The Institute of Statisitcal Mathematics
Research Memorandum No. 942 June 02, 2005

Consistency of Kernel Canonical Correlation Analysis

Kenji Fukumizu
Francis R. Bach
and
Arthur Gretton

Consistency of Kernel Canonical Correlation Analysis

Kenji Fukumizu* Francis R. Bach† Arthur Gretton‡

Abstract

While the kernel CCA has been applied in many problems, the convergence of the estimated function with finite sample to the true function has not been established yet. This paper gives a mathematical proof of the statistical convergence of kernel CCA and a related method (NOCCO) to provide theoretical justification for the methods. The result gives also a sufficient condition on the regularization coefficient in the methods to ensure convergence.

1 Introduction

Kernel methods ([17]) have been recently developed as a methodology of nonlinear data analysis with positive definite kernels. In kernel methods, data are represented as functions or elements in the reproducing kernel Hilbert spaces (RKHS), which are given by the positive definite kernels. Application of various linear methods in the Hilbert spaces are possible by the reproducing property, which makes the computation of the inner product in the Hilbert space tractable. Many methods have been proposed as a nonlinear extension of conventional linear methods, such as kernel principal component analysis ([16]), kernel Fisher discriminant analysis ([14]), and so on.

Kernel canonical correlation analysis (kernel CCA) has been proposed ([1], [13], [3]) as a nonlinear extension of canonical correlation analysis. Given two random variables X and Y, kernel CCA aims at extracting the information which is shared by the two random variables. More precisely, the purpose of kernel CCA is to provide nonlinear mappings $f(X)$ and $g(Y)$

*Institute of Statistical Mathematics, Japan. E-mail: fukumizu@ism.ac.jp

†Ecole National des Mines de Paris, France

‡Max Planck Institute for Biological Cybernetics, Germany

in RKHS such that their correlation is maximized. Kernel CCA have been successfully applied to various practical problems for extracting nonlinear relations of variables ([19], [10]).

As in many statistical methods, the desired functions given by population are in practice estimated from a finite sample. Thus, the convergence of the estimated functions to the population ones with increasing sample size is very important to justify the method. Since the goal of kernel CCA is to estimate a pair of functions, the convergence should be evaluated in an appropriate functional norm: thus we need tools from functional analysis to characterize the type of convergence.

The purpose of this paper is to rigorously prove the statistical consistency of kernel CCA and a related method. The latter uses a NOrmalized Cross-Covariance Operator, and we call it NOCCO for short. Both kernel CCA and NOCCO require a regularization coefficient, which is similar to Tikhonov regularization ([9]), to enforce smoothness of the functions in the finite sample case (thus avoiding a trivial solution) and to enable operation inversion, but the decay of this regularization with increased sample size has not yet been established. The main theorems in this paper give a sufficient condition on the decay of the regularization coefficient for the finite sample estimators converge to the desired functions in the population limit.

Another important issue in establishing the convergence is an appropriate distance measure for functions. For NOCCO, we obtain the convergence in the norm of RKHS. This result is very strong: if the positive definite kernels are continuous and bounded, the norm is stronger than the uniform norm in the space of continuous functions, and thus the estimated functions converge uniformly to the desired ones. For kernel CCA, we show the convergence in L_2 norm, which is a standard distance measure for functions.

There have been some relevant works on nonlinear extension of canonical correlation analysis. One of them is constrained covariance (COCO, [8]), which uses a different normalization of covariance. Another one is the nonlinear CCA for curves, which are represented by stochastic processes on an interval ([12]). The latter work includes also a consistency result. We will also discuss the relation between our results and these studies.

This paper is organized as follows. Section 2 reviews kernel CCA and related methods, and formulates them in terms of cross-covariance operators, which are basic tools to analyze correlation problems in functional spaces. In Section 3, we describe two main theorems, which show the convergence of kernel CCA and NOCCO. Section 4 is devoted to the proof of the main theorems. Some basic facts from functional analysis and general lemmas are summarized in Appendix.

2 Kernel Canonical Correlation Analysis

In this section, we briefly review the kernel CCA following Bach and Jordan ([3]), and reformulate it with covariance operators on RKHS. For the detail of positive definite kernels and RKHS, see Aronszajn ([2]).

In this paper, a Hilbert space means a separable Hilbert space, and an operator always means a linear operator. The operator norm of a bounded operator T is denoted by $\|T\|$. The null space and the range of an operator T are denoted by $\mathcal{N}(T)$ and $\mathcal{R}(T)$, respectively.

Throughout this paper, $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$ are measurable spaces, and $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$ and $(\mathcal{H}_{\mathcal{Y}}, k_{\mathcal{Y}})$ are reproducing kernel Hilbert spaces (RKHS) of functions on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with measurable positive definite kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$. We consider a random vector $(X,Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$ with the law P_{XY}. The marginal distribution of X and Y are denoted by P_X and P_Y, respectively. It is always assumed that the positive definite kernels satisfy

$$E_X[k_{\mathcal{X}}(X,X)] < \infty \quad \text{and} \quad E_Y[k_{\mathcal{Y}}(Y,Y)] < \infty. \tag{1}$$

Note that under this assumption $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ are continuously included in $L_2(P_X)$ and $L_2(P_Y)$, respectively, where $L_2(\mu)$ denotes the Hilbert space of square integrable functions with respect to the measure μ. This is easily verified by $E_X[f(X)^2] = E_X[\langle f, k_{\mathcal{X}}(\,\cdot\,,X)\rangle^2] \le E_X[\|f\|^2_{\mathcal{H}_{\mathcal{X}}} \|k_{\mathcal{X}}(\,\cdot\,,X)\|^2_{\mathcal{H}_{\mathcal{X}}}] = \|f\|^2_{\mathcal{H}_{\mathcal{X}}} E_X[k_{\mathcal{X}}(X,X)]$ for $f \in \mathcal{H}_{\mathcal{X}}$.

2.1 CCA in reproducing kernel Hilbert spaces

Classical CCA is the method of providing the linear mappings a^TX and b^TY that achieve maximum correlation. Kernel CCA extends this approach by looking for functions $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$ such that the random variables $f(X)$ and $g(Y)$ have maximal correlation. More precisely, the kernel CCA solves the following problem:

$$\max_{\substack{f\in\mathcal{H}_{\mathcal{X}}, g\in\mathcal{H}_{\mathcal{Y}} \\ f\neq 0, g\neq 0}} \frac{\mathrm{Cov}[f(X), g(Y)]}{\mathrm{Var}[f(X)]^{1/2}\mathrm{Var}[g(Y)]^{1/2}}. \tag{2}$$

In practice, we have to estimate the desired function from a finite sample. Given i.i.d. sample $(X_1,Y_1), \ldots, (X_n,Y_n)$ from distribution P_{XY}, an empirical solution of Eq. (2) is given by

$$\max_{\substack{f\in\mathcal{H}_{\mathcal{X}}, g\in\mathcal{H}_{\mathcal{Y}} \\ f\neq 0, g\neq 0}} \frac{\widehat{\mathrm{Cov}}[f(X), g(Y)]}{\big(\widehat{\mathrm{Var}}[f(X)] + \varepsilon_n\|f\|^2_{\mathcal{H}_{\mathcal{X}}}\big)^{1/2}\big(\widehat{\mathrm{Var}}[g(Y)] + \varepsilon_n\|g\|^2_{\mathcal{H}_{\mathcal{Y}}}\big)^{1/2}}, \tag{3}$$

where

$$\begin{aligned}\widehat{\mathrm{Cov}}[f(X),g(Y)] &= \frac{1}{n}\sum_{i=1}^{n}\Big(f(X_i)-\frac{1}{n}\textstyle\sum_{j=1}^{n}f(X_j)\Big)\Big(g(Y_i)-\frac{1}{n}\textstyle\sum_{j=1}^{n}g(Y_j)\Big),\\ \widehat{\mathrm{Var}}[f(X)] &= \frac{1}{n}\sum_{i=1}^{n}\Big(f(X_i)-\frac{1}{n}\textstyle\sum_{j=1}^{n}f(X_j)\Big)^2,\\ \widehat{\mathrm{Var}}[g(Y)] &= \frac{1}{n}\sum_{i=1}^{n}\Big(g(Y_i)-\frac{1}{n}\textstyle\sum_{j=1}^{n}g(Y_j)\Big)^2,\end{aligned}$$

and a positive constant ε_n is the regularization coefficient. As we shall see, the regularization terms $\varepsilon_n\|f\|^2_{\mathcal{H}_\mathcal{X}}$ and $\varepsilon_n\|g\|^2_{\mathcal{H}_\mathcal{Y}}$ make the problem well-formulated statistically, enforce smoothness, and enable operator inversion, as in Tikhonov regularization ([9]). For this smoothing effect, see also the discussion in Section 3, Leurgans et al. ([12]).

