# SLICED INVERSE REGRESSION IN METRIC SPACES

Joni Virta, Kuang-Yao Lee, and Lexin Li

*University of Turku*

*Temple University*

*University of California at Berkeley*

## S1. Proofs of the theoretical results

**Proof of Lemma 1**: As the metric space $(\Omega_X^0, d_X)$ is separable, so is $\mathcal{H}_X^0$ (Lukić and Beder, 2001, Lemma 4.3). Being a closed subspace of a separable Hilbert space, $\ker(\Sigma_{XX}^0)$ is also a separable Hilbert space, and hence admits a countable dense subset $\mathcal{K}$. Fixing $h \in \mathcal{K}$, we have $\mathrm{Var}\{h(X)\} = 0$, implying that there exists a set $\mathcal{S}_{Xh}$, such that $P_X(\mathcal{S}_{Xh}) = 1$ and that, for all $x \in \mathcal{S}_{Xh}$, we have $h(x) - \mathrm{E}\{h(X)\} = \langle h, \kappa_X(\cdot, x) - \mu_X \rangle_{\mathcal{H}_X^0} = 0$.

Denoting $\Omega_X = \cap_{h \in \mathcal{K}} \mathcal{S}_{Xh}$, the countability of $\mathcal{K}$ implies that $P_X(\Omega_X) = 1$. We next show that, for all $x \in \Omega_X$, we have $\kappa_X(\cdot, x) - \mu_X \in \ker(\Sigma_{XX}^0)^\perp = \mathcal{H}_X$. Taking an arbitrary $g \in \ker(\Sigma_{XX}^0)$, there exist a sequence of elements $h_j \in \mathcal{K}$, such that $\|g - h_j\|_{\mathcal{H}_X^0} \to 0$ as $j \to \infty$. Let $\mathbb{N}$ denote the collection of natural numbers. Then, for an arbitrary $x \in \Omega_X$ and $j \in \mathbb{N}$, we have that,

$$
\begin{aligned}
|\langle \kappa_X(\cdot, x) - \mu_X, g \rangle_{\mathcal{H}_X^0}| &\leq |\langle \kappa_X(\cdot, x) - \mu_X, g - h_j \rangle_{\mathcal{H}_X^0}| + |\langle \kappa_X(\cdot, x) - \mu_X, h_j \rangle_{\mathcal{H}_X^0}| \\
&\leq \|\kappa_X(\cdot, x) - \mu_X\|_{\mathcal{H}_X^0} \|g - h_j\|_{\mathcal{H}_X^0} + 0,
\end{aligned}
$$

which further implies that $\langle \kappa_X(\cdot, x) - \mu_X, g \rangle_{\mathcal{H}_X^0} = 0$. This completes the proof of Lemma 1. $\square$

---

**Proof of Lemma 2**: *(a).* The proof is structurally similar to that of Proposition 1 in Li and Song (2017), but we include it for completeness.

Fix $f \in \mathcal{H}_X$. Denote $\Sigma_{YY}^{\dagger}\Sigma_{YX} = R_{YX}$, and fix an arbitrary $g \in \mathcal{H}_Y$. Then,

$$\mathrm{Cov}\{(R_{YX}f)(Y), g(Y)\} = \langle \Sigma_{YY}^{\dagger}\Sigma_{YX}f, \Sigma_{YY}g\rangle_{\mathcal{H}_Y} = \langle \Sigma_{YX}f, g\rangle_{\mathcal{H}_Y} = \mathrm{Cov}\{f(X), g(Y)\}.$$

Consequently, for all $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$, we have,

$$\mathrm{Cov}\{f(X) - (R_{YX}f)(Y), g(Y)\} = 0, \tag{S1.1}$$

Consider an arbitrary $h \in L_2(P_Y)$. By Assumption 2, there exist a sequence $\{h_n\}$ of elements of $\mathcal{H}_Y$, such that $\mathrm{var}\{h(Y) - h_n(Y)\} \to 0$, as $n \to \infty$. Therefore, by (S1.1),

$$\begin{aligned}
&|\mathrm{Cov}\{f(X) - (R_{YX}f)(Y), h(Y)\}| \\
\le\; &|\mathrm{Cov}\{f(X) - (R_{YX}f)(Y), h(Y) - h_n(Y)\}| + |\mathrm{Cov}\{f(X) - (R_{YX}f)(Y), h_n(Y)\}| \\
\le\; &[\mathrm{Var}\{f(X) - (R_{YX}f)(Y)\}\,\mathrm{Var}\{h(Y) - h_n(Y)\}]^{1/2} + 0,
\end{aligned}$$

for all $n$. The first variance in the final expression above is finite, since $\mathrm{Var}\{f(X)\} \le \|\Sigma_{XX}\|_{\mathrm{OP}}\|f\|^2_{\mathcal{H}_X} < \infty$, and $\mathrm{Var}\{(R_{YX}f)(Y)\} \le \|R_{YX}\|_{\mathrm{OP}}\|\Sigma_{YX}\|_{\mathrm{OP}}\|f\|^2_{\mathcal{H}_X} < \infty$. This implies that (S1.1) holds also when $g$ is replaced with any $h \in L_2(P_Y)$.

