A Least Squares Formulation for a Class of Generalized Eigenvalue Problems in Machine Learning

Liang Sun SUN.LIANG@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

Many machine learning algorithms can be formulated as a generalized eigenvalue problem. One major limitation of such formulation is that the generalized eigenvalue problem is computationally expensive to solve especially for large-scale problems. In this paper, we show that under a mild condition, a class of generalized eigenvalue problems in machine learning can be formulated as a least squares problem. This class of problems include classical techniques such as Canonical Correlation Analysis (CCA), Partial Least Squares (PLS), and Linear Discriminant Analysis (LDA), as well as Hypergraph Spectral Learning (HSL). As a result, various regularization techniques can be readily incorporated into the formulation to improve model sparsity and generalization ability. In addition, the least squares formulation leads to efficient and scalable implementations based on the iterative conjugate gradient type algorithms. We report experimental results that confirm the established equivalence relationship. Results also demonstrate the efficiency and effectiveness of the equivalent least squares formulations on large-scale problems.

1. Introduction

A number of machine learning algorithms can be formulated as a generalized eigenvalue problem. Such techniques include Canonical Correlation Analysis (CCA), Partial Least Squares (PLS), Linear Discriminant Analysis (LDA), and Hypergraph Spectral Learning (HSL) (Hotelling, 1936; Rosipal & Krämer, 2006; Sun et al., 2008a; Tao et al., 2009; Ye, 2007). Although well-established algorithms in numerical linear algebra have been developed to solve generalized eigenvalue problems, they are in general computationally expensive and hence may not scale to large-scale machine learning problems. In addition, it is challenging to directly incorporate the sparsity constraint into the mathematical formulation of these techniques. Sparsity often leads to easy interpretation and a good generalization ability. It has been used successfully in linear regression (Tibshirani, 1996), and Principal Component Analysis (PCA) (d'Aspremont et al., 2004).

Multivariate Linear Regression (MLR) that minimizes the sum-of-squares error function, called least squares, is a classical technique for regression problems. It can also be applied for classification problems by defining an appropriate class indicator matrix (Bishop, 2006). The solution to the least squares problems can be obtained by solving a linear system of equations, and a number of algorithms, including the conjugate gradient algorithm, can be applied to solve it efficiently (Golub & Van Loan, 1996). Furthermore, the least squares formulation can be readily extended using the regularization technique. For example, the 1-norm and 2-norm regularization can be incorporated into the least squares formulation to improve sparsity and control model complexity (Bishop, 2006).

Motivated by the mathematical and numerical properties of the generalized eigenvalue problem and the least squares formulation, several researchers have attempted to connect these two approaches. In particular, it has been shown that there is close relationship between LDA, CCA, and least squares (Hastie et al., 2001; Bishop, 2006; Ye, 2007; Sun et al., 2008b). However, the intrinsic relationship between least squares and other techniques involving generalized eigenvalue problems mentioned above remains unclear.

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

In this paper, we study the relationship between the least squares formulation and a class of generalized eigenvalue problems in machine learning. In particular, we establish the equivalence relationship between these two formulations under a mild condition. As a result, various regularization techniques such as the 1-norm and 2-norm regularization can be readily incorporated into the formulation to improve model sparsity and generalization ability. In addition, this equivalence relationship leads to efficient and scalable implementations for these generalized eigenvalue problems based on the iterative conjugate gradient type algorithms such as LSQR (Paige & Saunders, 1982). We have conducted experiments using several benchmark data sets. The experiments confirm the equivalence relationship between these two models under the given assumption. Our results show that even when the assumption does not hold, the performance of these two models is still very close. Results also demonstrate the efficiency and effectiveness of the equivalent least squares models and their extensions.

Notations: The number of training samples, the data dimensionality, and the number of classes (or labels) are denoted by n, d, and k, respectively. $x_i \in \mathbb{R}^d$ denotes the ith observation, and $y_i \in \mathbb{R}^k$ encodes the label information for x_i. $X = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{d\times n}$ represents the data matrix, and $Y = [y_1, y_2, \cdots, y_n] \in \mathbb{R}^{k\times n}$ is the matrix representation for label information. $\{x_i\}_1^n$ is assumed to be centered, i.e., $\sum_{i=1}^n x_i = 0$. $\mathcal{S} \in \mathbb{R}^{n\times n}$ is a symmetric and positive semi-definite matrix, and e is a vector of all ones.

Organization: We present background and related work in Section 2, establish the equivalence relationship between the generalized eigenvalue problem and the least squares problem in Section 3, discuss extensions based on the established equivalence result in Section 4, present the efficient implementation in Section 5, report empirical results in Section 6, and conclude this paper in Section 7.

