Sparse Canonical Correlation Analysis

David R. Hardoon and John Shawe-Taylor
Centre for Computational Statistics and Machine Learning
Department of Computer Science
University College London
London, WC1E 6BT, UK
{D.Hardoon, jst}@cs.ucl.ac.uk

August 2007

Abstract

In this paper we present a novel method for solving Canonical Correlation Analysis (CCA) in a sparse convex framework using a least squares approach. The presented method focuses on the scenario when one is interested in (or limited to) a primal representation for the first view while having a dual representation for the second view. Sparse CCA (SCCA) minimises the number of features used in both the primal and dual projections while maximising the correlation between the two views.

The method is demonstrated on two paired corpuses of English-French and English-Spanish for mate-retreival and word generation tasks. We are able to observe, in the mate-retreival, that when the number of the original features is large SCCA outperforms Kernel CCA (KCCA), learning the common semantic space from a sparse set of features.

1 Introduction

Proposed by [12], CCA is a technique for finding pairs of basis vectors that maximises the correlation between a set of paired variables. The set of paired variables can be considered as two views of the same object, a perspective we adopt throughout the paper. Since the debut of CCA, a multitude of analyses, adaptations and applications have been proposed [13, 6, 7, 1, 5, 4, 2, 10, 11, 9, 18, 8].

The disadvantage of CCA and similar statistical methods is that the learned projections are a linear combination of all the features in the primal

representation and in the dual representation. This makes the interpretation of the solutions difficult. Studies by [20, 15, 3] and the more recent [17] have addressed this issue for Principle Component Analysis (PCA) and Partial Least Squares (PLS) by learning only the relevant features that maximise the variance for PCA and covariance for PLS. To the knowledge of the authors, this paper is the first attempt of giving a sparse CCA method.

We introduce a new convex least square variant of CCA which seeks a semantic projection that uses as few relevant features as possible to explain as much correlation as possible. In previous studies, CCA had either been formulated in the primal or dual (kernel) representation for both views. These formulations, coupled with the need for sparsity, could prove insufficient when one desires or is limited to a primal-dual representation, i.e. one wishes to learn the correlation of words in one language that map to documents in another. We address these possible scenarios by giving SCCA in a primal-dual framework in which one view is represented in the primal and the other in the dual (kernel defined) representation. We compare SCCA with KCCA on a bilingual English-French and English-Spanish data-set for a mate retrieval task. We show that in the mate retrieval task SCCA performs as well as KCCA when the number of original features is small. SCCA outperforms KCCA when the number of original features is large. This emphasises SCCA's ability to learn the semantic space from only the relevant features.

In Section 2 we give a brief review of CCA, and Section 3 formulates and defines SCCA. In Section 4 all the pieces are assembled to give the complete algorithm. The experiments on the paired bilingual data-sets is given in Section 5 and Section 6 concludes this paper.

2   Canonical Correlation Analysis

Consider the linear combination $x_a = \mathbf{w}_a'\mathbf{x}_a$ and $x_b = \mathbf{w}_b'\mathbf{x}_b$. Let $\mathbf{x}_a$ and $\mathbf{x}_b$ be two random variables from a multi-normal distribution, with zero mean. The correlation between x_a and x_b can be defined as

$$\max_{\mathbf{w}_a,\mathbf{w}_b} \rho = \mathbf{w}_a' C_{ab} \mathbf{w}_b \tag{1}$$

subject to $\mathbf{w}_a' C_{aa} \mathbf{w}_a = \mathbf{w}_b' C_{bb} \mathbf{w}_b = 1$. C_{aa} and C_{bb} are the non-singular within-set covariance matrices and C_{ab} is the between-sets covariance matrix.

The kernelising of CCA [6, 7] offers an alternative by first projecting the data into a higher dimensional feature space $\phi : \mathbf{x} = (x_1, \ldots, x_n) \rightarrow$

$\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x}))$ $(N \geq n)$ before performing CCA in the new feature space. The kernel variant of CCA is useful when the correlation is believed to exist in some non linear relationship. Given the kernel functions κ_a and κ_b let $K_{\boldsymbol{a}} = X_a X_a'$ and $K_{\boldsymbol{b}} = X_b X_b'$ be the kernel matrices corresponding to the two representations of the data, where X_a is the matrix whose rows are the vectors $\phi_a(\mathbf{x}_i)$, $i = 1, \ldots, \ell$ from the first representation while X_b is the matrix with rows $\phi_b(\mathbf{x}_i)$ from the second representation. The weights $\mathbf{w}_a$ and $\mathbf{w}_b$ can be expressed as a linear combination of the training examples $\mathbf{w}_a = X_a \boldsymbol{\alpha}$ and $\mathbf{w}_b = X_b \boldsymbol{\beta}$. Substitution into the primal CCA equation (1) gives the optimisation

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \rho = \boldsymbol{\alpha}' K_{\mathbf{a}} K_{\mathbf{b}} \boldsymbol{\beta}$$

subject to $\boldsymbol{\alpha}' K_{\mathbf{a}}^2 \boldsymbol{\alpha} = \boldsymbol{\beta}' K_{\mathbf{b}}^2 \boldsymbol{\beta} = 1$. This is the dual form of the primal CCA optimisation problem given above which can be cast as a generalised eigenvalue problem and for which the first k generalised eigenvectors can be found efficiently. Both CCA and KCCA can be formulated as an eigenproblem.