Note that a square-integrable random variable $Z$ is almost surely equal to the conditional expectation $\mathrm{E}\{f(X) \mid Y\}$, if

$$\mathrm{E}[\{f(X) - Z\}h(Y)] = 0, \tag{S1.2}$$

for all $h \in L_2(P_Y)$. A direct computation using (S1.1) shows that the choice $Z = (R_{YX}f)(Y) - \mathrm{E}\{(R_{YX}f)(Y)\} + \mathrm{E}\{f(X)\}$ satisfies (S1.2) for all $h \in L_2(P_Y)$, which implies the following holds almost surely,

$$\mathrm{E}\{f(X) \mid Y\} - \mathrm{E}\{f(X)\} = (\Sigma_{YY}^{\dagger}\Sigma_{YX}f)(Y) - \mathrm{E}\{(\Sigma_{YY}^{\dagger}\Sigma_{YX}f)(Y)\}. \tag{S1.3}$$

Finally, by Lemma 1, the right-hand side of (S1.3) is almost surely equal to the random variable $\langle \Sigma_{YY}^{\dagger}\Sigma_{YX}f, \kappa_Y(\cdot, Y) - \mu_Y\rangle_{\mathcal{H}_Y}$, where we take $\kappa_Y(\cdot, Y) - \mu_Y$ to equal the zero element for those values of $Y$ for which it is not a member of $\mathcal{H}_Y$. This proves the assertion (a).

*(b).* By definition,

$$\mathrm{E}\langle g, [\{\kappa_Y(\cdot, Y) - \mu_Y\} \otimes \{\kappa_Y(\cdot, Y) - \mu_Y\}]g'\rangle_{\mathcal{H}_Y} = \langle g, \Sigma_{YY}g'\rangle_{\mathcal{H}_Y},$$

for all $g, g' \in \mathcal{H}_Y$. This, in conjunction with part (a) of the lemma, implies that the left-hand side of the assertion (b) equals,

$$\begin{aligned}
&\mathrm{E}\langle \Sigma_{YY}^{\dagger}\Sigma_{YX}f, [\{\kappa_Y(\cdot, Y) - \mu_Y\} \otimes \{\kappa_Y(\cdot, Y) - \mu_Y\}]\Sigma_{YY}^{\dagger}\Sigma_{YX}f'\rangle_{\mathcal{H}_Y} \\
&= \langle \Sigma_{YY}^{\dagger}\Sigma_{YX}f, \Sigma_{YY}\Sigma_{YY}^{\dagger}\Sigma_{YX}f'\rangle_{\mathcal{H}_Y} = \langle f, \Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YX}f'\rangle_{\mathcal{H}_X}.
\end{aligned}$$

This proves the assertion (b), and completes the proof of Lemma 2. □

**Proof of Theorem 1**: By Theorem 1 in Douglas (1966), and Assumption 3, both $\Sigma_{YY}^{\dagger}\Sigma_{YX}$ and $\Sigma_{XX}^{\dagger}\Sigma_{XY}$ are bounded. Henceforth, $\Lambda_{\text{SIR}}$ is also bounded.

To prove the unbiasedness of $\Lambda_{\text{SIR}}$, we first note that the set $\mathcal{S}_{Y|X}$ is closed because measurability is preserved in taking point-wise limits, which in an RKHS is implied by the convergence in norm. Concurrently, $\mathcal{S}_{Y|X} = \overline{\text{span}}\{f \in \mathcal{H}_X \mid f \text{ is } \mathcal{G}_{Y|X}\text{-measurable }\}$, and we have the desired result of $\overline{\text{ran}}(\Lambda_{\text{SIR}}) \subseteq \mathcal{S}_{Y|X}$, as long as we can show that $\mathcal{S}_{Y|X}^{\perp} \subseteq \ker(\Lambda_{\text{SIR}}^*)$.

We begin by establishing this inclusion for the elements of $\text{ran}(\Sigma_{XX})$. Let $f = \Sigma_{XX}m$ for an arbitrary $m \in \mathcal{H}_X$. Suppose $\langle f, h\rangle_{\mathcal{H}_X} = 0$ for all $h \in \mathcal{S}_{Y|X}$, which implies that, $\text{Cov}\{m(X), h(X)\} = \langle \Sigma_{XX}m, h\rangle_{\mathcal{H}_X} = 0$ for all $h \in \mathcal{S}_{Y|X}$. Let $\mathcal{S}_{Y|X}^* = \{h \in L_2(P_X) \mid h \text{ is } \mathcal{G}_{Y|X}\text{-measurable}\}$. Then, by (Li, 2018, Theorem 13.3), we have that $\text{Cov}\{m(X), h(X)\} = 0$ for all $h \in \mathcal{S}_{Y|X}^*$. This in turn implies that

$$\langle m - \mathrm{E}\{m(X)\}, h\rangle_{L_2(P_X)} = 0,$$

for all $h \in \mathcal{S}^*_{Y|X}$, where $\mathrm{E}\{m(X)\}$ represents the constant function taking the value $\mathrm{E}\{m(X)\}$ everywhere. Therefore, following Lee et al. (2013, Lemma 1), we have that

$$\mathrm{E}\{m(X) \mid \mathcal{G}_{Y|X}\} - \mathrm{E}\{m(X)\} = 0, \quad \text{almost surely.} \tag{S1.4}$$

