Journal of Machine Learning Research 8 (2007) 361-383 Submitted 8/06; Revised 12/06; Published 2/07

Statistical Consistency of Kernel Canonical Correlation Analysis

Kenji Fukumizu FUKUMIZU@ISM.AC.JP
Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku
Tokyo 106-8569 Japan

Francis R. Bach FRANCIS.BACH@MINES.ORG
Centre de Morphologie Mathématique
Ecole des Mines de Paris
35, rue Saint-Honoré
77300 Fontainebleau, France

Arthur Gretton ARTHUR.GRETTON@TUEBINGEN.MPG.DE
Department Schölkopf
Max Planck Institute for Biological Cybernetics
Spemannstraße 38, 72076 Tübingen, Germany

Editor: Aapo Hyvarinen

Abstract

While kernel canonical correlation analysis (CCA) has been applied in many contexts, the convergence of finite sample estimates of the associated functions to their population counterparts has not yet been established. This paper gives a mathematical proof of the statistical convergence of kernel CCA, providing a theoretical justification for the method. The proof uses covariance operators defined on reproducing kernel Hilbert spaces, and analyzes the convergence of their empirical estimates of finite rank to their population counterparts, which can have infinite rank. The result also gives a sufficient condition for convergence on the regularization coefficient involved in kernel CCA: this should decrease as $n^{-1/3}$, where n is the number of data.

Keywords: canonical correlation analysis, kernel, consistency, regularization, Hilbert space

1. Introduction

Kernel methods (Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002) have recently been developed as a methodology for nonlinear data analysis with positive definite kernels. In kernel methods, data are represented as functions or elements in reproducing kernel Hilbert spaces (RKHS), which are associated with positive definite kernels. The application of various linear methods in these Hilbert spaces is possible due to the reproducing property, which makes computation of inner product in the Hilbert spaces tractable. Many methods have been proposed as nonlinear extensions of conventional linear methods, such as kernel principal component analysis (Schölkopf et al., 1998), kernel Fisher discriminant analysis (Mika et al., 1999), and so on.

Kernel canonical correlation analysis (kernel CCA) was proposed (Akaho, 2001; Melzer et al., 2001; Bach and Jordan, 2002) as a nonlinear extension of canonical correlation analysis with positive definite kernels. 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 $g(Y)$, where f and g belong to the respective RKHS $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, such that their correlation is maximized. Kernel CCA has been successfully applied in practice for extracting nonlinear relations between variables in genomic data (Yamanishi et al., 2003), fMRI brain images (Hardoon et al., 2004), chaotic time series (Suetani et al., 2006) and independent component analysis (Bach and Jordan, 2002).

As in many statistical methods, the target functions defined in the population case are in practice estimated from a finite sample. Thus, the convergence of the estimated functions to the population functions 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; we thus 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. In proving the consistency of kernel CCA, we show also the consistency of a pair of functions which may be used as an alternative method for expressing the nonlinear dependence of two variables. The latter method uses the eigenfunctions of 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 (Groetsch, 1984), to enforce smoothness of the functions in the finite sample case (thus avoiding a trivial solution) and to enable operator 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 to converge to the desired functions in the population limit.

Another important issue in establishing convergence is an appropriate distance measure for functions. For NOCCO, we obtain convergence in the norm of the associated 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 prove convergence in the L_2 norm, which is a standard distance measure for functions.

There are earlier studies relevant to the convergence of functional correlation analysis. Among others, Breiman and Friedman (1985) propose alternating conditional expectation (ACE), an iterative algorithm for functional CCA and more general regression, and demonstrate statistical consistency of the algorithm for an infinite amount of data.

Most relevant to this paper are several studies on the consistency of CCA with positive definite kernels, notably the work on nonlinear CCA for stochastic processes by Leurgans et al. (1993); He et al. (2003), who also provide consistency results. An alternative approach is to study the eigenfunctions of the cross-covariance operators, without normalising by the variance, as in the constrained covariance (COCO, Gretton et al., 2005b). We will discuss the relation between our results and these studies.

We begin our presentation in Section 2 with a review of kernel CCA and related methods, formulating them in terms of cross-covariance operators, which are the basic tools to analyze correlation in functional spaces. In Section 3, we describe the two main theorems, which respectively show the convergence of kernel CCA and NOCCO. Section 4 contains numerical results to illustrate the behavior of the methods. Section 5 is devoted to the proof of the main theorems. Some basic facts from functional analysis and general lemmas are summarized in the Appendix.

2. Kernel Canonical Correlation Analysis

In this section, we briefly review kernel CCA, following Bach and Jordan (2002), and reformulate it with covariance operators on RKHS. For the detail of positive definite kernels and RKHS, see Aronszajn (1950).

In this paper, a Hilbert space always means a separable Hilbert space, and an operator a linear operator. The operator norm of a bounded operator T is denoted by $\|T\|$ and defined as $\|T\| = \sup_{\|f\|=1} \|Tf\|$. The null space and the range of an operator $T : \mathcal{H}_1 \to \mathcal{H}_2$ are denoted by $\mathcal{N}(T)$ and $\mathcal{R}(T)$, respectively; that is, $\mathcal{N}(T) = \{f \in \mathcal{H}_1 \mid Tf = 0\}$ and $\mathcal{R}(T) = \{Tf \in \mathcal{H}_2 \mid f \in \mathcal{H}_1\}$.

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 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 law P_{XY}. The marginal distributions 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 (e.g., Greenacre, 1984) looks for linear mappings $a^T X$ and $b^T Y$ 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, kernel CCA solves the following problem:[1]

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

The maximizing functions f and g are decided up to scale.

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

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

[1] In Eq. (2) we assume $\mathrm{Var}[f(X)] \neq 0$ and $\mathrm{Var}[g(Y)] \neq 0$. See Section 2.2 for discussion on conditions under which an RKHS includes a function leading to null variance.

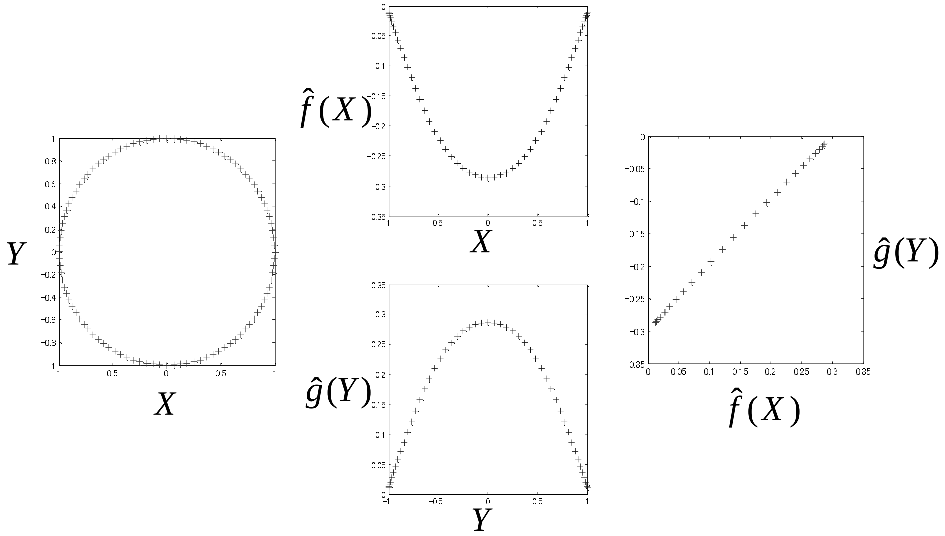


Figure 1: An example of kernel CCA. A Gaussian RBF kernel $k(x,y) = \exp\left(-\frac{1}{2\sigma^2}(x-y)^2\right)$ is used for both X and Y. Left: the original data. Center: derived functions $\widehat{f}(X_i)$ and $\widehat{g}(Y_i)$. Right: transformed data.