2.2 Cross-covariance operators on RKHS

The kernel CCA and related methods can be formulated by cross-covariance operators, which make theoretical analysis easier. Cross-covariance operators are used also to define practical methods for dependence of variables ([6], [7]). This subsection explains the basic properties of cross-covariance operators. For more details see [4], [6], and [7]. The *cross-covariance operator* [1] of (X,Y) is an operator from $\mathcal{H}_\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$, which is defined by

$$\langle g,\Sigma_{YX}f\rangle_{\mathcal{H}_\mathcal{Y}} = E_{XY}\big[(f(X)-E_X[f(X)])(g(Y)-E_Y[g(Y)])\big] \quad (=\mathrm{Cov}[f(X),g(Y)])$$

for all $f\in\mathcal{H}_\mathcal{X}$ and $g\in\mathcal{H}_\mathcal{Y}$. It is easy to see that the right hand side of the above equation is a bounded bilinear function on $\mathcal{H}_\mathcal{X}\times\mathcal{H}_\mathcal{Y}$. By Riesz's representation theorem, a bounded operator Σ_{YX} exists uniquely. The cross-covariance operator expresses all the covariance given by functions in RKHS as the bilinear functional. Thus, it contains all the information on relevance of X and Y expressed by the nonlinear functions in RKHS.

Obviously, $\Sigma_{YX}=\Sigma^*_{XY}$, where T^* denotes the adjoint of an operator T. If Y is equal to X, in particular, the self-adjoint operator Σ_{XX} is called the *covariance operator*. Note that $f\in\mathcal{N}(\Sigma_{XX})$ if and only if $\mathrm{Var}_X[f(X)]=0$. The null space $\mathcal{N}(\Sigma_{XX})$ is equal to $\{f\in\mathcal{H}_\mathcal{X}\mid f(X)=\text{constant almost everywhere}\}$. If there exists a probability density function

[1] Cross-covariance operator can be defined for Banach spaces, in general [4]. However, we confine our discussion on reproducing kernel Hilbert spaces.

for P_X and it is positive for all $x \in \mathcal{X}$, then $\mathcal{N}(\Sigma_{XX})$ is at most one-dimensional.

The expectation element $m_X \in \mathcal{H}_\mathcal{X}$ with respect to a random variable X is defined by

$$\langle f, m_X \rangle_{\mathcal{H}_\mathcal{X}} = E_X[f(X)] = E_X[\langle f, k_\mathcal{X}(\cdot, X)\rangle_{\mathcal{H}_\mathcal{X}}] \qquad (\forall f \in \mathcal{H}_\mathcal{X}). \tag{4}$$

The existence and uniqueness of m_X is proved again by Riesz's representation theorem. Using the expectation elements, the characterization of the covariance operator Σ_{YX} is rewritten by

$$\langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} = E_{XY}[\langle f, k_\mathcal{X}(\cdot, X) - m_X\rangle_{\mathcal{H}_\mathcal{X}} \langle k_\mathcal{Y}(\cdot, Y) - m_Y, g\rangle_{\mathcal{H}_\mathcal{Y}}].$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. random vectors on $\mathcal{X} \times \mathcal{Y}$ with distribution P_{XY}. The *empirical cross-covariance operator* $\widehat{\Sigma}^{(n)}_{YX}$ is defined by the cross-covariance operator with the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}\delta_{Y_i}$. By definition, for any $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$, the operator $\widehat{\Sigma}^{(n)}_{YX}$ gives the empirical covariance as follows;

$$\langle g, \widehat{\Sigma}^{(n)}_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} = \widehat{\mathrm{Cov}}[f(X), G(Y)]$$

Let Q_X and Q_Y be the orthogonal projection which maps $\mathcal{H}_\mathcal{X}$ onto $\overline{\mathcal{R}(\Sigma_{XX})}$ and $\mathcal{H}_\mathcal{Y}$ onto $\overline{\mathcal{R}(\Sigma_{YY})}$, respectively. It is known ([4], Theorem 1) that Σ_{YX} has a representation

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}, \tag{5}$$

where $V_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ is a unique bounded operator such that $\|V_{YX}\| \leq 1$ and $V_{YX} = Q_Y V_{YX} Q_X$. We often write V_{YX} by $\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ by abuse of notation, even when $\Sigma_{XX}^{-1/2}$ or $\Sigma_{YY}^{-1/2}$ are not appropriately defined as operators.

2.3 Representation of kernel CCA and related methods with cross-covariance operators

With cross-covariance operators for (X, Y), the kernel CCA problem can be formulated by

$$\sup_{f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y}} \langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} \quad \text{subject to} \quad \begin{cases} \langle f, \Sigma_{XX} f\rangle_{\mathcal{H}_\mathcal{X}} = 1, \\ \langle g, \Sigma_{YY} g\rangle_{\mathcal{H}_\mathcal{Y}} = 1. \end{cases} \tag{6}$$

As with classical CCA, the solution of the above kernel CCA problem is given by the eigenfunctions corresponding to the largest eigenvalue of the following generalized eigenproblem:

$$\begin{pmatrix} O & \Sigma_{XY} \\ \Sigma_{YX} & O \end{pmatrix} \begin{pmatrix} f \\ g \end{pmatrix} = \rho_1 \begin{pmatrix} \Sigma_{XX} & O \\ O & \Sigma_{YY} \end{pmatrix} \begin{pmatrix} f \\ g \end{pmatrix}. \tag{7}$$

For i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, the empirical estimator in Eq. (3) is represented by

$$\sup_{f \in \mathcal{H}_{\mathcal{X}}, g \in \mathcal{H}_{\mathcal{Y}}} \langle g, \widehat{\Sigma}_{YX}^{(n)} f \rangle_{\mathcal{H}_{\mathcal{Y}}} \quad \text{subject to} \quad \begin{cases} \langle f, (\widehat{\Sigma}_{XX}^{(n)} + \varepsilon_n I) f \rangle_{\mathcal{H}_{\mathcal{X}}} = 1, \\ \langle g, (\widehat{\Sigma}_{YY}^{(n)} + \varepsilon_n I) g \rangle_{\mathcal{H}_{\mathcal{Y}}} = 1, \end{cases} \tag{8}$$

and the substitute of Eq. (7) is

$$\begin{pmatrix} O & \widehat{\Sigma}_{XY}^{(n)} \\ \widehat{\Sigma}_{YX}^{(n)} & O \end{pmatrix} \begin{pmatrix} f \\ g \end{pmatrix} = \widehat{\rho}_1^{(n)} \begin{pmatrix} \widehat{\Sigma}_{XX}^{(n)} + \varepsilon_n I & O \\ O & \widehat{\Sigma}_{YY}^{(n)} + \varepsilon_n I \end{pmatrix} \begin{pmatrix} f \\ g \end{pmatrix}. \tag{9}$$

Let us assume that the operator V_{YX} given by Eq. (5) is compact[2], and let ϕ and ψ be the unit eigenfunctions of V_{YX} corresponding to the largest singular value; that is,

$$\langle \psi, V_{YX} \phi \rangle_{\mathcal{H}_{\mathcal{Y}}} = \max_{\substack{f \in \mathcal{H}_{\mathcal{X}}, g \in \mathcal{H}_{\mathcal{Y}} \\ \|f\|_{\mathcal{H}_{\mathcal{X}}} = \|g\|_{\mathcal{H}_{\mathcal{Y}}} = 1}} \langle g, V_{YX} f \rangle_{\mathcal{H}_{\mathcal{Y}}}. \tag{10}$$

Given that $\phi \in \mathcal{R}(\Sigma_{XX})$ and $\psi \in \mathcal{R}(\Sigma_{YY})$, it is easy to see from Eq. (7) that the solution of the kernel CCA is

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

In the empirical case, let $\widehat{\phi}_n \in \mathcal{H}_{\mathcal{X}}$ and $\widehat{\psi}_n \in \mathcal{H}_{\mathcal{Y}}$ be the unit eigenfunctions corresponding to the largest singular value of the finite rank operator

$$\widehat{V}_{YX}^{(n)} := \big(\widehat{\Sigma}_{YY}^{(n)} + \varepsilon_n I\big)^{-1/2} \widehat{\Sigma}_{YX}^{(n)} \big(\widehat{\Sigma}_{XX}^{(n)} + \varepsilon_n I\big)^{-1/2}.$$

From Eq. (9), the empirical estimators $\widehat{f}_n$ and $\widehat{g}_n$ of kernel CCA are equal to

$$\widehat{f}_n = (\widehat{\Sigma}_{XX}^{(n)} + \varepsilon_n I)^{-1/2} \widehat{\phi}_n, \qquad \widehat{g}_n = (\widehat{\Sigma}_{YY}^{(n)} + \varepsilon_n I)^{-1/2} \widehat{\psi}_n.$$

