A Least Squares Formulation for Canonical Correlation Analysis

Liang Sun LSUN27@ASU.EDU
Shuiwang Ji SHUIWANG.JI@ASU.EDU
Jieping Ye JIEPING.YE@ASU.EDU
Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287 USA

Abstract

Canonical Correlation Analysis (CCA) is a well-known technique for finding the correlations between two sets of multi-dimensional variables. It projects both sets of variables into a lower-dimensional space in which they are maximally correlated. CCA is commonly applied for supervised dimensionality reduction, in which one of the multi-dimensional variables is derived from the class label. It has been shown that CCA can be formulated as a least squares problem in the binary-class case. However, their relationship in the more general setting remains unclear. In this paper, we show that, under a mild condition which tends to hold for high-dimensional data, CCA in multi-label classifications can be formulated as a least squares problem. Based on this equivalence relationship, we propose several CCA extensions including sparse CCA using 1-norm regularization. Experiments on multi-label data sets confirm the established equivalence relationship. Results also demonstrate the effectiveness of the proposed CCA extensions.

1. Introduction

Canonical Correlation Analysis (CCA) (Hotelling, 1936) is commonly used for finding the correlations between two sets of multi-dimensional variables. It makes use of two views of the same set of objects and projects them into a lower-dimensional space in which they are maximally correlated. CCA has been applied successfully in various applications (Hardoon et al., 2004; Vert & Kanehisa, 2003). 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. In this setting, the data can be projected into a lower-dimensional space directed by the label information. Such formulation is particularly appealing in the context of dimensionality reduction for multi-label data (Yu et al., 2006).

Multivariate linear regression (MLR) that minimizes the sum-of-squares cost function is a well-studied technique for regression problems. It can also be applied for classification with an appropriate class indicator matrix (Bishop, 2006; Hastie et al., 2001). The solution to least squares problems can be obtained by solving a linear system of equations. A number of algorithms, including the conjugate gradient algorithm, can be applied to solve it efficiently (Golub & Loan, 1996). Furthermore, the least squares formulation can be readily extended using the regularization technique. For example, 1-norm regularization can be incorporated into the least squares formulation to control model complexity and improve sparseness (Tibshirani, 1996). Sparseness often leads to easy interpretation and a good generalization ability. It has been used successfully in several algorithms including Principal Component Analysis (d'Aspremont et al., 2004) and SVM (Zhu et al., 2003).

In contrast to least squares, CCA involves a generalized eigenvalue problem, which is computationally more expensive to solve. Furthermore, it is challenging to derive sparse CCA, as it involves a difficult sparse generalized eigenvalue problem. Convex relaxation of sparse CCA has been studied in (Sriperumbudur et al., 2007), where the exact sparse CCA formulation has been relaxed in several steps. On the other hand, interesting connection between least squares and CCA has been established in the literature. In particular, CCA has been shown to be equivalent to Fisher Linear Discriminant Analysis (LDA) for binary-class problems (Hastie et al., 1995). Meanwhile, it is well-known that LDA is equivalent to least squares in this case (Bishop, 2006; Hastie et al., 2001). Thus CCA can be

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

formulated as a least squares problem for binary-class problems. In practice, the multi-class and multi-label problems are more prevalent. It is therefore tempting to investigate the relationship between least squares and CCA in the more general setting.

In this paper, we study the relationship between CCA and least squares for multi-label problems. We show that, under a mild condition which tends to hold for high-dimensional data, CCA can be formulated as a least squares problem by constructing a specific class indictor matrix. Based on this equivalence relationship, we propose several CCA extensions including sparse CCA using the 1-norm regularization. Furthermore, the entire solution path for sparse CCA can be readily computed by the Least Angle Regression algorithm (LARS) (Efron et al., 2004). We evaluate the established theoretical results using a collection of multi-label data sets. Our experiments confirm the equivalence relationship between these two models under the given assumption. Results also show that, even when the assumption does not hold, they achieve very similar performance. Our experiments also demonstrate the effectiveness of the proposed CCA extensions.

Notations The number of training samples, the data dimensionality, and the number of labels are denoted by n, d, and k, respectively. $x_i \in \mathrm{R}^d$ denotes the ith observation and $y_i \in \mathrm{R}^k$ encodes the corresponding label information. Let $X = [x_1, \cdots, x_n] \in \mathrm{R}^{d\times n}$ be the data matrix and $Y = [y_1, \cdots, y_n] \in \mathrm{R}^{k\times n}$ be the class label matrix. We assume that both $\{x_i\}_1^n$ and $\{y_i\}_1^n$ are centered, i.e., $\sum_{i=1}^n x_i = 0$, and $\sum_{i=1}^n y_i = 0$.

2. Background and Related Work

In this section we give a brief overview of CCA and least squares as well as several related work.