We next show that (S1.4) leads to $\mathrm{E}\{m(X) \mid Y\} = \mathrm{E}\{m(X)\}$ almost surely. Let $\sigma(Y, \mathcal{G}_{Y|X})$ be the smallest $\sigma$-field containing both $\sigma(Y)$ and $\mathcal{G}_{Y|X}$. By rule of iterative expectation, we have, almost surely,

$$\mathrm{E}\{m(X) \mid Y\} = \mathrm{E}[\mathrm{E}\{m(X) \mid \sigma(Y, \mathcal{G}_{Y|X})\} \mid Y] = \mathrm{E}[\mathrm{E}\{m(X) \mid \mathcal{G}_{Y|X}\} \mid Y] = \mathrm{E}\{m(X)\},$$

where the second equality follows from the fact that $Y \perp\!\!\!\perp X \mid \mathcal{G}_{Y|X}$, and the third equality is by (S1.4).

Combining the above result with Lemma 2 leads to that, for all $g \in \mathcal{H}_X$,

$$0 = \langle m, \Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YX} g\rangle_{\mathcal{H}_X} = \langle f, \Lambda_{\mathrm{SIR}} g\rangle_{\mathcal{H}_X}.$$

In other words, $f \in \ker(\Lambda^*_{\mathrm{SIR}})$. Therefore, $\mathrm{ran}(\Sigma_{XX}) \cap \mathcal{S}^{\perp}_{Y|X} \subseteq \ker(\Lambda^*_{\mathrm{SIR}})$.

To extend this inclusion to hold in the full orthogonal complement $\mathcal{S}^{\perp}_{Y|X}$, we invoke Assumption 4, which implies that, for $f \in \mathcal{S}^{\perp}_{Y|X}$, there exist a sequence of elements $f_n$ of $\mathrm{ran}(\Sigma_{XX}) \cap \mathcal{S}^{\perp}_{Y|X}$, such that $\|f_n - f\|_{\mathcal{H}_X} \to 0$, as $n \to 0$. Because $\Lambda^*_{\mathrm{SIR}} f_n = 0$ for all $n$, we also have $\Lambda^*_{\mathrm{SIR}} f = 0$ by continuity. This completes the proof of Theorem 1. □

**Proof of Theorem 3**: We first present three auxiliary lemmas, under the same set of conditions of Theorem 3. We then prove Theorem 3 based on these lemmas.

The first auxiliary lemma shows that the sample covariance operators are root-$n$ consistent estimators of the corresponding population counterparts.

**Lemma S1.1.** *Suppose the conditions of Theorem 3 hold. Then,* $\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{\mathrm{HS}}$, $\|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_{\mathrm{HS}}$ *and* $\|\hat{\Sigma}_{YY} - \Sigma_{YY}\|_{\mathrm{HS}}$ *are of the order* $\mathcal{O}_p(1/\sqrt{n})$.

*Proof of Lemma S1.1.* Denote $h_X = \kappa_X(\cdot, X) - \mu_X$. By definition, the covariance operator $\Sigma_{XX}$ satisfies that,

$$\langle f, \Sigma_{XX} g\rangle_{\mathcal{H}_X} = \mathrm{E}(\langle f, h_X\rangle_{\mathcal{H}_X} \langle g, h_X\rangle_{\mathcal{H}_X}) = \mathrm{E}\{\langle f, (h_X \otimes h_X) g\rangle_{\mathcal{H}_X}\},$$

where $h_X$ is to be the zero element for those realizations $x \in \Omega$ not belonging to the almost sure set in Lemma 1.

