Mach Learn (2011) 83: 331–353
DOI 10.1007/s10994-010-5222-7

Sparse canonical correlation analysis

David R. Hardoon · John Shawe-Taylor

Received: 19 September 2008 / Revised: 11 March 2010 / Accepted: 17 October 2010 /
Published online: 6 November 2010

Abstract 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 compared to alternative sparse solutions as well as demonstrated on paired corpuses for mate-retrieval. We are able to observe, in the mate-retrieval, 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.

Keywords Sparsity · Canonical correlation analysis

1 Introduction

Proposed by Hotelling (1936), CCA is a technique for finding pairs of vectors that maximise 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

Editor: T. Jebara.

D.R. Hardoon (✉)
Data Mining Department, Institute for Infocomm Research (I2R), A*STAR, 1 Fusionopolis Way,
#21-01 Connexis, Singapore 138632, Singapore
e-mail: davidrh@me.com

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

D.R. Hardoon
e-mail: d.hardoon@cs.ucl.ac.uk

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

the debut of CCA, a multitude of analyses, adaptations and applications have been proposed (Ketterling 1971; Fyfe and Lai 2000, 2000; Akaho 2001; Friman et al. 2001a, 2001b; Bach and Jordan 2002; Hardoon and Shawe-Taylor 2003; Hardoon et al. 2004, 2006, 2007; Fukumizu et al. 2007; Szedmak et al. 2007).

The potential disadvantage of CCA and similar statistical methods, such as Principle Component Analysis (PCA) and Partial Least Squares (PLS), is that the learned projections are a linear combination of all the features in the primal and dual representations respectively. This makes the interpretation of the solutions difficult. Studies by Zou et al. (2004), Moghaddam et al. (2006), Dhanjal et al. (2006) and the more recent d'Aspremont et al. (2007), Sriperumbudur et al. (2007) have addressed this issue for PCA and PLS by learning only the relevant features that maximise the variance for PCA and covariance for PLS. Subsequent to Hardoon and Shawe-Taylor (2007) an application of sparse CCA has been proposed by Torres et al. (2007) where the authors imposed sparsity on the semantic space by penalising the cardinality of the solution vector (Weston et al. 2003). The SCCA presented in this paper is novel to the extent that instead of working with covariance matrices (Torres et al. 2007), which may be computationally intensive to compute when the dimensionality of the data is large, it deals directly with the training data.

In the Machine Learning (ML) community it is common practice to refer to the input space as the primal-representation and the kernel space as the dual-representation. In order to avoid confusion with the meanings of the terms primal and dual commonly used in the optimisation literature, we will use ML-primal to refer to the input space and ML-dual to refer to the kernel space for the remainder of the paper, though note that the references to primal and dual in the abstract refer to ML-primal and ML-dual.

Faced with real-world problems[1] combined with the need to understand or interpret the found solutions we introduce a new convex least squares 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 ML-primal (input) or ML-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 ML primal-dual representation, i.e. one wishes to learn the correlation of words in one language that map to documents in another. Further justification for this SCCA formulation is given in Sect. 2. We address these possible scenarios by formulating SCCA in a ML primal-dual framework in which one view is represented in the ML-primal and the other in the ML-dual (kernel defined) representation.

We compare our proposed SCCA solution to that of a quadratic program as well as an alternative sparse algorithm. We continue to compare SCCA with KCCA on two bilingual data-set for a mate retrieval task. In our final experiment we show that in the mate retrieval task SCCA performs as well as KCCA when the number of original features is small and SCCA outperforms KCCA when the number of original features is large. This emphasises SCCA's ability to learn the semantic space from a small number of relevant features.

In Sect. 2 we give a brief review of CCA, and Sect. 3 formulates and defines SCCA. In Sect. 4 we derive our optimisation problem and show how all the pieces are assembled to give the complete algorithm. We provide a quadratic program and sparse alternate solution to the SCCA optimisation in Sect. 5. In Sect. 6 we detail the bilingual datasets and continue to discuss our experiments in Sect. 7. Section 8 concludes this paper.

[1] Detailed motivation for such real-world problems is given in Sect. 3.

2 Canonical correlation analysis

We briefly review canonical correlation analysis and its ML-dual (kernel) variant to provide a smooth understanding of the transition to the sparse formulation. First, basic notation representation used in the paper is defined

$\mathbf{b}$ – boldface lower case letters represent vectors
s – lower case letters represent scalars
M – upper case letters represent matrices.

The correlation between $\mathbf{x}_a$ and $\mathbf{x}_b$ can be computed as

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

where $C_{aa} = X_a X_a'$ and $C_{bb} = X_b X_b'$ are the within-set covariance matrices and $C_{ab} = X_a X_b'$ is the between-sets covariance matrix, X_a is the matrix whose columns are the vectors $\mathbf{x}_i$, $i = 1, \ldots, \ell$ from the first representation while X_b is the matrix with columns $\mathbf{x}_i$ from the second representation. We are able to observe that scaling $\mathbf{w}_a, \mathbf{w}_b$ does not effect the quotient in (1), which is therefore equivalent to maximising $\mathbf{w}_a' C_{ab} \mathbf{w}_b$ subject to $\mathbf{w}_a' C_{aa} \mathbf{w}_a = \mathbf{w}_b' C_{bb} \mathbf{w}_b = 1$.

The kernelising of CCA (Fyfe and Lai 2000, 2000) offers an alternative by first projecting the data into a higher dimensional feature space $\boldsymbol{\phi}_t : \mathbf{x} = (x_1, \ldots, x_n) \rightarrow \boldsymbol{\phi}_t(\mathbf{x}) = (\boldsymbol{\phi}_1(\mathbf{x}), \ldots, \boldsymbol{\phi}_N(\mathbf{x}))$ $(N \geq n, t = a, b)$ before performing CCA in the new feature spaces. 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 $K_{\mathbf{a}} = X_a' X_a$ and $K_{\mathbf{b}} = X_b' X_b$ be the linear kernel matrices corresponding to the two representations of the data, where X_a is now the matrix whose columns are the vectors $\boldsymbol{\phi}_a(\mathbf{x}_i)$, $i = 1, \ldots, \ell$ from the first representation while X_b is the matrix with columns $\boldsymbol{\phi}_b(\mathbf{x}_i)$ 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 ML-primal CCA (1) gives the optimisation

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \rho = \frac{\boldsymbol{\alpha}' K_{\mathbf{a}} K_{\mathbf{b}} \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}' K_{\mathbf{a}}^2 \boldsymbol{\alpha} \boldsymbol{\beta} K_{\mathbf{b}}^2 \boldsymbol{\beta}}},$$

which is equivalent to maximising $\boldsymbol{\alpha}' K_{\mathbf{a}} K_{\mathbf{b}} \boldsymbol{\beta}$ subject to $\boldsymbol{\alpha}' K_{\mathbf{a}}^2 \boldsymbol{\alpha} = \boldsymbol{\beta}' K_{\mathbf{b}}^2 \boldsymbol{\beta} = 1$. This is the ML-dual form of the CCA optimisation problem given in (1) 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 symmetric eigenproblems.

A variety of theoretical analyses have been presented for CCA (Akaho 2001; Bach and Jordan 2002; Hardoon et al. 2004; Shawe-Taylor and Cristianini 2004; Fukumizu et al. 2007; Hardoon and Shawe-Taylor 2009). A common conclusion of some of these analyses is the need to regularise KCCA. For example the quality of the generalisation of the associated pattern function is shown in Hardoon and Shawe-Taylor (2009) to be controlled by the sum of the squares of the weight vector norms. 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 KCCA is the same as regularised CCA (since the former and latter are linear). Nonetheless using KCCA with a linear kernel can have advantages over CCA, the most important being speed when the number of features is larger than the number of samples.[2]

[2] The KCCA toolbox used was from http://www.davidroihardoon.com/Research/Code.html.