2. Background and Related Work

In this section, we present a class of generalized eigenvalue problems studied in this paper. The least squares formulation is briefly reviewed.

2.1. A Class of Generalized Eigenvalue Problems

We consider a class of generalized eigenvalue problems in the following form:

$$X\mathcal{S}X^T w = \lambda XX^T w, \tag{1}$$

where $X \in \mathbb{R}^{d\times n}$ represents the data matrix and $\mathcal{S} \in \mathbb{R}^{n\times n}$ is symmetric and positive semi-definite. The generalized eigenvalue problem in Eq. (1) is often reformulated as the following eigenvalue problem:

$$(XX^T)^\dagger X\mathcal{S}X^T w = \lambda w, \tag{2}$$

where $(XX^T)^\dagger$ is the pseudoinverse of XX^T. In general, we are interested in eigenvectors corresponding to nonzero eigenvalues. It turns out that many machine learning techniques can be formulated in the form of Eqs. (1) and (2).

2.2. Examples of Generalized Eigenvalue Problems

We briefly review several algorithms that involve a generalized eigenvalue problem in the general form of Eq. (1). They include Canonical Correlation Analysis, Partial Least Squares, Linear Discriminant Analysis, and Hypergraph Spectral Learning. For supervised learning methods, the label information is encoded in the matrix $Y = [y_1, y_2, \cdots, y_n] \in \mathbb{R}^{k\times n}$, where $y_i(j) = 1$ if x_i belongs to class j and $y_i(j) = 0$ otherwise.

Canonical Correlation Analysis

In CCA (Hotelling, 1936), two different representations, X and Y, 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. Denote the projection vector for X by $w_x \in \mathbb{R}^d$, and assume that YY^T is nonsingular. It can be verified that w_x is the first principal eigenvector of the following generalized eigenvalue problem:

$$XY^T(YY^T)^{-1}YX^T w_x = \lambda XX^T w_x. \tag{3}$$

Multiple projection vectors can be obtained simultaneously by computing the first ℓ principal eigenvectors of the generalized eigenvalue problem in Eq. (3). It can be observed that CCA is in the form of the generalized eigenvalue problem in Eq. (1) with $\mathcal{S} = Y^T(YY^T)^{-1}Y$.

Partial Least Squares

In contrast to CCA, Orthonormalized PLS (OPLS), a variant of PLS (Rosipal & Krämer, 2006), computes orthogonal score vectors by maximizing the covariance between X and Y. It solves the following generalized eigenvalue problem:

$$XY^TYX^T w = \lambda XX^T w. \tag{4}$$

It can be observed that Orthonormalized PLS involves a generalized eigenvalue problem in Eq. (1) with $\mathcal{S} = Y^TY$.

Hypergraph Spectral Learning

A hypergraph (Agarwal et al., 2006) is a generalization of the traditional graph in which the edges (a.k.a. hyperedges) are arbitrary non-empty subsets of the vertex set. HSL (Sun et al., 2008a) employs a hypergraph to capture the correlation information among different labels for improved classification performance in multi-label learning. It has been shown that given the normalized Laplacian $\mathcal{L}_H$ for the constructed hypergraph, HSL invovles the following generalized eigenvalue problem:

$$X\mathcal{S}X^T w = \lambda(XX^T)w, \text{ where } \mathcal{S} = I - \mathcal{L}_H. \tag{5}$$

It has been shown that for many existing definitions of the Laplacian $\mathcal{L}_H$, the resulting matrix S is symmetric and positive semi-definite, and it can be decomposed as $\mathcal{S} = HH^T$, where $H \in \mathbb{R}^{n\times k}$.

Linear Discriminant Analysis

LDA is a supervised dimensionality reduction technique. The transformation in LDA is obtained by maximizing the ratio of the inter-class distance to the intra-class distance. It is known that CCA is equivalent to LDA for multi-class problems. Thus, the $\mathcal{S}$ matrix can be derived similarly.

2.3. Least Squares for Regression and Classification

Least squares is a classical technique for both regression and classification (Bishop, 2006). In 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 that both the observations and the targets are centered, then the intercept can be eliminated. In this case, the weight matrix $W \in \mathbb{R}^{d\times k}$ can be computed by minimizing the following sum-of-squares error function:

$$\min_W \sum_{i=1}^n \|W^T x_i - t_i\|_2^2 = \|W^T X - T\|_F^2, \tag{6}$$

where $T = [t_1, \cdots, t_n]$ is the target matrix. It is well-known that the optimal solution W_{ls} is given by

$$W_{ls} = (XX^T)^\dagger XT^T. \tag{7}$$

Least squares can also be applied for classification problems. 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 $k \geq 2$. 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 the class indicator matrix (Hastie et al., 2001; Ye, 2007). In contrast to the generalized eigenvalue problem, the least squares problem can be solved efficiently using iterative conjugate gradient algorithms (Golub & Van Loan, 1996; Paige & Saunders, 1982).