2.1. Canonical Correlation Analysis

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. Let $X \in \mathbb{R}^{d\times n}$ and $Y \in \mathbb{R}^{k\times n}$ be the matrices for the two representations of the objects where x_i and y_i correspond to the ith object. CCA computes two projection vectors, $w_x \in \mathbb{R}^d$ and $w_y \in \mathbb{R}^k$, such that the following correlation coefficient:

$$\rho = \frac{w_x^T X Y^T w_y}{\sqrt{(w_x^T X X^T w_x)(w_y^T Y Y^T w_y)}} \tag{1}$$

is maximized. Since ρ is invariant to the scaling of w_x and w_y, CCA can be formulated equivalently as

$$\begin{aligned} \max_{w_x, w_y} \quad & w_x^T X Y^T w_y \\ \text{subject to} \quad & w_x^T X X^T w_x = 1, \;\; w_y^T Y Y^T w_y = 1. \end{aligned} \tag{2}$$

We assume that YY^T is nonsingular. It can be shown that w_x can be obtained by solving the following optimization problem:

$$\begin{aligned} \max_{w_x} \quad & w_x^T X Y^T \left(Y Y^T\right)^{-1} Y X^T w_x \\ \text{subject to} \quad & w_x^T X X^T w_x = 1. \end{aligned} \tag{3}$$

Both formulations in Eqs. (2) and (3) attempt to find the eigenvectors corresponding to top eigenvalues of the following generalized eigenvalue problem:

$$X Y^T (Y Y^T)^{-1} Y X^T w_x = \eta X X^T w_x, \tag{4}$$

where η is the eigenvalue corresponding to the eigenvector w_x. Multiple projection vectors under certain orthonormality constraints can be computed simultaneously by solving the following optimization problem (Hardoon et al., 2004):

$$\begin{aligned} \max_{W} \quad & \mathrm{trace}(W^T X Y^T (Y Y^T)^{-1} Y X^T W) \\ \text{subject to} \quad & W^T X X^T W = I, \end{aligned} \tag{5}$$

where $W \in \mathbb{R}^{d\times \ell}$ is the projection matrix, ℓ is the number of projection vectors, and I is the identity matrix. The solution to the optimization problem in Eq. (5) consists of the top ℓ eigenvectors of the generalized eigenvalue problem in Eq. (4).

In regularized CCA (rCCA), a regularization term λI with $\lambda > 0$ is added to XX^T in Eq. (5) to prevent the overfitting and avoid the singularity of XX^T (Bach & Jordan, 2003). Specifically, rCCA solves the following generalized eigenvalue problem:

$$X Y^T (Y Y^T)^{-1} Y X^T w_x = \eta (X X^T + \lambda I) w_x. \tag{6}$$

2.2. Least Squares for Regression and Classification

In regression, we are given a training set $\{(x_i, t_i)\}_{i=1}^n$, where $x_i \in \mathrm{R}^d$ is the observation and $t_i \in \mathrm{R}^k$ is the corresponding target. We assume that both the observations and the targets are centered. Thus the bias term can be ignored. In this case, the projection matrix W can be computed by minimizing the following sum-of-squares cost function:

$$\min_W \sum_{i=1}^n \|W^T x_i - t_i\|_2^2 = \|W^T X - T\|_F^2, \tag{7}$$

where $T = [t_1, \cdots, t_n]$. It is well known that the optimal projection matrix is given by

$$W_{LS} = (XX^T)^\dagger XT^T, \tag{8}$$

where the pseudo-inverse is used in case XX^T is singular. To improve the generality ability of the model, a penalty term based on 2-norm or 1-norm regularization is commonly applied (Hastie et al., 2001).

Least squares is also commonly applied for classification. In the general multi-class case, 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 $y_i \in \{1, 2, \cdots, k\}$ denotes the class label of the i-th sample, and $k > 2$. To apply the least squares formulation to the multi-class case, the 1-of-k binary coding scheme is usually employed to apply a vector-valued class code to each data point (Bishop, 2006). The solution to the least squares problem depends on the choice of class indicator matrix. Several class indicator matrices have been proposed in the literature (Hastie et al., 2001).

2.3. Related Work

The inherent relationship between least squares and several other models has been established in the past. In particular, LDA for two-class problems can be formulated as a least squares problem (Duda et al., 2000; Bishop, 2006). Moreover, this equivalence relationship can be extended to the multi-class case using a specific class indicator matrix (Ye, 2007). CCA has been shown to be equivalent to LDA for multi-class problems (Hastie et al., 1995). Thus, CCA is closely related to least squares in the multi-class case. We show in the next section that, under a mild condition, CCA can be formulated as a least squares problem for multi-label classifications when one of the views used in CCA is derived from the labels.