Note that all the empirical operators and the estimators described above can be expressed by *Gram matrices*. The solutions $\widehat{f}_n$ and $\widehat{g}_n$ are exactly

[2] See Appendix A for compact operators.

the same as the those given in Bach and Jordan ([3]). We now confirm it by rewriting $\widehat{V}_{YX}^{(n)}$ with Gram matrices. Let $u_i \in \mathcal{H}_\mathcal{X}$ and $v_i \in \mathcal{H}_\mathcal{Y}$ $(1 \leq i \leq n)$ be functions defined by

$$u_i = k_\mathcal{X}(\cdot, X_i) - \frac{1}{n}\sum_{j=1}^n k_\mathcal{X}(\cdot, X_j), \qquad v_i = k_\mathcal{Y}(\cdot, Y_i) - \frac{1}{n}\sum_{j=1}^n k_\mathcal{Y}(\cdot, Y_j).$$

Because $\mathcal{R}(\widehat{\Sigma}_{XX}^{(n)})$ and $\mathcal{R}(\widehat{\Sigma}_{YY}^{(n)})$ are spanned by $(u_i)_{i=1}^n$ and $(v_i)_{i=1}^n$, respectively, the eigenfunctions of $\widehat{V}_{YX}^{(n)}$ are given by a linear combination of u_i and v_i. Letting $\phi = \sum_{i=1}^n \alpha_i u_i$ and $\psi = \sum_{i=1}^n \beta_i v_i$, direct calculation of $\langle \psi, \widehat{V}_{YX}^{(n)}\phi\rangle_{\mathcal{H}_\mathcal{Y}}$ shows that the solutions $\widehat{\phi}_n$ and $\widehat{\psi}_n$ of NOCCO are given by the coefficients $\widehat{\alpha}$ and $\widehat{\beta}$ that achieve

$$\max_{\substack{\alpha,\beta\in\mathbb{R}^n \\ \alpha^T G_X \alpha = \beta^T G_Y \beta = 1}} \beta^T \big(G_Y + n\varepsilon_n I_n\big)^{-1/2} G_Y G_X \big(G_X + n\varepsilon_n I_n\big)^{-1/2}\alpha,$$

where G_X is the centralized Gram matrix defined by

$$(G_X)_{ij} = k_\mathcal{X}(X_i, X_j) - \frac{1}{n}\sum_{b=1}^n k_\mathcal{X}(X_i, X_b) - \frac{1}{n}\sum_{a=1}^n k_\mathcal{X}(X_a, X_j) + \frac{1}{n^2}\sum_{a=1}^n\sum_{b=1}^n k_\mathcal{X}(X_a, X_b)$$

and G_Y defined accordingly. The solution of kernel CCA problem is given by

$$\widehat{f}_n = (\widehat{\Sigma}_{XX}^{(n)} + \varepsilon_n I)^{-1/2}\widehat{\phi}_n = \sum_{i=1}^n \widehat{\xi}_i u_i, \qquad \widehat{g}_n = (\widehat{\Sigma}_{YY}^{(n)} + \varepsilon_n I)^{-1/2}\widehat{\psi}_n = \sum_{i=1}^n \widehat{\zeta}_i v_i,$$

where

$$\widehat{\xi} = \sqrt{n}(G_X + n\varepsilon_n I_n)^{-1/2}\widehat{\alpha} \quad \text{and} \quad \widehat{\zeta} = \sqrt{n}(G_Y + n\varepsilon_n I_n)^{-1/2}\widehat{\beta}.$$

Thus, the linear coefficients $\widehat{\xi}$ and $\widehat{\zeta}$ are the solution of

$$\max_{\substack{\xi,\zeta\in\mathbb{R}^n \\ \xi^T(G_X^2 + n\varepsilon_n G_X)\xi = \zeta^T(G_Y^2 + n\varepsilon_n G_Y)\zeta = n}} \zeta^T G_Y G_X \xi,$$

which is exactly the same as the one proposed in Bach and Jordan ([3]). Note that they approximate $(G_X^2 + n\varepsilon_n G_X)$ by $(G_X + \frac{n\varepsilon}{2}I_n)^2$ for computational simplicity.

There are additional, related methods to extract nonlinear dependence of two random variables. The Constrained Covariance (COCO, [8]) uses the unit eigenfunctions of the cross-covariance operator Σ_{YX}. Thus the solution of COCO is

$$\max_{\substack{f\in\mathcal{H}_\mathcal{X},g\in\mathcal{H}_\mathcal{Y}\\ \|f\|_{\mathcal{H}_\mathcal{X}}=\|g\|_{\mathcal{H}_\mathcal{Y}}=1}} \langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} = \max_{\substack{f\in\mathcal{H}_\mathcal{X},g\in\mathcal{H}_\mathcal{Y}\\ \|f\|_{\mathcal{H}_\mathcal{X}}=\|g\|_{\mathcal{H}_\mathcal{Y}}=1}} \mathrm{Cov}[f(X), g(Y)].$$

The consistency of COCO has been proved in [7]. Instead of normalizing the covariance by the variances, COCO normalizes the covariance by the RKHS norms of f and g. Kernel CCA is a more direct nonlinear extension of the ordinary CCA than COCO. COCO tends to find functions with large variance for $f(X)$ and $g(Y)$, which may not be the most correlated features. On the other hand, kernel CCA may encounter situations where it finds functions with moderately large covariance but very small variances for $f(X)$ or $g(Y)$, since Σ_{XX} and Σ_{YY} can have arbitrarily small eigenvalues.

A possible compromise of these methods is to use ϕ and ψ in Eq. (10), and their estimates $\widehat{\phi}_n$ and $\widehat{\psi}_n$. While the statistical meaning of this method is not as direct as the kenrel CCA, it can incorporate the normalization by Σ_{XX} and Σ_{YY}. We call this variant *NOrmalized Cross-Covariance Operator* (NOCCO). We will establish the consistency of kernel CCA and NOCCO in the next section.

3 Main theorems

First, the following theorem asserts the consistency of the estimator of NOCCO in the RKHS norm.

Theorem 1. *Let $(\varepsilon_n)_{n=1}^\infty$ be a sequence of positive numbers such that*

$$\lim_{n\to\infty} \varepsilon_n = 0, \qquad \lim_{n\to\infty} \frac{n^{-1/3}}{\varepsilon_n} = 0. \tag{11}$$

Assume V_{YX} is a compact operator and the eigenspaces which attain the singular value problem

$$\max_{\substack{\phi\in\mathcal{H}_\mathcal{X},\psi\in\mathcal{H}_\mathcal{Y}\\ \|\phi\|_{\mathcal{H}_\mathcal{X}}=\|\psi\|_{\mathcal{H}_\mathcal{Y}}=1}} \langle \psi, V_{YX}\phi\rangle_{\mathcal{H}_\mathcal{Y}}$$

are one-dimensional. Let $\widehat{\phi}_n$ and $\widehat{\psi}_n$ be the unit eigenfunctions for the largest singular value of $\widehat{V}_{YX}^{(n)}$. Then,

$$|\langle \widehat{\phi}_n, \phi\rangle_{\mathcal{H}_\mathcal{X}}| \to 1, \qquad |\langle \widehat{\psi}_n, \psi\rangle_{\mathcal{H}_\mathcal{Y}}| \to 1$$

in probability, as n goes to infinity.

The next main result shows the convergence of kernel CCA in the norm of $L_2(P_X)$ and $L_2(P_Y)$.

Theorem 2. *Let $(\varepsilon_n)_{n=1}^{\infty}$ be a sequence of positive numbers which satisfies Eq. (11). Assume that ϕ and ψ are included in $\mathcal{R}(\Sigma_{XX})$ and $\mathcal{R}(\Sigma_{YY})$, respectively, and that V_{YX} is compact. Then,*

$$\left\|(\widehat{f}_n - E_X[\widehat{f}_n(X)]) - (f - E_X[f(X)])\right\|_{L_2(P_X)} \to 0$$

and

$$\left\|(\widehat{g}_n - E_Y[\widehat{g}_n(Y)]) - (g - E_Y[g(Y)])\right\|_{L_2(P_Y)} \to 0$$

in probability, as n goes to infinity.

While we restrict our attention on the first eigenfunctions, it is not difficult to see the convergence of eigenspaces corresponding to the m-th largest eigenvalue by extending Lemma 9 in Appendix.