3. Generalized Eigenvalue Problem versus Least Squares Problem

In this section, we investigate the relationship between the generalized eigenvalue problem in Eq. (1) or its equivalent formulation in Eq. (2) and the least squares formulation. In particular, we show that under a mild condition[1], the eigenvalue problem in Eq. (2) can be formulated as a least squares problem with a specific target matrix.

3.1. Matrix Orthonormality Property

For convenience of presentation, we define matrices C_X and C_S as follows:

$$C_X = XX^T \in \mathbb{R}^{d\times d}, C_S = X\mathcal{S}X^T \in \mathbb{R}^{d\times d}.$$

The eigenvalue problem in Eq. (2) can then be expressed as

$$C_X^\dagger C_S w = \lambda w. \tag{8}$$

Recall that $\mathcal{S}$ is symmetric and positive semi-definite, thus it can be decomposed as

$$\mathcal{S} = HH^T, \tag{9}$$

where $H \in \mathbb{R}^{n\times s}$, and $s \leq n$. For most examples discussed in Section 2.2, the closed-form of H can be obtained and $s = k \ll n$. Since X is centered, i.e., $Xe = 0$, we have $XC = X$, where $C = I - \frac{1}{n}ee^T$ is the centering matrix satisfying $C^T = C$, and $Ce = 0$. It follows that

$$\begin{aligned} C_S &= X\mathcal{S}X^T = (XC)\mathcal{S}(XC)^T = X(C\mathcal{S}C^T)X^T \\ &= X(C^T\mathcal{S}C)X^T = X\tilde{\mathcal{S}}X^T, \end{aligned} \tag{10}$$

where $\tilde{\mathcal{S}} = C^T\mathcal{S}C$. Note that

$$e^T\tilde{\mathcal{S}}e = e^TC^T\mathcal{S}Ce = 0. \tag{11}$$

Thus, we can assume that $e^T\mathcal{S}e = 0$, that is, both columns and rows of $\mathcal{S}$ are centered. Before presenting the main results, we have the following lemmas.

[1] It states that $\{x_i\}_{i=1}^n$ are linearly independent before centering, i.e., rank$(X) = n - 1$ after the data is centered (of zero mean).

Lemma 1. *Assume that $e^T \mathcal{S} e = 0$. Let H be defined in Eq. (9). Let $HP = QR$ be the QR decomposition of H with column pivoting, where $Q \in \mathbb{R}^{n \times r}$ has orthonormal columns, $R \in \mathbb{R}^{r \times s}$ is upper triangular, $r = rank(H) \leq s$, and P is a permutation matrix. Then we have $Q^T e = 0$.*

Proof. Since P is a permutation matrix, we have $P^T P = I$. The result follows since $e^T \mathcal{S} e = e^T H H^T e = e^T Q R P^T P R^T Q^T e = e^T Q R R^T Q^T e = 0$ and RR^T is positive definite. □

Lemma 2. *Let $A \in \mathbb{R}^{m \times (m-1)}$ and $B \in \mathbb{R}^{m \times p}$ $(p \leq m)$ be two matrices satisfying $A^T e = 0$, $A^T A = I_{m-1}$, $B^T B = I_p$, and $B^T e = 0$. Let $F = A^T B$. Then $F^T F = I_p$.*

Proof. Define the orthogonal matrix A_x as $A_x = \left[A, \frac{1}{\sqrt{m}} e\right] \in \mathbb{R}^{m \times m}$. We have

$$I_m = A_x A_x^T = AA^T + \frac{1}{m} ee^T \Leftrightarrow AA^T = I_m - \frac{1}{m} ee^T.$$

Since $B^T e = 0$ and $B^T B = I_p$, we obtain

$$\begin{aligned} F^T F &= B^T A A^T B = B^T \left(I_m - \frac{1}{m} ee^T \right) B \\ &= B^T B - \frac{1}{m} (B^T e)(B^T e)^T = I_p. \end{aligned}$$

