

A Scalable Two-Stage Approach for a Class of Dimensionality Reduction Techniques

Liang Sun
Arizona State University
Tempe, AZ 85287, USA
sun.liang@asu.edu

Betul Ceran
Arizona State University
Tempe, AZ 85287, USA
betul@asu.edu

Jieping Ye
Arizona State University
Tempe, AZ 85287, USA
jieping.ye@asu.edu

ABSTRACT

Dimensionality reduction plays an important role in many data mining applications involving high-dimensional data. Many existing dimensionality reduction techniques can be formulated as a generalized eigenvalue problem, which does not scale to large-size problems. Prior work transforms the generalized eigenvalue problem into an equivalent least squares formulation, which can then be solved efficiently. However, the equivalence relationship only holds under certain assumptions without regularization, which severely limits their applicability in practice. In this paper, an efficient two-stage approach is proposed to solve a class of dimensionality reduction techniques, including Canonical Correlation Analysis, Orthonormal Partial Least Squares, Linear Discriminant Analysis, and Hypergraph Spectral Learning. The proposed two-stage approach scales linearly in terms of both the sample size and data dimensionality. The main contributions of this paper include (1) we rigorously establish the equivalence relationship between the proposed two-stage approach and the original formulation without any assumption; and (2) we show that the equivalence relationship still holds in the regularization setting. We have conducted extensive experiments using both synthetic and real-world data sets. Our experimental results confirm the equivalence relationship established in this paper. Results also demonstrate the scalability of the proposed two-stage approach.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications - Data Mining

General Terms

Algorithm

Keywords

Dimensionality reduction, generalized eigenvalue problem, least squares, regularization, scalability

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD'10, July 25–28, 2010, Washington, DC, USA.

Copyright 2010 ACM 978-1-4503-0055-110/07 ...\$10.00.

1. INTRODUCTION

Recent technological innovations have allowed us to collect massive amounts of data with a large number of features. One of the key issues in such data analysis is the *curse of dimensionality* [5], i.e., an enormous number of samples is required to perform accurate prediction on problems with large numbers of features. Dimensionality reduction, which extracts a small number of features by removing the irrelevant, redundant, and noisy information, is an effective way to overcome the curse of dimensionality. Many dimensionality reduction algorithms have been proposed in the past, including Canonical Correlation Analysis (CCA) [15], Partial Least Squares (PLS) [21], Linear Discriminant Analysis (LDA) [6] and Hypergraph Spectral Learning (HSL) [24]. A common characteristic of these algorithms is that they can be formulated as a generalized eigenvalue problem. Although well-established algorithms in numerical linear algebra exist to solve generalized eigenvalue problems [12, 22], they are in general computationally expensive to solve and hence may not scale to large-size problems.

There have been several recent attempts to improve the scalability of dimensionality reduction techniques [8, 9, 16, 24, 25, 26, 32, 33]. The key idea is to transform the generalized eigenvalue problem into an equivalent least squares formulation, which can be solved efficiently using existing algorithms such as the iterative conjugate gradient algorithm [12, 19, 20]. In particular, an equivalent least squares formulation has been developed for a class of dimensionality reduction techniques in [26]. However, it suffers from several drawbacks which limits its applicability in practice. First, the equivalent transformation relies on a key assumption that all the data points are linear independent. This assumption tends to hold for high-dimensional data, but it is likely to fail for (relatively) low-dimensional data. Secondly, the equivalence relationship between the least squares formulation and the original formulation does not hold when the regularization is employed. However, regularization has been shown to be critical in many data mining and machine learning algorithms including support vector machines [23].

In this paper, we propose an efficient two-stage approach to solve a class of dimensionality reduction techniques, including CCA, PLS, LDA and HSL. In the first stage we solve a least squares problem using the iterative conjugate gradient algorithm [12, 19, 20]. The distinct property of this stage is its low time complexity. In the second stage, the original data is projected onto a low-dimensional space, and then we solve a generalized eigenvalue problem with a significantly reduced size. The proposed two-stage approach

scales linearly in terms of both the sample size and data dimensionality, thus applicable for large-size problems. The main contributions of this paper include:

- We rigorously prove the equivalence relationship between the two-stage approach and the direct approach which solves the generalized eigenvalue problem directly. Compared with previous work, the two-stage approach does not require any assumption and can be applied in all cases.
- We show that the two-stage approach can be further extended to the regularization setting. The equivalence relationship is also rigorously proved in this case.

We have conducted extensive experiments to evaluate the proposed two-stage approach using both synthetic and real-world benchmark data sets. Our experimental results confirm the equivalence relationship established in this paper. Results also demonstrate the scalability of the proposed two-stage approach.

Organization The rest of the paper is organized as follows. Section 2 briefly reviews the class of dimensionality reduction techniques discussed in the paper. In Section 3, we present the two-stage approach, and establish the equivalence relationship. We further extend the two-stage approach to the regularization setting in Section 4. A comprehensive empirical study is reported in Section 5. Finally, we conclude in Section 6.

Notations Throughout the paper, all matrices are boldface uppercase, and vectors are boldface lowercase. n is the number of samples, d is the data dimensionality, and k is the number of classes (or labels). The i th sample is denoted as $\mathbf{x}_i \in \mathbb{R}^d$, and its corresponding label is denoted as $\mathbf{y}_i \in \mathbb{R}^k$. $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ represents the data matrix, and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] \in \mathbb{R}^{k \times n}$ is the matrix representation for label information. $\mathbf{S} \in \mathbb{R}^{n \times n}$ is a symmetric and positive semi-definite matrix. \mathbf{I}_p is the p -by- p identity matrix, and $\mathbf{1}_p$ is a vector of all ones with length p . Note that the subscript p may be omitted when the size is clear from the context.

2. A CLASS OF DIMENSIONALITY REDUCTION TECHNIQUES

The class of generalized eigenvalue problems considered in this paper exhibit the following form:

$$\mathbf{S}\mathbf{X}\mathbf{X}^T\mathbf{w} = \lambda\mathbf{X}\mathbf{X}^T\mathbf{w}, \quad (1)$$

where $\mathbf{S} \in \mathbb{R}^{n \times n}$ is a symmetric and positive semi-definite matrix. In general, we are interested in the principal eigenvectors corresponding to nonzero eigenvalues. The generalized eigenvalue problem in Eq. (1) is often reformulated as the following eigenvalue problem

$$(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{X}\mathbf{S}\mathbf{X}^T\mathbf{w} = \lambda\mathbf{w}, \quad (2)$$

where $(\mathbf{X}\mathbf{X}^T)^\dagger$ is the Moore-Penrose pseudoinverse of $\mathbf{X}\mathbf{X}^T$. In addition, the generalized eigenvalue problem in Eq. (1) can also be formulated as an optimization problem:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \text{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{S}\mathbf{X}^T\mathbf{W}) \\ \text{s. t.} \quad & \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}. \end{aligned} \quad (3)$$

In the following derivation, we generally use the formulation in Eq.(2).

Many existing dimensionality reduction techniques exhibit the form in Eq. (1) or (2). In particular, the matrix \mathbf{S} can be represented in the following form:

$$\mathbf{S} = \mathbf{H}\mathbf{H}^T, \quad (4)$$

where $\mathbf{H} \in \mathbb{R}^{n \times k}$ is often constructed from the label information in supervised learning.

To control the model complexity and avoid the singularity of $\mathbf{X}\mathbf{X}^T$, a regularization term $\gamma\mathbf{I}_d$ with $\gamma > 0$ is added to $\mathbf{X}\mathbf{X}^T$ in Eq. (1), leading to the following generalized eigenvalue problem:

$$\mathbf{X}\mathbf{S}\mathbf{X}^T\mathbf{w} = \lambda(\mathbf{X}\mathbf{X}^T + \gamma\mathbf{I}_d)\mathbf{w}. \quad (5)$$

We briefly review several algorithms involving the generalized eigenvalue problem in the form of Eq. (1). Specifically, they include Canonical Correlation Analysis, Orthonormalized Partial Least Squares, Hypergraph Spectral Learning, and Linear Discriminant Analysis. For supervised learning methods, the label information is encoded in the matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] \in \mathbb{R}^{k \times n}$, where $\mathbf{y}_i(j) = 1$ if \mathbf{x}_i belongs to class j and $\mathbf{y}_i(j) = 0$ otherwise.

