Mach Learn (2009) 74: 23–38
DOI 10.1007/s10994-008-5085-3

Convergence analysis of kernel Canonical Correlation Analysis: theory and practice

David R. Hardoon · John Shawe-Taylor

Received: 17 October 2007 / Revised: 4 March 2008 / Accepted: 3 September 2008 /
Published online: 8 October 2008
Springer Science+Business Media, LLC 2008

Abstract Canonical Correlation Analysis is a technique for finding pairs of basis vectors that maximise the correlation of a set of paired variables, these pairs can be considered as two views of the same object. This paper provides a convergence analysis of Canonical Correlation Analysis by defining a pattern function that captures the degree to which the features from the two views are similar. We analyse the convergence using Rademacher complexity, hence deriving the error bound for new data. The analysis provides further justification for the regularisation of kernel Canonical Correlation Analysis and is corroborated by experiments on real world data.

Keywords Canonical Correlation Analysis · Rademacher complexity · Kernel methods

1 Introduction

Proposed by H. Hotelling in 1936, Canonical Correlation Analysis (CCA) is a technique for finding pairs of basis vectors that maximise the correlation between the projections of the paired variables onto the corresponding basis vectors. Correlation is dependent on the chosen coordinate system, therefore even if there is a very strong linear relationship between two sets of multidimensional variables this relationship might not be visible as a correlation. CCA seeks a pair of linear transformations one for each of the pairs of variables such that when the variables are transformed the corresponding coordinates are maximally correlated. Kernel Canonical Correlation Analysis (KCCA) performs this analysis in a kernel defined feature space. First introduced by Fyfe and Lai (2000) and later by Akaho (2001)

Editor: Tony Jebara.

D.R. Hardoon (✉) · J. Shawe-Taylor
Centre for Computational Statistics and Machine Learning, Department of Computer Science,
University College London, Gower St., London WC1E 6BT, UK
e-mail: D.Hardoon@cs.ucl.ac.uk

J. Shawe-Taylor
e-mail: jst@cs.ucl.ac.uk

and Bach and Jordan (2002). KCCA has shown its potential usage in multimedia applications with emphasis on information retrieval. These applications include cross-language text retrieval (Vinokourov et al. 2002) where documents in one language are retrieved via a query from another language, as well as webpage classification (Vinokourov et al. 2003), in which different elements of webpage are used as a complex label structure. More recently content-based image retrieval (Hardoon et al. 2006) has retrieved images from a text query without reference to their original labelling. One can find further studies applying this technique such as those by Friman et al. (2003) where CCA was applied to functional magnetic resonance imaging analysis, and the more recent Hardoon et al. (2007). It has also been applied in independent component analysis (Bach and Jordan 2002) and blind signal separation (Fyfe and Lai 2000). A review of the method is given by Ketterling (1971).

CCA and KCCA solve the problem of finding a canonical correlation between two sets of variables. In this paper we consider the paired variables as two views of the same object, as the technique is applicable in cases where we hypothesise that both views individually contain all the relevant information. In such situations KCCA can identify the relevant subspaces in both views, projecting out irrelevant specifics from both views. For this reason we also refer to the projection space as the semantic space. We show that the empirical estimate of the correlation coefficient is a good estimate of the population correlation by using large deviation bounds.

In previous work (Hardoon et al. 2004) we show that using kernel CCA with no regularisation will be likely to produce perfect correlations between the two views. These correlations can therefore fail to distinguish between spurious features and those that capture the underlying semantics. Similarly, other studies have also dealt with these issues providing justification for regularisation (Bach and Jordan 2002; Kuss and Graepel 2002). Recently, Fukumizu et al. (2006) investigated the general problem of establishing a consistency of KCCA by providing rates for the regularisation parameter. However, as highlighted in the concluding remarks of Fukumizu et al. (2006), the practical problem of choosing the regularisation coefficient in practice remains largely unsolved.

Despite CCA's long history we have found no finite sample statistical analysis of the technique. An initial analysis and theoretical bound was given in Shawe-Taylor and Cristianini (2004), which was later corrected in Hardoon (2006). In this paper we provide a detailed theoretical analysis of KCCA and propose a finite sample statistical analysis of KCCA by using a regression formulation similar to the Alternating Conditional Expectations (ACE) method (Breiman and Friedman 1985). We show this to be tighter than the previously computed bound in Hardoon (2006) and show, through a feasibility experiment, that the derived bound can be used in practice to select the regularisation coefficient. This analysis aims to provide a better understanding of the technique's convergence by using Rademacher complexity to obtain an error bound for a new data sample. We find that the theoretical analysis provides a further justification for the regularisation of kernel CCA as previously proposed by Bach and Jordan (2002) but indicates that an a-posteriori normalisation of the features should be used (detailed in Sect. 4).

The paper is divided as follows, in Sect. 2 we give some background results and in Sect. 3 we briefly review the Canonical Correlation Analysis method. The crux of the matter and novelty of the paper is given in Sect. 4 where we develop the required mathematical machinery and derive the CCA generalisation bound. In Sect. 5 we describe a real-world feasibility experiment verifying the developed theory. Finally, we bring forward our concluding remarks in Sect. 6.

2 Background results

We begin by giving the definition of Rademacher complexity. Assume an underlying distribution $\mathcal{D}$ generating random vectors. We will frequently be considering estimating aspects of this distribution from a random sample S generated identically and independently (i.i.d.) by $\mathcal{D}$.

If $\mathcal{D}$ generates a random object x and

$$S = \{x_1, \dots, x_\ell\}$$

is a sample generated i.i.d. according to $\mathcal{D}$, we denote with $\mathbb{E}[f(x)] = \mathbb{E}_{\mathcal{D}}[f(x)]$ the true expectation of the function $f(x)$ and with $\hat{\mathbb{E}}[f(x)]$ we denote the empirical expectation of $f(x)$, where

$$\hat{\mathbb{E}}[f(x)] = \frac{1}{\ell}\sum_{i=1}^{\ell} f(x_i).$$

Similarly we will use $\mathbb{E}_\sigma$ to denote expectation w.r.t. a random vector σ and $\mathbb{E}_S$ to denote expectation over the generation of the random i.i.d. sample S.