where

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

$$\widehat{\mathrm{Var}}[f(X)] = \frac{1}{n}\sum_{i=1}^{n}\Big(f(X_i) - \frac{1}{n}\sum_{j=1}^{n} f(X_j)\Big)^2,$$

$$\widehat{\mathrm{Var}}[g(Y)] = \frac{1}{n}\sum_{i=1}^{n}\Big(g(Y_i) - \frac{1}{n}\textstyle\sum_{j=1}^{n} g(Y_j)\Big)^2,$$

and a positive constant ε_n is the regularization coefficient (Bach and Jordan, 2002). 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 (Groetsch, 1984). For this smoothing effect, see also the discussion by Leurgans et al. (1993, Section 3).

Figure 1 shows the result of kernel CCA for a synthetic data set. The nonlinear mappings clarify the strong dependency between X and Y. Note that the dependency of the original data cannot be captured by classical CCA, because they have no linear correlation.

2.2 Cross-covariance Operators on RKHS

Kernel CCA and related methods can be formulated using cross-covariance operators, which make the theoretical analysis easier. Cross-covariance operators have also been used to derive practical methods for measuring the dependence of random variables (Fukumizu et al., 2004; Gretton et al., 2005a). This subsection reviews the basic properties of cross-covariance operators. For more details, see Baker (1973), Fukumizu et al. (2004), and Gretton et al. (2005a). The *cross-covariance* *operator*[2] 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 $g \in \mathcal{H}_\mathcal{Y}$. By regarding the right hand side as a linear functional on the direct product $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$, Riesz's representation theorem (Reed and Simon, 1980, for example) guarantees the existence and uniqueness of a bounded operator Σ_{YX}. The cross-covariance operator expresses the covariance between functions in the RKHS as a bilinear functional, and contains all the information regarding the dependence of X and Y expressible by nonlinear functions in the RKHS.

Obviously, $\Sigma_{YX} = \Sigma_{XY}^*$, where T^* denotes the adjoint of an operator T. In particular, if Y is equal to X, 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 $\{f \in \mathcal{H}_\mathcal{X} \mid f(X) = \text{constant almost surely}\}$. Under the assumptions that $\mathcal{X}$ is a topological space with continuous kernel $k_\mathcal{X}$ and the support of P_X is $\mathcal{X}$, the null space $\mathcal{N}(\Sigma_{XX})$ is equal to $\mathcal{H}_\mathcal{X} \cap \mathbb{R}$, where $\mathbb{R}$ denotes the constant functions. For the Gaussian RBF kernel $k(x,y) = \exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right)$ defined on $\mathcal{X} \subset \mathbb{R}^m$, it is known (Steinwart et al., 2004) that if the interior of $\mathcal{X}$ is not empty, a nontrivial constant function is not included in the RKHS; thus $\mathcal{N}(\Sigma_{XX}) = \{0\}$ in such cases.

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

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

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

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

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* $\widehat{\Sigma}_{YX}^{(n)}$ is defined as 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 $\widehat{\Sigma}_{YX}^{(n)}$ gives the empirical covariance as follows;

$$\begin{aligned}
&\langle g, \widehat{\Sigma}_{YX}^{(n)} f\rangle_{\mathcal{H}_\mathcal{Y}} \\
&= \frac{1}{n}\sum_{i=1}^n \left\langle g, k_\mathcal{Y}(\cdot, Y_i) - \frac{1}{n}\sum_{s=1}^n k_\mathcal{Y}(\cdot, Y_s)\right\rangle_{\mathcal{H}_\mathcal{Y}} \left\langle k_\mathcal{X}(\cdot, X_i) - \frac{1}{n}\sum_{t=1}^n k_\mathcal{X}(\cdot, X_t), f\right\rangle_{\mathcal{H}_\mathcal{X}} \\
&= \widehat{\mathrm{Cov}}[f(X), G(Y)].
\end{aligned}$$

Obviously, the rank of $\widehat{\Sigma}_{YX}^{(n)}$ is finite, because $\mathcal{R}(\widehat{\Sigma}_{YX}^{(n)})$ and $\mathcal{N}(\widehat{\Sigma}_{YX}^{(n)})^\perp$ are included in the linear hull of $\{k_\mathcal{Y}(\cdot, Y_i) - \frac{1}{n}\sum_{s=1}^n k_\mathcal{Y}(\cdot, Y_s)\}_{i=1}^n$ and $\{k_\mathcal{X}(\cdot, X_i) - \frac{1}{n}\sum_{t=1}^n k_\mathcal{X}(\cdot, X_t)\}_{i=1}^n$, respectively.

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 (Baker, 1973, 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 $\|V_{YX}\| \le 1$ and $V_{YX} = Q_Y V_{YX} Q_X$. Note that the inverse of an operator may not exist in general, or may not be continuous if it exists. We often write V_{YX} by $\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$, however, by abuse of notation, even when $\Sigma_{XX}^{-1/2}$ or $\Sigma_{YY}^{-1/2}$ are not appropriately defined as operators.

[2] Cross-covariance operator have been defined for Banach spaces by Baker (1973). However, we confine our discussion to RKHS.

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 as

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

As with classical CCA (Anderson, 2003, for example), the solution of the above kernel CCA problem is given by the eigenfunctions corresponding to the largest eigenvalue of the following generalized eigenproblem:

$$\begin{cases} \Sigma_{YX} f = \rho_1 \Sigma_{YY} g, \\ \Sigma_{XY} g = \rho_1 \Sigma_{XX} f. \end{cases} \tag{6}$$

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

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

and Eq. (6) becomes

$$\begin{cases} \widehat{\Sigma}^{(n)}_{YX} f = \widehat{\rho}^{(n)}_1 \big(\widehat{\Sigma}^{(n)}_{YY} + \varepsilon_n I\big) g, \\ \widehat{\Sigma}^{(n)}_{XY} g = \widehat{\rho}^{(n)}_1 \big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big) f. \end{cases} \tag{7}$$

