Appl. Comput. Harmon. Anal. 41 (2016) 692–712

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Kernel-based conditional canonical correlation analysis via modified Tikhonov regularization

Jia Cai [a,*], Hongwei Sun [b]

[a] School of Mathematics and Statistics, Guangdong University of Finance & Economics, Guangzhou, Guangdong, 510320, China
[b] School of Science, University of Jinan, Jinan, Shandong, 250022, China

ARTICLE INFO

Article history:
Received 23 July 2014
Received in revised form 13 March 2015
Accepted 24 April 2015
Available online 29 April 2015
Communicated by Bin Han

MSC:
68T05
62H20

Keywords:
Kernel CCA
Conditional cross-covariance operator
Reproducing kernel Hilbert space

ABSTRACT

This paper proposes a new conditional kernel CCA (canonical correlation analysis) algorithm and exploits statistical consistency of it via modified Tikhonov regularization scheme, which is a continuous study of [11]. A new measure which characterizes consistency of learning ability is discussed based on the notion of distance between feature subspaces. The consistency analysis is conducted under the assumptions of normalized cross-covariance operators, which is mild and can be constructed by means of mean square contingency. Meantime, the relationship between this new measure and previous consistency scheme is investigated. Furthermore, we study conditional kernel CCA in a more general scenario by means of the trace operator.

1. Introduction

Data dependence analysis is one of the main concerns of statistical inference. Proposed by Hotelling [17], canonical correlation analysis (CCA) is a powerful method of multivariate data analysis, which aims at seeking a pair of linear transformations associated with the two sets of variables such that the projected variables are maximally correlated. The optimal pair of linear transformations can be solved by a generalized eigenvalue problem [2]. Due to linearity, CCA cannot capture nonlinear relations. Hence, kernelization of CCA (kernel CCA) was introduced [1,12]. A pattern function in a Euclidean space was defined to study convergence analysis of kernel CCA via Rademacher complexity [15], while statistical consistency of it was investigated under a decay condition of regularization parameters [10]. Cai and Sun [4] conducted convergence rates of it under AC condition. Currently, kernel CCA has been widely used in many fields of

* Corresponding author.
E-mail addresses: jiacai1999@gdufe.edu.cn (J. Cai), ss_sunhw@ujn.edu.cn (H. Sun).

http://dx.doi.org/10.1016/j.acha.2015.04.006
1063-5203/

science and technology, including: biology and neurology [14,22], bioinformatics [23], image retrieval [13], cross-language text retrieval [24], etc.

The notion of conditional correlation arises from the problems of, e.g., chaotic time series, graphical modeling of medical data [11] and causal learning [21]. Causal learning detects the causal structure of the events. Causal knowledge and beliefs play a significant role in much of our everyday cognition. A lot of theoretical analysis has been done to explain human causal learning. Thus, usually we need to consider the dependence between X and Y given another variable Z. This is essentially different from kernel CCA, which only focuses on describing the relations between two variables. Fukumizu et al. [11] proposed a new measure of conditional dependence based on the normalized conditional cross-covariance operators. However, to the best of our knowledge, no literature gives a systematic study about the model, convergence analysis and geometry structure of conditional kernel CCA, which involves three variables. Hence it is more complicated. This paper extends their work (Theorem 5) and aims at providing a suitable measure to characterize the consistency of estimated functions from i.i.d. sample to their population counterparts when they are not unique. Furthermore, the convergence rates of empirical normalized conditional cross-covariance operator (NCCCO) to the NCCCO are also addressed in the sense of Hilbert–Schmidt norm under mild conditions, which is the extension of Theorem 5 in [11]. Meantime, we generalize conditional kernel CCA to multiple setting by means of the trace operator, and the conclusion stated in Theorem 3 can be viewed as an extension of Theorem 3.1 in [5].

The rest of the paper is organized as follows. In Section 2, we give a brief review of kernel CCA problem and introduce a new notion of conditional kernel CCA. The key analysis and main results will also be investigated. Section 3 devotes to the extension of multiple conditional kernel CCA. Proof of main results goes to Section 4. Finally, some concluding remarks are given in Section 5.

2. Theoretical background and main results

In this section, we make a new and systematic study on conditional kernel CCA, and develop a new appropriate consistency analysis elaborately. Let us first review the kernel CCA problem.

2.1. Brief review of kernel CCA

Define the norm of a bounded linear operator $\mathbf{A}$ from a Banach space $(\mathcal{H}_1, \|\cdot\|_{\mathcal{H}_1})$ to another Banach space $(\mathcal{H}_2, \|\cdot\|_{\mathcal{H}_2})$ as $\|\mathbf{A}\| = \sup_{\|f\|_{\mathcal{H}_1}=1} \|\mathbf{A}f\|_{\mathcal{H}_2}$. The null space and the range of an operator $\mathbf{A}$ are denoted by $\mathcal{N}(\mathbf{A})$ and $\mathcal{R}(\mathbf{A})$ respectively, where $\mathcal{N}(\mathbf{A}) = \{f \in \mathcal{H}_1 | \mathbf{A}f = 0\}$ and $\mathcal{R}(\mathbf{A}) = \{\mathbf{A}f \in \mathcal{H}_2 | f \in \mathcal{H}_1\}$.

Throughout the paper, $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$ and $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$ are measurable spaces. $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$, $(\mathcal{H}_{\mathcal{Y}}, k_{\mathcal{Y}})$ and $(\mathcal{H}_{\mathcal{Z}}, k_{\mathcal{Z}})$ are RKHSs (see [7,8] and the references therein) of real-valued functions on $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ respectively, endowed with measurable positive semi-definite kernels $k_{\mathcal{X}}$, $k_{\mathcal{Y}}$ and $k_{\mathcal{Z}}$, respectively. We assume that they satisfy

$$\kappa_1 = \sup_{X \in \mathcal{X}} |k_{\mathcal{X}}(X, X)| < \infty, \quad \kappa_2 = \sup_{Y \in \mathcal{Y}} |k_{\mathcal{Y}}(Y, Y)| < \infty \quad \text{and} \quad \kappa_3 = \sup_{Z \in \mathcal{Z}} |k_{\mathcal{Z}}(Z, Z)| < \infty. \tag{2.1}$$

Given two random variables X and Y, kernel CCA aims at providing nonlinear mappings $f(X)$ and $g(Y)$ such that their correlation is maximized, where $f \neq 0$ and $g \neq 0$ belongs to $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, respectively. That is [10]:

$$\max_{f \in \mathcal{H}_{\mathcal{X}}, g \in \mathcal{H}_{\mathcal{Y}}} \frac{\mathrm{Cov}[f(X), g(Y)]}{\mathrm{Var}[f(X)]^{1/2} \mathrm{Var}[g(Y)]^{1/2}}. \tag{2.2}$$

Kernel CCA problem can be well-formulated using cross-covariance operators [3,10].

$$\langle g, \mathbf{\Sigma}_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} = \mathrm{Cov}[f(X), g(Y)] = \mathbb{E}_{XY}\Big[\langle f, k_\mathcal{X}(\cdot, X) - m_X\rangle_{\mathcal{H}_\mathcal{X}} \langle k_\mathcal{Y}(\cdot, Y) - m_Y, g\rangle_{\mathcal{H}_\mathcal{Y}}\Big], \forall f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y},$$

where the mean element $m_X \in \mathcal{H}_\mathcal{X}$ (similarly for m_Y) with respect to X is defined as

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

Cross-covariance operators are introduced on RKHSs (reproducing kernel Hilbert spaces) where the theory is much simpler while they are generally defined for random variables in Banach spaces [3]. Therefore

$$m_X = \mathbb{E}_X[k_\mathcal{X}(\cdot, X)], \qquad \mathbf{\Sigma}_{YX} = \mathbb{E}_{XY}[(k_\mathcal{X}(\cdot, X) - m_X) \otimes (k_\mathcal{Y}(\cdot, Y) - m_Y)].$$

It is easy to see that $\mathbf{\Sigma}_{YX} = \mathbf{\Sigma}^*_{XY}$, where $\mathbf{A}^*$ denotes the adjoint of an operator $\mathbf{A}$. If $Y = X$, $\mathbf{\Sigma}_{XX}$ is called the covariance operator, which is self-adjoint and positive. Hence, (2.2) can be reformulated as

$$\max_{f\in\mathcal{H}_\mathcal{X}, g\in\mathcal{H}_\mathcal{Y}} \frac{\langle g, \mathbf{\Sigma}_{YX} f\rangle}{\sqrt{\langle g, \mathbf{\Sigma}_{YY} g\rangle}\sqrt{\langle f, \mathbf{\Sigma}_{XX} f\rangle}}. \tag{2.3}$$

This yields the following conclusion immediately. The proof is simple and obvious, we will omit it here.

Proposition 1. *(1) For any $f, g \in \mathcal{H}_\mathcal{X}$, there holds*

$$\langle \mathbf{\Sigma}_{XX} f, g\rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}[(f(X)g(X))] - \mathbb{E}[f(X)]\mathbb{E}[g(X)];$$

(2) For any $g \in \mathcal{H}_\mathcal{Y}$, assume $\mathbb{E}_{Y|X}[g(Y)|X = \cdot] \in \mathcal{H}_\mathcal{X}$ and $k_\mathcal{X}(X, X)$, $k_\mathcal{Y}(Y, Y)$ satisfy Eq. (2.1), then there holds

$$\mathbf{\Sigma}_{XX}\mathbb{E}_{Y|X}[g(Y)|X = \cdot] = \mathbf{\Sigma}_{XY} g.$$

Remark 1. In essence, $\mathbb{E}_{Y|X}[g(Y)|X = \cdot] \in \mathcal{H}_\mathcal{X}$ can be satisfied as shown in Proposition 4, [9] without any reference to a specific g. The assumption for conclusion (2) in Proposition 1 could be relaxed to $\mathbb{E}_X[k_\mathcal{X}(X, X)] < \infty$ and $\mathbb{E}_Y[k_\mathcal{Y}(Y, Y)] < \infty$.

Employing the ideas of cross-covariance operators, a comprehensive description about conditional kernel CCA will be presented in the next section.

2.2. Conditional kernel CCA algorithm

In this paper, our purpose is to analyze the effect of variable Z to the dependence between X and Y. We define our conditional kernel CCA algorithm as

$$\max_{f\in\mathcal{H}_\mathcal{X}, g\in\mathcal{H}_\mathcal{Y}} \frac{\mathbb{E}_Z[\mathrm{Cov}_{XY|Z}[f(X), g(Y)]|Z]}{\mathrm{Var}[f(X)]^{1/2}\mathrm{Var}[g(Y)]^{1/2}}, \tag{2.4}$$

which are motivated from the expression forms of kernel CCA in [10] and the theory of conditional cross-covariance operator in [9]. Theoretical consistency about this algorithm will be given in the sequel. The reason why we did not consider conditional variances in the denominator, viz, $\mathrm{Var}[f(X)|Z]$ and $\mathrm{Var}[g(Y)|Z]$ will also be explained from the operator viewpoint. A one-step analysis shows that

$$\begin{aligned}\langle g, \mathbf{\Sigma}_{YX|Z} f\rangle_{\mathcal{H}_\mathcal{Y}} &= \mathbb{E}_Z[\mathrm{Cov}_{XY|Z}[f(X), g(Y)]|Z] \\ &= \mathbb{E}_Z\left[\mathbb{E}_{XY}[f(X)g(Y)|Z] - \mathbb{E}_{X|Z}(f(X)|Z)\mathbb{E}_{Y|Z}(g(Y)|Z)\right] \\ &= \mathbb{E}_{XY}[f(X)g(Y)] - \mathbb{E}_Z\Big([\mathbb{E}_{X|Z}(f(X)|Z)][\mathbb{E}_{Y|Z}(g(Y)|Z)]\Big) \\ &= \langle g, \mathbf{\Sigma}_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} - \langle \mathbf{\Sigma}_{ZZ}\mathbb{E}_{X|Z}(f(X)|Z), \mathbb{E}_{Y|Z}(g(Y)|Z)\rangle_{\mathcal{H}_\mathcal{Z}}.\end{aligned}$$

Applying Proposition 1, we get

$$\begin{aligned}\langle g, \mathbf{\Sigma}_{YX|Z} f\rangle_{\mathcal{H}_\mathcal{Y}} &= \langle g, \mathbf{\Sigma}_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} - \langle \mathbf{\Sigma}_{ZX} f, \mathbb{E}_{Y|Z}(g(Y)|Z)\rangle_{\mathcal{H}_\mathcal{Z}} \\ &= \langle g, \mathbf{\Sigma}_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} - \langle \mathbf{\Sigma}_{ZZ}^{-1}\mathbf{\Sigma}_{ZX} f, \mathbf{\Sigma}_{ZY} g\rangle_{\mathcal{H}_\mathcal{Z}} \\ &= \langle g, (\mathbf{\Sigma}_{YX} - \mathbf{\Sigma}_{YZ}\mathbf{\Sigma}_{ZZ}^{-1}\mathbf{\Sigma}_{ZX}) f\rangle_{\mathcal{H}_\mathcal{Y}}.\end{aligned}$$

The operator $\mathbf{\Sigma}_{ZZ}$ is typically not invertible, in which scenario, one can use the right inverse of $\mathbf{\Sigma}_{ZZ}$ on $(\mathcal{N}(\mathbf{\Sigma}_{ZZ}))^\perp$ to replace $\mathbf{\Sigma}_{ZZ}^{-1}$ as it did in [9]. Alternatively, one may represent $\mathbf{\Sigma}_{YX}$ as $\mathbf{\Sigma}_{YY}^{1/2}\mathbf{V}_{YX}\mathbf{\Sigma}_{XX}^{1/2}$, where $\mathbf{V}_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ is a unique bounded operator such that $\|\mathbf{V}_{YX}\| \leq 1$, and $\mathbf{V}_{YX} = Q_Y\mathbf{V}_{YX}Q_X$ (Theorem 1, [3]), here Q_X, Q_Y are the orthogonal projections that map $\mathcal{H}_\mathcal{X}$ onto $\overline{\mathcal{R}(\mathbf{\Sigma}_{XX})}$ and $\mathcal{H}_\mathcal{Y}$ onto $\overline{\mathcal{R}(\mathbf{\Sigma}_{YY})}$, respectively. It is called normalized cross-covariance operator (NOCCO). Similar properties hold for $\mathbf{\Sigma}_{YZ}$ and $\mathbf{\Sigma}_{ZX}$. Thus

$$\mathbf{\Sigma}_{YX|Z} = \mathbf{\Sigma}_{YY}^{1/2}(\mathbf{V}_{YX} - \mathbf{V}_{YZ}\mathbf{V}_{ZX})\mathbf{\Sigma}_{XX}^{1/2}.$$

Denote $\mathbf{V}_{YX|Z} = \mathbf{V}_{YX} - \mathbf{V}_{YZ}\mathbf{V}_{ZX}$, then

$$\mathbf{\Sigma}_{YX|Z} = \mathbf{\Sigma}_{YY}^{1/2}\mathbf{V}_{YX|Z}\mathbf{\Sigma}_{XX}^{1/2}.$$

If we rewrite (2.4) as

$$\max_{f\in\mathcal{H}_\mathcal{X}, g\in\mathcal{H}_\mathcal{Y}} \frac{\langle g, \mathbf{\Sigma}_{YX|Z} f\rangle}{\sqrt{\langle g, \mathbf{\Sigma}_{YY} g\rangle}\sqrt{\langle f, \mathbf{\Sigma}_{XX} f\rangle}},$$

and let $\tilde{f} = \mathbf{\Sigma}_{XX}^{1/2} f$, $\tilde{g} = \mathbf{\Sigma}_{YY}^{1/2} g$, the above expression can be reformulated as

$$\max_{\|\tilde{f}\|_{\mathcal{H}_\mathcal{X}}=1, \|\tilde{g}\|_{\mathcal{H}_\mathcal{Y}}=1} \langle \tilde{g}, \mathbf{V}_{YX|Z}\tilde{f}\rangle.$$

Recall in the theoretical analysis of kernel CCA, namely (2.3) which is equivalent to

$$\max_{\|\tilde{f}\|_{\mathcal{H}_\mathcal{X}}=1, \|\tilde{g}\|_{\mathcal{H}_\mathcal{Y}}=1} \langle \tilde{g}, \mathbf{V}_{YX}\tilde{f}\rangle.$$

Therefore, our model (2.4) was motivated by replacing $\mathbf{V}_{YX}$ with $\mathbf{V}_{YX|Z}$. One advantage of this approach is that conditional variances are circumvented and conditional dependence information is included in $\mathbf{V}_{YX|Z}$. Hence algorithm (2.4) is meaningful from the operator viewpoint.