3. Relationship between CCA and Least Squares

In this section we investigate the relationship between CCA and least squares in the multi-label case. We first define four matrices essential for our derivation:

$$\begin{aligned} H &= Y^T(YY^T)^{-\frac{1}{2}} \in \mathrm{R}^{n\times k}, && (9)\\ C_{XX} &= XX^T \in \mathrm{R}^{d\times d}, && (10)\\ C_{HH} &= XHH^TX^T \in \mathrm{R}^{d\times d}, && (11)\\ C_{DD} &= C_{XX} - C_{HH} \in \mathrm{R}^{d\times d}. && (12) \end{aligned}$$

Note that we assume $n \gg k$ and $\mathrm{rank}(Y) = k$ for multi-label problems. Thus, $(YY^T)^{-\frac{1}{2}}$ is well-defined. It follows from the definition above that the solution to CCA can be expressed as the eigenvectors corresponding to top eigenvalues of the matrix $C_{XX}^\dagger C_{HH}$.

3.1. Basic Matrix Properties

In this subsection, we study the basic properties of the matrices involved in the following discussion. Following the definition of H in Eq. (9), we have:

Lemma 1. *Let H be defined as in Eq. (9) and let $\{y_i\}_1^n$ be centered, i.e., $\sum_{i=1}^n y_i = 0$. Then we have*

(1). H has orthonormal columns, i.e., $H^TH = I_k$;

(2). $H^Te = 0$.

Given $H \in \mathrm{R}^{n\times k}$ with orthonormal columns, there exists $D \in \mathrm{R}^{n\times(n-k)}$ such that $[H, D] \in \mathrm{R}^{n\times n}$ is an orthogonal matrix (Golub & Loan, 1996), that is

$$I_n = [H, D][H, D]^T = HH^T + DD^T.$$

It follows that

$$C_{DD} = C_{XX} - C_{HH} = XDD^TX^T. \tag{13}$$

It can be verified from Eqs. (10), (11), and (13) that the matrices C_{XX}, C_{HH}, and C_{DD} are all positive semidefinite.

Let the Singular Value Decomposition (SVD) of X be

$$\begin{aligned} X &= U\Sigma V^T = [U_1, U_2]\ \mathrm{diag}(\Sigma_r, 0)\ [V_1, V_2]^T \\ &= U_1\Sigma_r V_1^T, && (14) \end{aligned}$$

where $r = \mathrm{rank}(X)$, U and V are orthogonal matrices, $\Sigma \in \mathrm{R}^{d\times n}$, $U_1 \in \mathrm{R}^{d\times r}$, $U_2 \in \mathrm{R}^{d\times(d-r)}$, $V_1 \in \mathrm{R}^{n\times r}$, $V_2 \in \mathrm{R}^{n\times(n-r)}$, and $\Sigma_r \in \mathrm{R}^{r\times r}$.

Since U_2 lies in the null space X^T, we have:

Lemma 2. *Let H, X, U_2, and D be defined as above. Then $H^TX^TU_2 = 0$ and $D^TX^TU_2 = 0$.*

3.2. Computing CCA via Eigendecomposition

Recall that the solution to CCA consists of the top ℓ eigenvectors of the matrix $C_{XX}^\dagger C_{HH}$. We next show how to compute the eigenvectors of $C_{XX}^\dagger C_{HH}$. Define the matrix $A \in \mathrm{R}^{r\times k}$ by

$$A = \Sigma_r^{-1}U_1^TXH. \tag{15}$$

Let the SVD of A be $A = P\Sigma_A Q^T$, where $P \in \mathrm{R}^{r\times r}$ and $Q \in \mathrm{R}^{k\times k}$ are orthogonal, and $\Sigma_A \in \mathrm{R}^{r\times k}$ is diagonal. Then

$$AA^T = P\Sigma_A\Sigma_A^TP^T. \tag{16}$$

The matrix $C_{XX}^{\dagger}C_{HH}$ can be diagonalized as follows:

$$\begin{aligned}
C_{XX}^{\dagger}C_{HH} &= U_1\Sigma_r^{-2}U_1^T XHH^TX^T \\
&= U_1\Sigma_r^{-1}AH^TX^TUU^T \\
&= U\begin{bmatrix} I_r \\ 0 \end{bmatrix}\Sigma_r^{-1}A[H^TX^TU_1, H^TX^TU_2]U^T \\
&= U\begin{bmatrix} \Sigma_r^{-1}AA^T\Sigma_r & 0 \\ 0 & 0 \end{bmatrix}U^T \\
&= U\begin{bmatrix} \Sigma_r^{-1}P & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} \Sigma_A\Sigma_A^T & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} P^T\Sigma_r & 0 \\ 0 & I \end{bmatrix}U^T
\end{aligned}$$

where the second equality follows sine U is orthogonal, the fourth equality follows sine $H^TX^TU_2 = 0$ as shown in Lemma 2, and the last equality follows from Eq. (16). Thus the solution to CCA, which consists of the top ℓ eigenvectors of matrix $C_{XX}^{\dagger}C_{HH}$, is given by

$$W_{CCA} = U_1\Sigma_r^{-1}P_l, \tag{17}$$

where P_ℓ contains the first ℓ columns of P.