Let us assume that the operator V_{YX} given by Eq. (5) is compact,[3] and let ϕ and ψ be the unit eigenfunctions of V_{YX} corresponding to the largest singular value;[4] that is,

$$\langle \psi, V_{YX}\phi\rangle_{\mathcal{H}_Y} = \max_{\substack{f\in\mathcal{H}_X, g\in\mathcal{H}_Y \\ \|f\|_{\mathcal{H}_X} = \|g\|_{\mathcal{H}_Y} = 1}} \langle g, V_{YX} f\rangle_{\mathcal{H}_Y}. \tag{8}$$

Given that $\phi \in \mathcal{R}(\Sigma_{XX})$ and $\psi \in \mathcal{R}(\Sigma_{YY})$, it is easy to see from Eq. (6) that the solution of the kernel CCA is given by the inverse images[5]

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

where f and g are determined up to an almost sure constant function. In the empirical case, let $\widehat{\phi}_n \in \mathcal{H}_X$ and $\widehat{\psi}_n \in \mathcal{H}_Y$ be the unit eigenfunctions corresponding to the largest singular value of the finite rank operator

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

From Eq. (7), the empirical estimators $\widehat{f}_n$ and $\widehat{g}_n$ of kernel CCA are

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

[3] See Appendix A for compact operators.

[4] While we presume that the eigenspaces are one dimensional in this section, we can easily relax it to multidimensional spaces by considering the eigenspaces corresponding to the largest eigenvalues. See the remarks after Theorem 2.

[5] The operators $\Sigma_{XX}^{1/2}$ and $\Sigma_{YY}^{1/2}$ may not be invertible, but their inverses are well-defined up to an almost sure constant function when applied to functions belonging to the respective ranges of $\Sigma_{XX}^{1/2}$ and $\Sigma_{YY}^{1/2}$.

The empirical operators and the estimators described above can be expressed using *Gram matrices*, as is often done in kernel methods. The solutions $\widehat{f}_n$ and $\widehat{g}_n$ are exactly the same as those given in Bach and Jordan (2002), as we confirm below. Let $u_i \in \mathcal{H}_\mathcal{X}$ and $v_i \in \mathcal{H}_\mathcal{Y}$ $(1 \le i \le n)$ 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 $\mathcal{R}(\widehat{\Sigma}^{(n)}_{XX})$ and $\mathcal{R}(\widehat{\Sigma}^{(n)}_{YY})$ are spanned by $(u_i)_{i=1}^n$ and $(v_i)_{i=1}^n$, respectively, the eigenfunctions of $\widehat{V}^{(n)}_{YX}$ 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 $\langle \psi, \widehat{V}^{(n)}_{YX}\phi\rangle_{\mathcal{H}_\mathcal{Y}}$ shows that the eigenfunctions $\widehat{\phi}_n$ and $\widehat{\psi}_n$ of $\widehat{\Sigma}^{(n)}_{YX}$ corresponding to the largest singular value are given by the coefficients $\widehat{\alpha}$ and $\widehat{\beta}$ that satisfy

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

where G_X is the centered Gram matrix,

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

with G_Y defined accordingly. The solution of the kernel CCA problem is

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

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 by Bach and Jordan (2002). Bach and Jordan approximate $(G_X^2 + n\varepsilon_n G_X)$ by $(G_X + \frac{n\varepsilon_n}{2}I_n)^2$ for computational simplicity. Note that our theoretical results in the next section still hold with this approximation, because this modification causes only higher order changes in $\widehat{\alpha}$ and $\widehat{\beta}$, which perturbs the empirical eigenfunctions $\widehat{\phi}_n$, $\widehat{\psi}_n$, $\widehat{f}_n$, and $\widehat{g}_n$ only in higher order.

There are additional, related methods to extract nonlinear dependence of two random variables with positive definite kernels. The Constrained Covariance (COCO, Gretton et al., 2005b) 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 by Gretton et al. (2005a). Unlike kernel CCA, 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 between these methods is to use ϕ and ψ in Eq. (8), and their estimates $\widehat{\phi}_n$ and $\widehat{\psi}_n$. While the statistical meaning of this approach is not as direct as kernel 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, and give experimental comparisons of these methods in Section 4.

3. Main Theorems

First, the following theorem asserts the consistency of the estimator of NOCCO in the RKHS norm, when the regularization parameter ε_n goes to zero slowly enough.

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

$$\lim_{n\to\infty}\varepsilon_n = 0, \qquad \lim_{n\to\infty}\frac{n^{-1/3}}{\varepsilon_n} = 0. \tag{9}$$

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

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

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,

$$|\langle\widehat{\phi}_n,\phi\rangle_{\mathcal{H}_X}| \to 1, \qquad |\langle\widehat{\psi}_n,\psi\rangle_{\mathcal{H}_Y}| \to 1$$

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. (9). 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 in the above theorems we confine our attention to the first eigenfunctions, it is not difficult to verify the convergence of eigenspaces corresponding to the m-th largest eigenvalue by extending Lemma 10 in the Appendix. See also the remark after the lemma.

The convergence of NOCCO in RKHS norm is a very strong result. If $\mathcal{X}$ and $\mathcal{Y}$ are topological spaces, 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.

The above theorems assume the compactness of V_{YX}, which requires that 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 correlation of $\Sigma_{XX}^{-1/2}\phi_i(X)$ and $\Sigma_{YY}^{-1/2}\psi_i(Y)$ decay 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 this case, $V_{YX} = I$ is not compact, and the problem in Theorem 1 is solved by an arbitrary function. In this situation, the kernel CCA problem in Theorem 2 does not have solutions if Σ_{XX} has arbitrarily small eigenvalues.

We give a useful sufficient condition that V_{YX} is Hilbert-Schmidt, which necessarily implies compactness. The condition is described in terms of mean square contingency, which is one of the standard criteria to measure the dependency of two random variables (Rényi, 1970). It is known (Buja, 1990) that the covariance operator considered on L^2 is Hilbert-Schmidt if the mean square contingency is finite. We modify the result to the case of the covariance operator on RKHS.

Assume that the measure spaces $(\mathcal{X},\mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Y},\mathcal{B}_{\mathcal{Y}})$ admit measures $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$, respectively, and that P_{XY} is absolutely continuous with respect to the product measure $\mu_{\mathcal{X}} \times \mu_{\mathcal{Y}}$ with a probability density function $p_{XY}(x,y)$. Let $\zeta(x,y)$ be a function on $\mathcal{X}\times\mathcal{Y}$ defined by

$$\zeta(x,y) = \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - 1,$$

where $p_X(x)$ and $p_Y(y)$ are the probability density functions of the marginal distributions P_X and P_Y, respectively. The *mean square contingency* $C(X,Y)$ is defined by

$$C(X,Y) = \left\{\int\int \zeta(x,y)^2 dP_X dP_Y\right\}^{1/2}.$$

It is easy to see $C(X,Y) = 0$ if and only if X and Y are independent. Obviously we have

$$C(X,Y)^2 = \int\int \frac{p_{XY}(x,y)^2}{p_X(x)p_Y(y)} d\mu_{\mathcal{X}} d\mu_{\mathcal{Y}} - 1 = \int \zeta(x,y) dP_{XY}.$$

Thus, $C(X,Y)^2$ is an upper bound of the mutual information $MI(X,Y) = \int \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} dP_{XY}$, because $\log(z+1) \leq z$ for $z > 0$.