2.1 Canonical Correlation Analysis

In Canonical Correlation Analysis (CCA) [4, 13, 15], two different representations, \mathbf{X} and \mathbf{Y} , of the same set of objects are given, and a projection is computed for each representation such that the correlation coefficient

$$\rho = \frac{\mathbf{w}_x^T\mathbf{X}\mathbf{Y}^T\mathbf{w}_y}{\sqrt{(\mathbf{w}_x^T\mathbf{X}\mathbf{X}^T\mathbf{w}_x)(\mathbf{w}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{w}_y)}} \quad (6)$$

is maximized in the dimensionality-reduced space, where $\mathbf{w}_x \in \mathbb{R}^d$ and $\mathbf{w}_y \in \mathbb{R}^k$ are projection vectors for \mathbf{X} and \mathbf{Y} , respectively. Assume that $\mathbf{Y}\mathbf{Y}^T$ is nonsingular. It can be verified that \mathbf{w}_x is the first principal eigenvector of the following generalized eigenvalue problem:

$$\mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}\mathbf{X}^T\mathbf{w}_x = \lambda\mathbf{X}\mathbf{X}^T\mathbf{w}_x. \quad (7)$$

Multiple projection vectors can be obtained simultaneously by computing the first ℓ principal eigenvectors of the generalized eigenvalue problem in Eq. (7). It can be observed that CCA is in the form of the generalized eigenvalue problem in Eq. (1) with $\mathbf{S} = \mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}$ and $\mathbf{H} = \mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1/2}$.

2.2 Orthonormalized Partial Least Squares

Partial least squares (PLS) [29] is a family of methods for modeling relations between two sets of variables. In this paper, the Orthonormalized Partial Least Squares (OPLS) [2, 30], a popular variant of PLS, is studied. In contrast to CCA, OPLS computes orthogonal score vectors by maximizing the covariance between \mathbf{X} and \mathbf{Y} . It solves the following generalized eigenvalue problem:

$$\mathbf{X}\mathbf{Y}^T\mathbf{Y}\mathbf{X}^T\mathbf{w} = \lambda\mathbf{X}\mathbf{X}^T\mathbf{w}. \quad (8)$$

It follows from Eq. (8) that orthonormalized PLS involves a generalized eigenvalue problem in Eq. (1) with $\mathbf{S} = \mathbf{Y}^T\mathbf{Y}$ and $\mathbf{H} = \mathbf{Y}^T$.

2.3 Hypergraph Spectral Learning

Hypergraph Spectral Learning (HSL) [24] is a dimensionality reduction technique for multi-label classification. A hypergraph [1] 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 employs a hypergraph

Algorithm 1 The Two-Stage Approach without Regularization

Input: \mathbf{X}, \mathbf{H}

Output: \mathbf{W}

Stage 1: Solve the following least squares problem:

$$\min_{\mathbf{W}_1} \|\mathbf{W}_1^T \mathbf{X} - \mathbf{H}^T\|_F^2. \quad (10)$$

Stage 2: Compute $\tilde{\mathbf{X}} = \mathbf{W}_1^T \mathbf{X}$, and solve the following optimization problem:

$$\begin{aligned} \max_{\mathbf{W}_2} \quad & \text{Tr}(\mathbf{W}_2^T \tilde{\mathbf{X}} \mathbf{H} \mathbf{H}^T \tilde{\mathbf{X}}^T \mathbf{W}_2) \\ \text{s. t.} \quad & \mathbf{W}_2^T \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T \mathbf{W}_2 = \mathbf{I}_\ell. \end{aligned} \quad (11)$$

Compute $\mathbf{W} = \mathbf{W}_1 \mathbf{W}_2$ as the final solution.

to capture the correlation information among different labels for improved classification performance in multi-label learning. Specifically, HSL constructs a hyperedge for each label, and includes all instances annotated with a common label into one hyperedge, thus capturing their joint similarity. Three different Laplacians have been proposed to capture the spectral properties of the hypergraph, including clique expansion [1], star expansion [1] and Zhou’s Laplacian [34]. It has been shown that given the normalized Laplacian \mathbf{L} for the constructed hypergraph, HSL involves the following generalized eigenvalue problem:

$$\mathbf{X} \mathbf{S} \mathbf{X}^T \mathbf{w} = \lambda (\mathbf{X} \mathbf{X}^T) \mathbf{w}, \text{ where } \mathbf{S} = \mathbf{I} - \mathbf{L}. \quad (9)$$

It has been shown that for all three Laplacian matrices, the resulting matrix \mathbf{S} is symmetric and positive semi-definite, and it can be represented as $\mathbf{S} = \mathbf{H} \mathbf{H}^T$, where $\mathbf{H} \in \mathbb{R}^{n \times k}$. Note that \mathbf{H} can be constructed from the label information in \mathbf{Y} explicitly without the matrix decomposition.

2.4 Linear Discriminant Analysis

As a supervised dimensionality reduction technique, Linear Discriminant Analysis (LDA) attempts to minimize the within-class variance while maximizing the between-class variance after the linear projection. It has been shown that the optimal linear projection consists of the top eigenvectors of $\mathbf{S}_t^\dagger \mathbf{S}_b$ corresponding to nonzero eigenvalues [11, 31], where \mathbf{S}_t is the total covariance matrix and \mathbf{S}_b is the between-class covariance matrix. The matrices \mathbf{S}_t and \mathbf{S}_b are defined as follows:

$$\mathbf{S}_t = \frac{1}{n} \mathbf{X} \mathbf{X}^T, \mathbf{S}_b = \frac{1}{n} \sum_{j=1}^k n_j \mathbf{c}^{(j)} \mathbf{c}^{(j)T}, \quad (12)$$

where $\mathbf{c}^{(j)}$ is the centroid of the j th class and we assume that $\sum_{i=1}^n \mathbf{x}_i = 0$. We also assume that the data matrix \mathbf{X} is partitioned into k classes as $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_k]$, where $\mathbf{X}_j \in \mathbb{R}^{d \times n_j}$ corresponds to the data points from the j th class, n_j is the size of the j th class, and $\sum_{j=1}^k n_j = n$. Note that $\mathbf{c}^{(j)} = \frac{1}{n_j} \mathbf{X}_j \mathbf{1}_j$, where $\mathbf{1}_j$ is a vector of all ones with length n_j . Thus, we have

$$\begin{aligned} n \mathbf{S}_b &= \sum_{j=1}^k \frac{1}{n_j} \mathbf{X}_j \mathbf{1}_j \mathbf{1}_j^T \mathbf{X}_j^T = \sum_{j=1}^k \mathbf{X}_j \left(\frac{1}{n_j} \mathbf{1}_j \mathbf{1}_j^T \right) \mathbf{X}_j^T \\ &= \sum_{j=1}^k \mathbf{X}_j \mathbf{S}_j \mathbf{X}_j^T = \mathbf{X} \mathbf{S} \mathbf{X}^T, \end{aligned}$$

where $\mathbf{S}_j = \frac{1}{n_j} \mathbf{1}_j \mathbf{1}_j^T$, and \mathbf{S} is defined as $\text{diag}(\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_k)$ for LDA. Therefore, $\mathbf{S}_t^\dagger \mathbf{S}_b$ can also be formulated in the following form:

$$\mathbf{S}_t^\dagger \mathbf{S}_b = (\mathbf{X} \mathbf{X}^T)^\dagger (\mathbf{X} \mathbf{S} \mathbf{X}^T). \quad (13)$$

It can be verified that $\mathbf{S} = \mathbf{H} \mathbf{H}^T$, where

$$\mathbf{H} = \text{diag} \left(\frac{1}{\sqrt{n_1}} \mathbf{1}_1, \frac{1}{\sqrt{n_2}} \mathbf{1}_2, \dots, \frac{1}{\sqrt{n_k}} \mathbf{1}_k \right) \in \mathbb{R}^{n \times k}. \quad (14)$$

3. THE TWO-STAGE APPROACH WITHOUT REGULARIZATION

In this section, we present our two-stage approach and show that the two-stage approach is equivalent to the direct approach which solves the generalized eigenvalue problem in Eq. (1) directly.

3.1 The Algorithm

In the two-stage approach, we first solve a least squares problem by regressing \mathbf{X} on \mathbf{H}^T . In other words, \mathbf{H}^T can be considered as the “latent target” encoded by the label information \mathbf{Y} . After projecting the data matrix \mathbf{X} onto the subspace, we solve the resulting generalized eigenvalue problem by replacing the data matrix in Eq. (1) with the projected data matrix. Note that the data dimensionality is reduced dramatically after the projection, thus the resulting generalized eigenvalue problem in the second step can be solved efficiently. The two-stage approach is summarized in Algorithm 1.

