Linear Algebra and its Applications 437 (2012) 1–13

Contents lists available at SciVerse ScienceDirect

Linear Algebra and its Applications

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

Restricted kernel canonical correlation analysis

Nina Otopal [*]

Institute of Mathematics, Physics and Mechanics, Jadranska 19, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 27 July 2010
Accepted 4 February 2012
Available online 12 March 2012

Submitted by H. Schneider

AMS classification:
62H20
46E22

Keywords:
Kernel
Canonical correlation
Non-negativity

ABSTRACT

Kernel canonical correlation analysis (KCCA) is a procedure for assessing the relationship between two sets of random variables when the classical method, canonical correlation analysis (CCA), fails because of the nonlinearity of the data. The KCCA method is mostly used in machine learning, especially for information retrieval and text mining. Because the data is often represented with non-negative numbers, we propose to incorporate the non-negativity restriction directly into the KCCA method. Similar restrictions have been studied in relation to the classical CCA and called restricted canonical correlation analysis (RCCA), so that we call the proposed method restricted kernel canonical correlation analysis (RKCCA). We also provide some possible approaches for solving the optimization problem to which our method translates. The motivation for introducing RKCCA is given in Section 2.

1. Introduction

Canonical correlation analysis (CCA) was developed by Hoteling in 1936 [1] as a procedure for assessing the relationship between two sets of random variables. He defined a *canonical correlation* (CC) as the maximum correlation between any two linear combinations of random variables from each of the two sets. The method has been applied to many different fields: educational testing problems, neural networks, and data mining. In many practical situations there are some natural restrictions (e.g., positivity, non-negativity, monotonicity) on coefficients in these linear combinations. Das and Sen [4] introduced a method, called *restricted canonical correlation analysis* (RCCA), in which these restrictions are incorporated into the problem of canonical correlation analysis. Further research on this method was done by Omladič and Omladič [5].

[*] Permanent address: Trg svobode 30, 1420 Trbovlje, Slovenia. Tel.: +386 40753579.
E-mail address: nina.otopal@imfm.si

0024-3795/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.laa.2012.02.014

Kernel methods have been developed as a methodology for nonlinear data analysis with positive definite kernels [2]. In kernel methods the data is mapped to a high dimensional Hilbert space corresponding to the chosen positive definite kernel, called *feature space*. The scalar product in feature space and most linear methods in statistics can be computed via this kernel. With the kernel approach the computational power of linear learning machines is increased. Many methods have been proposed as nonlinear extensions of linear methods: *kernel principal component analysis* (KPCA), *Bayesian kernel methods* (Bayesian KM), *kernel canonical correlation analysis* (KCCA), and many others [3].

In this paper we propose a method we call *restricted kernel canonical correlation analysis* (RKCCA). We assume additionally that coefficients in linear combinations of features with correlations that are maximized in KCCA are restricted to be non-negative. We call the solution of RKCCA *restricted kernel canonical correlation* (RKCC). We use, similar to Das and Sen [4], the Karush–Kuhn–Tucker theorem to prove the fact that the squared RKCC equals one of the squared canonical correlations between sub-vectors of two random vectors with known covariance matrices.

The idea of sub-vectors and sub-matrices was first used in [5] for transforming the problem of RKCCA into an optimization problem related to eigenvalues of some generalized eigenvalue problem. For each set of indices the maximal eigenvalue of the generalized eigenvalue problem with sub-matrices and sub-vectors corresponding to this set is considered to be a candidate for the global maximum. The largest of these eigenvalues equals the RKCC.

The paper is organized as follows. The possible applications of RKCCA are presented in Section 2. In Section 3, we describe the KCCA. In Section 4, we propose the RKCCA. In Sections 5 and 6, we discuss possible solutions to the some open problems. A discussion of the complexity of the proposed approach is given at the end of Section 5. The reader will be assumed to have familiarity with some basic theory of reproducing kernels. For those who do not, we postpone a shortcut to this theory until Section 7 so as to not overburden the introductory paragraphs with technicalities.

2. Motivation

The CCA is mostly used in machine learning, especially for information retrieval and text mining [6–8]; therefore, it is often natural to assume that the regression weights are non-negative. One simple example in text mining is when the coefficients are actually weights for words. Because in that case all inputs are non-negative numbers, it is sensible, for the sake of interpretation, to restrict the coefficients also to be non-negatives. In analyzing text documents, we get high-dimensional matrices, and it is almost impossible to invert these matrices or to solve the eigenvalue problem from CCA. This is the reason the kernel method is applied, thus replacing the usual CCA with KCCA. In the dual representation we get matrices of the size $N \times N$ with N equal to the size of the sample. If we used the usual KCCA in this case, it could easily happen that some of the coefficients giving the maximal correlation are negative. Because this makes no sense in solving the text mining problem, we would ignore this group of coefficients and try to find the next one. The next one does not necessarily have the same canonical correlation, so we would probably end up with very different sets of coefficients giving different canonical correlations. We should therefore carefully choose the optimal set depending on the wanted interpretation. If we incorporate non-negativity directly into the problem and solve the RKCCA, the result may become clearer, more powerful, and easier to interpret.