Theorem 3 *Suppose that the measurable spaces $(\mathcal{X},\mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Y},\mathcal{B}_{\mathcal{Y}})$ have measures $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$, respectively, so that P_{XY} is absolutely continuous with respect to the product measure $\mu_{\mathcal{X}} \times \mu_{\mathcal{Y}}$ with a probability density function $p_{XY}(x,y)$. If the mean square contingency $C(X,Y)$ is finite, that is, if*

$$\int\int \frac{p_{XY}(x,y)^2}{p_X(x)p_Y(y)} d\mu_{\mathcal{X}} d\mu_{\mathcal{Y}} < \infty,$$

then the operator $V_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ is Hilbert-Schmidt, and

$$\|V_{YX}\|_{HS} \leq C(X,Y) = \|\zeta\|_{L^2(P_X\times P_Y)}.$$

The proof is given in Section 5. The assumption that the mean square contingency is finite is very natural when we consider the dependence of two different random variables, as in the situation where kernel CCA is applied. It is interesting to see that Breiman and Friedman (1985) also discuss a similar condition for the existence of optimal functions for functional canonical correlation.

He et al. (2003) discuss the eigendecomposition of the operator V_{YX}. They regard random processes in L^2 spaces as data. In our case, transforms of the original random variables, $k_{\mathcal{X}}(\cdot,X)$ and $k_{\mathcal{Y}}(\cdot,Y)$, also induce processes in RKHS. He et al. (2003) give a condition for the eigendecomposition of V_{YX} in terms of the eigendecomposition of the data processes, while our condition is a more direct property of the original random variables.

Leurgans et al. (1993) 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 this Sobolev space is an 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,

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

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

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

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

4. Numerical Simulations

In this section, we show results of numerical simulations for kernel CCA and related methods. We use a synthetic data set for which the optimal nonlinear functions in the population kernel CCA (Eq. (2)) are explicitly known, and demonstrate the convergence behavior for various values of ε_n. For the quantitative evaluation of convergence, we consider only kernel CCA, because the exact solutions for NOCCO or COCO in population are not known in closed form.

To generate our test data, we provide two univariate random variables X and Y for which the true transforms $f(X)$ and $g(Y)$ are highly linearly correlated for some $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_{\mathcal{Y}}$. We generate a sample from P_{XY} as follows: first, we sample $Z_1, \ldots, Z_n$ uniformly on the unit interval $[0,1]$. Next, we derive two i.i.d. linearly correlated random samples U_i and V_i from these Z_i. Finally,

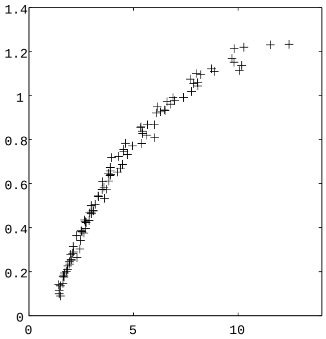


Figure 2: The plot of (R_i^X, R_i^Y) of the data used in the experiment.

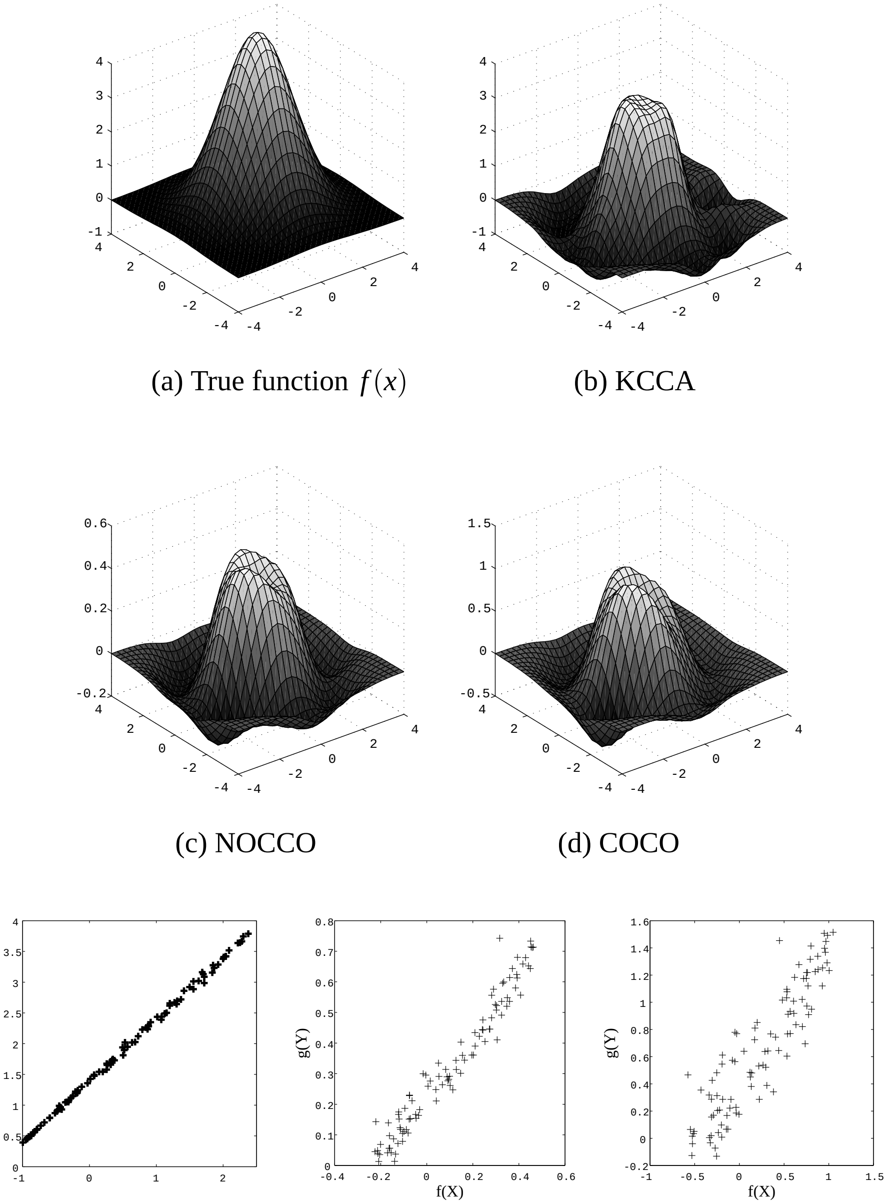
we transform these variables to radius data R_i^X and R_i^Y by the inverse of the Gaussian function $\exp(-aR^2)$ for some $a > 0$. The explicit form of these relations are

$$U_i = Z_i + 0.06 + e_i^X, \quad V_i = Z_i + 3 + e_i^Y,$$
$$R_i^X = \big(-4\log(U_i/1.5)\big)^{1/2}, \qquad R_i^Y = \big(-4\log(V_i/4.1)\big)^{1/2},$$

where e_i^X and e_i^Y are independent noise following a zero-mean saturated Gaussian distribution so that U_i and V_i are positive. See Figure 2 for an example data set. The samples X_i and Y_i are taken uniformly on the 2 dimensional circles with the radius R_i^X and R_i^Y, respectively. Thus, the maximum canonical correlation in population is attained by $f(x) = 1.5\exp(-\frac{1}{4}\|x\|^2)$ and $g(y) = 4.1\exp(-\frac{1}{4}\|y\|^2)$ up to scale and shift.