Since the covariance operator in a separable Hilbert space is a trace-class operator (Zwald et al., 2004), we have that $\|\Sigma_{XX}\|_{\mathrm{HS}} < \infty$. To show $\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{\mathrm{HS}} = \mathcal{O}_p(1/\sqrt{n})$, we note that,

$$\hat{\Sigma}_{XX} = \frac{1}{n} \sum_{i=1}^{n} (b_{X_i} \otimes b_{X_i}),$$

where $b_{X_i} = \kappa_X(\cdot, X_i) - (1/n)\sum_{j=1}^{n} \kappa_X(\cdot, X_j)$. Denoting $h_{X_i} = \kappa_X(\cdot, X_i) - \mu_X$, we further have that $b_{X_i} = h_{X_i} - \bar{h}_n$ where $\bar{h}_n = (1/n)\sum_{i=1}^{n} h_{X_i}$. Therefore,

$$\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{\mathrm{HS}} \le \left\| \frac{1}{n} \sum_{i=1}^{n} (h_{X_i} \otimes h_{X_i}) - \Sigma_{XX} \right\|_{\mathrm{HS}} + \|\bar{h}_n \otimes \bar{h}_n\|_{\mathrm{HS}}. \tag{S1.5}$$

For the second term on the right-hand-side of (S1.5), we have that,

$$\mathrm{E}\|\bar{h}_n \otimes \bar{h}_n\|_{\mathrm{HS}} = \mathrm{E}\|\bar{h}_n\|^2_{\mathcal{H}_X} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{E}\langle h_{X_i}, h_{X_j}\rangle_{\mathcal{H}_X}. \tag{S1.6}$$

For any pair of distinct indices $i \neq j$, $h_{X_i}$ and $h_{X_j}$ are independent. This means that,

$$\mathrm{E}\langle h_{X_i}, h_{X_j}\rangle_{\mathcal{H}_X} = \mathrm{E}_{X_j}\langle \mathrm{E}_{X_i}(h_{X_i}), h_{X_j}\rangle_{\mathcal{H}_X} = \mathrm{E}_{X_j}\langle 0, h_{X_j}\rangle_{\mathcal{H}_X} = 0, \tag{S1.7}$$

where $\mathrm{E}_{X_i}(\cdot)$ is the expectation with respect to the distribution of $X_i$. Consequently,

$$\mathrm{E}\|\bar{h}_n \otimes \bar{h}_n\|_{\mathrm{HS}} = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{E}\|h_{X_i}\|^2_{\mathcal{H}_X} = \frac{1}{n} \mathrm{E}\|h_{X_1}\|^2_{\mathcal{H}_X},$$

where $\mathrm{E}\|h_{X_1}\|^2_{\mathcal{H}_X} = \mathrm{E}\|h_{X_1}\|^2_{\mathcal{H}^0_X} = \mathrm{E}\{\kappa_X(X_1, X_1)\} - \|\mu_X\|^2_{\mathcal{H}^0_X} < \infty$, which holds by Assumption 8. Therefore, $\mathrm{E}\|\bar{h}_n \otimes \bar{h}_n\|_{\mathrm{HS}} = \mathcal{O}(1/n)$, which, by Markov's inequality, further implies that $\|\bar{h}_n \otimes \bar{h}_n\|_{\mathrm{HS}} = \mathcal{O}_p(1/n)$.

For the first term on the right-hand-side of (S1.5), we have that,

$$\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(h_{X_i} \otimes h_{X_i}) - \Sigma_{XX}\right\|^2_{\mathrm{HS}} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\langle H_i, H_j\rangle_{\mathrm{HS}}, \tag{S1.8}$$

where $H_i = (h_{X_i} \otimes h_{X_i}) - \Sigma_{XX}$. By the definition of the Hilbert-Schmidt norm,

$$\mathrm{E}\langle H_i, H_j\rangle_{\mathrm{HS}} = \mathrm{E}\langle h_{X_i} \otimes h_{X_i}, h_{X_j} \otimes h_{X_j}\rangle_{\mathrm{HS}} - \langle \Sigma_{XX}, \Sigma_{XX}\rangle_{\mathrm{HS}}.$$

Therefore, adopting he same strategy used to simplify (S1.6), we have $\mathrm{E}\langle h_{X_i} \otimes h_{X_i}, h_{X_j} \otimes h_{X_j}\rangle_{\mathrm{HS}} = \langle \Sigma_{XX}, \Sigma_{XX}\rangle_{\mathrm{HS}}$, for $i \neq j$. This allows us to ignore all pairs of distinct indices in (S1.8), and we obtain that,

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\langle H_i, H_j\rangle_{\mathrm{HS}} = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\|H_i\|^2_{\mathrm{HS}} = \frac{1}{n}\mathrm{E}\|H_1\|^2_{\mathrm{HS}}. \tag{S1.9}$$