3 Sparse CCA

The motivation for formulating a ML primal-dual SCCA is largely intuitive when faced with real-world problems combined with the need to understand or interpret the found solutions. Consider the following examples as potential case studies which would require ML primal-dual sparse multivariate analysis methods, such as the one proposed.

– Enzyme prediction; in this problem one would like to uncover the relationship between the enzyme sequence, or more accurately the sub-sequences within each enzyme sequence that are highly correlated with the possible combination of the enzyme reactants. We would like to find a sparse ML-primal weight representation on the enzyme sequence which correlates highly to sparse ML-dual feature vector of the reactants. This will allow a better understanding of the enzyme structure relationship to reactions.
– Bilingual analysis; when learning the semantic relationship between two languages, we may want to understand how one language maps from the word space (ML-primal) to the contextual document (ML-dual) space of another language. In both cases we do not want a complete mapping from all the words to all possible contexts but to be able to extract an interpretable relationship from a sparse word representation from one language to a particular and specific context (or sparse combination of) in the other language.
– Brain analysis; here, one would be interested in finding a (ML-primal) sparse voxel[3] activation map to some (ML-dual) non-linear stimulus activation (such as musical sequences, images and various other multidimensional input). The potential ability to find only the relevant voxels in the stimuli would remove the particularly problematic issue of thresholding the full voxel activation maps that are conventionally generated.

For the scope of this paper we limit ourselves to experiments with the bilingual texts problems.

Throughout the paper we only consider the setting when one is interested in a ML-primal representation for the first view and a ML-dual representation for the second view, although it is easily shown that the given derivations hold for the inverted case (i.e. a ML-dual representation for the first view and a ML-primal representation for the second view) which is therefore omitted. Furthermore, one could easily use this framework to learn the relationship between a ML-primal and ML-dual representation of the *same* data (i.e. explicitly learning a mapping).

Consider a sample from a pair of random vectors (i.i.d. assumptions hold) of the form $(\mathbf{x}_a^i, \mathbf{x}_b^i)$ each with zero mean (i.e. centred) where $i = 1, \dots, \ell$. Let X_a and X_b be matrices whose columns are the corresponding training samples and let $K_b = X_b' X_b$ be the kernel matrix of the second view and $\mathbf{w}_b$ be expressed as a linear combination of the training examples $\mathbf{w}_b = X_b \mathbf{e}$ (note that $\mathbf{e}$ is a general vector and should not be confused with notation sometimes used for unit coordinate vectors). The primal-dual CCA problem can be expressed as a primal-dual Rayleigh quotient

$$\begin{aligned}\rho &= \max_{\mathbf{w}_a, \mathbf{w}_b} \frac{\mathbf{w}_a' X_a X_b' \mathbf{w}_b}{\sqrt{\mathbf{w}_a' X_a X_a' \mathbf{w}_a \mathbf{w}_b X_b X_b' \mathbf{w}_b}} \\ &= \max_{\mathbf{w}_a, \mathbf{e}} \frac{\mathbf{w}_a' X_a X_b' X_b \mathbf{e}}{\sqrt{\mathbf{w}_a' X_a X_a' \mathbf{w}_a \mathbf{e}' X_b' X_b X_b' X_b \mathbf{e}}}\end{aligned}$$

[3] A voxel is a pixel representing the smallest three-dimensional point volume referenced in a functional Magnetic Resonance Imaging (fMRI) image of the brain. It is usually approximately 3 mm × 3 mm.

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

where we choose the primal weights $\mathbf{w}_a$ of the first representation and dual features $\mathbf{e}$ of the second representation such that the correlation ρ between the two vectors is maximised. As we are able to scale $\mathbf{w}_a$ and $\mathbf{e}$ without changing the quotient, the maximisation in (2) is equal to maximising $\mathbf{w}_a' X_a K_b \mathbf{e}$ subject to $\mathbf{w}_a' X_a' X_a \mathbf{w}_a = \mathbf{e}' K_b^2 \mathbf{e} = 1$. For simplicity let $X = X_a$, $\mathbf{w} = \mathbf{w}_a$ and $K = K_b$.

Having provided the initial primal-dual framework we proceed to reformulate the problem as a convex sparse least squares optimisation problem. We are able to show that maximising the correlation between the two vectors $K\mathbf{e}$ and $X'\mathbf{w}$ can be viewed as minimising the angle between them. Since the angle is invariant to rescaling, we can fix the scaling of one vector and then minimise the norm[4] 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* (2) *if and only if there exist* μ, γ *such that* $\mu\mathbf{w}, \gamma\mathbf{e}$ *are an optimal solution of* (3).

Furthermore, we note that the least squares problem in (3) is not a traditional one, as it has a constraint that makes it equivalent to an eigenvalue problem. Theorem 1 is well known in the statistics community and corresponds to the equivalence between one form of Alternating Conditional Expectation (ACE) and CCA (Breiman and Friedman 1985; Hastie and Tibshirani 1990). For an exact proof see Theorem 5.1 on p. 590 in Breiman and Friedman (1985).

Constraining the 2-norm of $K\mathbf{e}$ (or $X'\mathbf{w}$) will result in a non convex problem, i.e. we will not obtain a positive/negative-definite Hessian matrix. Motivated by the Rayleigh quotient solution for optimising CCA, whose resulting symmetric eigenproblem does *not* enforce the $\|K\mathbf{e}\|^2 = 1$ constraint, i.e. the optimal solution is invariant to rescaling of the solutions we replace the scaling of $\|K\mathbf{e}\|^2 = 1$ with the scaling of $\mathbf{e}$ to be $\|\mathbf{e}\|_\infty = 1$. We will readdress the resulting convexity when we achieve the final formulation.

After finding an optimal CCA solution, we are able to re-normalise $\mathbf{e}$ so that $\|K\mathbf{e}\|^2 = 1$ holds. We emphasise that even though K has been removed from the constraint the link to kernels (kernel tricks and RKHS) is represented in the choice of kernel K used for the dual-view, otherwise the presented method is a sparse linear CCA.[5] We can now focus on obtaining an optimal sparse solution for $\mathbf{w}, \mathbf{e}$.

It is obvious that when starting with $\mathbf{w} = \mathbf{e} = \mathbf{0}$ further minimising is impossible. To avoid this trivial solution and to ensure that the constraints hold in our starting condition[6] we set $\|\mathbf{e}\|_\infty = 1$ by fixing $e_k = 1$ for some fixed index $1 \le k \le \ell$ so that $\mathbf{e} = [e_1, \dots, e_{k-1}, e_k, e_{k+1}, \dots, e_\ell]$. To further obtain a sparse solution on $\mathbf{e}$ we constrain the 1-norm of the remaining coefficients $\|\tilde{\mathbf{e}}\|_1$, where we define $\tilde{\mathbf{e}} = [e_1, \dots, e_{k-1}, e_{k+1}, \dots, e_\ell]$.

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

[5] One should keep in mind that even kernel CCA is still linear CCA performed in kernel defined feature space.

[6] $\|\mathbf{e}\|_\infty = \max(|e_1|, \dots, |e_\ell|) = 1$, therefore there must be at least one e_i for some i that is equal to 1.

The motivation behind isolating a specific k and constraining the 1-norm of the remaining coefficients, other than ensuring a non-trivial solution, follows the intuition of wanting to find similarities between the samples given some basis for comparison. In the case of documents, this places the chosen document (indexed by k) in a semantic context defined by an additional (sparse) set of documents. This captures our previously stated goal of wanting to be able to extract an interpretable relationship from a sparse word representation from one language to a particular and specific context in the other language. The $j \in \mathbb{N}^{\ell}$ choices of k correspond to the $\mathbf{e}_j, \mathbf{w}_j$ projection vectors.

We discuss the optimal choice of k and ensuring orthogonality of the sparse projections in Sect. 4.2. Furthermore, turning the constraint $\|K\mathbf{e}\| = 1$ into $\|\mathbf{e}\| = 1$ links with similar procedures in supervised classification (e.g., Roth 2004).

