Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 02-021

Extension of Discriminant Analysis based on the Generalized Singular Value Decomposition

Peg Howland and Haesun Park

May 30, 2002

Extension of Discriminant Analysis based on the Generalized Singular Value Decomposition

Peg Howland* and Haesun Park†

March 1, 2002

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 utilized in application areas such as information retrieval to reduce the dimension of data while preserving its cluster structure. In the process, we establish relationships between the solutions obtained by various methods, which allow us to refine the optimization criteria and to improve the algorithms for achieving them.

1 Introduction

The goal of discriminant analysis is to map the original data into features that most effectively discriminate 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, for a data matrix $A \in \mathbb{R}^{m\times n}$, we want to find a linear transformation $G^T \in \mathbb{R}^{l\times m}$ that maps each column a_i, $1 \le i \le n$, of A in the m-dimensional space to a column y_i in the l-dimensional space :

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

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 are defined in discriminant analysis [3, 10]. For simplicity of discussion, we will assume that the columns of $A \in \mathbb{R}^{m\times n}$ are grouped into k clusters as

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

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

$$c^{(i)} = \frac{1}{n_i} A_i e^{(i)} \quad \text{where} \quad e^{(i)} = (1, \ldots, 1)^T \in \mathbb{R}^{n_i \times 1},$$

*The work of this author was supported in part by the National Science Foundation grant CCR-9901992 and the Guidant Fellowship. Dept. of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, MN 55455, U.S.A.(howland@cs.umn.edu)

†The work of this author was supported in part by the National Science Foundation grant CCR-9901992. Dept. of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, MN 55455, U.S.A.(hpark@cs.umn.edu), and Korea Institute for Advanced Study, 207-43 Cheongryangri-dong, Dongdaemun-gu, Seoul 130-012, KOREA (from Sept. 2002 to July 2003)

and the global centroid c is defined as

$$c = \frac{1}{n}Ae, \quad \text{where} \quad e = (1, \ldots, 1)^T \in \mathbb{R}^{n\times 1}.$$

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 [5] that the scatter matrices have the relationship

$$S_m = S_w + S_b. \tag{1}$$

Applying G^T to 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 [3, 10]. Since

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

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 is impossible, but it can be approximated by maximizing $\mathrm{trace}((S_w^Y)^{-1} S_b^Y)$. More general criteria include $\mathrm{trace}((S_2^Y)^{-1} S_1^Y)$ and $\ln(\det((S_2^Y)^{-1} S_1^Y))$, where S_1 and S_2 are chosen from S_w, S_b, and S_m. In this paper, we use the generalized singular value decomposition (GSVD) [4, 9, 11] to extend the applicability of these criteria to the case when S_2 is singular, as well as to establish the equivalence between alternate choices for S_1 and S_2. In addition, we describe the original two-class case in terms of the GSVD. We also address the optimization of the trace of an individual scatter matrix, and present an efficient method for achieving it. Finally, we present experimental results demonstrating the capabilities of the GSVD approach, as well as the potential for developing more efficient algorithms for trace optimization.

2 Generalized Singular Value Decomposition

After the GSVD was originally defined by Van Loan [11], Paige and Saunders [9] developed the following formulation for any two matrices with the same number of columns.

THEOREM 1 *Suppose two matrices $K_A \in \mathbb{R}^{m\times n}$ and $K_B \in \mathbb{R}^{p\times n}$ are given. Then for*

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

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

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

where

$$\underset{m\times t}{\Sigma_A} = \begin{pmatrix} I_A & & \\ & D_A & \\ & & 0_A \end{pmatrix}, \quad \underset{p\times t}{\Sigma_B} = \begin{pmatrix} O_B & & \\ & D_B & \\ & & I_B \end{pmatrix},$$

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

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

are identity matrices, where

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

$$0_A \in \mathbb{R}^{(m-r-s)\times(t-r-s)} \quad \textit{and} \quad 0_B \in \mathbb{R}^{(p-t+r)\times r}$$

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

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

satisfy

$$1 > \alpha_{r+1} \geq \cdots \geq \alpha_{r+s} > 0, \quad 0 < \beta_{r+1} \leq \cdots \leq \beta_{r+s} < 1, \tag{2}$$