In addition to text mining, the proposed method (RKCCA) could also be useful in functional magnetic resonance imaging (fMRI) analysis. The fMRI is a relatively new tool with the purpose of mapping the sensor, motor, and cognitive tasks to specific regions in the brain. The underlying mechanics of this technique are the regulation of the blood flow as an excess of oxygen is supplied to active neurons causing an increase in oxygenated blood surrounding the tissue of the active brain region. This effect is referred to as the BOLD (blood oxygenation level dependent) signal. Friman et al. [9] have shown that CCA has the ability to introduce several time-courses as the BOLD response has been shown to vary both between people and brain regions. Friman et al. [10] have shown that by using CCA with non-negativity restrictions (RCCA) instead of CCA the detection performance in fMRI analysis is increased. Also, Regnehed et al. [11] used RCCA for fMRI analysis. They have shown that adaptive spatial filtering combined with RCCA performs better than conventional GLM analysis. One factor that has limited the use of RCCA is the absence of an appropriate significance estimation method. In response to this issue, a completely data driven significance estimation method that adapts itself to the underlying data was introduced in [11]. The method was shown to provide accurate control over the false positive rate. On the other hand Hardoon et al. [8] presented a KCCA approach to measure the active regions of the brain using fMRI scans and their activity signal. In their study, KCCA was used to infer brain activity in functional MRI by learning a semantic representation of fMRI brain scans and their associated activity signal. The semantic space provides a common representation and enables a comparison between the fMRI and the activity signal. They compared the approach with CCA by localizing ’activity’ on a simulated null data set and proved that it performs better.

As Friman et al. [10] pointed out the detection performance in fMRI analysis is increased when using RCCA instead of CCA. This gives us strong motivation to introduce RKCCA. We expect this method to perform better compared to KCCA in the areas, where data call for non-negativity restrictions, such as fMRI and text mining as described above.

3. Kernel canonical correlation analysis

As mentioned in the introduction, the kernel canonical correlation analysis (KCCA) is a kernelized version of CCA. This method is used when CCA fails because of the nonlinearity of the data. Because the approach we propose is an extension of KCCA, we review the method in this section.

3.1. Nonregularized kernel canonical correlation analysis

Let us introduce some notation. Let $X \in \mathcal{X}$ be a random vector of the size n_X, which represents the first set of random variables and let $Y \in \mathcal{Y}$ be a random vector of the size n_Y, which represents the second set. If the dependence between X and Y is not linear, the usual CCA may give us only a small correlation coefficient because the method is linear. To avoid that problem, we use the kernel method described in Section 7.

Given positive definite kernels k_X and k_Y, we construct reproducing kernel Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$, also called feature spaces (see Section 7). We map random vectors X and Y into the according feature spaces

$$\begin{aligned} \Phi_X : \mathcal{X} &\longrightarrow \mathcal{H}_X \\ X &\longmapsto \Phi_X(X) = k_X(\cdot, X) \end{aligned} \qquad\qquad \begin{aligned} \Phi_Y : \mathcal{Y} &\longrightarrow \mathcal{H}_Y \\ Y &\longmapsto \Phi_Y(Y) = k_Y(\cdot, Y). \end{aligned}$$

The reproducing property of RKHS (see Section 7) gives us

$$\begin{aligned} \left\langle k_X(\cdot, X), k_X(\cdot, X')\right\rangle_{\mathcal{H}_X} &= \left\langle \Phi_X(X), \Phi_X(X')\right\rangle_{\mathcal{H}_X} = k_X(X, X') \\ \left\langle k_Y(\cdot, Y), k_Y(\cdot, Y')\right\rangle_{\mathcal{H}_Y} &= \left\langle \Phi_Y(Y), \Phi_Y(Y')\right\rangle_{\mathcal{H}_Y} = k_Y(Y, Y'). \end{aligned}$$

The point of the kernel approach is to take $k_X(X, X')$, with respect to $k_Y(Y, Y')$, instead of the scalar product $\langle \Phi_X(X), \Phi_X(X')\rangle_{\mathcal{H}_X}$, with respect to $\langle \Phi_Y(Y), \Phi_Y(Y')\rangle_{\mathcal{H}_Y}$, in feature spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ so the scalar products are calculated implicitly without ever really performing the computations in the high dimensional feature spaces.

The main ingredient of the *kernel canonical correlation analysis* (KCCA) is the *kernel canonical correlation*, defined as follows

$$KCC = \max_{f \in \mathcal{H}_X, g \in \mathcal{H}_Y} \frac{\operatorname{cov}(\langle f, \Phi_X(X)\rangle_{\mathcal{H}_X}, \langle g, \Phi_Y(Y)\rangle_{\mathcal{H}_Y})}{\sqrt{\operatorname{var}(\langle f, \Phi_X(X)\rangle_{\mathcal{H}_X})}\sqrt{\operatorname{var}(\langle g, \Phi_Y(Y)\rangle_{\mathcal{H}_Y})}}.$$

In praxis, the coefficient KCC, defined in terms of population (co)variances, is replaced by its empirical estimate because we only have access to a finite sample.

Let $\{x_1, x_2, \dots, x_N\}$ and $\{y_1, y_2, \dots, y_N\}$ be sets of empirical realizations of random vectors X and Y on the sample of size N. Here $x_i \in \mathbb{R}^{n_X}$ and $y_i \in \mathbb{R}^{n_Y}$ for each $i \in \{1, 2, \dots, N\}$. Let

