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
Available online 16 June 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 ϕ_x and ϕ_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 α and β, 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 β^T (the transpose of β) times Eq. (4) from α^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 α. This means we can get perfect correlation for any α and β without reference to any specific α, 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 ξ 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 θ_i ($i = 1, \dots, r$) be the i-th column of the matrix U_2. Therefore,

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

is equivalent to

$$\theta_i^T KLK\theta_i = 0 \qquad \text{and} \qquad \theta_i^T K^2\theta_i = 0.$$

Hence $KU_2 = 0$. This proves the first part of the results.

For the second part, since $U = (U_1 \ U_2) \in \mathbb{R}^{2m\times 2m}$ is an orthogonal matrix, then any vector in $\mathbb{R}^{2m\times 2m}$ can be represented by the linear combinations of the column vectors in U. Therefore, $\xi = U_1\eta + U_2\eta'$ for some $\eta \in \mathbb{R}^r, \eta' \in \mathbb{R}^{2m-r}$, and

$$KLK\xi = KLKU_1\eta \qquad \text{and} \qquad K^2\xi = K^2(U_1\eta + U_2\eta') = K^2U_1\eta.$$

Hence $KLK\xi = \mu K^2\xi$ can be converted into $KLKU_1\eta = \mu K^2U_1\eta$, which completes the proof for the second part of the results.

The conclusion of the above lemma also means that the null space of M is that of K. Then we have the following conclusion for the solution of kernel CCA problem.

Theorem 1. *Denote $M_2 = \Sigma_1^{-1/2}U_1^T K^2 U_1 \Sigma_1^{-1/2}$. Let E^d be the matrix containing the d eigenvectors of M_2, which corresponds to the d largest eigenvalues of the matrix M_2. Let $W = U_1\Sigma_1^{-1/2}E^d$, then the canonical variates can be derived as follows:*

$$W_x = W(1:m), \qquad W_y = W(m+1:2m),$$

which means W_x can be derived from the first m rows of the matrix W, and W_y can be obtained from the last m rows of the matrix W.

Proof. Let $M_1 = \Sigma_1^{-1/2}U_1^T KLKU_1\Sigma_1^{-1/2}$. Obviously,

$$\begin{aligned} M_1 + M_2 &= \Sigma_1^{-1/2}U_1^T KLKU_1\Sigma_1^{-1/2} + \Sigma_1^{-1/2}U_1^T K^2U_1\Sigma_1^{-1/2} \\ &= \Sigma_1^{-1/2}U_1^T(KLK + K^2)U_1\Sigma_1^{-1/2} = I. \end{aligned}$$

The conclusion of Lemma 1 indicates that finding the eigenvectors of $KLK\xi = \gamma K^2\xi$ can be reformulated as that of

$$KLKU_1\eta = \mu K^2U_1\eta. \tag{7}$$

Let $\zeta = \Sigma_1^{1/2}\eta$, then problem (7) can be expressed as

$$KLKU_1\Sigma_1^{-1/2}\zeta = \mu K^2U_1\Sigma_1^{-1/2}\zeta.$$

By multiplying $\Sigma_1^{-1/2}U_1^T$ on both sides, one can see that

$$\Sigma_1^{-1/2}U_1^T KLKU_1\Sigma_1^{-1/2}\zeta = \mu\Sigma_1^{-1/2}U_1^T K^2U_1\Sigma_1^{-1/2}\zeta,$$

which means $M_1\zeta = \mu M_2\zeta$. Furthermore, we come to

$$M_2\zeta = \frac{1}{1+\mu}\zeta. \tag{8}$$

Thus in order to solve problem (2), one only need to find d eigenvectors corresponding to the d largest eigenvalues of (8). This completes the proof by noting the relations between ξ and ζ.

Remark 1. From the above arguments, one can find the canonical variates in a fast and robust manner. Since the matrix Σ_1 only contains nonzero eigenvalues, it is invertible. Moreover, by this reformulation, the constraints $W_x^T K_x^2 W_x = I$ and $W_y^T K_y^2 W_y = I$ are satisfied naturally. We eliminate the influence of "zero eigenvalues" by considering elaborate partition of the eigenvalues of the matrix M. One can also consider SVD of K_x and K_y separately, which will lead to the approach discussed by Chu et al. (2013a, b). In Chu et al. (2013a), SVD method was used to find the solutions of CCA problem without regularization technique, and competitive performance was declared there. Here we use a uniform way to conduct the analysis. We believe the approach described here can detect the mutual information between X and Y without any regularization technique as done in Chu et al. (2013a, b). It is still an open problem whether this truncated trick will overcome over-fitting, and it is not easy to give a rigorous proof. This issue will be the future work.

Remark 2. For practical problems, if one chooses Gaussian kernel for K_x, then $rank(K_x) = m - 1$ after centering. Nonetheless, for CCA problems, we usually use distinct type of kernels: polynomial kernel, Gaussian kernel to deal with information retrieval related tasks. Hence, $rank(K) = min\{rank(K_x), rank(K_y)\} < m - 1$, this can be found in the experimental part of this paper. Moreover, we do not need to choose all the $d(d = rank(K))$ eigenvectors corresponding to the d largest eigenvalues in real-world applications. This is similar to truncate the matrix Σ_1, which is also similar to conduct PCA procedure before utilizing CCA technique. However, PCA procedure may discard dimensions that contain import correlation information between X and Y. How to optimally choose d remains open.

3.2. Robust kernel CCA algorithm

Now we are in position to delve into the algorithm.

Algorithm 1 Robust Kernel CCA (KCCA-ROB).

Input:

Training data $X = (x_1, \ldots, x_m) \in \mathbb{R}^{n_1\times m}$, $Y = (y_1, \ldots, y_m) \in \mathbb{R}^{n_2\times m}$;

Output:

Weight vectors $W_x \in \mathbb{R}^{m\times d}$, $W_y \in \mathbb{R}^{m\times d}$;

1: Compute K_x, K_y, construct matrices K, L;
2: Compute matrix factorizations (6);
3: Let $M_2 = \Sigma_1^{-1/2}U_1^T K^2 U_1 \Sigma_1^{-1/2}$, compute the d eigenvectors corresponding to the d largest eigenvalues of the matrix M_2, and construct E^d.
4: Compute $W = U_1\Sigma_1^{-1/2}E^d$.
5: **return** $W_x = W(1:m)$ and $W_y = W(m+1:2m)$;

4. Experiments

In this section, we evaluate the developed methods on both a simulated dataset and real-world datasets. All the experiments were performed on Matlab R2010a (CCA-PMD was conducted with R 3.2.4) on a computer with 4 Core 2.5 GHZ CPUs and 8GB RAM. The methods used for comparison are listed as the following.