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 as

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

subject to $\|\mathbf{e}\|_\infty = 1$. The expression $\|X'\mathbf{w} - K\mathbf{e}\|^2$ is quadratic in the variables $\mathbf{w}$ and $\mathbf{e}$ and is bounded from below (≥ 0) and hence is convex since it can be expressed as $\|X'\mathbf{w} - K\mathbf{e}\|^2 = C + g'\mathbf{w} + f'\mathbf{e} + [\mathbf{w}'\mathbf{e}']H[\mathbf{w}'\mathbf{e}']'$. If H were not positive definite taking multiple μ of the eigenvector $v' = [v_1'v_2']$ with negative eigenvalue λ would give $C + \mu g'v_1 + \mu f'v_2 + \mu^2\lambda$ creating arbitrarily large negative values. When minimising subject to linear constraints (1-norms are linear) this makes the whole optimisation convex.

While (4) is similar to Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1994) (Basis Pursuit Denoising (Chen et al. 1999)), (4) also has similarity to non-negative matrix factorization (Heiler and Schnor 2006), it is not a standard LASSO problem unless $\mathbf{e}$ is fixed. Here we are trying to find sparse solutions for *both* $\mathbf{w}, \mathbf{e}$.

4 Derivation and algorithm

We propose a novel method for solving the optimisation problem represented in (4), where the suggested algorithm minimises the gap between the primal and dual Lagrangian solutions using a greedy search on $\mathbf{w}, \mathbf{e}$. The proposed algorithm finds a sparse $\mathbf{w}, \mathbf{e}$ vectors, by iteratively solving between the ML primal and dual formulation in turn. The solving of alternating formulations is similar in flavour to Lee et al. (2006) who had proposed a sparse coding algorithm for learning basis functions that capture higher-level feature in unlabelled data by proposing an algorithm that alternates between two optimisation L_1 and L_2 constrained least squares problems.

We give the proposed algorithm as the following high-level pseudo-code. A more complete description will follow later;

```
– Repeat
    1. Use the dual Lagrangian variables to solve the ML-primal variables
    2. Check whether all constraints on ML-primal variables hold
    3. Use ML-primal variables to solve the dual Lagrangian variables
    4. Check whether all dual Lagrangian variable constraints hold
    5. Check whether 2 holds, IF not go to 1
– End
```

We have yet to address how to determine which elements in $\mathbf{w}, \mathbf{e}$ are to be non-zero. We will show that from the derivation given in Sect. 4.1 a lower and upper bound is computed. Combining the bound with the constraints provides us with a criterion for selecting the non-zero elements for both $\mathbf{w}$ and $\mathbf{e}$. The criteria being that only the respective indices which violate the bound and the various constraints need to be updated. Furthermore, we limit ourselves to positive entries in $\mathbf{e}$ as we expect to align with a positive subset of articles.

We proceed to give the derivation of our problem. The minimisation

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

subject to $\|\mathbf{e}\|_\infty = 1$ can be written as

$$\mathbf{w}'XX'\mathbf{w} + \mathbf{e}'K^2\mathbf{e} - 2\mathbf{w}'XK\mathbf{e} + \mu\|\mathbf{w}\|_1 + \gamma\|\tilde{\mathbf{e}}\|_1$$

subject to $\|\mathbf{e}\|_\infty = 1$, where μ, γ are fixed positive parameters.

To simplify our mathematical notation we revert to uniformly using $\mathbf{e}$ in place of $\tilde{\mathbf{e}}$, as k will be fixed in an outer loop so that the only requirement is that no update will be made for e_k, which can be enforced in the actual algorithm. We further emphasise that we are only interested in the positive spectrum of $\mathbf{e}$, which again can be easily enforced by updating any $e_i < 0$ to be $e_i = 0$.[7] Therefore we could rewrite the constraint $\|\mathbf{e}\|_\infty = 1$ as $0 \le e_i \le 1, \forall i \in \mathbb{R}^\ell$.

We are able to obtain the corresponding Lagrangian

$$\mathcal{L} = \mathbf{w}'XX'\mathbf{w} + \mathbf{e}'K^2\mathbf{e} - 2\mathbf{w}'XK\mathbf{e} + \mu\|\mathbf{w}\|_1 + \gamma\mathbf{e}'\mathbf{j} - \boldsymbol{\beta}'\mathbf{e},$$

subject to

$$\boldsymbol{\beta} \ge \mathbf{0},$$

where $\boldsymbol{\beta}$ is the dual Lagrangian variable on $\mathbf{e}$ and μ, γ are positive scale factors as discussed in Theorem 1 and $\mathbf{j}$ is the all ones vector. We note that as we algorithmically ensure that $\mathbf{e} \ge 0$ we are able to write $\gamma\|\mathbf{e}\|_1 = \gamma\mathbf{e}'\mathbf{j}$ as $\|\mathbf{e}\|_1 := \sum_{i=1}^{\ell} |e_i|$.

We further observe that μ, γ can be considered as the hyper-parameters (or regularisation parameters) common in the LASSO literature, controlling the trade-off between the function objective and the level of sparsity. We show that the scale parameters can be treated as a type of dual Lagrangian parameters to provide an underlying automatic determination of sparsity. We demonstrate that this approach obtains very good results and is discussed in detail in Sect. 7.1.

To simplify the 1-norm derivation we express $\mathbf{w}$ by its positive and negative components[8] such that $\mathbf{w} = \mathbf{w}^+ - \mathbf{w}^-$ subject to $\mathbf{w}^+, \mathbf{w}^- \ge 0$.

This allows us to rewrite the Lagrangian as

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

[7] We can also easily enforce the $\|\cdot\|_\infty$ constraint by updating any $e_i > 1$ to be $e_i = 1$.

[8] This means that $\mathbf{w}^+/\mathbf{w}^-$ will only have the positive/negative values of $\mathbf{w}$ and zero elsewhere.

The corresponding Lagrangian in (5) is subject to

$$\begin{aligned}\boldsymbol{\alpha}^+ &\geq \mathbf{0},\\ \boldsymbol{\alpha}^- &\geq \mathbf{0},\\ \boldsymbol{\beta} &\geq \mathbf{0}.\end{aligned}$$

The two new dual Lagrangian variables $\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-$ are to uphold the positivity constraints on $\mathbf{w}^+, \mathbf{w}^-$.

4.1 SCCA derivation

In this section we will show that the constraints on the dual Lagrangian variables will form the criterion for selecting the non-zero elements from $\mathbf{w}$ and $\mathbf{e}$. First we define further notations used. Given the data matrix $X \in \mathbb{R}^{m\times\ell}$ and Kernel matrix $K \in \mathbb{R}^{\ell\times\ell}$ as defined in Sect. 3, we define the following vectors

$$\begin{aligned}\mathbf{w}^+ &= \left[w_1^+, \dots, w_m^+\right],\\ \mathbf{w}^- &= \left[w_1^-, \dots, w_m^-\right],\\ \boldsymbol{\alpha}^+ &= \left[\alpha_1^+, \dots, \alpha_m^+\right],\\ \boldsymbol{\alpha}^- &= \left[\alpha_1^-, \dots, \alpha_m^-\right],\\ \mathbf{e} &= [e_1, \dots, e_\ell],\\ \boldsymbol{\beta} &= [\beta_1, \dots, \beta_\ell].\end{aligned}$$

Throughout this section let i be the index of either $\mathbf{w}, \mathbf{e}$ that needs to be updated. We use the notation $(\cdot)_i$ or $[\cdot]_i$ to refer to the ith index within a vector and $(\cdot)_{ii}$ to refer to the ith element on the diagonal of a matrix.