□

Let $R = U_R \Sigma_R V_R^T$ be the thin Singular Value Decomposition (SVD) of $R \in \mathbb{R}^{r \times s}$, where $U_R \in \mathbb{R}^{r \times r}$ is orthogonal, $V_R \in \mathbb{R}^{s \times r}$ has orthonormal columns, and $\Sigma_R \in \mathbb{R}^{r \times r}$ is diagonal. It follows that $(QU_R)^T (QU_R) = I_r$, and the SVD of $\mathcal{S}$ can be derived as follows:

$$\begin{aligned} \mathcal{S} &= HH^T = QRR^T Q^T = QU_R \Sigma_R^2 U_R^T Q^T \\ &= (QU_R) \Sigma_R^2 (QU_R)^T. \end{aligned} \tag{12}$$

Assume that the columns of X are centered, i.e., $Xe = 0$, and $rank(X) = n - 1$. Let

$$X = U \Sigma V^T = U_1 \Sigma_1 V_1^T$$

be the SVD of X, where U and V are orthogonal, $\Sigma \in \mathbb{R}^{d \times n}$, $U_1 \Sigma_1 V_1^T$ is the compact SVD of X, $U_1 \in \mathbb{R}^{d \times (n-1)}$, $V_1 \in \mathbb{R}^{n \times (n-1)}$, and $\Sigma_1 \in \mathbb{R}^{(n-1) \times (n-1)}$ is diagonal. Define M_1 as

$$M_1 = V_1^T (QU_R) \in \mathbb{R}^{(n-1) \times r}. \tag{13}$$

We can show that the columns of M_1 are orthonormal as summarized in the following lemma:

Lemma 3. *Let M_1 be defined as above. Then $M_1^T M_1 = I_r$.*

Proof. Since $Xe = 0$, we have $V_1^T e = 0$. Also note that $V_1^T V_1 = I_{n-1}$. Recall that $(QU_R)^T (QU_R) = I_r$ and $(QU_R)^T e = U_R^T Q^T e = 0$ from Lemma 1. It follows from Lemma 2 that $M_1^T M_1 = I_r$. □

3.2. The Equivalence Relationship

We first derive the solution to the eigenvalue problem in Eq. (2) in the following theorem:

Theorem 1. *Let U_1, Σ_1, V_1, Q, Σ_R, and U_R be defined as above. Assume that the columns of X are centered, i.e., $Xe = 0$, and $rank(X) = n - 1$. Then the nonzero eigenvalues of the problem in Eq. (2) are $diag(\Sigma_R^2)$, and the corresponding eigenvectors are $W_{eig} = U_1 \Sigma_1^{-1} V_1^T Q U_R$.*

We summarize the main result of this section in the following theorem:

Theorem 2. *Assume that the class indicator matrix $\tilde{T}$ for least squares classification is defined as*

$$\tilde{T} = U_R^T Q^T \in \mathbb{R}^{r \times n}. \tag{14}$$

Then the solution to the least squares formulation in Eq. (6) is given by

$$W_{ls} = U_1 \Sigma_1^{-1} V_1^T Q U_R. \tag{15}$$

Thus, the eigenvalue problem and the least squares problem are equivalent.

The proofs of the above two theorems are given in the Appendix.

Remark 1. *The analysis in (Ye, 2007; Sun et al., 2008b) is based on a key assumption that the H matrix in $\mathcal{S} = HH^T$ as defined in Eq. (9) has orthonormal columns, which is the case for LDA and CCA. However, this is in general not true, e.g., the H matrix in OPLS and HSL. The equivalence result established in this paper significantly improves previous work by relaxing this assumption.*

The matrix W_{eig} is applied for dimensionality reduction (projection) for all examples of Eq. (1). The weight matrix W_{ls} in least squares can also be used for dimensionality reduction. If $\tilde{T} = Q^T$ is used as the class indicator matrix, the weight matrix becomes $\tilde{W}_{ls} = U_1 \Sigma_1^{-1} V_1^T Q$. Thus, the difference between W_{eig} and $\tilde{W}_{ls}$ is the orthogonal matrix U_R. Note that the Euclidean distance is invariant to any orthogonal transformation. If a classifier, such as K-Nearest-Neighbor (KNN) and linear Support Vector Machines

Algorithm 1 Efficient Implementation via LSQR

```
  Input: $X$, $H$
  Compute the QR decomposition of $H$: $HP = QR$.
  Compute the SVD of $R$: $R = U_R \Sigma_R V_R^T$.
  Compute the class indicator matrix $T = U_R^T Q^T$.
  Regress $X$ onto $T$ using LSQR.
```

(SVM) (Schölkopf & Smola, 2002) based on the Euclidean distance, is applied on the dimensionality-reduced data via W_{eig} and $\tilde{W}_{ls}$, they will achieve the same classification performance.