The convergence of NOCCO in RKHS norm is a very strong result. If $\mathcal{X}$ and $\mathcal{Y}$ are topological space, and if the kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are continuous and bounded, all the functions in $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ are continuous and the RKHS norm is stronger than the uniform norm in $C(\mathcal{X})$ and $C(\mathcal{Y})$, where $C(\mathcal{Z})$ is the Banach space of all the continuous functions on a topological space $\mathcal{Z}$ with the supremum norm. In fact, for any $f \in \mathcal{H}_{\mathcal{X}}$, we have $\sup_{x\in\mathcal{X}}|f(x)| = \sup_{x\in\mathcal{X}}|\langle k_{\mathcal{X}}(\cdot,x), f\rangle_{\mathcal{H}_{\mathcal{X}}}| \leq \sup_{x\in\mathcal{X}}(k_{\mathcal{X}}(x,x))^{1/2}\|f\|_{\mathcal{H}_{\mathcal{X}}}$. In such cases, Theorem 1 implies $\widehat{\phi}_n$ and $\widehat{\psi}_n$ converge uniformly to ϕ and ψ, respectively. This uniform convergence is useful in practice, because in many applications the function value at each point is important.

For any complete orthonormal systems (CONS) $\{\phi_i\}_{i=1}^{\infty}$ of $\mathcal{H}_{\mathcal{X}}$ and $\{\psi_i\}_{i=1}^{\infty}$ of $\mathcal{H}_{\mathcal{Y}}$, the compactness assumption on V_{YX} in the above theorems requires that the correlation of $\Sigma_{XX}^{-1/2}\phi_i(X)$ and $\Sigma_{YY}^{-1/2}\psi_i(Y)$ decays to zero as $i \to \infty$. This is not necessarily satisfied in general. A trivial example is the case of variables with $Y = X$, in which $V_{YX} = I$ is not compact. In this case, the problem in Theorem 1 is solved by an arbitrary function. Moreover, the kernel CCA problem in Theorem 2 does not have solutions, if Σ_{XX} has arbitrarily small eigenvalues.

Leurgans et al. ([12]) discuss canonical correlation analysis on curves, which are represented by stochastic processes on an interval, and use the Sobolev space of functions with square integrable second derivative. Since the Sobolev space is a RKHS, their method is an example of kernel CCA in a specific RKHS. They also prove the consistency of estimators under the condition $n^{-1/2}/\varepsilon_n \to 0$. Although the proof can be extended to a general RKHS, the convergence is measured by that of the correlation,

$$\frac{\left|\langle \widehat{f}_n, \Sigma_{XX} f\rangle_{\mathcal{H}_\mathcal{X}}\right|}{\left(\langle \widehat{f}_n, \Sigma_{XX}\widehat{f}_n\rangle_{\mathcal{H}_\mathcal{X}}\right)^{1/2}\left(\langle f, \Sigma_{XX} f\rangle_{\mathcal{H}_\mathcal{X}}\right)^{1/2}} \quad \to \quad 1,$$

which is weaker than the L_2 convergence in Theorem 2. In fact, since the desired eigenfunction f is normalized so that $\langle f, \Sigma_{XX} f\rangle_{\mathcal{H}_\mathcal{X}} = 1$, from Theorem 2 it is easy to derive the above convergence of correlation. On the other hand, the convergence of correlation does not imply $\langle (\widehat{f}_n - f), \Sigma_{XX}(\widehat{f}_n - f)\rangle_{\mathcal{H}_\mathcal{X}}$. From the equality

$$\begin{aligned}\langle (\widehat{f}_n - f), \Sigma_{XX}(\widehat{f}_n - f)\rangle_{\mathcal{H}_\mathcal{X}} &= \left(\langle \widehat{f}_n, \Sigma_{XX}\widehat{f}_n\rangle_{\mathcal{H}_\mathcal{X}} - \langle f, \Sigma_{XX} f\rangle_{\mathcal{H}_\mathcal{X}}\right)^2 \\ &\quad + 2\left(1 - \frac{\langle \widehat{f}_n, \Sigma_{XX} f\rangle_{\mathcal{H}_\mathcal{X}}}{\|\Sigma_{XX}^{1/2}\widehat{f}_n\|_{\mathcal{H}_\mathcal{X}}\|\Sigma_{XX}^{1/2} f\|_{\mathcal{H}_\mathcal{X}}}\right)\|\Sigma_{XX}^{1/2}\widehat{f}_n\|_{\mathcal{H}_\mathcal{X}}\|\Sigma_{XX}^{1/2} f\|_{\mathcal{H}_\mathcal{X}},\end{aligned}$$

we require the convergence $\langle \widehat{f}_n, \Sigma_{XX}\widehat{f}_n\rangle_{\mathcal{H}_\mathcal{X}} \to \langle f, \Sigma_{XX} f\rangle_{\mathcal{H}_\mathcal{X}} = 1$ in order to guarantee the left hand side to converge to zero. However, with the normalization $\langle \widehat{f}_n, (\widehat{\Sigma}_{XX}^{(n)} + \varepsilon_n I)\widehat{f}_n\rangle_{\mathcal{H}_\mathcal{X}} = \langle f, \Sigma_{XX} f\rangle_{\mathcal{H}_\mathcal{X}} = 1$, the convergence of $\langle \widehat{f}_n, \Sigma_{XX}\widehat{f}_n\rangle_{\mathcal{H}_\mathcal{X}}$ is not clear. We use the assumption $n^{-1/3}/\varepsilon_n \to 0$ to prove $\langle (\widehat{f}_n - f), \Sigma_{XX}(\widehat{f}_n - f)\rangle_{\mathcal{H}_\mathcal{X}}$ in Theorem 2.

4 Proof of the main theorems

4.1 Hilbert-Schmidt norm of covariance operators

As preliminaries to the proof of main theorems, in this subsection we show some results on the Hilbert-Schmidt norm of cross-covariance operators. For convenience, we provide in Appendix the definition and some basic properties of Hilbert-Schmidt operators. See also [7].

In describing the result, we use the notion of random elements in a Hilbert space ([18], [4]). Let $\mathcal{H}$ be a Hilbert space equipped with Borel σ-field. A *random element* in the Hilbert space $\mathcal{H}$ is a measurable map $F : \Omega \to \mathcal{H}$ from a measurable space $(\Omega, \mathfrak{S})$. Let $\mathcal{H}$ be a RKHS on a measurable set $\mathcal{X}$ with a measurable positive definite kernel k. For a random variable X in $\mathcal{X}$, the map $k(\cdot, X)$ defines a random element in $\mathcal{H}$.

A random element F in a Hilbert space $\mathcal{H}$ is said to have *strong order* p $(0 < p < \infty)$ if $E\|F\|^p$ is finite. For a random element F of strong order one, the expectation of F is defined as the element m_F in $\mathcal{H}$ such that

$$\langle m_F, g\rangle_{\mathcal{H}} = E[\langle F, g\rangle_{\mathcal{H}}]$$

holds for all $g \in \mathcal{H}$. The existence and the uniqueness is proved by Riesz's representation theorem. The expectation m_F is denoted by $E[F]$. Then, the equality $\langle E[F], g\rangle_{\mathcal{H}} = E[\langle F, g\rangle_{\mathcal{H}}]$ is justified, which means the expectation and the inner product are interchangeable. If F and G have strong order two, $\langle F, G\rangle_{\mathcal{H}}$ is integrable. If further F and G are independent, the relation

$$E[\langle F, G\rangle_{\mathcal{H}}] = \langle E[F], E[G]\rangle_{\mathcal{H}} \tag{12}$$

holds.

It is easy to see that the above example $F = k(\cdot, X)$ in a RKHS $\mathcal{H}$ has strong order two, i.e. $E[\|F\|^2] < \infty$, under the assumption $E[k(X, X)] < \infty$. The expectation of $k(\cdot, X)$ is equal to m_X in Eq. (4) by definition. For two RKHS $\mathcal{H}_{\mathcal{X}}$ on $\mathcal{X}$ and $\mathcal{H}_{\mathcal{Y}}$ on $\mathcal{Y}$ with kernel $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$, respectively, under the condition Eq. (1), the random element $k_{\mathcal{X}}(\cdot, X)k_{\mathcal{Y}}(\cdot, Y)$ in the direct product $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ has strong order one.

The following lemma is straightforward from Lemma 1 in [7] and Eq. (12).

