.$$

They are bounded and self-adjoint.

But in general, $\mathcal{H}_X$ and $\mathcal{H}_Y$ don't have to be RKHSs, we can still have Σ_{XX} and Σ_{YY}.

Let $\mathcal{G}_X = \overline{\text{Range}(\Sigma_{XX})}$, and $\mathcal{G}_Y = \overline{\text{Range}(\Sigma_{YY})}$. (so $\mathcal{G}_X$ and $\mathcal{H}_Y$ may not be $\subseteq \mathcal{H}_X, \mathcal{H}_Y$, but definitely $\subseteq L^2_{P_X}, L^2_{P_Y}$.)

Under assumptions (A) and (B), we can define (similar to Fukumizu's RKHSs case):

$$\langle f, \Sigma_{XX} g \rangle_{\mathcal{G}_X} := \langle f, g \rangle_{L^2_{P_X}},$$

$$\langle f, \Sigma_{YY} g \rangle_{\mathcal{G}_Y} := \langle f, g \rangle_{L^2_{P_Y}},$$

$$\langle f, \Sigma_{YX} g \rangle_{\mathcal{G}_Y} := \langle f, g \rangle_{L^2_{P_Y}},$$

why not $L^2_{P_{XY}}$?

and we have

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} R_{YX} \Sigma_{XX}^{1/2}$$

Reminder: our central class is a $L^2_{P_X}$ object.

But in general, $\mathcal{H}_X$ and $\mathcal{H}_Y$ don't have to be RKHSs, we can still have Σ_{XX} and Σ_{YY}.

Let $\mathcal{G}_X = \overline{\text{Range}(\Sigma_{XX})}$, and $\mathcal{G}_Y = \overline{\text{Range}(\Sigma_{YY})}$. (so $\mathcal{G}_X$ and $\mathcal{H}_Y$ may not be $\subseteq \mathcal{H}_X, \mathcal{H}_Y$, but definitely $\subseteq L^2_{P_X}, L^2_{P_Y}$.)

Under assumptions (A) and (B), we can define (similar to Fukumizu's RKHSs case):

$$\langle f, \Sigma_{XX} g \rangle_{\mathcal{G}_X} := \langle f, g \rangle_{L^2_{P_X}},$$

$$\langle f, \Sigma_{YY} g \rangle_{\mathcal{G}_Y} := \langle f, g \rangle_{L^2_{P_Y}},$$

$$\langle f, \Sigma_{YX} g \rangle_{\mathcal{G}_Y} := \langle f, g \rangle_{L^2_{P_Y}},$$

why not $L^2_{P_{XY}}$?

and we have

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} R_{YX} \Sigma_{XX}^{1/2}$$

Reminder: our central class is a $L^2_{P_X}$ object.

But in general, $\mathcal{H}_X$ and $\mathcal{H}_Y$ don't have to be RKHSs, we can still have Σ_{XX} and Σ_{YY}.

Let $\mathcal{G}_X = \overline{\text{Range}(\Sigma_{XX})}$, and $\mathcal{G}_Y = \overline{\text{Range}(\Sigma_{YY})}$. (so $\mathcal{G}_X$ and $\mathcal{H}_Y$ may not be $\subseteq \mathcal{H}_X, \mathcal{H}_Y$, but definitely $\subseteq L^2_{P_X}, L^2_{P_Y}$.)

Under assumptions (A) and (B), we can define (similar to Fukumizu's RKHSs case):

$$\langle f, \Sigma_{XX} g \rangle_{\mathcal{G}_X} := \langle f, g \rangle_{L^2_{P_X}},$$

$$\langle f, \Sigma_{YY} g \rangle_{\mathcal{G}_Y} := \langle f, g \rangle_{L^2_{P_Y}},$$

$$\langle f, \Sigma_{YX} g \rangle_{\mathcal{G}_Y} := \langle f, g \rangle_{L^2_{P_Y}},$$

why not $L^2_{P_{XY}}$?

and we have

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} R_{YX} \Sigma_{XX}^{1/2}$$

Reminder: our central class is a $L^2_{P_X}$ object.

Assumptions:

(A) $\mathcal{H}_X$ and $\mathcal{H}_Y$ are dense in $L^2_{P_X}$ and $L^2_{P_Y}$ modulo constants;

(B) There are constants $C_1 > 0$ and $C_2 > 0$ such that $\mathrm{Var}\,(f(X)) \leq C_1 \|f\|_{\mathcal{H}_X}$ and $\mathrm{Var}\,(g(Y)) \leq C_1 \|g\|_{\mathcal{H}_Y}$.