3.2 Time Complexity Analysis

In our implementation, we apply the LSQR algorithm [19, 20], a conjugate gradient method for solving the least squares problem in the first stage. Previous studies have shown that LSQR is reliable even for ill-conditioned problems [20]. In addition, when the data matrix \mathbf{X} is sparse, the least squares problem can be solved very efficiently using LSQR. Note that the computational cost of each iteration of LSQR is $O(3n + 5d + 2dn)$ when \mathbf{X} is dense or $O(3n + 5d + 2z)$ when \mathbf{X} is sparse, where z is the number of nonzero entries in \mathbf{X} [20]. Since $\mathbf{H}^T \in \mathbb{R}^{n \times k}$, k least squares problems are solved simultaneously in the first stage of Algorithm 1, which implies that the total computational cost of the first stage is $O(Nk(3n + 5d + 2dn))$ using LSQR when \mathbf{X} is dense, where N is the total number of iterations. When the data matrix \mathbf{X} is sparse, the cost of LSQR is reduced to $O(Nk(3n + 5d + 2z))$. In the second stage, the cost of computing $\tilde{\mathbf{X}}$ is $O(ndk)$ when \mathbf{X} is dense or $O(kz)$ when \mathbf{X} is sparse. Since the size of $\tilde{\mathbf{X}}$ is significantly reduced, the cost of solving the optimization problem is $O(nk^2)$ in the second stage. The cost of combining \mathbf{W}_1 and \mathbf{W}_2 is $O(dk\ell)$, where ℓ ($\ell \leq k$) is the number of final projection vectors. Therefore, the total computational cost is $O(Nk(3n + 5d + 2z) + kz + nk^2 + dk^2)$ when \mathbf{X} is sparse.

3.3 Equivalence

Next we rigorously prove the equivalence relationship between the two-stage approach and the direct approach which solves the original eigenvalue problem in Eq.(2) directly.

Using the standard technique in linear algebra, the solution to the least squares problem in Eq. (10) is

$$\mathbf{W}_1 = (\mathbf{X} \mathbf{X}^T)^\dagger \mathbf{X} \mathbf{H} \in \mathbb{R}^{d \times k}. \quad (15)$$

Let the singular value decomposition (SVD) [12] of \mathbf{X} be

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T, \quad (16)$$

where $\mathbf{U} \in \mathbb{R}^{d \times d}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are orthogonal, $\mathbf{U}_1 \in \mathbb{R}^{d \times r}$ and $\mathbf{V}_1 \in \mathbb{R}^{n \times r}$ have orthonormal columns, $\mathbf{\Sigma} \in \mathbb{R}^{d \times n}$ and $\mathbf{\Sigma}_1 \in \mathbb{R}^{r \times r}$ are diagonal, and $r = \text{rank}(\mathbf{X})$. Then \mathbf{W}_1 can be represented as

$$\mathbf{W}_1 = \mathbf{U}_1\mathbf{\Sigma}_1^{-1}\mathbf{V}_1^T\mathbf{H}. \quad (17)$$

Then $\tilde{\mathbf{X}}$ can be represented as

$$\tilde{\mathbf{X}} = \mathbf{W}_1^T\mathbf{X} = \mathbf{H}^T\mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^T\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T = \mathbf{H}^T\mathbf{V}_1\mathbf{V}_1^T, \quad (18)$$

and

$$\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \mathbf{H}^T\mathbf{V}_1\mathbf{V}_1^T\mathbf{V}_1\mathbf{V}_1^T\mathbf{H} = \mathbf{H}^T\mathbf{V}_1\mathbf{V}_1^T\mathbf{H}.$$

Thus, the optimization problem in Eq. (11) can be simplified into the following form:

$$\begin{aligned} \max_{\mathbf{W}_2} \quad & \text{Tr}(\mathbf{W}_2^T\mathbf{H}^T\mathbf{V}_1\mathbf{V}_1^T\mathbf{H}\mathbf{H}^T\mathbf{V}_1\mathbf{V}_1^T\mathbf{H}\mathbf{W}_2) \\ \text{s. t.} \quad & \mathbf{W}_2^T\mathbf{H}^T\mathbf{V}_1\mathbf{V}_1^T\mathbf{H}\mathbf{W}_2 = \mathbf{I}_\ell. \end{aligned}$$

Denote

$$\mathbf{A} = \mathbf{H}^T\mathbf{V}_1 \in \mathbb{R}^{k \times r}. \quad (19)$$

Then the optimization problem in the second stage can be reformulated as follows:

$$\begin{aligned} \max_{\mathbf{W}_2} \quad & \text{Tr}(\mathbf{W}_2^T\mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{A}^T\mathbf{W}_2) \\ \text{s. t.} \quad & \mathbf{W}_2^T\mathbf{A}\mathbf{A}^T\mathbf{W}_2 = \mathbf{I}_\ell. \end{aligned} \quad (20)$$

Let the compact SVD of \mathbf{A} be

$$\mathbf{A} = \mathbf{U}_A\mathbf{\Sigma}_A\mathbf{V}_A^T, \quad (21)$$

where $\mathbf{U}_A \in \mathbb{R}^{k \times t}$, $\mathbf{\Sigma}_A \in \mathbb{R}^{t \times t}$, $\mathbf{V}_A \in \mathbb{R}^{r \times t}$, and $t = \text{rank}(\mathbf{A})$.

Based on the above definitions, the solution to the two-stage approach is summarized in the following theorem.

THEOREM 1. *The top ℓ ($\ell \leq \text{rank}(\mathbf{A})$) projection vectors computed by Eq. (11) are given by*

$$\mathbf{W}_2 = (\mathbf{U}_A\mathbf{\Sigma}_A^{-1})_\ell, \quad (22)$$

where $(\mathbf{U}_A\mathbf{\Sigma}_A^{-1})_\ell$ consists of the first ℓ columns of $(\mathbf{U}_A\mathbf{\Sigma}_A^{-1})$. Thus, the projection vectors computed by the two-stage approach are

$$\mathbf{W} = \mathbf{W}_1\mathbf{W}_2 = \mathbf{U}_1\mathbf{\Sigma}_1^{-1}\mathbf{V}_{A\ell}. \quad (23)$$

When $\ell = \text{rank}(\mathbf{A})$, \mathbf{W} can be simplified as

$$\mathbf{W} = \mathbf{U}_1\mathbf{\Sigma}_1^{-1}\mathbf{V}_A. \quad (24)$$

PROOF. Using the Lagrange dual function in optimization theory, the optimization problem in Eq. (20) can be reformulated as the following eigenvalue problem:

$$(\mathbf{A}\mathbf{A}^T)^\dagger \mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{A}^T\mathbf{w}_2 = \lambda\mathbf{w}_2. \quad (25)$$

Next we derive the eigendecomposition of the matrix

$(\mathbf{A}\mathbf{A}^T)^\dagger \mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{A}^T$ as follows:

$$\begin{aligned} & (\mathbf{A}\mathbf{A}^T)^\dagger \mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{A}^T \\ &= (\mathbf{U}_A\mathbf{\Sigma}_A^2\mathbf{U}_A^T)^\dagger \mathbf{U}_A\mathbf{\Sigma}_A^2\mathbf{U}_A^T\mathbf{U}_A\mathbf{\Sigma}_A^2\mathbf{U}_A^T \\ &= \mathbf{U}_A\mathbf{\Sigma}_A^{-2}\mathbf{U}_A^T\mathbf{U}_A\mathbf{\Sigma}_A^4\mathbf{U}_A^T \\ &= \mathbf{U}_A\mathbf{\Sigma}_A^2\mathbf{U}_A \\ &= [\mathbf{U}_A, \mathbf{U}_A^\perp] \begin{bmatrix} \mathbf{\Sigma}_A^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} [\mathbf{U}_A, \mathbf{U}_A^\perp]^T, \end{aligned}$$

where $[\mathbf{U}_A, \mathbf{U}_A^\perp] \in \mathbb{R}^{k \times k}$ is orthogonal. Thus, the eigenvectors corresponding to the top ℓ eigenvalues are given by

$$\mathbf{W}_2 = \mathbf{U}_{A\ell},$$

where $\mathbf{U}_{A\ell}$ consists of the first ℓ columns of \mathbf{U}_A . To ensure that the constraint in Eq. (20) is satisfied, we normalize the columns of \mathbf{W}_2 without affecting the range space of \mathbf{W}_2^T :

$$\mathbf{W}_2 = (\mathbf{U}_A\mathbf{\Sigma}_A^{-1})_\ell.$$

Combing Eqs. (17) and (22), we have

$$\mathbf{W} = \mathbf{W}_1\mathbf{W}_2 = \mathbf{U}_1\mathbf{\Sigma}_1^{-1}\mathbf{A}^T(\mathbf{U}_A\mathbf{\Sigma}_A^{-1})_\ell = \mathbf{U}_1\mathbf{\Sigma}_1^{-1}\mathbf{V}_{A\ell}.$$