Given an i.i.d. sample $(X_1, Y_1, Z_1), \cdots, (X_m, Y_m, Z_m)$ drawn from an unknown probability distribution ρ, classical methods aim at providing efficient estimates for algorithm (2.4) based on a finite sample. Recall [10]

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

Therefore

$$\widehat{\mathbf{\Sigma}}_{YX}^{(m)} = \frac{1}{m}\sum_{i=1}^{m}\Big(k_{\mathcal{Y}}(\cdot,Y_i) - \frac{1}{m}\sum_{s=1}^{m}k_{\mathcal{Y}}(\cdot,Y_s)\Big)\otimes\Big(k_{\mathcal{X}}(\cdot,X_i) - \frac{1}{m}\sum_{t=1}^{m}k_{\mathcal{X}}(\cdot,X_t)\Big).$$

Obviously, $\widehat{\mathbf{\Sigma}}_{YX}^{(m)}$ is a finite rank operator. ERM scheme with Tikhonov regularization yields an empirical estimate of (2.4). But here we consider a slightly different one:

$$\max_{f\in\mathcal{H}_{\mathcal{X}},g\in\mathcal{H}_{\mathcal{Y}}}\frac{\widehat{\mathbb{E}}_{\varepsilon_m}[\widehat{\mathrm{Cov}}_{XY|Z}[f(X),g(Y)]|Z]}{(\widehat{\mathrm{Var}}[f(X)]+\varepsilon_m\|f\|_{\mathcal{H}_{\mathcal{X}}}^2)^{1/2}(\widehat{\mathrm{Var}}[g(Y)]+\varepsilon_m\|g\|_{\mathcal{H}_{\mathcal{Y}}}^2)^{1/2}}, \tag{2.5}$$

where $\varepsilon_m > 0$ is the regularization coefficient, $\widehat{\mathrm{Var}}(f(X)) = \frac{1}{m}\sum_{i=1}^{m}(f(X_i) - \frac{1}{m}\sum_{t=1}^{m}f(X_t))^2$ ($\widehat{\mathrm{Var}}(g(Y))$ similarly), and

$$\widehat{\mathbb{E}}_{\varepsilon_m}[\widehat{\mathrm{Cov}}_{XY|Z}[f(X),g(Y)]|Z] = \langle g, \widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)} f\rangle = \langle g, \widehat{\mathbf{\Sigma}}_{YX}^{(m)} f\rangle - \langle g, \widehat{\mathbf{\Sigma}}_{YZ}^{(m)}(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)}+\varepsilon_m I)^{-1}\widehat{\mathbf{\Sigma}}_{ZX}^{(m)} f\rangle.$$

The operator $\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)}$ is not invertible, therefore modified Tikhonov regularization scheme was used to perform the approximation. That is, we use regularization terms not only in the denominator but also in the numerator, and we replace $(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{-1}$ with $(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)}+\varepsilon_m I)^{-1}$. The regularization terms make the problem well-formulated in order to avoid trivial learning [15]. Recall

$$\widehat{\mathbf{V}}_{YX}^{(m)} = (\widehat{\mathbf{\Sigma}}_{YY}^{(m)}+\varepsilon_m I)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}^{(m)}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)}+\varepsilon_m I)^{-1/2}.$$

Define $\widehat{\mathbf{V}}_{YX|Z}^{(m)} = \widehat{\mathbf{V}}_{YX}^{(m)} - \widehat{\mathbf{V}}_{YZ}^{(m)}\widehat{\mathbf{V}}_{ZX}^{(m)}$ (empirical NCCCO), then $\widehat{\mathbf{V}}_{YX|Z}^{(m)}$ is a good approximation of $\mathbf{V}_{YX|Z}$. Direct calculation yields that

$$\widehat{\mathbf{V}}_{YX|Z}^{(m)} = (\widehat{\mathbf{\Sigma}}_{YY}^{(m)}+\varepsilon_m I)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)}+\varepsilon_m I)^{-1/2}.$$

The solution of (2.5) can be expressed using the idea of Gram matrices. Let $u_i \in \mathcal{H}_{\mathcal{X}}$, $v_i \in \mathcal{H}_{\mathcal{Y}}$ and $w_i \in \mathcal{H}_{\mathcal{Z}}$ be functions defined by $u_i = k_{\mathcal{X}}(\cdot,X_i) - \frac{1}{m}\sum_{t=1}^{m}k_{\mathcal{X}}(\cdot,X_t)$, $v_i = k_{\mathcal{Y}}(\cdot,Y_i) - \frac{1}{m}\sum_{t=1}^{m}k_{\mathcal{Y}}(\cdot,Y_t)$, $\omega_i = k_{\mathcal{Z}}(\cdot,Z_i) - \frac{1}{m}\sum_{t=1}^{m}k_{\mathcal{Z}}(\cdot,Z_t)$, and G_X, G_Y, G_Z are centered Gram matrices, such that $(G_X)_{i,j} = \langle u_i,u_j\rangle_{\mathcal{H}_{\mathcal{X}}}$, $(G_Y)_{i,j} = \langle v_i,v_j\rangle_{\mathcal{H}_{\mathcal{Y}}}$, $(G_Z)_{i,j} = \langle \omega_i,\omega_j\rangle_{\mathcal{H}_{\mathcal{Z}}}$. Intuitively, employing the methods that discussed for (2.4), we can reformulate (2.5) as

$$\begin{aligned}
&\max_{f\in\mathcal{H}_{\mathcal{X}},g\in\mathcal{H}_{\mathcal{Y}}}\frac{\langle g,\widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)} f\rangle}{\sqrt{\langle g,(\widehat{\mathbf{\Sigma}}_{YY}^{(m)}+\varepsilon_m I)g\rangle}\sqrt{\langle f,(\widehat{\mathbf{\Sigma}}_{XX}^{(m)}+\varepsilon_m I)f\rangle}},\\
&= \max_{\|\phi\|_{\mathcal{H}_{\mathcal{X}}}=\|\psi\|_{\mathcal{H}_{\mathcal{Y}}}=1}\langle\psi,\widehat{\mathbf{V}}_{YX|Z}^{(m)}\phi\rangle\\
&= \max_{\|\phi\|_{\mathcal{H}_{\mathcal{X}}}=\|\psi\|_{\mathcal{H}_{\mathcal{Y}}}=1}\langle\psi,(\widehat{\mathbf{\Sigma}}_{YY}^{(m)}+\varepsilon_m I)^{-1/2}\big[\widehat{\mathbf{\Sigma}}_{YX}^{(m)} - \widehat{\mathbf{\Sigma}}_{YZ}^{(m)}(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)}+\varepsilon_m I)^{-1}\widehat{\mathbf{\Sigma}}_{ZX}^{(m)}\big](\widehat{\mathbf{\Sigma}}_{XX}^{(m)}+\varepsilon_m I)^{-1/2}\phi\rangle.
\end{aligned}$$

Cause $\mathcal{R}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)})$, $\mathcal{R}(\widehat{\mathbf{\Sigma}}_{YY}^{(m)})$, $\mathcal{R}(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})$ are spanned by $(u_i)_{i=1}^m$, $(v_i)_{i=1}^m$ and $(\omega_i)_{i=1}^m$, respectively, then the unit eigenfunction pair $(\hat{\phi},\hat{\psi})$ of $\widehat{\mathbf{V}}_{YX|Z}^{(m)}$ corresponding to the largest singular value can be given by a linear combination of u_i and v_i. Let $\hat{\phi} = \sum_{i=1}^{m}\hat{\alpha}_i u_i$, $\hat{\psi} = \sum_{i=1}^{m}\hat{\beta}_i v_i$. Recall

$$\widehat{\mathbf{\Sigma}}_{YX}^{(m)} = \frac{1}{m}\sum_{i=1}^{m}\Big(k_{\mathcal{Y}}(\cdot,Y_i) - \frac{1}{m}\sum_{s=1}^{m}k_{\mathcal{Y}}(\cdot,Y_s)\Big)\otimes\Big(k_{\mathcal{X}}(\cdot,X_i) - \frac{1}{m}\sum_{t=1}^{m}k_{\mathcal{X}}(\cdot,X_t)\Big) = \frac{1}{m}\sum_{i=1}^{m}v_i\otimes u_i,$$

hence $\widehat{\mathbf{\Sigma}}_{YX}^{(m)}u_j = \frac{1}{m}\sum_{i=1}^{m}(G_X)_{ij}v_i$. Moreover, $\langle\psi, \widehat{\mathbf{\Sigma}}_{YX}^{(m)}\phi\rangle = \langle\beta, \frac{1}{m}G_YG_X\alpha\rangle$ for all $\phi = \sum_{i=1}^{m}\alpha_iu_i$, $\psi = \sum_{i=1}^{m}\beta_iv_i$. Similar expressions hold for $\widehat{\mathbf{\Sigma}}_{YZ}^{(m)}$, $\widehat{\mathbf{\Sigma}}_{ZX}^{(m)}$ and $\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)}$. Therefore [10],

$$\langle\psi, (\widehat{\mathbf{\Sigma}}_{YY}^{(m)} + \varepsilon_m I)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}^{(m)}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{-1/2}\phi\rangle = \langle\beta, (G_Y + m\varepsilon_m I)^{-1/2}G_YG_X(G_X + m\varepsilon_m I)^{-1/2}\alpha\rangle.$$

Applying the conclusion in [9], we can get a similar result

$$\begin{aligned}&\langle\psi, (\widehat{\mathbf{\Sigma}}_{YY}^{(m)} + \varepsilon_m I)^{-1/2}\widehat{\mathbf{\Sigma}}_{YZ}^{(m)}(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)} + \varepsilon_m I)^{-1}\widehat{\mathbf{\Sigma}}_{ZX}^{(m)}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{-1/2}\phi\rangle\\ &\quad= \langle\beta, (G_Y + m\varepsilon_m I)^{-1/2}G_YG_Z(G_Z + m\varepsilon_m I)^{-1}G_ZG_X(G_X + m\varepsilon_m I)^{-1/2}\alpha\rangle.\end{aligned}$$

The coefficients $\hat{\alpha}$, $\hat{\beta}$ therefore should satisfy

$$\max_{\substack{\alpha,\beta\in\mathbb{R}^m\\ \alpha^TG_X\alpha=\beta^TG_Y\beta=1}} \langle\beta, (G_Y + m\varepsilon_m I)^{-1/2}\widetilde{G}_{XYZ}(G_X + m\varepsilon_m I)^{-1/2}\alpha\rangle,$$

where $\widetilde{G}_{XYZ} = G_YG_X - G_YG_Z(G_Z + m\varepsilon_m I)^{-1}G_ZG_X$. By the theory introduced in [2], we can see that the solution of (2.5) can be expressed as

$$\hat{f} = (\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{-1/2}\hat{\phi} = \sum_{i=1}^{m}\hat{\theta}_iu_i, \qquad \hat{g} = (\widehat{\mathbf{\Sigma}}_{YY}^{(m)} + \varepsilon_m I)^{-1/2}\hat{\psi} = \sum_{i=1}^{m}\hat{\iota}_iv_i.$$

Simple calculations lead to

$$\hat{\theta} = \sqrt{m}(G_X + m\varepsilon_m I)^{-1/2}\hat{\alpha}, \qquad \hat{\iota} = \sqrt{m}(G_Y + m\varepsilon_m I)^{-1/2}\hat{\beta},$$

where $\hat{\theta}$, $\hat{\iota}$ are the solutions of

$$\max_{\substack{\theta,\iota\in\mathbb{R}^m\\ \theta^T(G_X^2+m\varepsilon_mG_X)\theta=\iota^T(G_Y^2+m\varepsilon_mG_Y)\iota=m}} \iota^T\widetilde{G}_{XYZ}\theta.$$

Also note that we can use different regularization parameters for algorithm (2.5), but here we consider a simpler one.

The Hilbert–Schmidt norm of empirical NCCCO $\widehat{\mathbf{V}}_{YX|Z}^{(m)}$ and NCCCO $\mathbf{V}_{YX|Z}$ encodes the dependence structure of random variables X and Y given Z. Ref. [11] states the convergence of $\|\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{YX|Z}\|_{\mathrm{HS}}$, but the convergence rates of it remain open and will be elucidated under mild conditions. Another contribution of this paper is that a new consistency measure is proposed for the multidimensional feature learning (the unit eigenfunctions corresponding to the largest singular value of $\mathbf{V}_{YX|Z}$ are not unique), which generalizes the classic consistency measure involving variance. Furthermore, key analysis extends to a more general case (multiple setting).

