A Least Squares Formulation for Canonical Correlation Analysis

Liang Sun, Shuiwang Ji, and Jieping Ye
Department of Computer Science and Engineering
Arizona State University
Motivation

- Canonical Correlation Analysis (CCA) is commonly used for finding the correlations between two sets of multi-dimensional variables (Hotelling, 1936; Hardoon et al., 2004; Vert & Kanehisa, 2003).
  - Two representations $X \in \mathbb{R}^{d \times n}$ and $Y \in \mathbb{R}^{k \times n}$.
- One popular use of CCA is for supervised learning, in which one view is derived from the data and another view is derived from the class labels.
  - CCA involves an eigenvalue problem, which is computationally expensive to solve.
  - It is challenging to derive sparse CCA models.
CCA is closely related to linear regression ($Y \in \mathbb{R}^{1 \times n}$).

CCA is equivalent to Fisher Linear Discriminant Analysis (LDA) for binary-class problems. Fisher LDA can be formulated as a least squares problem for binary-class problems.

- It can be solved efficiently using conjugate gradient.
- Sparse models can be derived readily using 1-norm regularization (Lasso).

Multivariate linear regression (MLR) is a well-studied technique for regression problems.

- To apply MLR to multi-label classification, one key issue is how to define an appropriate class indicator matrix.

Can we extend their equivalence relationship to the general (multi-label) case?
Main Contributions

- We establish the equivalence relationship between CCA and multivariate linear regression for multi-label problems under a mild condition.
- Based on the equivalence relationship, several CCA extensions including sparse CCA are derived.
- The entire solution path for sparse CCA can be readily computed by Least Angle Regression algorithm (LARS).
- Our experiments confirm the equivalence relationship, and results also show that the performance of these wo models is very close even when the assumption is violated.
Outline

1. Background
2. CCA Versus Multivariate Linear Regression
3. CCA Extensions
4. Experiment
In CCA two different representations of the same set of objects are given, and a projection is computed for each representation such that they are maximally correlated in the dimensionality-reduced space.

Two representations $X \in \mathbb{R}^{d \times n}$ and $Y \in \mathbb{R}^{k \times n}$.

CCA attempts to maximize the following correlation coefficient w.r.t. $w_x$ and $w_y$:

$$
\rho = \frac{w_x^T X Y^T w_y}{\sqrt{(w_x^T X X^T w_x)(w_y^T Y Y^T w_y)}}. \quad (1)
$$
Background: CCA

- The optimization problem in CCA can be formulated as

\[
\max_{w_x, w_y} \quad w_x^T X Y^T w_y \tag{2}
\]
subject to \( w_x^T X X^T w_x = 1, \quad w_y^T Y Y^T w_y = 1 \).

- Assume \( Y Y^T \) is nonsingular, \( w_x \) can be obtained by computing the eigenvector corresponding to top eigenvalue of the following generalized eigenvalue problem:

\[
X Y^T (Y Y^T)^{-1} Y X^T w_x = \eta X X^T w_x. \tag{3}
\]
Multiple projection vectors under certain orthonormality constraints can be obtained by computing the top $\ell$ eigenvectors of the generalized eigenvalue problem in Eq. (3).

In regularized CCA (rCCA), a regularization term $\lambda I$ with $\lambda > 0$ is added to $XX^T$ to prevent the overfitting and avoid the singularity of $XX^T$. Specifically, rCCA solves the following generalized eigenvalue problem:

$$XY^T(YY^T)^{-1}YX^Tw_x = \eta(XX^T + \lambda I)w_x.$$  \hspace{1cm} (4)
Background: Multivariate Linear Regression

- We are given a training set \( \{(x_i, t_i)\}_{i=1}^n \), where \( x_i \in \mathbb{R}^d \) is the observation and \( t_i \in \mathbb{R}^k \) is the corresponding target. We assume both \( \{x_i\}_{i=1}^N \) and \( \{t_i\}_{i=1}^n \) are centered.

- In MLR, we compute a weight matrix \( W \) by minimizing the following sum-of-squares cost function:

\[
\min_W \sum_{i=1}^n \| W^T x_i - t_i \|_2^2 = \| W^T X - T \|_F^2. \tag{5}
\]

- The optimal weight matrix is given by

\[
W_{LS} = (XX^T)^{\dagger}XT^T. \tag{6}
\]

- To improve its generalization ability, a penalty term based on 2-norm or 1-norm regularization is commonly applied.
In the general multi-label classification, we are given a data set consisting of \( n \) samples \( \{(x_i, y_i)\}_{i=1}^{n} \), where \( x_i \in \mathbb{R}^d \), and \( y_i \) denotes the set of class labels of the \( i \)-th sample.