Theorem 4: Extended Covariance Operators

Under assumptions (A) and (B), there exist unique isomorphisms

$$\tilde{\Sigma}^{1/2}_{XX} : L^2_{P_X} \to \mathcal{G}_X, \quad \tilde{\Sigma}^{1/2}_{YY} : L^2_{P_Y} \to \mathcal{G}_Y$$

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

$$\tilde{\Sigma}^{1/2}_{XX}(f - \mathbb{E}\,[f]) = \Sigma^{1/2}_{XX} f, \quad \tilde{\Sigma}^{1/2}_{YY}(f - \mathbb{E}\,[g]) = \Sigma^{1/2}_{YY} g.$$

Furthermore, for any $f \in L^2_{P_X}$, $g \in L^2_{P_Y}$ we have

$$\langle \tilde{\Sigma}^{1/2}_{YY}(g), R_{YX} \tilde{\Sigma}^{1/2}_{XX}(f) \rangle_{\mathcal{G}_Y} = \mathrm{Cov}(f(X), g(Y)).$$

Examples:

1. For $f' \in \mathcal{G}_X$ and $g' \in \mathcal{G}_Y$, let $f = f' - \mathbb{E}[f']$ and $g = g' - \mathbb{E}[g']$. Then

$$\langle \widetilde{\Sigma}_{YY}^{1/2} g, R_{YX} \widetilde{\Sigma}_{XX}^{1/2} f \rangle_{\mathcal{G}_Y} = \textbf{Cov}\left(f(X), g(Y)\right).$$

2. For all $f, g \in L^2_{P_X}$ and $s, t \in L^2_{P_Y}$, we have

$$\langle \widetilde{\Sigma}_{XX}^{1/2} g, \widetilde{\Sigma}_{XX}^{1/2} f \rangle_{\mathcal{G}_X} = \textbf{Cov}\left(f(X), g(X)\right)_{P_X},$$
$$\langle \widetilde{\Sigma}_{YY}^{1/2} s, \widetilde{\Sigma}_{YY}^{1/2} t \rangle_{\mathcal{G}_Y} = \textbf{Cov}\left(s(Y), t(Y)\right)_{P_Y}.$$

Remark: The extended covariance operators will be used to characterize the residual class $L^2_{P_X} \ominus L^2_{P_Y}$.

Examples:

1. For $f' \in \mathcal{G}_X$ and $g' \in \mathcal{G}_Y$, let $f = f' - \mathbb{E}[f']$ and $g = g' - \mathbb{E}[g']$. Then

$$\langle \widetilde{\Sigma}_{YY}^{1/2} g, R_{YX} \widetilde{\Sigma}_{XX}^{1/2} f \rangle_{\mathcal{G}_Y} = \mathbf{Cov}\left(f(X), g(Y)\right).$$

2. For all $f, g \in L^2_{P_X}$ and $s, t \in L^2_{P_Y}$, we have

$$\langle \widetilde{\Sigma}_{XX}^{1/2} g, \widetilde{\Sigma}_{XX}^{1/2} f \rangle_{\mathcal{G}_X} = \mathbf{Cov}\left(f(X), g(X)\right)_{P_X},$$
$$\langle \widetilde{\Sigma}_{YY}^{1/2} s, \widetilde{\Sigma}_{YY}^{1/2} t \rangle_{\mathcal{G}_Y} = \mathbf{Cov}\left(s(Y), t(Y)\right)_{P_Y}.$$

Remark: The extended covariance operators will be used to characterize the residual class $L^2_{P_X} \ominus L^2_{P_Y}$.

Conditional Expectation Operators

Define

$$\mathbb{E}_{X|Y} := \widetilde{\Sigma}_{YY}^{-1/2} R_{YX} \tilde{\Sigma}_{XX}^{1/2}$$
$$L^2_{P_X} \to L^2_{P_Y}$$

Proposition 3

Under conditions (A) and (B), we have:

(1) $\forall f \in L^2_{P_X}$, $\mathbb{E}_{X|Y} f = \mathbb{E}\left[f(X)|Y\right]$;

(2) $\forall g \in L^2_{P_Y}$, $\mathbb{E}^*_{X|Y} f = \mathbb{E}\left[g(Y)|X\right]$.

- **SIR** finds the DR directions by applying PCA on
$$[\mathsf{Var}(X)]^{-1}\,\mathsf{Var}(\mathbb{E}[X|Y]).$$

- We want to use operators to define the variance of the conditional expectation in functional spaces.

Corollary 1

Under conditions (A) and (B), $\forall f, g \in L^2_{P_X}$,
$$\langle g, \mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y} f\rangle_{L^2_{P_X}} = \mathbf{Cov}(\mathbb{E}[g(X)|Y], \mathbb{E}[f(X)|Y]).$$

Moreover, $\mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y}$ is a bounded linear operator on $L^2_{P_X}$, and $\|\mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y}\| \leq 1$.