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 [9]

$$U^T K_A X = (\Sigma_A, 0) \quad \text{and} \quad V^T K_B X = (\Sigma_B, 0), \tag{3}$$

where

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

From the form in Eqn. (3) 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 $\alpha_i = 1$, $\beta_i = 0$ for $i = 1, \ldots, r$ and $\alpha_i = 0$, $\beta_i = 1$ for $i = r+s+1, \ldots, t$, we have, for $1 \leq i \leq t$,

$$\beta_i^2 K_A^T K_A x_i = \alpha_i^2 K_B^T K_B x_i, \tag{4}$$

where x_i represents the ith column of X. For the remaining $n - t$ columns of X, both $K_A^T K_A x_i$ and $K_B^T K_B x_i$ are zero, so Eqn. (4) is satisfied for arbitrary values of α_i and β_i when $t + 1 \leq i \leq n$. 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 Extension of 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. We conclude with an efficient method for optimizing the trace of an individual scatter matrix.

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

For now, we will focus our discussion on the criteria of optimizing $J_1(G) = \mathrm{trace}((G^T S_2 G)^{-1}(G^T S_1 G))$, 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 [4], there exists a nonsingular matrix $X \in \mathbb{R}^{m\times m}$ such that

$$X^T S_1 X = \Lambda = \mathrm{diag}(\lambda_1 \ldots \lambda_m) \quad \text{and} \quad X^T S_2 X = I_m. \tag{5}$$

Letting x_i denote the ith column of X, we have

$$S_1 x_i = \lambda_i S_2 x_i, \tag{6}$$

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 and $x_i^T S_1 x_i = \lambda_i$, $\lambda_i \geq 0$ for $1 \leq i \leq m$. From (5), 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 [4]. 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, even though J_1 is minimized for some choices of S_1 and S_2.) When $l \geq q$, 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

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

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

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

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 text processing in information retrieval, is that the matrix S_2 must be nonsingular. Defining the matrices

$$H_w = [A_1 - c^{(1)} e^{(1)^T}, A_2 - c^{(2)} e^{(2)^T}, \ldots, A_k - c^{(k)} e^{(k)^T}] \in \mathbb{R}^{m \times n}, \tag{8}$$

$$H_b = [(c^{(1)} - c) e^{(1)^T}, (c^{(2)} - c) e^{(2)^T}, \ldots, (c^{(k)} - c) e^{(k)^T}] \in \mathbb{R}^{m \times n}, \text{ and} \tag{9}$$

$$H_m = [a_1 - c, \ldots, a_n - c] = A - c e^T \in \mathbb{R}^{m \times n}, \tag{10}$$

the scatter matrices can be expressed as

$$S_w = H_w H_w^T, \quad S_b = H_b H_b^T, \quad \text{and} \quad S_m = H_m H_m^T. \tag{11}$$

For S_2 to be nonsingular, we can only allow the case $m \leq n$, since S_2 is the product of an $m \times n$ matrix and an $n \times m$ matrix. 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 (6) becomes

$$\beta_i^2 S_1 x_i = \alpha_i^2 S_2 x_i. \tag{12}$$

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

3.2 Extension of $\mathbf{J_1 = \text{trace}(S_2^{-1} S_1)}$ 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. (11) and the definition of H_b given in Eqn. (9), $\text{rank}(S_b) \leq k - 1$. To approximate G that satisfies both

$$\max_G \text{trace}(G^T S_b G) \quad \text{and} \quad \min_G \text{trace}(G^T S_w G), \tag{13}$$

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. (2), 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 1 [9]. The major steps are limited to the complete orthogonal decomposition 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. The steps for this case are summarized in Algorithm DiscGSVD, adapted from [7].

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. (12), 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 $x_i \in \text{null}(S_w) \cap \text{null}(S_b)$, Eqn. (12) 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

$$\text{trace}(G^T S_b G) = \sum g_j^T S_b g_j \quad \text{and} \quad \text{trace}(G^T S_w G) = \sum g_j^T S_w g_j,$$

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 (13). For this reason, we do not include these columns of X in our solution.

Algorithm 1 DiscGSVD

Given a data matrix $A \in \mathbb{R}^{m\times n}$ with k clusters, it computes the columns of the matrix $G \in \mathbb{R}^{m\times(k-1)}$, which preserves the cluster structure in the reduced dimensional space, using $J_1(G) = \mathrm{trace}((G^T S_w G)^{-1} G^T S_b G)$. It also computes 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 (8), respectively. (Using this equivalent but lower dimensional form of H_b reduces complexity.)

2. Compute the complete orthogonal decomposition of $K = \begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix} \in \mathbb{R}^{(k+n)\times m}$,

   which is $P^T K Q = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}$.