Lemma 3. *The cross-covariance operator Σ_{YX} is a Hilbert-Schmidt operator. Moreover, the Hilbert-Schmidt norm is given by*

$$\begin{aligned}
&\|\Sigma_{YX}\|_{HS}^2 \\
&= E_{YX}E_{\tilde{Y}\tilde{X}}\big[\langle k_{\mathcal{X}}(\cdot, X) - m_X, k_{\mathcal{X}}(\cdot, \tilde{X}) - m_X\rangle_{\mathcal{H}_{\mathcal{X}}}\langle k_{\mathcal{Y}}(\cdot, \tilde{Y}) - m_Y, k_{\mathcal{Y}}(\cdot, Y) - m_Y\rangle_{\mathcal{H}_{\mathcal{Y}}}\big] \\
&= \big\|E_{YX}[(k_{\mathcal{X}}(\cdot, X) - m_X)(k_{\mathcal{Y}}(\cdot, Y) - m_Y)]\big\|_{\mathcal{H}_{\mathcal{X}}\otimes\mathcal{H}_{\mathcal{Y}}}^2
\end{aligned} \tag{13}$$

where $(\tilde{X}, \tilde{Y})$ and (X, Y) are independently and identically distributed with distribution P_{XY}.

From the facts $\mathcal{H}_{\mathcal{X}} \subset L_2(P_X)$ and $\mathcal{H}_{\mathcal{Y}} \subset L_2(P_Y)$, the law of large numbers implies for each $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$

$$\lim_{n\to\infty} \langle g, \widehat{\Sigma}_{YX}^{(n)} f\rangle_{\mathcal{H}_{\mathcal{Y}}} = \langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_{\mathcal{Y}}}$$

in probability. Moreover, the central limit theorem shows the above convergence is of $O_p(n^{-1/2})$. The following lemma shows a tight uniform result saying that $\|\widehat{\Sigma}_{YX}^{(n)} - \Sigma_{YX}\|_{HS}$ converges to zero in the order of $O_p(n^{-1/2})$.

Lemma 4.

$$\big\|\widehat{\Sigma}_{YX}^{(n)} - \Sigma_{YX}\big\|_{HS} = O_p(n^{-1/2}) \quad (n \to \infty).$$

Proof. Write for simplicity $F = k_{\mathcal{X}}(\cdot, X) - E_X[k_{\mathcal{X}}(\cdot, X)]$, $G = k_{\mathcal{Y}}(\cdot, Y) - E_Y[k_{\mathcal{Y}}(\cdot, Y)]$, $F_i = k_{\mathcal{X}}(\cdot, X_i) - E_X[k_{\mathcal{X}}(\cdot, X)]$, $G_i = k_{\mathcal{Y}}(\cdot, Y_i) - E_Y[k_{\mathcal{Y}}(\cdot, Y)]$, and $\mathcal{F} = \mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$. Then, $F, F_1, \ldots, F_n$ are i.i.d. random elements in $\mathcal{H}_\mathcal{X}$, and a similar fact holds for $G, G_1, \ldots, G_n$ also. Lemma 3 implies

$$\|\widehat{\Sigma}_{YX}^{(n)}\|_{HS}^2 = \Big\| \frac{1}{n}\sum_{i=1}^n \Big(F_i - \frac{1}{n}\sum_{j=1}^n F_j\Big)\Big(G_i - \frac{1}{n}\sum_{j=1}^n G_j\Big)\Big\|_{\mathcal{F}}^2,$$

and the same argument as the proof of Lemma 3 shows

$$\langle \Sigma_{YX}, \widehat{\Sigma}_{YX}^{(n)} \rangle_{HS} = \Big\langle E[FG], \frac{1}{n}\sum_{i=1}^n \Big(F_i - \frac{1}{n}\sum_{j=1}^n F_j\Big)\Big(G_i - \frac{1}{n}\sum_{j=1}^n G_j\Big)\Big\rangle_{\mathcal{F}}.$$

From these equations, we have

$$\begin{aligned}
\big\|\widehat{\Sigma}_{YX}^{(n)} - \Sigma_{YX}\big\|_{HS}^2 &= \big\|\Sigma_{YX}\big\|_{HS}^2 - 2\big\langle \Sigma_{YX}, \widehat{\Sigma}_{YX}^{(n)}\big\rangle_{HS} + \big\|\widehat{\Sigma}_{YX}^{(n)}\big\|_{HS}^2 \\
&= \Big\| \frac{1}{n}\sum_{i=1}^n \Big(F_i - \frac{1}{n}\sum_{j=1}^n F_j\Big)\Big(G_i - \frac{1}{n}\sum_{j=1}^n G_j\Big) - E[FG]\Big\|_{\mathcal{F}}^2 \\
&= \Big\| \frac{1}{n}\Big(1 - \frac{1}{n}\Big)\sum_{i=1}^n F_i G_i - \frac{1}{n^2}\sum_{i=1}^n \sum_{j \neq i} F_i G_j - E[FG]\Big\|_{\mathcal{F}}^2.
\end{aligned}$$

From $\langle FG, \tilde{F}\tilde{G}\rangle_{\mathcal{F}} = \langle F, \tilde{F}\rangle_{\mathcal{H}_\mathcal{X}} \langle G, \tilde{G}\rangle_{\mathcal{H}_\mathcal{Y}}$ and $E[F_i] = E[G_j] = 0$, the following relations hold on the expectations with respect to $(X_1, Y_1), \ldots, (X_n, Y_n)$;

$$\begin{aligned}
&E[\langle F_i G_i, F_k G_k\rangle_{\mathcal{F}}] = \begin{cases} E[\|FG\|_{\mathcal{F}}^2] & \text{for } i = k, \\ \|E[FG]\|_{\mathcal{F}}^2 & \text{for } i \neq k, \end{cases} \\
&E[\langle F_i G_j, F_k G_\ell\rangle_{\mathcal{F}}] = 0 \qquad \text{for } i \neq j \text{ and } \{i, j\} \neq \{k, \ell\}, \\
&E[\langle F_i G_j, F_i G_j\rangle_{\mathcal{F}}] = E[\|F\|_{\mathcal{H}_\mathcal{X}}^2] E[\|G\|_{\mathcal{H}_\mathcal{Y}}^2] \qquad \text{for } i \neq j, \\
&E[\langle F_i G_j, F_j G_i\rangle_{\mathcal{F}}] = \|E[FG]\|_{\mathcal{F}}^2 \qquad \text{for } i \neq j,
\end{aligned}$$

and

$$E\Big\langle \sum_{i=1}^n \sum_{j \neq i} F_i G_j, E[FG]\Big\rangle_{\mathcal{F}} = 0.$$

Using these relations, we obtain

$$
\begin{aligned}
&E\big\|\widehat{\Sigma}_{YX}^{(n)} - \Sigma_{YX}\big\|_{HS}^2 \\
&= \frac{1}{n^2}\Big(1-\frac{1}{n}\Big)^2 \sum_{i=1}^n \sum_{k=1}^n E\langle F_iG_i, F_kG_k\rangle_{\mathcal{F}} + \frac{1}{n^4}\sum_{i=1}^n \sum_{j\neq i} \sum_{k=1}^n \sum_{\ell\neq k} E\langle F_iG_j, F_kG_\ell\rangle_{\mathcal{F}} \\
&\quad + \|E[FG]\|_{\mathcal{F}}^2 - \frac{2}{n^3}\Big(1-\frac{1}{n}\Big) \sum_{i=1}^n \sum_{k=1}^n \sum_{\ell\neq k} E\langle F_iG_i, F_kG_\ell\rangle_{\mathcal{F}} \\
&\quad - \frac{2}{n}\Big(1-\frac{1}{n}\Big) \sum_{i=1}^n E\langle F_iG_i, E[FG]\rangle_{\mathcal{F}} \\
&= \frac{1}{n^2}\Big(1-\frac{1}{n}\Big)^2 \sum_{i=1}^n E\langle F_iG_i, F_iG_i\rangle_{\mathcal{F}} + \frac{1}{n^2}\Big(1-\frac{1}{n}\Big)^2 \sum_{i=1}^n \sum_{k\neq i} E\langle F_iG_i, F_kG_k\rangle_{\mathcal{F}} \\
&\quad + \frac{1}{n^4}\sum_{i=1}^n \sum_{j\neq i} E\langle F_iG_j, F_iG_j\rangle_{\mathcal{F}} + \frac{1}{n^4}\sum_{i=1}^n \sum_{j\neq i} E\langle F_iG_j, F_jG_i\rangle_{\mathcal{F}} \\
&\quad + \|E[FG]\|_{\mathcal{F}}^2 - 0 - \frac{2}{n}\Big(1-\frac{1}{n}\Big) \sum_{i=1}^n E\langle F_iG_i, E[FG]\rangle_{\mathcal{F}} \\
&= \frac{1}{n}\Big(1-\frac{1}{n}\Big)^2 E\|FG\|_{\mathcal{F}}^2 + \Big(1-\frac{1}{n}\Big)^3 \|E[FG]\|_{\mathcal{F}}^2 \\
&\quad + \frac{n-1}{n^3} E[\|F\|_{\mathcal{H}_\mathcal{X}}^2] E[\|G\|_{\mathcal{H}_\mathcal{Y}}^2] + \frac{n-1}{n^3} E\|FG\|_{\mathcal{F}}^2 \\
&\quad + \|E[FG]\|_{\mathcal{F}}^2 - 2\Big(1-\frac{1}{n}\Big)\|E[FG]\|_{\mathcal{F}}^2,
\end{aligned}
$$

from which we see the terms of $O(1)$ are canceled, and

$$
E\big\|\widehat{\Sigma}_{YX}^{(n)} - \Sigma_{YX}\big\|_{HS}^2 = O(1/n).
$$

The proof is completed by Chebyshev's inequality. □

4.2 Preliminary lemmas

We prepare further preliminary lemmas for the proof of the main theorems.

