Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition *

Peg Howland† and Haesun Park‡

April 7, 2003

Abstract

Discriminant analysis has been used for decades to extract features that preserve class separability. It is commonly defined as an optimization problem involving covariance matrices that represent the scatter within and between clusters. The requirement that one of these matrices be nonsingular limits its application to data sets with certain relative dimensions. We examine a number of optimization criteria, and extend their applicability by using the generalized singular value decomposition to circumvent the nonsingularity requirement. The result is a generalization of discriminant analysis that can be applied even when the sample size is smaller than the dimension of the sample data. We use classification results from the reduced representation to compare the effectiveness of this approach with some alternatives, and conclude with a discussion of their relative merits.

1 Introduction

The goal of discriminant analysis is to combine features of the original data in a way that most effectively discriminates between classes. With an appropriate extension, it can be applied to our goal of reducing the dimension of a data matrix in a way that most effectively preserves its cluster structure. That is, we want to find a linear transformation G^T that maps an m-dimensional data point a to a vector y in the l-dimensional space:

$$G^T : a \in \mathbb{R}^{m\times 1} \to y \in \mathbb{R}^{l\times 1}.$$

*This work was supported in part by the National Science Foundation grants CCR-9901992 and CCR-0204109. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

†Dept. of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, MN 55455 (`howland@cs.umn.edu`). The work of this author was supported in part by the Guidant Fellowship.

‡Dept. of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, MN 55455 (`hpark@cs.umn.edu`), and Korea Institute for Advanced Study, 207-43 Cheongryangri-dong, Dongdaemun-gu, Seoul 130-012, KOREA (from Sept. 2001 to Aug. 2002).

Assuming that the given data are already clustered, we seek a transformation that optimally preserves this cluster structure in the reduced dimensional space.

For this purpose, first we need to formulate a measure of cluster quality. When cluster quality is high, each cluster is tightly grouped, but well separated from the other clusters. To quantify this, scatter matrices [Fuk90, TK99] are defined in discriminant analysis. For simplicity of discussion, we will assume that data vectors $a_1, \ldots, a_n$ form columns of a matrix $A \in \mathbb{R}^{m \times n}$, and are grouped into k clusters as

$$A = [A_1 \quad A_2 \quad \cdots \quad A_k] \quad \text{where} \quad A_i \in \mathbb{R}^{m \times n_i}, \quad \text{and} \quad \sum_{i=1}^{k} n_i = n. \tag{1}$$

Let N_i denote the set of column indices that belong to cluster i. The centroid $c^{(i)}$ is computed by taking the average of the columns in cluster i; i.e.,

$$c^{(i)} = \frac{1}{n_i} \sum_{j \in N_i} a_j$$

and the global centroid c is defined as

$$c = \frac{1}{n} \sum_{j=1}^{n} a_j.$$

Then the within-cluster, between-cluster, and mixture scatter matrices are defined as

$$\begin{aligned} S_W &= \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})(a_j - c^{(i)})^T, \\ S_B &= \sum_{i=1}^{k} \sum_{j \in N_i} (c^{(i)} - c)(c^{(i)} - c)^T = \sum_{i=1}^{k} n_i (c^{(i)} - c)(c^{(i)} - c)^T, \text{ and} \\ S_M &= \sum_{i=1}^{n} (a_i - c)(a_i - c)^T, \end{aligned}$$

respectively. It is easy to show [JD88] that the scatter matrices have the relationship

$$S_M = S_W + S_B. \tag{2}$$

Applying G^T to the matrix A transforms the scatter matrices to

$$S_W^Y = G^T S_W G, \quad S_B^Y = G^T S_B G, \quad \text{and} \quad S_M^Y = G^T S_M G,$$

where the superscript Y denotes values in the l-dimensional space.

There are several measures of cluster quality that involve the three scatter matrices [Fuk90, TK99]. Since

$$\text{trace}(S_W) = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})^T (a_j - c^{(i)}) = \sum_{i=1}^{k} \sum_{j \in N_i} \| a_j - c^{(i)} \|_2^2$$

measures the closeness of the columns within the clusters, and

$$\mathrm{trace}(S_B) = \sum_{i=1}^{k} \sum_{j \in N_i} (c^{(i)} - c)^T (c^{(i)} - c) = \sum_{i=1}^{k} \sum_{j \in N_i} \|c^{(i)} - c\|_2^2$$

measures the separation between clusters, an optimal transformation that preserves the given cluster structure would maximize trace(S_B^Y) and minimize trace(S_W^Y).

This simultaneous optimization can be approximated by finding a transformation G that maximizes $\mathrm{trace}((S_W^Y)^{-1} S_B^Y)$. However, this criterion cannot be applied when the matrix S_W is singular, a situation that occurs frequently in many applications. For example, in handling document data in information retrieval, it is often the case that the number of terms in the document collection is larger than the total number of documents (i.e., $m > n$ in the term-document matrix A), and therefore the matrix S_W is singular. Furthermore, in applications where the data points are in a very high dimensional space and collecting data is expensive, S_W is singular because the value for n must be kept relatively small.

One way to make classical discriminant analysis applicable to the data matrix $A \in \mathbb{R}^{m \times n}$ with $m > n$ (and hence S_W singular) is to perform dimension reduction in two stages. The discriminant analysis stage is preceded by a stage in which the cluster structure is ignored. The most popular method for the first part of this process is rank reduction by the singular value decomposition (SVD), the main tool in latent semantic indexing (LSI) [DDF$^+$90, BDO95]. In fact, this idea has recently been implemented by Torkkola [Tor01]. However, the overall performance of this two-stage approach will be sensitive to the reduced dimension in its first stage. LSI has no theoretical optimal reduced dimension, and its computational estimation is difficult without the potentially expensive process of trying many test cases. We discuss this alternative approach in greater detail in Section 4.2.

In this paper, we extend discriminant analysis in a way that provides the optimal reduced dimension theoretically, without introducing another stage as described above. We consider the set of criteria involving

$$\mathrm{trace}((S_2^Y)^{-1} S_1^Y) \quad \text{and} \quad \ln(\det((S_2^Y)^{-1} S_1^Y)), \tag{3}$$

where S_1 and S_2 are chosen from S_W, S_B, and S_M. Classical discriminant analysis expresses their solution in terms of a generalized eigenvalue problem when S_2 is nonsingular. By reformulating the problem in terms of the generalized singular value decomposition (GSVD) [VL76, PS81, GVL96], we extend the applicability to the case when S_2 is singular. We also establish the equivalence among alternative choices for S_1 and S_2. In addition to the two-stage approach described above, we present a second alternative approach that optimizes the trace of an individual scatter matrix, and show how this can be achieved efficiently. Finally, we present experimental results demonstrating the capabilities of the GSVD approach, and comparing its effectiveness to the alternatives.

2 Generalized Singular Value Decomposition

The following theorem introduces the GSVD as was originally defined by Van Loan [VL76].

THEOREM 1 *Suppose two matrices $K_A \in \mathbb{R}^{m\times n}$ with $m \geq n$ and $K_B \in \mathbb{R}^{p\times n}$ are given. Then there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{p\times p}$ and a nonsingular matrix $X \in \mathbb{R}^{n\times n}$ such that*

$$U^T K_A X = diag(\alpha_1, ..., \alpha_n) \quad and \quad V^T K_B X = diag(\beta_1, ..., \beta_q),$$

where $q = \min(p, n)$, $\alpha_i \geq 0$ for $1 \leq i \leq n$, and $\beta_i \geq 0$ for $1 \leq i \leq q$.

This formulation cannot be applied to the matrix pair K_A and K_B when the dimensions of K_A do not satisfy the assumed restrictions. Paige and Saunders [PS81] developed a more general formulation which can be defined for any two matrices with the same number of columns. We restate theirs as follows.

