Journal of Multivariate Analysis 150 (2016) 183–190

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Minimax convergence rates for kernel CCA

Zengyan Fan [a,*], Heng Lian [b]

[a] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371, Singapore

[b] School of Mathematics and Statistics, University of New South Wales, Sydney, 2052, Australia

ARTICLE INFO

Article history:
Received 7 September 2015
Available online 31 May 2016

AMS subject classifications:
62G20
62H25

Keywords:
Canonical correlation analysis
Covariance operator
Cross-covariance operator
Lower bound

ABSTRACT

Consistency of kernel canonical correlation analysis (kernel CCA) has been established while its optimal convergence rate remains unknown. In this paper we derive rigorous upper and lower bounds for the convergence rate of the weight functions in kernel CCA. In particular the optimal convergence rate is shown to only depend on the rate of decay of the eigenvalues of the covariance operators.

1. Introduction

Canonical correlation analysis (CCA), introduced by Harold Hotelling [10], is a popular statistical tool in multivariate analysis. In the population, given two sets of random variables $x = (x_1, \ldots, x_p)^\top$ and $y = (y_1, \ldots, y_q)^\top$, CCA will find linear combinations of components of x and y that have the maximum possible correlation coefficient. Specifically, CCA solves the problem

$$(a, b) = \underset{\mathrm{Var}(a^\top x)=\mathrm{Var}(b^\top y)=1}{\operatorname{argmax}} \mathrm{Cov}(a^\top x, b^\top y).$$

Due to the normalization constraints, the maximum is just the maximal correlation coefficient. Given that recently data are increasingly collected in a high dimensional space, CCA has found renewed interest and applications in various fields (Hardoon and Shawe-Taylor [8], Krafty and Hall [12]).

However, this classical CCA, including its regularized modifications that take into account the high dimensional nature of the variables, is somewhat limited by its linearity which also inherits from the fact that correlation only measures linear dependency. To address this problem, kernel CCA was proposed (Akaho [1], Melzer et al. [14], Bach and Jordan [2]) that uses the popular "kernel trick"(Scholkopf [15]) and maps the original low-dimensional input space where x and y reside to a high-dimensional (typically even infinite-dimensional) feature space. If the linear CCA is performed in the feature space, when mapped back to the input space, we effectively have a nonlinear method. In particular, kernel CCA is able to find nonlinear mappings $f(x)$ and $g(y)$ with maximal possible correlation.

Mathematically, the "kernel trick" can be formalized by using the concept of reproducing kernel Hilbert space (RKHS). Following Wahba [17], an RKHS is a Hilbert space $\mathcal{H}$ consisting of mappings on, say, $\{x : x \in \mathcal{X}\}$ with inner product $\langle . \rangle_{\mathcal{H}}$

* Corresponding author.
E-mail address: zengyanfan@hotmail.com (Z. Fan).

http://dx.doi.org/10.1016/j.jmva.2016.05.008
0047-259X/

where the point evaluation mapping $f \in \mathcal{H} \to f(x)$ is bounded for any x. By the Riesz representation theorem, there exists an element $K(.,x) \in \mathcal{H}$ with $\langle K(.,x), f\rangle = f(x)$. It can be shown that $K(x,y) = \langle K(.,x), K(.,y)\rangle_{\mathcal{H}}$ is a positive definite mapping and $\mathcal{H}$ is the closure of mappings of the form $\sum_{m=1}^{M} a_m K(.,x_m)$. In kernel CCA, we will find mappings $f \in \mathcal{H}_x$ and $g \in \mathcal{H}_y$ where $\mathcal{H}_x$ and $\mathcal{H}_y$ are two RKHSs.

Consistency of kernel CCA has been considered in Balakrishnan et al. [3]. Fukumizu et al. [6] rigorously proved the statistical consistency of weight functions based on i.i.d. data (x_i, y_i), $i = 1, \ldots, n$. The main objective of this paper is to consider the much more challenging problem of convergence rate of kernel CCA, for which we need specific assumptions on the rate of decay of the eigenvalues of the covariance operator, which we will introduce in the next section. Even though this assumption is hard to verify in practice, it is still of significant interest to give an estimation performance benchmark for this popular estimation method. In nonparametric statistics, one always assumes certain smoothness of the function to be estimated to get nontrivial convergence rate, which is also hard to verify. Our theoretical study reveals that performance of kernel CCA depends on decay of eigenvalues of the covariance operators, which is roughly speaking a "smoothness" parameter analogous to nonparametric regression. If the smoothness parameters α_X and α_Y (defined later) become larger, the optimal regularization parameters ϵ_X and ϵ_Y (defined later) become smaller, and the convergence rate is faster. The method of proof can also be of independent interest which can hopefully help to establish convergence rates for other kernel methods. We also provide a matching lower bound showing the upper bound cannot be improved in some sense, which uses a novel construction of suitable models.

The rest of the paper is organized as follows. In Section 2, we first clarify existence issues for kernel CCA which was not stated in Fukumizu et al. [6]. The following two sections established the upper bound and the lower bound respectively. In Section 5, we conclude the paper with a discussion.

2. Kernel CCA in the population

In this section we set up the background, introduce notations and clarify the existence of the maximizing pair of nonlinear mappings (weight functions) f and g.