Lemma 5. *Let ε_n be a positive number such that $\varepsilon_n \to 0$ ($n \to \infty$). Then, for i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, the following equality holds;*

$$
\big\|\widehat{V}_{YX}^{(n)} - (\Sigma_{YY} + \varepsilon_n I)^{-1/2}\Sigma_{YX}(\Sigma_{XX} + \varepsilon_n I)^{-1/2}\big\| = O_p(\varepsilon_n^{-3/2} n^{-1/2}).
$$

Proof. The operator in the left hand side is decomposed as

$$\begin{aligned}&\widehat{V}_{YX}^{(n)} - (\Sigma_{YY}+\varepsilon_n I)^{-1/2}\Sigma_{YX}(\Sigma_{XX}+\varepsilon_n I)^{-1/2}\\ &= \big\{(\widehat{\Sigma}_{YY}^{(n)}+\varepsilon_n I)^{-1/2} - (\Sigma_{YY}+\varepsilon_n I)^{-1/2}\big\}\widehat{\Sigma}_{YX}^{(n)}(\widehat{\Sigma}_{XX}^{(n)}+\varepsilon_n I)^{-1/2}\\ &\quad + (\Sigma_{YY}+\varepsilon_n I)^{-1/2}\big\{\widehat{\Sigma}_{YX}^{(n)} - \Sigma_{YX}\big\}(\widehat{\Sigma}_{XX}^{(n)}+\varepsilon_n I)^{-1/2}\\ &\quad + (\Sigma_{YY}+\varepsilon_n I)^{-1/2}\Sigma_{YX}\big\{(\widehat{\Sigma}_{XX}^{(n)}+\varepsilon_n I)^{-1/2} - (\Sigma_{XX}+\varepsilon_n I)^{-1/2}\big\}.\end{aligned} \tag{14}$$

From the equality

$$A^{-1/2} - B^{-1/2} = A^{-1/2}\big(B^{3/2}-A^{3/2}\big)B^{-3/2} + (A-B)B^{-3/2},$$

the first term in the right hand side of Eq. (14) is equal to

$$\big\{(\widehat{\Sigma}_{YY}^{(n)}+\varepsilon_n I)^{-1/2}\big(\Sigma_{YY}^{3/2}-\widehat{\Sigma}_{YY}^{(n)3/2}\big)+\big(\widehat{\Sigma}_{YY}^{(n)}-\Sigma_{YY}\big)\big\}\big(\widehat{\Sigma}_{YY}^{(n)}+\varepsilon_n I\big)^{-3/2}\widehat{\Sigma}_{YX}^{(n)}(\widehat{\Sigma}_{XX}^{(n)}+\varepsilon_n I)^{-1/2}.$$

From the facts $\big\|(\widehat{\Sigma}_{YY}^{(n)}+\varepsilon_n I)^{-1/2}\big\| \le \frac{1}{\sqrt{\varepsilon_n}}$, $\big\|\big(\widehat{\Sigma}_{YY}^{(n)}+\varepsilon_n I\big)^{-1/2}\widehat{\Sigma}_{YX}^{(n)}\big(\widehat{\Sigma}_{XX}^{(n)}+\varepsilon_n I\big)^{-1/2}\big\| \le 1$ and Lemma 7 in Appendix, the norm of the above operator is bounded from above by

$$\frac{1}{\varepsilon_n}\Big\{\frac{3}{\sqrt{\varepsilon_n}}\max\big\{\|\Sigma_{YY}\|^{3/2}, \|\widehat{\Sigma}_{YY}^{(n)}\|^{3/2}\big\}+1\Big\}\|\widehat{\Sigma}_{YY}^{(n)}-\Sigma_{YY}\|.$$

A similar bound applies also to the third term of Eq. (14). An upper bound of the second term of Eq. (14) is $\frac{1}{\varepsilon_n}\|\Sigma_{YX}-\widehat{\Sigma}_{YX}^{(n)}\|$. Thus, Lemma 4 and the facts $\|\widehat{\Sigma}_{XX}^{(n)}\| = \|\Sigma_{XX}\| + o_p(1)$, $\|\widehat{\Sigma}_{YY}^{(n)}\| = \|\Sigma_{YY}\| + o_p(1)$ complete the proof. □

Lemma 6. *Assume V_{YX} is compact. Then, for a sequence $\varepsilon_n \to 0$,*

$$\big\|(\Sigma_{YY}+\varepsilon_n I)^{-1/2}\Sigma_{YX}(\Sigma_{XX}+\varepsilon_n I)^{-1/2} - V_{YX}\big\| \ \to\ 0 \quad (n\to\infty).$$

Proof. An upper bound of the left hand side of the assertion is given by

$$\begin{aligned}&\big\|\big\{(\Sigma_{YY}+\varepsilon_n I)^{-1/2} - \Sigma_{YY}^{-1/2}\big\}\Sigma_{YX}(\Sigma_{XX}+\varepsilon_n I)^{-1/2}\big\|\\ &\qquad\qquad + \big\|\Sigma_{YY}^{-1/2}\Sigma_{YX}\big\{(\Sigma_{XX}+\varepsilon_n I)^{-1/2} - \Sigma_{XX}^{-1/2}\big\}\big\|.\end{aligned} \tag{15}$$

The first term of Eq. (15) is equal to

$$\big\|\big\{(\Sigma_{YY}+\varepsilon_n I)^{-1/2}\Sigma_{YY}^{1/2} - I\big\}V_{YX}\big\|. \tag{16}$$

Note that the range of V_{YX} is included in $\overline{\mathcal{R}(\Sigma_{YY})}$, as remarked in Section 2.2. Let v be an arbitrary element in $\mathcal{R}(V_{YX}) \cap \mathcal{R}(\Sigma_{YY})$. Then, there exists $u \in \mathcal{H}_\mathcal{Y}$ such that $v = \Sigma_{YY} u$. Noting that Σ_{YY} and $(\Sigma_{YY} + \varepsilon_n I)^{1/2}$ are commutative, we have

$$\begin{aligned}
&\big\| \big\{ (\Sigma_{YY} + \varepsilon_n I)^{-1/2} \Sigma_{YY}^{1/2} - I \big\} v \big\|_{\mathcal{H}_\mathcal{Y}} \\
&= \big\| \big\{ (\Sigma_{YY} + \varepsilon_n I)^{-1/2} \Sigma_{YY}^{1/2} - I \big\} \Sigma_{YY} u \big\|_{\mathcal{H}_\mathcal{Y}} \\
&= \big\| (\Sigma_{YY} + \varepsilon_n I)^{-1/2} \Sigma_{YY}^{1/2} \big\{ \Sigma_{YY}^{1/2} - (\Sigma_{YY} + \varepsilon_n I)^{1/2} \big\} \Sigma_{YY}^{1/2} u \big\|_{\mathcal{H}_\mathcal{Y}} \\
&\leq \big\| \Sigma_{YY}^{1/2} - (\Sigma_{YY} + \varepsilon_n I)^{1/2} \big\| \, \big\| \Sigma_{YY}^{1/2} u \big\|_{\mathcal{H}_\mathcal{Y}}.
\end{aligned}$$

Since $\Sigma_{YY} + \varepsilon_n I \to \Sigma_{YY}$ in norm means $(\Sigma_{YY} + \varepsilon_n I)^{1/2} \to \Sigma_{YY}^{1/2}$ in norm, the convergence

$$\big\{ (\Sigma_{YY} + \varepsilon_n I)^{-1/2} \Sigma_{YY}^{1/2} - I \big\} v \ \to \ 0 \qquad (n \to \infty) \tag{17}$$

holds for all $v \in \mathcal{R}(V_{YX}) \cap \mathcal{R}(\Sigma_{YY})$. Because V_{YX} is compact, Lemma 8 in Appendix shows Eq. (16) converges to zero. The convergence of the second term in Eq. (15) can be proved similarly. □

4.3 Proof of the main theorems

We are now in the position ready to prove the main theorems.