Definition 1 (Rademacher complexity) For a sample $S = \{x_1, \dots, x_\ell\}$ generated by a distribution $\mathcal{D}$ on a set X and a real-valued function class $\mathcal{F}$ with domain X, the *empirical Rademacher complexity* of $\mathcal{F}$ is the random variable

$$\hat{R}_\ell(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{l}\sum_{i=1}^{\ell}\sigma_i f(x_i)\right| : x_1, \dots, x_\ell\right],$$

where $\sigma = (\sigma_1, \dots, \sigma_\ell)$ are independent uniform $\{\pm 1\}$-valued (Rademacher) random variables. The *Rademacher complexity* of $\mathcal{F}$ is

$$R_\ell(\mathcal{F}) = \mathbb{E}_S[\hat{R}_\ell(\mathcal{F})] = \mathbb{E}_{S\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f(x_i)\right|\right].$$

The main application of Rademacher complexity is given in the following theorem (Bartlett and Mendelson 2002) quoted in the form given in Shawe-Taylor and Cristianini (2004).

Theorem 2 *Fix* $\delta \in (0, 1)$ *and let* $\mathcal{F}$ *be a class of functions mapping from* Z *to* $[0, 1]$. *Let* $(z_i)_{i=1}^{\ell}$ *be drawn independently according to a probability distribution* $\mathcal{D}$. *Then with probability at least* $1 - \delta$ *over random draws of samples of size* ℓ, *every* $f \in \mathcal{F}$ *satisfies*

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[f(z)] &\le \hat{\mathbb{E}}[f(z)] + R_\ell(\mathcal{F}) + \sqrt{\frac{\ln(\frac{2}{\delta})}{2\ell}}\\ &\le \hat{\mathbb{E}}[f(z)] + \hat{R}_\ell(\mathcal{F}) + 3\sqrt{\frac{\ln(\frac{2}{\delta})}{2\ell}}.\end{aligned}$$

Definition 3 $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the Euclidean inner product of the vectors $\mathbf{x}$, $\mathbf{y}$ also written $\mathbf{x}'\mathbf{y}$.

A kernel is a function κ, such that for all $x, z \in X$

$$\kappa(x, z) = \langle \phi(x), \phi(z) \rangle \tag{1}$$

where ϕ is a mapping from X to a feature space F

$$\phi : X \to F.$$

The application of Rademacher complexity bounds to kernel defined function classes is well documented (Bartlett and Mendelson 2002). The function class considered is

$$\mathcal{F}_B = \left\{ x \mapsto \langle \mathbf{w}, \phi(x) \rangle : \|\mathbf{w}\| \le B \right\},$$

where $\phi \colon x \mapsto \phi(x)$ is the feature space mapping corresponding to the kernel function in (1); We quote the relevant theorem.

Theorem 4 (Bartlett and Mendelson 2002) *If $\kappa : X \times X \to \mathbb{R}$ is a kernel, and $S = \{x_1, \ldots, x_\ell\}$ is a sample of points from X, then the empirical Rademacher complexity of the class $\mathcal{F}_B$ satisfies*

$$\hat{R}_\ell(\mathcal{F}_B) \le \frac{2B}{\ell} \sqrt{\sum_{i=1}^{\ell} \kappa(x_i, x_i)} = \frac{2B}{\ell} \sqrt{\operatorname{tr}(K)},$$

where K is the kernel matrix of the sample S.

Finally, we will need the following result again given in Bartlett and Mendelson (2002), see also Ambroladze and Shawe-Taylor (2004) for a direct proof.

Theorem 5 *Let $\mathcal{A}$ be a Lipschitz function with Lipschitz constant L mapping the reals to the reals satisfying $\mathcal{A}(0) = 0$. The Rademacher complexity of the class $\mathcal{A} \circ \mathcal{F}$ satisfies*

$$\hat{R}_\ell(\mathcal{A} \circ \mathcal{F}) \le 2L\hat{R}_\ell(\mathcal{F}).$$

Furthermore for any classes $\mathcal{F}$ and $\mathcal{G}$

$$\hat{R}_\ell(\mathcal{F} + \mathcal{G}) \le \hat{R}_\ell(\mathcal{F}) + \hat{R}_\ell(\mathcal{G}).$$

3 Canonical Correlation Analysis

Consider two multivariate projections $\phi_a(x)$ and $\phi_b(x)$ of a random object. These will be the two views of the object x. We seek to maximise the empirical correlation between $x_a = \mathbf{w}_a'\phi_a(x)$ and $x_b = \mathbf{w}_b'\phi_b(x)$ over the projection directions $\mathbf{w}_a$ and $\mathbf{w}_b$. Without loss of generality, we assume the mean in the feature space to be zero. The empirical correlation expression can be written as

$$\begin{aligned}\max \rho &= \frac{\hat{\mathbb{E}}[x_a x_b]}{\sqrt{\hat{\mathbb{E}}[x_a^2]\hat{\mathbb{E}}[x_b^2]}}\\ &= \frac{\hat{\mathbb{E}}[\mathbf{w}_a'\phi_a(x)\phi_b(x)'\mathbf{w}_b]}{\sqrt{\hat{\mathbb{E}}[\mathbf{w}_a'\phi_a(x)\phi_a(x)'\mathbf{w}_a]\hat{\mathbb{E}}[\mathbf{w}_b'\phi_b(x)\phi_b(x)'\mathbf{w}_b]}}\\ &= \frac{\mathbf{w}_a'C_{ab}\mathbf{w}_b}{\sqrt{\mathbf{w}_a'C_{aa}\mathbf{w}_a\mathbf{w}_b'C_{bb}\mathbf{w}_b}},\end{aligned}$$

where

$$C_{st} = \frac{1}{\ell}\sum_{i=1}^{\ell}\phi_s(x_i)\phi_t(x_i)', \quad \text{for } s,t \in \{a,b\}.$$

Since the quotient is invariant to rescaling of $\mathbf{w}_a$ and $\mathbf{w}_b$ we can impose the constraints $\mathbf{w}_a'C_{aa}\mathbf{w}_a = 1$ and $\mathbf{w}_b'C_{bb}\mathbf{w}_b = 1$.

Following (Bach and Jordan 2002; Hardoon et al. 2004) the dual form of CCA will be given by solving

$$\max_{\alpha,\beta} \rho = \alpha' K_a K_b \beta$$

subject to $\alpha' K_a K_a \alpha = 1$ and $\beta' K_b K_b \beta = 1$, where K_a and K_b are the kernel matrices for the first and second view respectively. Although we present only the first direction, further directions are computed similarly where $\alpha_i' K_a K_a \alpha_j = 0$ for $i \neq j$.