Correspondingly,

$$\begin{aligned}\mathrm{E}\|H_1\|^2_{\mathrm{HS}} &\leq \mathrm{E}\|h_{X_1} \otimes h_{X_1}\|^2_{\mathrm{HS}} + 2\mathrm{E}\|h_{X_1} \otimes h_{X_1}\|_{\mathrm{HS}}\|\Sigma_{XX}\|_{\mathrm{HS}} + \|\Sigma_{XX}\|^2_{\mathrm{HS}} \\ &= \mathrm{E}\|h_{X_1}\|^4_{\mathcal{H}_X} + 2\mathrm{E}\|h_{X_1}\|^2_{\mathcal{H}_X}\|\Sigma_{XX}\|_{\mathrm{HS}} + \|\Sigma_{XX}\|^2_{\mathrm{HS}},\end{aligned}$$

which is finite, because $\Sigma_{XX}$ is a Hilbert-Schmidt operator, and that, by Assumption 8, $\mathrm{E}\|\kappa_X(\cdot, X_1)\|^4_{\mathcal{H}_X} = \mathrm{E}\{\kappa_X(X_1, X_1)^2\} < \infty$, guaranteeing that $\mathrm{E}\|h_{X_1}\|^4_{\mathcal{H}_X}$ is finite. Together, (S1.8) and (S1.9), along with Markov's inequality, imply that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(h_{X_i} \otimes h_{X_i}) - \Sigma_{XX}\right\|_{\mathrm{HS}} = \mathcal{O}_p(1/\sqrt{n}).$$

Combining the results above, we obtain that $\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{\mathrm{HS}} = \mathcal{O}_p(1/\sqrt{n})$. The results for $\hat{\Sigma}_{XY}$ and $\hat{\Sigma}_{YY}$ can be established similarly. This completes the proof of Lemma S1.1. $\square$

The second auxiliary lemma shows that the inverse operators $G_{n1}^{-1}$ and $G_{n2}^{-1}$ are bounded in the operator norm, where $G_{n1}^{-1} = (\hat{\Sigma}_{XX} + \tau I)^{-1}$ and $G_{n2}^{-1} = (\Sigma_{XX} + \tau I)^{-1}$.

**Lemma S1.2.** *Suppose the conditions of Theorem 3 hold. Then,* $\|G_{n1}^{-1}\|_{\mathrm{OP}} \leq 1/\tau$, *and* $\|G_{n2}^{-1}\|_{\mathrm{OP}} \leq 1/\tau$.

*Proof of Lemma S1.2.* Note that $\|G_{n2}^{-1}\|_{\mathrm{OP}} = (1/\tau)\|(\Sigma_{XX}/\tau + I)^{-1}\|_{\mathrm{OP}} \leq 1/\tau$. This relation holds because, by the positive semi-definiteness of $T = \Sigma_{XX}/\tau$, we have, for arbitrary $f \in \mathcal{H}_X$,

$$\begin{aligned} \|(T+I)^{-1}f\|_{\mathcal{H}_X}^2 &\leq \langle (T+I)(T+I)^{-1}f, (T+I)^{-1}f\rangle_{\mathcal{H}_X} \\ &\leq \|f\|_{\mathcal{H}_X}\|(T+I)^{-1}f\|_{\mathcal{H}_X}, \end{aligned}$$

implying that $\|(T+I)^{-1}f\|_{\mathcal{H}_X} \leq \|f\|_{\mathcal{H}_X}$. Similarly, we can show that $\|G_{n1}^{-1}\|_{\mathrm{OP}} \leq 1/\tau$. This completes the proof of Lemma S1.2. □

The third auxiliary lemma establishes the convergence rate for the effect of replacing the pseudo-inverse with its regularized counterpart on the population level.

**Lemma S1.3.** *Suppose the conditions of Theorem 3 hold. Then,* $\|G_{n2}^{-1}\Sigma_{XY} - \Sigma_{XX}^{\dagger}\Sigma_{XY}\|_{\mathrm{OP}} = \mathcal{O}(\tau)$.

*Proof of Lemma S1.3.* By Assumption 9 and Theorem 1 of Douglas (1966), we have $\Sigma_{XY} = \Sigma_{XX}^2 C$, for some bounded operator $C : \mathcal{H}_X \to \mathcal{H}_X$. This further implies that,

$$\begin{aligned} \|G_{n2}^{-1}\Sigma_{XY} - \Sigma_{XX}^{\dagger}\Sigma_{XY}\|_{\mathrm{OP}} &\leq \|(G_{n2}^{-1}\Sigma_{XX} - I)\Sigma_{XX}\|_{\mathrm{OP}}\|C\|_{\mathrm{OP}} \\ &= \|-\tau G_{n2}^{-1}\Sigma_{XX}\|_{\mathrm{OP}}\|C\|_{\mathrm{OP}} = \tau\|\tau G_{n2}^{-1} - I\|_{\mathrm{OP}}\|C\|_{\mathrm{OP}}, \end{aligned}$$

where $\|\tau G_{n2}^{-1} - I\|_{\mathrm{OP}} \leq \tau\|G_{n2}^{-1}\|_{\mathrm{OP}} + \|I\|_{\mathrm{OP}} \leq 2$, by Lemma S1.2. Therefore,

$$\|G_{n2}^{-1}\Sigma_{XY} - \Sigma_{XX}^{\dagger}\Sigma_{XY}\|_{\mathrm{OP}} = \mathcal{O}(\tau).$$

This completes the proof of Lemma S1.3. □

Based on the above three auxiliary lemmas, we next prove Theorem 3.