2.3. Consistency analysis

Let (η_i, ξ_i), $i \in \mathbb{N}$ be the unit eigenfunction pairs of $\mathbf{V}_{YX|Z}$, and

$$\mathbf{V}_{YX|Z} = \sum_{i=1}^{\infty}\sigma_i\eta_i \otimes \xi_i,$$

where the singular values σ_i, $i \in \mathbb{N}$ are arranged in nonincreasing order, ξ_i, η_i $(i \in \mathbb{N})$ are the orthonormal systems of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, respectively, and

$$\sigma_1 = \langle \eta_i, \mathbf{V}_{YX|Z}\xi_i\rangle_{\mathcal{H}_\mathcal{Y}} = \max_{\substack{f\in\mathcal{H}_\mathcal{X}, g\in\mathcal{H}_\mathcal{Y} \\ ||f||_{\mathcal{H}_\mathcal{X}}=||g||_{\mathcal{H}_\mathcal{Y}}=1}} \langle g, \mathbf{V}_{YX|Z} f\rangle_{\mathcal{H}_\mathcal{Y}} \qquad (i=1,\cdots,l),$$

where σ_1 is the largest singular value of $\mathbf{V}_{YX|Z}$. Similarly, in the empirical case, let $\widehat{\xi}_i$, $i=1,\cdots,\bar{r}$ ($\widehat{\eta}_i$, similarly) be the unit orthonormal eigenfunctions of the finite rank operator $\widehat{\mathbf{V}}^{(m)}_{YX|Z}$, where the corresponding singular values $\widehat{\sigma}_i$, $i=1,\cdots,\bar{r}$ are arranged in nonincreasing order. Our purpose is to learn f, g, hence we can take the estimators as follows

$$\widehat{g} = (\widehat{\mathbf{\Sigma}}^{(m)}_{YY} + \varepsilon_m I)^{-1/2}\widehat{\eta}_1, \qquad \widehat{f} = (\widehat{\mathbf{\Sigma}}^{(m)}_{XX} + \varepsilon_m I)^{-1/2}\widehat{\xi}_1. \tag{2.6}$$

In our setting, we focus on the case that σ_1 with multiplicity, i.e., $\sigma_1 = \sigma_2 = \cdots = \sigma_l > \sigma_{l+1}$, $l\in\mathbb{N}$, $l\geq 1$. The target function is in the subspace $S_1 = \text{span}\{\xi_i\}_{i=1}^l$. Analogous discussion holds for η_i ($i=1,\cdots,l$). This leads to additional difficulty in the theoretical analysis, and makes it challenging to establish appropriate measure to characterize the statistical consistency. It is meaningless to consider the convergence of

$$\left\|\mathbf{\Sigma}^{1/2}_{XX}(\widehat{f}-f)\right\|_{\mathcal{H}_\mathcal{X}}, \qquad \left\|\mathbf{\Sigma}^{1/2}_{YY}(\widehat{g}-g)\right\|_{\mathcal{H}_\mathcal{Y}},$$

as the traditional way in [4] and [10], because the solutions (g,f) and $(\widehat{g},\widehat{f})$ are not unique. A natural approach is to consider the distance between feature subspaces spanned by the corresponding eigenfunctions. That is $d(S_1,S_2) = \sqrt{l - \sum_{i=1}^l\sum_{j=1}^l \langle \xi_i, \widehat{\xi}_j\rangle^2}$, where $S_2 = \text{span}\{\widehat{\xi}_j\}_{j=1}^l$.

This measure is motivated from the one that introduced in the community of document retrieval analysis. It is well known that semantic space models provide a numerical representation of words' meaning extracted from corpus of documents. Semantic association between words in a semantic space is crucial. When considering semantic space as a subspace of a more general Hilbert space, the relationship between semantic spaces are captured by means of subspace distance [25]. If we consider eigenfunctions of $\mathbf{V}_{YX|Z}$ and $\widehat{\mathbf{V}}^{(m)}_{YX|Z}$ as different word vectors from different corpus of documents, we therefore can use this measure to characterize the consistency of conditional kernel CCA. Our purpose is to learn the l unit eigenfunctions ξ_i ($i=1,\cdots,l$). So we define S_2 to be the span of l unit eigenfunctions $\widehat{\xi}_j$ ($j=1,\cdots,l$). Note that $\widehat{\xi}_j$ ($j=1,\cdots,l$) may corresponding to the second or third largest singular values of $\widehat{\mathbf{V}}^{(m)}_{YX|Z}$. In fact, let $\mathbf{P}$ be a projector operator, $\mathbf{P}_{S_1}$ and $\mathbf{P}_{S_2}$ are the projection onto S_1 and S_2, respectively. We have

$$d(S_1,S_2) = \sqrt{l - \sum_{i=1}^l\sum_{j=1}^l \langle \xi_i, \widehat{\xi}_j\rangle^2} = \sqrt{l - \text{trace}(\mathbf{P}_{S_1}\mathbf{P}_{S_2})}.$$

Hence the proposed distance measure is the generalization of the measure for one-dimensional setting (see below in Remark 4). Therefore the distance comparison between different spaces (especially for semantic spaces) will give insights to a better understanding about the geometry structure of feature subspaces for conditional kernel CCA. Main results concerning $d(S_1,S_2)$ will be given in the sequel.

Let $\{\lambda_s\}_{s=1}^\infty$, $\{\mu_s\}_{s=1}^\infty$, $\{\nu_s\}_{s=1}^\infty$ be the set of nonzero eigenvalues of $\mathbf{\Sigma}_{XX}$, $\mathbf{\Sigma}_{YY}$ and $\mathbf{\Sigma}_{ZZ}$ respectively, satisfying $\lambda_1\geq\lambda_2\geq\ldots>0$, $\mu_1\geq\mu_2\geq\ldots>0$ and $\nu_1\geq\nu_2\geq\ldots>0$. $\{\phi_s\}_{s\geq 1}$, $\{\psi_s\}_{s\geq 1}$ and $\{\varphi_s\}_{s\geq 1}$ are the corresponding unit eigenfunctions respectively. Then $\mathbf{\Sigma}_{XX} = \sum_{s=1}^\infty \lambda_s\phi_s\otimes\phi_s$, $\mathbf{\Sigma}_{YY} = \sum_{s=1}^\infty \mu_s\psi_s\otimes\psi_s$ and $\mathbf{\Sigma}_{ZZ} = \sum_{s=1}^\infty \nu_s\varphi_s\otimes\varphi_s$.

The following assumption establishes some connections between the normalized cross-covariance operators (NOCCOs) and the eigenvalues of covariance operators. It plays significant role in the theoretical analysis.

Approximation Assumption (AA): *Suppose that there exist $r, q, t > 0$ satisfying*

$$M_1 \triangleq \max\left\{\Big(\sum_{s=1}^{\infty}\frac{\|\mathbf{V}_{YX}\phi_s\|^2_{\mathcal{H}_\mathcal{Y}}}{\lambda_s^{2r}}\Big)^{1/2}, \Big(\sum_{s=1}^{\infty}\frac{\|\mathbf{V}_{XY}\psi_s\|^2_{\mathcal{H}_\mathcal{X}}}{\mu_s^{2r}}\Big)^{1/2}\right\} < \infty,$$

$$M_2 \triangleq \max\left\{\Big(\sum_{s=1}^{\infty}\frac{\|\mathbf{V}_{YZ}\varphi_s\|^2_{\mathcal{H}_\mathcal{Y}}}{\nu_s^{2q}}\Big)^{1/2}, \Big(\sum_{s=1}^{\infty}\frac{\|\mathbf{V}_{ZY}\psi_s\|^2_{\mathcal{H}_\mathcal{Z}}}{\mu_s^{2q}}\Big)^{1/2}\right\} < \infty,$$

$$M_3 \triangleq \max\left\{\Big(\sum_{s=1}^{\infty}\frac{\|\mathbf{V}_{ZX}\phi_s\|^2_{\mathcal{H}_\mathcal{Z}}}{\lambda_s^{2t}}\Big)^{1/2}, \Big(\sum_{s=1}^{\infty}\frac{\|\mathbf{V}_{XZ}\varphi_s\|^2_{\mathcal{H}_\mathcal{X}}}{\nu_s^{2t}}\Big)^{1/2}\right\} < \infty.$$

Remark 2. Firstly, we give some explanations about AA. The first condition involves M_1 was used to achieve a convergence rate of $\|\widehat{\mathbf{V}}_{YX}^{(m)} - \mathbf{V}_{YX}\|$, the others are similar. Let us give a detailed explanation about M_1. The first assumption condition is equivalent to that there exist operators $W_1, W_2 \in HS(\mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y})$ such that $V_{YX} = W_1\mathbf{\Sigma}_{XX}^r$ and $V_{XY} = \mathbf{\Sigma}_{YY}^r W_2$, where $HS(\mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y})$ means Hilbert–Schmidt operator from $\mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$. $\mathbf{\Sigma}_{XX}^r$, $\mathbf{\Sigma}_{YY}^r$ are the r-th power of $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{YY}$, respectively, take form $\mathbf{\Sigma}_{XX}^r = \sum_{s=1}^{\infty}\lambda_i^r\phi_s\otimes\phi_s$, $\mathbf{\Sigma}_{YY}^r = \sum_{s=1}^{\infty}\mu_i^r\psi_s\otimes\psi_s$. It can be rewritten as $V_{YX} \in HS(\mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y})\mathbf{\Sigma}_{XX}^r$ and $V_{XY} \in \mathbf{\Sigma}_{YY}^r HS(\mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y})$. It also means $\mathbf{V}_{YX}\mathbf{\Sigma}_{XX}^{-r}$ and $\mathbf{\Sigma}_{YY}^{-r}\mathbf{V}_{XY}$ are both Hilbert–Schmidt operators, which are a little bit stronger conditions than the ones proposed in [11]. But these conditions are mild and motivated from the discussions made in the community of learning theory. In the theoretical analysis of learning algorithms generated by regularization schemes, where approximation assumption $f_\rho \in L_K^\theta(L^2_{\rho_X})$ is often considered (see [7,18,19] and the references therein). The index θ ($\theta > 0$) characterizes the decay of the approximation error. The other conditions are analogous. Here r, q, t play the same role as θ. In order to derive convergence rates, AA are imposed on $\mathbf{V}_{YX}$, $\mathbf{V}_{YZ}$ and $\mathbf{V}_{ZX}$. In the sequel, one would see that these conditions are mild and can be constructed by means of mean square contingency.

Before proceeding to the details of AA, let us state the convergence rates of empirical NCCCO $\widehat{\mathbf{V}}_{YX|Z}^{(m)}$ to $\mathbf{V}_{YX|Z}$.

Theorem 1. *Assume that the compact operators $\mathbf{V}_{YX}$, $\mathbf{V}_{YZ}$, $\mathbf{V}_{ZX}$ satisfy AA. Take $\varepsilon_m = m^{-\theta}$ with $0 < \theta < \frac{1}{3}$. For any $0 < \delta < 1$, with confidence at least $1-\delta$, we have*

$$\big\|\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{YX|Z}\big\|_{\mathrm{HS}} \le \widetilde{C}m^{-\vartheta}\log(54/\delta),$$

where $\vartheta = \min\{\frac{1}{2} - \frac{3}{2}\theta, \theta, r\theta, q\theta, t\theta\}$, $\widetilde{C}$ is some constant independent of m or δ and will be presented explicitly in the next section.

Remark 3. Here we extend the conclusion of [11] and address the convergence rates under decay conditions on $\mathbf{V}_{YX}$, $\mathbf{V}_{YZ}$ and $\mathbf{V}_{ZX}$.

In fact, the conditions imposed on $\mathbf{V}_{YX}$, $\mathbf{V}_{YZ}$ and $\mathbf{V}_{ZX}$ are mild. One can construct Hilbert–Schmidt operators $\mathbf{V}_{YX}$, $\mathbf{V}_{YZ}$, $\mathbf{V}_{XZ}$ by means of mean square contingency which is closely related with mutual information [10]. Taking $\mathbf{V}_{YX}$ as an example, assume $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ admit measures $\mu_\mathcal{X}$ and $\mu_\mathcal{Y}$, respectively. $\rho(x, y)$ is absolutely continuous w.r.t $\mu_\mathcal{X} \times \mu_\mathcal{Y}$ with a probability density function $p_{XY}(x, y)$. Let $p_X(x)$, $p_Y(y)$ be the probability density functions of the marginal distributions ρ_X, ρ_Y, respectively. Let $\varpi(x, y) = \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - 1$ be a function defined on $\mathcal{X} \times \mathcal{Y}$. The mean square contingency $C(X, Y)$ is defined by $C(X, Y) = \{\int\int \varpi^2(x, y)d\rho_X(x)d\rho_Y(y)\}^{1/2}$. We will see that if $C(X, Y)$ is finite, then $\mathbf{V}_{YX}$ is Hilbert–Schmidt. Although the detailed proof was given in [10], we provide a sketched description for the reader's convenience. Let $\{\phi_i\}_{i=1}^{\infty}$ and $\{\psi_i\}_{i=1}^{\infty}$ be the complete orthonormal systems of $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$

respectively, such that $\langle \phi_j, \mathbf{\Sigma}_{XX}\phi_i\rangle = \lambda_i\delta_{ij}$, $\langle \psi_j, \mathbf{\Sigma}_{YY}\psi_i\rangle = \mu_i\delta_{ij}$, λ_i, μ_i are nonnegative eigenvalues of $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{YY}$, respectively. δ_{ij} is Kronecker's delta, therefore

$$\begin{aligned}
\|\mathbf{V}_{YX}\|^2_{\mathrm{HS}} &= \sum_{i,j=1}^{\infty} \langle \psi_j, \mathbf{\Sigma}_{YY}^{-1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1/2}\phi_i\rangle^2_{\mathcal{H}_\mathcal{Y}} \\
&= \sum_{i,j=1}^{\infty} \langle \frac{\psi_j}{\sqrt{\mu_j}}, \mathbf{\Sigma}_{YX}\frac{\phi_i}{\sqrt{\lambda_i}}\rangle^2_{\mathcal{H}_\mathcal{Y}} \\
&= \sum_{i,j=1}^{\infty} \Big\{\mathbb{E}_{XY}[\tilde{\phi}_i(X)\tilde{\psi}_j(Y)]\Big\}^2 \\
&= \sum_{i,j=1}^{\infty} \Big\{\int\int \tilde{\phi}_i(X)\tilde{\psi}_j(Y)\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}d\rho_X d\rho_Y\Big\}^2 \\
&\le \|\varpi + 1\|^2_{L^2(\rho_X\times\rho_Y)},
\end{aligned}$$

where $\tilde{\phi}_i = (\phi_i - \mathbb{E}_X[\phi_i(X)])/\sqrt{\lambda_i}$, $\tilde{\psi}_i = (\psi_i - \mathbb{E}_Y[\psi_i(Y)])/\sqrt{\mu_i}$. Simple calculation gives that

$$C^2(x,y) = \int\int \Big(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - 1\Big)^2 d\rho_X(x)d\rho_Y(y) = \int\int \frac{p^2_{XY}(x,y)}{p_X(x)p_Y(y)}d\mu_\mathcal{X} d\mu_\mathcal{Y} - 1 = \mathbb{E}_{XY}[\varpi(x,y)].$$

Hence $\mathbf{V}_{YX}$ is a Hilbert–Schmidt operator under the finiteness of mean square contingency $C(x,y)$. Similar argument can be elucidated for $\mathbf{V}_{YZ}$ and $\mathbf{V}_{ZX}$. Now we are in a position to state the result concerning the consistency of conditional kernel CCA. Recall $d^2(S_1,S_2) = l - \sum_{i=1}^{l}\sum_{j=1}^{l}\langle \xi_i, \widehat{\xi}_j\rangle^2$, then we can see

Theorem 2. *Assume that the compact operators $\mathbf{V}_{YX}$, $\mathbf{V}_{YZ}$, $\mathbf{V}_{ZX}$ satisfy AA. Take $\varepsilon_m = m^{-\theta}$ with $0 < \theta < \frac{1}{3}$. For any $0 < \delta < 1$, with confidence at least $1-\delta$, we have*

$$d(S_1,S_2) \le \widetilde{C}' m^{-\frac{\vartheta}{2}}\log(54/\delta),$$

provided that m satisfies

$$m > \Big(\max\Big\{\frac{16\widetilde{C}}{\tau}, \frac{8\widetilde{C}}{\sigma_1^2}\Big\}\log(54/\delta)\Big)^{1/\vartheta}.$$

Here $\widetilde{C}'$ is a constant depends on r, σ_1 and independent of m or δ. ϑ is defined the same as that in Theorem 1, τ is a lower bound between σ_1 and σ_{l+1} (the second largest singular value of $\mathbf{V}_{YX|Z}$).