THEOREM 2 *Suppose two matrices $K_A \in \mathbb{R}^{n\times m}$ and $K_B \in \mathbb{R}^{p\times m}$ are given. Then for*

$$K = \begin{pmatrix} K_A \\ K_B \end{pmatrix} \quad and \quad t = rank(K),$$

there exist orthogonal matrices $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{p\times p}$, $W \in \mathbb{R}^{t\times t}$, and $Q \in \mathbb{R}^{m\times m}$ such that

$$U^T K_A Q = \Sigma_A(\underbrace{W^T R}_{t}, \underbrace{0}_{m-t}) \quad and \quad V^T K_B Q = \Sigma_B(\underbrace{W^T R}_{t}, \underbrace{0}_{m-t}),$$

where

$$\underset{n\times t}{\Sigma_A} = \begin{pmatrix} I_A & & \\ & D_A & \\ & & O_A \end{pmatrix}, \quad \underset{p\times t}{\Sigma_B} = \begin{pmatrix} O_B & & \\ & D_B & \\ & & I_B \end{pmatrix},$$

and $R \in \mathbb{R}^{t\times t}$ is nonsingular with its singular values equal to the nonzero singular values of K. The matrices

$$I_A \in \mathbb{R}^{r\times r} \quad and \quad I_B \in \mathbb{R}^{(t-r-s)\times(t-r-s)}$$

are identity matrices, where

$$r = rank\begin{pmatrix} K_A \\ K_B \end{pmatrix} - rank(K_B) \quad and \quad s = rank(K_A) + rank(K_B) - rank\begin{pmatrix} K_A \\ K_B \end{pmatrix},$$

$$O_A \in \mathbb{R}^{(n-r-s)\times(t-r-s)} \quad and \quad O_B \in \mathbb{R}^{(p-t+r)\times r}$$

are zero matrices with possibly no rows or no columns, and

$$D_A = diag(\alpha_{r+1}, \ldots, \alpha_{r+s}) \quad and \quad D_B = diag(\beta_{r+1}, \ldots, \beta_{r+s})$$

satisfy

$$1 > \alpha_{r+1} \geq \cdots \geq \alpha_{r+s} > 0, \quad 0 < \beta_{r+1} \leq \cdots \leq \beta_{r+s} < 1, \tag{4}$$

and $\alpha_i^2 + \beta_i^2 = 1$ for $i = r+1, \ldots, r+s$.

This form of GSVD is related to that of Van Loan by writing [PS81]

$$U^T K_A X = (\Sigma_A, 0) \quad \text{and} \quad V^T K_B X = (\Sigma_B, 0), \tag{5}$$

where

$$\underset{m\times m}{X} = Q \begin{pmatrix} R^{-1}W & 0 \\ 0 & I \end{pmatrix}.$$

From the form in Eqn. (5) we see that

$$K_A = U(\Sigma_A, 0)X^{-1} \quad \text{and} \quad K_B = V(\Sigma_B, 0)X^{-1},$$

which imply that

$$K_A^T K_A = X^{-T} \begin{pmatrix} \Sigma_A^T \Sigma_A & 0 \\ 0 & 0 \end{pmatrix} X^{-1} \quad \text{and} \quad K_B^T K_B = X^{-T} \begin{pmatrix} \Sigma_B^T \Sigma_B & 0 \\ 0 & 0 \end{pmatrix} X^{-1}.$$

Defining

$$\alpha_i = 1,\ \beta_i = 0 \text{ for } i = 1, \dots, r$$

and

$$\alpha_i = 0,\ \beta_i = 1 \text{ for } i = r + s + 1, \dots, t,$$

we have, for $1 \le i \le t$,

$$\beta_i^2 K_A^T K_A x_i = \alpha_i^2 K_B^T K_B x_i, \tag{6}$$

where x_i represents the ith column of X. For the remaining $m - t$ columns of X, both $K_A^T K_A x_i$ and $K_B^T K_B x_i$ are zero, so Eqn. (6) is satisfied for arbitrary values of α_i and β_i when $t + 1 \le i \le m$. The columns of X are the generalized right singular vectors for the matrix pair (K_A, K_B). In terms of the generalized singular values, or the α_i/β_i quotients, r of them are infinite, s are finite and nonzero, and $t - r - s$ are zero.

3 Generalization of Linear Discriminant Analysis

In this section, several criteria from discriminant analysis are extended utilizing the GSVD. We establish the equivalence for various choices of scatter matrices, as well as for seemingly quite different criteria involving the trace and the determinant.

3.1 Optimization of $\mathbf{J_1} = \mathrm{trace}(\mathbf{S_2^{-1}S_1})$ Criteria

For now, we will focus our discussion on the criteria of optimizing

$$J_1(G) = \mathrm{trace}((G^T S_2 G)^{-1}(G^T S_1 G)), \tag{7}$$

where S_1 and S_2 are chosen from S_W, S_B, and S_M. When S_2 is assumed to be nonsingular, it is symmetric positive definite. According to results from the symmetric-definite generalized eigenvalue problem [GVL96], there exists a nonsingular matrix $X \in \mathbb{R}^{m\times m}$ such that

$$X^T S_1 X = \Lambda = \mathrm{diag}(\lambda_1 \ldots \lambda_m) \quad \text{and} \quad X^T S_2 X = I_m. \tag{8}$$

Letting x_i denote the ith column of X, we have

$$S_1 x_i = \lambda_i S_2 x_i, \tag{9}$$

which means that λ_i and x_i are an eigenvalue-eigenvector pair of $S_2^{-1}S_1$. Since S_1 is positive semidefinite, $\lambda_i \geq 0$ for $1 \leq i \leq m$. From (8), we see that only the largest $q = \mathrm{rank}(S_1)$ λ_i's can be nonzero. In addition, by using a permutation matrix to order Λ (and likewise X), we can assume that $\lambda_1 \geq \cdots \geq \lambda_q \geq \lambda_{q+1} = \cdots = \lambda_m = 0$.

We have

$$\begin{aligned} J_1(G) &= \mathrm{trace}((G^T S_2 G)^{-1} G^T S_1 G) \\ &= \mathrm{trace}((G^T X^{-T} X^{-1} G)^{-1} G^T X^{-T} \Lambda X^{-1} G) \\ &= \mathrm{trace}((\tilde{G}^T \tilde{G})^{-1} \tilde{G}^T \Lambda \tilde{G}), \end{aligned}$$

where $\tilde{G} = X^{-1}G$. The matrix $\tilde{G}$ has full column rank provided G does, so it has the reduced QR factorization $\tilde{G} = QR$, where $Q \in \mathbb{R}^{m\times l}$ has orthonormal columns and R is nonsingular [GVL96]. Hence

$$\begin{aligned} J_1(G) &= \mathrm{trace}((R^T R)^{-1} R^T Q^T \Lambda Q R) \\ &= \mathrm{trace}(R^{-1} Q^T \Lambda Q R) \\ &= \mathrm{trace}(Q^T \Lambda Q R R^{-1}) \\ &= \mathrm{trace}(Q^T \Lambda Q). \end{aligned}$$

This shows that once we have simultaneously diagonalized S_1 and S_2, the maximization of $J_1(G)$ depends only on an orthonormal basis for $\mathrm{range}(X^{-1}G)$; i.e.,

$$\max_G J_1(G) = \max_{Q^T Q = I} \mathrm{trace}(Q^T \Lambda Q) \leq \lambda_1 + \cdots + \lambda_q = \mathrm{trace}(S_2^{-1} S_1).$$

(Here we consider only maximization. Similar arguments will hold when J_1 is *minimized* for some choices of S_1 and S_2.) For any l satisfying $l \geq q$, this upper bound on $J_1(G)$ is achieved for

$$Q = \begin{pmatrix} I_l \\ 0 \end{pmatrix} \quad \text{or} \quad G = X \begin{pmatrix} I_l \\ 0 \end{pmatrix} R.$$

Note that the transformation G is not unique. That is, J_1 satisfies the invariance property $J_1(G) = J_1(GW)$ for any nonsingular matrix $W \in \mathbb{R}^{l\times l}$, since

$$\begin{aligned} J_1(GW) &= \text{trace}((W^T G^T S_2 GW)^{-1}(W^T G^T S_1 GW)) \\ &= \text{trace}(W^{-1}(G^T S_2 G)^{-1} W^{-T} W^T (G^T S_1 G)W) \\ &= \text{trace}((G^T S_2 G)^{-1}(G^T S_1 G)WW^{-1}) \\ &= J_1(G). \end{aligned}$$

Hence, the maximum $J_1(G)$ is also achieved for $G = X \begin{pmatrix} I_l \\ 0 \end{pmatrix}$. This means that

$$\text{trace}((G^T S_2 G)^{-1} G^T S_1 G) = \text{trace}(S_2^{-1} S_1) \tag{10}$$

whenever $G \in \mathbb{R}^{m\times l}$ consists of l eigenvectors of $S_2^{-1} S_1$ corresponding to the l largest eigenvalues.