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. $Y = G^T A$

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 experiment summarized in Section 4 shows that Algorithm DiscGSVD works very well when S_w is singular, thus extending its applicability beyond that of the original discriminant analysis.

3.3 Extension of $\mathbf{J_1 = \mathrm{trace}(S_2^{-1} S_1)}$ Criteria for Two-class Case

When optimizing $J_1 = \mathrm{trace}(S_w^{-1} S_b)$ for $k = 2$, the scatter matrix S_w takes the form

$$S_w = \Sigma_{j\in N_1}(a_j - c^{(1)})(a_j - c^{(1)})^T + \Sigma_{j\in N_2}(a_j - c^{(2)})(a_j - c^{(2)})^T = n_1\Sigma_1 + n_2\Sigma_2,$$

where Σ_i is the covariance matrix of class i. Hence S_w is the weighted sum of class covariance matrices, and $S_w = H_w H_w^T$, where

$$H_w = [A_1 - c^{(1)}e^{(1)^T}, A_2 - c^{(2)}e^{(2)^T}] \in \mathbb{R}^{m\times n}.$$

Also,

$$S_b = n_1(c^{(1)} - c)(c^{(1)} - c)^T + n_2(c^{(2)} - c)(c^{(2)} - c)^T = \frac{n_1 n_2}{n}(c^{(2)} - c^{(1)})(c^{(2)} - c^{(1)})^T,$$

using the fact that $c = \frac{1}{n}(n_1 c^{(1)} + n_2 c^{(2)})$. So $S_b = H_b H_b^T$, where

$$H_b = \sqrt{\frac{n_1 n_2}{n}}(c^{(2)} - c^{(1)}) \in \mathbb{R}^{m\times 1}.$$

Since $\mathrm{rank}(S_b) = 1$, we have, for S_w nonsingular,

$$S_w^{-1} S_b x_1 = S_w^{-1}\frac{n_1 n_2}{n}(c^{(2)} - c^{(1)})(c^{(2)} - c^{(1)})^T x_1 = \lambda_1 x_1,$$

where $\lambda_1 > 0$. This implies $x_1 = \gamma S_w^{-1}(c^{(2)} - c^{(1)})$ for some scalar γ. Thus the leading eigenvector is a scalar multiple of the difference of class means, normalized by the weighted sum of class covariances. Substituting this expression for x_1 yields

$$\lambda_1 = \frac{n_1 n_2}{n}(c^{(2)} - c^{(1)})^T S_w^{-1}(c^{(2)} - c^{(1)}) = \mathrm{trace}(S_w^{-1} S_b).$$

Since this trace is preserved in the reduced dimension, we get better separation of the two classes if the difference of class means is large relative to the weighted sum of class covariance matrices.

For the matrix pair (H_b^T, H_w^T), the GSVD takes the form

$$uH_b^T X = (\Sigma_b, 0) \quad \text{and} \quad V^T H_w^T X = (\Sigma_w, 0),$$

where $u = \pm 1$, $V \in \mathbb{R}^{n \times n}$ is orthogonal, $X \in \mathbb{R}^{m \times m}$ is nonsingular with column $x_1 = \gamma S_w^{-1}(c^{(2)} - c^{(1)})$,

$$\underset{1 \times t}{\Sigma_b} = (\sigma_b, \underbrace{0, \ldots, 0}_{t-1}) \quad \text{for} \quad 0 < \sigma_b < 1,$$

and $\Sigma_w \in \mathbb{R}^{n \times t}$ has trailing diagonal given by

$$\mathrm{diag}(\sigma_w, \underbrace{1, \ldots, 1}_{t-1}) \quad \text{for} \quad 0 < \sigma_w < 1.$$

σ_b and σ_w are determined by $\sigma_b/\sigma_w = \lambda_1$ and $\sigma_b^2 + \sigma_w^2 = 1$. When S_w is singular and $x_1 \in \mathrm{null}(S_w) - \mathrm{null}(S_b)$, then $\sigma_w = 0$ and $\sigma_b = 1$, so $H_b^T x_1 = \pm 1$ and $H_w^T x_1 = 0$.

In his 1936 paper [2], Fisher originated discriminant analysis by defining a slightly different criterion for the two-class case. For $g \in \mathbb{R}^{m \times 1}$, he used

$$J(g) = \frac{g^T S_b g}{g^T S_w g},$$

where the scatter matrices are given by

$$S_b = (c^{(2)} - c^{(1)})(c^{(2)} - c^{(1)})^T \quad \text{and} \quad S_w = \Sigma_1 + \Sigma_2.$$

If the classes are equally sized, i.e. $n_1 = n_2$, the Fisher criterion produces the same solution as that for J_1.