The corresponding Lagrangian is

$$\begin{aligned}\mathcal{L}(\lambda,\alpha,\beta) = \alpha' K_a K_b \beta - \frac{\lambda_\alpha}{2}\left(\alpha' K_a^2 \alpha - 1\right)\\ - \frac{\lambda_\beta}{2}\left(\beta' K_b^2 \beta - 1\right).\end{aligned}$$

Taking derivatives with respect to α and β we obtain

$$\frac{\partial \mathcal{L}}{\partial \alpha} = K_a K_b \beta - \lambda_\alpha K_a^2 \alpha = \mathbf{0}, \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = K_b K_a \alpha - \lambda_\beta K_b^2 \beta = \mathbf{0}. \tag{3}$$

Subtracting β' times equation (3) from α' times equation (2) we have

$$\begin{aligned}0 &= \alpha' K_a K_b \beta - \alpha' \lambda_\alpha K_a^2 \alpha - \beta' K_b K_a \alpha + \beta' \lambda_\beta K_b^2 \beta\\ &= \lambda_\beta \beta' K_b^2 \beta - \lambda_\alpha \alpha' K_a^2 \alpha\end{aligned}$$

which together with the constraints implies that $\lambda_\alpha - \lambda_\beta = 0$, let $\lambda = \lambda_\alpha = \lambda_\beta$. Considering the case where the kernel matrices K_a and K_b are invertible, we have

$$\begin{aligned}\beta &= \frac{K_b^{-1} K_b^{-1} K_b K_a \alpha}{\lambda}\\ &= \frac{K_b^{-1} K_a \alpha}{\lambda}\end{aligned}$$

substituting in (2) we obtain

$$K_a K_b K_b^{-1} K_a \alpha - \lambda^2 K_a K_a \alpha = 0.$$

Hence

$$K_a K_a \alpha - \lambda^2 K_a K_a \alpha = 0$$

or

$$I\alpha = \lambda^2 \alpha. \tag{4}$$

If we centre the data these arguments would need to be refined but the main property would hold. We are left with a standard eigenproblem of the form $A\mathbf{x} = \lambda\mathbf{x}$. We can deduce from (4) that $\lambda = \pm 1$ for every vector of α, we ignore the negative correlation; hence we can choose the projections α to be unit vectors j_i $i = 1, \dots, \ell$ while β are the columns of $\frac{1}{\lambda} K_b^{-1} K_a$. Hence when K_a and K_b are invertible, perfect correlations can be formed. Since kernel methods provide high dimensional representations such dependence is not uncommon, as for instance with the Gaussian kernel. It is therefore clear that a naive application of CCA in a kernel defined feature space will not provide useful results (Leurgans et al. 1993).

4 CCA convergence analysis

We would like to capture the notion that the features from one view are almost identical to the features from the second view. The function $g_{\mathbf{w}_a,\mathbf{w}_b}(x) = \|\mathbf{w}_a'\phi_a(x) - \mathbf{w}_b'\phi_b(x)\|^2$ measures this property, since if $g_{\mathbf{w}_a,\mathbf{w}_b}(x) \approx 0$ the feature $\mathbf{w}_a'\phi_a(x)$ that can be obtained from one view of the data is almost identical to the second view's feature $\mathbf{w}_b'\phi_b(x)$. Therefore such pairs of features are able to capture underlying semantic properties of the data that are present in both views. In practice we will project into a k-dimensional space using as projection eigenvectors corresponding to the top k correlation directions. In order to handle this case we introduce the matrix W_a whose columns are the first k vectors $\mathbf{w}_a^1, \dots, \mathbf{w}_a^k$, and W_b with the corresponding $\mathbf{w}_b^i$ $i = 1, \dots, k$.

We are able to obtain a convergence analysis of the function by simply viewing $g_{\mathbf{w}_a,\mathbf{w}_b}(x)$ as a regression function, albeit with special structure, attempting to learn the constant 0 function. In order to apply the Rademacher generalisation bound, we must compute the empirical expected value of

$$\begin{aligned} g_{a,b}(x) &:= \hat{\mathbb{E}}\big[\|\mathbf{W}_a'\phi_a(x) - \mathbf{W}_b'\phi_b(x)\|^2\big] \\ &= \frac{1}{\ell}\sum_{i}^{\ell} \big(\phi_a(x_i)'\mathbf{W}_a\mathbf{W}_a'\phi_a(x_i) - 2\phi_a(x_i)'\mathbf{W}_a\mathbf{W}_b'\phi_b(x_i) + \phi_b(x_i)'\mathbf{W}_b\mathbf{W}_b'\phi_b(x_i)\big), \end{aligned} \tag{5}$$

where

$$\begin{aligned} \phi_a'(x)\mathbf{W}_a\mathbf{W}_a'\phi_a(x) &= \mathrm{Tr}(\phi_a(x)'\mathbf{W}_a\mathbf{W}_a'\phi_a(x)) \\ &= \mathrm{Tr}(\mathbf{W}_a\mathbf{W}_a'\phi_a(x)\phi_a(x)') \\ &= (\mathbf{W}_a\mathbf{W}_a') \circ (\phi_a(x)\phi_a(x)'), \end{aligned}$$

and $\mathrm{Tr}(A)$ is the trace of matrix A such that $\mathrm{Tr}(A) = \sum_i A_{ii}$. We represent $g_{a,b}(x)$ as a linear function $\hat{f}(x)$ in an appropriately defined feature space F. Let $\hat{\phi}$ be the mapping into the feature space F given by

$$\hat{\phi}(x) = \left[\mathrm{vec}(\phi_a(x)\phi_a(x)'), \mathrm{vec}(\phi_b(x)\phi_b(x)'), \sqrt{2}\mathrm{vec}(\phi_a(x)\phi_b(x)')\right]',$$

where $\mathrm{vec}(A)$ creates a row vector out of the entries of the matrix A by concatenating its rows. We have assumed for simplicity that the feature space is finite dimensional. Similar results can be obtained for the infinite dimensional case. Note that if $\circ$ denotes the Frobenius inner product between matrices, we have

$$A \circ B = \langle \mathrm{vec}(A), \mathrm{vec}(B)\rangle = \mathrm{Tr}(A'B) = \sum_i \sum_j A_{ij} B_{ij}.$$