$\{\Phi_X(x_1), \Phi_X(x_2), \ldots, \Phi_X(x_N)\}$ and $\{\Phi_Y(y_1), \Phi_Y(y_2), \ldots, \Phi_Y(y_N)\}$ denote the corresponding images in feature spaces. Let K_X and K_Y denote the *Gram matrices* corresponding to kernels k_X and k_Y respectively. They are defined as follows

$$K_X[ij] = k_X(x_i, x_j) \text{ for each } i, j \in \{1, 2, \ldots, N\}$$

$$K_Y[ij] = k_Y(y_i, y_j) \text{ for each } i, j \in \{1, 2, \ldots, N\}.$$

We can assume with no loss of generality that the data is centered in the feature spaces

$$\sum_{i=1}^{N} \Phi_X(x_i) = 0, \quad \sum_{j=1}^{N} \Phi_Y(y_j) = 0.$$

In [12] it was shown that under the above assumptions the empirical (co)variances can be expressed as follows

$$\widehat{\text{cov}}(\langle f, \Phi_X(X)\rangle_{\mathcal{H}_X}, \langle g, \Phi_Y(Y)\rangle_{\mathcal{H}_Y}) = \widehat{\text{cov}}\,(f(X), g(Y)) = \frac{1}{N}\alpha^T K_X K_Y \beta.$$

$$\widehat{\text{var}}(\langle \Phi_X(X), f\rangle) = \widehat{\text{var}}(f(X)) = \frac{1}{N}\alpha^T K_X K_X \alpha$$

$$\widehat{\text{var}}(\langle \Phi_Y(Y), g\rangle) = \widehat{\text{var}}(g(Y)) = \frac{1}{N}\beta^T K_Y K_Y \beta.$$

The empirical estimate for kernel canonical correlation (KCC) is therefore equal to

$$\widetilde{KCC} = \max_{\alpha,\beta \in \mathbb{R}^N} \frac{\alpha^T K_X K_Y \beta}{(\alpha^T K_X^2 \alpha)^{\frac{1}{2}} (\beta^T K_Y^2 \beta)^{\frac{1}{2}}}.$$

By solving the maximization problem above with Lagrange multipliers we can transform it into the following generalized eigenvalue problem

$$\begin{bmatrix} 0 & K_X K_Y \\ K_Y K_X & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \rho \begin{bmatrix} K_X^2 & 0 \\ 0 & K_Y^2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

This is exactly the generalized eigenvalue problem of CCA on two vectors of dimension N with covariance matrix $\begin{bmatrix} K_X^2 & K_X K_Y \\ K_Y K_X & K_Y^2 \end{bmatrix}$.

3.2. Regularized kernel canonical correlation analysis

A more useful estimate of population KCC can be obtained via the *regularized KCCA* introduced in [12,13]

$$\max_{f \in \mathcal{H}_X, g \in \mathcal{H}_Y} \frac{\text{cov}(f(X), g(Y))}{\left(\text{var}(f(X)) + \kappa ||f||_{\mathcal{H}_X}^2\right)^{\frac{1}{2}} \left(\text{var}(g(Y)) + \kappa ||g||_{\mathcal{H}_Y}^2\right)^{\frac{1}{2}}}.$$

The additional parameter κ may improve the results if wisely picked. This parameter should be small and positive and should approach zero with an increasing sample size N.

Now let us derive the estimate for the *regularized kernel canonical correlation* defined above. First, we have to expand factors in the denominator

$$\begin{aligned}\mathrm{var}(f(X)) + \kappa||f||^2_{\mathcal{H}_X} &= \frac{1}{N}\alpha^T K_X^2\alpha + \kappa\alpha^T K_X\alpha \approx \frac{1}{N}\alpha^T\left(K_X + \frac{N\kappa}{2}I\right)^2\alpha\\ \mathrm{var}(g(X)) + \kappa||g||^2_{\mathcal{H}_Y} &= \frac{1}{N}\beta^T K_Y^2\beta + \kappa\beta^T K_Y\beta \approx \frac{1}{N}\beta^T\left(K_Y + \frac{N\kappa}{2}I\right)^2\beta.\end{aligned}$$

We use the above approximation to get the empirical estimate for the regularized kernel correlation coefficient

$$\widehat{KCC} = \max_{\alpha,\beta\in\mathbb{R}^N} \frac{\alpha^T K_X K_Y\beta}{\left(\alpha^T\left(K_X + \frac{N\kappa}{2}I\right)^2\alpha\right)^{\frac{1}{2}}\left(\beta^T\left(K_Y + \frac{N\kappa}{2}I\right)^2\beta\right)^{\frac{1}{2}}}.$$

Again, we can easily transform the corresponding problem into the following generalized eigenvalue problem

$$\begin{bmatrix} 0 & K_XK_Y \\ K_YK_X & 0\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \rho\begin{bmatrix}\left(K_X + \frac{N\kappa}{2}I\right)^2 & 0\\ 0 & \left(K_Y + \frac{N\kappa}{2}I\right)^2\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}.$$

This is exactly the generalized eigenvalue problem of CCA on two vectors with a known covariance matrix. Algorithms for solving this eigenvalue problem (incomplete Cholesky decomposition, partial Gram–Schmidt, etc.) and their computational complexities are described in [12,14].

4. Restricted kernel canonical correlation analysis

Das and Sen [4] have introduced a restricted canonical correlation analysis. We use their idea and propose a similar approach for the kernelized version of CCA. We call it a *restricted kernel canonical correlation analysis* (RKCCA). We consider here only the empirical RKCCA. We are searching for an estimate of the KCC under non-negativity restriction.