1. KCCA-PGSO. Solving kernel CCA via partial Gram–Schmidt orthogonalization method (Hardoon et al., 2004b). The code can be found in the website of Dr. David Hardoon.[1]
2. CCA-PMD. Implementing CCA by penalized matrix decomposition method (Waaijenborg et al., 2008). An R package implementing CCA-PMD method is available.[2]
3. KCCA-ALB. Accelerated kernel version of linearized Bregman method (Chu et al., 2013a), the codes were provided by Dr. Xiaowei Zhang from Bioinformatics Institute, Agency for Science, Technology and Research, Singapore.
4. KCCA-RK. Solving kernel CCA problem via randomized Kaczmarz method (Cai and Tang, 2015).
5. KCCA-ROB. Utilizing the robust kernel CCA described in Algorithm 1.

Once the data matrices X and Y are available, CCA-PMD will be implemented with the regularization parameters selected from the candidate set $\{0.01 : 0.01 : 1\}$. For the other four kernel methods, linear kernel of type $k_{\text{pol}}(x, y) = x^T y$ will be used in cross-language document retrieval, and Gaussian kernel $k_{\text{gau}}(x, y) = \exp\{\frac{\|x-y\|^2}{2\sigma^2}\}$ will be used in the experiment for simulated dataset. In content-based image retrieval task, linear kernel was used for the annotated text of the images and Gaussian kernel will be used for the images. The tuning parameters which leverage the sparsity and stability of KCCA-ALB are chosen as those in Chu et al. (2013a): $\delta = 0.9$, regularization parameters $\mu_x = \mu_y = 10$, precision parameters $\varepsilon_x = \varepsilon_y = 1e-5$. The maximum iteration counter J in KCCA-RK was selected based on Corollary 1 there.

[1] http://www.davidroihardoon.com/Professional/Code.html
[2] http://cran.r-project.org/web/packages/PMA.index.html.

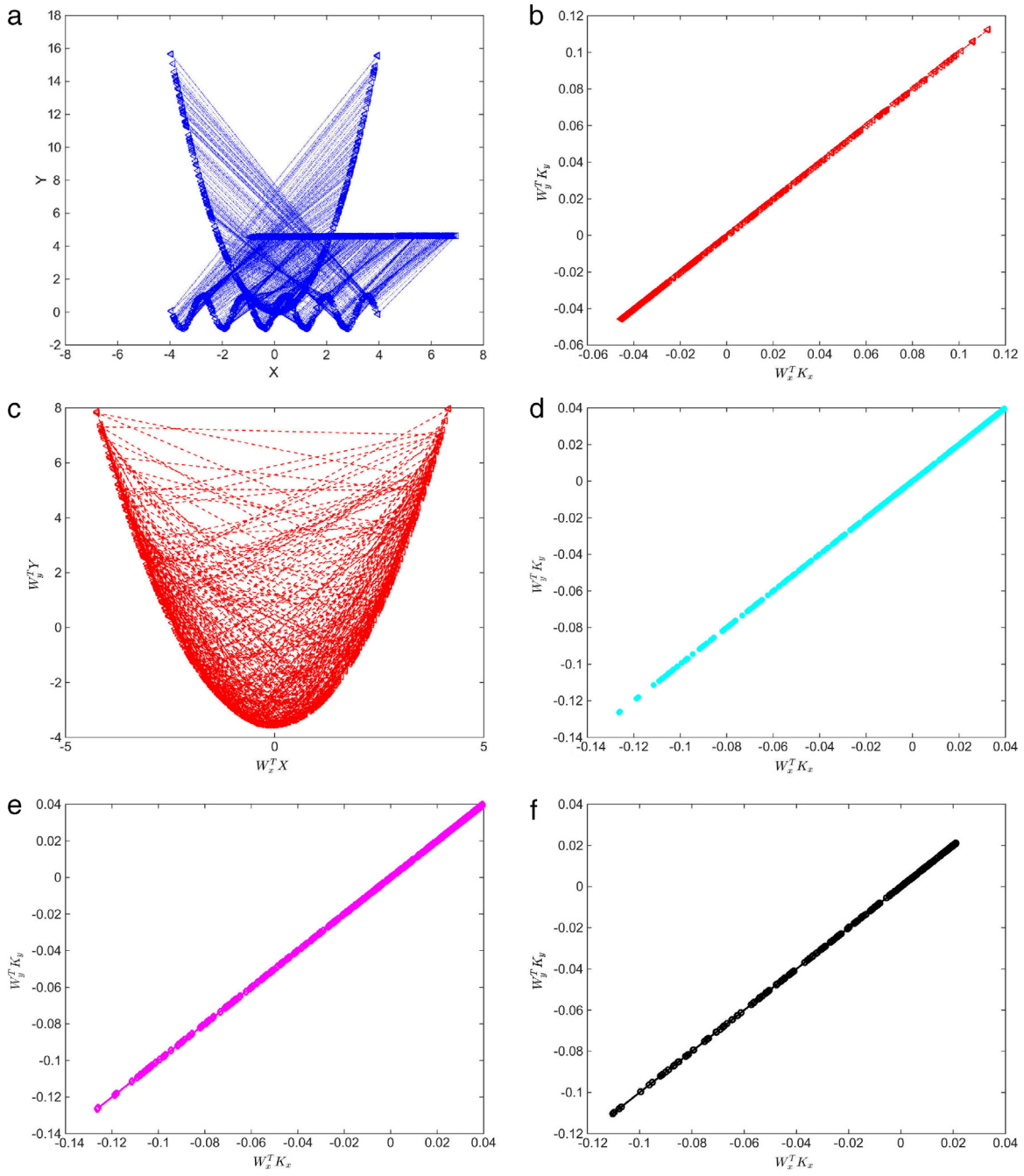


Fig. 1. Original signal and plots of the first pair of canonical variates computed by distinct CCA methods: (a) Original signal, (b) KCCA-PGSO, (c) CCA-PMD, (d) KCCA-ALB, (e) KCCA-RK, (f) KCCA-ROB.

4.1. Simulated dataset

Firstly, we use synthetic data to demonstrate the effectiveness of the proposed algorithm. Let Z be a random variable following the uniform distribution over interval $(-4, 4)$. 500 pairs of (X, Y) are generalized as below:

$$X : [-Z, Z, Z+3] \qquad Y : \left[Z^2, \sin(4Z), \log(Z+100)\right].$$

Apparently, X and Y are nonlinearly related, as shown in Fig. 1(a). Fig. 1(c) shows that some strong nonlinear relationships cannot be explained by CCA-PMD method. The correlation coefficient achieved by CCA-PMD stays at a low level. Instead, kernel CCA methods are all able to finely build the relationship between $(W_x^T K_x, W_y^T K_y)$, where the bandwidth parameter σ of the Gaussian kernel equals to the maximum distance between data points. In other words, KCCA-PGSO, KCCA-ALB, KCCA-RK, and KCCA-ROB can all find W_x and W_y with correlations one ($d = 1$) and the results are plotted in Fig. 1(b)–(f). They can interpret nonlinear relations for two sets of data.