When $\ell = \text{rank}(\mathbf{A})$, we have $\mathbf{V}_{A\ell} = \mathbf{V}_A$ and $\mathbf{W} = \mathbf{U}_1\mathbf{\Sigma}_1^{-1}\mathbf{V}_A$. This completes the proof of this theorem. \square

Note that the solution to the generalized eigenvalue problem in Eq. (1) consists of the principal eigenvectors of matrix $(\mathbf{X}\mathbf{X}^T)^\dagger(\mathbf{X}\mathbf{H}\mathbf{H}^T\mathbf{X}^T)$. We follow [25] to derive the eigendecomposition of $(\mathbf{X}\mathbf{X}^T)^\dagger(\mathbf{X}\mathbf{H}\mathbf{H}^T\mathbf{X}^T)$ and show the equivalence relationship between the two-stage approach and the direct approach. The results are summarized in the following theorem:

THEOREM 2. *The eigenvectors corresponding to the top ℓ ($\ell \leq \text{rank}(\mathbf{A})$) eigenvalues of $(\mathbf{X}\mathbf{X}^T)^\dagger(\mathbf{X}\mathbf{H}\mathbf{H}^T\mathbf{X}^T)$ are*

$$\mathbf{W}_0 = \mathbf{U}_1\mathbf{\Sigma}_1^{-1}\mathbf{V}_{A\ell}, \quad (26)$$

where $\mathbf{V}_{A\ell}$ consists of the first ℓ columns of \mathbf{V}_A . Thus, the two-stage approach produces the same solution as the direct approach which solves the original generalized eigenvalue problem directly.

PROOF. Given $\mathbf{V}_A \in \mathbb{R}^{r \times t}$ with orthonormal columns, there exists $\mathbf{V}_A^s \in \mathbb{R}^{r \times (r-t)}$ such that $[\mathbf{V}_A, \mathbf{V}_A^s] \in \mathbb{R}^{r \times r}$ is an orthogonal matrix [12]. Hereafter, we denote $[\mathbf{V}_A, \mathbf{V}_A^s] = \mathbf{V}_A^s$.

We can decompose $(\mathbf{X}\mathbf{X}^T)^\dagger(\mathbf{X}\mathbf{H}\mathbf{H}^T\mathbf{X}^T)$ as follows:

$$\begin{aligned} & (\mathbf{X}\mathbf{X}^T)^\dagger(\mathbf{X}\mathbf{H}\mathbf{H}^T\mathbf{X}^T) \\ &= \mathbf{U}_1\mathbf{\Sigma}_1^{-2}\mathbf{U}_1^T\mathbf{X}\mathbf{H}\mathbf{H}^T\mathbf{X}^T \\ &= \mathbf{U}_1\mathbf{\Sigma}_1^{-2}\mathbf{U}_1^T\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T\mathbf{H}\mathbf{H}^T\mathbf{V}_1\mathbf{\Sigma}_1\mathbf{U}_1^T \\ &= \mathbf{U}_1\mathbf{\Sigma}_1^{-1}\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}_1\mathbf{U}_1^T \\ &= \mathbf{U} \begin{bmatrix} \mathbf{I}_r \\ \mathbf{0} \end{bmatrix} \mathbf{\Sigma}_1^{-1}\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}_1 \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \end{bmatrix} \mathbf{U}^T \\ &= \mathbf{U} \begin{bmatrix} \mathbf{\Sigma}_1^{-1}\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^T \\ &= \mathbf{U} \begin{bmatrix} \mathbf{\Sigma}_1^{-1}\mathbf{V}_A^s & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_A^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_A^s{}^T\mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{U}^T, \end{aligned}$$

It is clear that the eigenvectors corresponding to the top ℓ eigenvalues of $(\mathbf{X}\mathbf{X}^T)^\dagger(\mathbf{X}\mathbf{H}\mathbf{H}^T\mathbf{X}^T)$ are given by

$$\mathbf{W}_0 = \mathbf{U}_1\mathbf{\Sigma}_1^{-1}\mathbf{V}_{A\ell}.$$

The equivalence relationship follows from Eqs. (23) and (26). This completes the proof of the theorem. \square

A consequence of Theorem 1 is that solving the optimization problem in Eq. (11) in the second stage amounts to computing the SVD of the matrix \mathbf{A} . Note that $\mathbf{A} \in \mathbb{R}^{k \times r}$, where $r = \text{rank}(\mathbf{X}) \leq \min\{d, n\}$, thus the computational cost of the SVD of \mathbf{A} is quite low. In practice we can perform SVD on \mathbf{A} directly instead of solving the optimization problem in Eq. (11).

REMARK 1. *A least squares formulation is proposed in [26] for a class of dimensionality reduction techniques with the same computational cost as the proposed two-stage approach. However, the analysis in [26] assumes that the data matrix \mathbf{X} is of full rank (before centering). This tends to fail for (relatively) low-dimensional data. In particular, when the number of data points is larger than the number of dimensions, this assumption is likely to be violated. The two-stage algorithm proposed in this paper significantly improves previous work by relaxing this assumption.*

4. THE TWO-STAGE APPROACH WITH REGULARIZATION

Regularization is commonly employed to control the model complexity and avoid overfitting. In this section we present the two-stage approach with regularization. We rigorously prove the equivalence relationship between the two-stage approach and the direct approach which solves the regularized generalized eigenvalue problem in Eq. (5).

4.1 The Algorithm

To handle the regularization in the generalized eigenvalue problem in Eq. (5), we solve a penalized least squares problem, or ridge regression [14] in the first step. Note that the ‘‘latent target’’ is the same as the one used in Algorithm 1; the difference is the regularization term included in the least squares problem. After projecting the data matrix \mathbf{X} onto the subspace, we compute an auxiliary matrix $\mathbf{D} \in \mathbb{R}^{k \times k}$ and its SVD. Intuitively, the SVD computation of \mathbf{D} amounts to solving the original optimization problem by replacing \mathbf{X} with $\tilde{\mathbf{X}}$. Note that the size of \mathbf{D} is very small, thus the cost of computing the SVD of \mathbf{D} is relatively low. The algorithm outline is summarized in Algorithm 2.

4.2 Time Complexity Analysis

Similar to Algorithm 1, the least squares problem in the first stage is solved using the LSQR algorithm [20] with the same cost. In the second stage, since $k \ll d$, the most expensive part is the computation of $\tilde{\mathbf{X}}$ with a cost of $O(kdn)$ if \mathbf{X} is dense or $O(kz)$ if \mathbf{X} is sparse where z is the number of nonzero entries in the data matrix \mathbf{X} . In addition, the cost of computing \mathbf{D} and \mathbf{W}_2 is $O(nk^2)$ and $O(k^3)$, respectively. The cost of combining \mathbf{W}_1 and \mathbf{W}_2 is $O(dk\ell)$, where ℓ ($\ell \leq k$) is the number of final projection vectors. Therefore, the total computational cost is $O(Nk(3n + 5d + 2z) + kz + nk^2 + dk^2)$ if \mathbf{X} is sparse.

4.3 Equivalence

Next we show that the two-stage approach with regularization produces the same solution as the direct approach which solves the regularized generalized eigenvalue problem in Eq. (5) directly.

Algorithm 2 The Two-Stage Approach with Regularization

Input: $\mathbf{X}, \mathbf{H}, \gamma$.