Now, a limitation of the J_1 criteria in many applications, including information retrieval, is that the matrix S_2 must be nonsingular. Recalling the partitioning of A into k clusters given in (1), we define the $m \times n$ matrices

$$H_W = [A_1 - c^{(1)} e^{(1)^T}, A_2 - c^{(2)} e^{(2)^T}, \dots, A_k - c^{(k)} e^{(k)^T}] \tag{11}$$
$$H_B = [(c^{(1)} - c)e^{(1)^T}, (c^{(2)} - c)e^{(2)^T}, \dots, (c^{(k)} - c)e^{(k)^T}] \tag{12}$$
$$H_M = [a_1 - c, \dots, a_n - c] = A - ce^T, \tag{13}$$

where $e^{(i)} = (1, \dots, 1)^T \in \mathbb{R}^{n_i \times 1}$ and $e = (1, \cdots, 1)^T \in \mathbb{R}^{n\times 1}$. Then the scatter matrices can be expressed as

$$S_W = H_W H_W^T, \quad S_B = H_B H_B^T, \quad \text{and} \quad S_M = H_M H_M^T. \tag{14}$$

For S_2 to be nonsingular, we can only allow the case $m \le n$, since S_2 is the product of an $m \times n$ matrix and an $n \times m$ matrix [Ort87]. Thus J_1 cannot be applied when the number of available data points is smaller than the dimension of the data. We seek a solution which does not impose this restriction, and which can be found without explicitly forming S_1 and S_2 from H_W, H_B, and H_M. Toward that end, we express λ_i as α_i^2/β_i^2, and the problem (9) generalizes to

$$\beta_i^2 S_1 x_i = \alpha_i^2 S_2 x_i. \tag{15}$$

This has the form of a problem that can be solved using the GSVD, as described in Section 2.

3.2 Generalization of $\mathbf{J_1} = \text{trace}(\mathbf{S_2^{-1} S_1})$ Criteria for Singular $\mathbf{S_2}$

Continuing with the J_1 criteria, we first consider the case where

$$(S_1, S_2) = (S_B, S_W).$$

From Eqn. (14) and the definition of H_B given in Eqn. (12), $\mathrm{rank}(S_B) \leq k-1$. To approximate G that satisfies both

$$\max_G \mathrm{trace}(G^T S_B G) \quad \text{and} \quad \min_G \mathrm{trace}(G^T S_W G), \tag{16}$$

we choose the x_i's which correspond to the $k-1$ largest λ_i's, where $\lambda_i = \alpha_i^2/\beta_i^2$. When the GSVD construction orders the singular value pairs as in Eqn. (4), the generalized singular values, or the α_i/β_i quotients, are in nonincreasing order. Therefore, the first $k-1$ columns of X are all we need. Our algorithm first computes the matrices H_B and H_W from the data matrix A. We then solve for a very limited portion of the GSVD of the matrix pair (H_B^T, H_W^T). This solution is accomplished by following the construction in the proof of Theorem 2 [PS81]. The major steps are limited to the complete orthogonal decomposition [GVL96, LH95] of

$$K = \begin{pmatrix} H_B^T \\ H_W^T \end{pmatrix},$$

which produces orthogonal matrices P and Q and a nonsingular matrix R, followed by the singular value decomposition of a leading principal submatrix of P, whose size is much smaller than that of the data matrix. The steps for this case are summarized in Algorithm LDA/GSVD, adapted from [HJP].

When $m > n$, the scatter matrix S_W is singular. Hence, we cannot even define the J_1 criterion, and discriminant analysis fails. Consider a generalized right singular vector x_i that lies in the null space of S_W. From Eqn. (15), we see that either x_i also lies in the null space of S_B, or the corresponding β_i equals zero. We will discuss each of these cases separately.

When

$$x_i \in \mathrm{null}(S_W) \cap \mathrm{null}(S_B),$$

Eqn. (15) is satisfied for arbitrary values of α_i and β_i. As explained in Section 2, this will be the case for the rightmost $m-t$ columns of X. To determine whether these columns should be included in G, consider

$$\mathrm{trace}(G^T S_B G) = \sum g_j^T S_B g_j \quad \text{and} \quad \mathrm{trace}(G^T S_W G) = \sum g_j^T S_W g_j,$$

where g_j represents the jth column of G. Since $x_i^T S_W x_i = 0$ and $x_i^T S_B x_i = 0$, adding the column x_i to G does not contribute to either maximization or minimization in (16). For this reason, we do not include these columns of X in our solution.

When

$$x_i \in \mathrm{null}(S_W) - \mathrm{null}(S_B),$$

then $\beta_i = 0$. As discussed in Section 2, this implies that $\alpha_i = 1$, and hence that the generalized singular value α_i/β_i is infinite. The leftmost columns of X will correspond to these. Including these columns in G increases $\mathrm{trace}(G^T S_B G)$, while leaving $\mathrm{trace}(G^T S_W G)$ unchanged. We conclude that, even when S_W is singular, the rule regarding which columns of X to include in G remain the same as for the nonsingular case. The experiments summarized in Section 5 demonstrate that Algorithm LDA/GSVD works very well even when S_W is singular, thus extending its applicability beyond that of classical discriminant analysis.

Algorithm 1 LDA/GSVD

Given a data matrix $A \in \mathbb{R}^{m \times n}$ with k clusters and an input vector $a \in \mathbb{R}^{m \times 1}$, compute the matrix $G \in \mathbb{R}^{m \times (k-1)}$ which preserves the cluster structure in the reduced dimensional space, using

$$J_1(G) = \mathrm{trace}((G^T S_W G)^{-1} G^T S_B G).$$

Also compute the $k-1$ dimensional representation y of a.

1. Compute H_B and H_W from A according to

$$H_B = [\sqrt{n_1}(c^{(1)} - c), \sqrt{n_2}(c^{(2)} - c), \ldots, \sqrt{n_k}(c^{(k)} - c)] \in \mathbb{R}^{m \times k},$$

   and (11), respectively. (Using this equivalent but lower dimensional form of H_B reduces complexity.)

2. Compute the complete orthogonal decomposition

$$P^T K Q = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{where} \quad K = \begin{pmatrix} H_B^T \\ H_W^T \end{pmatrix} \in \mathbb{R}^{(k+n) \times m}$$

3. Let $t = \mathrm{rank}(K)$.
4. Compute W from the SVD of $P(1:k, 1:t)$, which is $U^T P(1:k, 1:t) W = \Sigma_A$.
5. Compute the first $k-1$ columns of $X = Q \begin{pmatrix} R^{-1}W & 0 \\ 0 & I \end{pmatrix}$, and assign them to G.
6. $y = G^T a$

3.3 Equivalence of $\mathbf{J_1} = \mathrm{trace}(\mathbf{S_2^{-1} S_1})$ Criteria for Various $\mathbf{S_1}$ and $\mathbf{S_2}$

For the case when

$$(S_1, S_2) = (S_M, S_W),$$

if we follow the analysis in Section 3.1 literally, it appears that we would have to include $\mathrm{rank}(S_M) \not\leq k-1$ columns of X in G. However, using the relation (2), the generalized eigenvalue problem

$$S_M x_i = \lambda_i S_W x_i$$

can be rewritten as

$$S_B x_i = (\lambda_i - 1) S_W x_i, \quad \text{where} \quad \lambda_i \geq 1 \quad \text{for} \quad 1 \leq i \leq m.$$

In this case, the eigenvector matrix is the same as for the case of $(S_1, S_2) = (S_B, S_W)$, but the eigenvalue matrix is $\Lambda - I$. Since the same permutation can be used to put $\Lambda - I$ in nonincreasing order as was used for Λ, x_i corresponds to the ith largest eigenvalue of $S_W^{-1} S_B$. Therefore, when S_W is nonsingular, the solution is the same as for $(S_1, S_2) = (S_B, S_W)$.