To evaluate the performance of over-fitting, we artificially add noise to Y:

$$Y : [Z^2 + 0.7\varepsilon_1, \sin(4Z) + 0.8\varepsilon_2, \log(Z+100) + 0.9\varepsilon_3],$$

where $\varepsilon_1, \varepsilon_2$ and ε_3 are three noise vectors with

$$\varepsilon_j(i) \sim \mathcal{N}(0, 0.1^2), \quad \forall i = 1, \ldots, 500, j = 1, 2, 3.$$

We again use kernel CCA methods and the same kernel as noise-free cases. If over-fitting happens, the model will follow the trend induced by noise. In other words, over-fitting will make the solution different

Table 1
Comparison results obtained by KCCA-PGSO, CCA-PMD, KCCA-ALB, KCCA-RK and KCCA-ROB methods.

	KCCA-PGSO	CCA-PMD	KCCA-ALB	KCCA-RK	KCCA-ROB
Sumcorr	2.971	0.2194	3	3	2.972
$\frac{\|W_x^T K_x^2 W_x - I_d\|_F}{\sqrt{d}}$	2.2338e−03	1.4142	3.5729e−04	1.1163e−05	3.5264e−08
$\frac{\|W_y^T K_y^2 W_y - I_d\|_F}{\sqrt{d}}$	2.5681e−02	1.0745	9.6884e−04	1.2725e−08	3.5263e−08
diff_x	2.8383	0.6667	11642.666	11639.7307	0.7204
diff_y	18.097	0.6755	20291.5426	13.8405	0.9737

(a) IAPR TC-12 dataset.

(b) Ground Truth Image database.

Fig. 2. The sample images of the datasets used in our experiments.

from the noise-free case. Denote the results without and with noise as $\bar{W}_x, \bar{W}_y$ and $\tilde{W}_x, \tilde{W}_y$. Then the difference can be measured by the average relative difference, i.e.,

$$\mathrm{diff}_x = \frac{1}{d}\sum_i \frac{|\bar{W}_x(i) - \tilde{W}_x(i)|}{|\bar{W}_x(i)|} \text{ and } \mathrm{diff}_y = \frac{1}{d}\sum_i \frac{|\bar{W}_y(i) - \tilde{W}_y(i)|}{|\bar{W}_y(i)|}.$$

We choose $d = 3$ here. In Table 1, we report the correlation and the violation of orthogonality constraints, which are measured by $\frac{\|W_x^T K_x^2 W_x - I_d\|_F}{\sqrt{d}}$ and $\frac{\|W_y^T K_y^2 W_y - I_d\|_F}{\sqrt{d}}$. Here F means Frobenius norm. Notice that for CCA-PMD method, we use X^t, Y^t (test datasets) but the same notations as other methods. The reported results are the average of 10 trials. Accordingly, we can find that the proposed method have nonlinear information retrieval capability as the other kernel CCA, while, it has the minimum distance compared with the other kernel CCA methods and enjoys good stability to noise, which also means that it avoids over-fitting. Hence, it also demonstrates that KCCA-ALB and KCCA-RK methods will find canonical variates, which match the noises, since they achieve perfect correlation coefficient 3 ($d = 3$). It is unreasonable that they still achieved such high correlation coefficient under the effect of noise.

4.2. Real-world datasets

The first part of this subsection will testify the feasibility of the proposed algorithm on two famous image datasets: IAPR TC-12 dataset (Grubinger et al., 2006) and Ground Truth Image database.[3] In the second part, we will apply KCCA-ROB to the task of cross-language document retrieval on two text datasets: English–Spanish dataset and English–French dataset from JRC-Acquis database (Steinberger et al., 2006).

[3] http://www.washingtong.edu/research/imagedatabase/groundtruth/

4.2.1. Content-Based image retrieval

In this experiment, two datsets will be used for testing the efficiency of the proposed algorithm: IAPR TC-12 dataset, Ground Truth Image database (see Fig. 2). IAPR TC-12 dataset has been extensively used in image retrieval related tasks. Because the number of instances in IAPR TC-12 is too large and out of the memory for most computers, we will choose a subset from it, which contains 446 images that have distinct subjects, i.e., waterfall, snow, rock, bed etc. Among them, 141 images were randomly chosen as training data, while the rest 305 images were selected as test data. Ground Truth Image database, which consists of 21 datasets of outdoor scene images, was created at the University of Washington. We use 852 images from 19 datasets, and choose 217 images as training data, and use the rest for test. For the annotated text associated with images, we use the bag-of-words approach to obtain the text features. After removing stop words, stemming, we get a 909 × 446 (resp. 189 × 852) term-document matrix for IAPR TC-12 dataset (resp. Ground Truth Image database). For image features (color, texture), 4 × 5 Gabor filters were used to extract texture features, and each image was divided into 8 × 8 patches to extract HSV (Hue–Saturation–Value) color representation. Then the texture features and color features were concatenated to form the features of an image. For kernel CCA methods, we utilize linear kernel K_x for the text features and Gaussian kernel for the image features, the parameter σ is chosen as the minimal distance between different images. Firstly, we convert texts into a matrix M (annotated text of the images) for given queries in texts, then construct centered linear test kernel K_x^{ts} based upon queries and training texts (Chu et al., 2013b). Next, we project them into the lower dimensional space by computing $W_x^T K_x^{ts}$ for all the test images I^{ts}, and construct the centered test kernel matrix K_y^{ts} based on training images and test images, and project them into a lower dimensional space by computing $W_y^T K_y^{ts}$. Hence we will measure the retrieval precision by computing the distance function

$$h_1(i,j) = \|W_x^T K_x^{ts}(:,i) - W_y^T K_y^{ts}(:,j)\|_2, \quad \text{for } i,j = 1,\ldots,N,$$

for each fixed i, where N is the number of test documents. Thus from the value of $h_1(i,j)$ (for each fixed i), one can get the average

Table 2
Comparison results by KCCA-PGSO, CCA-PMD, KCCA-ALB, KCCA-RK and KCCA-ROB using IAPR TC-12 dataset.