Let us introduce some notation

$$\begin{aligned}P &= K_XK_Y\\ Q &= \left(K_X + \frac{N\kappa}{2}I\right)^2\\ R &= \left(K_Y + \frac{N\kappa}{2}I\right)^2.\end{aligned}$$

The empirical KCCA can now be written as follows

$$\max_{\alpha\in\mathbb{R}^N,\beta\in\mathbb{R}^N}\frac{\alpha^T P\beta}{\sqrt{\alpha^T Q\alpha\beta^T R\beta}}.$$

Let $\mathbb{R}^N_+$ be the space of all vectors $\xi \in \mathbb{R}^N$ for which $\xi_i \geq 0$ for all $i = 1, 2, \ldots, N$. In the RKCCA, we are searching for such vectors $a = (a_1, a_2, \ldots, a_N) \in \mathbb{R}^N_+$ and $b = (b_1, b_2, \ldots, b_N) \in \mathbb{R}^N_+$ with all non-negative components for which the following holds

$$\frac{a^T Pb}{\sqrt{a^T Qa}\sqrt{b^T Rb}} = \max_{\alpha\in\mathbb{R}^N_+,\beta\in\mathbb{R}^N_+}\frac{\alpha^T P\beta}{\sqrt{\alpha^T Q\alpha}\sqrt{\beta^T R\beta}}. \tag{1}$$

Because the quotient in (1) does not change if we multiply either α or β by a positive constant, we can restrict ourselves to the vectors α and β satisfying $\sqrt{\alpha^T Q\alpha} = \sqrt{\beta^T R\beta} = 1$. The set of vectors satisfying all these conditions is compact, and the function we are maximizing is a continuous function on this set. Consequently, it is clear that its supremum is attained. So the maximum in (1) always exists.

Let us introduce new notation

$$z = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \in \mathbb{R}^{2N}, \quad P_\star = \begin{bmatrix} 0 & \frac{1}{2}P \\ \frac{1}{2}P^T & 0 \end{bmatrix}$$
$$Q_\star = \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix}, \quad R_\star = \begin{bmatrix} 0 & 0 \\ 0 & R \end{bmatrix}.$$

We call the following estimate a *restricted kernel canonical correlation* (RKCC)

$$\text{RKCC} = \sup_{z \in \mathbb{R}_+^{2N}} \frac{z^T P_\star z}{\sqrt{z^T Q_\star z z^T R_\star z}}. \tag{2}$$

We have proved that the supremum in the case of non-negativity restriction is always attained. On the other side, in the related problem of positivity restriction, a maximum may not exist. Let $\mathbb{R}_{0+}^N$ be the space of all such vectors ξ for which the condition $\xi_i > 0$ for all $i \in \{1, 2, \ldots, N\}$ holds. We define a kernel canonical correlation under positivity restriction as

$$\text{RKCC}^+ = \sup_{z \in \mathbb{R}_{0+}^{2N}} \frac{z^T P_\star z}{\sqrt{z^T Q_\star z z^T R_\star z}}. \tag{3}$$

When the maximum in (3) exists, it has to be the same as the RKCC, which is the maximum in (2).

5. Karush–Kuhn–Tucker

In this section, we use the Karush–Kuhn–Tucker theorem to show that the RKCC equals an unconstrained solution to a modified CCA problem on two random vectors with known covariance matrix where one or several variables have been excluded.

Theorem 1. *The optimal solution of the maximization problem (2), that is,* RKCC $= \rho$, *must satisfy*

$$\begin{aligned} & P\beta - \rho Q\alpha + \Lambda_1 = 0; \quad \Lambda_1 = (\lambda_1, \lambda_2, \ldots, \lambda_N)^T \\ & P^T\alpha - \rho R\beta + \Lambda_2 = 0; \quad \Lambda_2 = (\lambda_{N+1}, \lambda_{N+2}, \ldots, \lambda_{2N})^T \\ & \alpha_i \geq 0, \quad i = 1, 2, \ldots, N \\ & \beta_j \geq 0, \quad j = 1, 2, \ldots, N \\ & \lambda_i \geq 0, \quad i = 1, 2, \ldots, 2N + 4 \\ & \lambda_i \alpha_i = 0 = \lambda_{N+j}\beta_j, \quad i = 1, 2, \ldots, N, \; j = 1, 2, \ldots, N \\ & \alpha^T Q\alpha = 1 = \beta^T R\beta. \end{aligned} \tag{4}$$

Proof. The Karush–Kuhn–Tucker theorem [15] states that an optimal solution of the problem

$$\begin{aligned} & \max \quad f(x) \\ & \text{subject to} \quad g_i(x) \leq b_i \text{ for all } i = 1, 2, \ldots, m \quad x \in \mathbb{R}^n, \end{aligned}$$

where $f(x)$ is a differentiable function for which Abadie's constraint qualification holds, must satisfy

$$\begin{aligned} \nabla f(x) - \sum_{i=1}^{m} \lambda_i \nabla g_i(x) = 0, & \\ g_i(x) \leq b_i, & \quad i = 1, 2, \ldots, m, \\ \lambda_i \geq 0, & \quad i = 1, 2, \ldots, m, \\ \lambda_i[g_i(x) - b_i] = 0, & \quad i = 1, 2, \ldots, m. \end{aligned}$$

Let us first rewrite the optimization problem of RKCCA in such a way that we will be able to use the Karush–Kuhn–Tucker theorem directly

$$\begin{aligned} \max \quad & f(z) = z^T P^\star z \\ \text{subject to} \quad & g_i(z) = -z_i \leq 0 \text{ for all } i = 1, 2, \ldots, 2N \\ & g_{2N+1} = z^T Q^\star z \leq 1 \\ & g_{2N+2} = -z^T Q^\star z \leq -1 \\ & g_{2N+3} = z^T R^\star z \leq 1 \\ & g_{2N+4} = -z^T R^\star z \leq -1. \end{aligned}$$

Because in this case all gradients $\nabla g_i(z),\ i = 1, 2, \ldots, 2N + 4$ are linearly independent, the linear independence constraint qualification holds. This implies that the Abadi'e constraint qualification also holds [15]. Therefore, the statement we are proving follows directly from the Karush–Kuhn–Tucker theorem, stated above. □

In the case of positivity restriction it holds that $\lambda_i = 0$, for all $i = 1, 2, \ldots, 2N$ in (4). Therefore the solution $\text{RKCC}^+ = \rho^+$ (if it exists) must satisfy

$$\begin{aligned} & P\beta - \rho^+ Q\alpha = 0 \\ & P^T\alpha - \rho^+ R\beta = 0. \end{aligned} \tag{5}$$

So, we have transformed the problem to the generalized eigenvalue problem

$$\begin{bmatrix} 0 & P \\ P^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \rho^+ \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

equal to one of the regularized KCCA. As we have shown before this is also equal to the CCA on two random vectors with covariance matrix

$$\Sigma = \begin{bmatrix} Q & P \\ P^T & R \end{bmatrix}. \tag{6}$$

Thus, if a solution to the optimization problem with positivity restriction exists, it has to be equal to one of the canonical correlations on two random vectors of dimension N with covariance matrix Σ (6).