Remark 4. Here we propose a new measure to describe the consistency for multidimensional conditional kernel CCA problem. When $m \to \infty$, $d(S_1,S_2) \to 0$. In fact, our consistency measure is the generalization of the one described in [10]. That is, $\big\|\mathbf{\Sigma}_{XX}^{1/2}(\widehat{f} - f)\big\|_{\mathcal{H}_\mathcal{X}}$. Note that

$$\begin{aligned}
\|\mathbf{\Sigma}_{XX}^{1/2}(\widehat{f} - f)\|_{\mathcal{H}_\mathcal{X}} &\le \|\mathbf{\Sigma}_{XX}^{1/2}\big\{(\widehat{\mathbf{\Sigma}}^{(m)}_{XX} + \varepsilon_m I)^{-1/2} - (\mathbf{\Sigma}_{XX} + \varepsilon_m I)^{-1/2}\big\}\widehat{\xi}\|_{\mathcal{H}_\mathcal{X}} \\
&\quad + \|\mathbf{\Sigma}_{XX}^{1/2}(\mathbf{\Sigma}_{XX} + \varepsilon_m I)^{-1/2}(\widehat{\xi} - \xi)\|_{\mathcal{H}_\mathcal{X}} + \|\mathbf{\Sigma}_{XX}^{1/2}(\mathbf{\Sigma}_{XX} + \varepsilon_m I)^{-1/2}\xi - \xi\|_{\mathcal{H}_\mathcal{X}} \\
&\le \|\mathbf{\Sigma}_{XX}^{1/2}\big\{(\widehat{\mathbf{\Sigma}}^{(m)}_{XX} + \varepsilon_m I)^{-1/2} - (\mathbf{\Sigma}_{XX} + \varepsilon_m I)^{-1/2}\big\}\widehat{\xi}\|_{\mathcal{H}_\mathcal{X}} + \|\widehat{\xi} - \xi\|_{\mathcal{H}_\mathcal{X}} \\
&\quad + \|\mathbf{\Sigma}_{XX}^{1/2}(\mathbf{\Sigma}_{XX} + \varepsilon_m I)^{-1/2}\xi - \xi\|_{\mathcal{H}_\mathcal{X}} \\
&:= I + II + III.
\end{aligned}$$

Terms I and III are determined by the properties of $\mathbf{\Sigma}_{XX}$ and $\widehat{\mathbf{\Sigma}}^{(m)}_{XX}$, the bounds of them can be achieved easily [4]. Statistical consistency of conditional kernel CCA are mainly determined by Term II. When $\langle \xi, \widehat{\xi}\rangle > 0$, $\|\xi - \widehat{\xi}\|^2 = 2 - 2\langle \xi, \widehat{\xi}\rangle$. If $l = 1$, our consistency measure corresponds to the one-dimensional learning: $d(S_1, S_2)$ takes form $d(S_1, S_2) = \sqrt{1 - \langle \xi, \widehat{\xi}\rangle^2}$. Therefore our analysis is a generalization of $\|\mathbf{\Sigma}^{1/2}_{XX}(\widehat{f} - f)\|_{\mathcal{H}_\mathcal{X}}$. In the sequel, the relationship between $d(S_1, S_2)$ and $\left\|\widehat{\mathbf{V}}^{(m)}_{YX|Z} - \mathbf{V}_{YX|Z}\right\|_{\mathrm{HS}}$ will be addressed. Hence we conclude that $\left\|\widehat{\mathbf{V}}^{(m)}_{YX|Z} - \mathbf{V}_{YX|Z}\right\|$ can be regarded as a surrogate for testing the statistical learning ability of conditional kernel CCA. Furthermore, $\left\|\widehat{\mathbf{V}}^{(m)}_{YX} - \mathbf{V}_{YX}\right\|$ can be used for measuring the convergence rates of kernel CCA when Z is null.

3. Extension to multiple setting

In the last section, we confined the consistency analysis to the spaces spanned by the eigenfunctions corresponding to the largest singular values. In dimension reduction or information retrieval related problems, one often consider the k-largest singular values for high-dimensional data processing problems [6]. Multiple CCA was widely considered in the literature [5,13,20]. Employing the eigenspaces spanned by the eigenfunctions corresponding to the largest singular values only, is not enough for most practical problems [5], especially in the coming of big data era. In this section, we will address an algorithm for multiple conditional kernel CCA, which extends the results of the last section. Recall conditional kernel CCA problem for the population case can be formulated as

$$\max_{f\in\mathcal{H}_\mathcal{X}, g\in\mathcal{H}_\mathcal{Y}} \frac{\langle g, \mathbf{\Sigma}_{YX|Z} f\rangle}{\sqrt{\langle f, \mathbf{\Sigma}_{XX} f\rangle}\sqrt{\langle g, \mathbf{\Sigma}_{XX} g\rangle}} = \max_{\|\xi\|_{\mathcal{H}_\mathcal{X}}=1, \|\eta\|_{\mathcal{H}_\mathcal{Y}}=1} \langle \eta, V_{YX|Z}\xi\rangle.$$

Applying the ideas for multiple CCA [13], multiple version of conditional kernel CCA can be normally formulated as

$$\begin{aligned}(\eta_k, \xi_k) = &\operatorname*{argmax}_{\|\xi\|_{\mathcal{H}_\mathcal{X}}=1, \|\eta\|_{\mathcal{H}_\mathcal{Y}}=1} \langle \eta, V_{YX|Z}\xi\rangle\\ s.t. \quad &\xi_k \perp \{\xi_1, \cdots, \xi_{k-1}\}\\ &\eta_k \perp \{\eta_1, \cdots, \eta_{k-1}\}.\end{aligned}$$

Accordingly, we can approximate multiple conditional kernel CCA via ERM scheme with modified Tikhonov regularization and come to

$$\begin{aligned}(g_k, f_k) = &\operatorname*{argmax}_{f\in\mathcal{H}_\mathcal{X}, g\in\mathcal{H}_\mathcal{Y}} \langle g, \widehat{\mathbf{\Sigma}}^{(m)}_{YX|Z} f\rangle\\ s.t. \quad &\langle f, (\widehat{\mathbf{\Sigma}}^{(m)}_{XX} + \varepsilon_m I) f\rangle = 1, \langle g, (\widehat{\mathbf{\Sigma}}^{(m)}_{YY} + \varepsilon_m I) g\rangle = 1\\ &(\widehat{\mathbf{\Sigma}}^{(m)}_{XX} + \varepsilon_m I)^{1/2} f \perp \{(\widehat{\mathbf{\Sigma}}^{(m)}_{XX} + \varepsilon_m I)^{1/2} f_1, \cdots, (\widehat{\mathbf{\Sigma}}^{(m)}_{XX} + \varepsilon_m I)^{1/2} f_{k-1}\}\\ &(\widehat{\mathbf{\Sigma}}^{(m)}_{YY} + \varepsilon_m I)^{1/2} g \perp \{(\widehat{\mathbf{\Sigma}}^{(m)}_{YY} + \varepsilon_m I)^{1/2} g_1, \cdots, (\widehat{\mathbf{\Sigma}}^{(m)}_{YY} + \varepsilon_m I)^{1/2} g_{k-1}\}.\end{aligned} \tag{3.1}$$

Here $k = 1, \cdots, d$, $d \le \bar{r}$ ($\bar{r}$ are the numbers of nonzero singular values of $\widehat{\mathbf{V}}^{(m)}_{YX|Z}$). For $F = (f_1, \cdots, f_d)$, if we define $\widehat{\mathbf{\Sigma}}^{(m)}_{YX|Z} F = (\widehat{\mathbf{\Sigma}}^{(m)}_{YX|Z} f_1, \cdots, \widehat{\mathbf{\Sigma}}^{(m)}_{YX|Z} f_d)$. The above problem can be reformulated in terms of the trace operator. Detailed proof will be postponed to Section 4.

Theorem 3. *Let (g_k, f_k) $(k = 1, \cdots, d)$ be a solution of the k-th problem (3.1) with $1 \le d \le \bar{r}$, then (G, F) is the solution of*

$$\begin{aligned}&\max_{\substack{F=(f_1,\cdots,f_d)\\ G=(g_1,\cdots,g_d)}} \operatorname{Trace}(G^T\widehat{\mathbf{\Sigma}}^{(m)}_{YX|Z}F)\\ &s.t. \qquad F^T(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)F=I,\\ &\qquad\quad G^T(\widehat{\mathbf{\Sigma}}^{(m)}_{YY}+\varepsilon_m I)G=I. \end{aligned} \tag{3.2}$$

On the other hand, If (G,F) is a solution of problem (3.2), and (g_k,f_k) $(k=1,\cdots,d)$ is the solution of the k-th problem (3.1), then there exist orthogonal matrices Q_3, Q_4 such that

$$(g_1,\cdots,g_d)=GQ_3, \qquad (f_1,\cdots,f_d)=FQ_4.$$

When $d=1$, (3.2) reduces to (2.5). Theorem 3 reveals that solutions of problem (3.1) and that of problem (3.2) are equivalent subject to orthogonal matrices. When Z is null, the conclusion here is similar to the one given as Theorem 3.1 in [5] for CCA problem. We will analyze the consistency of the above algorithm. Assume $\widehat{\sigma}_1\geq\widehat{\sigma}_2\geq\cdots\widehat{\sigma}_{\bar{r}}>0$ $(1\leq\bar{r}<m)$ are the nonzero singular values of $\widehat{\mathbf{V}}^{(m)}_{YX|Z}$, $(\widehat{\eta}_{i,j},\widehat{\xi}_{i,j})$ $(i=1,\cdots,\tilde{r},j=1,\cdots,\gamma_i)$ are the unit eigenfunction pairs of $\widehat{\sigma}_{\sum_{t=1}^{i-1}\gamma_t+j}$ (denote $\sum_{t=1}^{0}\gamma_t=0$), $\bar{r}=\sum_{i=1}^{\tilde{r}}\gamma_i$. Accordingly, $\sigma_1\geq\sigma_2\cdots$ are the singular values of $\mathbf{V}_{YX|Z}$ with eigenfunction pairs $(\eta_{i,j},\xi_{i,j})$ $(i=1,\cdots,j=1\cdots,l_i)$. That is,

$$\begin{aligned}&\mathbf{V}_{YX|Z}\xi_{i,j}=\sigma_{\sum_{t=1}^{i-1}l_t+j}\eta_{i,j}\ (i=1,\cdots,j=1\cdots,l_i),\\ &\widehat{\mathbf{V}}^{(m)}_{YX|Z}\widehat{\xi}_{i,j}=\widehat{\sigma}_{\sum_{t=1}^{i-1}\gamma_t+j}\widehat{\eta}_{i,j}\ (i=1,\cdots,\bar{r},j=1\cdots,\gamma_i).\end{aligned}$$

Similarly, denote $\sum_{t=1}^{0}l_t=0$. We only discuss the subspaces spanned by $\xi_{i,j}$ $(i=1,\cdots,j=1,\cdots,l_i)$ and $\widehat{\xi}_{i,j}$ $(i=1,\cdots,\tilde{r},j=1,\cdots,\gamma_i)$, respectively. Analogous argument can be addressed for the ones spanned by $\eta_{i,j}$ and $\widehat{\eta}_{i,j}$. Employing the ideas for consistency argument in the last section, denote $l=\sum_{i=1}^{\tilde{r}}l_i$, let $S_1'=\operatorname{span}\{\{\xi_{i,j}\}_{j=1}^{l_i}\}_{i=1}^{\tilde{r}}$, $S_2'=\operatorname{span}\{\{\widehat{\xi}_{i,j}\}_{j=1}^{l_i}\}_{i=1}^{\tilde{r}}$, according to the definition for distance of feature subspaces, we get

$$\begin{aligned}d^2(S_1',S_2')&=l-\sum_{i=1}^{\tilde{r}}\sum_{j=1}^{\tilde{r}}\sum_{s,t=1}^{l_i}\langle\xi_{i,s},\widehat{\xi}_{j,t}\rangle^2\\ &\leq l_1-\sum_{s=1}^{l_1}\sum_{t=1}^{l_1}\langle\xi_{1,s},\widehat{\xi}_{1,t}\rangle^2+l_2-\sum_{s=1}^{l_2}\sum_{t=1}^{l_2}\langle\xi_{2,s},\widehat{\xi}_{2,t}\rangle^2+\cdots+l_{\tilde{r}}-\sum_{s=1}^{l_{\tilde{r}}}\sum_{t=1}^{l_{\tilde{r}}}\langle\xi_{\tilde{r},s},\widehat{\xi}_{\tilde{r},t}\rangle^2.\end{aligned}$$

The above expression implies that we only need to consider the distance of subspaces between S_1 and S_2, which are spanned by $\{\xi_{1,1},\cdots,\xi_{1,l_1}\}$ and $\{\widehat{\xi}_{1,1},\cdots,\widehat{\xi}_{1,l_1}\}$, respectively. The other terms can be investigated analogously. Then the consistency problem of multiple conditional kernel CCA reduces to that of "single" conditional kernel CCA.