Proof of Theorem 1. From Lemmas 5 and 6, $\widehat{V}_{YX}^{(n)}$ converges to V_{YX} in norm. Because ϕ and ψ are the eigenfunction of the largest eigenvalue for $V_{YX} V_{XY}$ and $V_{XY} V_{YX}$, respectively, and the similar facts hold for $\widehat{\phi}_n$ and $\widehat{\psi}_n$, the assertion is obtained by Lemma 9 in Appendix. □

Proof of Theorem 2. We show only the convergence of $\widehat{f}_n$. Without loss of generality, we can assume $\widehat{\phi}_n \to \phi$ in $\mathcal{H}_\mathcal{X}$. The squared $L_2(P_X)$ distance of $\widehat{f}_n - E_X[\widehat{f}_n(X)]$ and $f - E_X[f(X)]$ is given by

$$\big\| \Sigma_{XX}^{1/2} (\widehat{f}_n - f) \big\|_{\mathcal{H}_\mathcal{X}}^2 = \big\| \Sigma_{XX}^{1/2} \widehat{f}_n \big\|_{\mathcal{H}_\mathcal{X}}^2 - 2 \langle \phi, \Sigma_{XX}^{1/2} \widehat{f}_n \rangle_{\mathcal{H}_\mathcal{X}} + \|\phi\|_{\mathcal{H}_\mathcal{X}}^2.$$

Thus, it suffices to show $\Sigma_{XX}^{1/2} \widehat{f}_n$ converges to ϕ in $\mathcal{H}_\mathcal{X}$ in probability. We have

$$\begin{aligned}
\big\| \Sigma_{XX}^{1/2} \widehat{f}_n - \phi \big\|_{\mathcal{H}_\mathcal{X}} \leq & \big\| \Sigma_{XX}^{1/2} \big\{ \big(\widehat{\Sigma}_{XX}^{(n)} + \varepsilon_n I \big)^{-1/2} - \big(\Sigma_{XX} + \varepsilon_n I \big)^{-1/2} \big\} \widehat{\phi}_n \big\|_{\mathcal{H}_\mathcal{X}} \\
& + \big\| \Sigma_{XX}^{1/2} \big(\Sigma_{XX} + \varepsilon_n I \big)^{-1/2} \big(\widehat{\phi}_n - \phi \big) \big\|_{\mathcal{H}_\mathcal{X}} \\
& + \big\| \Sigma_{XX}^{1/2} \big(\Sigma_{XX} + \varepsilon_n I \big)^{-1/2} \phi - \phi \big\|_{\mathcal{H}_\mathcal{X}}.
\end{aligned} \tag{18}$$

Using the same argument as the bound of the first term of Eq. (14), the first term in Eq. (18) is shown to converge to zero. The second term obviously converges to zero. Using the assumption $\phi \in \mathcal{R}(\Sigma_{XX})$, the same argument as the proof of Eq. (17) in Lemma 6 ensures the convergence of the third term to zero, which completes the proof. □

5 Concluding remarks

We have established the statistical convergence of kernel CCA and NOCCO, showing that the finite sample estimators of the relevant nonlinear mappings converge to the desired population functions. This convergence is proved in the RKHS norm for NOCCO, and in the L_2 norm for kernel CCA. These results give a theoretical justification for using the empirical estimates of NOCCO and kernel CCA in practice.

We have also derived a sufficient condition, $n^{1/3}\varepsilon_n \to \infty$, for the decay of the regularization coefficient ε_n, which ensures the convergence described above. As [12] suggests, the order of the sufficient condition seems to depend on the function norm used to determine convergence. An interesting consideration is whether the order $n^{1/3}\varepsilon_n \to \infty$ can be improved for convergence in the L_2 or RKHS norm.

We put an assumption of compactness for V_{YX} to derive convergence results in Theorem 1 and 2. However, practical characterization of this requirement in terms of the random variables X and Y has not been clarified. It is shown in Baker [4] that for Gaussian random elements in $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ with variance and covariance Σ_{XX}, Σ_{YY}, and Σ_{XY}, the operator V_{YX} is Hilbert-Schmidt and $\|V_{YX}\| < 1$ if and only if the mutual information of the Gaussian elements in RKHS is finite. Thus, if we consider Gaussian random elements ξ_X and ξ_Y with the same variance covariance operators with X and Y, the finiteness of the mutual information of ξ_X and ξ_Y works as a sufficient condition for the compactness of V_{YX}. However, the meaning of the mutual information in terms of the original random variables X and Y is not clear. It is a very interesting problem to derive practical sufficient conditions of the compactness.

Another question that remains to be addressed is when to use kernel CCA, COCO, or NOCCO in practice. The answer probably depends on the statistical properties of the data. It might consequently be helpful to determine the relation between the spectral properties of the data distribution and the solutions of these methods.

Acknowledgements

The first author is supported by JSPS KAKENHI 15700241 and research grant from Inamori Foundation. The authors thank Matthias Hein for his helpful comments.

References

[1] S. Akaho. A kernel method for canonical correlation analysis. In *Proceedings of International Meeting on Psychometric Society (IMPS2001)*, 2001.

[2] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 69(3):337–404, 1950.

[3] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[4] C. R. Baker. Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.*, 186:273–289, 1973.

[5] N. Dunford and J. T. Schwartz. *Linear Operators, Part II.* Interscience, 1963.

[6] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

[7] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. Technical Report 140, Max-Planck-Institut für biologische Kybernetik, 2005.

[8] A. Gretton, A. Smola, O. Bousquet, R. Herbrich, B. Schölkopf, and N. Logothetis. Behaviour and convergence of the constrained covariance. Technical Report 128, Max-Planck-Institut für biologische Kybernetik, 2004.

[9] C. W. Groetsch. *The theory of Tikhonov regularization for Fredholm equations of the first kind.* Pitman, 1984.

[10] D. R. Hardoon, J. Shawe-Taylor, and O. Friman. KCCA for fMRI analysis. In *Proceedings of Medical Image Understanding and Analysis (London)*, 2004.

[11] P. D. Lax. *Functional Analysis.* Wiley, 2002.

[12] S. Leurgans, R. Moyeed, and B. Silverman. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society, Series B*, 55(3):725–740, 1993.

[13] T. Melzer, M. Reiter, and H. Bischof. Nonlinear feature extraction using generalized canonical correlation analysis. In *Proceedings of International Conference on Artificial Neural Networks (ICANN)*, pages 353–360, 2001.

[14] S. Mika, G. Ratsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing*, volume IX, pages 41–48. IEEE, 1999.

[15] M. Reed and B. Simon. *Functional Analysis.* Academic Press, 1980.

[16] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[17] B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, 2002.

[18] N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan. *Probability Distributions on Banach Spaces.* D. Reidel Publishing Company, 1987.

[19] Y. Yamanishi, J.-P. Vert, A. Nakaya, and M. Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19:323i–330i, 2003.

A Basics from functional analysis

We briefly give definitions and basic properties of compact, trace class, and Hilbert-Schmidt operators. For complete references, see, for example, Reed & Simon ([15]), Dunford & Schwartz ([5]), and Lax [11], among others.

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces. A bounded operator $T : \mathcal{H}_1 \to \mathcal{H}_2$ is called *compact* if for every bounded sequence $\{f_n\} \subset \mathcal{H}_1$ the image $\{Tf_n\}$ has a subsequence which converges in $\mathcal{H}_2$. By Heine-Borel theorem, finite rank operators are necessarily compact. Among many useful properties of compact operators, singular value decomposition is available. Let $T : \mathcal{H}_1 \to \mathcal{H}_2$ be a compact operator. Then, there exist $N \in \mathbb{N} \cup \{\infty\}$, non-increasing sequence of positive numbers $\{\lambda_i\}_{i=1}^N$, and (not necessarily complete) orthonormal systems $\{\phi_i\}_{i=1}^N \subset \mathcal{H}_1$ and $\{\psi_i\}_{i=1}^N \subset \mathcal{H}_2$ so that

$$T = \sum_{i=1}^{\infty} \lambda_i \langle \phi_i, \cdot \rangle_{\mathcal{H}} \psi_i.$$

If $N = \infty$, then $\lambda_i \to 0$ $(i \to \infty)$ and the infinite series in the above equation converges in norm.

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces. A bounded operator $T : \mathcal{H}_1 \to \mathcal{H}_2$ is called *Hilbert-Schmidt* if $\sum_{i=1}^{\infty} \|T\varphi_i\|_{\mathcal{H}_2}^2 < \infty$ for a CONS $\{\varphi_i\}_{i=1}^{\infty}$ of $\mathcal{H}_1$. It is known that this value is independent of the choice of a CONS. For a Hilbert-Schmidt operator T, the Hilbert-Schmidt norm $\|T\|_{HS}$ is defined by

$$\|T\|_{HS}^2 = \sum_{i=1}^{\infty} \|T\varphi_i\|_{\mathcal{H}_2}^2. \tag{19}$$

For two Hilbert-Schmidt operators T_1 and T_2, the Hilbert-Schmidt inner product is defined by

$$\langle T_1, T_2 \rangle_{HS} = \sum_{i=1}^{\infty} \langle T_1 \varphi_i, T_2 \varphi_i \rangle_{\mathcal{H}_2}.$$

With this inner product, the set of all Hilbert-Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$ is a Hilbert space. Obviously,

$$\|T\| \leq \|T\|_{HS}$$

if T Hilbert-Schmidt.