Furthermore,

$$\langle \mathrm{vec}(\mathbf{u}_1\mathbf{u}_2'), \mathrm{vec}(\mathbf{v}_1\mathbf{v}_2')\rangle = \mathbf{u}_1\mathbf{u}_2' \circ \mathbf{v}_1\mathbf{v}_2' = \mathbf{v}_1'\mathbf{u}_1\mathbf{u}_2'\mathbf{v}_2, \tag{6}$$

for $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{v}_1$, $\mathbf{v}_2$ appropriately dimensioned vectors. The kernel $\hat{\kappa}$ corresponding to the feature mapping $\hat{\phi}$ is therefore given by

$$\begin{aligned}\hat{\kappa}(x,z) &= (\phi_a(x)'\phi_a(z))^2 + (\phi_b(x)'\phi_b(z))^2 + 2(\phi_a(x)'\phi_a(z))(\phi_b(x)'\phi_b(z)) \\ &= (\kappa_a(x,z) + \kappa_b(x,z))^2.\end{aligned}$$

Again using (6) it can be verified that the weight vector

$$\hat{\mathbf{W}} = \left[\mathrm{vec}(\mathbf{W}_a\mathbf{W}_a'), \mathrm{vec}(\mathbf{W}_b\mathbf{W}_b'), -\sqrt{2}\mathrm{vec}(\mathbf{W}_a\mathbf{W}_b')\right]',$$

satisfies

$$\begin{aligned}\langle \hat{\mathbf{W}}, \hat{\phi}(x)\rangle &= \mathrm{Tr}(\mathbf{W}_a\mathbf{W}_a'\phi_a(x)\phi_a(x)') + \mathrm{Tr}(\mathbf{W}_b\mathbf{W}_b'\phi_b(x)\phi_b(x)') - 2\mathrm{Tr}(\mathbf{W}_b\mathbf{W}_a'\phi_a(x)\phi_b(x)') \\ &= \phi_a(x)'\mathbf{W}_a\mathbf{W}_a'\phi_a(x) + \phi_b(x)'\mathbf{W}_b\mathbf{W}_b'\phi_b(x) - 2\phi_a(x)'\mathbf{W}_a\mathbf{W}_b'\phi_b(x) \\ &= \|\mathbf{W}_a'\phi_a(x)\|^2 + \|\mathbf{W}_b'\phi_b(x)\|^2 - 2\phi_a(x)\mathbf{W}_a\mathbf{W}_b'\phi_b(x) \\ &= \|\mathbf{W}_a'\phi_a(x) - \mathbf{W}_b'\phi_b(x)\|^2.\end{aligned}$$

It follows that $\hat{\mathbf{W}}$ realises the function $g_{a,b}(x)$ in the feature space defined by $\hat{\phi}(x)$. Furthermore again using (6) the norm of $\hat{\mathbf{W}}$ can be computed as

$$\begin{aligned}\|\hat{\mathbf{W}}\|^2 = \hat{\mathbf{W}}\hat{\mathbf{W}}' &= \mathrm{Tr}(\mathbf{W}_a\mathbf{W}_a'\mathbf{W}_a\mathbf{W}_a') \\ &\quad + \mathrm{Tr}(\mathbf{W}_b\mathbf{W}_b'\mathbf{W}_b\mathbf{W}_b') + 2\mathrm{Tr}(\mathbf{W}_b\mathbf{W}_a'\mathbf{W}_a\mathbf{W}_b') \\ &= \sum_{i,j}\left[\left(\mathbf{W}_i^{a'}\mathbf{W}_j^a\right)^2 + 2\mathbf{W}_i^{a'}\mathbf{W}_j^a\mathbf{W}_i^{b'}\mathbf{W}_j^b + \left(\mathbf{W}_i^{b'}\mathbf{W}_j^b\right)^2\right] \\ &= \sum_{i,j}\left(\mathbf{W}_i^{a'}\mathbf{W}_j^a + \mathbf{W}_i^{b'}\mathbf{W}_j^b\right)^2 \\ &= \|\mathbf{W}_a'\mathbf{W}_a + \mathbf{W}_b'\mathbf{W}_b\|_F^2.\end{aligned}$$

We are now ready to present our main theoretical result.

Theorem 6 *Fix A in $\mathbb{R}^+$. If we obtain features given by $\mathbf{W}_a^i, \mathbf{W}_b^i$ $i=1,\dots,k$ with $\|\mathbf{W}_a'\mathbf{W}_a + \mathbf{W}_b'\mathbf{W}_b\|_F \leq A$ with correlations $\rho_i = \mathbf{w}_a^{i'} C_{ab}\mathbf{w}_b^i$ and $\mathbf{w}_a^{i'} C_{aa}\mathbf{w}_a^i = 1 = \mathbf{w}_b^{i'} C_{bb}\mathbf{w}_b^i$, on a paired training set $S = \{\mathbf{x}_i, i=1,\dots,\ell\}$ of size ℓ in the feature space defined by the bounded kernels κ_a and κ_b drawn i.i.d. according to a distribution $\mathcal{D}$, then with probability greater than $1-\delta$ over the generation of S, the expected value of $g_{a,b}(x)$ on new data is bounded by*

$$\mathbb{E}_{\mathcal{D}}[g_{a,b}] \leq \hat{\mathbb{E}}_{\mathcal{D}}[g_{a,b}] + 4A\frac{1}{\ell}\sqrt{\sum_{i=1}^{\ell}(\kappa_a(x_i,x_i)+\kappa_b(x_i,x_i))^2} + 3RA\sqrt{\frac{\ln(\frac{2}{\delta})}{2\ell}} \qquad (7)$$

where

$$R = \max_{x\in\operatorname{supp}(\mathcal{D})}(\kappa_a(x,x)+\kappa_b(x,x))$$

Proof Let the kernel functions from the two corresponding feature projections be $\kappa_a(x,z) = \langle\phi_a(x),\phi_a(z)\rangle$ and $\kappa_b(x,z) = \langle\phi_b(x),\phi_b(z)\rangle$. By the above analysis, $g_{a,b}$ lies in the function class,

$$\mathcal{F}_A = \left\{x \to \langle\hat{\mathbf{W}},\hat{\phi}(x)\rangle : \|\hat{\mathbf{W}}\| \leq A\right\},$$

for $i=1,\dots,k$. We apply Theorem 2 to the loss class

$$\hat{\mathcal{F}} = \left\{\hat{f} : x \mapsto \mathcal{A}f(x) | f \in \mathcal{F}_A\right\} \subseteq \mathcal{A}\circ\mathcal{F}_A$$

where $\mathcal{A}$ is the function

$$\mathcal{A}(x) = \begin{cases} 0 & \text{if } x \leq 0; \\ \frac{x}{RA} & \text{if } 0 \leq x \leq RA; \\ 1 & \text{otherwise.}\end{cases}$$

Note that this ensures that the range of the function class is [0, 1].