In some cases, the number of nonzero eigenvalues, i.e. r, is large (comparable to n). It is common to use the top eigenvectors corresponding the largest $\ell < r$ eigenvalues as in PCA. From Theorems 1 and 2, if we keep the top ℓ singular vectors of $\mathcal{S}$ as the class indicator matrix, the equivalence relationship between the generalized eigenvalue problem and the least squares problem holds, as summarized below:

Corollary 1. *The top $\ell < r$ eigenvectors in Eq. (2) can be computed by solving a least squares problem with the top ℓ singular vectors of $\mathcal{S}$ employed as the class indicator matrix.*

4. Extensions

Based on the established equivalence relationship, the original generalized eigenvalue problem in Eq. (1) can be extended using the regularization technique. Regularization is commonly used to control the complexity of the model and improve the generalization performance. By using the target matrix $\tilde{T}$ in Eq. (14), we obtain the 2-norm regularized least squares formulation by minimizing the following objective function: $L_2 = \|W^T X - \tilde{T}\|_F^2 + \gamma \sum_{j=1}^{r} \|w_j\|_2^2$, where $W = [w_1, \cdots, w_r]$ and $\gamma > 0$ is the regularization parameter. We can then apply the LSQR algorithm, a conjugate gradient type method proposed in (Paige & Saunders, 1982) for solving large-scale sparse least-squares problems.

It is known that model sparsity can often be achieved by applying the L_1-norm regularization (Donoho, 2006; Tibshirani, 1996). This has been introduced into the least squares formulation and the resulting model is called lasso (Tibshirani, 1996). Based on the established equivalence relationship between the original generalized eigenvalue problem and the least squares formulation, we derive the 1-norm regularized least squares formulation by minimizing the following objective function: $L_1 = \|W^T X - \tilde{T}\|_F^2 + \gamma \sum_{j=1}^{r} \|w_j\|_1$, for some tuning parameter $\gamma > 0$ (Tibshirani, 1996). The lasso can be solved efficiently using the state-of-the-art algorithms (Friedman et al., 2007; Hale et al., 2008). In addition, the entire solution path for all values of γ can be obtained by applying the Least Angle Regression algorithm (Efron et al., 2004).

5. Efficient Implementation via LSQR

Recall that we deal with the generalized eigenvalue problem in Eq. (1), although in our theoretical derivation an equivalent eigenvalue problem in Eq. (2) is used instead. Large-scale generalized eigenvalue problems are known to be much harder than regular eigenvalue problems. There are two options to transform the problem in Eq. (1) into a standard eigenvalue problem (Saad, 1992): (i) factor XX^T; and (ii) employ the standard Lanczos algorithm for the matrix $(XX^T)^{-1} X \mathcal{S} X^T$ using the XX^T inner product. The second option has its own issue for singular matrices, which is the case for high-dimensional problems with a small regularization. Thus, in this paper we factor XX^T and solve a symmetric eigenvalue problem using the Lanczos algorithm.

The equivalent least squares formulation leads to an efficient implementation. The pseudo-code of the algorithm is given in Algorithm 1.

The complexity of the QR decomposition in the first step is $O(nk^2)$. Note that k is the number of classes, and $k \ll n$. The SVD of R costs $O(k^3)$. In the last step, we solve k least squares problems. In our implementation, we use the LSQR algorithm proposed in (Paige & Saunders, 1982), which is a conjugate gradient method for solving large-scale least squares problems. In practice, the original data matrix $X \in \mathbb{R}^{d \times n}$ may be sparse in many applications such as text document modeling. Note that the centering of X is necessary in some techniques such as CCA. However, X is no longer sparse after centering. In order to keep the sparsity of X, the vector x_i is augmented by an additional component as $\tilde{x}_i^T = [1, x_i^T]$, and the extended X is denoted as $\tilde{X} \in \mathbb{R}^{(d+1) \times n}$. This new component acts as the bias for least squares.

For dense data matrix, the overall computational cost of each iteration of LSQR is $O(3n + 5d + 2dn)$ (Paige & Saunders, 1982). Since the least squares problems are solved k times, the overall cost of LSQR is $O(Nk(3n + 5d + 2dn))$, where N is the total number of iterations. When the matrix $\tilde{X}$ is sparse, the cost is significantly reduced. Let the number of nonzero elements in $\tilde{X}$ be z, then the overall cost of LSQR is reduced to $O(Nk(3n + 5d + 2z))$. In summary, the total time complexity for solving the least squares formulation via LSQR is $O(nk^2 + Nk(3n + 5d + 2z))$.