d		10	30	50	70	90	110	Full (134)
ART (in second)	KCCA-PGSO	1.23	3.15	3.18	5.9	6.88	8.38	8.81
	CCA-PMD	2270.31	2656.4	3156.64	3510.18	3996.97	4545.32	5415.78
	KCCA-ALB	7.49	19.95	32.23	46.12	70.68	68.88	101.48
	KCCA-RK	2.11	4.88	7.91	10.75	13.51	15.81	18.91
	KCCA-ROB	1.39	2.37	3.23	6.52	7. 65	8.49	9.65
Sumcorr	KCCA-PGSO	0.818	2.5775	4.5635	6.1976	7.6128	9.3969	11.3614
	CCA-PMD	0.4487	1.3166	2.053	2.8421	3.3505	3.9721	4.8358
	KCCA-ALB	0.8103	2.081	3.5937	5.9342	7.1055	8.7893	11.7883
	KCCA-RK	0.8006	2.0625	3.5701	5.8842	7.0155	8.7006	11.6772
	KCCA-ROB	**1.2274**	**3.1394**	**4.721**	**6.4872**	**7.9077**	**10.3142**	**12.8754**
Average AROC	KCCA-PGSO	0.9089	0.9133	**0.9153**	**0.921**	**0.9175**	0.9153	0.9119
	CCA-PMD	0.508	0.509	0.5077	0.5074	0.5083	0.5084	0.508
	KCCA-ALB	0.896	0.9022	0.9087	0.9162	0.9171	0.9149	**0.9147**
	KCCA-RK	0.899	0.9007	0.9063	0.9141	**0.9175**	0. 9072	0.9127
	KCCA-ROB	**0.9096**	**0.9201**	0.9126	0.9153	0.9138	**0.9203**	0.9113
$\frac{\|W_x^T K_x^2 W_x - I_d\|_F}{\sqrt{d}}$	KCCA-PGSO	3.1821e−011	2.7209e−011	2.4071e−011	2.3075e−011	2.4193e−011	2.5132e−011	3.5701e−011
	CCA-PMD	0.3992	0.7078	1.0845	1.1975	1.3613	1.3497	1.4367
	KCCA-ALB	1.1987e−01	1.6362e−01	1.9979e−01	2.2113e−01	2.6792e−01	3.0443e−01	3.4104e−01
	KCCA-RK	6.2857e−013	7.4874e−013	8.8459e−013	9.5173e−013	1.0993e−012	1.2761e−012	1.4218e−012
	KCCA-ROB	1.3247e−011	1.1486e−011	9.2765e−012	8.0945e−012	7.4536e−012	7.5166e−012	8.9497e−012
$\frac{\|W_y^T K_y^2 W_y - I_d\|_F}{\sqrt{d}}$	KCCA-PGSO	2.2087e−013	7.1422e−013	8.8361e−013	1.041e−012	1.2213e−012	1.3424e−012	1.4163e−012
	CCA-PMD	1.5341	2.6971	3.6293	4.0793	4.5772	4.6441	5.0964
	KCCA-ALB	2.4833e−05	4.4118e−05	5.6663e−05	6.7266e−05	6.7688e−05	7.9674e−05	8.2138e−05
	KCCA-RK	3.6946e−015	5.6721e−015	7.2335e−015	8.3003e−015	9.4002e−015	1.0383e−014	1.1326e−014
	KCCA-ROB	1.3092e−011	1.1492e−011	9.2585e−012	8.0889e−012	7.4177e−012	7.4781e−012	8.8858e−012

Table 3
Comparison results by KCCA-PGSO, CCA-PMD, KCCA-ALB, KCCA-RK and KCCA-ROB using Ground Truth Image database.

d		10	30	50	70	90	110	Full (124)
ART (in second)	KCCA-PGSO	6.52	11.96	27.07	33.77	42.14	46.06	50.26
	CCA-PMD	866.49	1817.26	2246.99	3155.25	3658.49	3788.52	5121.06
	KCCA-ALB	70.18	174.38	286.08	403.77	568.77	569.6	953.41
	KCCA-RK	5.88	11.34	17.24	22.43	26.56	28.29	35.45
	KCCA-ROB	6.86	12.47	26.88	36.58	40.89	48.08	52.42
Sumcorr	KCCA-PGSO	0.9993	**3.8756**	6.5453	9.0005	12.0144	14.9712	16.3914
	CCA-PMD	**1.3214**	2.9029	5.3039	6.976	8.3549	9.1724	9.6803
	KCCA-ALB	1.1777	3.7582	5.829	8.1986	10.543	12.9253	14.9187
	KCCA-RK	1.1778	3.7581	5.8282	8.1973	10. 5415	12.9243	14.9183
	KCCA-ROB	0.6750	3.557	**7.4378**	**11.406**	**14.8173**	**17.9962**	**19.3144**
Average AROC	KCCA-PGSO	0.5775	**0.6467**	0.6822	0.6923	0.7061	0.7156	0.7112
	CCA-PMD	0.5336	0.5392	0.5557	0.5598	0.5614	0.5638	0.5644
	KCCA-ALB	**0.6029**	0.646	0.6531	0.6713	0.6784	0.6922	0.697
	KCCA-RK	0.6028	0.646	0.6532	0.6712	0.6784	0.6923	0.6968
	KCCA-ROB	0.5538	0.6379	**0.6929**	**0.719**	**0.7285**	**0.7344**	**0.7288**
$\frac{\|W_x^T K_x^2 W_x - I_d\|_F}{\sqrt{d}}$	KCCA-PGSO	1.8267e−013	2.1491e−013	2.2159e−013	2.2334e−013	2.2355e−013	2.7796e−013	3.7456e−013
	CCA-PMD	0.9944	1.7017	2.3602	2.9342	3.2631	3.41	3.482
	KCCA-ALB	4.2553e−03	8.3141e−03	1.287e−02	1.2881e−02	1.3909e−02	1.3518e−02	1.5437e−02
	KCCA-RK	6.03e−014	1.0032e−013	1.4949e−013	1.7611e−013	1.8672e−013	2.0211e−013	2.1359e−013
	KCCA-ROB	8.4663e−013	9.9303e−013	8.8908e−013	8.3288e−013	7.8744e−013	7.948e−013	9.3806e−013
$\frac{\|W_y^T K_y^2 W_y - I_d\|_F}{\sqrt{d}}$	KCCA-PGSO	5.6568e−014	1.2218e−013	1.5113e−013	1.6797e−013	1.9331e−013	2.2645e−013	2.665e−013
	CCA-PMD	0.8356	1.3044	1.6595	1.761	1.8494	1.8935	1.8773
	KCCA-ALB	4.3197e−06	1.1423e−05	1.401e−05	1.4454e−05	2.0299e−05	1.5527e−05	1.8379e−05
	KCCA-RK	1.8904e−015	3.1959e−015	4.2058e−015	4.9538e−015	5.6589e−015	6.281e−015	6.6927e−015
	KCCA-ROB	8.3984e−013	9.9055e−013	8.8685e−013	8.3229e−013	7.8594e−013	7.9323e−013	9.3272e−013

area under the receiver operating characteristic (ROC) curve (AROC, see Bradley (1997); Fawcett (2006)) for each i. The comparison results for the performance of KCCA-PGSO, CCA-PMD, KCCA-ALB, KCCA-RK and KCCA-ROB are listed in Tables 2 and 3 and Figs. 3 and 4. For KCCA-RK method, $J = 28$ (resp. 27) for IAPR TC-12 dataset (resp. Ground Truth Image database), and the choices of parameters for the other methods are the same as those described in the last section.