Taking derivatives of (5) in respect to $\mathbf{w}^+$, $\mathbf{w}^-$, $\mathbf{e}$ and equating to zero gives

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{w}^+} &= 2XX'\left(\mathbf{w}^+ - \mathbf{w}^-\right) - 2X'K\mathbf{e} - \boldsymbol{\alpha}^+ + \mu\mathbf{j} = \mathbf{0},\\ \frac{\partial\mathcal{L}}{\partial\mathbf{w}^-} &= -2XX'\left(\mathbf{w}^+ - \mathbf{w}^-\right) + 2X'K\mathbf{e} - \boldsymbol{\alpha}^- + \mu\mathbf{j} = \mathbf{0},\\ \frac{\partial\mathcal{L}}{\partial\mathbf{e}} &= 2K^2\mathbf{e} - 2KX'\mathbf{w} - \boldsymbol{\beta} + \gamma'\mathbf{j} = \mathbf{0},\end{aligned} \tag{6}$$

adding the first two equations gives

$$\begin{aligned}\boldsymbol{\alpha}^+ &= 2\mu\mathbf{j} - \boldsymbol{\alpha}^-,\\ \boldsymbol{\alpha}^- &= 2\mu\mathbf{j} - \boldsymbol{\alpha}^+,\end{aligned}$$

implying a lower and upper component-wise bound on $\boldsymbol{\alpha}^-, \boldsymbol{\alpha}^+$ of

$$\begin{aligned}\mathbf{0} \leq \boldsymbol{\alpha}^- &\leq 2\mu\mathbf{j},\\ \mathbf{0} \leq \boldsymbol{\alpha}^+ &\leq 2\mu\mathbf{j}.\end{aligned}$$

We use the bound on $\boldsymbol{\alpha}$ to indicate which indices of the vector $\mathbf{w}$ need to be updated by only updating the w_i's whose corresponding α_i violates the bound (i.e. the active sets of w_i and α_i respectively). Similarly, we only update e_i that has a corresponding β_i value smaller than 0.

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

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

We wish to compute the update rule for the selected indices of $\mathbf{w}$. Taking the second derivatives of (5) in respect to $\mathbf{w}^+$ and $\mathbf{w}^-$, gives

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

so for the $\mathbf{i}_i$, the unit vector with entry 1, we have an exact Taylor series expansion t^+ and t^- respectively for w_i^+ and w_i^- as

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

giving us the exact update for w_i^+ by setting

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

Therefore the update for w_i^+ is $\Delta w_i^+ = t^+$. We also compute the exact update for w_i^- as

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

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

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

Therefore we find that the new value of w_i should be

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

We must also consider the update of w_i when α_i is within the constraints and $w_i \neq 0$, i.e. previously α_i had violated the constraints triggering the updated of w_i to be non zero. Notice from (6) that

$$2\big(XX'\big)_{ii} w_i + 2\sum_{j\neq i}\big(XX'\big)_{ij} w_j = 2\big(X'K\mathbf{e}\big)_i - \alpha_i + \mu.$$

It is easy to observe that the only component which can change is $2(XX')_{ii}w_i$, therefore as we need to update w_i towards zero. Hence when $w_i > 0$ the absolute value of the update is

$$\begin{aligned} 2\big(XX'\big)_{ii}\Delta w_i &= 2\mu - \alpha_i, \\ \Delta w_i &= \frac{2\mu - \alpha_i}{2(XX')_{ii}} \end{aligned}$$

else when $w_i < 0$ then the update is the negation of

$$\begin{aligned} 2\big(XX'\big)_{ii}\Delta w_i &= 0 - \alpha_i, \\ \Delta w_i &= \frac{-\alpha_i}{2(XX')_{ii}} \end{aligned}$$

so that the update rule is $\hat{w}_i \leftarrow w_i - \Delta w_i$. In the updating of w_i we ensure that $w_i, \hat{w}_i$ do not have opposite signs, i.e. we will always stop at zero before updating in any new direction.

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

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

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

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

giving us the following update rule for e_i

$$\begin{aligned} \frac{\partial \hat{\mathcal{L}}(\mathbf{e} + t\mathbf{i}_i)}{\partial t} &= \big(2K^2\mathbf{e} - 2KX'\mathbf{w} - \boldsymbol{\beta} + \gamma'\mathbf{j}\big)_i + 4K_{ii}^2 t = 0 \\ \implies \quad t &= \frac{1}{4K_{ii}^2}\big[2KX'\mathbf{w} - 2K^2\mathbf{e} + \boldsymbol{\beta} - \gamma'\mathbf{j}\big]_i, \end{aligned}$$

the update for $\mathbf{e}$ is $\Delta e_i = t$. The new value of e_i will be

$$\hat{e}_i \leftarrow e_i + \frac{1}{4K_{ii}^2}\big[2KX'\mathbf{w} - 2K^2\mathbf{e} + \boldsymbol{\beta} - \gamma'\mathbf{j}\big]_i,$$

again ensuring that $0 \leq \hat{e}_i \leq 1$.

Algorithm 1 The SCCA algorithm

```
Input: Data matrix $\mathbf{X} \in \mathbb{R}^{m\times\ell}$, Kernel matrix $\mathbf{K} \in \mathbb{R}^{\ell\times\ell}$ and the value $k$.
  % Initialisation:
  $\mathbf{w} = \mathbf{0}$, $\mathbf{j} = \mathbf{1}$, $\mathbf{e} = \mathbf{0}$, $e_k = 1$
  $\mu = \frac{1}{M}\sum_i^M |(2XK\mathbf{e})_i|$, $\gamma = \frac{1}{\ell}\sum_i^\ell |(2K^2\mathbf{e})_i|$
  $\boldsymbol{\alpha}^- = 2XK\mathbf{e} + \mu\mathbf{j}$
  $I = (\boldsymbol{\alpha} < \mathbf{0}) \parallel (\boldsymbol{\alpha} > 2\mu\mathbf{j})$

  repeat
    % Update the found weight values:
    Converge over $\mathbf{w}$ using Algorithm 2

    % 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
    Converge over $\mathbf{e}$ using Algorithm 3

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

  $\mathbf{e} = \frac{\mathbf{e}}{\|K\mathbf{e}\|}$, $\mathbf{w} = \frac{\mathbf{w}}{\|X'\mathbf{w}\|}$

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

4.2 SCCA algorithm

Observe that in the initial condition when $\mathbf{w} = \mathbf{0}$ from (6) we are able to treat the scale parameters μ, γ as dual Lagrangian variables and set them to

$$\mu = \frac{1}{m}\sum_i^m \left|\left(2XK\mathbf{e}\right)_i\right|,$$

$$\gamma = \frac{1}{\ell}\sum_i^\ell \left|\left(2K^2\mathbf{e}\right)_i\right|.$$

We emphasise that this is to provide an underlying automatic determination of sparsity and show in Sect. 7.1 that this method works well in practice. Combining all the pieces we give the SCCA algorithm as pseudo-code in Algorithm 1, which takes k as a parameter. In order to choose the optimal value of k we need to run the algorithm with all values of k and select the solution (and respective k), in each iteration, which gives the best (minimum) objective value.

Finally, to ensure orthogonality of the extracted features (Shawe-Taylor and Cristianini 2004) for each $\mathbf{e}_j$ and corresponding $\mathbf{w}_j$, we compute the residual matrices X_j, $j = 1, \ldots, \ell$

Algorithm 2 The SCCA algorithm—Convergence over **w**

```
repeat
  for $i = 1$ to length of $I$ do
    if $\alpha_{I_i} > 2\mu$ then
      $\alpha_{I_i} = 2\mu$
      $\hat{w}_{I_i} \leftarrow w_{I_i} + \frac{1}{2(XX')_{I_i,I_i}}[2(XK\mathbf{e})_{I_i} - 2(XX'\mathbf{w})_{I_i} - \alpha^{-}_{I_i} + \mu]$
    else if $\alpha_{I_i} < 0$ then
      $\alpha_{I_i} = 0$
      $\hat{w}_{I_i} \leftarrow w_{I_i} + \frac{1}{2(XX')_{I_i,I_i}}[2(XK\mathbf{e})_{I_i} - 2(XX'\mathbf{w})_{I_i} - \alpha^{-}_{I_i} + \mu]$
    else
      if $w_{I_i} > 0$ then
        $\hat{w}_{I_i} \leftarrow w_{I_i} - \frac{2\mu - \alpha_{I_i}}{2(XX')_{I_i,I_i}}$
      else if $w_{I_i} < 0$ then
        $\hat{w}_{I_i} \leftarrow w_{I_i} + \frac{\alpha_{I_i}}{2(XX')_{I_i,I_i}}$
      end if
    end if
    if $\mathrm{sign}(w_{I_i}) \neq \mathrm{sign}(\hat{w}_{I_i})$ then
      $w_{I_i} = 0$
    else
      $w_{I_i} = \hat{w}_{I_i}$
    end if
  end for
until convergence over $\mathbf{w}$
```

Algorithm 3 The SCCA algorithm—Convergence over **e**