6. Experiments

In this section, we report experimental results that validate the established equivalence relationship. We also demonstrate the efficiency of the proposed least squares extensions.

Experimental Setup The techniques involved can be divided into two categories: (1) CCA, PLS, and HSL for multi-label learning; and (2) LDA for multi-class learning. We use both multi-label [yeast (Elisseeff & Weston, 2001) and Yahoo (Kazawa et al., 2005)] and multi-class [USPS (Hull, 1994)] data sets in the experiments. For the Yahoo data set, we preprocess them using the feature selection method proposed in (Yang & Pedersen, 1997). The statistics of the data sets are summarized in Table 1.

For each data set, a transformation matrix is learned from the training set, and it is then used to project the test data onto a lower-dimensional space. The linear Support Vector Machine (SVM) is applied for classification. The Receiver Operating Characteristic (ROC) score and classification accuracy are used to evaluate the performance of multi-label and multi-class tasks, respectively. Ten random partitions of the data sets into training and test sets are generated in the experiments, and the mean performance over all labels and all partitions are reported. All experiments are performed on a PC with Intel Core 2 Duo T7200 2.0G CPU and 2G RAM.

For the generalized eigenvalue problem in Eq. (1), a regularization term is commonly added as $(XX^T + \gamma I)$ to cope with the singularity problem of XX^T. We name the resulting regularized method using a prefix "r" before the corresponding method, e.g., "rStar"[2]. The equivalent least squares formulations are named using a prefix "LS" such as "LS-Star", and the resulting 1-norm and 2-norm regularized formulations are named by adding subscripts 1 and 2, respectively, e.g., "LS-Star$_1$" and "LS-Star$_2$". All algorithms were implemented in Matlab.

Evaluation of the Equivalence Relationship In this experiment, we show the equivalence relationship between the generalized eigenvalue problem and its corresponding least squares formulation for all techniques discussed in this paper. We observe that for all data sets, when the data dimensionality d is larger than the sample size n, rank$(X) = n - 1$ is likely to hold. We also observe that the generalized eigenvalue problem and its corresponding least squares formulation achieve the same performance when rank$(X) = n - 1$ holds. These are consistent with the theoretical results in Theorems 1 and 2.

Table 1. Statistics of the test data sets: n is number of data points, d is the data dimensionality, and k is the number of labels (classes).

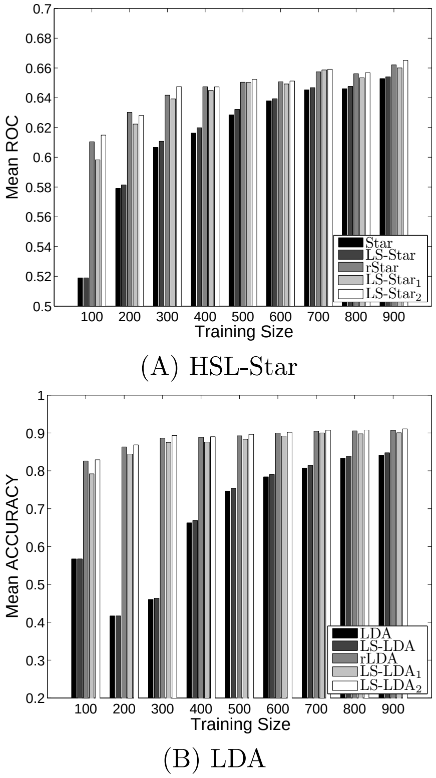
Data Set	n	d	k
Yeast	2417	103	14
Yahoo\Arts&Humanities	3712	23146	26
USPS	9298	256	10

We also compare the performance of the two formulations when the assumption in Theorems 1 and 2 is violated. Figure 1 shows the performance of different formulations when the size of training set varies from 100 to 900 with a step size about 100 on the yeast data set and the USPS data set. Due to the space constraint, only the results from HSL based on the star expansion and LDA are presented. We can observe from the figure that when n is small, the assumption in Theorem 1 holds and the two formulations achieve the same performance; when n is large, the assumption in Theorem 1 does not hold and the two formulations achieve different performance, although the difference is always very small in the experiment. We can also observe from Figure 1 that the regularized methods outperform the unregularized ones, which validates the effectiveness of regularization.

Evaluation of Scalability In this experiment, we compare the scalability of the original generalized eigenvalue problem and the equivalent least squares formulation. Since regularization is commonly employed in practice, we compare the regularized version of the generalized eigenvalue problem and its corresponding 2-norm regularized least squares formulation. The least squares problem is solved by the LSQR algorithm (Paige & Saunders, 1982).