The theoretical analysis shown in [11, 16] suggests the need 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 weight vector norms. We refer the reader to [11, 16] for a detailed analysis and the regularised form of KCCA. Although there are advantages in using KCCA, which have been demonstrated in various experiments across the literature, we clarify that when using a linear kernel in both views, regularised CCA is the same as regularised linear KCCA (since the former and latter are linear). Using KCCA with a linear kernel has advantages over CCA, the most important being speed.[1]

3   Sparse CCA

Consider a sample from a pair of multivariate random vectors of the form $(\mathbf{x}_a^i, \mathbf{x}_b^i)$ each with zero mean where $i = 1, \ldots, \ell$. Let X_a and X_b be a matrix whose rows are the corresponding training samples and let $K_b = X_b X_b'$ be the kernel matrix of the second view. The primal-dual CCA problem can be expressed as

$$\rho(\mathbf{w}_a, \mathbf{e}) = \max_{\mathbf{w}_a, \mathbf{e}} \frac{\mathbf{w}_a' X_a K_b \mathbf{e}}{\sqrt{\mathbf{w}_a' X_a X_a' \mathbf{w}_a \mathbf{e}' K_b^2 \mathbf{e}}}, \tag{2}$$

---

[1] The KCCA toolbox used was from http://homepage.mac.com/davidrh/Code.html

3

where we choose $\mathbf{w}_a$ and $\mathbf{e}$ such that the correlation ρ between the two vectors is maximised. For simplicity let $X = X_a$, $\mathbf{w} = \mathbf{w}_a$ and $K = K_b$.

We can see this as minimising the angle between two vectors $K\mathbf{e}$ and $X'\mathbf{w}$. Since the angle is invariant to rescaling, we can fix the scaling of one vector and then minimise the norm[2] between the two vectors

$$\min_{\mathbf{w},\mathbf{e}} \|X'\mathbf{w} - K\mathbf{e}\|^2 \tag{3}$$

subject to $\|K\mathbf{e}\|^2 = 1$. This intuition is formulated in the following theorem,

Theorem 1. *Vectors $\mathbf{w}, \mathbf{e}$ are an optimal solution of equation (2) if and only if there exist λ, μ such that $\lambda\mathbf{w}, \mu\mathbf{e}$ is an optimal solution of equation (3).*

Proof. Let $\mu^2\|K\mathbf{e}\|^2 = 1$. Now let λ minimise

$$\|\lambda X'\mathbf{w} - \mu K\mathbf{e}\|^2.$$

Suppose $\lambda\mathbf{w}, \mu\mathbf{e}$ is not an optimal solution. Then there exists $\hat{\mathbf{w}}, \hat{\mathbf{e}}$ that

$$\|X'\hat{\mathbf{w}} - K\hat{\mathbf{e}}\|^2 < \|\lambda X'\mathbf{w} - \mu K\mathbf{e}\|^2$$

with $\|K\hat{\mathbf{e}}\| = 1$.

Without loss of generality assume that scaling $\hat{\mathbf{w}}$ minimises the norm, so $K\hat{\mathbf{e}} - X'\hat{\mathbf{w}} \perp X'\hat{\mathbf{w}}$ and $\mu K\mathbf{e} - \lambda X'\mathbf{w} \perp \lambda X'\mathbf{w}$. Hence $\hat{\mathbf{w}}XX'\mathbf{w} = \hat{\mathbf{w}}'XK\mathbf{e}$ and $\lambda^2\mathbf{w}'XX'\mathbf{w} = \lambda\mu\mathbf{w}'XK\mathbf{e}$. Multiplying the two norms out, we have

$$\hat{\mathbf{w}}XX'\mathbf{w} - 2\hat{\mathbf{w}}XK\hat{\mathbf{e}} + 1 < \lambda^2\mathbf{w}'XX'\mathbf{w} - 2\lambda\mu\mathbf{w}XK\mathbf{e} + 1 \Rightarrow -\hat{\mathbf{w}}XK\mathbf{e} < -\lambda\mu\mathbf{w}XK\mathbf{e}.$$

It follows that

$$\frac{\hat{\mathbf{w}}'XK\mathbf{e}}{\sqrt{\hat{\mathbf{w}}'XX'\hat{\mathbf{w}}\mathbf{e}'K^2\mathbf{e}}} = \sqrt{\mathbf{w}XK\mathbf{e}} > \sqrt{\hat{\lambda}\mu\mathbf{w}'XK\mathbf{e}} = \frac{\hat{\lambda}\mu\mathbf{w}'XK\mathbf{e}}{\sqrt{\hat{\lambda}^2\mathbf{w}'XX'\mathbf{w}\mu^2\mathbf{e}'K^2\mathbf{e}}}$$

contradicting the optimality of $\mathbf{w}, \mathbf{e}$.