When $m > n$, the scatter matrix S_W is singular. For a generalized right singular vector $x_i \in \text{null}(S_W)$, $S_M x_i = S_B x_i$. Hence, we include the same columns in G as we did in the case of $(S_1, S_2) = (S_B, S_W)$. Alternatively, we can show that the solutions are the same by deriving a GSVD of the matrix pair (H_M^T, H_W^T) that has the same generalized right singular vectors as (H_B^T, H_W^T). To do this, we establish the following two properties of H_B and H_W.

PROPERTY 1 $H_W H_B^T = 0$.

Proof. From

$$H_M = H_W + H_B,$$

we have

$$\begin{aligned} S_M &= H_M H_M^T = (H_W + H_B)(H_W + H_B)^T \\ &= H_W H_W^T + H_B H_B^T + H_B H_W^T + H_W H_B^T \\ &= S_M + H_B H_W^T + H_W H_B^T, \end{aligned}$$

which implies

$$H_B H_W^T + H_W H_B^T = 0.$$

In fact, each of these products is zero, since

$$\begin{aligned} H_W H_B^T &= \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})(c^{(i)} - c)^T \\ &= \sum_{i=1}^{k} (n_i c^{(i)} c^{(i)^T} - n_i c^{(i)} c^T - n_i c^{(i)} c^{(i)^T} + n_i c^{(i)} c^T) = 0. \qquad \square \end{aligned}$$

PROPERTY 2 *For* $K = (H_B, H_W)^T \in \mathbb{R}^{2n \times m}$, $t = \text{rank}(K) \leq n$.

Proof. For $K^T = (H_B, H_W) \in \mathbb{R}^{m \times 2n}$, we have

$$\text{rank}(K^T) + \dim(\text{null}(K^T)) = 2n, \quad \text{or} \quad \dim(\text{null}(K^T)) = 2n - t.$$

Hence, $t \leq n$ if and only if $\dim(\text{null}(K^T)) \geq n$. Suppose $z_1 \in \text{null}(H_B)$ and $z_2 \in \text{null}(H_W)$. Then

$$(H_B, H_W) \begin{pmatrix} z_1 \\ 0 \end{pmatrix} = (H_B, H_W) \begin{pmatrix} 0 \\ z_2 \end{pmatrix} = 0.$$

This shows that

$$\dim(\text{null}(H_B, H_W)) \geq \dim(\text{null}(H_B)) + \dim(\text{null}(H_W)).$$

Property 1 implies

$$\dim(\text{null}(H_W)) \geq \text{rank}(H_B^T).$$

Combining this with

$$\dim(\text{null}(H_B)) = n - \text{rank}(H_B),$$

we have

$$\dim(\text{null}(H_B, H_W)) \geq n - \text{rank}(H_B) + \text{rank}(H_B^T) = n. \qquad \square$$

Now we proceed with the GSVD derivation. For the case of $(S_1, S_2) = (S_B, S_W)$, consider the GSVD of the pair (H_B^T, H_W^T), which is given by

$$U^T H_B^T X = (\Sigma_B, 0) \quad \text{and} \quad V^T H_W^T X = (\Sigma_W, 0),$$

where

$$\Sigma_B \quad \text{and} \quad \Sigma_W \in \mathbb{R}^{n \times t}, \quad \Sigma_B^T \Sigma_B + \Sigma_W^T \Sigma_W = I_t, \quad \text{and} \quad t = \text{rank} \begin{pmatrix} H_B^T \\ H_W^T \end{pmatrix}.$$

Then we have

$$H_M^T = U(\Sigma_B, 0)X^{-1} + V(\Sigma_W, 0)X^{-1} = U(\Sigma_B + U^T V \Sigma_W, 0)X^{-1}.$$

In addition,

$$H_W H_B^T = X^{-T} \begin{pmatrix} \Sigma_W^T \\ 0 \end{pmatrix} V^T U (\Sigma_B, 0) X^{-1} = 0_m$$

implies

$$\Sigma_W^T V^T U \Sigma_B = 0_t.$$

Hence

$$\begin{aligned} (\Sigma_B + U^T V \Sigma_W)^T (\Sigma_B + U^T V \Sigma_W) &= \Sigma_B^T \Sigma_B + \Sigma_W^T (V^T U U^T V) \Sigma_W + \Sigma_W^T V^T U \Sigma_B + \Sigma_B^T U^T V \Sigma_W \\ &= \Sigma_B^T \Sigma_B + \Sigma_W^T \Sigma_W = I_t, \end{aligned}$$

which means $\Sigma_B + U^T V \Sigma_W$ has orthonormal columns. This can only be true if $\Sigma_B + U^T V \Sigma_W$ has no more columns than rows, i.e. if $t \leq n$ as shown above in Property 2.

There exists $\hat{U}_2$ such that $(\Sigma_B + U^T V \Sigma_W, \hat{U}_2) \in \mathbb{R}^{n \times n}$ is orthogonal. Hence

$$H_M^T = U(\Sigma_B + U^T V \Sigma_W, \hat{U}_2) \begin{pmatrix} I_t & 0 \\ 0 & 0 \end{pmatrix} X^{-1},$$

and we can write

$$\hat{U}^T H_M^T X = (\Sigma_M, 0),$$

where

$$\hat{U} = U(\Sigma_B + U^T V \Sigma_W, \hat{U}_2) \quad \text{is orthogonal and} \quad \Sigma_M = \begin{pmatrix} I_t \\ 0 \end{pmatrix}.$$

Together with

$$V^T H_W^T X = (\Sigma_W, 0),$$

this forms a GSVD of the matrix pair (H_M^T, H_W^T), which has the same generalized right singular vectors as (H_B^T, H_W^T). As expected, each of the t nontrivial generalized singular values is infinite, finite and greater than one, or equal to one. Note that this form of GSVD for (H_M^T, H_W^T) does not satisfy the condition $\Sigma_M^T \Sigma_M + \Sigma_W^T \Sigma_W = I$ of the Paige and Saunders [PS81] formulation because each $\lambda_i \geq 1$. However, the invariance property and nonuniqueness of the right singular vector matrix X can be used to convert it to the Paige and Saunders form.

Note that if S_W is nonsingular, in the m-dimensional space,

$$\mathrm{trace}(S_W^{-1} S_M) = \mathrm{trace}(S_W^{-1}(S_W + S_B)) = m + \mathrm{trace}(S_W^{-1} S_B), \tag{17}$$

and in the l-dimensional space,

$$\mathrm{trace}((S_W^Y)^{-1} S_M^Y) = \mathrm{trace}((S_W^Y)^{-1}(S_W^Y + S_B^Y)) = l + \mathrm{trace}((S_W^Y)^{-1} S_B^Y). \tag{18}$$

This confirms that the solutions are the same for both $(S_1, S_2) = (S_B, S_W)$ and $(S_1, S_2) = (S_M, S_W)$. From Section 3.1, when G includes the eigenvectors of $S_W^{-1} S_B$ corresponding to the $l \geq k-1$ largest eigenvalues, then

$$\mathrm{trace}(S_W^{-1} S_B) = \mathrm{trace}((S_W^Y)^{-1} S_B^Y).$$

By subtracting (18) from (17) for any $l \geq k-1$, we get

$$\mathrm{trace}((S_W^Y)^{-1} S_M^Y) + (m - l) = \mathrm{trace}(S_W^{-1} S_M). \tag{19}$$

In other words, each additional eigenvector beyond the leftmost $k-1$ will add one to $\mathrm{trace}((S_W^Y)^{-1} S_M^Y)$. This shows that we do not preserve the cluster structure when measured by $\mathrm{trace}(S_W^{-1} S_M)$, although we do preserve $\mathrm{trace}(S_W^{-1} S_B)$. According to Eqn. (19), $\mathrm{trace}(S_W^{-1} S_M)$ will be preserved only if we include all m eigenvectors of $S_W^{-1} S_M$. This, together with Section 3.1, show indirectly that $\mathrm{rank}(S_M) = m$. That is, S_M is nonsingular whenever S_W is.