- $\langle f, \mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y} f\rangle_{L^2_{P_X}}$ generalizes $\mathsf{Var}(\mathbb{E}[X|Y])$.

GSIR

- **SIR** finds the DR directions by applying PCA on
$$[\mathrm{Var}\,(X)]^{-1}\,\mathrm{Var}\,(\mathbb{E}\,[X|Y])\,.$$

- We want to use operators to define the variance of the conditional expectation in functional spaces.

Corollary 1

Under conditions (A) and (B), $\forall f, g \in L^2_{P_X}$,
$$\langle g, \mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y} f\rangle_{L^2_{P_X}} = \mathbf{Cov}\,(\mathbb{E}\,[g(X)|Y]\,, \mathbb{E}\,[f(X)|Y])\,.$$
Moreover, $\mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y}$ is a bounded linear operator on $L^2_{P_X}$, and $\|\mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y}\| \leq 1$.

- $\langle f, \mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y} f\rangle_{L^2_{P_X}}$ generalizes $\mathrm{Var}\,(\mathbb{E}\,[X|Y])$.

GSIR

- **SIR** finds the DR directions by applying PCA on
$$[\mathrm{Var}(X)]^{-1}\,\mathrm{Var}(\mathbb{E}[X|Y]).$$

- We want to use operators to define the variance of the conditional expectation in functional spaces.

Corollary 1

Under conditions (A) and (B), $\forall f, g \in L^2_{P_X}$,
$$\langle g, \mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y} f\rangle_{L^2_{P_X}} = \mathbf{Cov}(\mathbb{E}[g(X)|Y], \mathbb{E}[f(X)|Y]).$$
Moreover, $\mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y}$ is a bounded linear operator on $L^2_{P_X}$, and $\|\mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y}\| \leq 1$.

- $\langle f, \mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y} f\rangle_{L^2_{P_X}}$ generalizes $\mathrm{Var}(\mathbb{E}[X|Y])$.

Relate operators with central class:

Theorem 5

If conditions (A) and (B) are satisfied and $\mathcal{M}_{\mathcal{G}^*}$ **is complete**, then

$$\overline{\text{Range}\left(\mathbb{E}^*_{X|Y}\mathbb{E}_{X|Y}\right)} = \mathcal{M}_{\mathcal{G}^*}.$$

Remark:

1) $L^2_{P_X}$ inner product absorbs the marginal variance in the predictor vector.
2) Sample estimator of the directions from $\mathcal{M}_{\mathcal{G}^*}$ is called GSIR.

$$T : \mathcal{H}_X \to L^2_{P_X}, \quad f \to f - \mathbb{E}\left[f\right],$$
$$T_j : \overline{\text{Range}(T)} \to \mathbb{R}, \quad g \to \mathbb{E}\left[g(X)|Y \in J_i\right],$$
where $\{J_i\}_{i=1,\dots,h}$ is a partition of Ω_Y (i.e. slicing)
$\mu_1, \dots, \mu_h \in \overline{\text{Range}(T)}$ are Riesz representations of $T_j's$.

Then use

$$\text{span}\left\{\Sigma_{XX}^{-1}\mu_1, \cdots, \Sigma_{XX}^{-1}\mu_h\right\} \subseteq \mathscr{C}_{Y|X} \subseteq \mathcal{M}_{\mathcal{G}^*}.$$

Sample Estimator for GSIR

$$\mathbb{E}^*_{X|Y}\widehat{\mathbb{E}_{X|Y}}$$
$$= (G_X + \epsilon_X I_n)^{-3/2} G_X^{3/2} (G_Y + \epsilon_Y I_n)^{-1} G_Y^2 (G_Y + \epsilon_Y I_n)^{-1} G_X^{3/2} (G_X + \epsilon_X I_n)^{-3/2}$$

where

G_X : centered Gram matrix indroced by pd function $k_X(\cdot, \cdot)$.

Then

$$\hat{f}_i = \hat{\phi}_i^T (G_Y + \epsilon_Y I_n)^{-1},$$

where $\hat{\phi}_i$ is the i^{th} leading eigen-vector of $\mathbb{E}^*_{X|Y}\widehat{\mathbb{E}_{X|Y}}$.

GSAVE

Define:

$$\mathbb{E}^{(nc)}_{Y|X} : {L^2_{P_X}}^{(nc)} \to {L^2_{P_Y}}^{(nc)} \quad \text{such that}$$
$$\langle g, \mathbb{E}^{(nc)}_{Y|X} f \rangle_{{L^2_{P_X}}^{(nc)}} = \mathbb{E}\left[g(Y)f(X)\right]$$

Then there exists an operator

$$V_{X|Y} : \Omega_Y \to \mathcal{B}(L^2_{P_X})$$
to represent $\left(\mathbb{E}^{(nc)}_{Y|X}[fg] - \mathbb{E}^{(nc)}_{Y|X}[f]\mathbb{E}^{(nc)}_{Y|X}[g]\right)(y)$.