Numerical results using IAPR TC-12 dataset are depicted in Table 2 and Fig. 3, where "Full" means all canonical variates are used, "Sumcorr" stands for summation of canonical correlation between test data, "ART" denotes the average running time. From Table 2 and Fig. 3, we can see that the summation of correlation coefficients for test data all increase when d becomes large. All the methods obey the orthogonality constraints very well, except CCA-PMD and KCCA-ALB methods. AROC values achieved by the proposed algorithm, namely KCCA-ROB are almost the highest, while those obtained by CCA-PMD stay at a lower level, around 0.5. Moreover, CCA-PMD method is very costly, the reason is that it needs to find the optimal regularization parameters by permuting each row of the data matrices X and Y, which is obviously very costly.

Experimental results using Ground Truth Image database are described in Table 3 and Fig. 4. We can see that when $d \geq 50$, KCCA-ROB obtained the largest AROC values and summation of correlation coefficients, and it obeys the orthogonal constraints very well. CCA-PMD does not obey orthogonal constraints and it is very time-consuming to find the solutions.

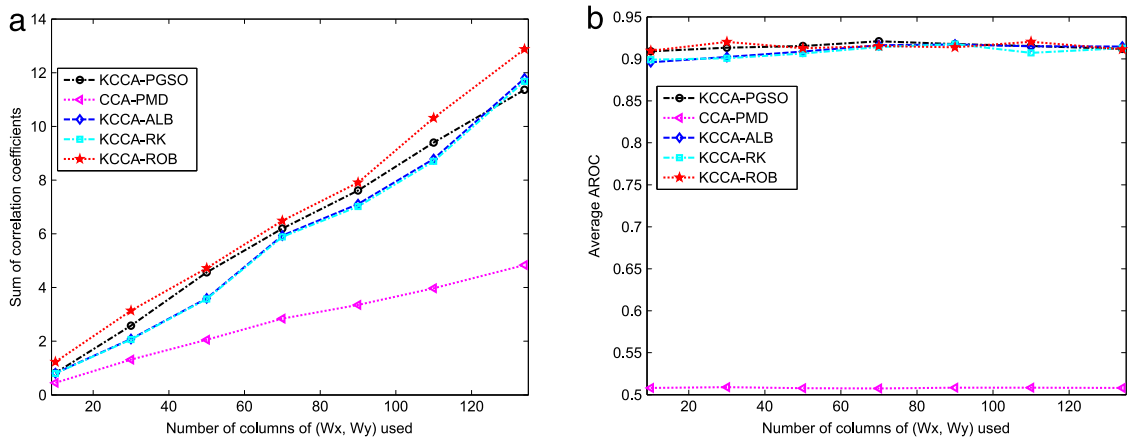


Fig. 3. Experimental results achieved by KCCA-PGSO, CCA-PMD, KCCA-ALB, KCCA-RK and KCCA-ROB using IAPR TC-12 dataset: (a) summation of correlation coefficients, (b) average AROC.

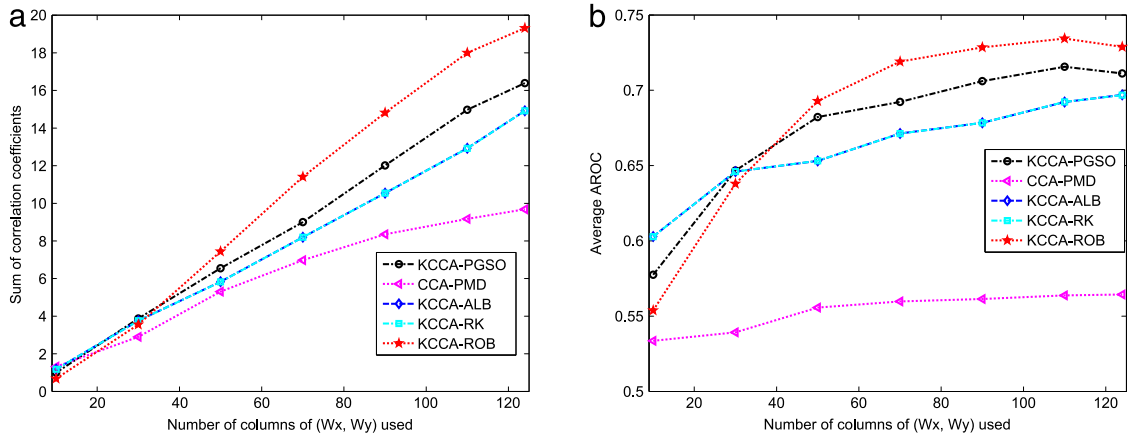


Fig. 4. Experimental results achieved by KCCA-PGSO, CCA-PMD, KCCA-ALB, KCCA-RK and KCCA-ROB using Ground Truth Image database: (a) summation of correlation coefficients, (b) average AROC.

4.2.2. *Cross-Language document retrieval*

Previously, kernel CCA has been proven effective for cross-language document retrieval (CLDR). In this subsection, we apply KCCA-ROB method to CLDR task and present comparison results (the samples can be found in Fig. 5). For given collection of documents in one language, CLDR aims at retrieving the most relevant document in the target language when given a query in another language. We first extract the "body" part of the XML file, and then obtain a bag-of-words representation by using Term Frequency Inverse Document Frequency (TFIDF) approach (for more details, please see Kowalski and Maybury (2002); Salton and McGill (1986); Shawe-Taylor and Cristianini (2004)), which is efficient in the document retrieval task. After removing numbers, stop-words (English, Spanish, French, respectively), and rare words (appearing less than three times), we obtain a 11651 $\times$ 800 term-document "English" matrix and a 19560 $\times$ 800 "Spanish" matrix for the English–Spanish dataset, and a 13776 $\times$ 1000 term-document "English" matrix and a 20115 $\times$ 1000 "French" matrix for the English–French dataset. For KCCA-RK algorithm, the iteration counter parameter $J = 27$ (resp. $J = 28$) for English–Spanish dataset (resp. English–French dataset), the choices of the parameters for the other methods are the same as those described previously.