3.3. Equivalence of CCA and Least Squares

Recall from Eq. (8) that for a given class indicator matrix T, the solution to the least squares problem is given by

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

We define the class indicator matrix $\tilde{T}$ as follows:

$$\tilde{T} = (YY^T)^{-\frac{1}{2}}Y = H^T, \tag{18}$$

In this case, the solution to the least squares problem is given by

$$\begin{aligned}
W_{LS} &= (XX^T)^{\dagger}XH = U_1\Sigma_r^{-2}U_1^TXH \\
&= U_1\Sigma_r^{-1}A = U_1\Sigma_r^{-1}P\Sigma_AQ^T.
\end{aligned} \tag{19}$$

It is clear from Eqs. (17) and (19) that the difference between CCA and least squares lies in Σ_A and Q^T.

We next show that all diagonal elements of Σ_A are one under a mild condition, that is, $\mathrm{rank}(X) = n-1$ and $\mathrm{rank}(Y) = k$. Note that the first condition is equivalent to requiring that the original data points are linearly independent before centering, which tends to hold for high-dimensional data.

Before presenting the main result summarized in Theorem 1 below, we have the following lemma:

Lemma 3. *Assume*

$$rank(C_{XX}) + s = rank(C_{HH}) + rank(C_{DD}),$$

for some non-negative integer s. Then for the matrix $\hat{\Sigma}_A = \Sigma_A\Sigma_A^T = diag(a_1, a_2, \cdots, a_r) \in \mathrm{R}^{r\times r}$, we have

$$1 = \cdots = a_{f-s} > a_{f-s+1} \geqslant \cdots \geqslant a_f > a_{f+1} = \cdots = 0.$$

where $f = rank(\Sigma_A)$.

Proof. Define the matrix $J \in \mathrm{R}^{d\times d}$ as follows:

$$J = U\begin{bmatrix} \Sigma_r^{-1}P & 0 \\ 0 & I_{d-r} \end{bmatrix}. \tag{20}$$

It follows from the definition of C_{XX}, C_{HH}, and C_{DD} in Eqs. (10)-(12) that

$$\begin{aligned}
J^TC_{XX}J &= \mathrm{diag}(I_r, 0), \\
J^TC_{HH}J &= \mathrm{diag}(\Sigma_A\Sigma_A^T, 0) \\
&= \mathrm{diag}(a_1, \cdots, a_r, 0, \cdots, 0), \\
J^TC_{DD}J &= J^TC_{XX}J - J^TC_{HH}J \\
&= \mathrm{diag}(b_1, \cdots, b_r, 0, \cdots, 0),
\end{aligned} \tag{21}$$

where $b_i = 1 - a_i$, for $i = 1, \cdots, r$. Note that since J is nonsingular, we have

$$\mathrm{rank}(C_{XX}) = \mathrm{rank}(J^TC_{XX}J) = r.$$

It follows from our assumption that

$$\mathrm{rank}(J^TC_{HH}J) + \mathrm{rank}(J^TC_{DD}J) = r + s. \tag{22}$$

Since both $J^TC_{HH}J$ and $J^TC_{DD}J$ are diagonal, there are a total of $r+s$ nonzero elements in $J^TC_{HH}J$ and $J^TC_{DD}J$. Note that $f = \mathrm{rank}(\Sigma_A) = \mathrm{rank}(\hat{\Sigma}_A)$, thus $a_1 \geqslant \cdots \geqslant a_f > 0 = a_{f+1} = \cdots = a_r$. It follows from Eq. (21) that

$$a_i + b_i = 1, b_r \geqslant \cdots \geqslant b_1 \geqslant 0, \tag{23}$$

This implies that at least one of a_i or b_i is positive for $1 \leqslant i \leqslant r$. To satisfy the rank equality in Eq. (22), we therefore must have

$$\begin{aligned}
1 &= a_1 = a_2 = \cdots = a_{f-s} > a_{f-s+1} \geqslant \cdots \geqslant a_f \\
&> a_{f+1} = \cdots = a_r = 0, \\
0 &= b_1 = b_2 = \cdots = b_{f-s} < b_{f-s+1} \leqslant \cdots \leqslant b_f \\
&< b_{f+1} = \cdots = b_r = 1.
\end{aligned}$$

This completes the proof of the lemma. □

Theorem 1. *Assume that $rank(X) = n-1$ and $rank(Y) = k$ for multi-label problems. Then we have*

$$\begin{aligned}
rank(C_{XX}) &= n-1, &(24)\\
rank(C_{HH}) &= k, &(25)\\
rank(C_{DD}) &= n-k-1. &(26)
\end{aligned}$$

Thus $s = 0$, where s is defined in Lemma 3, and

$$1 = a_1 = \cdots = a_k > a_{k+1} = \cdots = a_r = 0.$$

This implies that all diagonal elements of Σ_A are ones.