```
repeat
  for $i = 1$ to length of $J$ do
    if $J_i \neq k$ then
      $e_{J_i} \leftarrow e_{J_i} + \frac{1}{4K^2_{J_i J_i}}[2(KX'\mathbf{w})_{J_i} - 2(K^2\mathbf{e})_{J_i} - \gamma]$
      if $e_{J_i} < 0$ then
        $e_{J_i} = 0$
      else if $e_{J_i} > 1$ then
        $e_{J_i} = 1$
      end if
    end if
  end for
until convergence over $\mathbf{e}$
```

by projecting the columns of the data onto the orthogonal complement of $X'_j(X_j X'_j \mathbf{w}_j)$, a procedure known as deflation,

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

where U is a matrix with columns $\mathbf{u}_j = X_j X'_j \mathbf{w}_j$ and P is a matrix with columns $\mathbf{p}_j = \frac{X_j X'_j \mathbf{u}_j}{\mathbf{u}'_j X_j X'_j \mathbf{u}_j}$. The extracted projection directions can be computed (following Shawe-Taylor

Algorithm 4 The SCCA algorithm with deflation

```
Input: Data matrix $\mathbf{X} \in \mathbb{R}^{m\times\ell}$, Kernel matrix $\mathbf{K} \in \mathbb{R}^{\ell\times\ell}$.
  $X_1 = X$, $K_1 = K$
  for $j = 1$ to $\ell$ do
    $k$ = select optimal $k$ (elaborated in text)
    $[\mathbf{e}_j, \mathbf{w}_j]$ = SCCA_Algorithm 1 $(X_j, K_j, k)$

    $\tau_j = K_j'(K_j'\mathbf{e}_j)$
    $\mathbf{u}_j = X_j X_j' \mathbf{w}_j$
    $\mathbf{p}_j = \frac{X_j X_j' \mathbf{u}_j}{\mathbf{u}_j' X_j X_j' \mathbf{u}_j}$

    if $j < \ell$ then
      $K_{j+1} = (I - \frac{\tau_j \tau_j'}{\tau_j' \tau_j}) K_j (I - \frac{\tau_j \tau_j'}{\tau_j' \tau_j})$
      $X_{j+1} = X_j(I - \mathbf{u}_j \mathbf{p}_j')$
    end if
  end for
```

and Cristianini 2004) 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. The deflation procedure is illustrated in pseudocode in Algorithm 4, for a detailed review on deflation we refer the reader to Shawe-Taylor and Cristianini (2004).

5 Alternate formulations

For the sake of completeness we compare our above solution to two alternative approaches.[9] In the first, we formulate the SCCA minimisation as a quadratic program using the CVX[10] Matlab toolbox, given in Algorithm 6, where we are able to observe that we are alternating between the ML-primal and ML-dual optimisation problems.

In the second approach, we provide an alternative sparse (LARS-based)[11] solution for the SCCA problem. Here again we alternate between the ML-primal and ML-dual optimisations problems. Furthermore, the LARS-based formulation has slightly more differences from Algorithm 6 (as well as our original solution) as we are no longer able to uphold the $\|\mathbf{e}\|_\infty \leq 1$ constraint as well as moving $\mu\|\mathbf{w}\|_1, \gamma\|\tilde{\mathbf{e}}\|_1$ from the optimisation objective to its constraints as $\|\mathbf{w}\|_1 \leq \mu$, $\|\mathbf{e}\|_1 \leq \gamma$ respectively. The latter has been shown by Tibshirani (1994) to be equivalent. We highlight to the reader that both alternate approaches also require the setting of k.

[9] We were unable to compare to the primal sparse CCA method of Torres et al. (2007) as it required a commercial license of MOSEK.

[10] http://www.stanford.edu/~boyd/cvx/

[11] We use the implementation by Karl Skoglund, IMM, DTU, kas@imm.dtu.dk.

Algorithm 5 The SCCA algorithm—Quadratic program using CVX

```
repeat
    $\mathbf{y} = K\mathbf{e}$
    cvx_begin
    variable $\mathbf{w}(m)$
    minimize($\|X'\mathbf{w} - \mathbf{y}\|^2 + \mu\|\mathbf{w}\|_1$)
    cvx_end

    $\mathbf{y} = X'\mathbf{w}$
    cvx_begin
    variable $\mathbf{e}(\ell)$
    minimize($\|\mathbf{y} - K\mathbf{e}\|^2 + \gamma\|\tilde{\mathbf{e}}\|_1$)
    subject to
    $\mathbf{e}_j = 1$
    $\|\mathbf{e}\|_\infty \le 1$
    cvx_end
until convergence over $\mathbf{e}$, $\mathbf{w}$
```

Algorithm 6 The SCCA algorithm—Alternate solver using LARS