3.4 Equivalence of $\mathbf{J_1 = \mathrm{trace}(S_2^{-1} S_1)}$ 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 (1), the generalized eigenvalue problem $S_m x_i = \lambda_i S_w x_i$ can be rewritten as $S_b x_i = (\lambda_i - 1) S_w x_i$, where $\lambda_i \geq 1$ for $1 \leq i \leq m$. 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 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 $x_i \in \mathrm{null}(S_w)$, $S_m x_i = S_b x_i$. Hence, we include the same columns in G as we did in the case of $(S_1, S_2) = (S_b, S_w)$.

To show that the solutions are the same in terms of the GSVD, we use the fact that $H_m = H_w + H_b$. From

$$\begin{aligned} S_m &= H_m H_m^T = (H_w + H_b)(H_w + H_b)^T \\ &= H_w H_w^T + H_b H_b^T + H_b H_w^T + H_w H_b^T \\ &= S_m + H_b H_w^T + H_w H_b^T, \end{aligned}$$

we see that $H_b H_w^T + H_w H_b^T = 0$. In fact, each of these products is zero, since

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

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 Σ_b and $\Sigma_w \in \mathbb{R}^{n \times t}$, $\Sigma_b^T \Sigma_b + \Sigma_w^T \Sigma_w = I_t$, and $t = \text{rank} \begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix}$. 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 $t \leq n$. Since t is defined as $\text{rank}(K)$ where $K = \begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix} \in \mathbb{R}^{2n \times m}$, it is clear that $t \leq m$. However, it is not obvious that $t \leq n$ when $m > n$. We now use the property that $H_w H_b^T = 0$ to show that $t \leq n$. For $K^T = (H_b, H_w) \in \mathbb{R}^{m \times 2n}$, we have $\text{rank}(K^T) + \dim(\text{null}(K^T)) = 2n$, or $\dim(\text{null}(K^T)) = 2n - t$. 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

$$\dim(\text{null}(H_b, H_w)) \geq \dim(\text{null}(H_b)) + \dim(\text{null}(H_w)).$$

Here $H_w H_b^T = 0$ 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

$$\dim(\text{null}(H_b, H_w)) \geq n - \text{rank}(H_b) + \text{rank}(H_b^T) = n.$$

Having shown that $t \leq n$, we proceed with the GSVD derivation.

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 $\hat{U} = U(\Sigma_b + U^T V \Sigma_w, \hat{U}_2)$ is orthogonal and $\Sigma_m = \begin{pmatrix} I_t \\ 0 \end{pmatrix}$. 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 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 in the m-dimensional space,

$$\mathrm{trace}((S_w)^{-1}S_m) = \mathrm{trace}((S_w)^{-1}(S_w + S_b)) = m + \mathrm{trace}((S_w)^{-1}S_b).$$

and in the l-dimensional space,

$$\mathrm{trace}((S_w^Y)^{-1}S_m^Y) = \mathrm{trace}((S_w^Y)^{-1}(S_w^Y + S_b^Y)) = l + \mathrm{trace}((S_w^Y)^{-1}S_b^Y).$$

This confirms that the solutions are the same for both cases. By subtracting these equations for any $l \geq k-1$, we get

$$\mathrm{trace}((S_w^Y)^{-1}S_m^Y) + (m-l) = \mathrm{trace}(S_w^{-1}S_m).$$

In other words, the $m-l$ eigenvectors corresponding to zero eigenvalues for $S_w^{-1}S_b$ correspond to the eigenvalues of one for $S_w^{-1}S_m$. 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. (7), $\mathrm{trace}(S_w^{-1}S_m)$ will be preserved if we include all $\mathrm{rank}(S_m) = m$ eigenvectors of $S_w^{-1}S_m$.

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 (1) 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 $x_i \in \mathrm{null}(S_w)$. For that x_i, $S_m x_i = S_b x_i$, so the same reasoning applies regarding the effect on $\mathrm{trace}(G^T S_m G)$ and $\mathrm{trace}(G^T S_w G)$. 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).