Suppose for $\mathbf{w}, \mathbf{e}$ there exists λ, μ such that $\lambda\mathbf{w}, \mu\mathbf{e}$ are an optimal solution of equation (3) satisfying $\rho(\hat{\mathbf{w}}, \hat{\mathbf{e}}) > \rho(\mathbf{w}, \mathbf{e})$, rescale $\hat{\mathbf{w}}, \hat{\mathbf{e}}$ as in the first part with $\hat{\lambda}, \hat{\mu}$ and a reverse inequality follows for the norms

$$\|\hat{\lambda}X'\hat{\mathbf{w}} - \hat{\mu}K\hat{\mathbf{e}}\|^2 < \|\lambda X'\mathbf{w} - \mu K\mathbf{e}\|^2$$

contradicting the optimality of $\lambda\mathbf{w}, \mu\mathbf{e}$. $\square$

---

[2]We define $\|\cdot\|$ to be the $2-$norm.

Rather than constraining the 2−norm of $K\mathbf{e}$ which will result in a non convex problem we fix the ∞-norm of the vector $\mathbf{e}$. This will be achieved by fixing each index $\mathbf{e}_k = 1$ in turn and constraining the 1−norm of the remaining coefficients. We are also now able to constrain the 1−norm of $\mathbf{w}$ without effecting the convexity of the problem. This gives the final optimisation

$$\min_{\mathbf{w},\mathbf{e}} \|X'\mathbf{w} - K\mathbf{e}\|^2 + \mu\|\mathbf{w}\|_1 + \gamma\|\mathbf{e}\|_1 \tag{4}$$

subject to $\mathbf{e}_k = 1$.

4 Derivation & Algorithm

The minimisation in equation (4) added with its constraints gives the corresponding Lagrangian

$$\begin{aligned}
\mathcal{L} &= (\mathbf{w}^+ - \mathbf{w}^-)'XX'(\mathbf{w}^+ - \mathbf{w}^-) + \mathbf{e}'K^2\mathbf{e} - 2(\mathbf{w}^+ - \mathbf{w}^-)'XK\mathbf{e} \quad (5)\\
&\quad - \boldsymbol{\alpha}^{-'}\mathbf{w}^- - \boldsymbol{\alpha}^{+'}\mathbf{w}^+ - \boldsymbol{\beta}'\mathbf{e} + \gamma(\mathbf{e}'\mathbf{j}) + \mu((\mathbf{w}^+ + \mathbf{w}^-)'\mathbf{j}),
\end{aligned}$$

where $\mathbf{j}$ is the all ones vector. The corresponding Lagrangian in equation (5) is subject to

$$\begin{aligned}
\mathbf{w} &= \mathbf{w}^+ - \mathbf{w}^-\\
\mu &\geq 0\\
\gamma &\geq 0\\
\boldsymbol{\alpha}^+ &\geq \mathbf{0}\\
\boldsymbol{\alpha}^- &\geq \mathbf{0}\\
\boldsymbol{\beta} &\geq \mathbf{0}.
\end{aligned}$$

We will show that the constraints on the Lagrangian variables will form the baseline criterion for selecting the non-zero elements from $\mathbf{w}$ and $\mathbf{e}$.

Taking derivatives in respect to $\mathbf{w}^+$, $\mathbf{w}^-$, $\mathbf{e}$ and equating to zero gives

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{w}^+} &= 2XX'(\mathbf{w}^+ - \mathbf{w}^-) - 2X'K\mathbf{e} - \boldsymbol{\alpha}^+ + \mu\mathbf{j} = \mathbf{0}\\
\frac{\partial\mathcal{L}}{\partial\mathbf{w}^-} &= -2XX'(\mathbf{w}^+ - \mathbf{w}^-) + 2X'K\mathbf{e} - \boldsymbol{\alpha}^- + \mu\mathbf{j} = \mathbf{0}\\
\frac{\partial\mathcal{L}}{\partial\mathbf{e}} &= 2K^2\mathbf{e} - 2KX'\mathbf{w} - \boldsymbol{\beta} + \gamma'\mathbf{j} = \mathbf{0},
\end{aligned}$$

adding the first two equations gives $\boldsymbol{\alpha}^+ = 2\mu\mathbf{j} - \boldsymbol{\alpha}^-$, implying a lower and upper bound on $\boldsymbol{\alpha}^-$ (and $\boldsymbol{\alpha}^+$) of $\mathbf{0} \leq \boldsymbol{\alpha}^- \leq 2\mu\mathbf{j}$. We use the bound on $\boldsymbol{\alpha}$

as an indication as to which $\mathbf{w}$'s are to be updated by only updating the $\mathbf{w}_i$'s whose corresponding $\boldsymbol{\alpha}_i$ violates the bound. Similarly, we only update $\mathbf{e}_i$ that has a corresponding $\boldsymbol{\beta}_i$ value smaller than 0.