4. Proof of main results

In this section, main results Theorems 1, 2 and 3 will be proved. We firstly give the proof of Theorem 1.

4.1. Proof of Theorem 1

Recall that

$$\mathbf{V}_{YX|Z}=\mathbf{V}_{YX}-\mathbf{V}_{YZ}\mathbf{V}_{ZX}, \quad \widehat{\mathbf{V}}^{(m)}_{YX|Z}=\widehat{\mathbf{V}}^{(m)}_{YX}-\widehat{\mathbf{V}}^{(m)}_{YZ}\widehat{\mathbf{V}}^{(m)}_{ZX}.$$

Hence

$$\begin{aligned}\big\|\widehat{\mathbf{V}}^{(m)}_{YX|Z}-\mathbf{V}_{YX|Z}\big\|_{\mathrm{HS}} &= \big\|\widehat{\mathbf{V}}^{(m)}_{YX}-\mathbf{V}_{YX}-(\widehat{\mathbf{V}}^{(m)}_{YZ}-\mathbf{V}_{YZ})\widehat{\mathbf{V}}^{(m)}_{ZX}-\mathbf{V}_{YZ}(\widehat{\mathbf{V}}^{(m)}_{ZX}-\mathbf{V}_{ZX})\big\|_{\mathrm{HS}}\\ &\le \big\|\widehat{\mathbf{V}}^{(m)}_{YX}-\mathbf{V}_{YX}\big\|_{\mathrm{HS}}+\big\|\widehat{\mathbf{V}}^{(m)}_{YZ}-\mathbf{V}_{YZ}\big\|_{\mathrm{HS}}\big\|\widehat{\mathbf{V}}^{(m)}_{ZX}\big\|+\big\|\widehat{\mathbf{V}}^{(m)}_{ZX}-\mathbf{V}_{ZX}\big\|_{\mathrm{HS}}\big\|\mathbf{V}_{YZ}\big\|.\end{aligned}$$

Note that $\big\|\mathbf{V}_{YZ}\big\|\le 1$ and $\big\|\widehat{\mathbf{V}}^{(m)}_{ZX}\big\| = \big\|(\widehat{\mathbf{\Sigma}}^{(m)}_{ZZ}+\varepsilon_m I)^{-1/2}\widehat{\mathbf{\Sigma}}^{(m)}_{ZX}(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}\big\|\le 1$ (the proof of this bound will be given as Proposition 2 in Appendix A). This yields that

$$\big\|\widehat{\mathbf{V}}^{(m)}_{YX|Z}-\mathbf{V}_{YX|Z}\big\|_{\mathrm{HS}}\le \big\|\widehat{\mathbf{V}}^{(m)}_{YX}-\mathbf{V}_{YX}\big\|_{\mathrm{HS}}+\big\|\widehat{\mathbf{V}}^{(m)}_{YZ}-\mathbf{V}_{YZ}\big\|_{\mathrm{HS}}+\big\|\widehat{\mathbf{V}}^{(m)}_{ZX}-\mathbf{V}_{ZX}\big\|_{\mathrm{HS}}.$$

We only give the estimation of $\big\|\widehat{\mathbf{V}}^{(m)}_{YX}-\mathbf{V}_{YX}\big\|_{\mathrm{HS}}$, the others are analogous. Since

$$\begin{aligned}\widehat{\mathbf{V}}^{(m)}_{YX}-\mathbf{V}_{YX} &= \Big\{(\widehat{\mathbf{\Sigma}}^{(m)}_{YY}+\varepsilon_m I)^{-1/2}-(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}\Big\}\widehat{\mathbf{\Sigma}}^{(m)}_{YX}(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}\\ &\quad+(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}(\widehat{\mathbf{\Sigma}}^{(m)}_{YX}-\mathbf{\Sigma}_{YX})(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}\\ &\quad+(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}\mathbf{\Sigma}_{YX}\Big\{(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}-(\mathbf{\Sigma}_{XX}+\varepsilon_m I)^{-1/2}\Big\}\\ &\quad+(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}[\mathbf{\Sigma}_{YX}(\mathbf{\Sigma}_{XX}+\varepsilon_m I)^{-1/2}-\mathbf{\Sigma}^{1/2}_{YY}\mathbf{V}_{YX}]\\ &\quad+[(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}\mathbf{\Sigma}^{1/2}_{YY}-I]\mathbf{V}_{YX}.\end{aligned}\tag{4.1}$$

Thus we can see that

Lemma 1. *For any $0<\delta<1$, with confidence at least $1-\delta$, we have*

$$\big\|\{(\widehat{\mathbf{\Sigma}}^{(m)}_{YY}+\varepsilon_m I)^{-1/2}-(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}\}\widehat{\mathbf{\Sigma}}^{(m)}_{YX}(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}\big\|_{\mathrm{HS}}\le\frac{24\kappa_2^2(\kappa_2+1)}{\varepsilon_m^{3/2}m^{1/2}}\log(6/\delta),$$

$$\big\|(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}\mathbf{\Sigma}_{YX}\{(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}-(\mathbf{\Sigma}_{XX}+\varepsilon_m I)^{-1/2}\}\big\|_{\mathrm{HS}}\le\frac{24\kappa_1^2(\kappa_1+1)}{\varepsilon_m^{3/2}m^{1/2}}\log(6/\delta).$$

Proof. Note that

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

then

$$\begin{aligned}&\big\|\{(\widehat{\mathbf{\Sigma}}^{(m)}_{YY}+\varepsilon_m I)^{-1/2}-(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}\}\widehat{\mathbf{\Sigma}}^{(m)}_{YX}(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}\big\|_{\mathrm{HS}}\\ &\quad=\big\|(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}\{(\widehat{\mathbf{\Sigma}}^{(m)}_{YY}+\varepsilon_m I)^{3/2}-(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{3/2}+(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{1/2}(\mathbf{\Sigma}_{YY}-\widehat{\mathbf{\Sigma}}^{(m)}_{YY})\}\\ &\qquad\times(\widehat{\mathbf{\Sigma}}^{(m)}_{YY}+\varepsilon_m I)^{-3/2}\widehat{\mathbf{\Sigma}}^{(m)}_{YX}(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}\big\|_{\mathrm{HS}}.\end{aligned}$$

Recall $\widehat{\mathbf{\Sigma}}^{(m)}_{YY}=\frac{1}{m}\sum_{i=1}^m(k_{\mathcal{Y}}(\cdot,Y_i)-\frac{1}{m}\sum_{t=1}^m k_{\mathcal{Y}}(\cdot,Y_t))\bigotimes(k_{\mathcal{Y}}(\cdot,Y_i)-\frac{1}{m}\sum_{t=1}^m k_{\mathcal{Y}}(\cdot,Y_t))$. Applying Lemma 8 in [10], for any $0<\delta<1$,

$$\begin{aligned}&\big\|\{(\widehat{\mathbf{\Sigma}}^{(m)}_{YY}+\varepsilon_m I)^{-1/2}-(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}\}\widehat{\mathbf{\Sigma}}^{(m)}_{YX}(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}\big\|_{\mathrm{HS}}\\ &\quad\le\frac{4}{\varepsilon_m^{3/2}}\max\big\{\big\|\mathbf{\Sigma}_{YY}+\varepsilon_m I\big\|^{1/2},\big\|\widehat{\mathbf{\Sigma}}^{(m)}_{YY}+\varepsilon_m I\big\|^{1/2}\big\}\big\|\widehat{\mathbf{\Sigma}}^{(m)}_{YY}-\mathbf{\Sigma}_{YY}\big\|_{\mathrm{HS}}\\ &\quad\le\frac{24\kappa_2^2(\kappa_2+1)}{\varepsilon_m^{3/2}m^{1/2}}\log(6/\delta),\end{aligned}$$

holds true with confidence at least $1-\delta$. The last inequality follows by applying Lemma 5 in the appendix for the special case $Y = X$. Similarly, for any $0 < \delta < 1$,

$$\begin{aligned}
&\big\|(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}\mathbf{\Sigma}_{YX}\{(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}-(\mathbf{\Sigma}_{XX}+\varepsilon_m I)^{-1/2}\}\big\|_{\mathrm{HS}}\\
&\quad=\big\|(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}\mathbf{\Sigma}^{1/2}_{YY}\mathbf{V}^{1/2}_{YX}\mathbf{\Sigma}^{1/2}_{XX}\{(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}-(\mathbf{\Sigma}_{XX}+\varepsilon_m I)^{-1/2}\}\big\|_{\mathrm{HS}}\\
&\quad\le\big\|\mathbf{\Sigma}^{1/2}_{XX}\{(\widehat{\mathbf{\Sigma}}^{(m)}_{XX}+\varepsilon_m I)^{-1/2}-(\mathbf{\Sigma}_{XX}+\varepsilon_m I)^{-1/2}\}\big\|_{\mathrm{HS}}\\
&\quad\le\frac{24\kappa_1^2(\kappa_1+1)}{\varepsilon_m^{3/2}m^{1/2}}\log(6/\delta),
\end{aligned}$$

holds with confidence at least $1-\delta$. □

Next we give the estimation of the last two terms in Eq. (4.1).

Lemma 2. *Assume that the compact operator $\mathbf{V}_{YX}$ satisfy AA, then*

$$\begin{aligned}
\big\|\mathbf{V}_{YX}[\mathbf{\Sigma}^{1/2}_{XX}(\mathbf{\Sigma}_{XX}+\varepsilon_m I)^{-1/2}-I]\big\|_{\mathrm{HS}}&\le C_1\varepsilon_m^{\min\{r,1\}},\\
\big\|[(\mathbf{\Sigma}_{YY}+\varepsilon_m I)^{-1/2}\mathbf{\Sigma}^{1/2}_{YY}-I]\mathbf{V}_{YX}\big\|_{\mathrm{HS}}&\le C_1\varepsilon_m^{\min\{r,1\}},
\end{aligned}$$

where C_1 is some constant independent of m.

Proof. According to AA, there exist operators $W_1, W_2 \in HS(\mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}})$ such that $\mathbf{V}_{YX} = W_1\mathbf{\Sigma}^r_{XX}$ and $\mathbf{V}_{XY} = \mathbf{\Sigma}^r_{YY}W_2$. We only prove the first inequality, the second is similar.

$$\big\|\mathbf{V}_{YX}[\mathbf{\Sigma}^{1/2}_{XX}(\mathbf{\Sigma}_{XX}+\varepsilon_m I)^{-1/2}-I]\big\|^2_{\mathrm{HS}}=\sum_{s=1}^{\infty}\|W_1\mathbf{\Sigma}^r_{XX}(\mathbf{\Sigma}^{1/2}_{XX}(\mathbf{\Sigma}_{XX}+\varepsilon_m I)^{-1/2}-I)\phi_s\|^2_{\mathcal{H}_{\mathcal{Y}}}.$$

Recall λ_s, ϕ_s are the eigenpairs of $\mathbf{\Sigma}_{XX}$. Spectrum theorem of compact operators yields that

$$W_1\mathbf{\Sigma}^r_{XX}(\mathbf{\Sigma}^{1/2}_{XX}(\mathbf{\Sigma}_{XX}+\varepsilon_m I)^{-1/2}-I)\phi_s=\frac{\lambda_s^r(\lambda_s^{1/2}-(\lambda_s+\varepsilon_m)^{1/2})}{(\lambda_s+\varepsilon_m)^{1/2}}W_1\phi_s.$$

Therefore

$$\big\|\mathbf{V}_{YX}[\mathbf{\Sigma}^{1/2}_{XX}(\mathbf{\Sigma}_{XX}+\varepsilon_m I)^{-1/2}-I]\big\|^2_{\mathrm{HS}}\le\sum_{s=1}^{\infty}\frac{\lambda_s^{2r}\varepsilon_m^2}{(2\lambda_s+\varepsilon_m)^2}\|W_1\phi_s\|^2_{\mathcal{H}_{\mathcal{Y}}}.$$

Simple calculation shows that when $0 < r < 1$,

$$\frac{\lambda_s^{2r}\varepsilon_m^2}{(2\lambda_s+\varepsilon_m)^2}\le\frac{1}{4}\varepsilon_m^{2r}(\frac{r}{2})^{2r-2}(1-r)^{2-2r},$$

and if $r \ge 1$, then

$$\frac{\lambda_s^{2r}\varepsilon_m^2}{(2\lambda_s+\varepsilon_m)^2}\le\frac{1}{4}\kappa_1^{4r-4}\varepsilon_m^2.$$

Combining the above estimations, we prove the result. □

Proof of Theorem 1. Employing Lemmas 1, 2 and 5, for any $0 < \delta < 1$, with confidence at least $1 - \delta$, we see

$$\begin{aligned}\big\|\widehat{\mathbf{V}}_{YX}^{(m)} - \mathbf{V}_{YX}\big\|_{\mathrm{HS}} \le{}& \frac{24\kappa_2^2(\kappa_2+1)}{\varepsilon_m^{3/2}m^{1/2}}\log(18/\delta) + \frac{24\kappa_1^2(\kappa_1+1)}{\varepsilon_m^{3/2}m^{1/2}}\log(18/\delta) + \frac{6\kappa_1\kappa_2}{\varepsilon_m^{3/2}m^{1/2}}\log(18/\delta)\\ &+ 2C_1\varepsilon_m^{\min\{r,1\}}.\end{aligned}$$

This yields that

$$\big\|\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{YX|Z}\big\|_{\mathrm{HS}} \le \widetilde{C}m^{-\vartheta}\log(54/\delta),$$

holds true with confidence at least $1 - \delta$ by taking $\varepsilon_m = m^{-\theta}$. □