We perform kernel CCA, NOCCO, and COCO with Gaussian RBF kernel $k(x,y) = \exp(-\|x-y\|^2)$ on the data. Note that the true functions f and g for kernel CCA are included in RKHS with this kernel. The graphs of resulting functions for X, the true function $f(x)$, and the transformed data are shown in Figure 3. We see that the functions obtained by Kernel CCA, NOCCO, and COCO have a similar shape to $f(x)$. Note that, because the data exist only around the area $f(x) \le 1.0$, the estimation accuracy in the flat area including the origin is low. In the plots (e)-(g), kernel CCA gives linearly correlated feature vectors, while NOCCO and COCO do not aim at obtaining linearly correlated vectors. However, we see that these two methods also give the features that contain the sufficient information on the dependency between X and Y.

Next, we conduct numerical simulations to verify the convergence rate of kernel CCA. Figure 4 shows the convergence to the true functions for various decay rates of the regularization coefficient $\varepsilon_n = 0.001 \times n^{-a}$ with $a = 0.1 \sim 0.8$. For the estimated functions $\widehat{f}_n$ and $\widehat{g}_n$ with the data sizes $n = 10$, 25, 50, 75, 100, 250, 500, 750, and 1000, the $L^2(P_X)$ and $L^2(P_{\mathcal{Y}})$ distances between the estimated and true functions are evaluated by generating 10000 samples from the true distribution. The curves show an average over 30 experiments with different random data. It should be noted that, although the theoretical sufficient condition for convergence requires a slower order of ε_n than $n^{-1/3}$, faster orders give better convergence in these simulations. The convergence is best at $a = 0.6$, and becomes worse for faster decay rates; the optimum rate likely depends on the statistical properties of the data. It might therefore be interesting to find the best rate or best value of ε_n for the given data, although this is beyond the scope of the present paper.



(a) True function $f(x)$

(b) KCCA

(c) NOCCO

(d) COCO

(e) KCCA

(f) NOCCO

(g) COCO

Figure 3: The true function $f(x)$ and the estimated functions based on 100 data points are shown in (a)-(d). The plots of the transformed data $(\widehat{f}_n(X_i), \widehat{g}_n(Y_i))$ are given in (e)-(g). Note that in (e)-(g) the clear linear correlation is seen in (e), only because it is the criterion of the kernel CCA; the other two methods use different criterion, but still show strong correlation.

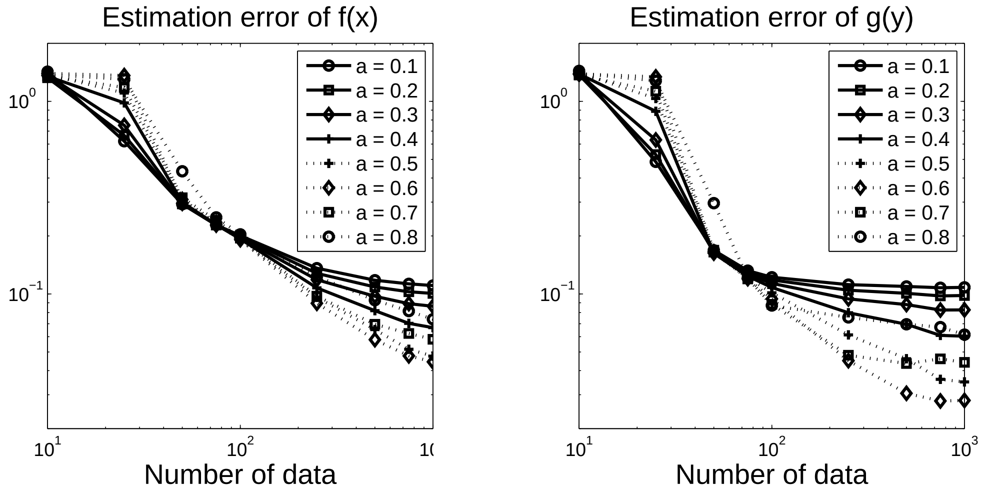


Figure 4: L^2 distances between the true function and its estimate using kernel CCA. The regularization parameter is $\varepsilon_n = 0.001 \times n^{-a}$, where the decay rate a ranges from 0.1 to 0.8.

5. Proof of the Main Theorems

In this section, we prove the main theorems in Section 3.

5.1 Hilbert-Schmidt Norm of Covariance Operators

Preliminary to the proofs, in this subsection we show some results on the Hilbert-Schmidt norm of cross-covariance operators. For convenience, we provide the definition and some basic properties of Hilbert-Schmidt operators in the Appendix. See also Gretton et al. (2005a).

We begin with a brief introduction to random elements in a Hilbert space (Vakhania et al., 1987; Baker, 1973). 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 an 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 of the mean element is a consequence of 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, $E[|\langle F, G \rangle_{\mathcal{H}}|]$ is finite. If further F and G are independent, the relation

$$E[\langle F, G \rangle_{\mathcal{H}}] = \langle E[F], E[G] \rangle_{\mathcal{H}} \tag{10}$$

holds.

It is easy to see that the example $F = k(\cdot, X)$ in an RKHS $\mathcal{H}$ has strong order two, that is, $E[\|F\|^2] < \infty$, 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 kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$, respectively, under the conditions 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 Gretton et al. (2005a) and Eq. (10). See Appendix for definitions of the Hilbert-Schmidt operator and Hilbert-Schmidt norm.

Lemma 4 *The cross-covariance operator Σ_{YX} is a Hilbert-Schmidt operator, and its Hilbert-Schmidt norm is given by*

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

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 order[6] $O_p(n^{-1/2})$. The following lemma shows the tight uniform result that $\|\widehat{\Sigma}_{YX}^{(n)} - \Sigma_{YX}\|_{HS}$ converges to zero in the order of $O_p(n^{-1/2})$.