For given query documents in one language, say English, we first convert the documents into a matrix M, and then construct centered linear test kernel K_x^{ts} based on queries and training texts, then project K_x^{ts} into the lower dimensional space by computing $W_x^T K_x^{ts}$. For the test documents in French (Spanish), we construct test kernel matrix K_y^{ts} and project it in the lower dimensional space by computing $W_y^T K_y^{ts}$. Then we compute the distance function

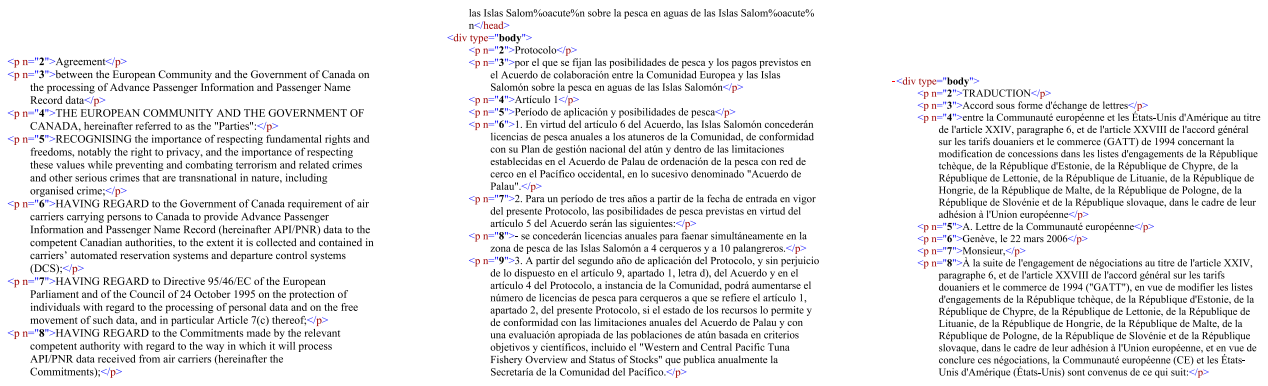
$$h_2(i,j) = \|W_x^T K_x^{ts}(:,i) - W_x^T K_x^{ts}(:,j)\|, \quad \text{for} \quad i,j = 1,\ldots,N$$

for each fixed i, where N is the number of the test documents. For each fixed i, we can get the precision from the value $h_2(i,j)$ by using the average area under the ROC.

Similar to the description for the CBIR task, K_x^{ts} (resp. K_y^{ts}) is replaced with X^t (resp. Y^t, test data) for CCA-PMD method. We randomly choose 200 pairs of documents as training data, and the rest 600 (800 for English–French dataset) pairs of documents are used for test. Tables 4 and 5 and Figs. 6 and 7 present the retrieval accuracy of KCCA-PGSO, CCA-PMD, KCCA-ALB, KCCA-RK and KCCA-ROB methods.

Numerical results for English–Spanish dataset are depicted in Table 4 and Fig. 6. We can see that the proposed KCCA-ROB method has the largest AROC values and Sumcorr (summation of correlation coefficients) when $d \geq 40$, and satisfies the orthogonal constraints very well. CCA-PMD method is very costly to find the canonical variates and it does not obey the orthogonal constraints. Hence in the experiment for the English–French dataset, we will not compare with CCA-PMD method, which is too time-consuming. The experimental results for the English–French dataset are summarized in Table 5 and Fig. 7. Except $d = 10$, KCCA-ROB method has the largest AROC values and Sumcorr. Its result also satisfies the orthogonal constraints very well.

To summarize, we claim that the proposed kernel CCA algorithm is comparable with the existing state-of-the-art kernel CCA methods. The canonical variates were found by solving a generalized eigenvalue–eigenvector problem. Therefore, it satisfies the orthogonal constraints very well and can efficiently remove the redundant information. The



(a) English. (b) Spanish. (c) French.

Fig. 5. The samples of JRC-Acquis database that used in our experiments.

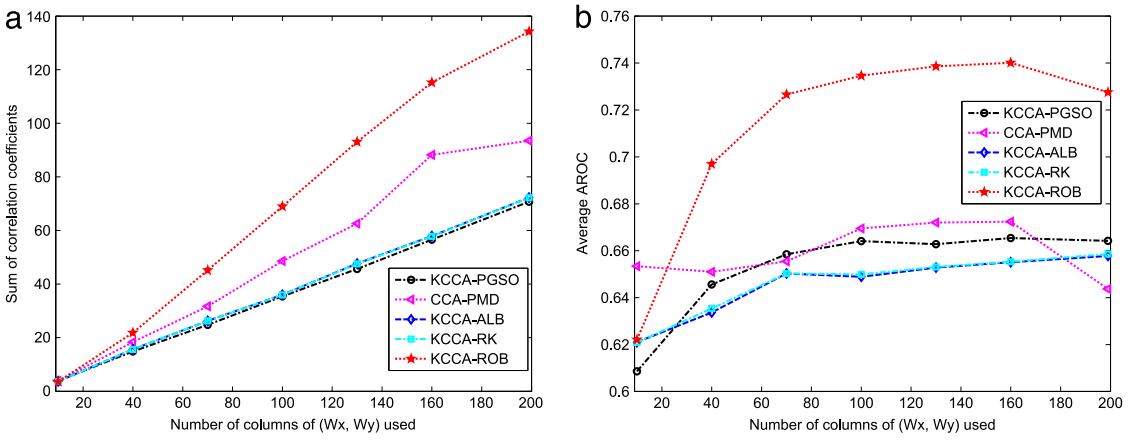


Fig. 6. Experimental results achieved by KCCA-PGSO, CCA-PMD, KCCA-ALB, KCCA-RK and KCCA-ROB using English–Spanish dataset: (a) summation of correlation coefficients, (b) average AROC.

Table 4
Comparison results by KCCA-PGSO, CCA-PMD, KCCA-ALB, KCCA-RK and KCCA-ROB using English–Spanish dataset.