We first establish the closeness of $G_{n1}^{-1}\hat{\Sigma}_{XY}$ to the operator $G_{n2}^{-1}\Sigma_{XY}$. Note that

$$\begin{aligned}
&\|G_{n1}^{-1}\hat{\Sigma}_{XY} - G_{n2}^{-1}\Sigma_{XY}\|_{\mathrm{OP}} \\
\leq\ &\|G_{n1}^{-1}\|_{\mathrm{OP}}\|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_{\mathrm{OP}} + \|(G_{n1}^{-1} - G_{n2}^{-1})\Sigma_{XY}\|_{\mathrm{OP}}.
\end{aligned} \tag{S1.10}$$

To simplify (S1.10), we observe that, by Assumption 9, we have $\Sigma_{XY} = \Sigma_{XX}D$ for some bounded operator $D$. This implies that

$$\begin{aligned}
\|(G_{n1}^{-1} - G_{n2}^{-1})\Sigma_{XY}\|_{\mathrm{OP}} &= \|G_{n1}^{-1}(G_{n1} - G_{n2})G_{n2}^{-1}\Sigma_{XY}\|_{\mathrm{OP}} \\
&\leq \|G_{n1}^{-1}\|_{\mathrm{OP}}\|G_{n1} - G_{n2}\|_{\mathrm{OP}}\|G_{n2}^{-1}\Sigma_{XY}\|_{\mathrm{OP}} \\
&\leq (1/\tau)\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{\mathrm{OP}}\|G_{n2}^{-1}\Sigma_{XY}\|_{\mathrm{OP}} \\
&\leq \mathcal{O}_p(1/\{\tau\sqrt{n}\})\|(\Sigma_{XX} + \tau I)^{-1}\Sigma_{XX}\|_{\mathrm{OP}}\|D\|_{\mathrm{OP}} \\
&= \mathcal{O}_p(1/\{\tau\sqrt{n}\}),
\end{aligned}$$

where the last equality holds because the largest eigenvalue of the operator $(\Sigma_{XX} + \tau I)^{-1}\Sigma_{XX}$ is bounded from above by one.

Therefore, together with Lemmas S1.1 and S1.2, we have that

$$\|G_{n1}^{-1}\hat{\Sigma}_{XY} - G_{n2}^{-1}\Sigma_{XY}\|_{\mathrm{OP}} = \mathcal{O}_p(1/\{\tau\sqrt{n}\}). \tag{S1.11}$$

Combining (S1.11) with Lemma S1.3 leads to

$$\|G_{n1}^{-1}\Sigma_{XY} - \Sigma_{XX}^{\dagger}\Sigma_{XY}\|_{\mathrm{OP}} = \mathcal{O}_p(\tau + 1/\{\tau\sqrt{n}\}).$$

The sample convergence of $(\hat{\Sigma}_{YY} + \tau I)^{-1}\hat{\Sigma}_{YX}$ can be established similarly. This completes the proof of Theorem 3.

□

**Proof of Theorem 4**: Denote $\mathbb{I}_{ik} = \mathbb{I}(Y_i = k)$, $N_k = \sum_{i=1}^n \mathbb{I}_{ik}$, and $h_{X_i} = \kappa_X(\cdot, X_i) - \mu_X$. We have that

$$\hat{\gamma}_{X|k} = \frac{1}{N_k}\sum_{i=1}^n \mathbb{I}_{ik}h_{X_i} - \frac{1}{n}\sum_{i=1}^n h_{X_i} = \frac{1}{N_k}\sum_{i=1}^n \mathbb{I}_{ik}h_{X_i} - \bar{h}_n.$$

Consequently,

$$\begin{aligned}\mathrm{E}\|\hat{\gamma}_{X|k}-\gamma_{X|k}\|^2_{\mathcal{H}_X} = \mathrm{E}\left\|\frac{1}{N_k}\sum_{i=1}^{n}\mathbb{I}_{ik}h_{X_i}-\gamma_{X|k}\right\|^2_{\mathcal{H}_X} + \mathrm{E}\|\bar{h}_n\|^2_{\mathcal{H}_X} \\ - 2\mathrm{E}\left\langle\frac{1}{N_k}\sum_{i=1}^{n}\mathbb{I}_{ik}h_{X_i}-\gamma_{X|k},\bar{h}_n\right\rangle_{\mathcal{H}_X}.\end{aligned} \tag{S1.12}$$

The first term of the right-hand-side of (S1.12) is equal to,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left(\frac{1}{N_k^2}\mathbb{I}_{ik}\mathbb{I}_{jk}\langle h_{X_i}-\gamma_{X|k},h_{X_j}-\gamma_{X|k}\rangle_{\mathcal{H}_X}\right). \tag{S1.13}$$