Output: \mathbf{W}

Stage 1: Solve the following least squares problem:

$$\min_{\mathbf{W}_1} \|\mathbf{W}_1^T \mathbf{X} - \mathbf{H}^T\|_F^2 + \gamma \|\mathbf{W}_1\|_F^2. \quad (27)$$

Stage 2: Compute $\tilde{\mathbf{X}} = \mathbf{W}_1^T \mathbf{X}$ and $\mathbf{D} = \tilde{\mathbf{X}}\mathbf{H}$. Compute the compact SVD of $\mathbf{D} = \mathbf{U}_D \mathbf{\Sigma}_D \mathbf{U}_D^T$ and set $\mathbf{W}_2 = \mathbf{U}_D \mathbf{\Sigma}_D^{-1/2}$. Compute $\mathbf{W} = \mathbf{W}_1 \mathbf{W}_2$ as the final solution.

Following the standard techniques in linear algebra, the solution to the least squares problem with regularization in Eq. (27) is

$$\mathbf{W}_1 = (\mathbf{X}\mathbf{X}^T + \gamma\mathbf{I})^\dagger \mathbf{X}\mathbf{H} = \mathbf{U}_1 (\mathbf{\Sigma}_1^2 + \gamma\mathbf{I})^{-1} \mathbf{\Sigma}_1 \mathbf{V}_1^T \mathbf{H}. \quad (28)$$

Then $\tilde{\mathbf{X}}$ can be represented as

$$\tilde{\mathbf{X}} = \mathbf{W}_1^T \mathbf{X} = \mathbf{H}^T \mathbf{V}_1 \mathbf{\Sigma}_1 (\mathbf{\Sigma}_1^2 + \gamma\mathbf{I})^{-1} \mathbf{\Sigma}_1 \mathbf{V}_1^T. \quad (29)$$

Thus the matrix $\mathbf{D} \in \mathbb{R}^{k \times k}$ can be represented as:

$$\mathbf{D} = \tilde{\mathbf{X}}\mathbf{H} = \mathbf{H}^T \mathbf{V}_1 \mathbf{\Sigma}_1 (\mathbf{\Sigma}_1^2 + \gamma\mathbf{I})^{-1} \mathbf{\Sigma}_1 \mathbf{V}_1^T \mathbf{H} = \mathbf{B}\mathbf{B}^T, \quad (30)$$

where the matrix $\mathbf{B} \in \mathbb{R}^{k \times r}$ is defined as

$$\mathbf{B} = \mathbf{H}^T \mathbf{V}_1 \mathbf{\Sigma}_1 (\mathbf{\Sigma}_1^2 + \gamma\mathbf{I})^{-1/2}. \quad (31)$$

The solution to the two-stage approach in Algorithm 2 is summarized in the following theorem:

THEOREM 3. *Let the compact SVD of \mathbf{B} be*

$$\mathbf{B} = \mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B^T, \quad (32)$$

where $\mathbf{U}_B \in \mathbb{R}^{k \times q}$, $\mathbf{\Sigma}_B \in \mathbb{R}^{q \times q}$, $\mathbf{V}_B \in \mathbb{R}^{r \times q}$, and $q = \text{rank}(\mathbf{B})$. The top ℓ ($\ell \leq \text{rank}(\mathbf{B})$) projection vectors computed by Algorithm 2 are given by

$$\mathbf{W} = \mathbf{W}_1 \mathbf{W}_2 = \mathbf{U}_1 (\mathbf{\Sigma}_1^2 + \gamma\mathbf{I})^{-1/2} \mathbf{V}_{B\ell}, \quad (33)$$

where

$$\mathbf{W}_2 = (\mathbf{U}_B \mathbf{\Sigma}_B^{-1})_\ell, \quad (34)$$

and $\mathbf{V}_{B\ell}$ consists of the first ℓ columns of \mathbf{V}_B , $(\mathbf{U}_B \mathbf{\Sigma}_B^{-1})_\ell$ consists of the first ℓ columns of $\mathbf{U}_B \mathbf{\Sigma}_B^{-1}$. When $\ell = \text{rank}(\mathbf{B})$, \mathbf{W} can be simplified as

$$\mathbf{W} = \mathbf{U}_1 (\mathbf{\Sigma}_1^2 + \gamma\mathbf{I})^{-1/2} \mathbf{V}_B. \quad (35)$$

PROOF. Note that $\mathbf{D} = \mathbf{B}\mathbf{B}^T$. The SVD of \mathbf{D} can be obtained from the SVD of \mathbf{B} as follows:

$$\mathbf{D} = \mathbf{U}_B \mathbf{\Sigma}_B^2 \mathbf{U}_B^T = \mathbf{U}_D \mathbf{\Sigma}_D \mathbf{U}_D^T. \quad (36)$$

It follows from Algorithm 2 that

$$\mathbf{W}_2 = \mathbf{U}_B \mathbf{\Sigma}_B^{-1}.$$

Recall that \mathbf{W}_1 can also be represented using \mathbf{B} as:

$$\mathbf{W}_1 = \mathbf{U}_1 (\mathbf{\Sigma}_1^2 + \gamma\mathbf{I})^{-1} \mathbf{\Sigma}_1 \mathbf{V}_1^T \mathbf{H} = \mathbf{U}_1 (\mathbf{\Sigma}_1^2 + \gamma\mathbf{I})^{-1/2} \mathbf{B}^T.$$

We can thus derive \mathbf{W} as follows:

$$\begin{aligned}\mathbf{W} &= \mathbf{W}_1 \mathbf{W}_2 \\ &= \mathbf{U}_1 (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} \mathbf{B}^T \mathbf{U}_B \boldsymbol{\Sigma}_B^{-1} \\ &= \mathbf{U}_1 (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} \mathbf{V}_B \boldsymbol{\Sigma}_B \mathbf{U}_B^T \mathbf{U}_B \boldsymbol{\Sigma}_B^{-1} \\ &= \mathbf{U}_1 (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} \mathbf{V}_B.\end{aligned}$$

If only the first ℓ projection vectors are required, then the resulting \mathbf{W} is given by

$$\mathbf{W} = \mathbf{U}_1 (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} \mathbf{V}_{B\ell}.$$

This completes the proof. \square

We follow [27] for the eigendecomposition of the matrix $(\mathbf{X}\mathbf{X}^T + \gamma \mathbf{I})^{-1} (\mathbf{X}\mathbf{H}\mathbf{H}^T \mathbf{X}^T)$. The eigendecomposition is summarized in the following theorem, based on which we also obtain the equivalence relationship between the two-stage approach and the direct approach.

THEOREM 4. *The eigenvectors corresponding to the top ℓ eigenvalues of matrix $(\mathbf{X}\mathbf{X}^T + \gamma \mathbf{I})^{-1} (\mathbf{X}\mathbf{H}\mathbf{H}^T \mathbf{X}^T)$ are given by*

$$\mathbf{W}_0 = \mathbf{U}_1 (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} \mathbf{V}_{B\ell}, \quad (37)$$

where $\mathbf{V}_{B\ell}$ consists of the first ℓ ($\ell \leq \text{rank}(\mathbf{B})$) columns of \mathbf{V}_B . Thus, the two-stage approach in Algorithm 2 is equivalent to the direct approach which solves the generalized eigenvalue problem with regularization directly.

PROOF. Given $\mathbf{V}_B \in \mathbb{R}^{r \times q}$ with orthonormal columns, there exists $\mathbf{V}_B^\perp \in \mathbb{R}^{r \times (r-q)}$ such that $[\mathbf{V}_B, \mathbf{V}_B^\perp] \in \mathbb{R}^{r \times r}$ is an orthogonal matrix. Hereafter, we denote $[\mathbf{V}_B, \mathbf{V}_B^\perp] = \mathbf{V}_B^s$.