For the case

$$(S_1, S_2) = (S_W, S_M),$$

we want to minimize $\mathrm{trace}(S_M^{-1} S_W)$. Once again, the relation (2) can be used to rewrite the generalized eigenvalue problem. $S_W x_i = \lambda_i S_M x_i$ becomes

$$S_B x_i = (\frac{1}{\lambda_i} - 1) S_W x_i,$$

for $\lambda_i \neq 0$. The eigenvector matrix is the same, but the eigenvalue matrix is $\Lambda^{-1} - I$. When $\lambda_1 \geq \cdots \geq \lambda_m$, we have

$$\frac{1}{\lambda_1} - 1 \leq \cdots \leq \frac{1}{\lambda_m} - 1,$$

so the same permutation can be used to put $\Lambda^{-1} - I$ in nondecreasing order as put Λ in nonincreasing order. After permuting, x_i corresponds to both the ith smallest eigenvalue of $S_M^{-1} S_W$ and to the ith largest eigenvalue of $S_W^{-1} S_B$. Therefore, for a given value of l, we use the first l eigenvectors, just as we did for $(S_1, S_2) = (S_B, S_W)$.

Again we consider a generalized right singular vector $x_i \in \text{null}(S_W)$. For that x_i, $S_M x_i = S_B x_i$, so the same reasoning applies regarding the effect on $\text{trace}(G^T S_M G)$ and $\text{trace}(G^T S_W G)$. Therefore, the solution is the same as for $(S_1, S_2) = (S_B, S_W)$, even in the singular case. The GSVD of the matrix pair (H_W^T, H_M^T) can be derived from that of (H_B^T, H_W^T) in the same way as shown above for (H_M^T, H_W^T). However, since we are minimizing in this case, the generalized singular values are in nondecreasing order, taking on reciprocal values of those for (H_M^T, H_W^T).

Having shown the equivalence of the J_1 criteria for various (S_1, S_2), we conclude that

$$(S_1, S_2) = (S_B, S_W)$$

should be used for the sake of computational efficiency. The LDA/GSVD algorithm reduces computational complexity further by using a lower dimensional form of H_B rather than that presented in Eqn. (12), and it avoids a potential loss of information [GVL96, page 239, Example 5.3.2] by not explicitly forming S_B and S_W as cross-products of H_B and H_W.

3.4 Generalization of $\mathbf{J_2} = \ln(\det(\mathbf{S_2^{-1} S_1}))$ Criteria for Singular $\mathbf{S_2}$

Consider

$$J_2(G) = \ln(\det((G^T S_2 G)^{-1} G^T S_1 G)) = \ln(\det(G^T S_1 G)) - \ln(\det(G^T S_2 G)),$$

where the scatter matrices S_1 and S_2 are nonsingular. It can be shown [Fuk90] that

$$\frac{\partial J_2(G)}{\partial G} = 2 S_1 G (G^T S_1 G)^{-1} - 2 S_2 G (G^T S_2 G)^{-1},$$

and setting this to zero yields

$$S_2^{-1} S_1 G = G (G^T S_2 G)^{-1} (G^T S_1 G) = G((S_2^Y)^{-1} S_1^Y). \tag{20}$$

If we simultaneously diagonalize S_1^Y and S_2^Y, we get

$$Z^T S_1^Y Z = \Lambda = \text{diag}(\lambda_1 \ldots \lambda_l) \quad \text{and} \quad Z^T S_2^Y Z = I_l,$$

where $Z \in \mathbb{R}^{l \times l}$ is nonsingular. Hence

$$(S_2^Y)^{-1} S_1^Y = Z \Lambda Z^{-1}$$

and Eqn. (20) becomes $S_2^{-1}S_1GZ = GZ\Lambda$, where $GZ \in \mathbb{R}^{m\times l}$ consists of l eigenvectors of $S_2^{-1}S_1$. By the same argument we made for J_1, $J_2(G) = J_2(GZ)$, and so

$$\begin{aligned} J_2(G) &= \ln(\det((GZ)^T S_1(GZ))) - \ln(\det((GZ)^T S_2(GZ))) \\ &= \ln(\det(((GZ)^T S_2(GZ))^{-1}(GZ)^T S_2(GZ)\Lambda)) \\ &= \ln(\det(\Lambda)) = \ln\lambda_1 + \cdots + \ln\lambda_l. \end{aligned}$$

This shows that an optimum G satisfies the same generalized eigenvalue problem as for J_1, and that we should choose the eigenvectors that correspond to the l largest (smallest) eigenvalues of $S_2^{-1}S_1$ if we are maximizing (minimizing) J_2.

We now extend the optimization of J_2 to singular matrices for the case where $(S_1, S_2) = (S_B, S_W)$. Consider a generalized right singular vector $x_i \in \operatorname{null}(S_B)$. If x_i is included in G, then $G^T S_B G$ has a zero column which forces its determinant to zero. Since we want to maximize $\ln(\det(G^T S_B G))$, we restrict G to the $l = \operatorname{rank}(S_B)$ generalized right singular vectors that correspond to the largest generalized singular values. However, these leftmost $\operatorname{rank}(S_B)$ vectors may include a vector $x_i \in \operatorname{null}(S_W) - \operatorname{null}(S_B)$. Including this x_i will force $\ln(\det(G^T S_W G))$, which we want to minimize, to $-\infty$. Therefore, in the singular case, we include the leftmost $\operatorname{rank}(S_B)$ generalized right singular vectors, just as we did for trace optimization of J_1.

4 Alternative Approaches

4.1 Orthogonal Centroid Method

Simpler criteria for preserving cluster structure, such as $\min \operatorname{trace}(G^T S_W G)$ and $\max \operatorname{trace}(G^T S_B G)$, involve only one of the scatter matrices. A straightforward minimization of $\operatorname{trace}(G^T S_W G)$ seems meaningless since the optimum always reduces the dimension to one, even when the solution is restricted to the case when G has orthonormal columns. On the other hand, with the same restriction, maximization of $\operatorname{trace}(G^T S_B G)$ produces an equivalent solution to the Orthogonal Centroid method, which is introduced and shown to give promising reduced dimensional classification results in [PJR], and is summarized in Algorithm 2.

Let

$$J_3(G) = \operatorname{trace}(G^T S_B G).$$

If we let $G \in \mathbb{R}^{m\times l}$ be any matrix with full column rank, then essentially there is no upper bound and maximization is also meaningless. Now, let us restrict the solution to the case when G has orthonormal columns. Then there exists $\hat{G} \in \mathbb{R}^{m\times(m-l)}$ such that $\begin{pmatrix} G, & \hat{G} \end{pmatrix}$ is an orthogonal matrix. In addition, since S_B is positive semidefinite, we have

$$\operatorname{trace}(G^T S_B G) \le \operatorname{trace}(G^T S_B G) + \operatorname{trace}(\hat{G}^T S_B \hat{G}) = \operatorname{trace}(S_B).$$

If the SVD of H_B is given by $H_B = U\Sigma V^T$, then $S_B U = U\Sigma\Sigma^T$. Hence the columns of U form an orthonormal set of eigenvectors of S_B corresponding to the nonincreasing eigenvalues on the diagonal of

$\Lambda = \Sigma\Sigma^T$. For $r = \mathrm{rank}(S_B)$, if we let U_r denote the first r columns of U and $\Lambda_r = \mathrm{diag}(\lambda_1 \ldots \lambda_r)$, we have

$$\begin{aligned} J_3(U_r) &= \mathrm{trace}(U_r^T S_B U_r) \\ &= \mathrm{trace}(U_r^T U_r \Lambda_r) \\ &= \lambda_1 + \cdots + \lambda_r \\ &= \mathrm{trace}(S_B). \end{aligned}$$

This means that we preserve $\mathrm{trace}(S_B)$ if we take U_r as G.