4.2. Proof of Theorem 2

Now we are in the position to give the proof of Theorem 2.
Denote $\widehat{\mathbf{A}} = \widehat{\mathbf{V}}_{XY|Z}^{(m)}\widehat{\mathbf{V}}_{YX|Z}^{(m)}$, $\mathbf{A} = \mathbf{V}_{XY|Z}\mathbf{V}_{YX|Z}$, we have

$$\begin{aligned}\big\|\widehat{\mathbf{A}} - \mathbf{A}\big\|_{\mathrm{HS}} &= \big\|\widehat{\mathbf{V}}_{XY|Z}^{(m)}(\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{YX|Z}) + (\widehat{\mathbf{V}}_{XY|Z}^{(m)} - \mathbf{V}_{XY|Z})\mathbf{V}_{YX|Z}\big\|_{\mathrm{HS}}\\ &\le \big\|\widehat{\mathbf{V}}_{XY|Z}^{(m)}\big\| \cdot \big\|\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{YX|Z}\big\|_{\mathrm{HS}} + \big\|\widehat{\mathbf{V}}_{XY|Z}^{(m)} - \mathbf{V}_{XY|Z}\big\|_{\mathrm{HS}} \cdot \big\|\mathbf{V}_{YX|Z}\big\|.\end{aligned}$$

From the definitions of $\widehat{\mathbf{V}}_{XY|Z}^{(m)}$ and $\mathbf{V}_{XY|Z}$, it is obvious that $\|\widehat{\mathbf{V}}_{XY|Z}^{(m)}\| \le 2$, $\|\mathbf{V}_{XY|Z}\| \le 2$. Therefore a rigorous bound shows

$$\big\|\widehat{\mathbf{V}}_{XY|Z}^{(m)}\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{XY|Z}\mathbf{V}_{YX|Z}\big\|_{\mathrm{HS}} \le 2\big\|\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{YX|Z}\big\|_{\mathrm{HS}} + 2\big\|\widehat{\mathbf{V}}_{XY|Z}^{(m)} - \mathbf{V}_{XY|Z}\big\|_{\mathrm{HS}}.$$

Moreover, note that

$$\begin{aligned}\big\|\widehat{\mathbf{V}}_{XY|Z}^{(m)} - \mathbf{V}_{XY|Z}\big\|_{\mathrm{HS}} &= \big\|(\widehat{\mathbf{V}}_{XY|Z}^{(m)} - \mathbf{V}_{XY|Z})^*\big\|_{\mathrm{HS}}\\ &= \big\|(\widehat{\mathbf{V}}_{XY}^{(m)} - \widehat{\mathbf{V}}_{XZ}^{(m)}\widehat{\mathbf{V}}_{ZY}^{(m)} - \mathbf{V}_{XY} + \mathbf{V}_{XZ}\mathbf{V}_{ZY})^*\big\|_{\mathrm{HS}}\\ &= \big\|\widehat{\mathbf{V}}_{YX}^{(m)} - \widehat{\mathbf{V}}_{YZ}^{(m)}\widehat{\mathbf{V}}_{ZX}^{(m)} - \mathbf{V}_{YX} + \mathbf{V}_{YZ}\mathbf{V}_{ZX}\big\|_{\mathrm{HS}}\\ &= \big\|\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{YX|Z}\big\|_{\mathrm{HS}}.\end{aligned}$$

Hence

$$\big\|\widehat{\mathbf{A}} - \mathbf{A}\big\|_{\mathrm{HS}} \le 4\big\|\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{YX|Z}\big\|_{\mathrm{HS}}.$$

Firstly, we describe the relationship between σ_i, $\widehat{\sigma}_i$ $(1 \le i \le \bar{r})$ and $\mathbf{A}$, $\widehat{\mathbf{A}}$. From now on, we rearrange $\{\xi_{1,1}, \cdots, \xi_{1,l_1}, \xi_{2,1}, \cdots, \xi_{2,l_2}, \cdots\}$ as $\{\xi_1, \cdots, \xi_{l_1}, \xi_{l_1+1}, \cdots, \xi_{l_1+l_2}, \cdots,\}$, so that two indexes are changed to one for simplicity. Analogously, $\{\widehat{\xi}_{1,1}, \cdots, \widehat{\xi}_{1,\gamma_1}, \widehat{\xi}_{2,1}, \cdots, \widehat{\xi}_{2,\gamma_2}, \cdots, \widehat{\xi}_{\bar{r},\gamma_{\bar{r}}}\}$ are rearranged as $\{\widehat{\xi}_1, \cdots, \widehat{\xi}_{\gamma_1}, \widehat{\xi}_{\gamma_1+1}, \cdots, \widehat{\xi}_{\gamma_1+\gamma_2}, \cdots, \widehat{\xi}_{\bar{r}}\}$. For any $1 \le k \le \bar{r}$, denote $H_k = \mathrm{span}\{\xi_1, \cdots, \xi_k\}$, and $\widehat{H}_k = \mathrm{span}\{\widehat{\xi}_1, \cdots, \widehat{\xi}_k\}$, then we see that

Lemma 3. *For any $1 \le k \le \bar{r}$, there holds $|\sigma_k^2 - \hat{\sigma}_k^2| \le \|\mathbf{A} - \widehat{\mathbf{A}}\|$.*

Proof. Since $\dim \widehat{H}_k = k > \dim H_{k-1} = k-1$, there exists $\widehat{\xi}^* \in \widehat{H}_k$, s.t., $\widehat{\xi}^* \perp H_{k-1}$ and $\|\widehat{\xi}^*\| = 1$. Then,

$$\|\mathbf{A} - \widehat{\mathbf{A}}\| \ge \|(\mathbf{A} - \widehat{\mathbf{A}})\widehat{\xi}^*\| \ge \|\widehat{\mathbf{A}}\widehat{\xi}^*\| - \|\mathbf{A}\widehat{\xi}^*\| \ge \hat{\sigma}_k^2 - \sigma_k^2.$$

Similarly, we can prove that $\|\mathbf{A} - \widehat{\mathbf{A}}\| \ge \sigma_k^2 - \hat{\sigma}_k^2$. The conclusion therefore holds. □

Assume that $\sigma_{l_1}^2 - \sigma_{l_1+1}^2 = \tau > 0$, then for any $1 \le i \le l_1$,

$$\widehat{\sigma}_i^2 - \widehat{\sigma}_{l_1+1}^2 = \widehat{\sigma}_i^2 - \sigma_i^2 + \sigma_i^2 + \sigma_{l_1}^2 - \sigma_{l_1+1}^2 + \sigma_{l_1+1}^2 - \widehat{\sigma}_{l_1+1}^2 \ge \tau - 2\|\mathbf{A} - \widehat{\mathbf{A}}\|.$$

Lemma 4. *Suppose that $\mathbf{P}_{H_{l_1}}$ and $\mathbf{P}_{\widehat{H}_{l_1}}$ are the projection operators onto subspaces H_{l_1} and $\widehat{H}_{l_1}$ respectively, and $\|\mathbf{A} - \widehat{\mathbf{A}}\| \le \min\{\frac{\sigma_1^2}{2}, \frac{\tau}{4}\}$. Then for any $0 < \delta < 1$, with confidence at least $1 - \delta$, we have*

$$\|(I - \mathbf{P}_{\widehat{H}_{l_1}})\xi_i\|^2 \le \frac{8\|\mathbf{A} - \widehat{\mathbf{A}}\|}{\tau}, \qquad \forall 1 \le i \le l_1;$$
$$\|(I - \mathbf{P}_{H_{l_1}})\widehat{\xi}_i\|^2 \le \frac{8\|\mathbf{A} - \widehat{\mathbf{A}}\|}{\tau}, \qquad \forall 1 \le i \le l_1,$$

provided that $m > \left(\max\left\{\frac{16\tilde{C}}{\tau}, \frac{8\tilde{C}}{\sigma_1^2}\right\}\log(54/\delta)\right)^{1/\vartheta}$.

Proof. We only prove the first inequality. For any $1 \le i \le l_1$,

$$\begin{aligned}\sigma_1^2 = \|\mathbf{A}\xi_i\| &\le \|(\mathbf{A} - \widehat{\mathbf{A}})\xi_i\| + \|\widehat{\mathbf{A}}\xi_i\| \\ &\le \|\mathbf{A} - \widehat{\mathbf{A}}\| + \sqrt{\|\widehat{\mathbf{A}}\mathbf{P}_{\widehat{H}_{l_1}}\xi_i\|^2 + \|\widehat{\mathbf{A}}(I - \mathbf{P}_{\widehat{H}_{l_1}})\xi_i\|^2} \\ &\le \|\mathbf{A} - \widehat{\mathbf{A}}\| + \sqrt{\widehat{\sigma}_1^4\|\mathbf{P}_{\widehat{H}_{l_1}}\xi_i\|^2 + \widehat{\sigma}_{l+1}^4\|(I - \mathbf{P}_{\widehat{H}_{l_1}})\xi_i\|^2}.\end{aligned}$$

This estimate gives that

$$\left(\sigma_1^2 - \|\mathbf{A} - \widehat{\mathbf{A}}\|\right)^2 \le \widehat{\sigma}_1^4 - (\widehat{\sigma}_1^4 - \widehat{\sigma}_{l+1}^4)\|(I - \mathbf{P}_{\widehat{H}_{l_1}})\xi_i\|^2,$$

and

$$(\widehat{\sigma}_1^4 - \widehat{\sigma}_{l_1+1}^4)\|(I - \mathbf{P}_{\widehat{H}_{l_1}})\xi_i\|^2 \le \widehat{\sigma}_1^4 - \left(\sigma_1^2 - \|\mathbf{A} - \widehat{\mathbf{A}}\|\right)^2,$$

this yields that

$$\begin{aligned}\|(I - \mathbf{P}_{\widehat{H}_{l_1}})\xi_i\|^2 &\le \frac{(\widehat{\sigma}_1^2 + \sigma_1^2 - \|\mathbf{A} - \widehat{\mathbf{A}}\|)(\widehat{\sigma}_1^2 - \sigma_1^2 + \|\mathbf{A} - \widehat{\mathbf{A}}\|)}{(\widehat{\sigma}_1^2 + \widehat{\sigma}_{l_1+1}^2)(\widehat{\sigma}_1^2 - \widehat{\sigma}_{l_1+1}^2)} \\ &\le \frac{2(\widehat{\sigma}_1^2 + \sigma_1^2 - \|\mathbf{A} - \widehat{\mathbf{A}}\|)\|\mathbf{A} - \widehat{\mathbf{A}}\|}{\widehat{\sigma}_1^2(\tau - 2\|\mathbf{A} - \widehat{\mathbf{A}}\|)} \\ &\le \frac{4\|\mathbf{A} - \widehat{\mathbf{A}}\|}{\tau - 2\|\mathbf{A} - \widehat{\mathbf{A}}\|}.\end{aligned}$$

Recall

$$\left\|\widehat{\mathbf{A}}-\mathbf{A}\right\|_{\mathrm{HS}} \leq 4\left\|\widehat{\mathbf{V}}_{YX|Z}^{(m)}-\mathbf{V}_{YX|Z}\right\|_{\mathrm{HS}},$$

and note that $m > \left(\max\left\{\frac{16\tilde{C}}{\tau}, \frac{8\tilde{C}}{\sigma_1^2}\right\}\log(54/\delta)\right)^{1/\vartheta}$, combining with Theorem 1 yields that $\left\|\widehat{\mathbf{A}}-\mathbf{A}\right\| \leq \left\|\widehat{\mathbf{A}}-\mathbf{A}\right\|_{\mathrm{HS}} \leq \frac{\tau}{4}$ and $\left\|\widehat{\mathbf{A}}-\mathbf{A}\right\| \leq \frac{\sigma_1^2}{2}$.

For the second inequality, when $\left\|\widehat{\mathbf{A}}-\mathbf{A}\right\| \leq \frac{\sigma_1^2}{2}$, Lemma 3 results in

$$\left\|\widehat{\mathbf{A}}-\mathbf{A}\right\| \geq \sigma_1^2 - \widehat{\sigma}_1^2.$$

Therefore,

$$\widehat{\sigma}_1^2 \geq \sigma_1^2 - \left\|\widehat{\mathbf{A}}-\mathbf{A}\right\| \geq \frac{\sigma_1^2}{2} \geq \left\|\widehat{\mathbf{A}}-\mathbf{A}\right\|.$$

Also note that

$$\begin{aligned}\widehat{\sigma}_1^2 &= \|\widehat{\mathbf{A}}\widehat{\xi}_i\| \leq \|(\widehat{\mathbf{A}}-\mathbf{A})\widehat{\xi}_i\| + \|\mathbf{A}\widehat{\xi}_i\| \\ &\leq \|\widehat{\mathbf{A}}-\mathbf{A}\| + \sqrt{\|\mathbf{A}\mathbf{P}_{H_{l_1}}\widehat{\xi}_i\|^2 + \|\mathbf{A}(I-\mathbf{P}_{H_{l_1}})\widehat{\xi}_i\|^2}.\end{aligned}$$

Following the same steps as above mentioned for proving the first inequality, we get that

$$\|(I-\mathbf{P}_{H_{l_1}})\widehat{\xi}_i\|^2 \leq \frac{8\|\mathbf{A}-\widehat{\mathbf{A}}\|}{\tau}.$$

This completes the proof. □

Now we are in the position to give the proof of Theorem 2.