d		10	40	70	100	130	160	Full (199)
ART (in second)	KCCA-PGSO	5.33	13.25	32.94	38.55	48.16	52. 22	62. 4
	CCA-PMD	5831.39	10355.16	15038.16	19262.39	26480.61	27018.41	36122.76
	KCCA-ALB	34.8	103. 76	201. 12	269. 89	337.35	388.71	486. 43
	KCCA-RK	5.11	12.94	20. 91	28.66	33.06	38. 91	47.22
	KCCA-ROB	5.61	13.92	31.75	41. 22	49. 16	56. 41	62. 43
Sumcorr	KCCA-PGSO	**3.9565**	14.8442	24.7958	35. 407	45.6485	56.5642	70.7576
	CCA-PMD	3.8703	18.3518	31.7076	48.6208	62. 5945	88.2141	93.5201
	KCCA-ALB	3.679	15.6388	26.2627	35.9796	47.652	57.901	72.2618
	KCCA-RK	3.6816	15.6409	26.2561	35.9467	47.5613	57.8	72. 2026
	KCCA-ROB	3.5916	**21.8342**	**45.1757**	**68.9818**	**93.0477**	**115. 2215**	**134.1786**
Average AROC	KCCA-PGSO	0.6086	0.6456	0.6585	0.6641	0.6628	0.6654	0.6642
	CCA-PMD	**0.6534**	0.651	0.6555	0.6695	0.672	0.6724	0.6437
	KCCA-ALB	0.621	0.6337	0.6503	0.6489	0.6528	0.6551	0.6578
	KCCA-RK	0.6209	0.6354	0.6503	0.6499	0.6531	0.6553	0.6586
	KCCA-ROB	0.6221	**0.697**	**0.7266**	**0.7346**	**0.7386**	**0.7401**	**0.7275**
$\frac{\|W_x^T K_x^2 W_x - I_d\|_F}{\sqrt{d}}$	KCCA-PGSO	7.16e−011	2.3397e−010	2.0462e−010	2.0004e−010	1.8798e−010	1.9488e−010	2.6959e−010
	CCA-PMD	1.5106e−02	4.8282e−02	1.0988e−01	2.116e−01	6.038e−01	1.0261	1.3851
	KCCA-ALB	1.1712e−01	3.5425e−01	4.3551e−01	6.2563e−01	7.1981e−01	8.3463e−01	9.4962e−01
	KCCA-RK	3.8702e−013	7.6273e−013	9.2934e−013	1.08e−012	1.2448e−012	1.4151e−012	1.8909e−012
	KCCA-ROB	1.4222e−09	7.1243e−10	5.3855e−010	4.5058e−010	3.9519e−010	3.5622e−010	3.2398e−02
$\frac{\|W_y^T K_y^2 W_y - I_d\|_F}{\sqrt{d}}$	KCCA-PGSO	5.1875e−011	1.2613e−010	1.0523e−010	1.0045e−010	9.2886e−010	9.2208e−011	1.2272e−010
	CCA-PMD	1.9535e−02	3.9609e−02	1.0497e−01	2.0579e−02	5.7454e−01	1.016	1.4175
	KCCA-ALB	1.2019e−01	2.187e−01	2.7011e−01	3.2981e−01	3.6735e−01	4.2326e−01	5.1074e−01
	KCCA-RK	2.6564e−013	5.7806e−013	8.2014e−013	9.5771e−013	1.1379e−012	1.2364e−012	1.5118e−012
	KCCA-ROB	1.4222e−09	7.1243e−10	5.3855e−010	4.5058e−010	3.9519e−010	3.5622e−010	2.2434e−02

Table 5
Comparison results by KCCA-PGSO, KCCA-ALB, KCCA-RK and KCCA-ROB using English–French dataset.

d		10	40	70	100	130	160	Full (199)
ART (in second)	KCCA-PGSO	8.9	23.06	53.55	70.32	77.1	90.13	111.8
	KCCA-ALB	16.33	56.77	114.24	151.99	200.84	242.32	303.74
	KCCA-RK	7.83	16.91	27.06	33.8	40.33	48.93	58.44
	KCCA-ROB	9.48	22.03	54.51	69.93	77.34	96.15	111.95
Sumcorr	KCCA-PGSO	3.0692	13.1021	24.7483	37.0364	49.663	61.7811	75.9635
	KCCA-ALB	2.9675	13.0943	23.0459	33.7683	42.4268	52.8954	65.4913
	KCCA-RK	2.9016	12.9512	22.7371	33.3353	42.1895	52.6148	65.1794
	KCCA-ROB	**4.1918**	**21.3183**	**45.6851**	**69.3886**	**93.28**	**116.1166**	**136.4105**
Average AROC	KCCA-PGSO	**0.6541**	0.6572	0.6652	0.679	0.6836	0.6867	0.6829
	KCCA-ALB	0.6087	0.6492	0.655	0.667	0.6629	0.6653	0.6655
	KCCA-RK	0.6142	0.6522	0.6572	0.6678	0.6648	0.6672	0.6667
	KCCA-ROB	0.6432	**0.7099**	**0.7468**	**0.7585**	**0.761**	**0.7645**	**0.7647**
$\frac{\|W_x^T K_x^2 W_x - I_d\|_F}{\sqrt{d}}$	KCCA-PGSO	2.9223e−010	3.316e−010	3.4002e−010	3.2865e−010	3.3281e−010	3.49e−010	4.3729e−010
	KCCA-ALB	3.147e−01	7.4718e−01	8.3275e−01	1.0607	1.1675	1.222	1.3719
	KCCA-RK	7.0208e−013	1.8811e−012	2.2803e−012	2.5019e−012	2.9212e−012	2.9909e−012	3.2009e−012
	KCCA-ROB	7.5374e−010	3.8024e−010	2.8745e−010	2.405e−010	2.1093e−010	1.9014e−010	1.0023e−01
$\frac{\|W_y^T K_y^2 W_y - I_d\|_F}{\sqrt{d}}$	KCCA-PGSO	2.6388e−010	3.8482e−010	3.8029e−010	3.7537e−010	3.7634e−010	3.8872e−010	4.935e−010
	KCCA-ALB	6.886ee-01	1.4504	1.7832	2.1034	2.3746	2.4542	2.6516
	KCCA-RK	1.3395e−011	9.5578e−012	6.5405e−012	7.3307e−012	8.1206e−012	5.6834e−012	8.9241e−012
	KCCA-ROB	7.5375e−010	3.8025e−010	2.8745e−010	2.405e−010	2.1093e−010	1.9014e−010	1.0025e−01

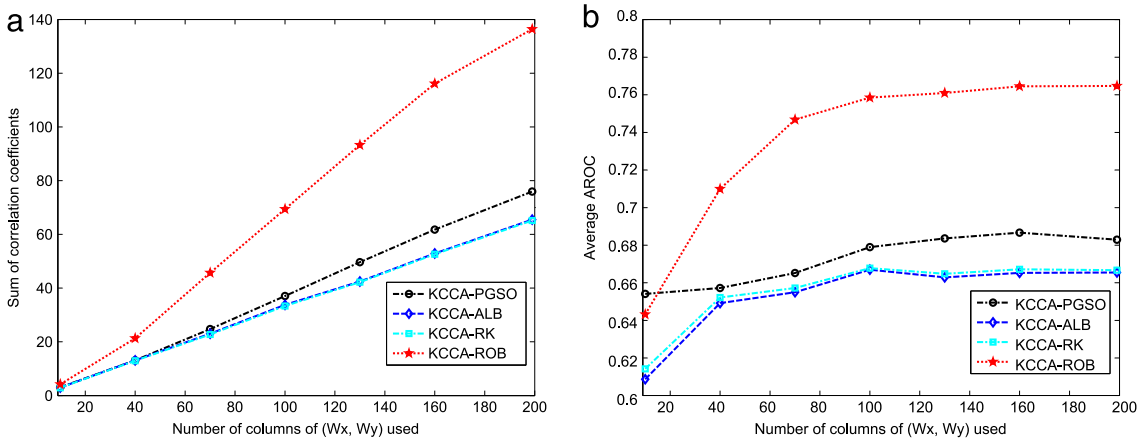


Fig. 7. Experimental results achieved by KCCA-PGSO, KCCA-ALB, KCCA-RK and KCCA-ROB using English–French dataset: (a) summation of correlation coefficients, (b) average AROC.

reason why CCA-PMD is too costly is due to the permutation procedure, which was used to find the optimal regularization parameters. From the above argument about experimental results, one can see that KCCA-ALB is also slightly costly, the reason is because KCCA-ALB aims at finding the sparse solutions of canonical variates, and it needs to shrink the obtained value at each instant. However, it is hard to get both high sparsity and high precision (Xu et al., 2012).