The only way the maximum in the case of positivity restriction (3) may not exist is if some of the optimal coefficients related to the supremum are equal to zero. This gives us a way to characterize the RKCCA with sub-vectors and sub-matrices.

We need some new notation

$$\begin{aligned} & [n] = \{1, 2, \ldots, n\} \text{ for all } n \in \mathbb{N} \\ & I_n = \{\mathbf{a} : \varnothing \neq \mathbf{a} \subseteq [n],\ \text{with elements in } \mathbf{a} \text{ written in natural order}\} \\ & |\mathbf{a}| \ldots \text{cardinality of } \mathbf{a}. \end{aligned}$$

For a vector $x \in \mathbb{R}^N$ and a set of indices $\mathbf{a} \in I_N$, let the sub-vector of x consisting of those components of x whose indices belong to $\mathbf{a}$ be denoted by x_a. Similarly, we introduce a notation $S_{a:b}$ for the sub-matrix of an $N \times N$ matrix S consisting of those rows with indices that are in $\mathbf{a} \in I_N$ and those columns with indices that are in $\mathbf{b} \in I_N$.

The RKCC can be written in terms of sub-vectors and sub-matrices. As was shown in the previous section, the regularized KCCA can be viewed as the CCA on two random vectors with covariance matrix

Σ (6). Let us denote those two vectors by $X^{(1)}$ and $X^{(2)}$. Now we can write the RKCC as

$$\begin{aligned}\mathrm{RKCC} &= \max_{\alpha,\beta\in\mathbb{R}_N^+} \operatorname{corr}\left(\alpha^T X^{(1)}, \beta^T X^{(2)}\right)\\ &= \max_{\mathbf{a},\mathbf{b}\in I_N} \sup_{\alpha\in\mathbb{R}_{0+}^{|\mathbf{a}|},\ \beta\in\mathbb{R}_{0+}^{|\mathbf{b}|}} \operatorname{corr}\left(\alpha^T X_a^{(1)}, \beta^T X_b^{(2)}\right)\\ &= \max_{\mathbf{a},\mathbf{b}\in I_N} \widetilde{\max}_{\alpha\in\mathbb{R}_{0+}^{|\mathbf{a}|},\ \beta\in\mathbb{R}_{0+}^{|\mathbf{b}|}} \operatorname{corr}\left(\alpha^T X_a^{(1)}, \beta^T X_b^{(2)}\right).\end{aligned}$$

Here $\widetilde{\max}$ stands for the maximum when it exists. If it does not exist, we ignore the corresponding subset of indices. So, $\widetilde{\max}$ is really the maximum of (3) for matrices $P_{a:b}$, $Q_{a:a}$, $R_{b:b}$. When this maximum exists, the maximal correlation ρ^+ has to satisfy (5), which in our case equals to

$$\begin{aligned}&P_{a:b}\beta - \rho^+ Q_{a:a}\alpha = 0\\ &P_{a:b}^T\alpha - \rho^+ R_{b:b}\beta = 0.\end{aligned}$$

Denote by $cc^2(P, Q, R)$ the set of squared canonical correlations corresponding to the covariance matrix $\begin{bmatrix} Q & P \\ P^T & R \end{bmatrix}$. The considerations above can now be rewritten into

$$\mathrm{RKCC}^2 \in \bigcup_{\mathbf{a},\mathbf{b}\in I_N} \left\{cc^2(P_{a:b}, Q_{a:a}, R_{b:b})\right\}.$$

The squared RKCC equals one of the ordinary squared canonical correlations between sub-vectors $X_a^{(1)}$ and $X_b^{(2)}$ for some sets of indices $\mathbf{a}, \mathbf{b} \in I_N$. The statement here involves RKCC^2 and not RKCC because Eq. (5) can be solved only for ρ^{+2}.

To find the RKCC, one must solve the ordinary CCA problem for all possible sub-matrices $Q_{a:a}$, $P_{a:b}$, $R_{b:b}$ and pick the largest correlation for which the regression weights fulfill the non-negativity constraints. For N-dimensional input variables, there are $(2^N - 1)^2$ such problems to solve. Hence, the solution of the RKCCA has an unpleasant property of growing exponentially with the size of the sample. Two important properties of the proposed approach are that we are guaranteed to find the global optimum and that we can find this optimum algebraically, i.e., no iterative numerical search is required.

6. Generalized eigenvalue problem

It was already shown in Section 4 that a maximum in the following equation

$$\mathrm{RKCC} = \rho = \frac{a^T P b}{\sqrt{a^T Q a}\sqrt{b^T R b}} = \max_{\alpha\in\mathbb{R}_+^N,\beta\in\mathbb{R}_+^N} \frac{\alpha^T P \beta}{\sqrt{\alpha^T Q \alpha}\sqrt{\beta^T R \beta}} \tag{7}$$

always exists. Here, we translate this maximization into an optimization problem related to eigenvalues of some generalized eigenvalue problem.