Denote $v_{ij}=\langle h_{X_i}-\gamma_{X|k},h_{X_j}-\gamma_{X|k}\rangle_{\mathcal{H}_X}$. Then, for any distinct $i\neq j$, conditioning on the values of the corresponding indicators implies the summand in (S1.13) equals

$$\mathrm{E}\left(\frac{1}{N_k^2}v_{ij}\;\middle|\;\mathbb{I}_{ik}\mathbb{I}_{jk}=1\right)\pi_k^2 = \mathrm{E}\left(\frac{1}{N_k^2}\;\middle|\;\mathbb{I}_{ik}\mathbb{I}_{jk}=1\right)\mathrm{E}\left(v_{ij}\;\middle|\;\mathbb{I}_{ik}\mathbb{I}_{jk}=1\right)\pi_k^2,$$

where $\pi_k=P(Y=k)$. Using a similar argument as in (S1.7) and the definition of $\gamma_{X|k}$, we have that $\mathrm{E}\left(v_{ij}\;\middle|\;\mathbb{I}_{ik}\mathbb{I}_{jk}=1\right)=0$, implying that (S1.13) is further equal to

$$\begin{aligned}&\sum_{i=1}^{n}\mathrm{E}\left(\frac{1}{N_k^2}\mathbb{I}_{ik}\|h_{X_i}-\gamma_{X|k}\|^2_{\mathcal{H}_X}\right) \\ &= n\pi_k\mathrm{E}\left(\frac{1}{N_k^2}\mid\mathbb{I}_{1k}=1\right)\mathrm{E}\left(\|h_{X_1}-\gamma_{X|k}\|^2_{\mathcal{H}_X}\mid\mathbb{I}_{1k}=1\right).\end{aligned} \tag{S1.14}$$

Correspondingly, we have that

$$\begin{aligned}\pi_k\mathrm{E}(\|h_{X_1}-\gamma_{X|k}\|^2_{\mathcal{H}_X}\mid\mathbb{I}_{1k}=1) &= \pi_k\{\mathrm{E}(\|h_{X_1}\|^2_{\mathcal{H}_X}\mid\mathbb{I}_{1k}=1)-\|\gamma_{X|k}\|^2_{\mathcal{H}_X}\} \\ &\leq \mathrm{E}(\|h_{X_1}\|^2_{\mathcal{H}_X})<\infty.\end{aligned}$$

We next tackle the term $\mathrm{E}(1/N_k^2\mid\mathbb{I}_{1k}=1)$. Conditioning on $\{\mathbb{I}_{1k}=1\}$, the distribution of $(n-1)^2/N_k^2$ is that of $(n-1)^2/(B+1)^2$, where $B$ follows a Binomial$(n-1,\pi_k)$ distribution. By Theorem 1 of Shi et al. (2010), the expected value of $(n-1)^\alpha/(B+c)^\alpha$

is of the order $\mathcal{O}(1)$ for any $c, \alpha > 0$. This further implies that (S1.14) and, consequently, the first term on the right-hand-side of (S1.12), is of the order $\mathcal{O}(1/n)$.

The second term of the right-hand-side of (S1.12), as shown in Theorem 3 and by Assumption 8, is of the order $\mathcal{O}(1/n)$.

The third term of the right-hand-side of (S1.12) can be expressed as

$$-\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}\left(\frac{1}{N_k}\mathbb{I}_{ik}\langle h_{X_i} - \gamma_{X|k}, h_{X_j}\rangle\right). \tag{S1.15}$$

Conditioning on the indicators, we can rewrite the contribution of any index pair of $i \neq j$ to the sum as

$$\sum_{\ell=1}^{K} \pi_k\pi_\ell \mathrm{E}(1/N_k \mid \mathbb{I}_{ik}\mathbb{I}_{j\ell} = 1)\mathrm{E}(\langle h_{X_i} - \gamma_{X|k}, h_{X_j}\rangle \mid \mathbb{I}_{ik}\mathbb{I}_{j\ell} = 1),$$

where the final expected value is equal to zero, following a similar argument to (S1.7). Therefore, only the index pairs of $i = j$ contribute to the sum (S1.15), which then takes the form,

$$\begin{aligned}
&-\frac{2}{n}\sum_{i=1}^{n} \mathrm{E}\left(\frac{1}{N_k}\mathbb{I}_{ik}\langle h_{X_i} - \gamma_{X|k}, h_{X_i}\rangle\right) \\
=&-2\pi_k \mathrm{E}(1/N_k \mid \mathbb{I}_{ik} = 1)\mathrm{E}(\langle h_{X_i} - \gamma_{X|k}, h_{X_i}\rangle \mid \mathbb{I}_{ik} = 1).
\end{aligned}$$

Now, $\mathrm{E}(1/N_k \mid \mathbb{I}_{ik} = 1) = \mathcal{O}(1/n)$ by Theorem 1 of Shi et al. (2010). Moreover, we can show that the term $\pi_k \mathrm{E}(\langle h_{X_i} - \gamma_{X|k}, h_{X_i}\rangle \mid \mathbb{I}_{ik} = 1)$ is of the order $\mathcal{O}(1)$. Consequently, the third term on the right-hand-side of (S1.12) is of the order $\mathcal{O}(1/n)$.