We are able to rewrite the derivative with respect to $\mathbf{w}^+$ in terms of $\boldsymbol{\alpha}^-$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{w}^+} &= 2XX'(\mathbf{w}^+ - \mathbf{w}^-) - 2X'K\mathbf{e} - 2\mu\mathbf{j} + \boldsymbol{\alpha}^- + \mu\mathbf{j} \\
&= 2XX'(\mathbf{w}^+ - \mathbf{w}^-) - 2X'K\mathbf{e} - \mu\mathbf{j} + \boldsymbol{\alpha}^-.
\end{aligned}
$$

Taking second derivatives in respect to $\mathbf{w}^+$ and $\mathbf{w}^-$, gives

$$
\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w}^+} &= 2XX' \\
\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w}^-} &= -2XX',
\end{aligned}
$$

so for the $\mathbf{j}_i$, the unit vector with entry 1, we have an exact Taylor series expansion

$$
\begin{aligned}
\mathcal{L}(\mathbf{w}^+ + t^+\mathbf{j}_i) &= \mathcal{L} + \frac{\partial \mathcal{L}}{\partial \mathbf{w}_i^+}t^+ + \frac{\partial^2 \mathcal{L}}{\partial \mathbf{w}_i^+}(t^+)^2 \\
\mathcal{L}(\mathbf{w}^- + t^-\mathbf{j}_i) &= \mathcal{L} + \frac{\partial \mathcal{L}}{\partial \mathbf{w}_i^-}t^- + \frac{\partial^2 \mathcal{L}}{\partial \mathbf{w}_i^-}(t^-)^2
\end{aligned}
$$

giving us the exact update for $\mathbf{w}^+$ by setting

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{w}^+ + t^+\mathbf{j}_i)}{\partial t^+} &= \left(2XX'(\mathbf{w}^+ - \mathbf{w}^-) - 2X'K\mathbf{e} - \boldsymbol{\alpha}^+ + \mu\mathbf{j}\right)_i + 4(XX')_{ii}t^+ = 0 \\
\Rightarrow t^+ &= \frac{1}{4(XX')_{ii}}\left[2X'K\mathbf{e} - 2XX'(\mathbf{w}^+ - \mathbf{w}^-) - \boldsymbol{\alpha}^- + \mu\mathbf{j}\right]_i,
\end{aligned}
$$

and for $\mathbf{w}^-$ as

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{w}^- + t^-\mathbf{j}_i)}{\partial t^-} &= \left(-2XX'(\mathbf{w}^+ - \mathbf{w}^-) + 2X'K\mathbf{e} - \boldsymbol{\alpha}^- + \mu\mathbf{j}\right)_i + 4(XX')_{ii}t^- = 0 \\
\Rightarrow t^- &= -\frac{1}{4(XX')_{ii}}\left[2X'K\mathbf{e} - 2XX'(\mathbf{w}^+ - \mathbf{w}^-) - \boldsymbol{\alpha}^- + \mu\mathbf{j}\right]_i.
\end{aligned}
$$

Recall that $\mathbf{w} = (\mathbf{w}^+ - \mathbf{w}^-)$, hence the update rule for $\mathbf{w}$ is

$$
\mathbf{w}_i \leftarrow \mathbf{w}_i + (\Delta\mathbf{w}_i^+ - \Delta\mathbf{w}_i^-).
$$

Algorithm 1 The SCCA algorithm

input: Data matrix $\mathbf{X} \in \mathbb{R}^{N \times \ell}$, Kernel matrix $\mathbf{K} \in \mathbb{R}^{\ell \times \ell}$ and $\mathbf{e}_k = 1$.

  % Initialisation:
  $\mathbf{w} = \mathbf{0}, \mathbf{j} = \mathbf{1}$

  $\mu = \frac{1}{M} \sum_i^M |(2XK\mathbf{e})_i|$
  $\gamma = \frac{1}{N} \sum_i^N |(2K^2\mathbf{e})_i|$