Lemma 5

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

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 $G, G_1, \ldots, G_n$. Lemma 4 implies

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

and the same argument as in the proof of the lemma yields

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

From these equations, we have

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

[6] A random variable Z_n is said to be of order $O_p(a_n)$ if for any $\varepsilon > 0$ there exists $M > 0$ such that $\sup_n \Pr(|Z_n| > Ma_n) < \varepsilon$. See, for example, van der Vaart (1998).

which provides a bound

$$\big\|\widehat{\Sigma}_{YX}^{(n)}-\Sigma_{YX}\big\|_{HS}\le\Big\|\frac{1}{n}\sum_{i=1}^{n}F_iG_i-E[FG]\Big\|_{\mathcal{F}}+2\Big\|\Big(\frac{1}{n}\sum_{i=1}^{n}F_i\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}G_i\Big)\Big\|_{\mathcal{F}}. \tag{11}$$

Let $Z_i=F_iG_i-E[FG]$. Since the variance of a sum of independent random variables is equal to the sum of their variances, we obtain

$$E\Big\|\frac{1}{n}\sum_{i=1}^{n}Z_i\Big\|_{\mathcal{F}}^2=\frac{1}{n}E\|Z_1\|_{\mathcal{F}}^2, \tag{12}$$

which is of order $O(1/n)$ because $E\|Z_1\|_{\mathcal{F}}^2<\infty$. From the inequality

$$\begin{aligned}E\Big\|\Big(\frac{1}{n}\sum_{i=1}^{n}F_i\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}G_i\Big)\Big\|_{\mathcal{F}}&=E\Big[\Big\|\frac{1}{n}\sum_{i=1}^{n}F_i\Big\|_{\mathcal{H}_{\mathcal{X}}}\Big\|\frac{1}{n}\sum_{i=1}^{n}G_i\Big\|_{\mathcal{H}_{\mathcal{Y}}}\Big]\\&\le\Big(E\Big\|\frac{1}{n}\sum_{i=1}^{n}F_i\Big\|_{\mathcal{H}_{\mathcal{X}}}^2\Big)^{1/2}\Big(E\Big\|\frac{1}{n}\sum_{i=1}^{n}G_i\Big\|_{\mathcal{H}_{\mathcal{Y}}}^2\Big)^{1/2},\end{aligned}$$

in a similar way to Eq. (12), we have $E\big\|\big(\frac{1}{n}\sum_{i=1}^{n}F_i\big)\big(\frac{1}{n}\sum_{i=1}^{n}G_i\big)\big\|_{\mathcal{F}}=O(1/n)$.

From Eq. (11), we have $E\big\|\widehat{\Sigma}_{YX}^{(n)}-\Sigma_{YX}\big\|_{HS}=O(1/\sqrt{n})$, and the proof is completed by Chebyshev's inequality. ■

5.2 Preliminary Lemmas

For the proof of the main theorems, we show the empirical estimate $\widehat{V}_{YX}^{(n)}$ converges in norm to the normalized cross-covariance operator $V_{YX}=\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ for an appropriate order of the regularization coefficient ε_n. We divide the task into two lemmas: the first evaluates the difference between the empirical estimate $\widehat{V}_{YX}^{(n)}$ and a regularized version of V_{YX}, and the second asserts that the regularized version converges to V_{YX} if ε_n goes to zero at the appropriate rate.

Lemma 6 *Let ε_n be a positive number such that $\varepsilon_n\to0$ ($n\to\infty$). Then, for the i.i.d. sample $(X_1,Y_1),\ldots,(X_n,Y_n)$, we have*

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

Proof The operator in the left hand side is decomposed as

$$\begin{aligned}&\widehat{V}_{YX}^{(n)}-(\Sigma_{YY}+\varepsilon_nI)^{-1/2}\Sigma_{YX}(\Sigma_{XX}+\varepsilon_nI)^{-1/2}\\&=\big\{(\widehat{\Sigma}_{YY}^{(n)}+\varepsilon_nI)^{-1/2}-(\Sigma_{YY}+\varepsilon_nI)^{-1/2}\big\}\widehat{\Sigma}_{YX}^{(n)}(\widehat{\Sigma}_{XX}^{(n)}+\varepsilon_nI)^{-1/2}\\&\quad+(\Sigma_{YY}+\varepsilon_nI)^{-1/2}\big\{\widehat{\Sigma}_{YX}^{(n)}-\Sigma_{YX}\big\}(\widehat{\Sigma}_{XX}^{(n)}+\varepsilon_nI)^{-1/2}\\&\quad+(\Sigma_{YY}+\varepsilon_nI)^{-1/2}\Sigma_{YX}\big\{(\widehat{\Sigma}_{XX}^{(n)}+\varepsilon_nI)^{-1/2}-(\Sigma_{XX}+\varepsilon_nI)^{-1/2}\big\}.\end{aligned} \tag{13}$$

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. (13) is equal to

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

From $\big\|(\widehat{\Sigma}_{YY}^{(n)}+\varepsilon_n I)^{-1/2}\big\|\le\frac{1}{\sqrt{\varepsilon_n}}$, $\big\|\big(\widehat{\Sigma}_{YY}^{(n)}+\varepsilon_n I\big)^{-1/2}\widehat{\Sigma}_{YX}^{(n)}\big(\widehat{\Sigma}_{XX}^{(n)}+\varepsilon_n I\big)^{-1/2}\big\|\le 1$, and Lemma 8 in the Appendix, the norm of the above operator is bounded from above by

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

A similar bound also applies to the third term of Eq. (13). An upper bound on the second term of Eq. (13) is $\frac{1}{\varepsilon_n}\|\Sigma_{YX}-\widehat{\Sigma}_{YX}^{(n)}\|$. Thus, the proof is completed using $\|\widehat{\Sigma}_{XX}^{(n)}\|=\|\Sigma_{XX}\|+o_p(1)$, $\|\widehat{\Sigma}_{YY}^{(n)}\|=\|\Sigma_{YY}\|+o_p(1)$, and Lemma 5. ■

In the next theorem, the compactness assumption on V_{YX} plays an essential role.

Lemma 7 *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

$$\big\|\big\{(\Sigma_{YY}+\varepsilon_n I)^{-1/2}-\Sigma_{YY}^{-1/2}\big\}\Sigma_{YX}(\Sigma_{XX}+\varepsilon_n I)^{-1/2}\big\| \\ +\big\|\Sigma_{YY}^{-1/2}\Sigma_{YX}\big\{(\Sigma_{XX}+\varepsilon_n I)^{-1/2}-\Sigma_{XX}^{-1/2}\big\}\big\|. \quad (14)$$

The first term of Eq. (14) is upper bounded by

$$\big\|\big\{(\Sigma_{YY}+\varepsilon_n I)^{-1/2}\Sigma_{YY}^{1/2}-I\big\}V_{YX}\big\|. \quad (15)$$

Note that the range of V_{YX} is included in $\overline{\mathcal{R}(\Sigma_{YY})}$, as pointed out 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

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

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

$$\big\{(\Sigma_{YY}+\varepsilon_n I)^{-1/2}\Sigma_{YY}^{1/2}-I\big\}v\ \to\ 0\qquad(n\to\infty) \quad (16)$$

holds for all $v \in \mathcal{R}(V_{YX}) \cap \mathcal{R}(\Sigma_{YY})$. Because V_{YX} is compact, Lemma 9 in the Appendix shows Eq. (15) converges to zero. The convergence of the second term in Eq. (14) can be proved similarly. ■

Note that the assertion of the above theorem does not necessarily hold without the compactness assumption. In fact, if $Y = X$ and the RKHS is infinite dimensional, $V_{YX} = I$ is not compact, and the norm in the left hand of the assertion is $\|\Sigma_{XX}(\Sigma_{XX}+\varepsilon_n I)^{-1} - I\|$. Since Σ_{XX} has arbitrarily small positive eigenvalues, it is easy to see that this norm is equal to one for all n.