B Lemmas used in the proofs

We show three lemmas used in the proofs in Section 4. Although they may be basic facts, we show the complete proofs for convenience.

Lemma 7. *Suppose A and B are positive self-adjoint operators on a Hilbert space such that $0 \leq A \leq \lambda I$ and $0 \leq B \leq \lambda I$ hold for a positive constant λ. Then,*

$$\|A^{3/2} - B^{3/2}\| \leq 3\lambda^{3/2} \|A - B\|.$$

Proof. Without loss of generality we can assume $\lambda = 1$. Let

$$f(z) = (1-z)^{3/2} \qquad \text{and} \qquad g(z) = (1-z)^{1/2}$$

be functions on $\{z \mid |z| \leq 1\}$, and

$$f(z) = \sum_{n=1}^{\infty} b_n z^n \qquad \text{and} \qquad g(z) = \sum_{n=0}^{\infty} c_n z^n$$

These series converge absolutely for $|z| \leq 1$. In fact, because direct differentiation derives $b_0 = 1$, $b_1 = -\frac{3}{2}$, and $b_n > 0$ for $n \geq 2$, the inequality

$$\begin{aligned}\sum_{n=0}^{N} |b_n| &= 1 + \frac{3}{2} + \sum_{n=2}^{N} b_n = 1 + \frac{3}{2} + \lim_{x \uparrow 1} \sum_{n=2}^{N} b_n x^n \\ &\leq 1 + \frac{3}{2} + \lim_{x \uparrow 1} \left\{ f(x) - 1 + \frac{3}{2} \right\} = 3\end{aligned}$$

shows the convergence of $\sum_{n=0}^{\infty} b_n z^n$ for $|z| = 1$. The bound $\sum_{n=0}^{\infty} |c_n| \leq 2$ can be proved similarly.

From $0 \leq I - A, I - B \leq I$, we have $f(A) = A^{3/2}$, $f(B) = B^{3/2}$, and thus,

$$\|A^{3/2} - B^{3/2}\| \leq \sum_{n=0}^{\infty} |b_n| \|(I - A)^n - (I - B)^n\|.$$

It is easy to see $\|T^n - S^n\| \leq n\|T - S\|$ by induction for any operators T and S with $\|T\| \leq 1$ and $\|S\| \leq 1$. From $f'(z) = -\frac{3}{2} g(z)$, the relation $nb_n = -\frac{3}{2} c_n$ holds for all n. Thus,

$$\|A^{3/2} - B^{3/2}\| \leq \sum_{n=0}^{\infty} n|b_n| \|A - B\| = \frac{3}{2} \sum_{n=0}^{\infty} |c_n| \|A - B\| \leq 3\|A - B\|$$

holds, which proves the lemma. □

The following lemma is a slight extension of Exercise 9, Section 21.2 in Lax ([11]).

Lemma 8. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces, and $\mathcal{H}_0$ be a dense linear subspace of $\mathcal{H}_2$. Suppose A_n and A are bounded operators on $\mathcal{H}_2$, and B is a compact operator from $\mathcal{H}_1$ to $\mathcal{H}_2$ such that*

$$A_n u \to A u$$

for all $u \in \mathcal{H}_0$, and

$$\sup_n \|A_n\| \leq M$$

for some $M > 0$. Then, $A_n B$ converges to AB in norm.

Proof. First, we prove that $A_n u \to Au$ holds for an arbitrary $u \in \mathcal{H}_2$. For any $\varepsilon > 0$, there is $u_0 \in \mathcal{H}_0$ so that $\|u - u_0\|_{\mathcal{H}_2} \le \varepsilon/(3(M + \|A\|))$. For $u_0 \in \mathcal{H}_0$, there is $N \in \mathbb{N}$ such that $\|A_n u_0 - Au_0\|_{\mathcal{H}_2} \le \varepsilon/3$ for all $n \ge N$. Then, for all $n \ge N$ we have

$$\|A_n u - Au\|_{\mathcal{H}_2} \le \|A_n\|\|u - u_0\|_{\mathcal{H}_2} + \|A_n u_0 - Au_0\|_{\mathcal{H}_2} + \|A\|\|u - u_0\|_{\mathcal{H}_2} \le \varepsilon.$$

Next, assume that the operator norm $\|A_n B - AB\|$ does not converge to zero. Then, there exist $\delta > 0$ and a subsequence (n') such that $\|A_{n'}B - AB\| \ge 2\delta$. For each n' there exists $v_{n'} \in \mathcal{H}_1$ such that $\|v_{n'}\|_{\mathcal{H}_1} = 1$ and $\|A_{n'}Bv_{n'} - ABv_{n'}\|_{\mathcal{H}_2} \ge \delta$. Let $u_{n'} = Bv_{n'}$. Because B is compact and $\|v_{n'}\|_{\mathcal{H}_1} = 1$, there is a subsequence $u_{n''}$ and u_* in $\mathcal{H}_2$ such that $u_{n''} \to u_*$. We have

$$\begin{aligned}
&\|A_{n''}u_{n''} - Au_{n''}\|_{\mathcal{H}_2} \\
&\le \|A_{n''}(u_{n''} - u_*)\|_{\mathcal{H}_2} + \|(A_{n''} - A)u_*\|_{\mathcal{H}_2} + \|A(u_{n''} - u_*)\|_{\mathcal{H}_2} \\
&\le (M + \|A\|)\|u_{n''} - u_*\|_{\mathcal{H}_2} + \|(A_{n''} - A)u_*\|_{\mathcal{H}_2},
\end{aligned}$$

which converges to zero as $n'' \to \infty$. This contradicts the choice of $v_{n'}$. □

Lemma 9. *Let A be a compact positive operator on a Hilbert space $\mathcal{H}$, and A_n ($n \in \mathbb{N}$) be bounded positive operators on $\mathcal{H}$ such that A_n converges to A in norm. Assume that the eigenspace of A corresponding to the largest eigenvalue is one-dimensional spanned by a unit eigenvector ϕ, and the maximum of the spectrum of A_n is attained by a unit eigenvector f_n. Then,*

$$|\langle f_n, \phi\rangle_{\mathcal{H}}| \to 1 \quad (n \to \infty).$$

Proof. Because A is compact and positive, the eigen decomposition

$$A = \sum_{i=1}^{\infty} \rho_i \phi_i \langle \phi_i, \cdot\,\rangle$$

holds, where $\rho_1 > \rho_2 \ge \rho_3 \ge \cdots \ge 0$ are eigenvalues and $\{\phi_i\}$ is the corresponding eigenvectors so that $\{\phi_i\}$ is the CONS of $\mathcal{H}$.

Let $\delta_n = |\langle f_n, \phi_1\rangle|$. We have

$$\begin{aligned}
\langle f_n, Af_n\rangle &= \rho_1 \langle f_n, \phi_1\rangle^2 + \sum_{i=2}^{\infty} \rho_i \langle \phi_i, f_n\rangle^2 \\
&\le \rho_1 \langle f_n, \phi_1\rangle^2 + \rho_2\bigl(1 - \langle f_n, \phi_1\rangle^2\bigr) = \rho_1 \delta_n^2 + \rho_2(1 - \delta_n^2).
\end{aligned}$$

On the other hand, the convergence

$$\begin{aligned}|\langle f_n, Af_n\rangle - \langle \phi_1, A\phi_1\rangle| &\le |\langle f_n, Af_n\rangle - \langle f_n, A_n f_n\rangle| + |\langle f_n, A_n f_n\rangle - \langle \phi_1, A\phi_1\rangle| \\ &\le \|A - A_n\| + \big|\|A_n\| - \|A\|\big| \quad \to \quad 0\end{aligned}$$

implies that $\langle f_n, Af_n\rangle$ must converges to ρ_1. By $\rho_1 > \rho_2$, this concludes $\delta_n \to 1$. □

Note that from the norm convergence $Q_n A_n Q_n \to QAQ$, where Q_n and Q are the orthogonal projections onto the orthogonal complement to ϕ_n and ϕ, respectively, we have the convergence of the eigenvector corresponding to the second eigenvalue. It is not difficult to obtain the convergence of the eigenspaces corresponding to the m-th eigenvalue in a similar way.