So we have:

$$\langle f, V_{X|Y} f \rangle_{L^2_{P_X}} = \mathrm{Var}\left(f(X)|Y\right)$$

GSAVE

Define:

$$\mathbb{E}^{(nc)}_{Y|X} : {L^2_{P_X}}^{(nc)} \to {L^2_{P_Y}}^{(nc)} \quad \text{such that}$$
$$\langle g, \mathbb{E}^{(nc)}_{Y|X} f \rangle_{{L^2_{P_X}}^{(nc)}} = \mathbb{E}\left[g(Y)f(X)\right]$$

Then there exists an operator

$$V_{X|Y} : \Omega_Y \to \mathcal{B}(L^2_{P_X})$$
$$\text{to represent } \left(\mathbb{E}^{(nc)}_{Y|X}[fg] - \mathbb{E}^{(nc)}_{Y|X}[f]\mathbb{E}^{(nc)}_{Y|X}[g]\right)(y).$$

So we have:

$$\langle f, V_{X|Y} f \rangle_{L^2_{P_X}} = \mathrm{Var}\left(f(X)|Y\right)$$

Define:

$$\mathbb{E}^{(nc)}_{Y|X} : {L^2_{P_X}}^{(nc)} \to {L^2_{P_Y}}^{(nc)} \quad \text{such that}$$
$$\langle g, \mathbb{E}^{(nc)}_{Y|X} f \rangle_{{L^2_{P_X}}^{(nc)}} = \mathbb{E}\left[g(Y)f(X)\right]$$

Then there exists an operator

$$V_{X|Y} : \Omega_Y \to \mathcal{B}(L^2_{P_X})$$
$$\text{to represent } \left(\mathbb{E}^{(nc)}_{Y|X}[fg] - \mathbb{E}^{(nc)}_{Y|X}[f]\mathbb{E}^{(nc)}_{Y|X}[g]\right)(y).$$

So we have:

$$\langle f, V_{X|Y} f \rangle_{L^2_{P_X}} = \mathrm{Var}\left(f(X)|Y\right)$$

Define

$$V : \langle f, Vg \rangle = \mathbf{Cov}\left(f(X), g(X)\right),$$
$$S = \mathbb{E}\left[\left(V - V_{X|Y}\right)^2\right].$$

S generalizes $\Sigma^{-1}\mathbb{E}\left[\mathrm{Var}\left(X\right) - \mathrm{Var}\left(X|Y\right)\right]^2 \Sigma^{-1}$.

Then,

$$\mathscr{C}_{X|Y} \subseteq \overline{\mathrm{Range}(S)} \subseteq \mathcal{M}_{\mathcal{G}^*}$$

Remark: GSAVE is expected to discover functions outside $\mathscr{C}_{X|Y}$.

Define

$$V : \langle f, Vg \rangle = \textbf{Cov}\left(f(X), g(X)\right),$$
$$S = \mathbb{E}\left[\left(V - V_{X|Y}\right)^2\right].$$

S generalizes $\Sigma^{-1}\mathbb{E}\left[\text{Var}\left(X\right) - \text{Var}\left(X|Y\right)\right]^2 \Sigma^{-1}$.

Then,

$$\mathscr{C}_{X|Y} \subseteq \overline{\text{Range}(S)} \subseteq \mathcal{M}_{\mathcal{G}^*}$$

Remark: GSAVE is expected to discover functions outside $\mathscr{C}_{X|Y}$.

Sample Estimator for GSAVE

$$\widehat{\mathbb{E}}^{(nc)}_{X|Y} = \left(L_Y L_Y^T\right)^+ \left(L_Y L_X^T\right)$$

where

$L_X = (1_n, K_X)^T$, i.e. non-centered Gram matrix plus a intercept column

$$\mathscr{L}_Y(y) = (1, k_Y(y, Y_1), \ldots, k_Y(y, Y_n))^T$$
$$C_Y(y) = L_Y^T \left(L_Y L_Y^T\right)^+ \mathscr{L}_Y(y)$$
$$\Lambda(y) = \mathsf{diag}\left(C_Y(y)\right) - C_Y(y) C_Y^T(y)$$

Then

$$\hat{S} = \frac{1}{n}\sum_{i=1}^{n} \left(L_X Q L_X^T + \epsilon_X I_{n+1}\right)^{-1/2} L_X Q \Gamma_i Q \Gamma_i Q L_X^T \left(L_X Q L_X^T + \epsilon_X I_{n+1}\right)^{-1/2}$$

where

$$\Gamma_i = \left(Q/n - \Lambda(Y_i)\right), \quad Q = I_n - 1_n 1_n^T / n$$