  $\boldsymbol{\alpha}^- = 2X'K\mathbf{e} + \mu\mathbf{j}$
  $I = (\boldsymbol{\alpha} < \mathbf{0}) \,\&\, (\boldsymbol{\alpha} > 2\mu\mathbf{j})$

  repeat
    % Update the found weight values:
    repeat
      for $i = 1$ to length of I **do**
        if $\boldsymbol{\alpha}_{I_i} > 2\mu$ **then**
          $\boldsymbol{\alpha}_{I_i} = 2\mu$
          $\hat{\mathbf{w}}_{I_i} \leftarrow \mathbf{w}_{I_i} + \frac{1}{2(XX')_{I_i,I_i}} \left[2(X'K\mathbf{e})_{I_i} - 2(XX'\mathbf{w})_{I_i} - \boldsymbol{\alpha}^-_{I_i} + \mu \right]$
        else if $\boldsymbol{\alpha}_{I_i} < 0$ **then**
          $\boldsymbol{\alpha}_{I_i} = 0$
          $\hat{\mathbf{w}}_{I_i} \leftarrow \mathbf{w}_{I_i} + \frac{1}{2(XX')_{I_i,I_i}} \left[2(X'K\mathbf{e})_{I_i} - 2(XX'\mathbf{w})_{I_i} - \boldsymbol{\alpha}^-_{I_i} + \mu \right]$
        else
          if $\mathbf{w}_{I_i} > 0$ **then**
            $\hat{\mathbf{w}}_{I_i} \leftarrow \mathbf{w}_{I_i} - \frac{2\mu - \boldsymbol{\alpha}_{I_i}}{2(XX')_{I_i,I_i}}$
          else if $\mathbf{w}_{I_i} < 0$ **then**
            $\hat{\mathbf{w}}_{I_i} \leftarrow \mathbf{w}_{I_i} + \frac{\boldsymbol{\alpha}_{I_i}}{2(XX')_{I_i,I_i}}$
          end if
        end if
        if $sign(\mathbf{w}_{I_i}) \neq sign(\hat{\mathbf{w}}_{I_i})$ **then**
          $\mathbf{w}_{I_i} = 0$
        else
          $\mathbf{w}_{I_i} = \hat{\mathbf{w}}_{I_i}$
        end if
      end for
    until convergence over $\mathbf{w}$

    % Find the dual values that are to be updated
    $\boldsymbol{\beta} = 2K^2\mathbf{e} - 2KX\mathbf{w} + \gamma\mathbf{j}$
    $J = (\boldsymbol{\beta} < \mathbf{0})$

    % Update the found dual projection values
    repeat
      for $i = 1$ to length of J **do**
        if $J_i \neq k$ **then**
          $\mathbf{e}_{J_i} \leftarrow \mathbf{e}_{J_i} + \frac{1}{4K^2_{J_i J_i}} \left[2(KX'\mathbf{w})_{J_i} - 2(K^2\mathbf{e})_{J_i} - \gamma \right]$
          if $\mathbf{e}_{J_i} < 0$ **then**
            $\mathbf{e}_{J_i} = 0$
          else if $\mathbf{e}_{J_i} > 1$ **then**
            $\mathbf{e}_{J_i} = 1$
          end if
        end if
      end for
    until convergence over $\mathbf{e}$

    % Find the weight values that are to be updated
    $\boldsymbol{\alpha}^- = 2X'K\mathbf{e} - 2XX'\mathbf{w} + \mu\mathbf{j}$
    $I = (\boldsymbol{\alpha} < \mathbf{0}) \,\&\, (\boldsymbol{\alpha} > 2\mu\mathbf{j})$

  until convergence
Output: Feature directions $\mathbf{w}, \mathbf{e}$

We find that $t := t^+ - t^-$ therefore the new value of $\mathbf{w}$ should be $\mathbf{w} + t\mathbf{j}_i$,

$$\mathbf{w}_i \leftarrow \mathbf{w}_i + \frac{1}{2(XX')_{ii}} \left[2X'K\mathbf{e} - 2XX'\mathbf{w} - \boldsymbol{\alpha}^- + \mu\mathbf{j} \right]_i.$$

We must also consider the update of $\mathbf{w}_i$ when $\boldsymbol{\alpha}_i$ is within the constraints and $\mathbf{w}_i \neq 0$. Notice that

$$2(XX')_{ii}\mathbf{w}_i + 2\sum_{j\neq i}(XX')_{ij}\mathbf{w}_j = 2(X'K\mathbf{e})_i - \boldsymbol{\alpha}_i + \mu.$$

It is easy to observe that the only component which can change is $2(XX')_{ii}\mathbf{w}_i$, therefore as we need to update $\mathbf{w}_i$ towards zero when $\mathbf{w}_i > 0$

$$
\begin{aligned}
2(XX')_{ii}\Delta\mathbf{w}_i &= 2\mu - \alpha_i \\
\Delta\mathbf{w}_i &= \frac{2\mu - \alpha_i}{2(XX')_{ii}}
\end{aligned}
$$

else when $\mathbf{w}_i < 0$ then

$$
\begin{aligned}
2(XX')_{ii}\Delta\mathbf{w}_i &= -\alpha_i \\
\Delta\mathbf{w}_i &= \frac{-\alpha_i}{2(XX')_{ii}}
\end{aligned}
$$

where the update rule is $\hat{\mathbf{w}}_i \leftarrow \mathbf{w}_i - \Delta\mathbf{w}_i$, but ensuring that $\mathbf{w}_i, \hat{\mathbf{w}}_i$ do not switch sign, i.e. we will always pause on zero before updating in any new direction.