Assume there are \( k \) labels. The 1-of-\( k \) binary coding scheme \( T \in \mathbb{R}^{k \times n} \) is commonly employed to apply a vector-valued class code to each data point.

- Sample class indicator matrix: \( T_{ij} = 1 \) if \( x_j \) contains the \( i \)-th class label, and \( T_{ij} = 0 \) otherwise.

The solution to the least squares problem depends on the choice of class indicator matrix.
The optimal projection matrix $W_{CCA}$ in CCA consists of the top eigenvectors of

$$(XX^T)^\dagger \left( XY^T (YY^T)^{-1} YX^T \right).$$

The optimal weight matrix in MLR is given by

$$W_{LS} = (XX^T)^\dagger XT^T.$$

Establish the equivalence relationship between these two.
Notations and Definitions

- To simplify the discussion, we define the following matrices:

\[
H = Y^T (YY^T)^{-\frac{1}{2}} \in \mathbb{R}^{n \times k}, \quad (7)
\]
\[
C_{XX} = XX^T \in \mathbb{R}^{d \times d}, \quad (8)
\]
\[
C_{HH} = XHH^T X^T \in \mathbb{R}^{d \times d}, \quad (9)
\]
\[
C_{DD} = C_{XX} - C_{HH} \in \mathbb{R}^{d \times d}. \quad (10)
\]

- Denote SVD of \( X \) by

\[
X = U\Sigma V^T = [U_1, U_2] \text{ diag}(\Sigma_r, 0) [V_1, V_2]^T = U_1 \Sigma_r V_1^T,
\]

where \( r = \text{rank}(X), \) \( U_1 \in \mathbb{R}^{d \times r}, \) \( V_1 \in \mathbb{R}^{n \times r}. \)

- The optimal projection matrix \( W_{CCA} \) of CCA consists of the top eigenvectors of \( C_{XX}^{\dagger} C_{HH}. \)
Computing CCA via Eigendecomposition

- Define $A \in \mathbb{R}^{r \times k}$ as $A = \Sigma_r^{-1}U_1^T XH$.
- Let the SVD of $A$ be $A = P \Sigma_A Q^T$.
- The eigendecomposition of $C_{XX}^\dagger C_{HH}$ can be derived as follows:

$$C_{XX}^\dagger C_{HH} = U_1 \Sigma_r^{-2} U_1^T XHH^T X^T = U_1 \Sigma_r^{-1} A H^T X^T U U^T$$

$$= U \begin{bmatrix} I_r & \Sigma_r^{-1} A \end{bmatrix} \begin{bmatrix} XH^T X & \Sigma_r^{-1} A \end{bmatrix} U^T$$

$$= U \begin{bmatrix} \Sigma_r^{-1} A A^T \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} U^T$$

$$= U \begin{bmatrix} \Sigma_r^{-1} P & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_r \end{bmatrix} \begin{bmatrix} P^T \Sigma_r & 0 \\ 0 & I \end{bmatrix} U^T.$$
The optimal projection matrix in CCA is given by

$$W_{CCA} = U_1 \Sigma_r^{-1} P_\ell, \quad (11)$$

where $P_\ell$ contains the first $\ell$ columns of $P$.

In MLR, we define the class indicator matrix $T$ as

$$T = (YY^T)^{-\frac{1}{2}} Y = H^T,$

and the optimal solution is given by

$$W_{LS} = (XX^T)^\dagger XH = U_1 \Sigma_r^{-2} U_1^T XH = U_1 \Sigma_r^{-1} P \Sigma_A Q^T. \quad (12)$$

Our main results show that all diagonal elements of $\Sigma_A$ are ones provided that $\text{rank}(X) = n - 1$. 
Regularization is commonly used to control the complexity of the model and improve the generalization performance.

Based on the least squares formulation of CCA, we obtain the 2-norm regularized least squares CCA formulation (called “LS-CCA2”) that minimizes the following objective function:

\[ L_2(W, \lambda) = \sum_{j=1}^{k} \left( \sum_{i=1}^{n} (x_i^T w_j - T_{ij})^2 + \lambda \|w_j\|_2^2 \right), \]

where \( W = [w_1, \cdots, w_k] \), and \( \lambda > 0 \) is the regularization parameter.
CCA Extensions: Sparse CCA