Proof of Theorem 2. Theorem 1 gives that when $m > \left(\max\left\{\frac{16\tilde{C}}{\tau}, \frac{8\tilde{C}}{\sigma_1^2}\right\}\log(54/\delta)\right)^{1/\vartheta}$, $\left\|\widehat{\mathbf{A}}-\mathbf{A}\right\|_{\mathrm{HS}} \leq \min\{\frac{\sigma_1^2}{2}, \frac{\tau}{4}\}$. A rigorous estimate shows that

$$\begin{aligned}\|\mathbf{A}-\widehat{\mathbf{A}}\|_{\mathrm{HS}}^2 &= \sum_{i=1}^{\infty}\left\|(\mathbf{A}-\widehat{\mathbf{A}})\xi_i\right\|^2 \\ &= \sum_{i=1}^{\infty}\Big\|\sum_{j=1}^{\infty}(\sigma_i^2-\widehat{\sigma}_j^2)\langle\xi_i, \widehat{\xi}_j\rangle\widehat{\xi}_j\Big\|^2 \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(\sigma_i^2-\widehat{\sigma}_j^2)^2\langle\xi_i, \widehat{\xi}_j\rangle^2 \\ &= \sum_{i=1}^{\infty}\sigma_i^4 + \sum_{j=1}^{\infty}\widehat{\sigma}_j^4 - 2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sigma_i^2\widehat{\sigma}_j^2\langle\xi_i, \widehat{\xi}_j\rangle^2.\end{aligned}$$

It can be decomposed further as

$$\|\mathbf{A}-\widehat{\mathbf{A}}\|_{\mathrm{HS}}^2 = \sum_{i=1}^{\infty}\sigma_i^4 + \sum_{j=1}^{\infty}\widehat{\sigma}_j^4 - 2\sum_{i=1}^{l_1}\sum_{j=1}^{l_1}\sigma_i^2\widehat{\sigma}_j^2\langle\xi_i, \widehat{\xi}_j\rangle^2 - 2\sum_{i=1}^{l_1}\sum_{j=l_1+1}^{\infty}\sigma_i^2\widehat{\sigma}_j^2\langle\xi_i, \widehat{\xi}_j\rangle^2$$

$$
\begin{aligned}
&- 2\sum_{i=l_1+1}^{\infty}\sum_{j=1}^{l_1}\sigma_i^2\widehat{\sigma}_j^2\langle\xi_i,\,\widehat{\xi}_j\rangle^2 - 2\sum_{i=l_1+1}^{\infty}\sum_{j=l_1+1}^{\infty}\sigma_i^2\widehat{\sigma}_j^2\langle\xi_i,\,\widehat{\xi}_j\rangle^2\\
&\geq \sum_{i=1}^{\infty}\sigma_i^4 + \sum_{j=1}^{\infty}\widehat{\sigma}_j^4 - 2\sum_{i=1}^{l_1}\sum_{j=1}^{l_1}\sigma_i^2\widehat{\sigma}_j^2\langle\xi_i,\,\widehat{\xi}_j\rangle^2 - 2\sigma_1^2\widehat{\sigma}_{l_1+1}^2\sum_{i=1}^{l_1}\|(I-\mathbf{P}_{\widehat{H}_{l_1}})\xi_i\|^2\\
&\quad - 2\widehat{\sigma}_1^2\sigma_{l_1+1}^2\sum_{j=1}^{l_1}\|(I-\mathbf{P}_{H_{l_1}})\widehat{\xi}_j\|^2 - \sum_{i=l_1+1}^{\infty}\sum_{j=l_1+1}^{\infty}(\sigma_i^4+\widehat{\sigma}_j^4)\langle\xi_i,\,\widehat{\xi}_j\rangle^2,
\end{aligned}
$$

which, together with Lemma 4 yields that

$$
\begin{aligned}
\|\mathbf{A}-\widehat{\mathbf{A}}\|_{\mathrm{HS}}^2 &\geq l_1\sigma_1^4 + \sum_{j=1}^{l_1}\widehat{\sigma}_j^4 - 2\sum_{i=1}^{l_1}\sum_{j=1}^{l_1}\sigma_i^2\widehat{\sigma}_j^2\langle\xi_i,\,\widehat{\xi}_j\rangle^2 - 2\sigma_1^2\widehat{\sigma}_{l_1+1}^2\frac{8l_1\|\mathbf{A}-\widehat{\mathbf{A}}\|}{\tau} - 2\widehat{\sigma}_1^2\sigma_{l_1+1}^2\frac{8l_1\|\mathbf{A}-\widehat{\mathbf{A}}\|}{\tau}\\
&= 2l_1\sigma_1^4 + \sum_{j=1}^{l_1}(\widehat{\sigma}_j^4-\sigma_j^4) - 2\sigma_1^4\sum_{i=1}^{l_1}\sum_{j=1}^{l_1}\langle\xi_i,\widehat{\xi}_j\rangle^2 - 2\sum_{i=1}^{l_1}\sum_{j=1}^{l_1}\sigma_1^2(\widehat{\sigma}_j^2-\sigma_1^2)\langle\xi_i,\widehat{\xi}_j\rangle^2\\
&\quad - \frac{16l_1(\sigma_1^2\widehat{\sigma}_{l_1+1}^2+\widehat{\sigma}_1^2\sigma_{l_1+1}^2)}{\tau}\|\mathbf{A}-\widehat{\mathbf{A}}\|\\
&\geq 2l_1\sigma_1^4 - 2\sigma_1^4\sum_{i=1}^{l_1}\sum_{j=1}^{l_1}\langle\xi_i,\widehat{\xi}_j\rangle^2 - \Big(\frac{16l_1(\sigma_1^2\widehat{\sigma}_{l_1+1}^2+\widehat{\sigma}_1^2\sigma_{l_1+1}^2)}{\tau} + 4l_1\sigma_1^2\Big)\|\mathbf{A}-\widehat{\mathbf{A}}\|_{\mathrm{HS}},
\end{aligned}
$$

holds true with confidence at least $1-\delta$. Recall $d^2(S_1,S_2) = l_1 - \sum_{i=1}^{l_1}\sum_{j=1}^{l_1}\langle\xi_i,\widehat{\xi}_j\rangle^2$ and note that $\|\widehat{\mathbf{A}}-\mathbf{A}\|_{\mathrm{HS}} \leq 4\|\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{YX|Z}\|_{\mathrm{HS}}$. Hence

$$
2\sigma_1^4 d^2(S_1,S_2) \leq \Big(\frac{64l_1(\sigma_1^2\widehat{\sigma}_{l_1+1}^2+\widehat{\sigma}_1^2\sigma_{l_1+1}^2)}{\tau} + 16l_1\sigma_1^2 + \tau\Big)\|\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{YX|Z}\|_{\mathrm{HS}}.
$$

Combining this with Theorem 1, the proof is completed by replacing l_1 with l. □

4.3. Proof of Theorem 3

This section is devoted to prove Theorem 3. The ideas of proof are inspired from the ones for Theorem 3.1 in [5]. Recall $\widehat{\mathbf{V}}_{YX|Z}^{(m)} = (\widehat{\mathbf{\Sigma}}_{YY}^{(m)}+\varepsilon_m I)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)}+\varepsilon_m I)^{-1/2}$. Obviously, it is a finite rank operator. Hence we have $\widehat{\mathbf{V}}_{YX|Z}^{(m)} = \sum_{i=1}^{\bar{r}}\widehat{\sigma}_i\widehat{\eta}_i\otimes\widehat{\xi}_i$, where $\widehat{\xi}_i$, $\widehat{\eta}_i$ are the unit eigenfunctions corresponding to $\widehat{\sigma}_i$ $(i=1,\cdots,\bar{r})$. Therefore, problem (3.2) can be converted into

$$
\begin{aligned}
&\max_{\widetilde{F},\widetilde{G}}\ \mathrm{Trace}(\widetilde{G}^T\widehat{\mathbf{V}}_{YX|Z}^{(m)}\widetilde{F})\\
&s.t.\quad \widetilde{G}^T\widetilde{G}=I, \widetilde{F}^T\widetilde{F}=I,
\end{aligned}
\tag{4.2}
$$

where $\widetilde{F} = (\widehat{\mathbf{\Sigma}}_{XX}^{(m)}+\varepsilon_m I)^{1/2}F$, $\widetilde{G} = (\widehat{\mathbf{\Sigma}}_{YY}^{(m)}+\varepsilon_m I)^{1/2}G$. Moreover, note that $\mathrm{Trace}(\widetilde{G}^T\widehat{\mathbf{V}}_{YX|Z}^{(m)}\widetilde{F}) = \sum_{i=1}^{d}\tilde{g}_i^T\widehat{\mathbf{V}}_{YX|Z}^{(m)}\tilde{f}_i$, from matrix theory [16], we can see that $\sum_{i=1}^{d}\widehat{\sigma}_i = \max\mathrm{Trace}(\widetilde{G}^T\widehat{\mathbf{V}}_{YX|Z}^{(m)}\widetilde{F})$, where $\widetilde{G}$, $\widetilde{F}$ satisfy the above mentioned constraints, and $\widetilde{G}=(\tilde{g}_1,\cdots,\tilde{g}_d)$, $\widetilde{F}=(\tilde{f}_1,\cdots,\tilde{f}_d)$. Now we give the proof of Theorem 3, and only need to prove the equivalence of (3.1) and (4.2).

Proof.

- If $(\widetilde{G}_1, \widetilde{F}_1)$ is a solution of (4.2), then eigenvalues of $\widetilde{G}_1^T \widehat{\mathbf{V}}_{YX|Z}^{(m)} \widetilde{F}_1$ are positive.
  Let $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \tilde{\sigma}_s \geq \tilde{\sigma}_{s+1} \cdots \geq \tilde{\sigma}_d$ be the eigenvalues of $\widetilde{G}_1^T \widehat{\mathbf{V}}_{YX|Z}^{(m)} \widetilde{F}_1$. Firstly, we prove that they are nonnegative. Otherwise, if $\tilde{\sigma}_s < 0$, let $\widetilde{G}_1^T \widehat{\mathbf{V}}_{YX|Z}^{(m)} \widetilde{F}_1 = Q_3 C Q_4^T$, where Q_3, Q_4 are orthogonal matrices, $C = \begin{pmatrix} \tilde{\sigma}_1 & & \\ & \ddots & \\ & & \tilde{\sigma}_d \end{pmatrix}$. Denote $B = \begin{pmatrix} I_{s-1} & O \\ O & -I_{d-s+1} \end{pmatrix}$, then

$$\mathrm{Trace}[(\widetilde{G}_1 Q_3)^T \widehat{\mathbf{V}}_{YX|Z}^{(m)} \widetilde{F}_1 Q_4 B] = -\tilde{\sigma}_s - \tilde{\sigma}_{s+1} - \cdots - \tilde{\sigma}_d + \sum_{i=1}^{s-1} \tilde{\sigma}_i > \mathrm{Trace}(\widetilde{G}_1^T \widehat{\mathbf{V}}_{YX|Z}^{(m)} \widetilde{F}_1).$$

  Since $(\widetilde{G}_1 Q_3)^T (\widetilde{G}_1 Q_3) = I$, $(\widetilde{F}_1 Q_4 B)^T (\widetilde{F}_1 Q_4 B) = I$. This is a contradiction with the fact that $\mathrm{Trace}(\widetilde{G}_1^T \widehat{\mathbf{V}}_{YX|Z}^{(m)} \widetilde{F}_1)$ is the maximum objective function value of problem (4.2). Now we prove that $\widehat{\sigma}_1, \cdots, \widehat{\sigma}_d$ are the eigenvalues of $\widetilde{G}_1^T \widehat{\mathbf{V}}_{YX|Z}^{(m)} \widetilde{F}_1$. Firstly, since $\sum_{i=1}^d \widehat{\sigma}_i = \max \mathrm{Trace}(\widetilde{G})^T \widehat{\mathbf{V}}_{YX|Z}^{(m)} \widetilde{F} = \mathrm{Trace}(\widetilde{G}_1)^T \widehat{\mathbf{V}}_{YX|Z}^{(m)} \widetilde{F}_1$, and $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_d$ are the eigenvalues of $(\widetilde{G}_1)^T \widehat{\mathbf{V}}_{YX|Z}^{(m)} \widetilde{F}_1$. Then by induction, we have $\tilde{\sigma}_i \leq \widehat{\sigma}_i$, $i = 1, \cdots, d$ (see [16]), and $\mathrm{Trace}(\widetilde{G}_1^T \widehat{\mathbf{V}}_{YX|Z}^{(m)} \widetilde{F}_1) = \sum_{i=1}^d \widehat{\sigma}_i = \sum_{i=1}^d \tilde{\sigma}_i \leq \sum_{i=1}^d \widehat{\sigma}_i$. This implies $\tilde{\sigma}_i = \widehat{\sigma}_i$, $i = 1, \cdots, d$.
- $\langle g_i, \widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)} f_i \rangle = \widehat{\sigma}_i$ $(i = 1, \cdots, d)$.
  If (g_k, f_k) $(k = 1, \cdots, d)$ is a solution of the k-th problem (3.1). Assume $\tilde{g}_k$ ($\tilde{f}_k$ respectively) are the unit orthonormal eigenfunctions of $\widehat{\mathbf{V}}_{YX|Z}^{(m)}$ $(k = 1, \cdots, d)$, take $\widehat{g}_k = (\widehat{\mathbf{\Sigma}}_{YY}^{(m)} + \varepsilon_m I)^{-1/2} \tilde{g}_k$, $\widehat{f}_k = (\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{-1/2} \tilde{f}_k$, then $\langle \widehat{f}_k, (\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I) \widehat{f}_k \rangle = 1$, $\langle \widehat{g}_k, (\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I) \widehat{g}_k \rangle = 1$, $\forall k = 1 \cdots, d$, and

$$(\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{1/2} \widehat{f}_k \perp (\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{1/2} \widehat{f}_j, \qquad (\widehat{\mathbf{\Sigma}}_{YY}^{(m)} + \varepsilon_m I)^{1/2} \widehat{g}_k \perp (\widehat{\mathbf{\Sigma}}_{YY}^{(m)} + \varepsilon_m I)^{1/2} \widehat{g}_j (k \neq j).$$

  Hence

$$\langle g_k, \widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)} f_k \rangle \geq \langle \widehat{g}_k, \widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)} \widehat{f}_k \rangle = \widehat{\sigma}_k, \qquad \forall k = 1, \cdots, d. \tag{4.3}$$

  Next, we will prove that $\widehat{\sigma}_k = \langle g_k, \widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)} f_k \rangle$ by induction. When $d = 1$, it is obvious that $\langle g_1, \widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)} f_1 \rangle = \widehat{\sigma}_1$. Assume $1 \leq k < d$, and $\langle g_i, \widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)} f_i \rangle = \widehat{\sigma}_i$ for all $1 \leq i \leq k$, we shall prove $\langle g_{k+1}, \widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)} f_{k+1} \rangle = \widehat{\sigma}_{k+1}$.
  For any (g, f) satisfying

$$\langle f, (\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I) f \rangle = 1, (\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{1/2} f \perp \{(\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{1/2} f_1, \cdots, (\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{1/2} f_k\},$$
$$\langle g, (\widehat{\mathbf{\Sigma}}_{YY}^{(m)} + \varepsilon_m I) g \rangle = 1, (\widehat{\mathbf{\Sigma}}_{YY}^{(m)} + \varepsilon_m I)^{1/2} g \perp \{(\widehat{\mathbf{\Sigma}}_{YY}^{(m)} + \varepsilon_m I)^{1/2} g_1, \cdots, (\widehat{\mathbf{\Sigma}}_{YY}^{(m)} + \varepsilon_m I)^{1/2} g_k\}.$$

  Then $\breve{F} = (f_1, \cdots, f_k, f)$, $\breve{G} = (g_1, \cdots, g_k, g)$ satisfy $\breve{F}^T (\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I) \breve{F} = I$, $\breve{G}^T (\widehat{\mathbf{\Sigma}}_{YY}^{(m)} + \varepsilon_m I) \breve{G} = I$, this yields that

$$\sum_{i=1}^{k+1} \widehat{\sigma}_i \geq \mathrm{Trace}(\breve{G}^T \widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)} \breve{F}) = \sum_{i=1}^{k} \widehat{\sigma}_i + \langle g, \widehat{\mathbf{\Sigma}}_{YX|Z}^{(m)} f \rangle.$$

Thus, $\langle g, \widehat{\mathbf{\Sigma}}^{(m)}_{YX|Z} f\rangle \leq \widehat{\sigma}_{k+1}$, which means $\langle g_{k+1}, \widehat{\mathbf{\Sigma}}^{(m)}_{YX|Z} f_{k+1}\rangle \leq \widehat{\sigma}_{k+1}$. Combining this with Eq. (4.3), we get

$$\langle g_{k+1}, \widehat{\mathbf{\Sigma}}^{(m)}_{YX|Z} f_{k+1}\rangle = \widehat{\sigma}_{k+1}.$$

From the above argument, we can see that if (g_k, f_k) $(\forall k = 1, \cdots, d)$ is a solution of the k-th problem (3.1), then $G = (g_1, \cdots, g_d)$, $F = (f_1, \cdots, f_d)$ is the solution of problem (3.2).

