Sparse Canonical Correlation Analysis
(Double Barrelled LASSO)

David R. Hardoon and John Shawe-Taylor
D.Hardoon@cs.ucl.ac.uk

University College London
Department of Computer Science
Centre for Computational Statistics & Machine Learning

Workshop on Learning from Multiple Sources, NIPS
Whistler, Vancouver, 13 December 2008
Talk Outline

- Brief review of Canonical Correlation Analysis
- Motivation
- (Machine Learning) Primal - Dual formulation
- Sparse Canonical Correlation Analysis
- LASSO Analogy
- Experiments
Brief Introduction to Canonical Correlation Analysis

• We are trying to find a linear combination such that the correlation between the two views $X_a, X_b$ are maximised (primal cca)

$$
\max_{w_a, w_b} \rho = \frac{w'_a C_{ab} w_b}{\sqrt{w'_a C_{aa} w'_a w'_b C_{bb} w_b}}.
$$

• Kernel CCA - finding a non-linear combination (dual cca) by letting $w = X \alpha$

$$
\max_{\alpha_a, \alpha_b} \rho = \frac{\alpha'_a K_a K_b \alpha_b}{\sqrt{\alpha'_a K_a^2 \alpha_a \alpha'_b K_b^2 \alpha_b}}.
$$
Motivation
Motivation

• Why have a sparse CCA formulation?
Motivation

• Why have a sparse CCA formulation?

• Why have a machine learning primal-dual sparse CCA formulation?
Motivation

• Why have a sparse CCA formulation?
  • Why have a machine learning primal-dual sparse CCA formulation?

• Largely intuitive when faced with real-world problems combined with the need to understand or interpret the found solutions.
Motivation

- Why have a sparse CCA formulation?
  - Why have a machine learning primal-dual sparse CCA formulation?
- Largely intuitive when faced with real-world problems combined with the need to understand or interpret the found solutions.
- Consider the following cases;
Motivation

• Why have a sparse CCA formulation?
  • Why have a machine learning primal-dual sparse CCA formulation?
• Largely intuitive when faced with real-world problems combined with the need to understand or interpret the found solutions.
• Consider the following cases;
  • Enzyme prediction
Motivation

- Why have a sparse CCA formulation?
  - Why have a machine learning primal-dual sparse CCA formulation?
- Largely intuitive when faced with real-world problems combined with the need to understand or interpret the found solutions.
- Consider the following cases;
  - Enzyme prediction
  - Bilingual analysis
Motivation

• Why have a sparse CCA formulation?
  • Why have a machine learning primal-dual sparse CCA formulation?

• Largely intuitive when faced with real-world problems combined with the need to understand or interpret the found solutions.

• Consider the following cases;
  • Enzyme prediction
  • Bilingual analysis
  • Brain analysis
Motivation

• Why have a sparse CCA formulation?
  • Why have a machine learning primal-dual sparse CCA formulation?

• Largely intuitive when faced with real-world problems combined with the need to understand or interpret the found solutions.

• Consider the following cases;
  • Enzyme prediction
  • Bilingual analysis
  • Brain analysis
  • etc...
Primal - Dual Problem

• We are able to reformulate the CCA problem as a (machine learning) primal-dual representation. This will allow us to learn the correlation between words in one language and documents in another.

\[
\begin{align*}
\max_{w_a,w_b} & \quad \frac{w'_a X_a X'_b w_b}{\sqrt{w'_a X'_a X'_a w_a w_b X_b X'_b w_b}} \\
\max_{w_a,e} & \quad \frac{w'_a X_a X'_b X_b e}{\sqrt{w'_a X'_a X'_a w_a e' X'_b X_b X'_b X_b e}} \\
\max_{w_a,e} & \quad \frac{w'_a X_a K_b e}{\sqrt{w'_a X'_a X'_a w_a e' K_b^2 e}},
\end{align*}
\]
Primal - Dual Problem

- We are able to reformulate the CCA problem as a (machine learning) primal-dual representation. This will allow us to learn the correlation between words in one language and documents in another.

\[
\max_{w_a, e} w'_a X_a K_b e \\
\left\| X_a w_a \right\|^2 = 1 \quad \left\| K_b e \right\|^2 = 1
\]
Sparse CCA

- Maximising the correlation between two vectors can be viewed as minimising the angle between them (Breiman & Friedman 1985, Hastie & Tibshirani 1990)

\[
\max_{\mathbf{w}_a, \mathbf{e}} \mathbf{w}'_a \mathbf{X}_a \mathbf{K}_b \mathbf{e}
\]

- Subject to:
  \[\| \mathbf{X}_a \mathbf{w}_a \|^2 = 1 \quad \| \mathbf{K}_b \mathbf{e} \|^2 = 1\]

- Not there yet...
Sparse CCA

• Maximising the correlation between two vectors can be viewed as minimising the angle between them (Breiman & Friedman 1985, Hastie & Tibshirani 1990)

\[
\min_{\mathbf{w}, \mathbf{e}} \| X' \mathbf{w} - K \mathbf{e} \|^2
\]

• Subject to: 

\[ \| X_a \mathbf{w}_a \|^2 = 1 \quad \| K_b \mathbf{e} \|^2 = 1 \]