Proof. Denote $e^T = [1, 1, \cdots, 1] \in \mathrm{R}^{1\times n}$, $H = [h_1, \cdots, h_k]$, and $D = [h_{k+1}, \cdots, h_n]$. Note that X is column centered, i.e., $\sum_{i=1}^{n} x_i = 0$. It follows from Lemma 1 that $H^T e = 0$, that is,

$$h_i^T e = 0, \text{ for } 1 \leqslant i \leqslant k. \tag{27}$$

Since $[H, D]$ is an orthogonal matrix, $\{h_1, \cdots, h_n\}$ form a basis for R^n. Thus we can represent $e \in \mathrm{R}^n$ as

$$e = \sum_{i=1}^{n} \alpha_i h_i, \text{ where } \alpha_i \in \mathrm{R}. \tag{28}$$

It follows from the orthogonality of $[H, D]$ and Eq. (27) that e can be expressed as $e = \sum_{i=k+1}^{n} \alpha_i h_i$, and

$$0 = Xe = X\left(\sum_{i=k+1}^{n} \alpha_i h_i\right) = \sum_{i=k+1}^{n} \alpha_i (Xh_i). \tag{29}$$

Since not all α_i's are zero, the $n-k$ columns of XD are linearly dependent, thus $\mathrm{rank}(XD) \leqslant n-k-1$. According to the property of matrix rank, we have

$$\begin{aligned} &\mathrm{rank}(XD) \geqslant \mathrm{rank}(X) + \mathrm{rank}(D) - n \\ &= (n-1) + (n-k) - n = n-k-1. \end{aligned} \tag{30}$$

Thus, $\mathrm{rank}(XD) = n-k-1$ holds.

For matrix XH, we have

$$\begin{aligned} &\mathrm{rank}(X) = \mathrm{rank}(X[H, D]) \leqslant \mathrm{rank}(XH) + \mathrm{rank}(XD) \\ &\Leftrightarrow n-1 \leqslant \mathrm{rank}(XH) + n - k - 1 \\ &\Leftrightarrow \mathrm{rank}(XH) \geqslant k. \end{aligned}$$

On the other hand, since $XH \in \mathrm{R}^{d\times k}$, $\mathrm{rank}(XH) \leqslant k$. Thus we have $\mathrm{rank}(XH) = k$ and

$$\begin{aligned} \mathrm{rank}(C_{XX}) &= \mathrm{rank}(X) = n-1, \\ \mathrm{rank}(C_{HH}) &= \mathrm{rank}(XH) = k, \\ \mathrm{rank}(C_{DD}) &= \mathrm{rank}(XD) = n-k-1. \end{aligned}$$

It follows that $s = 0$. On the other hand,

$$f = \mathrm{rank}(A) = \mathrm{rank}(\Sigma_r^{-1} U_1^T XH) = \mathrm{rank}(XH) = k.$$

Hence,

$$1 = a_1 = a_2 = \cdots = a_k > 0 = a_{k+1} = \cdots = a_r,$$

and all diagonal elements of Σ_A are one. This completes the proof of the theorem. □

Since $\mathrm{rank}(\Sigma_A) = k$, $C_{XX}^{\dagger} C_{HH}$ contains k nonzero eigenvalues. If we choose $\ell = k$, then

$$W_{CCA} = U_1 \Sigma_r^{-1} P_k. \tag{31}$$

The only difference between W_{LS} and W_{CCA} lies in the orthogonal matrix Q^T in W_{LS}.

In practice, we can use both W_{CCA} and W_{LS} to project the original data onto a lower-dimensional space before classification. For any classifiers based on Euclidean distance, the orthogonal transformation Q^T will not affect the classification performance since the Euclidean distance is invariant of any orthogonal transformations. Some well-known algorithms with this property include the K-Nearest-Neighbor (KNN) algorithm (Duda et al., 2000) based on the Euclidean distance and the linear Support Vector Machines (SVM) (Schölkopf & Smola, 2002). In the following, the least squares CCA formulation is called "LS-CCA".

4. Extensions of CCA

Based on the equivalence relationship established in the last section, the classical CCA formulation can be extended using the regularization technique.