We can diagonalize $(\mathbf{X}\mathbf{X}^T + \gamma \mathbf{I})^{-1} (\mathbf{X}\mathbf{H}\mathbf{H}^T \mathbf{X}^T)$ as follows:

$$\begin{aligned}& (\mathbf{X}\mathbf{X}^T + \gamma \mathbf{I})^{-1} (\mathbf{X}\mathbf{H}\mathbf{H}^T \mathbf{X}^T) \\ &= \mathbf{U}_1 (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1} \boldsymbol{\Sigma}_1 \mathbf{V}_1^T \mathbf{H}\mathbf{H}^T \mathbf{V}_1 \boldsymbol{\Sigma}_1 \mathbf{U}_1^T \\ &= \mathbf{U}_1 (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} \boldsymbol{\Sigma}_1 \mathbf{V}_1^T \mathbf{H}\mathbf{H}^T \\ & \quad \mathbf{V}_1 \boldsymbol{\Sigma}_1 (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{1/2} \mathbf{U}_1^T \\ &= \mathbf{U}_1 (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} \mathbf{B}^T \mathbf{B} (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{1/2} \mathbf{U}_1^T \\ &= \mathbf{U} \begin{bmatrix} \mathbf{I}_r & \\ \mathbf{0} & \end{bmatrix} (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} \mathbf{B}^T \mathbf{B} (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{1/2} \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \end{bmatrix} \mathbf{U}^T \\ &= \mathbf{U} \begin{bmatrix} (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} \mathbf{B}^T \mathbf{B} (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^T \\ &= \mathbf{U} \begin{bmatrix} (\boldsymbol{\Sigma}_1^2 + \lambda \mathbf{I})^{-1/2} \mathbf{V}_B^s & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_B^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \\ & \quad \begin{bmatrix} \mathbf{V}_B^{sT} (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{U}^T.\end{aligned}$$

Thus, the eigenvectors corresponding to the top ℓ eigenvalues of $(\mathbf{X}\mathbf{X}^T + \gamma \mathbf{I})^{-1} (\mathbf{X}\mathbf{H}\mathbf{H}^T \mathbf{X}^T)$ are given by $\mathbf{U}_1 (\boldsymbol{\Sigma}_1^2 + \gamma \mathbf{I})^{-1/2} \mathbf{V}_{B\ell}$. The equivalence between the two-stage approach and the direct approach follows from Eqs. (33) and (37). This completes the proof of the theorem. \square

REMARK 2. *The least squares algorithm proposed in [26] only works for dimensionality reduction techniques without regularization. This drawback limits its applicability in practice since the regularized algorithms are expected to be more effective in practice due to its better generalization performance. The two-stage algorithm proposed in this paper significantly improves previous work by extending the equivalence to the regularization setting.*

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

Data Set	Type	n	d	k
Syn1	Multi-class	1000	100	5
Syn2	Multi-class	1000	5000	5
Syn3	Multi-label	1000	100	5
Syn4	Multi-label	1000	5000	5
Ionosphere	Multi-class	351	34	2
Optical digits	Multi-class	5620	64	10
Satimage	Multi-class	6435	36	6
USPS	Multi-class	9298	256	10
Wine	Multi-class	178	13	3
Scene	Multi-label	2407	294	6
Yeast	Multi-label	2417	103	14
news20	Multi-class	15935	62061	20
rcv1v2	Multi-label	3000	47236	101

REMARK 3. *The two-stage approach can be extended to the kernel-induced feature space. The equivalence relationship still holds for all dimensionality reduction techniques discussed above with and without regularization. Due to the space constraint, we skip the detailed proof in this paper.*

5. EXPERIMENTS

We have performed extensive experiments to verify the established equivalence relationship and demonstrate the scalability of the proposed two-stage approach. All the experiments are performed on a PC with Intel Core 2 Duo T9500 2.6G CPU and 4GB RAM. We implement all algorithms in Matlab. All Matlab codes and synthetic data sets are available at www.public.asu.edu/~lsun27/Code/TwoStage.html.

5.1 Experiment Setup

The dimensionality reduction techniques discussed in this paper can be divided into two categories: 1) LDA for multi-class classification; 2) HSL, CCA, and OPLS for multi-label classification. We utilize both multi-class and multi-label data sets, including synthetic and real-world data sets, in the experiments. Two synthetic data sets for multi-class classification as well as two synthetic data sets for multi-label classification are generated. In the synthetic data sets, each entry of the data matrix \mathbf{X} is generated independently from the standard Gaussian distribution $\mathcal{N}(0, 1)$. The number of classes is $k = 5$, and the labels are generated uniformly with random. Five real-world data sets from the UCI machine learning repository [3] and two benchmark data sets in multi-label classification [7, 10] are used in our experiments. To investigate the scalability of the two-stage approach, two large-scale data sets news20 [17] and rcv1v2 [18] are used. The statistics of all data sets are summarized in Table 1.

To distinguish different techniques tested in the experiments, we name the regularized techniques using a prefix “r” before the corresponding technique, e.g., “rLDA”. The two-stage versions are named using a prefix “2S” (“2S” means two stage) such as “2SLDA” and “2SrLDA”, which are the two-stage versions of LDA and regularized LDA, respectively.

5.2 Performance Comparison

In this experiment, we compare the performance of different approaches for all techniques using both synthetic and

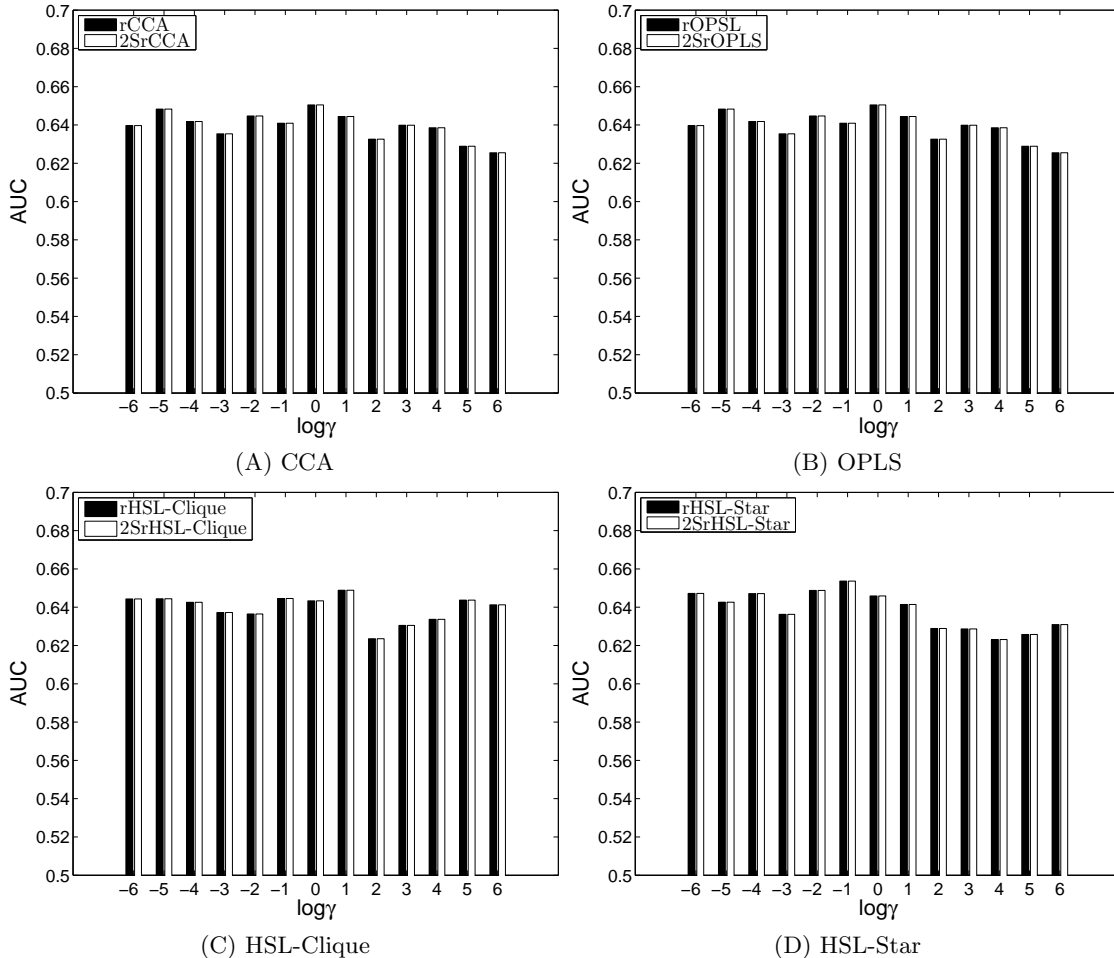


Figure 1: Comparison of different approaches in terms of average AUC for different techniques on the Yeast data set as the regularization parameter γ varies.

real-world data sets. Denote \mathbf{W}_0 as the solution of the generalized eigenvalue problem in Eq. (2) or (5) by solving it directly, and \mathbf{W} as the solution of Eq. (2) or (5) by solving it using the two-stage approach.