Let us introduce new notation

$$A = \begin{bmatrix} 0 & P \\ P^T & 0 \end{bmatrix} \in \mathbb{R}^{2N\times 2N}, \quad B = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \in \mathbb{R}^{2N\times 2N}, \quad z = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}_+^{2N}.$$

It is easy to show (see [5]) that if a and b are vectors of coefficients satisfying (7), it must hold for z that

$$\rho = \frac{z^T A z}{z^T B z} = \max_{w\in\mathbb{R}_+^{2N}} \frac{w^T A w}{w^T B w}. \tag{8}$$

With no loss of generality, we restrict ourselves to vectors $z \in \mathbb{R}_+^{2N}$ satisfying the additional condition $z^T Bz = 1$. Because the set of vectors satisfying these conditions is compact and the function we are maximizing is a continuous function, the maximum in (8) always exists. It has to be the same as the maximum in (7), which is the RKCC.

We will use the notation of sub-vectors and sub-matrices from Section 5. Let $z \in \mathbb{R}_+^{2N}$ be a solution of the maximization problem (8), and let $\mathbf{a} \in I_{2N}$ be the set of indices for which $z_i > 0$. It is clear that in this case the vector $z_a \in \mathbb{R}_{0+}^{|\mathbf{a}|}$ is a solution of the following maximization problem

$$\frac{z_a^T A_{a:a} z_a}{z_a^T B_{a:a} z_a} = \max_{w_a \in \mathbb{R}_+^{|\mathbf{a}|}} \frac{w_a^T A_{a:a} w_a}{w_a^T B_{a:a} w_a}. \tag{9}$$

Theorem 2. *Let $z \in \mathbb{R}_+^{2N}$ and $z_a \in \mathbb{R}_{0+}^{|\boldsymbol{a}|}$ be as defined above. Then the RKCC, denoted by ρ, equals the maximal generalized eigenvalue of the generalized eigenvalue problem $A_{a:a} z_a = \rho B_{a:a} z_a$ with z_a equal to the corresponding eigenvector.*

Proof. Recall

$$A = \begin{bmatrix} 0 & P \\ P^T & 0 \end{bmatrix} = \begin{bmatrix} 0 & K_X K_Y \\ K_X K_Y^T & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} = \begin{bmatrix} \left(K_X + \frac{N\kappa}{2} I\right)^2 & 0 \\ 0 & \left(K_Y + \frac{N\kappa}{2} I\right)^2 \end{bmatrix}. \tag{10}$$

Because K_X and K_Y are positive definite kernel matrices, the matrix A (therefore also $A_{a:a}$) is a real symmetric matrix and the matrix B (therefore also $B_{a:a}$) is a positive definite matrix. So, we have a similar situation as Omladič and Omladič in [5], and we can use their idea.

First, let us note that eigenvalues in the generalized eigenvalue problem $A_{a:a} z_a = \rho B_{a:a} z_a$ are the same as eigenvalues in the spectral decomposition of the following symmetric matrix

$$B_{a:a}^{-1/2} A_{a:a} B_{a:a}^{-1/2} = \sum_r \rho_r P_r. \tag{11}$$

Here eigenvalues are denoted by ρ_r (and indexed in decreasing order) and the corresponding spectral idempotents (which are symmetric and have the total sum equal to identity matrix I) are denoted by P_r. If we denote the eigenvector corresponding to some eigenvalue of the matrix $B_{a:a}^{-1/2} A_{a:a} B_{a:a}^{-1/2}$ by x_a (and assume with no loss of generality that $x_a^T x_a = 1$), the relation between both eigenvectors is

$$z_a = B_{a:a}^{-1/2} x_a.$$

Therefore, it is enough to consider the maximal eigenvalue ρ_0 of the matrix $B_{a:a}^{-1/2} A_{a:a} B_{a:a}^{-1/2}$.

Notice that (11) implies

$$\rho = \frac{z_a^T A_{a:a} z_a}{z_a^T B_{a:a} z_a} = \sum_r \rho_r \frac{z_a^T B_{a:a}^{1/2} P_r B_{a:a}^{1/2} z_a}{z_a^T B_{a:a} z_a}$$

and that this is a convex combination of eigenvalues ρ_r of matrix $B_{a:a}^{-1/2} A_{a:a} B_{a:a}^{-1/2}$. Thus, the RKCC $= \rho$ in not greater than the maximal eigenvalue ρ_0.

To finish the proof, we have to show that $\rho_0 \leq \rho$, so that $\rho_0 = \rho$. Denote $\gamma = z_a^T B_{a:a} z_a$ and introduce

$$y_a = z_a \cos\varphi + B_{a:a}^{-1/2} x_a \sin\varphi.$$

It is clear that if φ is close enough to zero, y_a will be close to z_a and it will have strictly positive entries on the set **a**. From (6) we get

$$\begin{aligned} y_a^T B_{a:a} y_a &= \gamma \cos^2\varphi + 2 z_a^T B_{a:a}^{1/2} x_a \cos\varphi \sin\varphi + \sin^2\varphi \\ y_a^T A_{a:a} y_a &= \rho\gamma \cos^2\varphi + 2 z_a^T B_{a:a}^{1/2} B_{a_a}^{-1/2} A_{a:a} B_{a:a}^{-1/2} x_a \cos\varphi \sin\varphi \\ &\quad + x_a^T B_{a:a}^{-1/2} A_{a:a} B_{a:a}^{-1/2} x_a \sin^2\varphi \\ &= \rho\gamma \cos^2\varphi + 2\rho_0 z_a^T B_{a:a}^{1/2} x_a \cos\varphi \sin\varphi + \rho_0 \sin^2\varphi. \end{aligned}$$