Applying Theorem 2 to the pattern function $\hat{g}_{a,b} = \mathcal{A}\circ g_{a,b} \in \hat{\mathcal{F}}$ or equivalently $\hat{g}_{a,b} = g_{a,b}\frac{1}{RA}$ we can conclude that with probability $1-\delta$,

$$\mathbb{E}_{\mathcal{D}}[\hat{g}_{a,b}(x)] \leq \hat{\mathbb{E}}[\hat{g}_{a,b}(x)] + \hat{R}_\ell(\hat{\mathcal{F}}) + 3\sqrt{\frac{\ln(\frac{2}{\delta})}{2\ell}}. \qquad (8)$$

Note that $0 \leq g_{a,b}(x) \leq RA$ so $\hat{g}_{a,b}(x) = g_{a,b}(x)$ on the support of $\mathcal{D}$.

Using Theorems 4 and 5 gives

$$\hat{R}_\ell(\hat{\mathcal{F}}) \leq \frac{4A}{\ell RA}\sqrt{\sum_{i=1}^{\ell}(\kappa_a(x_i,x_i)+\kappa_b(x_i,x_i))^2}.$$

Multiplying (8) through with RA and using (5) gives the result. □

The theorem indicates, in an indirect way through $\mathcal{A}$, that the empirical value of the pattern function will be close to its expectation provided that the norms of the direction vectors are controlled and the dimension k of the projection space is small compared with ℓ. Hence, we must trade-off between finding good correlations while not allowing the norms to become too large. Theorem 6 suggests to regularise KCCA as it shows that the quality of the generalisation of the associated pattern function is controlled by the sum of the squares of the norms of the weight vectors $\mathbf{w}_a$ and $\mathbf{w}_b$. We regularise by penalising the norms of the weight vectors

$$\max_{\mathbf{w}_a,\mathbf{w}_b} \rho(\mathbf{w}_a, \mathbf{w}_b) = \frac{\mathbf{w}'_a C_{ab} \mathbf{w}_b}{\sqrt{\mathbf{w}'_a((1-\tau_a)C_{aa} + \tau_a I)\mathbf{w}_a \mathbf{w}'_b((1-\tau_b)C_{bb} + \tau_b I)\mathbf{w}_b}}$$

where τ_a and τ_b control the flexibility in the two feature spaces. Note that the analysis assumes that the vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ that satisfy

$$\mathbf{w}'_a C_{aa} \mathbf{w}_a = 1,$$
$$\mathbf{w}'_b C_{bb} \mathbf{w}_b = 1.$$

The re-normalisation is imperative as the regularised version of CCA is *not* a true CCA,[1] since it has vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ so that

$$\mathbf{w}'_a((1-\tau_a)C_{aa} + \tau_a I)\mathbf{w}_a = 1,$$
$$\mathbf{w}'_b((1-\tau_b)C_{bb} + \tau_b I)\mathbf{w}_b = 1.$$

The scaling values associated with the solutions for different regularisation values will result in the pattern function being non comparable.[2] In other words, if we do not re-normalise so that the true CCA conditions hold, the value ρ for $\tau > 0$ is not a correlation value. This is true for both the primal and dual cases. Previous works with regularised KCCA have neglected to ensure this condition is satisfied for the chosen projections.

Following Hardoon et al. (2004), the dual form of CCA with regularisation is

$$\max_{\alpha,\beta} \rho(\alpha, \beta) = \alpha' K_a K_b \beta,$$

subject to

$$(1-\tau_a)\alpha' K_a^2 \alpha + \tau_a \alpha' K_a \alpha = 1,$$
$$(1-\tau_b)\beta' K_b^2 \beta + \tau_b \beta' K_b \beta = 1.$$

The corresponding Lagrangian is

$$\begin{aligned}\mathcal{L}(\lambda_\alpha, \lambda_\beta, \alpha, \beta) = \alpha' K_a K_b \beta &\\ &- \frac{\lambda_\alpha}{2}((1-\tau_a)\alpha' K_a^2 \alpha + \tau_b \alpha' K_a \alpha - 1)\\ &- \frac{\lambda_\beta}{2}((1-\tau_a)\beta' K_b^2 \beta + \tau_b \beta' K_b \beta - 1).\end{aligned}$$

[1] Hardoon (2006) has shown that CCA with a regularisation $\tau = 1$ results in solving a Partial Least Squares (PLS) for the first direction.

[2] The scale of the weight vectors is only irrelevant with respect to the Rayleigh quotient being optimised.

Taking derivatives with respect to α and β gives

$$\frac{\partial \mathcal{L}}{\partial \alpha} = K_a K_b \beta - \lambda_\alpha ((1-\tau_a) K_a^2 \alpha + \tau_a K_a \alpha), \tag{9}$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = K_b K_a \alpha - \lambda_\beta ((1-\tau_b) K_b^2 \beta + \tau_b K_b \beta). \tag{10}$$

Subtracting β' times the second equation from α' times the first we have

$$\begin{aligned} 0 &= \alpha' K_a K_b \beta - \lambda_\alpha \alpha' ((1-\tau_a) K_a^2 \alpha + \tau_a K_a \alpha) \\ &\quad - \beta' K_b K_a \alpha + \lambda_\beta \beta' ((1-\tau_b) K_b^2 \beta + \tau_b K_b \beta), \\ &= \lambda_\beta \beta' ((1-\tau_b) K_b^2 \beta + \tau_b K_b \beta) \\ &\quad - \lambda_\alpha \alpha' ((1-\tau_a) K_a^2 \alpha + \tau_a K_a \alpha), \end{aligned}$$

which together with the constraints shows that $\lambda_\alpha - \lambda_\beta = 0$. Let $\lambda = \lambda_\alpha = \lambda_\beta$. Consider the case where K_a and K_b are invertible, we have

$$\begin{aligned} \beta &= \frac{((1-\tau_b) K_b + \tau_b I)^{-1} K_b^{-1} K_b K_a \alpha}{\lambda} \\ &= \frac{((1-\tau_b) K_b + \tau_b I)^{-1} K_a \alpha}{\lambda} \end{aligned}$$

substituting into (9) gives

$$K_b ((1-\tau_b) K_b + \tau_b I)^{-1} K_a \alpha = \lambda^2 ((1-\tau_a) K_a + \tau_a I) \alpha. \tag{11}$$

We are able to observe that by using regularisation we no longer obtain perfect correlation, as in (4). Although this is not a symmetric eigenproblem, it is easy to show (Hardoon et al. 2004) that by computing incomplete Cholesky decompositions of the kernel matrices we are able to reformulate the problem into a standard symmetric eigenproblem.