- Sparseness can often be achieved by penalizing the $L_1$-norm of the variables.
- The sparse $1$-norm least squares CCA formulation (called “LS-CCA$_1$”) can be derived by minimizing the following objective function:

$$L_1(W, \lambda) = \sum_{j=1}^{k} \left( \sum_{i=1}^{n} (x_i^T w_j - T_{ij})^2 + \lambda \| w_j \|_1 \right).$$

- The optimal $w_j^*$, for $1 \leq j \leq k$, is given by

$$w_j^* = \arg \min_{w_j} \left( \sum_{i=1}^{n} (x_i^T w_j - T_{ij})^2 + \lambda \| w_j \|_1 \right). \quad (13)$$
The optimal $w_j^*$ can be computed equivalently as

$$w_j^* = \arg \min_{\|w_j\|_1 \leq \tau} \sum_{i=1}^{n} (x_i^T w_j - T_{ij})^2.$$  \hspace{1cm} (14)

The solution can be readily computed by the Least Angle Regression algorithm (LARS), which computes the entire solution path for all values of $\tau$, with essentially the same computational cost as fitting the model with a single $\tau$ value.
Experiment – Experimental Setup

- Two types of data used in the experiment:
  - The gene expression pattern image data of *Drosophila*.
  - The scene data set.

- Five methods are investigated:
  - CCA
  - Regularized CCA (rCCA)
  - LS-CCA
  - LS-CCA$_2$
  - LS-CCA$_1$

- Linear SVM is applied for classification for each label after the CCA projection.

- The Receiver Operating Characteristic (ROC) value is computed for each label and the averaged performance over all labels is reported.
Equivalence Relationship

- The assumption \( \text{rank}(X) = n - 1 \) holds in all cases when the data dimensionality \( d \) is larger than the sample size \( n \).
- Our results show that when \( d > n \), all diagonal elements of \( \Sigma_A \) are ones and CCA and LS-CCA achieve the same classification performance, which confirms our theoretical analysis.
Performance Comparison

Table: Comparison of different CCA methods in terms of mean ROC scores. $n_{tot}$ denotes the total number of images in the data set, and $k$ denotes the number of terms (labels). Ten different splittings of the data into training (of size $n$) and test (of size $n_{tot} - n$) sets are applied for each data set. For the regularized algorithms, the value of the parameter is chosen via cross-validation. The proposed sparse CCA model (LS-CCA$_1$) performs the best for this data set.

<table>
<thead>
<tr>
<th>$n_{tot}$</th>
<th>$k$</th>
<th>CCA</th>
<th>LS-CCA</th>
<th>rCCA</th>
<th>LS-CCA$_2$</th>
<th>LS-CCA$_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>863</td>
<td>10</td>
<td>0.542</td>
<td>0.542</td>
<td>0.617</td>
<td>0.619</td>
<td><strong>0.722</strong></td>
</tr>
<tr>
<td>1041</td>
<td>15</td>
<td>0.534</td>
<td>0.534</td>
<td>0.602</td>
<td>0.603</td>
<td><strong>0.707</strong></td>
</tr>
<tr>
<td>1138</td>
<td>20</td>
<td>0.538</td>
<td>0.538</td>
<td>0.609</td>
<td>0.610</td>
<td><strong>0.714</strong></td>
</tr>
<tr>
<td>1222</td>
<td>25</td>
<td>0.540</td>
<td>0.540</td>
<td>0.603</td>
<td>0.605</td>
<td><strong>0.704</strong></td>
</tr>
<tr>
<td>1349</td>
<td>30</td>
<td>0.548</td>
<td>0.548</td>
<td>0.606</td>
<td>0.608</td>
<td><strong>0.709</strong></td>
</tr>
</tbody>
</table>
Sensitivity Study

- We vary the training sample size to investigate LS-CCA and its variants in comparison with CCA.
- $d = 384$ for the gene data set, and $d = 294$ for the scene data set.

Figure: Gene data set

Figure: Scene data set
Figure: The entire collection of solution paths for a subset of the coefficients from the first weight vector $w_1$ on the scene data set. The x-axis denotes the sparseness coefficient, and the y-axis denotes the value of the coefficients (of $w_1$).
Conclusion and Future Work

• **Conclusion**
  - We establish the equivalence relationship between CCA and multivariate linear regression under a mild condition.
  - Several CCA extensions including sparse CCA are proposed based on the equivalence relationship.
  - Experiments confirm the equivalence relationship. Results also show that the performance of CCA and MLR is very close even when the assumption is violated.

• **Future Work**
  - Investigate semi-supervised CCA by incorporating unlabeled data into the CCA framework.
  - Extensions to the nonlinear case using the kernel trick.
  - More extensive investigation of the sparse CCA model in biological applications.