To verify whether both approaches produce equivalent projections, we compute $\|\mathbf{W}_0\mathbf{W}_0^T - \mathbf{W}\mathbf{W}^T\|_2$ under different values of the regularization parameter γ . We vary the value of γ from 0 to $1e6$. It follows from [27] that $\|\mathbf{W}_0\mathbf{W}_0^T - \mathbf{W}\mathbf{W}^T\|_2 = 0$ if and only if $\mathbf{W}_0 = \mathbf{W}\mathbf{R}$, where \mathbf{R} is an orthogonal matrix. Thus, both \mathbf{W} and \mathbf{W}_0 project the original data onto the same low-dimensional space. Note that a direct comparison between \mathbf{W} and \mathbf{W}_0 is possible only when the generalized eigenvalue problem in Eq. (2) or (5) admits a unique solution. This is not always the case, e.g., when two eigenvalues coincide.

The values of $\|\mathbf{W}_0\mathbf{W}_0^T - \mathbf{W}\mathbf{W}^T\|_2$ for all data sets are summarized in Table 2. Note that two different variants of HSL, HSL-Clique and HSL-Star, which compute the Laplacian using different expansion schemes, are tested. From Table 2 it can be observed that for all values of the regularization parameter γ , $\|\mathbf{W}_0\mathbf{W}_0^T - \mathbf{W}\mathbf{W}^T\|_2$ is always very small, which confirms the equivalence relationship between \mathbf{W} and \mathbf{W}_0 for projection.

Next, we investigate the classification performance of different techniques. We compare the performance of different approaches on the multi-label data set Yeast [10]. The data set is randomly partitioned into a training set and a test data set with equal size. After the projection matrix is learned from the training set, the test data set is projected onto the low-dimensional space. In our experiments, the linear support vector machines (SVM) is applied for classification. The average Area Under ROC Curve (AUC) over all labels are summarized in Figure 1 for all techniques. The regularization parameter γ varies from $1e-6$ to $1e6$. From Figure 1 we conclude that the proposed two-stage approach always produces the same classification results as the direct approach in all cases.

A similar experiment is performed on the multi-class data set wine [3] for LDA. The classification accuracies of LDA under different values of γ are summarized in Figure 2, and similar observations can be made.

5.3 Scalability Comparison

In this experiment, we study the scalability of the two-stage approach in comparison with the direct approach. Since regularization is commonly used in practice, we compare

Table 2: The value of $\|\mathbf{W}\mathbf{W}^T - \mathbf{W}_0\mathbf{W}_0^T\|_2$ under different values of the regularization parameter γ on the synthetic and real-world data sets. \mathbf{W}_0 is the solution of the generalized eigenvalue problem in Eq. (5) by solving it directly, and \mathbf{W} is the solution of Eq. (5) by solving it using the two-stage approach. Each row corresponds to a specific technique and each column corresponds to a specific value of the regularization parameter γ .

Data	Technique	0	1.0e-006	1.0e-004	1.0e-002	1.0e+000	1.0e+002	1.0e+004	1.0e+006
Syn1	LDA	2.9e-018	3.6e-018	3.4e-018	3.1e-018	2.6e-018	2.5e-018	3.1e-019	3.0e-021
Syn2	LDA	5.8e-019	1.4e-018	1.2e-018	8.9e-019	1.2e-018	9.9e-019	2.3e-019	2.9e-021
Syn3	CCA	4.9e-018	8.4e-018	7.0e-018	6.5e-018	9.5e-018	6.0e-018	5.1e-019	7.2e-021
	OPLS	4.6e-018	5.0e-018	8.7e-018	5.0e-018	6.6e-018	6.1e-018	5.4e-019	5.0e-021
	HSL-Clique	1.0e-017	1.8e-017	1.2e-017	1.2e-017	1.5e-017	1.4e-017	2.9e-018	2.5e-020
Syn4	HSL-Star	1.4e-017	2.4e-017	9.3e-018	2.6e-017	2.1e-017	5.0e-017	9.8e-019	1.3e-020
	CCA	1.3e-018	5.2e-018	3.2e-018	1.8e-018	1.3e-018	1.8e-018	4.2e-019	5.9e-021
	OPLS	1.0e-018	1.1e-018	1.3e-018	1.5e-018	1.3e-018	1.3e-018	2.9e-019	5.9e-021
	HSL-Clique	2.7e-018	2.9e-018	2.7e-018	5.0e-018	3.2e-018	2.7e-018	8.9e-019	1.4e-020
Scene	HSL-Star	2.5e-018	3.7e-018	2.9e-018	5.7e-018	4.1e-018	2.9e-018	1.1e-018	3.1e-020
	CCA	2.4e-015	2.1e-015	6.1e-015	3.7e-015	1.2e-015	1.8e-016	6.0e-018	9.0e-020
	OPLS	2.0e-015	3.4e-015	3.8e-015	2.5e-015	1.1e-015	2.3e-016	1.1e-017	1.4e-019
	HSL-Clique	4.5e-015	9.1e-015	2.6e-014	1.2e-014	3.6e-015	1.3e-015	5.9e-017	1.0e-018
Yeast	HSL-Star	4.6e-015	3.3e-014	2.1e-014	7.7e-015	1.1e-014	2.5e-016	1.0e-016	6.5e-019
	CCA	1.6e-012	1.5e-011	1.2e-012	1.4e-015	6.9e-016	5.9e-017	1.7e-018	1.4e-020
	OPLS	4.1e-012	1.6e-011	3.7e-012	1.2e-014	1.5e-015	3.7e-016	3.2e-018	2.9e-020
	HSL-Clique	1.5e-012	1.4e-011	3.7e-012	3.9e-015	1.6e-015	2.7e-016	5.1e-018	2.5e-020
Wine	HSL-Star	2.1e-012	1.0e-011	2.4e-012	1.1e-014	9.4e-015	1.1e-015	1.5e-017	4.4e-019
	LDA	5.9e-017	2.1e-016	2.3e-016	2.1e-016	3.2e-017	2.2e-018	1.3e-020	2.0e-020
Satimage	LDA	4.6e-016	2.2e-015	8.4e-016	7.3e-016	7.7e-016	8.1e-017	3.9e-017	6.2e-019
Ionosphere	LDA	8.5e-018	1.0e-017	4.3e-018	2.1e-017	6.8e-018	6.6e-018	6.6e-020	1.1e-021
Optical digits	LDA	6.2e-018	7.2e-018	6.7e-018	5.7e-018	1.9e-018	1.5e-019	5.9e-020	5.6e-021
USPS	LDA	7.0e-015	3.0e-014	2.6e-014	6.6e-015	1.1e-016	3.0e-018	4.1e-019	6.6e-021

the scalability of different algorithms with regularization. In terms of the implementation, the least squares problem in the first stage of the two-stage approach is solved using the LSQR algorithm [19, 20]. For the direct approach which solves the generalized eigenvalue problem directly, the Lanczos algorithm [12] is applied. It is well-known that solving large-scale generalized eigenvalue problems is much more difficult than the regular eigenvalue problems [22, 28]. In order to transform the generalized eigenvalue problem into the regular eigenvalue problem, we can factor $\mathbf{X}\mathbf{X}^T$ or apply the standard Lanczos algorithm for the matrix $(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{S}\mathbf{X}^T$ using the $\mathbf{X}\mathbf{X}^T$ inner product [22]. Due to the issue of singularity for high-dimensional data set with small regularization for the second method, in this paper we follow the procedure in [26], which factors $\mathbf{X}\mathbf{X}^T$ and solves a symmetric eigenvalue problem using the Lanczos algorithm.

The computation time (in log scale) of different techniques on the large-scale data set rcv1v2 is shown in Figures 4 and 5. In Figure 4, we increase the sample size from 500 to 3000 with a step size 500, and the dimensionality is fixed at 5000. The computation time of both approaches increases as the sample size increases. However, it can be observed that the

computation time of the two-stage approach is significantly less than that of the direct approach. In Figure 5 we fix the sample size at 3000 and increase the dimensionality from 500 to 5000 with a step size 500. Similar observations can be made from Figure 5.

We perform a similar experiment on the news20 data set for LDA (multi-class classification). The experimental results are summarized in Figure 3. In Figure 3 (left), we fix the dimensionality at 5000 and increase the sample size from 500 to 3000 with a step size 500. In Figure 3 (right), the dimensionality is increased from 500 to 3000 with a step size 500 while the sample size is fixed at 5000. It can be observed from these figures that the two-stage approach is much more efficient than the direct approach.