We continue by taking second derivatives of the Lagrangian in equation (5) with respect to $\mathbf{e}$

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{e}} = 2K^2,$$

so for the $\mathbf{j}_i$ the unit vector with entry 1 we have an exact Taylor series expansion

$$\mathcal{L}(\mathbf{e} + t\mathbf{j}_i) = \mathcal{L} + \frac{\partial \mathcal{L}}{\partial \mathbf{e}_i}t + \frac{\partial^2 \mathcal{L}}{\partial \mathbf{e}_i}(t)^2$$

giving us the following update rule for $\mathbf{e}_i$

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{e} + t\mathbf{j}_i)}{\partial t} &= (2K^2\mathbf{e} - 2KX'\mathbf{w} - \boldsymbol{\beta} + \boldsymbol{\gamma}'\mathbf{j})_i + 4K_{ii}^2 t = 0 \\
\Rightarrow t &= \frac{1}{4K_{ii}^2} \left[2KX'\mathbf{w} - 2K^2\mathbf{e} + \boldsymbol{\beta} - \boldsymbol{\gamma}'\mathbf{j} \right]_i.
\end{aligned}
$$

8

The new value of $\mathbf{e}$ should be $\mathbf{e} + t\mathbf{j}_i$

$$\mathbf{e}_i \leftarrow \mathbf{e}_i + \frac{1}{4K_{ii}^2} \left[2KX'\mathbf{w} - 2K^2\mathbf{e} + \boldsymbol{\beta} - \gamma'\mathbf{j} \right]_i.$$

We give the complete algorithm as pseudo-code in Algorithm 1.

To ensure orthogonality of the extracted features [16] for each $\mathbf{e}_k$ and corresponding $\mathbf{w}$, we compute the residual matrices X_j, $j = 1, \ldots, k$ by projecting the columns of the data onto the orthogonal complement of $X_j'(X_jX_j'\mathbf{w}_j)$, a procedure known as deflation,

$$X_{j+1} = X_j \left(I - \mathbf{u}_j\mathbf{p}_j' \right),$$

where U is a matrix with columns $\mathbf{u}_j = X_jX_j'\mathbf{w}_j$ and P is a matrix with columns $\mathbf{p}_j = \frac{X_jX_j'\mathbf{u}_j}{\mathbf{u}_j'X_jX_j'\mathbf{u}_j}$. The extracted projection directions can be computed (following [16]) as $U(P'U)^{-1}$. Similarly we deflate for the dual view

$$K_{j+1} = \left(I - \frac{\tau_j\tau_j'}{\tau_j'\tau_j} \right) K_j \left(I - \frac{\tau_j\tau_j'}{\tau_j'\tau_j} \right),$$

where $\tau_j = K_j'(K_j'\mathbf{e}_j)$ and compute the projection directions as $B(T'KB)^{-1}T$ where B is a matrix with columns $K_j\mathbf{e}_j$ and T has columns τ_j.

5 Experiments

In the following experiments we use two paired English-French and English-Spanish corpuses. The English-French corpus consists of 300 samples with 2637 English features and 2951 French features while the English-Spanish corpus consists of $1,000$ samples with $40,629$ English features and $57,796$ Spanish features. The features represent the number of words in each language. Both corpuses are pre-processed with Term Frequency Inverse Document Frequency (TFIDF) followed by zero-meaning and normalisation. The linear kernel was used for the dual view. The KCCA regularisation parameter was heuristically fixed to be 0.03, while SCCA had no parameters to tune.

5.1 Mate Retrieval

Our experiment is of mate-retrieval, in which a document from the test corpus of one language is considered as the query and only the mate document

from the other language is considered relevant. In the following experiments the results are an average of retrieving the mate for both English and French (English and Spanish) and have been repeated 10 times with a random train-test split.

We compute the mate-retrieval by projecting the query document as well as the paired (other language) test documents into the learnt semantic space where the inner product between the projected data is computed. Let q be the query in one language and K_s the kernel matrix of the inner product between the second language's testing and training documents

$$ l = \left\langle \frac{q'\mathbf{w}}{\|q'\mathbf{w}\|}, \frac{K_s\mathbf{e}}{\|K_s\mathbf{e}\|} \right\rangle . $$

The resulting inner products l are then sorted by value. We measure the success of the mate-retrieval task using average precision, this assesses where the correct mate within the sorted inner products l is located. 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 start by giving the results for the English-French mate-retrieval as shown in Figure 1. The left plot depicts the average precision ($\pm$ standard deviation) when 50 documents are used for training and the remaining 250 are used as test queries. The right plot in Figure 1 gives the average precision ($\pm$ standard deviation) when 100 documents are used for training and the remaining 200 for testing. It is interesting to observe that even though SCCA does not learn the common semantic space using all the features (plotted in Figure 2) for either primal or dual views (although SCCA will use full dual features when using the full number of projections) its error is extremely similar to that of KCCA and in fact converges with it when a sufficient number of projections are used. It is important to emphasise that KCCA uses the full number of documents (50 and 100) and the full number of words (an average of 2794 for both languages) to learn the common semantic space. For example, following the left plot in Figure 1 and the additional plots in Figure 2 we are able to observe that when 35 projections are used KCCA and SCCA show a similar error. However, SCCA uses approximately 142 words and 42 documents to learn the semantic space, while KCCA uses 2794 words and 50 documents.