5 Experiments

In the following experiment we demonstrate how one regularisation parameter $\tau = \tau_a = \tau_b$ will control the flexibility and remove spurious features. This is shown by viewing the effect of the regularisation parameter τ on the pattern function $g_{a,b}(x)$, as defined in the previous section. We expect the regularisation to remove spurious features and hence allow for a better similarity of the two views which in turn translates into a lower value of the pattern function. In the experiments we increase the value of τ from 0 to 1 by increments of 0.05. We use the *ESP-Game* images and associated keywords as found on the ESP-Game webpage.[3] The two views of the data are obtained from the images and keywords.

We chose the combined data of image and keywords for the experiment as we believe that finding a common feature space between images and text is a non-trivial task. The ESP-Game is a website where images are displayed for users to annotate with keywords. The

[3] http://www.espgame.org.

goal is to have two or more users choose the same keyword at the same time, which is then added to the image annotation with a score representing the number of combined times the keyword has been chosen. The overall database contains several thousands of images and associated keywords. We reduce the overall number of examples by including only images that have at least 5 keywords each with a score of 10 or more and that are not grayscale. We further minimise the number of examples used by extracting the images that have at least one of the keywords *house* and *water*. This reduces our overall examples to 1682, which we divide evenly to obtain 841 training and 841 testing examples.

The extracted features were: image Hue Saturation Values (HSV) colour, image Gabor texture (Kolenda et al. 2002) and text term frequencies, which form a vector indexed by terms with entries for each word that appears in the text describing an image. Let a reference the first view derived from the image part of the data and let b reference the text part. Following previous work (Hardoon and Shawe-Taylor 2003) we compute the kernel κ_a for the first view by applying a Gaussian kernel, defined as follows

$$\kappa_a(x, y) = \exp\left(-\frac{\|\psi(x) - \psi(y)\|^2}{2\sigma^2}\right),$$

where σ is the minimum distance between the different images and $\psi(x)$ is a concatenation of the Gabor texture and HSV feature vectors. The kernel κ_b for the second view was a linear kernel on the normalised term frequency vectors.

The weight matrices $\mathbf{W}_a$ and $\mathbf{W}_b$ can be written as a linear combination of the training examples, $\mathbf{W}_a = \phi_a(S)\Delta_a$ and $\mathbf{W}_b = \phi_b(S)\Delta_b$ where $\phi_a(S)$ is the matrix with columns $\phi_a(x_i)$ and similarly $\phi_b(S)$. As we wish to apply the pattern function in the kernel space we evaluate the pattern function on the test examples as

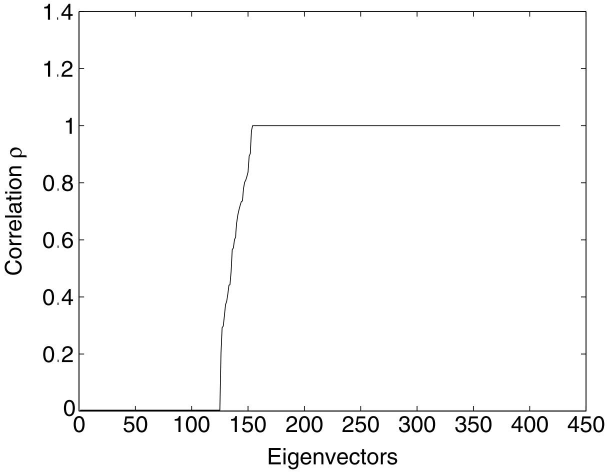
$$\begin{aligned} \frac{1}{\ell_t} g_{a,b}(x^t) &= \frac{1}{\ell_t}\sum_{i=1}^{\ell_t} \|\mathbf{W}'_a\phi_a(x_i^t) - \mathbf{W}'_b\phi_b(x_i^t)\|^2 \\ &= \frac{1}{\ell_t}\|\Delta'_a K_a^t - \Delta'_b K_b^t\|_F^2, \end{aligned} \tag{12}$$

where x_i^t are the test examples (ℓ_t is the number of test samples) and K_a^t, K_b^t are the two kernel matrices whose rows are indexed by the training examples and columns are indexed by the test examples.

We first demonstrate the case where no regularisation is used: $\tau = 0$. The obtained correlation values are plotted in Fig. 1 where we are able to observe that for the eigenvectors that contribute towards the information of the two views, these are the top eigenvectors, do indeed exhibit "perfect" correlation. Perfect correlations are obtained only for a limited number of eigenvectors as our kernel matrices are not full rank, $\text{rank}(K_a) = 838$ and $\text{rank}(K_b) = 759$. The perfect correlation can give spurious features, as no control on the flexibility of the features is provided (Hardoon and Shawe-Taylor 2003). For the experiment we chose to use the last 100 eigenvectors in α and β corresponding to the largest 100 eigenvalues for the feature projection. As indicated above and in order to make a fair comparison of the various Rayleigh quotient solutions we must rescale each $\mathbf{w}_a^i$, $\mathbf{w}_b^i$ so that $\mathbf{w}_a^{i\,\prime} C_{aa}\mathbf{w}_a^i = 1 = \mathbf{w}_b^{i\,\prime} C_{bb}\mathbf{w}_b^i$ as for the $\tau = 0$ case.

