Engineering Applications of Artificial Intelligence 64 (2017) 33–42

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

# Robust kernel canonical correlation analysis with applications to information retrieval

Jia Cai [a,*], Xiaolin Huang [b,c]

[a] School of Mathematics and Statistics, Guangdong University of Finance & Economics, Guangzhou, Guangdong, 510320, China
[b] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
[c] MOE Key Laboratory of System Control and Information Processing, 800 Dongchuan Road, Shanghai, 200240, China

ARTICLE INFO

*MSC:*
68T05
62H20

*Keywords:*
Kernel CCA
Singular value decomposition
Reproducing kernel Hilbert space
Cross-language document retrieval
Content-based image retrieval

ABSTRACT

Canonical correlation analysis (CCA) is a powerful statistical tool quantifying correlations between two sets of multidimensional variables. CCA cannot detect nonlinear relationship, and it is costly to derive canonical variates for high-dimensional data. Kernel CCA, a nonlinear extension of the CCA method, can efficiently exploit nonlinear relations and reduce high dimensionality. However, kernel CCA yields the so called over-fitting phenomenon in the high-dimensional feature space. To handle the shortcomings of kernel CCA, this paper develops a novel robust kernel CCA algorithm (KCCA-ROB). The derived method begins with reformulating the traditional generalized eigenvalue–eigenvector problem into a new framework. Under this novel framework, we develop a stable and fast algorithm by means of singular value decomposition (SVD) method. Experimental results on both a simulated dataset and real-world datasets demonstrate the effectiveness of the developed method.



## 1. Introduction

With the rapid development of science and technology, we are confronted with the challenging problem of finding relationship from a large amount of data. It has been a long history for analyzing this ubiquitous relationship. Canonical correlation analysis (CCA) (Hotelling, 1936), as such a paradigm, is a powerful statistical tool for detecting the latent mutual information between two sets of multidimensional variates. The two sets of multidimensional variables can be regarded as two distinct objects or two views of the same object. CCA aims at finding a pair of linear transformations, such that the transformed variables in the lower dimensional space are maximally correlated. Hence it has been widely used in a variety of distinct fields: cross-language document retrieval (Vinokourov et al., 2002), genomic data analysis (Yamanishi et al., 2003), functional magnetic resonance imaging (Hardoon et al., 2004a), multi-view learning (Farquhar et al., 2005; Kakade and Foster, 2007; Sun, 2013) etc. Kettenring (1971) extended CCA to the setting of more than two sets. Generalized CCA was proposed by Tenenhaus and Tenenhaus (2014) for studying multiblock data analysis. Furthermore, tensor CCA was introduced (Luo et al., 2015) to handle the data with arbitrary number of views. Sparse CCA algorithms were investigated by Chu et al. (2013a), Waaijenborg et al. (2008), Witten et al. (2009).

Nonetheless, the major drawback of CCA is that it cannot capture nonlinear relations among variables. Especially for the data that are not in the forms of vectors, for instance, in images, microarray data and so on. Therefore deep CCA (Andrew et al., 2013) was introduced to tackle this issue by employing the idea of deep learning method. However, how many layers should be selected is still an open problem. Another commonly used technique for the nonlinear extension of CCA is the kernel trick, resulting in kernel CCA. The main idea of kernel CCA is to map the variables into a higher-dimensional feature space, and then apply CCA in the RKHSs (reproducing kernel Hilbert spaces, see Cucker and Zhou (2007); De Vito et al. (2004); Zhou (2003) and the references therein). Kernel CCA can achieve dimension reduction results and detect nonlinear relationships. Hence, it has been extensively used in biology and neurology (Hardoon et al., 2004a; Vert and Kanehisa, 2002), content-based image retrieval (Hardoon et al., 2004b), natural language processing (Vinokourov et al., 2002). In the theoretical analysis of kernel CCA, convergence analysis was studied by Hardoon and Shawe-Taylor (2009) via Rademacher complexity. Fukumizu et al. (2007) conducted statistical consistency of kernel CCA from the cross-covariance operator viewpoint. Cai and Sun (2011) investigated it under the AC condition, which is an assumption about the relationship between the eigenvalues of cross-covariance operator and covariance operators.