The computation time of the two formulations on the high-dimensional multi-label Yahoo data set is shown in Figure 2 (top figure), where the data dimensionality increases and the training sample size is fixed at 1000. Only the results from CCA and HSL are presented due to the space constraint. It can be observed that the computation time for both algorithms increases steadily as the data dimensionality increases. However, the computation time of the least squares formulation is substantially less than that of the original one. We also evaluate the scalability of the two formulations in terms of the training sample size. Figure 2 (bottom figure) shows the computation time of the two formulations on the Yahoo data set as the training sample size increases with the data dimensionality fixed at 5000. We can also observe that the least squares formulation is much more scalable than the original one.

[2] rStar denotes the regularized HSL formulation when the star expansion (Agarwal et al., 2006) is used to form the hypergraph Laplacian.



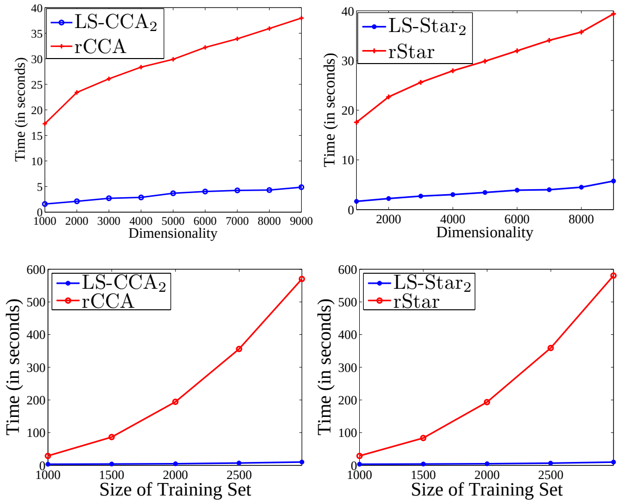
(A) HSL-Star

(B) LDA

Figure 1. Comparison of different formulations in terms of the ROC score/accuracy for different techniques on the yeast data set and the USPS data set. HSL is applied on the yeast data set, and LDA is applied on the USPS data set. For regularized algorithms, the optimal value of γ is estimated from $\{1e-6, 1e-4, 1e-2, 1, 10, 100, 1000\}$ using cross validation.

7. Conclusions and Future Work

In this paper, we study the relationship between a class of generalized eigenvalue problems in machine learning and the least squares formulation. In particular, we show that a class of generalized eigenvalue problems in machine learning can be reformulated as a least squares problem under a mild condition, which generally holds for high-dimensional data. The class of problems include CCA, PLS, HSL, and LDA. Based on the established equivalence relationship, various regularization techniques can be employed to improve the generalization ability and induce sparsity in the resulting model. In addition, the least squares formulation results in the efficient implementation based on the iterative conjugate gradient type algorithms such as LSQR. Our experimental results confirm the established equivalence relationship. Results also show that the performance of the least squares formulation and the original generalized eigenvalue problem is very close even when the assumption is violated. Our experiments also demonstrate the effectiveness and scalability of the least squares extensions.



(A) CCA (B) HSL-Star

Figure 2. Computation time (in seconds) of the generalized eigenvalue problem and the corresponding least squares formulation on the Yahoo\Arts&Humanities data set as the data dimensionality (top row) or the training sample size (bottom row) increases. The x-axis represents the data dimensionality (top) or the training sample size (bottom) and the y-axis represents the computation time.

Unlabeled data can be incorporated into the least squares formulation through the graph Laplacian, which captures the local geometry of the data (Belkin et al., 2006). We plan to investigate the effectiveness of this semi-supervised framework for the class of generalized eigenvalue problems studied in this paper. The equivalence relationship will in general not hold for low-dimensional data. However, it is common to map the low-dimensional data into a high-dimensional feature space through a nonlinear feature mapping induced by the kernel (Schölkopf & Smola, 2002). We plan to study the relationship between these two formulations in the kernel-induced feature space.

Acknowledgments

This work was supported by NSF IIS-0612069, IIS-0812551, CCF-0811790, and NGA HM1582-08-1-0016.

References

Agarwal, S., Branson, K., & Belongie, S. (2006). Higher order learning with graphs. *International Conference on Machine Learning* (pp. 17–24).

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, *7*, 2399–2434.

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

d'Aspremont, A., Ghaoui, L., Jordan, M., & Lanckriet,

G. (2004). A direct formulation for sparse PCA using semidefinite programming. *Neural Information Processing Systems* (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.

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

Elisseeff, A., & Weston, J. (2001). A kernel method for multi-labelled classification. *Neural Information Processing Systems* (pp. 681–687).