Now we show that this solution is equivalent to the solution of the Orthogonal Centroid method, which does not involve the computation of eigenvectors. Defining the centroid matrix

$$C = \begin{pmatrix} c^{(1)} & c^{(2)} & \cdots & c^{(k)} \end{pmatrix}$$

as in Algorithm 2, C has the reduced QR decomposition $C = Q_k R$, where $Q_k \in \mathbb{R}^{m \times k}$ has orthonormal columns and $R \in \mathbb{R}^{k \times k}$ [GVL96]. Suppose x is an eigenvector of S_B corresponding to the eigenvalue $\lambda \neq 0$. Then

$$S_B x = \sum_{i=1}^{k} n_i (c^{(i)} - c)(c^{(i)} - c)^T x = \lambda x.$$

This means $x \in \mathrm{span}\{c^{(i)} - c | 1 \leq i \leq k\}$, and hence $x \in \mathrm{span}\{c^{(i)} | 1 \leq i \leq k\}$. Accordingly,

$$\mathrm{range}(U_r) \subseteq \mathrm{range}(C) \subseteq \mathrm{range}(Q_k),$$

which implies that

$$U_r = Q_k W$$

for some matrix $W \in \mathbb{R}^{k \times r}$ with orthonormal columns. This yields

$$\begin{aligned} J_3(U_r) &= \mathrm{trace}(W^T Q_k^T S_B Q_k W) \\ &\leq \mathrm{trace}(Q_k^T S_B Q_k) \\ &= J_3(Q_k), \end{aligned}$$

where the inequality follows from the positive semidefiniteness of $Q_k^T S_B Q_k$. Hence

$$J_3(Q_k) = \mathrm{trace}(S_B),$$

and Q_k plays the same role as U_r. In other words, instead of computing the eigenvectors, we simply need to compute Q_k, which is much cheaper. Therefore, by computing a reduced QR decomposition of the centroid matrix, we obtain a solution that maximizes $\mathrm{trace}(G^T S_B G)$ over all G with orthonormal columns.

Algorithm 2 Orthogonal Centroid Method

Given a data matrix $A \in \mathbb{R}^{m\times n}$ with k clusters and an input vector $a \in \mathbb{R}^{m\times 1}$, compute a k-dimensional representation y of a.

1. Compute the centroid $c^{(i)}$ of the ith cluster, $1 \le i \le k$.
2. Set $C = \begin{pmatrix} c^{(1)} & c^{(2)} & \cdots & c^{(k)} \end{pmatrix}$.
3. Compute the matrix Q_k in the reduced QR decomposition $C = Q_k R$.
4. $y = Q_k^T a$.

4.2 Two-stage Approach

As mentioned in the Introduction, another approach for dealing with the singularity of S_W when $m > n$ uses LSI/SVD as a first stage, followed by the discriminant analysis stage. The goal of the first stage is to reduce the dimension of the data matrix enough so that the new S_W is nonsingular, and classical LDA can be performed. LSI/SVD uses the truncated SVD to find a rank-l approximation of A. That is, if $l \le \mathrm{rank}(A)$, then

$$A \approx U_l \Sigma_l V_l^T$$

where the columns of U_l are the first l left singular vectors, Σ_l is a diagonal matrix with the l largest singular values in nonincreasing order along its diagonal, and the columns of V_l are the first l right singular vectors. LSI/SVD typically uses $\Sigma_l V_l^T$ as the reduced dimensional representation of A, or equivalently, it computes the l-dimensional representation of $a \in \mathbb{R}^{m\times 1}$ as $y = U_l^T a$.

It is well known that the truncated SVD provides the closest approximation to A in Frobenius or L_2 norm. However, unless performed locally on each cluster as in [Hul94, SHP95], LSI ignores the cluster structure while reducing the dimension to l. Since there is no theoretical optimum value of l, potentially expensive testing may be required to determine the intermediate reduced dimensional representation of A that should be the input for the LDA stage.

5 Experimental Results

In this section, we demonstrate the effectiveness of the LDA/GSVD and Orthogonal Centroid algorithms, which use the J_1 criterion with $(S_1, S_2) = (S_B, S_W)$ and the J_3 criterion with $G^T G = I$, respectively. For LDA/GSVD, we confirm its mathematical equivalence to J_1 using an alternative choice of (S_1, S_2), and we illustrate the discriminatory power of J_1 via two-dimensional projections. Just as important, we validate our extension of J_1 to the singular case. For Orthogonal Centroid, its preservation of $\mathrm{trace}(S_B)$ is shown to be a very effective compromise for the simultaneous optimization of two traces approximated by J_1. Our final tests show the sensitivity of the two-stage approach to the reduced dimension in its first stage, thus strengthening our contention that the single-stage LDA/GSVD is a more effective approach.

5.1 Equivalence of J_1 for (S_B, S_W) and (S_M, S_W)

Table 1: Traces and Misclassification Rates (in %) with L_2 norm similarity

Method	Full	trace$(S_W^{-1}S_B)$	trace$(S_W^{-1}S_M)$	
Dim	150×2000	6×2000	6×2000	7×2000
trace(S_W)	299700	1.97	1.48	1.98
trace(S_B)	22925	4.03	3.04	3.04
trace(S_M)	322630	6.00	4.52	5.02
trace$(S_W^{-1}S_B)$	12.6	12.6	12.6	12.6
trace$(S_W^{-1}S_M)$	162.6	18.6	18.6	19.6
centroid	2.6 %	2.2 %	2.0 %	2.0 %
5nn	18.7 %	2.2 %	2.2 %	2.4 %
15nn	10.1 %	1.8 %	1.9 %	2.1 %

In Table 1, we use clustered data that are artificially generated by an algorithm adapted from [JD88, Appendix H]. The data consist of 2000 vectors in a space of dimension 150, with $k = 7$ clusters. LDA/GSVD reduces the dimension from 150 to $k - 1 = 6$. We compare the LDA/GSVD criterion, $J_1 = \mathrm{trace}(S_W^{-1}S_B)$, with the alternative J_1 criterion, $\mathrm{trace}(S_W^{-1}S_M)$. The trace values confirm our theoretical findings, namely that the generalized eigenvectors that optimize the alternative J_1 also optimize LDA/GSVD's J_1, and including an additional eigenvector increases $\mathrm{trace}(S_W^{-1}S_M)$ by one.

We also report misclassification rates for a centroid-based classification method [HJP] and the k nearest neighbor (knn) classification method [TK99], which are summarized in Algorithms 3 and 4. (Note that the classification parameter of knn differs from the number of clusters k.) These are obtained using the L_2 norm, or Euclidean distance, similarity measure. While these rates differ slightly with the choice of S_B or S_M, and the reduction to six or seven rows using the latter, they establish no advantage of using S_M over S_B, even when we include an additional eigenvector to bring us closer to the preservation of $\mathrm{trace}(S_W^{-1}S_M)$. These results bolster our argument that the correct choice of J_1 is optimized in our LDA/GSVD algorithm, since it limits the GSVD computation to a composite matrix with $k + n$ rows, rather than one with $2n$ rows.

5.2 Discriminatory Power of J_1

To further illustrate the power of the J_1 criterion, we apply it to the same 2000 data vectors as in Section 5.1, this time reducing the dimension from 150 to two. Even though the optimal reduced dimension is six, J_1 does surprisingly well at discriminating among seven classes, as seen in Figure 1. As expected, the alternative J_1 does equally well in Figure 2. In contrast, Figure 3 shows that the truncated SVD, as used in LSI, is not the best discriminator.