In Fig. 2 we plot the pattern function $g_{a,b}$ as defined in (5) and (12) for various values of the regularisation parameter τ respectively on the training and testing data. The value of the pattern function amounts to the error between the similarity of the two views once projected into the common feature space. We are able to observe from the plots that when there is no

Fig. 1 Correlation values for $\tau = 0$ on the training data



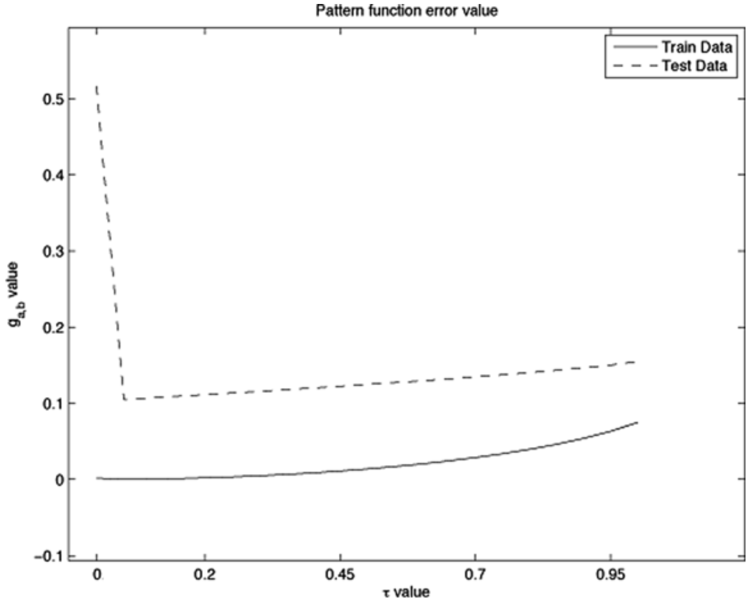


Fig. 2 The pattern function $g_{a,b}$ for different values of τ on the training & test data, normalised by the respective number of samples

control on the flexibility the error of the pattern function on the training data is 0 as we obtain perfect correlations. As some of these features are spurious the error on the testing data is relatively high. When increasing the regularisation value to extract features better defining the underlying semantics while reducing those which are spurious, the pattern function error will decrease on the testing data.

As we are introducing a penalty parameter, the performance on the training data will gradually reduce. Once an optimal value τ is found for the testing data, further increasing τ towards 1 will cause underfitting which will gradually increase the error. In Fig. 2 we are

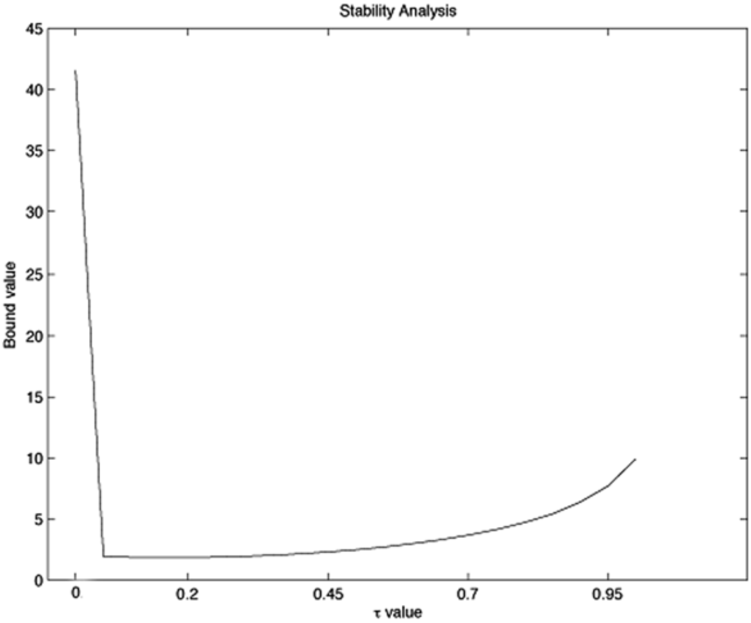


Fig. 3 In this figure we plot the new bound values on $\mathbb{E}_{\mathcal{D}}[g_{a,b}]$ for different values of τ

able to view that the error between the two views is minimal when $\tau = 0.05$. This means that with that value of τ we are able to, with higher accuracy, capture the notion that the features from one view are almost identical to the features on the second view. Hence the optimal regularisation parameter using the pattern function for testing is $\tau = 0.05$.

We plot the bound on $\mathbb{E}_{\mathcal{D}}[g_{a,b}]$ from (7) in Theorem 6. Observing that the value of the bound in Fig. 3 gives rise to the possibility of model selection i.e. we can choose the value of $\tau = 0.15$ that corresponds to the minimal bound value. From Figs. 2 and 3 we can see that the smallest bound value gives rise to a τ value that yields close to optimal performance. In Fig. 4 we plot the previously suggested bound on $\mathbb{E}_{\mathcal{D}}[g_{a,b}]$ (Shawe-Taylor and Cristianini 2004; Hardoon 2006) and it is immediately apparent that this bound is by several factors looser than the newly proposed bound and also does not allow for model selection.

Finally, we test to see whether the regularisation parameter computed by the bound is indeed an optimal, or near optimal, value with respect to a Content Based Information Retrieval (CBIR) real-world task. We asses the accuracy of retrieving the *exact* test images' paired document of keywords, also known as mate-retrieval (for a detailed description of using KCCA for CBIR we refer the reader to Hardoon et al. 2006). If the classifier is accurate the test document belonging to the test image should be near the top of the resulting list. The quality of the ordering is studied by computing average precision values. Let I_j be the index location of the retrieved mate from query q_j, the average precision p is computed as

$$p = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{I_j},$$

where M is the number of query documents. We plot the average precision in Fig. 5 where we are able to observe that the optimal regularisation parameter is $\tau = 0.1$ close to the

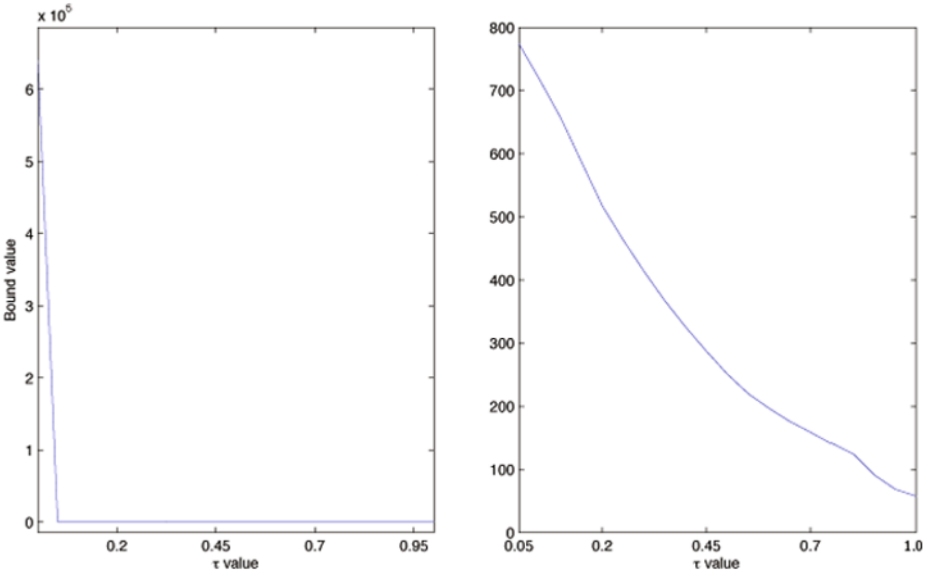


Fig. 4 In this figure we plot the previous (corrected) bound on $\mathbb{E}_{\mathcal{D}}[g_{a,b}]$ (Shawe-Taylor and Cristianini 2004) for different values of τ. The *right hand* figure is identical to the *left hand* figure excluding $\tau = 0$