```
repeat
    $\mathbf{y} = K\mathbf{e}$
    $\min_{\mathbf{w}} \|X'\mathbf{w} - \mathbf{y}\|$ s.t. $\|\mathbf{w}\|_1 \le \mu$
    $\mathbf{y} = X'\mathbf{w}$
    $\min_{\mathbf{e}} \|\mathbf{y} - K\mathbf{e}\|$ s.t. $\|\mathbf{e}\|_1 \le \gamma$
until convergence over $\mathbf{e}$, $\mathbf{w}$
```

5.1 Simulated data comparison

We construct simulated data by generating a 2 dimensional background cluster (background noise) constituting of 200 samples drawn independently from a uniform distribution over a 10×10 cube centered at the origin. We then proceed to generated a paired 2 dimensional cluster, each constituting of 20 samples drawn independently from Gaussian distributions centered around $\{(4, 2)\}$ and $\{(-2, 2)\}$ for the two views respectively. We use a Gaussian kernel with the smoothing parameter set to $\sigma = 2$ (arbitrarily set). Finally, in order to maintain similarity between the executions of the algorithms, we set γ and μ as illustrated in Sect. 4.2 for all three approaches. We compute the optimisation value, for all methods, as specific in (4) (after re-normalising the resulting weight ML-primal-dual weight vectors), i.e.

$$f = \|X'\mathbf{w} - K\mathbf{e}\|^2 + \mu\|\mathbf{w}\|_1 + \gamma\|\mathbf{e}\|_1.$$

In the following comparison we run the SCCA, SCCA CVX and SCCA LARS approaches on the simulated data and list in Table 1 their performance, averaged for all k. We are able to observe that SCCA is significantly computationally faster (for the case of nonlinear kernel for ML-dual) as well as resulting in significantly smaller objective function value and a higher correlation value when compared to the alternate CVX approach. Finally, the correlation and objective values achieved by the SCCA LARS approach indicate that it

Table 1 Comparison of SCCA to two alternate approaches of SSCA CVX and SCCA LARS. We list the run time (seconds), number of elements selected for the ML-primal **w** and ML-primal **e**, optimisation objective value and the correlation value

Method	Run-time (s)	# **w**	# **e**	f	Correlation value
SCCA	0.0052	1	1	0.4262	0.2371
SCCA CVX	2.9709	2	120	53.51	0.1161
SCCA LARS	93.750	2	118	4.175e+7	1.0000

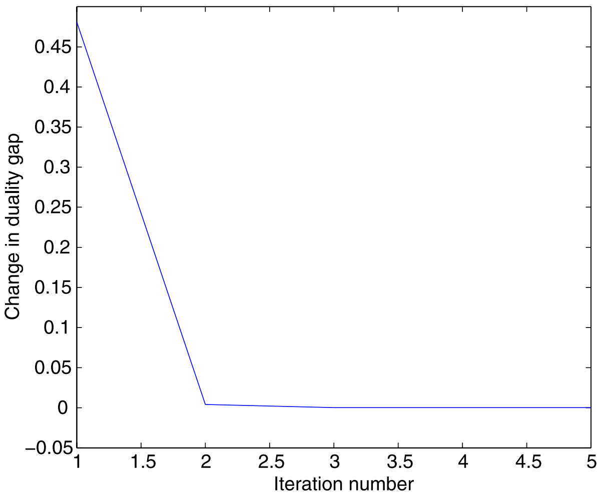

Fig. 1 We visualise the average change in the duality gap for each of the SCCA algorithm iterations on the simulated data. As anticipated this converges to zero while the algorithm converges to a solution

has over-fitted on the simulated-data. In the experiments section we compare and evaluate between SCCA and its alternate methods on a mate-retrieval task. Finally, we demonstrate the convergence properties of the proposed SCCA algorithm by visualising the duality-gap in Fig. 1 where we are able to observe how the average change in the gap convergences to zero as the algorithm convergences to its solution.

6 Data description and experimental setup

In the following experiments we use data from two datasets;

- The Danish-German corpus from the europal dataset (Koehn 2005)[12] where we have a total of 150 samples, consisting of aligned documents, with 12,679 Danish features and 26,028 German features.
- Two paired English-French and English-Spanish corpora from the jrc-acquis dataset (Ralf et al. 2006). The English-French corpus consists of 300 samples with 2,637 English features and 2,951 French features while the English-Spanish corpus consists of 1,000 samples with 40,629 English features and 57,796 Spanish features.

[12] http://people.csail.mit.edu/~koehn/publications/europarl.ps

The features represent the number of words in each language. The corpora are preprocessed into a Term Frequency Inverse Document Frequency (TFIDF) representation followed by zero-meaning (centring) and normalisation. In our lingual based experiments the linear kernel was used for the dual view.

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 paired language is considered relevant. In the following experiments the results are an average of retrieving the mate for both language 1 and language 2 and has 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.

7 Experiments

7.1 Hyperparameter validation

In the following section we validated, on the jrc-acquis data, our approach for automatically determining the regularisation parameter (hyper-parameter) μ (or alternatively γ). The SCCA problem

$$\min_{\mathbf{w},\mathbf{e}} \|X'\mathbf{w} - K\mathbf{e}\|^2 + \mu\|\mathbf{w}\|_1 + \gamma\|\tilde{\mathbf{e}}\|_1, \tag{7}$$

subject to $\|\mathbf{e}\|_\infty = 1$ can be simplified to a general LASSO solver by removing the optimisation over $\mathbf{e}$, resulting in

$$\min_{\mathbf{w}} \|X'\mathbf{w} - \mathbf{k}\|^2 + \mu\|\mathbf{w}\|_1,$$

where, given our paired data, $\mathbf{k}$ is the inner product between the query and the training samples and X is the second paired data samples. This simplified formulation is trivially solved by Algorithm 1 by ignoring the loops that adapt $\mathbf{e}$. The simplification of (7) allows us to focus on showing that μ is close to optimal, which is also true for γ, and therefore omitted.

The hyper-parameters control the level of sparsity. Therefore, we test the level of sparsity as a function of the hyper-parameter value. We proceed by creating a new document d^* from

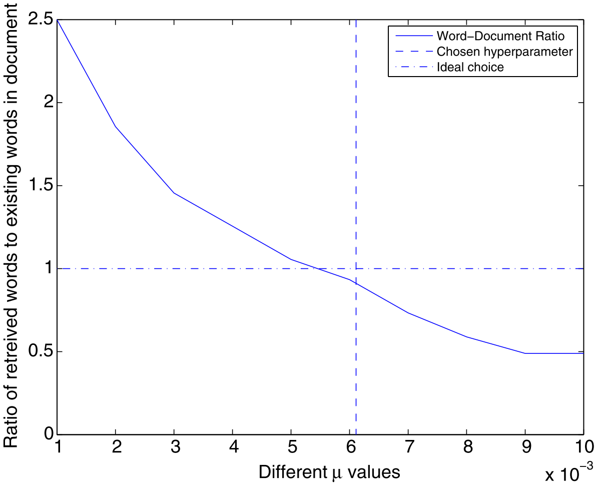

Fig. 2 Document generation for the English-French corpus (visualisation for a single query): We plot the ratio of total number of selected words to the total number of words in the original document. *The horizontal line* defines the optimal choice where the total number of selected words is identical to the total number of words in the original document. *The vertical line* represents the result using the automatic setting of the hyper-parameter. We are able to observe that the automatic selection of μ is a good approximation for selecting the level of sparsity

Table 2 French-English Corpus: The ratio of the total number of selected words to the actual total number of words in the paired test document, averaged over all queries. The optimal average ratio if we always generate an 'ideal' document is 1

	Average selection ratio
Automatic setting of μ	1.01 ± 0.54
Non-sparse method	28.15 ± 15.71

a paired language that best matches our query[13] and observe how the change in μ affects the total number of words being selected. An "ideal" μ would generate a new document, in the paired language, and select an equal number of words in the query's actual paired document. Recall that the data has been mean corrected (centred) and therefore no longer sparse.

We set μ to be in the range of $[0.001, \dots, 1]$ with an increment of 0.001 and use a leave-paired document-out routine for the English-French corpus, which is repeated for all 300 documents. Figure 2 illustrates, for a single query, the effective change in μ on the level of sparsity. We plot the ratio of the total number of selected words to the total number of words in the original document. An ideal choice of μ would choose a ratio of 1 (the horizontal lines) i.e. create a document with exactly the same number of words as the original document or in other words select a μ such that the cross would lie on the plot. We are able to observe that the method for automatically choosing μ (the vertical line) is able to create a new document with a close approximation to the total number of words in the original document.