3.5 $\mathbf{J_2 = \ln(\det(S_2^{-1}S_1))}$ Criteria

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 [3] that

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

and setting this to zero yields

$$S_2^{-1}S_1 G = G(G^T S_2 G)^{-1}(G^T S_1 G) = G((S_2^Y)^{-1}S_1^Y). \tag{14}$$

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

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

where $Z \in \mathbb{R}^{l \times l}$ is nonsingular. Hence $(S_2^Y)^{-1}S_1^Y = Z\Lambda Z^{-1}$ and Eqn. (14) becomes $S_2^{-1}S_1 GZ = GZ\Lambda$, 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 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 $x_i \in \text{null}(S_b)$. 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 $l = \text{rank}(S_b)$ generalized right singular vectors that correspond to the largest generalized singular values. However, these leftmost $\text{rank}(S_b)$ vectors may include an $x_i \in \text{null}(S_w) - \text{null}(S_b)$. 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 $\text{rank}(S_b)$ generalized right singular vectors, just as we did for trace optimization in J_1.

3.6 Efficient Solution for Maximizing $\mathbf{J_3} = \text{trace}(\mathbf{S_b})$

Simpler criteria for preserving cluster structure, such as $\min \text{trace}(G^T S_w G)$ and $\max \text{trace}(G^T S_b G)$, involve only one of the scatter matrices. A straightforward minimization of $\text{trace}(G^T S_w G)$ 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 $\text{trace}(G^T S_b G)$ produces an equivalent solution to the CentroidQR method, which is introduced and shown to give promising reduced dimensional classification results in [8, 7], and summarized in Algorithm 2.

Let $J_3(G) = \text{trace}(G^T S_b G)$. If we let G be any full rank matrix, then essentially there is no bound and maximizing is also meaningless. Now, let us restrict the solution to the case when G has orthonormal columns. Suppose the eigenvalues of the symmetric positive semidefinite matrix S_b are sorted as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. Then since $\text{rank}(S_b) \leq k-1$, $\text{trace}(S_b) = \lambda_1 + \lambda_2 + \cdots + \lambda_{k-1}$. For $G^T G = I$,

$$\text{trace}(G^T S_b G) \leq \lambda_1 + \cdots + \lambda_{k-1} = \text{trace}(S_b).$$

The symmetry of S_b guarantees the existence of an orthonormal set of eigenvectors. If the columns of G are the $k-1$ eigenvectors that correspond to the $k-1$ largest eigenvalues,

$$\text{trace}(G^T S_b G) = \text{trace}(G^T G \Lambda_{k-1}) = \text{trace}(S_b),$$

where $\Lambda_{k-1} = \text{diag}(\lambda_1 \ldots \lambda_{k-1})$. Thus the upper bound is achieved for this G.

If the SVD of H_b is $H_b = U \Sigma V^T$, then $S_b = U \Sigma \Sigma^T U^T$. This means that G consists of the first $k-1$ columns of U, or $G = U_{k-1}$. Since increasing the dimension of G by including more eigenvectors does not change the trace value, we obtain the same trace value with $G = U_k$.

Now we show that this solution is equivalent to the solution of the CentroidQR method, which does not involve the computation of eigenvectors. Let $C = \left(c^{(1)} \cdots c^{(k)}\right)$ be the centroid matrix whose columns are the centroids of the k clusters. Assuming the centroids are linearly independent, C has the reduced QR decomposition [4] $C = Q_k R$, where $Q_k \in \mathbb{R}^{m \times k}$ has orthonormal columns and R is nonsingular. Suppose x is an eigenvector of S_b. Then

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

This means $x \in \text{span}\{c^{(i)} - c | 1 \leq i \leq k\}$ and hence $x \in \text{span}\{c^{(i)} | 1 \leq i \leq k\}$. Accordingly, $\text{range}(U_k) = \text{range}(C) = \text{range}(Q_k)$ which implies that $Q_k = U_k W$ for some orthogonal matrix W. Since J_3 is invariant under any orthogonal transformation in the same space, Q_k plays the same role as U_k, and therefore, as U_{k-1}. 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 can obtain a solution that maximizes $\text{trace}(G^T S_b G)$ without computing the eigenvectors.

4 Numerical Results

Having shown the equivalence of the J_1 criteria for various (S_1, S_2), as well as the equivalence of J_1 to J_2, we conclude that the J_1 criterion with $(S_1, S_2) = (S_b, S_w)$ should be used for efficiency. In this section, we extract

Algorithm 2 CentroidQR

Given a data matrix $A \in \mathbb{R}^{m \times n}$ with k clusters, it computes a k-dimensional representation Y of A.