* Corresponding author.
*E-mail addresses:* jiacai1999@gdufe.edu.cn (J. Cai), xiaolinhuang@sjtu.edu.cn (X. Huang).

http://dx.doi.org/10.1016/j.engappai.2017.05.016
Received 26 August 2016; Received in revised form 17 May 2017; Accepted 23 May 2017
0952-1976/

One crucial problem of the kernel CCA is the so called over-fitting phenomenon. One way is to use the regularization technique to handle it, and cross validation (CV) method was used to select the optimal regularization parameter. However, it is time-consuming to utilize CV to select the tuning parameter and the parameter selected by CV does not necessarily lead to the best performance for the test dataset. How to select appropriate kernels is another problem. Zhu et al. (2012) proposed a mixed kernel CCA, which combines polynomial kernel and Gaussian kernel for the purpose of dimension reduction. Hardoon et al. (2004b) utilized a partial Gram–Schmidt orthogonalization to solve the kernel CCA issue. We still need to choose the optimal regularization parameter, however. Motivated by the idea of Xing et al. (2016) for the CCA problem, this paper will focus on the stability analysis of kernel CCA, and develop a novel robust kernel CCA algorithm for information retrieval related tasks. Numerical experiments on both simulated dataset and real-world datasets, including content-based image retrieval and cross-language document retrieval, demonstrate the effectiveness and the feasibility of the algorithm. The rest of the paper is organized as follows. We review the CCA and the kernel CCA in Section 2. Section 3 is dedicated to the depiction of the new algorithm. Section 4 gives the experimental results. We conclude this paper and discuss future works in Section 5.

## 2. Background

In this section, we will give a brief review of CCA and kernel CCA. Let $x \in \mathbb{R}^{n_1}$ and $y \in \mathbb{R}^{n_2}$ be two random variables. Given $m$ observations $\{x_i, y_i\}_{i=1}^m$. Denote $X = (x_1, \dots, x_m) \in \mathbb{R}^{n_1 \times m}, Y = (y_1, \dots, y_m) \in \mathbb{R}^{n_2 \times m}$. One usually assumes that $X$ and $Y$ are centralized ($\sum_{i=1}^m x_i = 0, \sum_{i=1}^m y_i = 0$) without loss of generality (w.l.o.g.). Then CCA solves

$$\begin{aligned} \max_{w_x, w_y} \quad & w_x^T X Y^T w_y \\ s.t. \quad & w_x^T X X^T w_x = 1, \\ & w_y^T Y Y^T w_y = 1. \end{aligned}$$

When $n_1$ or $n_2$ is very large, it is time-consuming to find canonical variates $w_x$ and $w_y$. Obviously, CCA cannot detect nonlinear relations. To handle this issue, kernel CCA was introduced. It starts to construct feature mappings $\phi_x$ and $\phi_y$ such that $X$ and $Y$ can be converted into

$$\Phi_x = (\phi_x(x_1), \dots, \phi_x(x_m)) \in \mathbb{R}^{\mathcal{N}_1 \times m}, \quad \Phi_y = (\phi_y(y_1), \dots, \phi_y(y_m)) \in \mathbb{R}^{\mathcal{N}_2 \times m},$$

where $\mathcal{N}_1$ (resp. $\mathcal{N}_2$) is the dimension of reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\mathcal{X}$ (resp. $\mathcal{H}_\mathcal{Y}$), maybe infinite dimension. Applying the so-called kernel trick, we can introduce $k_\mathcal{X}(x_1, x_2)$ such that $k_\mathcal{X}(x_1, x_2) = \langle \phi_x(x_1), \phi_x(x_2) \rangle_{\mathcal{H}_\mathcal{X}}$ ($k_\mathcal{Y}(y_1, y_2) = \langle \phi_y(y_1), \phi_y(y_2) \rangle_{\mathcal{H}_\mathcal{Y}}$), where $\langle , \rangle$ is the inner product in respective hypothesis space. Denote the Gram matrices $K_x = \langle \Phi_x, \Phi_x \rangle = (k_\mathcal{X}(x_i, x_j))_{i,j=1}^m$, $K_y = \langle \Phi_y, \Phi_y \rangle = (k_\mathcal{Y}(y_i, y_j))_{i,j=1}^m$. Assume that $K_x$ and $K_y$ are centralized w.l.o.g. unless otherwise specified. For more details about data centering in RKHS, see Schölkopf and Smola (2002). Kernel CCA seeks linear transformations in the RKHS by taking $w_x = \Phi_x \alpha = \sum_{i=1}^m \alpha_i \phi_x(x_i)$, $w_y = \Phi_y \beta = \sum_{i=1}^m \beta_i \phi_y(y_i)$. Therefore, kernel CCA takes the form

$$\begin{aligned} \max_{\alpha, \beta} \quad & \alpha^T K_x K_y \beta \\ s.t. \quad & \alpha^T K_x^2 \alpha = 1, \\ & \beta^T K_y^2 \beta = 1, \end{aligned} \tag{1}$$

where $\alpha = (\alpha_1, \dots, \alpha_m)^T, \beta = (\beta_1, \dots, \beta_m)^T$. The expression (1) implies that kernel CCA can be viewed as the dual of the original CCA problem. One can see that kernel CCA can reduce dimensionality efficiently. Similar to the forms of multiple CCA (Chu et al., 2013a; Hardoon et al., 2004b), multiple kernel CCA can be defined as the following (see Chu et al. (2013b))

$$\begin{aligned} \max_{W_x, W_y} \quad & \mathrm{Trace}(W_x^T K_x K_y W_y) \\ s.t. \quad & W_x^T K_x^2 W_x = I, \qquad W_x \in \mathbb{R}^{m \times d}, \\ & W_y^T K_y^2 W_y = I, \qquad W_y \in \mathbb{R}^{m \times d}, \end{aligned} \tag{2}$$

where $W_x = (\alpha^1, \dots, \alpha^d), W_y = (\beta^1, \dots, \beta^d)$ consist of dual vectors for $X$ and $Y$, respectively.