- Let $(\widetilde{G}, \widetilde{F})$ be a solution of problem (4.2), then there exist orthogonal matrices Q_3, Q_4 such that

$$\widetilde{G}^T \widehat{\mathbf{V}}^{(m)}_{YX|Z} \widetilde{F} = Q_3 C Q_4^T.$$

Hence

$$(\widetilde{G}Q_3)^T \widehat{\mathbf{V}}^{(m)}_{YX|Z} \widetilde{F} Q_4 = C.$$

Then $\tilde{g}_k^T \widehat{\mathbf{V}}^{(m)}_{YX|Z} \tilde{f}_k = \widehat{\sigma}_k$, $k = 1, \cdots, d$, $(\tilde{g}_k, \tilde{f}_k)$ $(k = 1, \cdots, d)$ satisfies constraints of the k-th problem (3.1), and thus is the solution of it. Therefore, if $(\widetilde{G}, \widetilde{F})$ is a solution of problem (4.2), then there exists orthogonal matrices Q_3, Q_4 such that $(\tilde{g}_k, \tilde{f}_k)$ is the solution of the k-th problem (3.1), and

$$(\tilde{g}_1, \cdots, \tilde{g}_d) = \widetilde{G}Q_3, \qquad (\tilde{f}_1, \cdots, \tilde{f}_d) = \widetilde{F}Q_4.$$

This completes the proof. □

5. Conclusions

In this paper, we introduce a new conditional kernel CCA algorithm motivated by the conditional dependence measure presented in [11] and the discussion about kernel CCA in [10]. The algorithm and theoretical analysis for conditional CCA are elegantly conducted under mild conditions on $\mathbf{V}_{YX}$, $\mathbf{V}_{YZ}$ and $\mathbf{V}_{ZX}$. We demonstrate that these conditions are closely related with mean square contingency as indicated in Section 2. Meantime, the convergence rates of empirical NCCCO to NCCCO are conducted under the above conditions in the sense of Hilbert–Schmidt norm, which is the extension of Theorem 5 in [11]. Moreover, the multiple extension of conditional kernel CCA has also been addressed in Section 3, which can be viewed as a generalization of Theorem 3.1 in [5].

There are some practical problems that remain to be addressed for conditional kernel CCA. One is how to choose the regularization constant ε_m in practice. The final convergence rates of our algorithm are "dragged slow" due to the sufficient condition of ε_m. That is $\varepsilon_m = m^{-\alpha}$, $0 < \alpha < \frac{1}{3}$. This problem should be studied more in our future work. Moreover, how to find simpler conditions than AA and improve the convergence rates of conditional kernel CCA will be investigated in the future. Another important unsolved problem is the choice of kernel. Kernel method is efficient for detecting nonlinear relations between variables. Successful applications of kernel-based algorithms are widespread in the community of learning theory. Thus, in order to improve the learning rates of conditional kernel CCA, how to choose an optimal combination of kernels is crucial in the literature of CCA related problems. A combination of Gaussian kernel and polynomial kernel was studied in [26] for kernel CCA problem, which shows good performance in the community of kernel learning. But the theoretical analysis of it is still not clear and this will be investigated in the future work.

Acknowledgments

The work described in this paper is supported partially by National Natural Science Foundation of China (No. 11401112), Nature Science Foundation of Shandong province, China (No. ZR2014AM010), Foundation

for Distinguished Young Talents in Higher Education of Guangdong (No. 2013LYM0032), Science and Technology Innovation Project of Guangdong (No. 2013KJCX0083) and Guangdong University of Finance & Economics Grant (No. 12GJPY11001). We would like to thank Prof. Ding Xuan Zhou, Dr. Xin Guo for useful discussions which have helped to improve the presentation of the paper.

Appendix A

We need the following lemma to bound $\|\widehat{\mathbf{V}}_{YX|Z}^{(m)} - \mathbf{V}_{YX|Z}\|_{\mathrm{HS}}$.

Lemma 5. *For any $0 < \delta < 1$, with confidence at least $1-\delta$, we have*

$$\|\widehat{\mathbf{\Sigma}}_{YX}^{(m)} - \mathbf{\Sigma}_{YX}\|_{\mathrm{HS}} \le \frac{6\kappa_1\kappa_2\log(6/\delta)}{\sqrt{m}}, \quad \|\widehat{\mathbf{\Sigma}}_{YZ}^{(m)} - \mathbf{\Sigma}_{YZ}\|_{\mathrm{HS}} \le \frac{6\kappa_2\kappa_3\log(6/\delta)}{\sqrt{m}},$$
$$\|\widehat{\mathbf{\Sigma}}_{ZX}^{(m)} - \mathbf{\Sigma}_{ZX}\|_{\mathrm{HS}} \le \frac{6\kappa_1\kappa_3\log(6/\delta)}{\sqrt{m}}.$$

The other two inequalities can be derived by following the same ideas as shown in the proof of the first inequality. More details can be found in [4].

Proposition 2. *The cross-covariance operator $\widehat{\mathbf{\Sigma}}_{ZX}^{(m)}$ can be represented as $\widehat{\mathbf{\Sigma}}_{ZX}^{(m)} = (\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{1/2}\widetilde{\mathbf{V}}_{ZX}^{(m)}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)})^{1/2}$, where $\widetilde{\mathbf{V}}_{ZX}^{(m)}$ is a bounded linear operator such that $\widetilde{\mathbf{V}}_{ZX}^{(m)} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Z}}$ and $\|\widetilde{\mathbf{V}}_{ZX}^{(m)}\| \le 1$. If $\widehat{\mathbf{V}}_{ZX}^{(m)} = (\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)} + \varepsilon_m I)^{-1/2}\widehat{\mathbf{\Sigma}}_{ZX}^{(m)}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{-1/2}$, then $\|\widehat{\mathbf{V}}_{ZX}^{(m)}\| \le 1$.*

Proof. Let s be any fixed element in $\mathcal{R}((\widehat{\mathbf{\Sigma}}_{XX}^{(m)})^{1/2})$ with f any element of $\mathcal{H}_{\mathcal{X}}$ satisfying $(\widehat{\mathbf{\Sigma}}_{XX}^{(m)})^{1/2}f = s$. Define a linear functional h_s on $\mathcal{R}((\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{1/2})$ by

$$\begin{aligned} h_s\big((\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{1/2}g\big) &= \langle g, \widehat{\mathbf{\Sigma}}_{ZX}^{(m)}f\rangle \\ &= \frac{1}{m}\sum_{i=1}^{m}\Big(f(X_i) - \frac{1}{m}\sum_{t=1}^{m}f(X_t)\Big)\Big(g(Z_i) - \frac{1}{m}\sum_{t=1}^{m}g(Z_t)\Big) \\ &\le \Big(\langle f, \widehat{\mathbf{\Sigma}}_{XX}^{(m)}f\rangle\Big)^{1/2}\Big(\langle g, \widehat{\mathbf{\Sigma}}_{ZZ}^{(m)}g\rangle\Big)^{1/2} \\ &= \|(\widehat{\mathbf{\Sigma}}_{XX}^{(m)})^{1/2}f\| \cdot \|(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{1/2}g\| \qquad \forall g \in \mathcal{H}_{\mathcal{Z}}. \end{aligned}$$

Hence $\Big|h_s\big((\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{1/2}g\big)\Big| \le \|s\| \cdot \|(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{1/2}g\|$, h_s is bounded on $\mathcal{R}((\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{1/2})$, and thus can be extended by continuity to a bounded linear functional on $\overline{\mathcal{R}(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})}$, the extension has norm $\le \|s\|$. By Riesz' theorem, there exists a unique element h such that $h_s(w) = \langle h, w\rangle$, $\forall\ w \in \overline{\mathcal{R}(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})}$ and $\|h\| \le \|s\|$.

Define a map $\widehat{\mathbf{V}}_{ZX}^{(m)} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Z}}$ by $\widehat{\mathbf{V}}_{ZX}^{(m)}s = h$, then $\widehat{\mathbf{V}}_{ZX}^{(m)}$ is defined for all s in $\mathcal{R}((\widehat{\mathbf{\Sigma}}_{XX}^{(m)})^{1/2})$. It is linear, single-valued and bounded because $\|\widehat{\mathbf{V}}_{ZX}^{(m)}s\| \le \|s\|$. Thus $\widehat{\mathbf{V}}_{ZX}^{(m)}$ can be extended by continuity to a bounded linear operator $\widetilde{\mathbf{V}}_{ZX}^{(m)}$ defined on $\overline{\mathcal{R}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)})}$, and $h_s(w) = \langle\widetilde{\mathbf{V}}_{ZX}^{(m)}s, w\rangle$. We can extend the domain of $\widetilde{\mathbf{V}}_{ZX}^{(m)}$ to all of $\mathcal{H}_{\mathcal{X}}$ by defining $\widetilde{\mathbf{V}}_{ZX}^{(m)}f = 0$ for $f \in \Big(\overline{\mathcal{R}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)})}\Big)^{\perp}$.

Hence for any $f \in \mathcal{H}_{\mathcal{X}}$, $s = (\widehat{\mathbf{\Sigma}}_{XX}^{(m)})^{1/2}f$ and for any $g \in \mathcal{H}_{\mathcal{Z}}$, we have

$$h_s\big((\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{1/2}g\big) = \langle g, \widehat{\mathbf{\Sigma}}_{ZX}^{(m)}f\rangle = \langle(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{1/2}g, \widetilde{\mathbf{V}}_{ZX}^{(m)}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)})^{1/2}f\rangle.$$

Thus

$$\widehat{\mathbf{\Sigma}}_{ZX}^{(m)} = (\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{1/2}\widetilde{\mathbf{V}}_{ZX}^{(m)}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)})^{1/2},$$

and $\|\widetilde{\mathbf{V}}_{ZX}^{(m)}\| \leq 1$, then

$$\begin{aligned}\|\widehat{\mathbf{V}}_{ZX}^{(m)}\| &= \|(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)} + \varepsilon_m I)^{-1/2}\widehat{\mathbf{\Sigma}}_{ZX}^{(m)}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{-1/2}\| \\ &= \|(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)} + \varepsilon_m I)^{-1/2}(\widehat{\mathbf{\Sigma}}_{ZZ}^{(m)})^{1/2}\widetilde{\mathbf{V}}_{ZX}^{(m)}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)})^{1/2}(\widehat{\mathbf{\Sigma}}_{XX}^{(m)} + \varepsilon_m I)^{-1/2}\| \\ &\leq 1.\end{aligned}$$

The conclusion follows. □

References

[1] S. Akaho, A kernel method for canonical correlation analysis, in: International Meeting of Psychometric Society, Osaka, Japan, 2001.
[2] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, 3rd ed., John Wiley & Sons, New York, 2003.
[3] C.R. Baker, Joint measures and cross-covariance operators, Trans. Amer. Math. Soc. 186 (1973) 273–289.
[4] J. Cai, H.W. Sun, Convergence rate of kernel canonical correlation analysis, Sci. China Math. 54 (2011) 2161–2170.
[5] D. Chu, L. Liao, M. Ng, et al., Sparse canonical correlation analysis: new formulation and algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 3050–3065.
[6] N. Cristianini, J. Shawe-Taylor, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, UK, 2004.
[7] F. Cucker, D.X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, Cambridge, UK, 2007.
[8] E. De Vito, L. Rosasco, A. Caponnetto, et al., Some properties of regularized kernel methods, J. Mach. Learn. Res. 5 (2004) 1363–1390.
[9] K. Fukumizu, F.R. Bach, A. Gretton, Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces, J. Mach. Learn. Res. 5 (2004) 73–99.
[10] K. Fukumizu, F.R. Bach, A. Gretton, Statistical consistency of kernel canonical correlation analysis, J. Mach. Learn. Res. 8 (2007) 361–383.
[11] K. Fukumizu, A. Gretton, X. Sun, et al., Kernel measures of conditional dependence, in: NIPS, vol. 20, 2008, pp. 489–496.
[12] P.L. Lai, C. Fyfe, Kernel and nonlinear canonical correlation analysis, Int. J. Neural Syst. 10 (2001) 365–374.
[13] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (2004) 2639–2664.
[14] D.R. Hardoon, J. Shawe-Taylor, O. Friman, KCCA for fMRI analysis, in: Proceedings of Medical Image Understanding and Analysis, 2004.
[15] D.R. Hardoon, J. Shawe-Taylor, Convergence analysis of kernel canonical correlation analysis: theory and practice, Mach. Learn. 74 (2009) 23–38.
[16] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, 1991.
[17] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 312–377.
[18] T. Hu, J. Fan, Q. Wu, et al., Regularization schemes for minimum error entropy principle, Anal. Appl. 13 (2015) 437, http://dx.doi.org/10.1142/S0219530514500110.
[19] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2007) 153–172.
[20] L. Sun, S. Ji, J. Ye, Canonical correlation analysis for multi-label classification: a least squares formulation, extensions and analysis, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 194–200.
[21] X. Sun, D. Janzing, B. Schölkopf, et al., A kernel-based causal learning algorithm, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 855–862.
[22] J.P. Vert, M. Kanehisa, Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA, in: NIPS, 2002.
[23] Y. Yamanishi, J.P. Vert, A. Nakaya, et al., Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis, Bioinformatics 19 (1) (2003) i323–i330.
[24] A. Vinokourov, J. Shawe-Taylor, N. Cristianini, Inferring a semantic representation of text via cross-language correlation analysis, in: NIPS, vol. 15, 2002, pp. 1473–1480.
[25] L. Wang, X. Wang, J. Feng, Subspace distance analysis with application to adaptive bayesian algorithm for face recognition, Pattern Recognit. 39 (2006) 456–464.
[26] X.F. Zhu, Z. Huang, H.T. Shen, et al., Dimensionality reduction by mixed kernel canonical correlation analysis, Pattern Recognit. 45 (2012) 3003–3016.