Regularization is commonly used to control the complexity of the model and improve the generalization performance. It is the key to some popular algorithms such as SVM. Linear regression using the 2-norm regularization, called ridge regression (Hoerl & Kennard, 1970; Rifkin et al., 2003), minimizes the penalized sum-of-squares cost function. By using the target matrix $\tilde{T}$ in Eq. (18), we obtain the 2-norm regularized least squares CCA formulation (called "LS-CCA$_2$") by minimizing the following objective function:

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

where $W = [w_1, \cdots, w_k]$, and $\lambda > 0$ is the regularization parameter.

In mathematical programming, it is known that sparseness can often be achieved by penalizing the L_1-norm of the variables (Donoho, 2006; Tibshirani, 1996). It has been introduced into the least squares formulation and the resulting model is called lasso (Tibshirani, 1996). Based on the established equivalence relationship between CCA and least squares, we derive the 1-norm least squares CCA formulation (called "LS-CCA$_1$") by minimizing the following objective function:

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

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

$$w_j^* = \arg\min_{w_j} \left(\sum_{i=1}^{n} (x_i^T w_j - \tilde{T}_{ij})^2 + \lambda \|w_j\|_1 \right), \tag{32}$$

which can be reformulated as:

$$w_j^* = \arg \min_{\|w_j\|_1 \leq \tau} \sum_{i=1}^{n} (x_i^T w_j - \tilde{T}_{ij})^2, \qquad (33)$$

for some tuning parameter $\tau > 0$ (Tibshirani, 1996). Furthermore, the solution can be readily computed by the Least Angle Regression algorithm (Efron et al., 2004). One key feature of LARS is that it computes the entire solution path for all values of τ, with essentially the same computational cost as fitting the model with a single τ value.

If the value of τ is large enough, the constraint in Eq. (33) is not effective, resulting in an unconstrained optimization problem. We can thus consider τ from a finite range $[0, \hat{\tau}]$, for some $\hat{\tau} > 0$. Define $\gamma = \tau/\hat{\tau}$ so that $\tau = \hat{\tau}\gamma$ with $0 \leq \gamma \leq 1$. The estimation of τ is equivalent to the estimation of γ. Cross-validation is commonly used to estimate the optimal value from a large candidate set $S = \{\gamma_1, \gamma_2, \cdots, \gamma_{|S|}\}$, where $|S|$ denotes the size of S. If the value of γ is sufficiently small, many of the coefficients in W will become exactly zero, which leads to a sparse CCA model. We thus call γ the "sparseness coefficient".

5. Experiments

We use a collection of multi-label data sets to experimentally verify the equivalence relationship established in this paper. We also evaluate the performance of various CCA extensions.

5.1. Experimental Setup

We use two types of data in the experiment. The gene expression pattern image data[1] describe the gene expression patterns of *Drosophila* during development (Tomancak & *et al.*, 2002). Each image is annotated with a variable number of textual terms (labels) from a controlled vocabulary. We apply Gabor filters to extract a 384-dimensional feature vector from each image. We use five data sets with different numbers of terms (class labels). We also use the scene data set (Boutell et al., 2004) which contains 2407 samples of 294-dimension and 6 labels. In all the experiments, ten random splittings of data into training and test sets are generated and the averaged performance is reported.

In the experiment, five methods including CCA and its regularized version rCCA in Eq. (6), as well as LS-CCA and its regularization versions LS-CCA$_2$ and LS-CCA$_1$ are compared. These CCA methods are used to project the data into a lower-dimensional space in which a linear SVM is applied for classification for each label. The Receiver Operating Characteristic (ROC) value is computed for each label and the averaged performance over all labels is reported.

[1]All images were extracted from the FlyExpress database at http://www.flyexpress.net.

5.2. Gene Expression Pattern Image Data

In this experiment we first evaluate the equivalence relationship between CCA and least squares. For all cases, we set the data dimensionality d larger than the sample size d, i.e., $n/d > 1$. The condition in Theorem 1 holds in all cases. We observe that for all splittings of all of the five data sets, $\text{rank}(C_{XX})$ equals $\text{rank}(C_{HH}) + \text{rank}(C_{DD})$), and the ratio of the maximal to the minimal diagonal element of Σ_A is 1, which implies that all diagonal elements of Σ_A are the same, i.e., ones. Our experimental evidences are consistent with the theoretical results presented in Section 3.3.

5.2.1. Performance Comparison

In Table 1, we report the mean ROC scores over all terms and all splittings for each data set. The main observations include: (1) CCA and LS-CCA achieve the same performance for all data sets, which is consistent with our theoretical results; (2) The regularized CCA extensions including rCCA, LS-CCA$_2$, and LS-CCA$_1$ perform much better than their counterparts CCA and LS-CCA without the regularization; and (3) LS-CCA$_2$ is comparable to rCCA in all data sets, while LS-CCA$_1$ achieves the best performance in all cases. These further justify the use of the proposed least squares CCA formulations for multi-label classifications.