1. Compute the centroid $c^{(i)}$ of the ith cluster, $1 \leq i \leq k$.
2. Set $C = \begin{pmatrix} c^{(1)} & c^{(2)} & \cdots & c^{(k)} \end{pmatrix}$.
3. Compute the reduced QR decomposition of C, which is $C = Q_k R$.
4. Solve $\min_Y \|Q_k Y - A\|_F$ (in fact, $Y = Q_k^T A$).

from [7] a brief summary of an experiment for the singular case. We compare trace values and misclassification rates for the full document set with those for the reduced dimensional representations produced by both DiscGSVD and CentroidQR algorithms, using the J_1 criterion with $(S_1, S_2) = (S_b, S_w)$ and the J_3 criterion with $G^T G = I$, respectively.

The documents consist of five categories of abstracts from the MEDLINE [1] database. Each category has 40 documents. There are 7519 terms after preprocessing with stopping and stemming algorithms [6]. For this 7519×200 term-document matrix, S_w is singular and the original discriminant analysis breaks down. However, our improved DiscGSVD method circumvents this singularity problem.

Table 1: 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 2: Traces and Misclassification Rate with L_2 norm similarity

Method		Full	CentroidQR	DiscGSVD
Dim		7519×200	5×200	4×200
trace values	$\mathrm{trace}(S_w)$	73048	4210	0.05
	$\mathrm{trace}(S_b)$	<u>6229</u>	<u>6229</u>	4
	$\frac{\mathrm{trace}(S_b)}{\mathrm{trace}(S_w)}$	0.09	1.5	<u>80</u>
misclassification rate in %	centroid	5	5	1
	1nn	40	3	1

Algorithm DiscGSVD dramatically reduces the dimension 7519 to 4, or one less than the number of clusters. The CentroidQR method reduces the dimension to 5. Table 2 shows classification results using the L_2 norm similarity measure. DiscGSVD produces the lowest misclassification rate using both a centroid-based and the nearest neighbor [5] classification methods. Because the J_1 criterion is not defined in this case, we compute the ratio $\frac{\mathrm{trace}(S_b)}{\mathrm{trace}(S_w)}$ as a rough optimality measure. We observe that the ratio is strikingly higher for the DiscGSVD reduction than for the other methods. These experimental results confirm that the DiscGSVD algorithm effectively extends the applicability of the J_1 criterion to cases that the original discriminant analysis cannot handle. In addition, the CentroidQR algorithm preserves $\mathrm{trace}(S_b)$ from the full dimension without the expense of computing eigenvectors. Taken together, the results

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

for these two methods demonstrate the potential for dramatic and efficient dimension reduction without compromising cluster structure.

5 Future Work

In practice, a disadvantage of methods that involve the GSVD is that its computation is costly. By taking advantage of the fact that the solution only requires a small portion of the generalized right singular vectors, and the relationship between the GSVD and the generalized QR decomposition, we hope to produce a more efficient algorithm for optimizing J_1, as we did in Section 3.6 for J_3 using the QR decomposition.

As we observed in Section 3.3 for the two-class case, the J_1 criterion should work well when the majority of discrimination information is captured in the difference of means rather than in differences in class covariance. Additional experiments are planned to determine the situations for which J_1 is effective for text data, with the goal of formulating improved criteria for cluster-preserving dimension reduction.

References

[1] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*, second edition, John Wiley and Sons, Inc., 2001.

[2] R.A. Fisher. The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7 Part II:179-188, 1936.

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

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

[5] A.K. Jain, and R.C. Dubes. *Algorithms for clustering data*, Prentice Hall, 1988.

[6] G. Kowalski. *Information Retrieval System: Theory and Implementation*, Kluwer Academic Publishers, 1997.

[7] H. Park, M. Jeon and P. Howland. Cluster structure preserving dimension reduction based on the generalized singular value decomposition, *SIAM Journal on Matrix Analysis and Applications*, accepted for publication.

[8] H. Park, M. Jeon, and J.B.Rosen. Lower dimensional representation of text data based on centroids and least squares, *BIT*, submitted to *BIT*, September 2001.

[9] C.C. Paige and M.A. Saunders. Towards a generalized singular value decomposition, *SIAM J. Numer. Anal.*, 18:398-405, 1981.

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

[11] C.F. Van Loan. Generalizing the singular value decomposition, *SIAM J. Numer. Anal.*, 13:76-83, 1976.