6. CONCLUSIONS

In this paper we propose an efficient two-stage approach for a class of dimensionality reduction techniques, including Canonical Correlation Analysis, Orthonormalized Partial Least Squares, Hypergraph Spectral Learning and Linear Discriminant Analysis. We rigorously prove the equivalence relationship between the two-stage approach and the

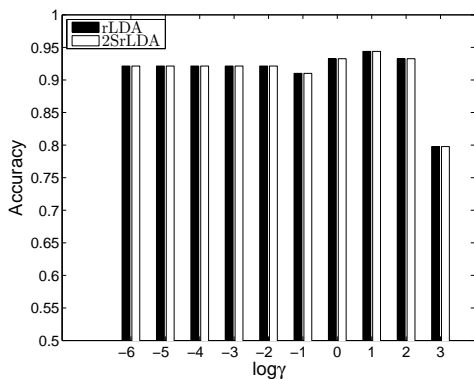


Figure 2: Comparison of different approaches in terms of classification accuracy for LDA on the wine data set as the regularization parameter γ varies.

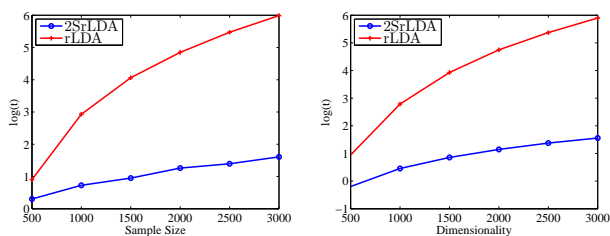


Figure 3: Scalability comparison on the news20 data set as the sample size (left) and dimensionality (right) increase. The horizontal axis is the sample size (left) or the dimensionality (right), and the vertical axis is $\log(t)$, where t is the computation time (in seconds).

direct approach which solves the generalized eigenvalue problem directly. Compared with previous work, one appealing feature of the two-stage approach is that no assumption is required for the equivalence relationship. In addition, the two-stage approach can be further extended to the regularization setting. We show that the proposed two-stage approach scales linearly in terms of both the sample size and data dimensionality. We have performed extensive experiments on both synthetic and real-world data sets. Our experimental results confirm the established equivalence relationship. Results also demonstrate the scalability of the proposed two-stage approach.

The current two-stage approach assumes that the complete data set for training is given in advance, and learning is carried out in one batch. However, in many real applications the data come as a stream. We plan to explore the online algorithm for the class of dimensionality reduction techniques.

Acknowledgements

This work was supported by NSF IIS-0612069, IIS-0812551, IIS-0953662, NGA HM1582-08-1-0016, the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the US Army.

7. REFERENCES

- [1] S. Agarwal, K. Branson, and S. Belongie. Higher order learning with graphs. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 17–24, 2006.
- [2] J. Arenas-Garcia and G. Camps-Valls. Efficient kernel orthonormalized PLS for remote sensing applications. *IEEE Transactions on Geoscience and Remote Sensing*, 46(10):2872–2881, 2008.
- [3] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007.
- [4] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2003.
- [5] R. E. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ, 1961.
- [6] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, 2006.
- [7] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.
- [8] D. Cai. *Spectral Regression: A Regression Framework for Efficient Regularized Subspace Learning*. PhD thesis, Department of Computer Science, University of Illinois at Urbana-Champaign, 2009.
- [9] D. Cai, X. He, and J. Han. SRDA: An efficient algorithm for large-scale discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering*, 20(1):1–12, 2008.
- [10] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems 13 (NIPS)*, pages 681–687, 2001.
- [11] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, NY, 1990.
- [12] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Press, Baltimore, MD, 3rd edition, 1996.
- [13] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.
- [14] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2nd edition, 2009.
- [15] H. Hotelling. Relations between two sets of variables. *Biometrika*, 28:312–377, 1936.
- [16] P. Howland and H. Park. Two-stage methods for linear discriminant analysis: Equivalent results at a lower cost. Technical report, Georgia Institute of Technology, 2009.
- [17] K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning (ICML)*, pages 331–339, 1995.
- [18] D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

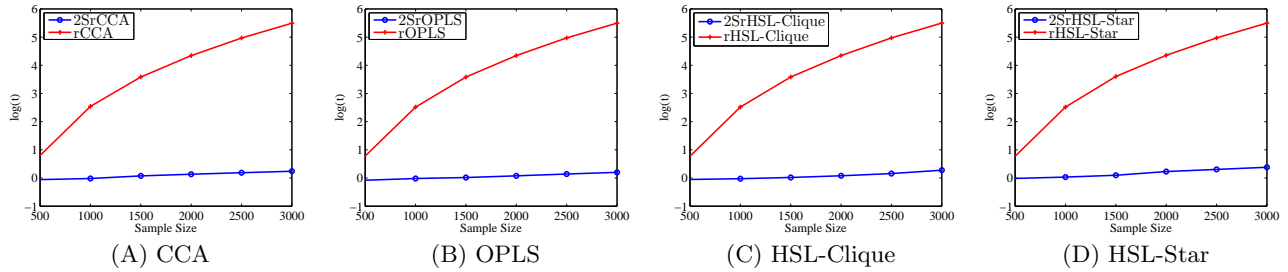


Figure 4: Scalability comparison on the rcv1v2 data set as the sample size increases. The horizontal axis is the sample size, and the vertical axis is $\log(t)$, where t is the computation time (in seconds).

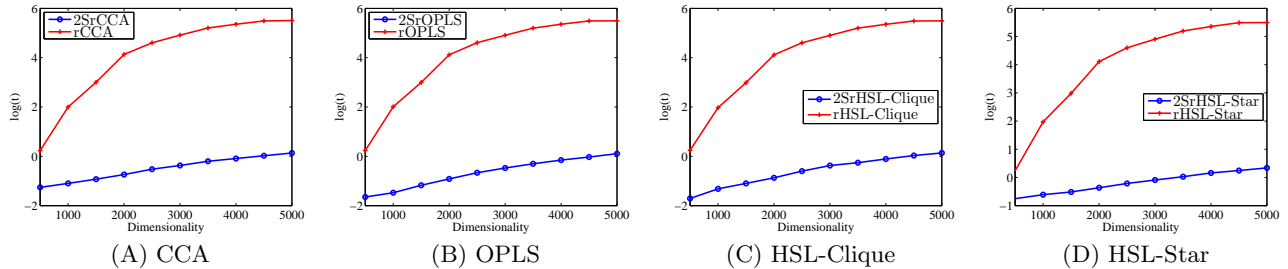


Figure 5: Scalability comparison on the rcv1v2 data set as the dimensionality increases. The horizontal axis is the dimensionality, and the vertical axis is $\log(t)$, where t is the computation time (in seconds).

- [19] C. C. Paige and M. A. Saunders. Algorithm 583: LSQR: Sparse linear equations and least squares problems. *ACM Transactions on Mathematical Software*, 8(2):195–209, 1982.
- [20] C. C. Paige and M. A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, 8(1):43–71, 1982.
- [21] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection Techniques, Lecture Notes in Computer Science*, pages 34–51, 2006.
- [22] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Halsted Press, New York, NY, 1992.
- [23] B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, MA, 2002.
- [24] L. Sun, S. Ji, and J. Ye. Hypergraph spectral learning for multi-label classification. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 668–676, 2008.
- [25] L. Sun, S. Ji, and J. Ye. A least squares formulation for canonical correlation analysis. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 1024–1031, 2008.
- [26] L. Sun, S. Ji, and J. Ye. A least squares formulation for a class of generalized eigenvalue problems in machine learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 977–984, 2009.
- [27] L. Sun, S. Ji, S. Yu, and J. Ye. On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1230–1235, 2009.
- [28] D.S. Watkins. *Fundamentals of matrix computations*. John Wiley & Sons, Inc., New York, NY, 1991.
- [29] J. Wold and *et al.* *Chemometrics, mathematics and statistics in chemistry*. Reidel Publishing Company, Dordrecht, Holland, 1984.
- [30] K. Worsley, J.-B. Poline, K. J. Friston, and A.C. Evans. Characterizing the response of PET and fMRI data using multivariate linear models. *Neuroimage*, 6(4):305–319, 1997.
- [31] J. Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6:483–502, 2005.
- [32] J. Ye and Q. Li. A two-stage linear discriminant analysis via QR-decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:929–941, 2005.
- [33] Z. Zhang, G. Dai, and M. Jordan. A flexible and efficient algorithm for regularized fisher discriminant analysis. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 632–647. Springer, 2009.
- [34] D. Zhou, J. Huang, and B. Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems 18 (NIPS)*, pages 1601–1608, 2006.