Table 1. 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.

n_{tot}	k	CCA	LS-CCA	rCCA	LS-CCA$_2$	LS-CCA$_1$
863	10	0.542	0.542	0.617	0.619	**0.722**
1041	15	0.534	0.534	0.602	0.603	**0.707**
1138	20	0.538	0.538	0.609	0.610	**0.714**
1222	25	0.540	0.540	0.603	0.605	**0.704**
1349	30	0.548	0.548	0.606	0.608	**0.709**

5.2.2. Sensitivity Study

In this experiment, we investigate the performance of LS-CCA and its variants in comparison with CCA when the condition in Theorem 1 does not hold, which

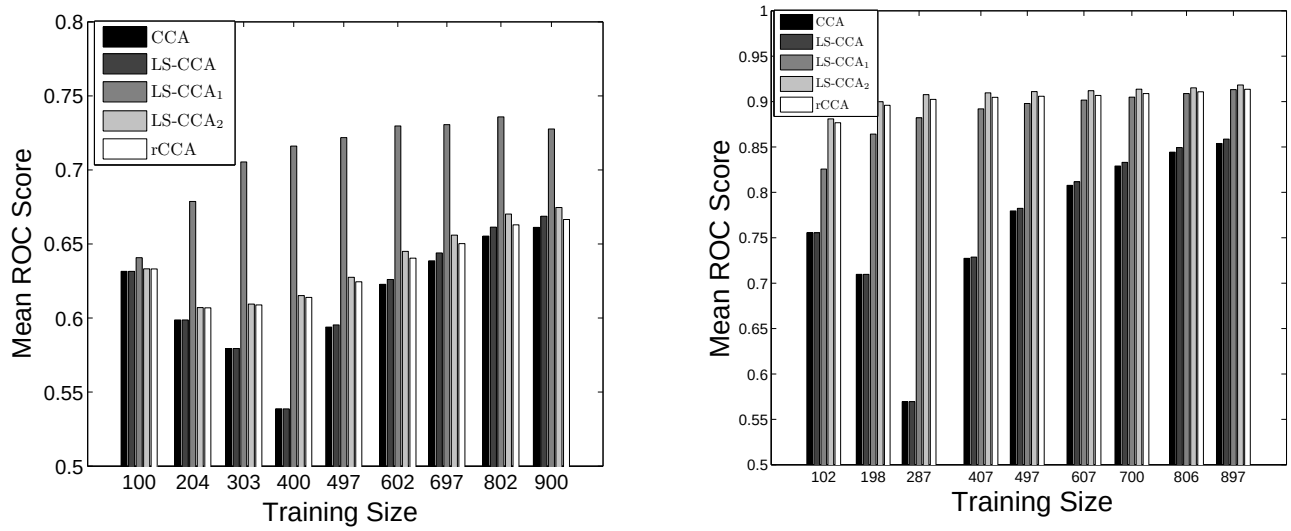

Figure 1. Comparison of all algorithms on gene data set (left) and scene data set (right) in terms of mean ROC scores.

is the case in many applications. Specifically, we use a gene data set with the dimensionality fixed at $d = 384$, while the size of the training set varies from 100 to 900 with a step size about 100.

The performance of different algorithms as the size of training set increases is presented in Figure 1 (left graph). We can observe that in general, the performance of all algorithms increases as the training size increases. When n is small, the condition in Theorem 1 holds, thus CCA and LS-CCA are equivalent, and they achieve the same performance. When n further increases, CCA and LS-CCA achieve different ROC scores, although the difference is always very small in our experiment. Similar to the last experiment, we can observe from the figure that the regularized methods perform much better than CCA and LS-CCA, and LS-CCA$_2$ is comparable to rCCA. The sparse formulation LS-CCA$_1$ performs the best for this data set.

5.3. Scene Data Set

We conduct a similar set of experiments on the scene data. As in the gene data set, the equivalence relationship holds when the condition in Theorem 1 holds.

For the performance comparison and sensitivity study, we generate a sequence of training sets with the size n ranging from 100 to 900 with a step size around 100. The results are summarized in Figure 1 (right graph). Similar to the gene data set, CCA and LS-CCA achieve the same performance when n is small, and they differ slightly when n is large. We can also observe from the figure that the regularized algorithms including rCCA, and LS-CCA$_2$, and LS-CCA$_1$ perform much better than CCA and LS-CCA without regularization, and LS-CCA$_2$ performs slightly better than others in this data set.

5.4. The Entire CCA Solution Path

In this experiment, we investigate the sparse CCA model, i.e., LS-CCA$_1$ using the scene data set. Recall that the sparseness of the weight vectors w_i's depends on the sparseness coefficient γ between 0 and 1.