• Not there yet...
Sparse CCA

- Maximising the correlation between two vectors can be viewed as minimising the angle between them (Breiman & Friedman 1985, Hastie & Tibshirani 1990)

\[
\min_{w,e} \| X'w - Ke \|^2
\]

- Subject to:

\[
\| Ke \|^2 = 1
\]

- Not there yet...
Sparse CCA

- Maximising the correlation between two vectors can be viewed as minimising the angle between them (Breiman & Friedman 1985, Hastie & Tibshirani 1990)

\[
\min_{\mathbf{w},\mathbf{e}} \|X'\mathbf{w} - K\mathbf{e}\|^2
\]

- Subject to:

\[
\|\mathbf{e}\|_\infty = 1
\]

- Not there yet...
• Note that as $\mathbf{e} = \{e_1, \ldots, e_k, \ldots, e_\ell\}$

• We are able to drop $\|\mathbf{e}\|_\infty = 1$ by setting $e_k = 1$ for a given $k = 1, \ldots, \ell$
• Note that as \( \mathbf{e} = \{e_1, \ldots, e_k, \ldots, e_\ell\} \)

• We are able to drop \( \|\mathbf{e}\|_\infty = 1 \) by setting \( e_k = 1 \) for a given \( k = 1, \ldots, \ell \)

• We impose sparsity using the 1-norm on the ml-primal weights \( \mathbf{w} \) and on the remaining ml-dual weights \( \hat{\mathbf{e}} = \{e_1, \ldots, e_{k-1}, e_{k+1}, \ldots, e_\ell\} \)
• What do we mean by setting $e_k$

• We apply the following naive procedure

• solve for $\mathbf{e}^1 = (1, e_2, e_3, \ldots, e_\ell)$ deflate

• solve for $\mathbf{e}^2 = (e_1, 1, e_3, \ldots, e_\ell)$ deflate, etc.

  \[\vdots\]

  $\mathbf{e}^\ell = (e_1, e_2, \ldots, e_{\ell-1}, 1)$

• A alternative would be to solve sparse KPCA $k = \max_i \frac{I_i'K^2I_i}{I_i'KI_i}$ after each deflation where $\mathbf{I}$ is the identity matrix
• Note that as $\mathbf{e} = \{e_1, \ldots, e_k, \ldots, e_\ell\}$

• We are able to drop $\|\mathbf{e}\|_\infty = 1$ by setting $e_k = 1$ for a given $k = 1, \ldots, \ell$

• We impose sparsity using the 1-norm on the ml-primal weights $\mathbf{w}$ and on the remaining ml-dual weights $\hat{\mathbf{e}} = \{e_1, \ldots, e_{k-1}, e_{k+1}, \ldots, e_\ell\}$
• Note that as $\mathbf{e} = \{e_1, \ldots, e_k, \ldots, e_\ell\}$

• We are able to drop $\|\mathbf{e}\|_\infty = 1$ by setting $e_k = 1$ for a given $k = 1, \ldots, \ell$

• We impose sparsity using the 1-norm on the ml-primal weights $\mathbf{w}$ and on the remaining ml-dual weights $\hat{\mathbf{e}} = \{e_1, \ldots, e_{k-1}, e_{k+1}, \ldots, e_\ell\}$

• Resulting in our final (convex) optimisation being

$$
\min_{\mathbf{w}, \mathbf{e}} \|X'\mathbf{w} - K\mathbf{e}\|_2^2 + \mu \|\mathbf{w}\|_1 + \gamma \|\hat{\mathbf{e}}\|_1
$$
• Note that as \( \mathbf{e} = \{e_1, \ldots, e_k, \ldots, e_\ell\} \)

• We are able to drop \( \|\mathbf{e}\|_\infty = 1 \) by setting \( e_k = 1 \) for a given \( k = 1, \ldots, \ell \)

• We impose sparsity using the 1-norm on the ml-primal weights \( \mathbf{w} \)
  and on the remaining ml-dual weights \( \hat{\mathbf{e}} = \{e_1, \ldots, e_{k-1}, e_{k+1}, \ldots, e_\ell\} \)

• Resulting in our final (convex) optimisation being

\[
\min_{\mathbf{w, e}} \|X'\mathbf{w} - Ke\|_2^2 + \mu \|\mathbf{w}\|_1 + \gamma \|\hat{\mathbf{e}}\|_1
\]

• Great - but how do we solve this in practice?
Algorithmic Solution

• We limit ourselves to the positive spectrum of $\mathbf{e}$ (we are interested in positive contributors to the semantic context)

• We represent $\mathbf{w} = \mathbf{w}^+ - \mathbf{w}^-$

• Result in the Lagrangian

$$\mathcal{L} = (\mathbf{w}^+ - \mathbf{w}^-)' \mathbf{X} \mathbf{X}' (\mathbf{w}^+ - \mathbf{w}^-) + \mathbf{e}' \mathbf{K}^2 \mathbf{e} - 2(\mathbf{w}^+ - \mathbf{w}^-)' \mathbf{X} \mathbf{K} \mathbf{e} - \alpha^+ \mathbf{w}^+ - \alpha^- \mathbf{w}^- + \mu(\mathbf{w}^+ + \mathbf{w}^-)' \mathbf{j} + \gamma \mathbf{e}' \mathbf{j} - \beta' \mathbf{e}$$