So if $\rho_0 > \rho$, we can choose such $\varphi \in [0, \frac{\pi}{2}]$ close to zero that y_a will have strictly positive entries on **a** and that the quotient $\frac{y_a^T A_{a:a} y_a}{y_a^T B_{a:a} y_a}$ will be strictly greater than ρ. This contradicts the local maximality of ρ; hence, ρ_0 cannot be strictly greater than ρ. Thus, they are equal. □

Theorem 3. *Under the above assumptions, it holds that* $(A - \rho B)z \leq 0$ *and the set of indices on which this vector is strictly negative is disjoint with the set* ***a****.*

Proof. It is enough to consider the case when the set **a** contains all but one index. So let $\mathbf{a} = \{1, 2, \dots, 2N-1\}$ and write

$$A = \begin{bmatrix} A_{a:a} & c \\ c^T & \delta \end{bmatrix}, \quad B = \begin{bmatrix} B_{a:a} & d \\ d^T & \epsilon \end{bmatrix} \quad \text{and} \quad z = \begin{bmatrix} z_a \\ 0 \end{bmatrix}.$$

The previous theorem implies that

$$(A - \rho B)z = \begin{bmatrix} 0 \\ (c - \rho d)^T z_a \end{bmatrix}.$$

Thus, the set of indices where vector $(A - \rho B)z$ is strictly negative is disjoint with the set **a**.

We have proved the second part of the statement. We still have to prove that $(c - \rho d)^T z_a \leq 0$. With no loss of generality we assume that $z^T B z = 1$ and we define a vector

$$x = \begin{bmatrix} z_a \cos\varphi \\ \sin\varphi \end{bmatrix}$$

for any $\varphi \in [0, \pi/2]$. Clearly it holds that $x \geq 0$ and

$$\begin{aligned} x^T B x &= \cos^2\varphi + 2 d^T z_a \cos\varphi \sin\varphi + \epsilon \sin^2\varphi \\ x^T A x &= \rho \cos^2\varphi + 2 c^T z_a \cos\varphi \sin\varphi + \delta \sin^2\varphi. \end{aligned}$$

It is easy to see that the first derivative of $\frac{x^T A x}{x^T B x}$ as a function of φ at $\varphi = 0$ is equal to $2(c - \rho d)^T z_a$. So if $(c - \rho d)^T z_a$ were strictly positive, the quotient $\frac{x^T A x}{x^T B x}$ which equals ρ at $\varphi = 0$ would be strictly increasing as a function of φ, contradicting the maximality of ρ. Thus, it must hold that $(c - \rho d)^T z_a \leq 0$. □

Using these theorems we can solve the statistical problem of RKCCA with the following search process:

- For every set of indices $\mathbf{a} \in I_{2N}$ find the maximal generalized eigenvalue ρ and the corresponding eigenvector z_a of the generalized eigenvalue problem $A_{a:a} z_a = \rho B_{a:a} z_a$ for sub-matrices $A_{a:a}$ and $B_{a:a}$.
- Consider the solutions $z \in \mathbb{R}_+^{2N}$ such that $z_a \in \mathbb{R}_{0+}^{|\mathbf{a}|}$ and $(A - \rho B)z \leq 0$.
- Choose the maximal among the obtained eigenvalues to get the global solution of the optimization problem.

The RKCC $= \rho$ is the maximal generalized eigenvalue that corresponds to the optimal set of indices $\mathbf{a} \in I_{2N}$.

It is possible that the complexity of the proposed algorithm can be reduced using one of the standard optimization techniques. We leave this question for further research.

7. Theory of reproducing kernels and kernel approach

For those readers who are not familiar with the theory of reproducing kernels, here we present some basic notions. We will use [17] to define reproducing kernel and point out some of its properties that are important for this paper.

Consider a linear class $\mathcal{F}$ of functions $f(x)$ defined in a set E. We shall suppose that $\mathcal{F}$ is a real class, that is, it admits multiplication by real constants.

Suppose further that for $f \in \mathcal{F}$ a norm $||f||$ is defined given by an ordinary quadratic form $Q(f)$

$$||f||^2 = Q(f).$$

Here functional $Q(f)$ is called ordinary quadratic if for real ξ_1, ξ_2 and for $f_1, f_2 \in \mathcal{F}$ it holds that $Q(\xi_1 f_1 + \xi_2 f_2) = \xi_1^2 Q(f_1) + 2\xi_1\xi_2 Q(f_1, f_2) + \xi_2^2 Q(f_2)$, where $Q(f_1, f_2)$ is the corresponding bilinear form. This bilinear form will be denoted by

$$\langle f_1, f_2 \rangle = Q(f_1, f_2)$$

and called the scalar product corresponding to the norm $||f||$. Clearly

$$||f||^2 = \langle f, f \rangle.$$

The class $\mathcal{F}$ with the norm, $||\quad||$, forms a normed real vector space. If this space is complete, it is a Hilbert space.