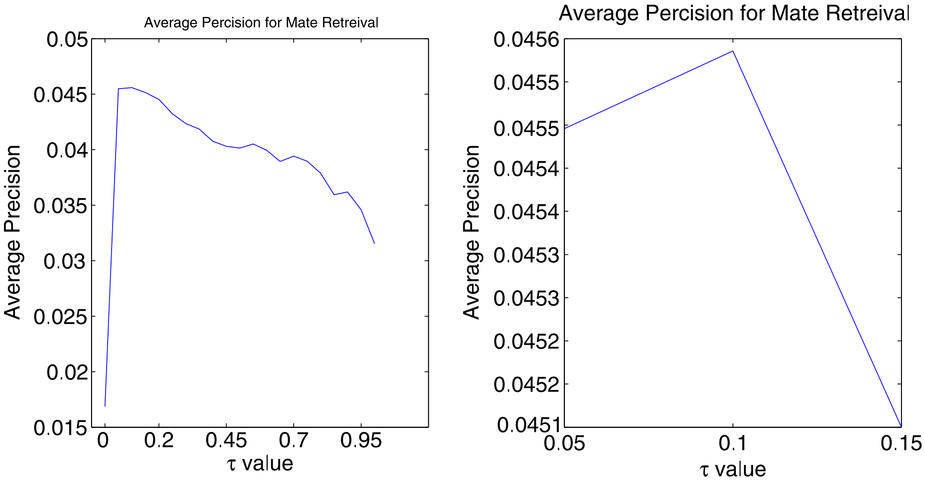


Fig. 5 The left figure is the average precision on the test data for different τ values while the *right hand* figure is a zoomed-in plot of the *left* figure

optimal parameter computed by the pattern function and by the bound. Hence showing that the regularisation parameter computed a-priori by the bound is near optimal with respect to the quality of the pattern learned and real-world application.

6 Conclusions

Kernel Canonical Correlation Analysis has been shown to be a powerful tool for extracting patterns between two complex views of data, although these patterns may be too flexible without proper regularisation. In this paper we have provided an in-depth investigation of the statistical convergence of kernel Canonical Correlation Analysis, showing that the error bound on a new example indicates that the empirical value of the pattern function will be close to its expectation provided that the norms of the two direction vectors are controlled. We have shown via the theoretical analysis a justification for regularisation, which is further validated in our experiments. The analysis has brought up a problem with the previous applications of regularised kernel CCA that did not re-normalised the projections to account for true CCA conditions and have used the subsequent ρ values as an indication of correlation. Only when the feature vectors are correctly normalised is the pattern function minimised and hence the projections most closely match. We plan to further investigate the application of the bound as a method for regularisation in model-selection.

Acknowledgements We would like to acknowledge the financial support of EU Projects LAVA, No. IST-2001-34405 and PASCAL network of excellence, No. IST-2002-506778.

References

Akaho, S. (2001). A kernel method for canonical correlation analysis. In *International meeting of psychometric society*, Osaka.

Ambroladze, A., & Shawe-Taylor, J. (2004). Complexity of pattern classes and Lipschitz property. In *Proceedings of the conference on algorithmic learning theory, ALT'04*.

Bach, F., & Jordan, M. (2002). Kernel independent component analysis. *Journal of Machine Leaning Research*, *3*, 1–48.

Bartlett, P. L., & Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, *3*, 463–482.

Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression. *Journal of the American Statistical Association*, *80*, 580–598.

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

Fukumizu, K., Bach, F. R., & Gretton, A. (2006). Consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, *8*, 361–383.

Fyfe, C., & Lai, P. (2000). ICA using kernel canonical correlation analysis. In *Proc. int. workshop on independent component analysis and blind signal separation*.

Hardoon, D. R. (2006). *Semantic models for machine learning*. Ph.D. thesis, University of Southampton.

Hardoon, D. R., & Shawe-Taylor, J. (2003). KCCA for different level precision in content-based image retrieval. In *Proceedings of third international workshop on content-based multimedia indexing*, IRISA, Rennes, France.

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

Hardoon, D. R., Saunders, C., Szedmak, S., & Shawe-Taylor, J. (2006). A correlation approach for automatic image annotation. In *Springer LNAI* (Vol. 4093, pp. 681–692).

Hardoon, D. R., Mourao-Miranda, J., Brammer, M., & Shawe-Taylor, J. (2007). Unsupervised analysis of fmri data using kernel canonical correlation. *NeuroImage*, *37*(4), 1250–1259.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 312–377.

Ketterling, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, *58*, 433–451.

Kolenda, T., Hansen, L. K., Larsen, J., & Winther, O. (2002). Independent component analysis for understanding multimedia content. In H. Bourlard, T. Adali, S. Bengio, J. Larsen, & S. Douglas (Eds.), *Proceedings of IEEE workshop on neural networks for signal processing XII* (pp. 757–766). New York: IEEE Press. Martigny, Valais, Switzerland, Sept. 4–6, 2002.

Kuss, M., & Graepel, T. (2002). *The geometry of kernel canonical correlation analysis*. Technical report, Max Planck Institute for Biological Cybernetics.

Leurgans, S. E., Moyeed, R. A., & Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *Journal at the Royal Statistical Society*, *55*, 725–740.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.

Vinokourov, A., Shawe-Taylor, J., & Cristianini, N. (2002). Inferring a semantic representation of text via cross-language correlation analysis. In *Advances of neural information processing systems 15*.

Vinokourov, A., Hardoon, D. R., & Shawe-Taylor, J. (2003). Learning the semantics of multimedia content with application to web image retrieval and classification. In *Proceedings of fourth international symposium on independent component analysis and blind source separation*, Nara, Japan.