In Table 2 we are able to show that the average ratio of total number of selected words for each document generated in the paired language is very close to the ideal level of sparsity, while a non-sparse method (as expected) generates a document with an average of ≈28 times the number of words from the original document. Now that we have established the automatic setting of the hyper-parameters, we proceed in testing how 'good' the selected words in the form of mate-retreiveal experiments.

[13] I.e. given a query in French we want to generate a document in English that best matches the query. The generated document can then be compared to the actual paired English document.

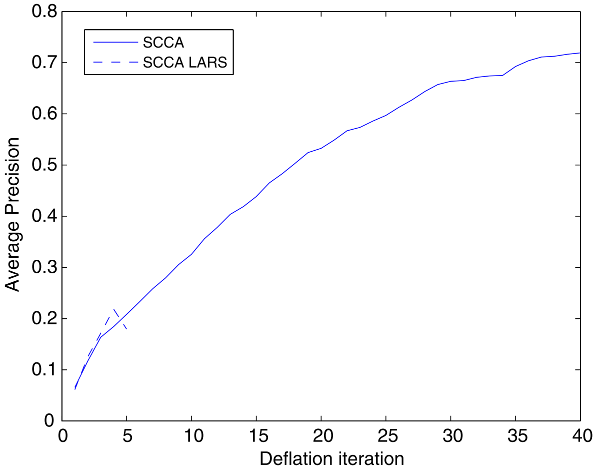

Fig. 3 We plot the average precision (y axis) on the 100 test documents as a function of the deflation iteration (x axis) for up to the maximum of 40 components. As anticipated, increasing the deflation steps improves on the average precision since a richer semantic space is constructed. SCCA LARS became numerically unstable after iteration 5 and as a result did not converge

Table 3 We give the averaged average precision (as detailed in Fig. 3) and reported the number of features used #{**w**, **e**} at the last deflation step across all the computed primal and dual weight vectors. Finally, we list the run-time for each deflation step for a chosen optimal k

Method	(averaged) Average precision	# **w**	# **e**	Run-time (s)
SCCA	0.4915 ± 0.0325	163.5	40	0.9046
SCCA LARS	0.1273 ± 0.0340	19.2	50	0.7442

7.2 Sparse CCA for mate-retrieval

In this section we discuss an experiment where we use the europal German-Danish paired bilingual corpus to evaluate our SCCA solver to the proposed alternatives on a mate retrieval task. Due to the large number of features we were unable to evaluate the SCCA CVX as the resulting quadratic program was too computationally exhaustive to run in practice.

We randomly split the German-Danish samples into 50 training and 100 testing documents and use the procedure outlined in Sect. 7.1 to set the hyper-parameters for both SCCA and SCCA LARS. The reported results are an averaged over 10 repetitions of randomly splitting the data into train and test sets. Furthermore, we highlight that we use the exact same procedure for the SCCA LARS approach as with our proposed algorithm (i.e. we run Algorithm 4 where *SCCA_Algorithm* is replaced with Algorithm 6). Finally, we select k by transversing, in each deflation iteration, through all possible k values (excluding those previously selected) and choosing k with the associated smallest objective function value. This is done for both approaches. Due to the kernel matrix rank we limit the maximum number of deflation iterations to 40. We highlight that Algorithm 4 with SCCA LARS was not able to converge beyond 5 deflation iterations as the LASSO solver became numerically unstable.

Our results are given in Fig. 3 and in Table 3 where we are able to demonstrate that our sparse CCA active set approach, despite being slightly slower in run-time per each deflation step, is able to significantly outperform the SCCA LARS alternative as well as being more numerically stable. In Table 3 we report the average number of features (both primal and dual) at the final deflation step. It is interesting to observe that SCCA LARS makes use

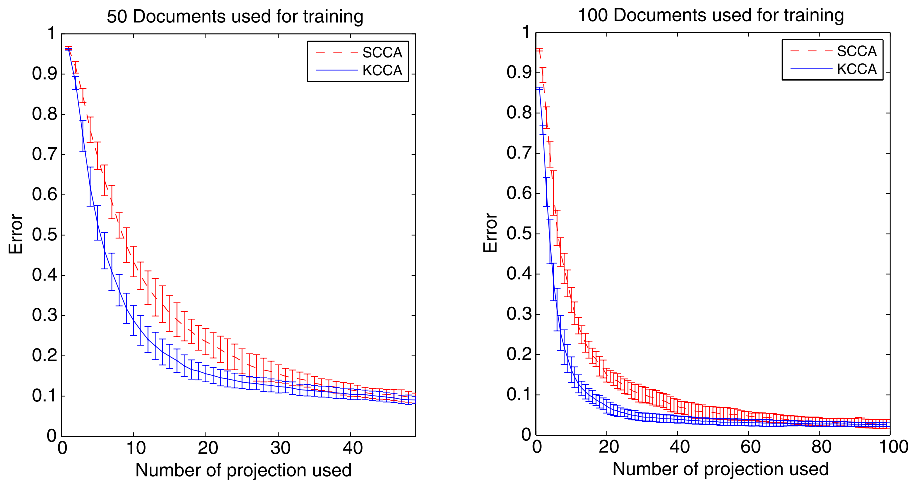

Fig. 4 English-French: The average precision error $(1 - p)$ with ± standard division error bars for SCCA and KCCA for different number of projections used for the mate-retrieval task. *The left figure* is for 50 training and 250 testing documents while *the right figure* is for 100 training and 200 testing documents

of all 50 dual features[14] at iteration 5 (as detailed above, the method became numerically unstable beyond this point) whereas our SCCA solver still only used 40 dual features[15] at final (40th) iteration. Both approaches resulted with a very sparse primal weight vector from the original 12,679 Danish and 26,028 German features.

7.3 KCCA–SCCA comparison

In previous studies (Vinokourov et al. 2003; Hardoon and Shawe-Taylor 2003; Hardoon et al. 2003; Szedmak et al. 2007) KCCA has been shown to work well for mate-retrieval therefore in the following experiment we compare KCCA to SCCA on the mate-retrieval task. The best test performance for the KCCA regularisation parameter for the paired corpora was found to be 0.03. We used this value to ensure that KCCA was not at a disadvantage since SCCA had no parameters to tune. Finally, we adopt a simplistic strategy of picking the values of k in numerical order $k = 1, \dots, \ell$.

We start by giving the results for the English-French mate-retrieval as shown in Fig. 4. The left plot depicts the average precision (± standard deviation) when 50 documents are used for training and the remaining 250 are used as test queries. The right plot in Fig. 4 gives the average precision (± 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 (average plotted in Fig. 5) for either ML 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 2,794 for both languages) to learn the common semantic space. For example,

[14] Averaged across $\mathbf{e}_i^{lars}$, $\mathbf{w}_i^{lars}$ for $i = 1, \dots, 5$.

[15] Averaged across $\mathbf{e}_i^{scca}$, $\mathbf{w}_i^{scca}$ for $i = 1, \dots, 40$.

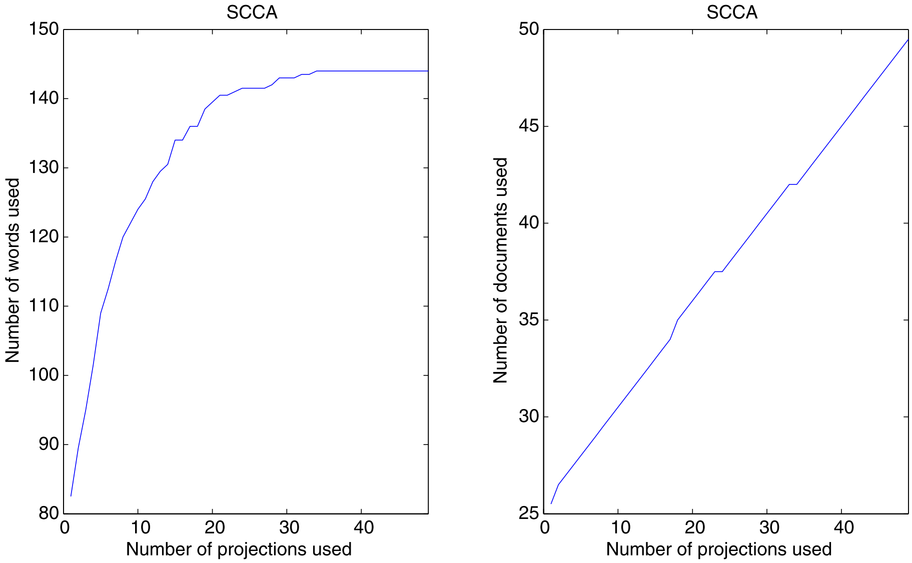

Fig. 5 English-French: Level of Sparsity—the following figure is an extension of Fig. 4 which uses 50 documents for training. *The left figure* plots the average number of words used while *the right figure* plots the average number of documents used with the number of projections. For reference, KCCA uses all the words (average of 2794) and documents (50) for all number of projections

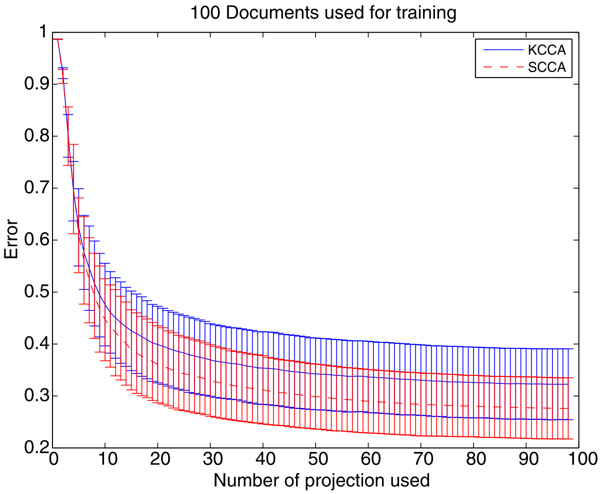

Fig. 6 English-Spanish: The average precision error $(1 - p)$ with ± standard division error bars of SCCA and KCCA for different number of projections used for the mate-retrieval task. We use 100 documents for training and 900 for testing documents

following the left plot in Fig. 4 and the additional plots in Fig. 5 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 2,794 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 Fig. 6 where we are clearly able to observe SCCA outperforming

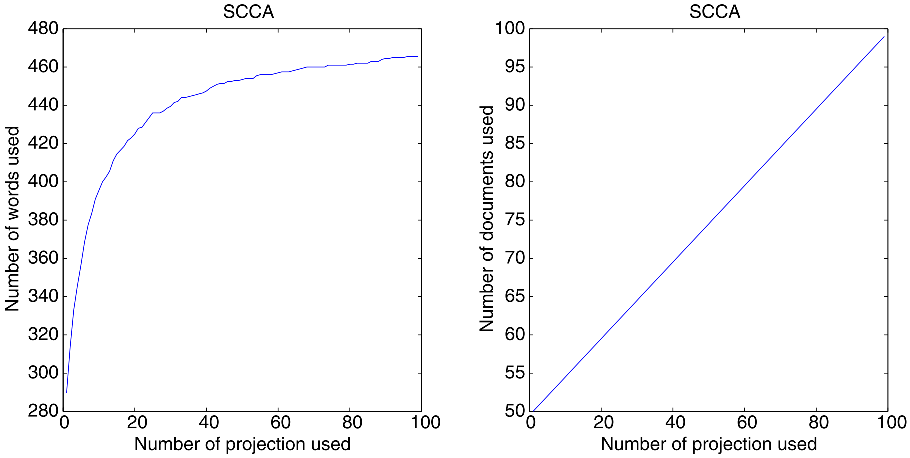

Fig. 7 English-Spanish: Level of Sparsity—the following figure is an extension of Fig. 6 which uses 100 documents for training. *The left figure* plots the average number of words used and while *the right figure* plots the average number of documents used with increasing number of projections. For reference, KCCA uses all the words (average of 49,212) and documents (100) for all number of projections