5.3 Proof of the Main Theorems

We are now in a position to prove Theorems 1 and 2.

Proof of Theorem 1 From Lemmas 6 and 7, $\widehat{V}_{YX}^{(n)}$ converges to V_{YX} in norm. Because ϕ and ψ are the eigenfunctions corresponding to the largest eigenvalue of $V_{YX}V_{XY}$ and $V_{XY}V_{YX}$, respectively, and a similar fact holds for $\widehat{\phi}_n$ and $\widehat{\psi}_n$, the assertion is obtained by Lemma 10 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 $\mathcal{H}_X$. The squared $L_2(P_X)$ distance between $\widehat{f}_n - E_X[\widehat{f}_n(X)]$ and $f - E_X[f(X)]$ is given by

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

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

$$\begin{aligned}\big\|\Sigma_{XX}^{1/2}\widehat{f}_n - \phi\big\|_{\mathcal{H}_X} \le{}& \big\|\Sigma_{XX}^{1/2}\big\{\big(\widehat{\Sigma}_{XX}^{(n)}+\varepsilon_n I\big)^{-1/2} - \big(\Sigma_{XX}+\varepsilon_n I\big)^{-1/2}\big\}\widehat{\phi}_n\big\|_{\mathcal{H}_X} \\ &+ \big\|\Sigma_{XX}^{1/2}\big(\Sigma_{XX}+\varepsilon_n I\big)^{-1/2}\big(\widehat{\phi}_n - \phi\big)\big\|_{\mathcal{H}_X} \\ &+ \big\|\Sigma_{XX}^{1/2}\big(\Sigma_{XX}+\varepsilon_n I\big)^{-1/2}\phi - \phi\big\|_{\mathcal{H}_X}. \end{aligned} \qquad (17)$$

Using the same argument as in the bound of the first term of Eq. (13), the first term in Eq. (17) 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 in the proof of Eq. (16) in Lemma 7 ensures the convergence of the third term to zero, which completes the proof. ■

With the definition of mean square contingency, Theorem 3 can be proved as follows.

Proof of Theorem 3 Since under the assumptions $E_X[k_{\mathcal{X}}(X,X)] < \infty$ and $E_Y[k_{\mathcal{Y}}(Y,Y)] < \infty$ the operators Σ_{XX} and Σ_{YY} are compact and self-adjoint, there exist complete orthonormal systems $\{\varphi_i\}_{i=1}^{\infty}$ and $\{\psi_i\}_{i=1}^{\infty}$ for $\mathcal{H}_X$ and $\mathcal{H}_{\mathcal{Y}}$, respectively, such that $\langle \varphi_j, \Sigma_{XX}\varphi_i\rangle_{\mathcal{H}_X} = \lambda_i\delta_{ij}$ and $\langle \psi_j, \Sigma_{YY}\psi_i\rangle_{\mathcal{H}_{\mathcal{Y}}} = \nu_i\delta_{ij}$, where λ_i and ν_i are nonnegative eigenvalues and δ_{ij} is Kronecker's delta. Let $\tilde{\phi}_i = (\varphi_i - E_X[\varphi_i(X)])/\sqrt{\lambda_i}$ and $\tilde{\psi}_i = (\psi_i - E_Y[\psi_i(Y)])/\sqrt{\nu_i}$. It follows that $(\tilde{\phi}_i, \tilde{\phi}_j)_{L^2(P_X)} = \delta_{ij}$ and $(\tilde{\psi}_i, \tilde{\psi}_j)_{L^2(P_Y)} = \delta_{ij}$, where $(\cdot,\cdot)_{L^2(P_X)}$ and $(\cdot,\cdot)_{L^2(P_Y)}$ denote the inner product of $L^2(P_X)$ and $L^2(P_Y)$, respectively. We have

$$\sum_{i,j=1}^{\infty}\langle\psi_j,\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}\varphi_i\rangle_{\mathcal{H}_{\mathcal{Y}}}^2 = \sum_{i,j=1}^{\infty}\Big\langle\frac{\psi_i}{\sqrt{\nu_j}},\Sigma_{YX}\frac{\varphi_i}{\sqrt{\lambda_i}}\Big\rangle_{\mathcal{H}_{\mathcal{Y}}}^2 = \sum_{i,j=1}^{\infty}E_{XY}[\tilde{\phi}_i(X)\tilde{\psi}_j(Y)]^2.$$

Note that we do not need to consider the eigenfunctions with respect to the zero eigenvalue in the sum, because $\overline{\mathcal{R}(V_{YX})} = \overline{\mathcal{R}(\Sigma_{YY})} = \mathcal{N}(\Sigma_{YY})^{\perp}$ and $\mathcal{N}(V_{YX}) = \mathcal{N}(\Sigma_{XX})$.

Since the set $\{\tilde{\phi}_i\tilde{\psi}_j\}$ is orthonormal in $L^2(P_X\times P_Y)$, we obtain

$$\sum_{i,j=1}^{\infty}E_{XY}[\tilde{\phi}_i(X)\tilde{\psi}_j(Y)]^2 = \sum_{i,j=1}^{\infty}\Big\{\int\int\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\tilde{\phi}_i(x)\tilde{\psi}_j(y)dP_XdP_Y\Big\}^2 \le \|\zeta+1\|_{L^2(P_X\times P_Y)}^2,$$

which is finite by assumption. ■

6. 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 Leurgans et al. (1993) suggest, the order of the sufficient condition seems to depend on the functional norm used to determine convergence. An interesting question is whether the theoretical order $n^{1/3}\varepsilon_n\to\infty$ can be improved for convergence in the L_2 or RKHS norm.

A result relevant to the convergence of kernel principal component analysis (KPCA) has recently been obtained by Zwald and Blanchard (2006). They show a probabilistic upper bound on the difference between the projectors onto the D-dimensional population eigenspaces and the empirical eigenspaces. Since KPCA needs no inversion operation, the theoretical analysis is easier than for kernel CCA. That said, it would be very interesting to consider the applicability of the methods developed by Zwald and Blanchard (2006) to kernel CCA.

There are some practical problems that remain to be addressed when applying kernel CCA and related methods. One of the problems is how to choose the regularization coefficient ε_n in practice. As the numerical simulations in Section 4 show, the order $n^{-1/3}$ is only a sufficient condition for convergence in general cases, and the optimal ε_n to estimate the true functions may depend on statistical properties of the given data, such as spectral distribution of Gram matrices. This problem should be studied more in future to make the methods more applicable.

The choice of kernel is another important unsolved problem. The kernel defines the meaning of ”nonlinear correlation” through an assumed class of functions, and thus determines how to measure the dependence structure of the data. If a parameterized family of kernels such as the Gaussian RBF kernel is provided, then cross-validation might be a reasonable method to select the best kernel (see Leurgans et al., 1993), however this remains to be established.