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, 302–332.

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

Hale, E., Yin, W., & Zhang, Y. (2008). Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM Journal on Optimization*, *19*, 1107–1130.

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

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

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*, 550–554.

Kazawa, H., Izumitani, T., Taira, H., & Maeda, E. (2005). Maximal margin labeling for multi-topic text categorization. *Neural Information Processing Systems* (pp. 649–656).

Paige, C. C., & Saunders, M. A. (1982). LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, *8*, 43–71.

Rosipal, R., & Krämer, N. (2006). Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection Techniques, Lecture Notes in Computer Science* (pp. 34–51).

Saad, Y. (1992). *Numerical methods for large eigenvalue problems.* New York, NY: Halsted Press.

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

Sun, L., Ji, S., & Ye, J. (2008a). Hypergraph spectral learning for multi-label classification. *ACM SIGKDD International Conference On Knowledge Discovery and Data Mining* (pp. 668–676).

Sun, L., Ji, S., & Ye, J. (2008b). A least squares formulation for canonical correlation analysis. *International Conference on Machine Learning* (pp. 1024–1031).

Tao, D., Li, X., Wu, X., & Maybank, S. (2009). Geometric mean for subspace selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*, 260–274.

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

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *International Conference on Machine Learning* (pp. 412–420).

Ye, J. (2007). Least squares linear discriminant analysis. *International Conference on Machine Learning* (pp. 1087–1094).

Appendix

Proof of Theorem 1

Proof. It follows from Lemma 3 that the columns of $M_1 \in \mathbb{R}^{(n-1)\times r}$ are orthonormal. Hence there exists $M_2 \in \mathbb{R}^{(n-1)\times(n-1-r)}$ such that $M = [M_1, M_2] \in \mathbb{R}^{(n-1)\times(n-1)}$ is orthogonal (Golub & Van Loan, 1996). We can derive the eigen-decomposition of $C_X^\dagger C_S$ as:

$$\begin{aligned}
&C_X^\dagger C_S = \left(XX^T\right)^\dagger X\mathcal{S}X^T \\
=&(U_1\Sigma_1^{-2}U_1^T)U_1\Sigma_1V_1^T\left(QU_R\right)\Sigma_R^2\left(QU_R\right)^T V_1\Sigma_1U_1^T \\
=&U_1\Sigma_1^{-1}V_1^T\left(QU_R\right)\Sigma_R^2\left(QU_R\right)^T V_1\Sigma_1U_1^T \\
=&U_1\Sigma_1^{-1}M_1\Sigma_R^2M_1^T\Sigma_1U_1^T \\
=&U\begin{bmatrix} I_{n-1} \\ 0 \end{bmatrix}\Sigma_1^{-1}\begin{bmatrix} M_1 & M_2 \end{bmatrix}\begin{bmatrix} \Sigma_R^2 & 0 \\ 0 & 0_{n-1-r} \end{bmatrix}\begin{bmatrix} M_1^T \\ M_2^T \end{bmatrix} \\
&\Sigma_1[I_{n-1}, 0]U^T \\
=&U\begin{bmatrix} I_{n-1} \\ 0 \end{bmatrix}\Sigma_1^{-1}M\begin{bmatrix} \Sigma_R^2 & 0 \\ 0 & 0_{n-1-r} \end{bmatrix}M^T\Sigma_1[I_{n-1}, 0]U^T \\
=&U\begin{bmatrix} \Sigma_1^{-1}M & \\ & I \end{bmatrix}\begin{bmatrix} \Sigma_R^2 & \\ & 0_{n-r} \end{bmatrix}\begin{bmatrix} M^T\Sigma_1 & \\ & I \end{bmatrix}U^T. \qquad (16)
\end{aligned}$$

There are r nonzero eigenvalues, which are $\text{diag}(\Sigma_R^2)$, and the corresponding eigenvectors are

$$W_{eig} = U_1\Sigma_1^{-1}M_1 = U_1\Sigma_1^{-1}V_1^TQU_R. \qquad (17)$$

□

Proof of Theorem 2

Proof. When $\tilde{T}$ is used as the class indicator matrix, it follows from Eq. (7) that the solution to the least squares problem is

$$\begin{aligned}
W_{ls} &= (XX^T)^\dagger X\tilde{T}^T = (XX^T)^\dagger XQU_R \\
&= U_1\Sigma_1^{-2}U_1^TU_1\Sigma_1V_1^TQU_R = U_1\Sigma_1^{-1}V_1^TQU_R.
\end{aligned}$$

It follows from Eq. (17) that $W_{ls} = W_{eig}$. □