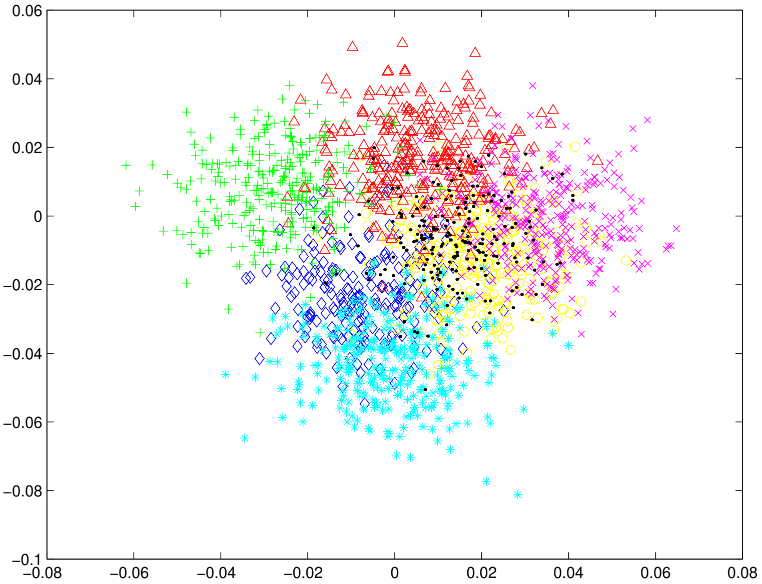

Figure 1: Max trace$(S_W^{-1} S_B)$ projection onto two dimensions.

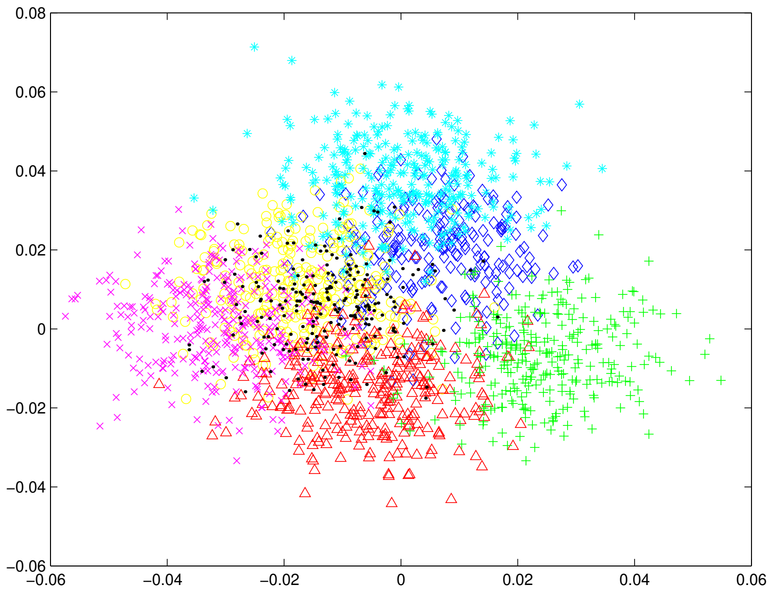

Figure 2: Max trace$(S_W^{-1} S_M)$ projection onto two dimensions.

Algorithm 3 Centroid-based Classification

Given a data matrix A with k clusters and k corresponding centroids, $c^{(i)}$ for $1 \le i \le k$, find the index j of the cluster to which a vector q belongs.

- find the index j such that $sim(q, c^{(i)})$, $1 \le i \le k$, is minimum (or maximum), where $sim(q, c^{(i)})$ is the similarity measure between q and $c^{(i)}$.

  (For example, $sim(q, c^{(i)}) = \|q - c^{(i)}\|_2$ using the L_2 norm, and we take the index with the minimum value. Using the cosine measure, $sim(q, c^{(i)}) = cos(q, c^{(i)}) = \frac{q^T c^{(i)}}{\|q\|_2 \|c^{(i)}\|_2}$, and we take the index with the maximum value.)

Algorithm 4 k Nearest Neighbor (knn) Classification

Given a data matrix $A = [a_1 \quad \ldots \quad a_n]$ with k clusters, find the cluster to which a vector q belongs.

1. From the similarity measure $sim(q, a_j)$ for $1 \le j \le n$, find the k nearest neighbors of q. (We use k to distinguish the algorithm parameter from the number of clusters k.)
2. Among these k vectors, count the number belonging to each cluster.
3. Assign q to the cluster with the greatest count in the previous step.

5.3 Comparison to Orthogonal Centroid Method in Singular Case

Another set of experiments validates our extension of J_1 to the singular case. For this purpose, we use five categories of abstracts from the MEDLINE [1] database (see Table 2). Each category has 40 documents. There are 7519 terms after preprocessing with stemming and removal of stop words [Kow97]. Since 7519 exceeds the number of documents (200), S_W is singular and classical discriminant analysis breaks down. However, our LDA/GSVD method circumvents this singularity problem.

The LDA/GSVD algorithm dramatically reduces the dimension 7519 to four, or one less than the number of clusters. The Orthogonal Centroid method reduces the dimension to five. Table 3 shows classification results using the L_2 norm similarity measure. LDA/GSVD produces the lowest misclassification rate using both centroid-based and nearest neighbor classification methods. Because the J_1 criterion is not defined in this case, we compute the ratio $\mathrm{trace}(S_B)/\mathrm{trace}(S_W)$ as a rough optimality measure. We observe that the ratio is strikingly higher for LDA/GSVD reduction than for the other methods. These experimental results confirm that the LDA/GSVD algorithm effectively extends the applicability of the J_1 criterion to cases that classical discriminant analysis cannot handle. In addition, the Orthogonal Centroid algorithm preserves $\mathrm{trace}(S_B)$ from the full dimension without the expense of computing eigenvectors. Taken together, the results for these two methods demonstrate the potential for dramatic and efficient dimension reduction without compromising cluster structure.

[1] http://www.ncbi.nlm.nih.gov/PubMed

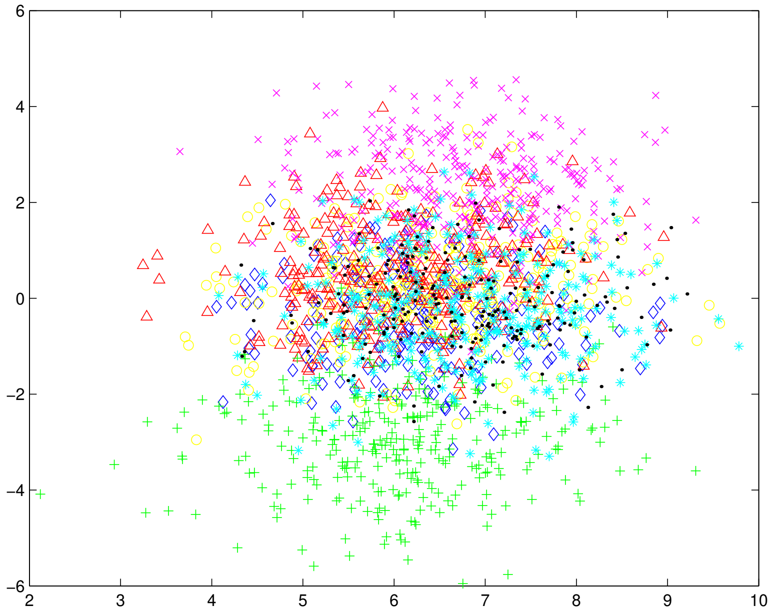

Figure 3: Two-dimensional representation using $\Sigma_2 V_2^T$ from truncated SVD.