KCCA throughout. We believe this to be a good example of when too many features hinder the learnt semantic space, also explaining the difference in the results obtained from the English-French corpus as the number of features are significantly smaller in that case. The average level of SCCA sparsity is plotted in Fig. 7. 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.

Despite these already impressive results our intuition is that even better results are attainable if the hyper-parameters would be tuned to give optimal results. The question of hyper-parameter optimality is left for future research. Although, it seems that the main gain of SCCA is sparsity and interpretability of the features.

8 Conclusions

Despite being introduced in 1936, CCA has proven to be an inspirational methodology for 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 sparse CCA method formulated as a convex least squares approach. We also provide a different perspective of solving CCA by using a ML primal-dual formulation which focuses on the scenario when one is interested in (or limited to) a ML-primal representation for the first view while having a ML-dual representation for the second view. A greedy optimisation algorithm is derived. Furthermore, we give two alternate solutions for SCCA; the first as a quadratic program and the second as a LARS based solver.

The method is demonstrated on a bi-lingual English-French and English-Spanish paired corpora for mate retrieval. 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 ML primal-dual views.

The paper's reason d'être is to propose a new efficient algorithm for solving the sparse CCA problem. We believe that while addressing this problem new and interesting questions which need to be addressed have surfaced

- Theoretically justified approach to compute the hyperparameters μ, γ.
- Extending SCCA to a ML primal-primal (ML dual-dual) framework.
- Theoretical analysis of consistency.

We believe this work to be an initial stage for a new sparse framework to be explored and extended.

Acknowledgements David R. Hardoon would like to acknowledge support by the EPSRC project Le Strum, EP-D063612-1. The authors' would also like to thank Zakria Hussain and Nic Schraudolph for insightful discussions. This publication only reflects the authors views.

References

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

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

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

Chen, S. S., Donoho, D. L., & Saunders, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, *20*(1), 33–61.

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., & Lanckriet, G. (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, *49*(3), 434–448.

Dhanjal, C., Gunn, S. R., & Shawe-Taylor, J. (2006). Sparse feature extraction using generalised partial least squares. In *Proceedings of the IEEE international workshop on machine learning for signal processing* (pp. 27–32).

Friman, O., Borga, M., Lundberg, P., & Knutsson, H. (2001a). A correlation framework for functional MRI data analysis. In *Proceedings of the 12th Scandinavian conference on image analysis*, Bergen, Norway, June 2001.

Friman, O., Carlsson, J., Lundberg, P., Borga, M., & Knutsson, H. (2001b). Detection of neural activity in functional MRI using canonical correlation analysis. *Magnetic Resonance in Medicine*, *450*(2), 323–330.

Fukumizu, K., Bach, F. R., & Gretton, A. (2007). Consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, *8*, 361–383.

Fyfe, C., & Lai, P. L. (2000). ICA using kernel canonical correlation analysis. In *Proc. int. workshop on independent component analysis and blind signal separation (ICA 2000)* (pp. 279–284).

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., & Shawe-Taylor, J. (2007). *Sparse canonical correlation analysis* (Technical report). UK: University College London.

Hardoon, D. R., & Shawe-Taylor, J. (2009). Convergence analysis of kernel canonical correlation analysis: Theory and practice. *Machine Learning*, *74*(1), 23–38.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2003). *Canonical correlation analysis; an overview with application to learning methods* (Technical Report CSD-TR-03-02). Royal Holloway University of London.

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). Berlin: Springer.

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.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London/Boca Raton: Chapman & Hall/CRC Press.

Heiler, M., & Schnor, C. (2006). Learning sparse representations by non-negative matrix factorization and sequential cone programming. *Journal of Machine Learning Research*, *7*, 1385–1407.

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.

Koehn, P. (2005). Europarl: A multilingual corpus for evaluation of machine translation. In *Conference proceedings: the tenth machine translation summit* (pp. 79–86).

Lai, P. L., & Fyfe, C. (2000). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, *10*(5), 365–377.

Lee, H., Battle, A., Raina, R., & Ng, A. Y. (2006). Efficient sparse coding algorithms. In *Proceedings of the 20th annual conference on neural information process systems (NIPS)*.

Moghaddam, B., Weiss, Y., & Avidan, S. (2006). Spectral bounds for sparse pca: Exact and greedy algorithms. In *Neural information processing systems (NIPS 06)*.

Ralf, S., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., & Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th international conference on language resources and evaluation (LREC'2006)*.

Roth, V. (2004). The generalized lasso. *IEEE Transactions on Neural Networks*, *15*(1), 16–28.

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

Sriperumbudur, B. K., Torres, D., & Lanckriet, G. (2007). Sparse eigen methods by d.c. programming. In C. Brodley & A. Danyluk (Eds.), *Proceedings of 2nd international conference on machine learning* (pp. 831–838). San Mateo: Morgan Kaufmann.

Szedmak, S., De Bie, T., & Hardoon, D. R. (2007). A metamorphosis of canonical correlation analysis into multivariate maximum margin learning. In *15th European symposium on artificial neural networks (ESANN)*.

Tibshirani, R. (1994). *Regression shrinkage and selection via the lasso* (Technical report). University of Toronto.

Torres, D., Turnbull, D., Barrington, L., & Lanckriet, G. (2007). Identifying words that are musically meaningful. In *Proceedings of the 8th international conference on music information retrieval*.

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.

Weston, J., Elisseeff, A., Scholkopf, B., & Tipping, M. (2003). Use of the zero norm with linear models and kernel method. *Journal of Machine Learning Research*, *3*, 1439–1461.

Zou, H., Hastie, T., & Tibshirani, R. (2004). *Sparse principal component analysis* (Technical report). Statistics department, Stanford University.