Let $\mathcal{F}$ be a class of functions defined in a set E, forming a real Hilbert space. The function $k(x, y)$ of $x, y \in E$ is called a *reproducing kernel* (RK) of $\mathcal{F}$, if

(1) For every y, $k(x, y)$ as a function of $x \in E$ belongs to $\mathcal{F}$.
(2) The *reproducing* property: for every $y \in E$ and every $f \in \mathcal{F}$

$$f(y) = \langle f(\cdot), k(\cdot, y) \rangle.$$

Theorem 4 (Properties of reproducing kernels)**.**

(1) If a reproducing kernel exists, it is unique.
(2) For the existence of a reproducing kernel $k(x, y)$ it is necessary and sufficient that for every y of the set E, $f(y)$ be continuous functional of f running through the Hilbert space $\mathcal{F}$.
(3) $k(x, y)$ is a positive semidefinite kernel, that is, the quadratic form in $\xi_1, \xi_2, \ldots, \xi_n$

$$\sum_{i,j=1}^{n} k(y_i, y_j)\xi_i\xi_j$$

is non-negative for all $y_1, y_2, \ldots, y_n$ in E.
(4) For every positive semidefinite kernel $k(x, y)$ there corresponds one and only one class of functions with a uniquely determined quadratic form that creates a Hilbert space and admits $k(x, y)$ as a RK. This Hilbert space is called reproducing kernel Hilbert space (RKHS).

Proof. See [17]. □

Let us now connect this theory with the kernel approach commonly used in machine learning. We start with a positive definite kernel k and a random vector $X \in \mathcal{X}$. By the above theorem, we are able to construct a RKHS. We denote it by $\mathcal{H}$. Furthermore we denote by Φ the following map

$$\Phi : X \in \mathcal{X} \longrightarrow k(\cdot, X) \in \mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}.$$

Here $\mathbb{R}^{\mathcal{X}}$ denote the space of functions from $\mathcal{X}$ to $\mathbb{R}$. In machine learning the RKHS $\mathcal{H}$ is called a *feature space* and the map Φ is called a *feature map*.

It was shown in [16] that any function $f \in \mathcal{H}$ can be written as

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, X_i)$$

for an arbitrary $n \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$ and $X_i \in \mathcal{X}, i \in \{1, 2, \ldots, n\}$. The scalar product of a function $f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, X_i) \in \mathcal{H}$ with a function $g(\cdot) = \sum_{j=1}^{n} \beta_j k(\cdot, X_j) \in \mathcal{H}$ is given by

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \beta_j k(X_i, X_j).$$

The reproducing property is now written as

$$\langle k(\cdot, X), f \rangle = f(X) \text{ for all } f \in \mathcal{H}$$

or in a special case

$$\left\langle \Phi(X), \Phi(X') \right\rangle = k(X, X')$$

for each pair $(X, X') \in \mathcal{X} \times \mathcal{X}$.

8. Discussion

The main idea of sub-vectors and sub-matrices could also be used to derive restricted kernel PCA and some other restricted kernel methods.

Similarly, we could solve the problem of partially restricted KCCA where only some coefficients are restricted to be non-negative and the others are without restrictions.

Acknowledgements

The author is thankful to M. Omladič for the idea of combining the kernel method with the restricted canonical correlation analysis and to J. Shawe-Taylor for pointing out the importance of non-negativity in applications of kernel methods to machine learning.

Research on this work was supported by a fellowship from the Agency for Research and Development (ARRS) of the Republic of Slovenia.

References

[1] H. Hoteling, Relations between two sets of variates, Biometrika 28 (1936) 321–377.

[2] K. Fukumizu, F.R. Bach, A. Gretton, Statistical consistency of kernel canonical correlation analysis, J. Mach. Learning Res. 8 (2007) 361–383.

[3] B. Schoelkopf, A.J. Smola, Learning with Kernels, The MIT Press, Cambridge, Massachusetts, London, England, 2002.

[4] S. Das, P.K. Sen, Restricted canonical correlations, Linear Algebra Appl. 210 (1994) 29–47.

[5] M. Omladič, V. Omladič, More on restricted canonical correlations, Linear Algebra Appl. 321 (2000) 285–293.

[6] Y. Li, J. Shawe-Tylor, Using KCCA for Japanese–English cross-language information retrieval, J. Intell. Inform. Syst. 2 (2006) 117–133.

[7] A. Vinokourov, D.R. Hardoon, J. Shawe-Tylor, Learning the semantics of multimedia content with application to web image retrieval and classification, in: Fourth International Symposium on Independent Component Analysis and Blind Source Separation, 2003.

[8] D. Hardoon, J. Mourao-Miranda, M. Brammer, J. Shawe-Taylor, Unsupervised analysis of fMRI data using kernel canonical correlation, NeuroImage 37 (2007) 1250–1259.

[9] O. Friman, J. Carlsson, P. Lundberg, et al., Detection of neural activity in fMRI using canonical correlation analysis, Magn. Reson. Med. 45 (2001) 323–330.

[10] O. Friman, M. Borga, P. Lundberg, H. Knutsson, Adaptive analysis of fMRI data, NeuroImage 19 (2003) 837–845.

[11] M. Ragnehed, M. Engstrďżm̈, H. Knutsson, B. Soederfeldt, P. Lundberg, Restricted canonical correlation analysis in functional MRI-validation and a novel thresholding technique, J. Magn. Reson. Imaging 29 (2009) 146–154.

[12] F.R. Bach, M.I. Jordan, Kernel independent component analysis, J. Mach. Learning Res. 3 (2002) 1–48.

[13] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, B. Schökopf, Kernel methods for measuring independence, J. Mach. Learning Res. 6 (2005) 2075–2129.

[14] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (2004) 2639–2664.

[15] N. Andreasson, A. Evgrafov, M. Patriksson, An Introduction to Continuous Optimization, Studentlitteratur AB, Lund, Sweeden, 2005.

[16] T. Hofmann, B. Schölkopf, A.J. Smola, Kernel methods in machine learning, Ann. Stat. 3 (2008) 1171–1220.

[17] N. Aronszajn, Theory of reproducing kernels, Trans. Am. Math. Soc. 68 (1950) 337–404.