One of the methods related to kernel CCA is independent component analysis (ICA), since Bach and Jordan (2002) use kernel CCA in their kernel ICA algorithm. The theoretical results developed in this paper will work as a basis for analyzing the properties of the kernel ICA algorithm; in particular, for demonstrating statistical consistency of the estimator. Since ICA estimates the demixing matrix as a parameter, however, we need to consider covariance operators parameterized by this matrix, and must discuss how convergence of the objective function depends on the parameter. It is not a straightforward task to obtain consistency of kernel ICA from the results of this paper. Extending our results to the parametric case is an interesting topic for future work.

Acknowledgments

The first author was supported by JSPS KAKENHI 15700241, Research Grant of Inamori Foundation, Scientific Grant of Mitsubishi Foundation, and Transdisciplinary Research Integration Projects of ROIS. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. The authors thank Dr. Shotaro Akaho, Yukito Iba, Hiromichi Suetani, and Matthias Hein for their helpful comments, and thank editor and anonymous reviewers for their comments to improve the manuscript. Part of this work was done while the first author was visiting Max Planck Institute for Biological Cybernetics.

Appendix A. Basics from Functional Analysis

We briefly give definitions and basic properties of compact and Hilbert-Schmidt operators. For complete references, see, for example, Reed and Simon (1980), Dunford and Schwartz (1963), and Lax (2002), 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 the Heine-Borel theorem, finite rank operators are necessarily compact. Among many useful properties of compact operators, singular value decomposition is available. For a compact operator $T : \mathcal{H}_1 \to \mathcal{H}_2$, there exist $N \in \mathbb{N} \cup \{\infty\}$, a 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 $\{\psi_i\}_{i=1}^N \subset \mathcal{H}_2$ such that

$$T = \sum_{i=1}^{N} \lambda_i \langle \phi_i, \cdot \rangle_{\mathcal{H}_1} \psi_i.$$

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 sum is independent of the choice of a CONS. 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 which the set of all Hilbert-Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$ is a Hilbert space. The Hilbert-Schmidt norm $\|T\|_{HS}$ is defined by $\|T\|_{HS}^2 = \langle T, T \rangle_{HS} = \sum_{i=1}^\infty \|T\varphi_i\|_{\mathcal{H}_2}^2$ as usual. Obviously, for a

Hilbert-Schmidt operator T, we have

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

Appendix B. Lemmas Used in the Proofs

We show three lemmas used in the proofs in Section 5. Although they may be basic facts, we show the complete proofs for convenience.

Lemma 8 *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,*

$$\|A^{3/2} - B^{3/2}\| \leq 3\lambda^{1/2}\|A - B\|.$$

Proof Without loss of generality we can assume $\lambda = 1$. Define functions f and g on $\{z \mid |z| \leq 1\}$ by $f(z) = (1-z)^{3/2}$ and $g(z) = (1-z)^{1/2}$. Let

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

be the power series expansions. They converge absolutely for $|z| \leq 1$. In fact, because direct differentiation yields $b_0 = 1$, $b_1 = -\frac{3}{2}$, 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 $f(I - A) = A^{3/2}$, $f(I - B) = B^{3/2}$, and thus,

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

It is easy to see $\|T^n - S^n\| \leq n\|T - S\|$ by induction for 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. Therefore, we obtain

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

■

The following lemma is a slight extension of Exercise 9, Section 21.2 in Lax (2002).

Lemma 9 *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*

$$A_n u \to Au$$

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 $\|u - u_0\|_{\mathcal{H}_2} \le \varepsilon/(2(M + \|A\|))$. For $u_0 \in \mathcal{H}_0$, there is $N \in \mathbb{N}$ such that $\|A_n u_0 - Au_0\|_{\mathcal{H}_2} \le \varepsilon/2$ 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 $\|A_{n'} B v_{n'} - ABv_{n'}\|_{\mathcal{H}_2} \ge \delta$. 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 10 *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 eigendecomposition

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

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

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

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. These two facts, together with $\rho_1 > \rho_2$, result in $\delta_n \to 1$. ■

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 complements of ϕ_n and ϕ, respectively, we have convergence of the eigenvector corresponding to the second eigenvalue. It is not difficult to obtain convergence of the eigenspaces corresponding to the m-th eigenvalue in a similar way.

References

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

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, third edition, 2003.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 69(3):337–404, 1950.

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

Charles R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–598, 1985.

Andreas Buja. Remarks on functional canonical variates, alternating least squares methods and ACE. *The Annals of Statistics*, 18(3):1032–1069, 1990.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

Nelson Dunford and Jacob T. Schwartz. *Linear Operators, Part II*. Interscience, 1963.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

Micheal J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, London, 1984.

Arthur Gretton, Olivier Bousquet, Alexander J. Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *16th International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005a.

Arthur Gretton, Alexander J. Smola, Olivier Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schölkopf, and N. Logothetis. Kernel constrained covariance for dependence measurement. In *AISTATS*, volume 10, 2005b.

Charles W. Groetsch. *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. Pitman, 1984.

David R. Hardoon, John Shawe-Taylor, and Ola Friman. KCCA for fMRI analysis. In *Proceedings of Medical Image Understanding and Analysis (London)*, 2004.

Guozhong He, Hans-Georg Müller, and Jane-Ling Wang. Functional canonical analysis for square integrable stochastic procersses. *Journal of Multivariate Analysis*, 85:54–77, 2003.

Peter D. Lax. *Functional Analysis*. Wiley, 2002.

Sue E. Leurgans, Rana A. Moyeed, and Bernard W. Silverman. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society, Series B*, 55(3):725–740, 1993.

Thomas Melzer, Michael Reiter, and Horst Bischof. Nonlinear feature extraction using generalized canonical correlation analysis. In *Proceedings of International Conference on Artificial Neural Networks (ICANN)*, pages 353–360, 2001.

Sebastian Mika, Gunnar Rätsch, Jason Weston, Bernhard Schölkopf, and Klaus-Robert 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.

Michael Reed and Barry Simon. *Functional Analysis*. Academic Press, 1980.

Alfréd Rényi. *Probability Theory*. Horth-Holland, 1970.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, 2002.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

Ingo Steinwart, Don Hush, and Clint Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. Technical Report LA-UR-04-8274, Los Alamos National Laboratory, 2004.

Hiromichi Suetani, Yukito Iba, and Kazuyuki Aihara. Detecting hidden synchronization of chaotic dynamical systems: A kernel-based approach. *Journal of Physics A: Mathematical and General*, 39:10723–10742, 2006.

Nikolai N. Vakhania, Vazha I. Tarieladze, and Sergei A. Chobanyan. *Probability Distributions on Banach Spaces*. D. Reidel Publishing Company, 1987.

Ard W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Yoshihiro Yamanishi, Jean-Philippe Vert, Akihiro Nakaya, and Minoru Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19:323i–330i, 2003.

Laurent Zwald and Gilles Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, 2006.