The second mate-retrieval experiment uses the English-Spanish paired corpus. In each run we randomly split the 1000 samples into 100 training and 900 testing paired documents. The results are plotted in Figure 3 where we are clearly able to observe SCCA outperforming KCCA throughout. We believe this to be a good example of when too many features hinder the learnt semantic space. The level of SCCA sparsity is plotted in Figure 4. In comparison to KCCA which uses all words $(49, 212)$ SCCA uses a maximum of 460 words.

The performance of SCCA, especially in the latter English-Spanish experiment, shows that we are indeed able to extract meaningful semantics between the two languages, using only the relevant features.
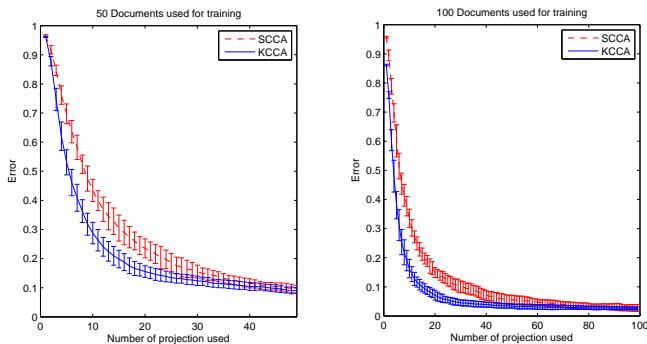
Figure 1: English-French: The average precision error $(1\text{-}p)$ with $\pm$ standard division error bars for SCCA and KCCA for different number of projections used for the mate-retrieval task. The left figure is for 50 training and 250 testing documents while the right figure is for 100 training and 200 testing documents.

6 Conclusions

Despite being introduced in 1936, CCA has proven to be a bedrock of new and continuing research. In this paper we analyse the formulation of CCA and address the issues of sparsity as well as convexity by presenting a novel SCCA method formulated as a convex least squares approach. We also provide a different perspective of solving CCA by using a primal-dual formulation which focuses on the scenario when one is interested in (or limited to) a primal representation for the first view while having a dual represen-
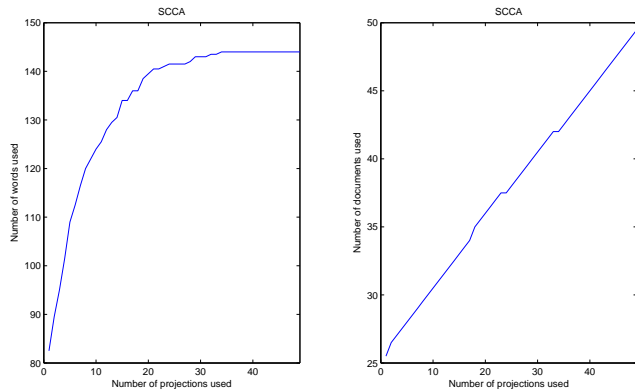
Figure 2: English-French: Level of Sparsity - The following figure is an extension of Figure 1 which uses 50 documents for training. The left figure plots the number of words used while the right figure plots the number of documents used with the number of projections. For reference, KCCA uses all the words (average of 2794) and documents (50) for all number of projections.
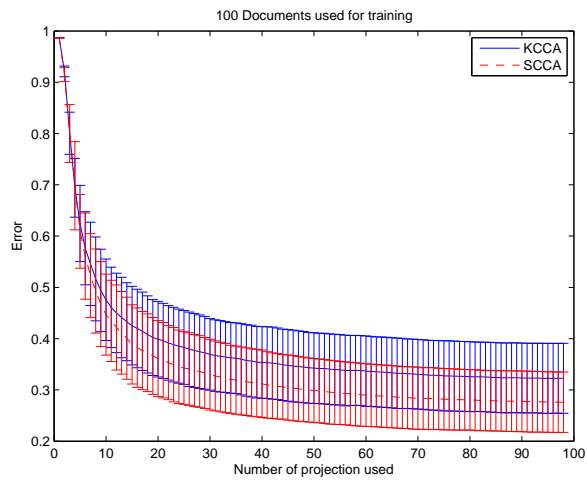