5. Conclusion

Many kernel CCA methods have the over-fitting problem, although they can efficiently detect nonlinear relations and handle the curse of dimensionality. To resolve this issue, we first reformulate the original kernel CCA under a novel framework. In the light of the new framework, we develop a robust kernel CCA algorithm via SVD method. It remains open whether the truncated method could overcome the over-fitting problem. The effectiveness of the proposed algorithm is supported by experimental results on both simulated data and real-world data. In the experiments, KCCA-ROB method achieves high efficiency in terms of AROC, running time, orthogonal constraints when comparing with the existing CCA and kernel CCA methods. However, there are many interesting topics to study in the future: (1) Sparsity is often incorporated into CCA. Is it possible to design sparse kernel CCA algorithm for big data, and how to balance sparsity and stability for CCA? (2) For high-dimensional large data, how to develop distributed kernel CCA algorithms to accelerate the computational efficiency. The applications of CCA in the more challenging tasks of cross-language document retrieval, content-base image retrieval etc. will also be studied in the future.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 11401112, 61603248), National Statistical Science Research Program (2016LZ47) and China Scholarship Council. The author would like to show sincere gratitude to Dr. Xiaowei Zhang from Bioinformatics Institute, Agency for Science, Technology and Research, Singapore for providing the KCCA-ALB code and the preprocessed Ground Truth Image data, Dr. Xin Guo from Hong Kong Polytechnic University for useful discussions about the preprocessing of XML file, which have helped to improve the presentation of the paper.

References

Andrew, G., Arora, R., Bilmes, J., Livescu, K., 2013. Deep canonical correlation analysis. ICML 28, 1247–1255.
Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30, 1145–1159.
Cai, J., Sun, H.W., 2011. Convergence rate of kernel canonical correlation analysis. Sci. China Math. 54, 2161–2170.

Cai, J., Tang, Y., 2015. A new information retrieval algorithm via randomized Kaczmarz kernel canonical correlation analysis. Unpublished results.
Chu, D., Liao, L., Ng, M., Zhang, X.W., 2013a. Sparse canonical correlation analysis: new formulation and algorithm. IEEE Trans. Pattern Anal. Mach. Intell. 35, 3050–3065.
Chu, D., Liao, L., Ng, M., Zhang, X.W., 2013b. Sparse kernel canonical correlation analysis. In: Proceedings of International Multiconference of Engineers and Computer Scientists, IMECS 2013.
Cucker, F., Zhou, D.X., 2007. Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, Cambridge.
De Vito, E., Rosasco, L., Caponnetto, A., Piana, M., Veri, A., 2004. Some properties of regularized kernel methods. J. Mach. Learn. Res. 5, 1363–1390.
Farquhar, J., Hardoon, D.R., Meng, H.Y., Szedmak, S., 2005. Two view learning: SVM-2K, theory and practice. NIPS 355–362.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874.
Fukumizu, K., Bach, F.R., Gretton, A., 2007. Statistical consistency of kernel canonical correlation analysis. J. Mach. Learn. Res. 8, 361–383.
Grubinger, M., Clough, P.D., Müller, H., Deselaers, T., 2006. The iapr benchmark: a new evaluation resource for visual information systems. In: International Conference on Language Resources and Evaluation, LREC, pp. 13–23.
Hardoon, D.R., Shawe-Taylor, J., Friman, O., 2004a. KCCA for fMRI analysis. In: Proceedings of Medical Image Understanding and Analysis, MIUA 2004.
Hardoon, D.R., Szedmak, S., Shawe-Taylor, J., 2004b. Canonical correlation analysis: an overview with application to learning methods. Neural Comput. 16, 2639–2664.
Hardoon, D.R., Shawe-Taylor, J., 2009. Convergence analysis of kernel canonical correlation analysis: theory and practice. Mach. Learn. 74, 23–38.
Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 312–377.
Kakade, S.M., Foster, D.P., 2007. Multi-view regression via canonical correlation analysis. COLT 82–96.
Kettenring, J.R., 1971. Canonical analysis of several sets of variables. Biometrika 58, 433–451.
Kowalski, G., Maybury, M.T., 2002. Information Storage and Retrieval Systems: Theory and Implementation, second ed. Kluwer Academic Publishers.
Luo, Y., Tao, D., Ramamohanarao, K., Xu, C., Wen, Y., 2015. Tensor canonical correlation analysis for multi-view dimension reduction. IEEE Trans. Knowl. Data Eng. 27, 3111–3124.
Salton, G., McGill, M., 1986. Introduction to Modern Information Retrieval. McGraw-Hill, Inc. New York.
Schölkopf, B., Smola, A., 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond. MIT Press.
Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge.
Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. et al., 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC, pp. 2142–2147.
Sun, S.L., 2013. A survey of multi-view machine learning. Neural Comput. Appl. 23, 2031–2038.
Tenenhaus, A., Tenenhaus, M., 2014. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. European J. Oper. Res. 238, 391–403.
Vert, J.P., Kanehisa, M., 2002. Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. NIPS 1449–1456.
Vinokourov, A., Shawe-Taylor, J., Cristianini, N., 2002. Inferring a semantic representation of text via cross-language correlation analysis. NIPS 1473–1480.
Waaijenborg, S., de Witt Hamer, P.C.V., Zwinderman, A.H., 2008. Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. Stat. Appl. Genet. Mol. Biol. 7, article 3.
Witten, D.M., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534.
Xing, X., Wang, K., Yan, T., Lv, Z., 2016. Complete canonical correlation analysis with application to multi-view gait recognition. Pattern Recognit. 50, 107–117.
Xu, H., Caramanis, C., Mannor, S., 2012. Sparse algorithms are not stable: a no-free-lunch theorem. IEEE Trans. Pattern Anal. Mach. Intell. 34, 187–193.
Yamanishi, Y., Vert, J.P., Nakaya, A., Kanehisa, M., 2003. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics 19, i323–i330.
Zhou, D.X., 2003. Capacity of reproducing kernel spaces in learning theory. IEEE Trans. Inform. Theory. 49, 1743–1752.
Zhu, X.F., Huang, Z., Shen, H.T., Cheng, J., C.S., Xu., 2012. Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recognit. 45, 3003–3016.