We assume $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$ are measurable spaces and (x, y) are random elements taking values on $\mathcal{X} \times \mathcal{Y}$ with joint distribution P_{XY}. Let $\mathcal{H}_x$ and $\mathcal{H}_y$ be RKHSs with kernels K_1 and K_2 respectively with $\mathrm{E}_x K_1(x,x) < \infty$ and $\mathrm{E}_y K_2(y,y) < \infty$ where the subscript denotes the random variable over which we take the expectation. This assumption will ensure that the covariance operators introduced below are trace-class operators.

Define the two covariance operators by $\langle f, \Sigma_X f\rangle_{\mathcal{H}_x} = \mathrm{Var}\{f(x)\}$, $\forall f \in \mathcal{H}_x$ and $\langle g, \Sigma_Y g\rangle_{\mathcal{H}_y} = \mathrm{Var}\{g(y)\}$, $\forall g \in \mathcal{H}_y$. Since $\mathrm{Var}\{f(x)\} = \mathrm{E}\langle f, K_1(.,x) - \mathrm{E}_x K_1(.,x)\rangle^2_{\mathcal{H}_x} \le \|f\|^2_{\mathcal{H}_x}\{\mathrm{E}_x K_1(x,x) - \|\mathrm{E}_x K_1(.,x)\|^2_{\mathcal{H}_x}\} < \infty$, covariance operators are well-defined bounded operators. In fact we can write $\Sigma_X = \mathrm{E}_x[\{K_1(.,x) - \mathrm{E}_x K_1(.,x)\} \otimes \{K_1(.,x) - \mathrm{E}_x K_1(.,x)\}]$ and $\Sigma_Y = \mathrm{E}_y[\{K_2(.,y) - \mathrm{E}_y K_2(.,y)\} \otimes \{K_2(.,y) - \mathrm{E}_y K_2(.,y)\}]$ where for $f_1, f_2 \in \mathcal{H}_x$, $f_1 \otimes f_2$ is the operator such that $(f_1 \otimes f_2)h = \langle f_2, h\rangle_{\mathcal{H}_x} f_1$, for example. Since $\mathrm{E}_x K_1(x,x) < \infty$, we have $\mathrm{tr}(\Sigma_X) = \mathrm{E}_x\langle K_1(.,x) - \mathrm{E}_x K_1(.,x), K_1(.,x) - \mathrm{E}_x K_1(.,x)\rangle_{\mathcal{H}_x} < \infty$ and thus Σ_X and Σ_Y are trace-class operators and in particular are Hilbert–Schmidt operators. Furthermore, we define the cross-covariance operators $\Sigma_{XY} = \Sigma_{YX}^{\top} = \mathrm{E}_{x,y}[\{K_1(.,x) - \mathrm{E}_x K_1(.,x)\} \otimes \{K_2(.,y) - \mathrm{E}_y K_2(.,y)\}]$ where $()^{\top}$ denotes the adjoint operator. By the reproducing property, we can easily see $\langle f, \Sigma_{XY} g\rangle_{\mathcal{H}_x} = \mathrm{Cov}\{f(x), g(y)\}$.

Since constants are irrelevant we only speak of mappings modulo constants. That is we regard f and $f + c$, $c \in R$ to be the same mapping and $f \in A$ for a set A means that there exists some constant c such that $f + c \in A$. Since $\mathrm{Var}\{f(x)\} = 0$ when f is a constant, this also has the technical convenience that now the null space of the covariance operator is $\{0\}$. Also due to this reason, we do not explicitly subtract the means for various mappings in the mathematical expressions in the rest of the paper.

Since Σ_X is a self-adjoint Hilbert–Schmidt operator, by spectral theorem we can write

$$\Sigma_X = \sum_{j=1}^{\infty} \lambda_j \phi_j \otimes \phi_j,$$

where $\lambda_1 \ge \lambda_2 \ge \cdots > 0$ are the eigenvalues and ϕ_j are the eigenfunctions with $\langle \phi_j, \phi_k\rangle_{\mathcal{H}_x} = \delta_{jk}$ ($\delta_{jk} = 1$ if $j = k$ and 0 otherwise). Since $\{\phi_j\}$ is an orthonormal basis in $\mathcal{H}_x$, we can write

$$K_1(.,x) = \sum_j \langle K_1(.,x), \phi_j\rangle_{\mathcal{H}_x} \phi_j =: \sum_j \phi_j(x)\phi_j,$$

and we have $\mathrm{E}_x \phi_j(x)\phi_k(x) = \mathrm{Cov}(\phi_j(x), \phi_k(x)) = \langle \phi_j, \Sigma_X \phi_k\rangle_{\mathcal{H}_x} = \lambda_k \delta_{jk}$. Similarly, we can write $\Sigma_Y = \sum_{j=1}^{\infty} \mu_j \psi_j \otimes \psi_j$, $K_2(.,y) = \sum_j \psi_j(y)\psi_j$ with $\mathrm{E}_y \psi_j(y)\psi_k(y) = \mu_k \delta_{jk}$.

In the population, kernel CCA solves the problem

$$\begin{aligned} \rho &= \sup_{\mathrm{Var}(f(x))=\mathrm{Var}(g(y))=1} \mathrm{Cov}\{f(x), g(y)\} \\ &= \sup_{\langle f, \Sigma_X f\rangle_{\mathcal{H}_x} = \langle g, \Sigma_Y g\rangle_{\mathcal{H}_y} = 1} \langle f, \Sigma_{XY} g\rangle_{\mathcal{H}_x}. \end{aligned} \tag{1}$$