• Subject to $\alpha^+ \geq 0 \quad \alpha^- \geq 0 \quad \beta \geq 0$
Setting Sparsity

- Taking derivatives and equating to zero gives

\[
\frac{\partial L}{\partial w^+} = 2XX'(w^+ - w^-) - 2XKe - \alpha^+ + \mu j
\]

\[
\frac{\partial L}{\partial w^-} = -2XX'(w^+ - w^-) + 2XKe - \alpha^- + \mu j
\]

- Notice that adding the two equations give the following property

\[
\alpha^- = 2\mu j - \alpha^+ \\
\downarrow \\
0 \leq \alpha^- \leq 2\mu j
\]

\[
\alpha^+ = 2\mu j - \alpha^- \\
\downarrow \\
0 \leq \alpha^+ \leq 2\mu j
\]

- We use the above bound as an indication of which \( w_i \) need to be updated
- Similarly, we only update \( e_i \) that has a corresponding \( \beta_i < 0 \)
LASSO Analogy

- The SCCA problem can be simplified to a general Least Absolute Shrinkage and Selection Operator (LASSO) problem by removing the optimisation over $e$

\[
\min_{w,e} \|X'w - Ke\|^2 + \mu\|w\|_1 + \gamma\|\tilde{e}\|_1,
\]

\[
\min_w \|X'w - k\|^2 + \mu\|w\|_1,
\]

- (Tibshirani 1994, Chen et. al 1999, etc.)

- Hence the ‘double barrelled LASSO’
Experiments: Mate retrieval

- English - French
  - 300 Samples (documents)
  - 2637 English features (words), 2951 French features (words)

- English - Spanish
  - 1000 Samples
  - 40629 English features (words), 57796 Spanish features (words)

- Both processed with TFIDF followed by zero-meaning (centring)
We compare KCCA to SCCA for a mate retrieval procedure

\[ p = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{I_j} \]

- Compute average precision
- \( M \) is the number of query documents
- KCCA regularisation set a-priori (yes, we cheated)
Hyperparameter Parametrisation

• We propose an automatic approach for setting sparsity
Hyperparameter Parametrisation

• We propose an automatic approach for setting sparsity

• During the initial step \[ 2XX'w - 2X'Ke - \alpha^+ + \mu j = 0 \]
Hyperparameter Parametrisation

• We propose an automatic approach for setting sparsity

• During the initial step  \( 2XX'w - 2X'Ke - \alpha^+ + \mu j = 0 \)
We propose an automatic approach for setting sparsity

\[ 2XX'w - 2X'Ke - \alpha^+ + \mu j = 0 \]

During the initial step

\[ \mu = \frac{1}{\ell} \sum_{i=1}^{\ell} [2X'Ke]_i \]

Allowing us to set
Hyperparameter Parametrisation

- We propose an automatic approach for setting sparsity
  
  - During the initial step
    \[ 2XX'w - 2X'Ke - \alpha^+ + \mu j = 0 \]
  
  - Allowing us to set
    \[ \mu = \frac{1}{\ell} \sum_{i=1}^{\ell} [||2X'Ke||]_i \]

- The simplification of SCCA allows us to focus on showing that \( \mu \) is close to optimal, which is also true for \( \gamma \).
Hyperparameter Parametrisation

• We propose an automatic approach for setting sparsity

• During the initial step

\[ 2X'w - 2X'Ke - \alpha^+ + \mu j = 0 \]

• Allowing us to set

\[ \mu = \frac{1}{\ell} \sum_{i=1}^{\ell} [\|2X'Ke\|]_i \]

• The simplification of SCCA allows us to focus on showing that \( \mu \) is close to optimal, which is also true for \( \gamma \).

• We compute the level of sparsity as a function of the hyperparameter value.
The effective change in $\mu$ on the level of sparsity

Ratio of retrieved words to existing words in document

Different $\mu$ values

Word–Document Ratio
Chosen hyperparameter
Ideal choice

$x \times 10^{-3}$
Table 1: 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Average Selection Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic setting of $\mu$</td>
<td>$1.01 \pm 0.54$</td>
</tr>
<tr>
<td>Non-sparse method</td>
<td>$28.15 \pm 15.71$</td>
</tr>
</tbody>
</table>

We now proceed in testing how “good” the selected words are in the form of a mate-retrieval experiment.
French - English Corpus
EF: Words - Documents Retrieval
English - Spanish Corpus
ES: Words - Documents Retrieval

![Graph 1: Number of words used vs. Number of projection used](image1)

![Graph 2: Number of documents used vs. Number of projection used](image2)
Conclusion

• Proposed a novel approach for a sparse primal-dual cca

  • Proposed an automatic approach to the hyperparameter setting

• Open questions and future directions

  • Setting $e_k$ using CCA/KPCA bounds

  • Incorporating the negative spectrum of the dual features
Thank you. -- Questions?