Obviously, problem (1) can be solved by means of Lagrangian method. Define

$$L(\lambda_1, \lambda_2, \alpha, \beta) = \alpha^T K_x K_y \beta - \frac{\lambda_1}{2}(\alpha^T K_x^2 \alpha - 1) - \frac{\lambda_2}{2}(\beta^T K_y^2 \beta - 1),$$

Taking derivatives with respect to $\alpha$ and $\beta$, we can see that

$$\frac{\partial L}{\partial \alpha} = K_x K_y \beta - \lambda_1 K_x^2 \alpha = 0, \tag{3}$$

$$\frac{\partial L}{\partial \beta} = K_y K_x \alpha - \lambda_2 K_y^2 \beta = 0, \tag{4}$$

Subtracting $\beta^T$ (the transpose of $\beta$) times Eq. (4) from $\alpha^T$ times Eq. (3) yields that

$$\lambda_2 = \lambda_2 \beta^T K_y^2 \beta = \lambda_1 \alpha^T K_x^2 \alpha = \lambda_1, \tag{5}$$

which implies that $\lambda_1 = \lambda_2$. If $K_x$ and $K_y$ are invertible, then Eq. (4) leads to $\beta = \frac{K_y^{-1} K_x \alpha}{\lambda_1}$. Substitute this into Eq. (3), one can see that $\lambda_1^2 \alpha = I\alpha$, which means $\lambda_1 = 1$. Thus $\lambda_1 = 1$ for every vector $\alpha$. This means we can get perfect correlation for any $\alpha$ and $\beta$ without reference to any specific $\alpha$, and over-fitting phenomenon arises in high-dimensional feature space. Hence a natural question is how to exploit nonlinear relations and circumvent the potential over-fitting problem.

In the next section, we will solve this problem from another viewpoint, which is inspired from the idea of Xing et al. (2016).

## 3. Robust kernel CCA algorithm and main results

### 3.1. Reformulation of kernel CCA

Recall that $K_x K_y \beta = \lambda_1 K_x^2 \alpha$, $K_y K_x \alpha = \lambda_2 K_y^2 \beta$. Simple calculations lead to

$$-K_x K_y \beta + K_x^2 \alpha = \mu K_x^2 \alpha,$$

and

$$-K_y K_x \alpha + K_y^2 \beta = \mu K_y^2 \beta,$$

where $\mu = 1 - \lambda_1 = 1 - \lambda_2$. Denote

$$\xi = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad K = \begin{pmatrix} K_x & 0 \\ 0 & K_y \end{pmatrix}, \quad L = \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}.$$

Therefore, kernel CCA problem can be formulated as a compact generalized eigenvalue problem:

$$KLK\xi = \mu K^2 \xi.$$

Let the reduced SVD (singular value decomposition) of $M = KLK + K^2 \in \mathbb{R}^{2m \times 2m}$ be

$$\begin{aligned} M &= (U_1 \ U_2) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} (U_1^T \ U_2^T) \\ &= U_1 \Sigma_1 U_1^T, \end{aligned} \tag{6}$$

where $U_1 \in \mathbb{R}^{2m \times r}$, $U_2 \in \mathbb{R}^{2m \times (2m-r)}$, $\Sigma_1 \in \mathbb{R}^{r \times r}$. We first have the following properties of kernel CCA problems.

**Lemma 1.** *For the matrix $K$, we have $KU_2 = 0$, and finding the optimal projection vector $\xi$ can be converted into that of $U_1\eta$ for some $\eta \neq 0$.*

**Proof.** Recall that $M = U_1 \Sigma_1 U_1^T$, then $U_2^T(KLK + K^2)U_2 = U_2^T U_1 \Sigma_1 U_1^T U_2 = 0$. On the other hand,

$$KLK + K^2 = K \cdot \frac{\sqrt{2}}{2} L \left( K \cdot \frac{\sqrt{2}}{2} L \right)^T + K^2,$$

which means $KLK$ and $K^2$ are both positive semi-definite matrices. Let $\theta_i$ ($i = 1, \dots, r$) be the i-th column of the matrix $U_2$. Therefore,

$$U_2^T(KLK + K^2)U_2 = U_2^T U_1 \Sigma_1 U_1^T U_2 = 0$$