Applying the Markov's inequality obtains that $\|\hat{\gamma}_{X|k} - \gamma_{X|k}\|_{\mathcal{H}_X} = \mathcal{O}_p(1/\sqrt{n})$.

We next show that $\sum_{k=1}^{K}(N_k/n)(\hat{\gamma}_{X|k} \otimes \hat{\gamma}_{X|k})$ converges in the Hilbert-Schmidt norm to $\Gamma_{XX|Y} = \sum_{k=1}^{K}\pi_k(\gamma_{X|k} \otimes \gamma_{X|k})$. Note that the norm of the difference $\sum_{k=1}^{K}(N_k/n)(\hat{\gamma}_{X|k} \otimes \hat{\gamma}_{X|k}) - \sum_{k=1}^{K}\pi_k(\gamma_{X|k} \otimes \gamma_{X|k})$ is upper-bounded by

$$
\begin{aligned}
&\left\|\sum_{k=1}^{K}(N_k/n)(\hat{\gamma}_{X|k}\otimes\hat{\gamma}_{X|k}) - \sum_{k=1}^{K}\pi_k(\gamma_{X|k}\otimes\gamma_{X|k})\right\|_{\mathrm{HS}} \\
\leq\ & \sum_{k=1}^{K}\|(N_k/n)(\hat{\gamma}_{X|k}\otimes\hat{\gamma}_{X|k}) - \pi_k(\gamma_{X|k}\otimes\gamma_{X|k})\|_{\mathrm{HS}} \\
\leq\ & \sum_{k=1}^{K}\{(N_k/n)\|(\hat{\gamma}_{X|k}\otimes\hat{\gamma}_{X|k}) - (\gamma_{X|k}\otimes\gamma_{X|k})\|_{\mathrm{HS}} + |(N_k/n)-\pi_k|\|\gamma_{X|k}\otimes\gamma_{X|k}\|_{\mathrm{HS}}\} \\
\leq\ & \sum_{k=1}^{K}\{(N_k/n)\|(\hat{\gamma}_{X|k}-\gamma_{X|k})\otimes\hat{\gamma}_{X|k}\|_{\mathrm{HS}} \\
& + (N_k/n)\|\{\gamma_{X|k}\otimes(\hat{\gamma}_{X|k}-\gamma_{X|k})\}\|_{\mathrm{HS}} + |(N_k/n)-\pi_k|\|\gamma_{X|k}\otimes\gamma_{X|k}\|_{\mathrm{HS}}\}. \qquad \text{(S1.16)}
\end{aligned}
$$

Note that $\|(\hat{\gamma}_{X|k}-\gamma_{X|k})\otimes\hat{\gamma}_{X|k}\|^2_{\mathrm{HS}} = \|\hat{\gamma}_{X|k}-\gamma_{X|k}\|^2_{\mathcal{H}_X}\|\hat{\gamma}_{X|k}\|^2_{\mathcal{H}_X} = \mathcal{O}_p(1/n)$, because $\|\hat{\gamma}_{X|k}\|^2_{\mathcal{H}_X}$ converges to the finite constant $\|\gamma_{X|k}\|^2_{\mathcal{H}_X}$ as $n\to\infty$. Similarly, we can show that $\|\gamma_{X|k}\otimes(\hat{\gamma}_{X|k}-\gamma_{X|k})\|^2_{\mathrm{HS}} = \mathcal{O}_p(1/n)$, and that $|(N_k/n)-\pi_k| = \mathcal{O}_p(1/\sqrt{n})$, which follows from the standard Central Limit Theorem. Substituting all terms with their rates in (S1.16), we have that $\|\sum_{k=1}^{K}(N_k/n)(\hat{\gamma}_{X|k}\otimes\hat{\gamma}_{X|k})-\sum_{k=1}^{K}\pi_k(\gamma_{X|k}\otimes\gamma_{X|k})\|_{\mathrm{HS}} = \mathcal{O}_p(1/\sqrt{n})$.

We further employ the proof of Theorem 3 to deal with the inclusion of the pseudo-inverses. This completes the proof of Theorem 4. □

## References


Douglas, R. G. (1966). On majorization, factorization, and range inclusion of operators on Hilbert space. *Proceedings of the American Mathematical Society 17*(2), 413–415.

Lee, K.-Y., B. Li, and F. Chiaromonte (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics 41*(1), 221–249.

Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R.* Chapman and Hall, CRC.

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

Lukić, M. and J. Beder (2001). Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society 353*(10), 3945–3969.

Shi, X., Y. Wu, and Y. Liu (2010). A note on asymptotic approximations of inverse moments of nonnegative random variables. *Statistics & Probability Letters 80*(15-16), 1260–1264.

Zwald, L., O. Bousquet, and G. Blanchard (2004). Statistical properties of kernel principal component analysis. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science) 3120*, 594–608.