Table 2: MEDLINE Data Set

class	category	no. of documents
1	heart attack	40
2	colon cancer	40
3	diabetes	40
4	oral cancer	40
5	tooth decay	40
	dimension	7519×200

Table 3: Traces and Misclassification Rate with L_2 norm similarity

Method		Full	Orthogonal Centroid	LDA/GSVD
Dim		7519×200	5×200	4×200
trace values	trace(S_W)	73048	4210	0.05
	trace(S_B)	6229	6229	3.95
	$\frac{\text{trace}(S_B)}{\text{trace}(S_W)}$	0.09	1.5	79
misclassification rate in %	centroid	5	5	1
	1nn	40	3	1

Table 4: Traces and Misclassification Rate (in %) with L_2 norm similarity

Method	Full	LDA/GSVD	LSI$\rightarrow$ 200 LDA/GSVD	LSI$\rightarrow$ 200 LDA/EIG
Dim	7519×200	4×200	4×200	4×200
$\mathrm{trace}(S_W)$	73048	0.05	0.05	3.17×10^{-54}
$\mathrm{trace}(S_B)$	6229	3.95	3.95	6.11×10^{-25}
$\frac{\mathrm{trace}(S_B)}{\mathrm{trace}(S_W)}$	0.09	79	79	1.92×10^{29}
centroid	5%	1%	1%	0%
1nn	40%	1%	0%	0%
3nn	51%	1.5%	1.5%	19%

5.4 Comparison to Two-stage Approach in Singular Case

A final set of experiments also uses the MEDLINE database of Table 2. The results are summarized in Tables 4 and 5. Table 4 compares our LDA/GSVD method with two-stage approaches whose LSI stage reduces the data dimension to $n = 200$. Although the data matrix is square after the LSI stage, S_W remains singular. We first note that if the second stage uses the GSVD, then the final trace values are identical to those of the single stage of LDA/GSVD, and the misclassification rates are almost identical. However, if MATLAB's eig function is used for the second stage, the trace values are scaled quite differently, and the classification results are slightly better for centroid and 1nn classification and considerably worse for 3nn classification. Clearly, the intermediate reduction to a square matrix produces widely varying results depending on which LDA algorithm is used in the second stage.

In Table 5, the dimension reduction methods vary only in the intermediate dimension after the LSI stage. Since $\mathrm{rank}(S_W) = 195$, we include it in our range of LSI dimensions, and conclude with LSI to the LDA/GSVD optimum dimension of 4. Our rough optimality measure, $\mathrm{trace}(S_B)/\mathrm{trace}(S_W)$, declines as the LSI dimension decreases, and misclassification rates increase over the same range. These tests clearly show the sensitivity of the two-stage approach to the dimension chosen in the LSI stage.

These tests of the two-stage approach also bring up several issues in its usage. First of all, what LSI dimension will result in nonsingular S_W? Second, when choosing the generalized eigenvectors to include as columns of the G matrix in the LDA stage, what is the meaning of negative generalized eigenvalues? This is in contrast to the GSVD approach, for which we have infinite, finite and positive, zero, and arbitrary generalized singular values, and a rationale for the inclusion or exclusion of the corresponding generalized singular vectors in the solution matrix G. Third, which algorithm should be used for the LDA stage, particularly when the LSI dimension is close to n so that the LSI representation is square but S_W is singular. These issues will be explored in [HP].

Table 5: Traces and Misclassification Rate (in %) with L_2 norm similarity

Method	LSI$\rightarrow$ 195 LDA/EIG	LSI$\rightarrow$ 150 LDA/EIG	LSI$\rightarrow$ 50 LDA/EIG	LSI$\rightarrow$ 20 LDA/EIG	LSI$\rightarrow$ 4 LDA/EIG
Dim	4×200	4×200	4×200	4×200	4×200
$\mathrm{trace}(S_W)$	14.07	313	1446	2963	6962
$\mathrm{trace}(S_B)$	850.60	2903	4555	5124	3473
$\frac{\mathrm{trace}(S_B)}{\mathrm{trace}(S_W)}$	60.42	9.27	3.15	1.73	0.50
centroid	1%	5%	6%	8%	34.5%
1nn	2%	3.5%	4%	8%	24%
3nn	1%	2.5%	3.5%	7.5%	33.5%

6 Conclusion

Our experimental results verify that the J_1 criterion, when applicable, effectively optimizes classification in the reduced dimensional space, while our LDA/GSVD extends the applicability to cases that classical discriminant analysis cannot handle. In addition, our LDA/GSVD algorithm avoids the numerical problems inherent in explicitly forming the scatter matrices.

A disadvantage of methods that involve the GSVD is that its computation is costly. Computationally, the most expensive part of Algorithm LDA/GSVD is Step 2, where a complete orthogonal decomposition is needed. Assuming $k \leq n$, $t \leq m$, and $t = \mathcal{O}(n)$, the complete orthogonal decomposition of K costs $\mathcal{O}(nmt)$ when $m \leq n$, and $\mathcal{O}(m^2t)$ when $m > n$ [GVL96]. Therefore, a fast algorithm needs to be developed for step 2.

For Orthogonal Centroid, the most expensive step is the reduced QR decomposition of C, which costs $\mathcal{O}(mk^2)$ [GVL96]. By solving a simpler eigenvalue problem and avoiding the computation of eigenvectors, Orthogonal Centroid is significantly cheaper than LDA/GSVD. Our experiments show it to be a very reasonable compromise.

Compared to the two-stage approach of LSI followed by LDA, our one-stage LDA/GSVD avoids the potentially costly experimentation involved in determining the dimension for LSI. Short of experimenting with various LSI dimensions, one could reduce the data to dimension n so that the matrix is square, but classification results after the LDA stage may vary widely depending on the LDA method chosen. Our preliminary results show that use of the GSVD may have numerical advantages in this context as well.

Finally, it bears repeating that dimension reduction is only a preprocessing stage. Since classification and document retrieval will be the dominating parts computationally, the expense of dimension reduction should be weighed against its effectiveness in reducing the cost involved in those processes.

References

[BDO95] M.W. Berry, S.T. Dumais, and G.W. O'Brien. Using linear algebra for intelligent information retrieval. SIAM Review, 37(4):573–595, 1995.

[DDF+90] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. J. American Society for Information Science, 41:391–407, 1990.

[Fuk90] K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, second edition, 1990.

[GVL96] G.H. Golub and C.F. Van Loan. Matrix Computations. Johns Hopkins University Press, third edition, 1996.

[HJP] P. Howland, M. Jeon, and H. Park. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. To appear in SIAM J. Matrix Anal. Appl.

[HP] P. Howland and H. Park. An improved algorithm for linear discriminant analysis. in preparation.

[Hul94] D. Hull. Improving text retrieval for the routing problem using latent semantic indexing. In Proc. SIGIR, pages 282–291, 1994.

[JD88] A.K. Jain and R.C. Dubes. Algorithms for Clustering Data. Prentice Hall, 1988.

[Kow97] G. Kowalski. Information Retrieval Systems : Theory and Implementation. Kluwer Academic Publishers, 1997.

[LH95] C.L. Lawson and R.J. Hanson. Solving Least Squares Problems. SIAM, 1995.

[Ort87] J. Ortega. Matrix Theory: A Second Course. Plenum Press, 1987.

[PJR] H. Park, M. Jeon, and J.B. Rosen. Lower dimensional representation of text data based on centroids and least squares. To appear in BIT.

[PS81] C.C. Paige and M.A. Saunders. Towards a generalized singular value decomposition. SIAM J. Numer. Anal., 18(3):398–405, 1981.

[SHP95] H. Schütze, D. Hull, and J. Pedersen. A comparison of classifiers and document representations for the routing problem. In Proc. SIGIR, pages 229–237, 1995.

[TK99] S. Theodoridis and K. Koutroumbas. Pattern Recognition. Academic Press, 1999.

[Tor01] K. Torkkola. Linear discriminant analysis in document classification. In IEEE ICDM Workshop on Text Mining, 2001.

[VL76] C.F. Van Loan. Generalizing the singular value decomposition. SIAM J. Numer. Anal., 13(1):76–83, 1976.