Figure 2 shows the entire collection of solution paths for a subset of the coefficients from the first weight vector w_1. The x-axis denotes the sparseness coefficient γ, and the y-axis denotes the value of the coefficients. The vertical lines denote (a subset of) the turning point of the path, as the solution path for each of the coefficients is piecewise linear (Efron et al., 2004). We can observe from Figure 2 that when $\gamma = 1$, most of the coefficients are non-zero, i.e., the model is dense. When the value of the sparseness coefficient γ decreases (from the right to the left side along the x-axis), more and more coefficients become exactly zero. All coefficients become zero when $\gamma = 0$.

6. Conclusion and Future Work

In this paper we show that CCA for multi-label classifications can be formulated as a least squares problems under a mild condition, which tends to hold for high-dimensional data. Based on the equivalence relationship established in this paper, we propose several CCA extensions including sparse CCA. We have conducted experiments on a collection of multi-label data

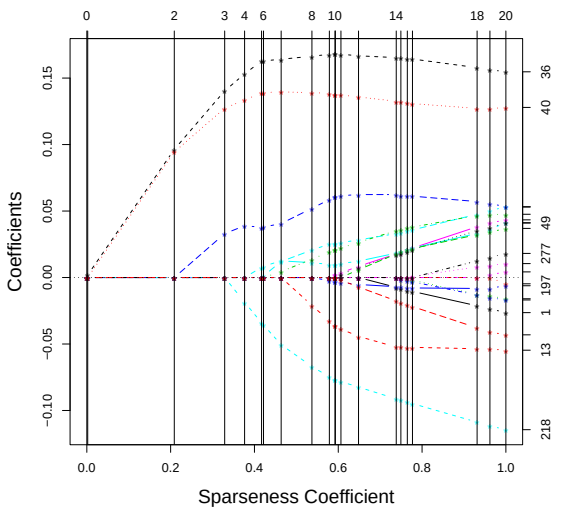

Figure 2. The entire collection of solution paths for a subset of the coefficients from the first weight vector w_1 on the scene data set. The x-axis denotes the sparseness coefficient γ, and the y-axis denotes the value of the coefficients.

sets to validate the proposed equivalence relationship. Our experimental results show that the performance of the proposed least squares formulation and CCA is very close even when the condition does not hold. Results also demonstrate the effectiveness of the proposed CCA extensions.

The proposed least squares formulation facilitates the incorporation of the unlabeled data into the CCA framework through the graph Laplacian, which captures the local geometry of the data (Belkin et al., 2006). We plan to examine the effectiveness of this semi-supervised CCA model for learning from both labeled and unlabeled data. The proposed sparse CCA performs well for the gene data set. We plan to analyze the biological relevance of the features extracted via the sparse CCA model.

Acknowledgments

This research is sponsored in part by funds from the Arizona State University and the National Science Foundation under Grant No. IIS-0612069.

References

Bach, F. R., & Jordan, M. I. (2003). Kernel independent component analysis. *J. Mach. Learn. Res.*, *3*, 1–48.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, *7*, 2399–2434.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition*, *37*, 1757–1771.

d'Aspremont, A., Ghaoui, L., Jordan, M., & Lanckriet, G. (2004). A direct formulation for sparse PCA using semidefinite programming. *NIPS* (pp. 41–48).

Donoho, D. (2006). For most large underdetermined systems of linear equations, the minimal 11-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, *59*, 907–934.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. New York: John Wiley and Sons, Inc.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*, 407.

Golub, G. H., & Loan, C. F. V. (1996). *Matrix computations*. Baltimore, MD: Johns Hopkins Press.

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

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. *Annals of Statistics*, *23*, 73–102.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

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

Rifkin, R., Yeo, G., & Poggio, T. (2003). Regularized least-squares classification. *Advances in Learning Theory: Methods, Model and Applications, NATO Science Series III: Computer and Systems Sciences* (pp. 131–154). Amsterdam: VIOS Press.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

Sriperumbudur, B. K., Torres, D. A., & Lanckriet, G. R. G. (2007). Sparse eigen methods by D.C. programming. *ICML* (pp. 831–838).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, *58*, 267–288.

Tomancak, P., & *et al.* (2002). Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biology*, *3*.

Vert, J.-P., & Kanehisa, M. (2003). Graph-driven feature extraction from microarray data using diffusion kernels and kernel cca. *NIPS* (pp. 1425–1432).

Ye, J. (2007). Least squares linear discriminant analysis. *ICML* (pp. 1087–1094).

Yu, S., Yu, K., Tresp, V., & Kriegel, H.-P. (2006). Multi-output regularized feature projection. *IEEE Trans. Knowl. Data Eng.*, *18*, 1600–1613.

Zhu, J., Rosset, S., Hastie, T., & Tibshirani, R. (2003). 1-norm support vector machines. *NIPS* (pp. 49–56).