Figure 3: English-Spanish: The average precision error (1-p) with $\pm$ standard division error bars of SCCA and KCCA for different number of projections used for the mate-retrieval task. We use 100 documents for training and 900 for testing documents.
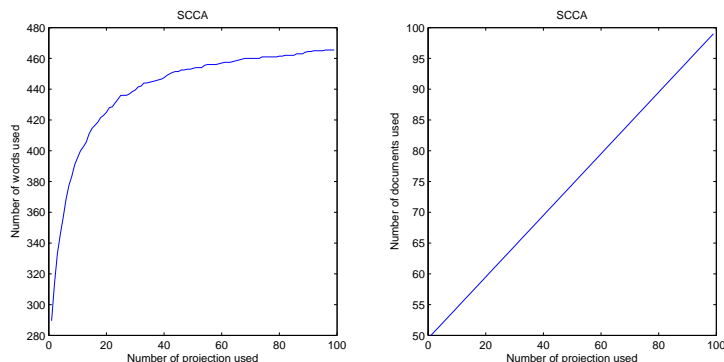
Figure 4: English-Spanish: Level of Sparsity - The following figure is an extension of Figure 3 which uses 100 documents for training. The left figure plots the number of words used and while the right figure plots the number of documents used with increasing number of projections. For reference, KCCA uses all the words (average of $49,212$) and documents (100) for all number of projections.

tation for the second view. The method is demonstrated on a bi-lingual English-French and English-Spanish paired corpuses for mate retrieval experiments. The true capacity of SCCA becomes visible when the number of features becomes extremely large as SCCA is able to learn the common semantic space using a very sparse representation of the primal-dual views. We believe this works as an initial stage for a new sparse framework for CCA to be explored and extended.

Acknowledgements

References

[1] Shotaro Akaho. A kernel method for canonical correlation analysis. In *International Meeting of Psychometric Society*, Osaka, 2001.

[2] Francis Bach and Michael Jordan. Kernel independent component analysis. *Journal of Machine Leaning Research*, 3:1–48, 2002.

[3] Charanpal Dhanjal, Steve R. Gunn, and John Shawe-Taylor. Sparse feature extraction using generalised partial least squares. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, pages 27–32, 2006.

[4] O. Friman, M. Borga, P. Lundberg, and H. Knutsson. A correlation framework for functional MRI data analysis. In *Proceedings of the 12th Scandinavian Conference on Image Analysis*, Bergen, Norway, June 2001. SCIA.

[5] O. Friman, J. Carlsson, P. Lundberg, M. Borga, and H. Knutsson. Detection of neural activity in functional MRI using canonical correlation analysis. *Magnetic Resonance in Medicine*, 45(2):323–330, February 2001.

[6] C. Fyfe and P. Lai. ICA using kernel canonical correlation analysis. In *Proc. of Int. Workshop on Independent Component Analysis and Blind Signal Separation*, 2000.

[7] Colin Fyfe and Pei Ling Lai. Kernel and nonlinear canonical correlation analysis. In *IEEE-INNS-ENNS International Joint Conference on Neural Networks*, volume 4, 2000.

[8] David R. Hardoon, Janaina Mourao-Miranda, Michael Brammer, and John Shawe-Taylor. Unsupervised analysis of fmri data using kernel canonical correlation. *NeuroImage*, In Press:–, 2007.

[9] David R. Hardoon, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. A correlation approach for automatic image annotation. In *Springer LNAI 4093*, pages 681–692, 2006.

[10] David R. Hardoon and John Shawe-Taylor. KCCA for different level precision in content-based image retrieval. In *Proceedings of Third International Workshop on Content-Based Multimedia Indexing*, IRISA, Rennes, France, 2003.

[11] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 16:2639–2664, 2004.

[12] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:312–377, 1936.

[13] J. R. Ketterling. Canonical analysis of several sets of variables. *Biometrika*, 58:433–451, 1971.

[14] Jure Leskovec, Marko Grobelnik, and Natasa Milic-Frayling. Learning semantic sub-graphs for document summarization. *Proceedings of the 7th International Multi-Conference Information Society*, B:18–25, 2004.

[15] Baback Moghaddam, Yair Weiss, and Shai Avidan. Spectral bounds for sparse pca: Exact and greedy algorithms. In *Neural Information Processing Systems (NIPS 06)*, 2006.

[16] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[17] Bharath K. Sriperumbudur, David Torres, and Gert Lanckriet. Sparse eigen methods by d.c. programming. In *To appear in International Conference on Machine Learning 2007*, 2007.

[18] Sandor Szedmak, Tijl De Bie, and David R. Hardoon. A metamorphosis of canonical correlation analysis into multivariate maximum margin learning. In *15'th European Symposium on Artificial Neural Networks (ESANN)*, 2007.

[19] Hirao Tsutomu, Kazawa Hideto, Isozaki Hideki, Maeda Eisaku, and Matsumoto Yuji. Machine learning approach to multi-document summarization. *Journal of Natural Language Processing*, 10:81–108, 2003.

[20] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. Technical report, Statistics department, Stanford University, 2004.