SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 165–179

STRUCTURE PRESERVING DIMENSION REDUCTION FOR CLUSTERED TEXT DATA BASED ON THE GENERALIZED SINGULAR VALUE DECOMPOSITION*

PEG HOWLAND†, MOONGU JEON‡, AND HAESUN PARK†

Abstract. In today's vector space information retrieval systems, dimension reduction is imperative for efficiently manipulating the massive quantity of data. To be useful, this lower-dimensional representation must be a good approximation of the full document set. To that end, we adapt and extend the discriminant analysis projection used in pattern recognition. This projection preserves cluster structure by maximizing the scatter between clusters while minimizing the scatter within clusters. A common limitation of trace optimization in discriminant analysis is that one of the scatter matrices must be nonsingular, which restricts its application to document sets in which the number of terms does not exceed the number of documents. We show that by using the generalized singular value decomposition (GSVD), we can achieve the same goal regardless of the relative dimensions of the term-document matrix. In addition, applying the GSVD allows us to avoid the explicit formation of the scatter matrices in favor of working directly with the data matrix, thus improving the numerical properties of the approach. Finally, we present experimental results that confirm the effectiveness of our approach.

Key words. dimension reduction, discriminant analysis, pattern recognition, trace optimization, scatter matrix, generalized eigenvalue problem, generalized singular value decomposition, text classification

AMS subject classifications. 15A09, 68T10, 62H30, 65F15, 15A18

PII. S0895479801393666

1. Introduction.

The vector space–based information retrieval system, originated by Salton [13, 14], represents documents as vectors in a vector space. The document set comprises an $m \times n$ term-document matrix $A = (a_{ij})$, in which each column represents a document and each entry a_{ij} represents the weighted frequency of term i in document j. A major benefit of this representation is that the algebraic structure of the vector space can be exploited [1]. Modern document sets are huge [3], so we need to find a lower-dimensional representation of the data. To achieve higher efficiency in manipulating the data, it is often necessary to reduce the dimension severely. Since this may result in loss of information, we seek a representation in the lower-dimensional space that best approximates the document collection in the full space [8, 12].

The specific method we present in this paper is based on the discriminant analysis projection used in pattern recognition [4, 15]. Its goal is to find the mapping that transforms each column of A into a column in the lower-dimensional space, while preserving the cluster structure of the full data matrix. This is accomplished by

*Received by the editors August 13, 2001; accepted for publication (in revised form) by L. Eldén October 22, 2002; published electronically May 15, 2003. This research was supported in part by National Science Foundation (NSF) grant CCR-9901992. 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 NSF.

http://www.siam.org/journals/simax/25-1/39366.html

†Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 (howland@cs.umn.edu, hpark@cs.umn.edu). A part of this work was carried out while the third author was visiting the Korea Institute for Advanced Study, Seoul, Korea, for her sabbatical leave, from September 2001 to July 2002.

‡Department of Computer Science, University of California, Santa Barbara, CA 93106 (jeon@cs.ucsb.edu).

forming scatter matrices from A, the traces of which provide measures of the quality of the cluster relationship. After defining the optimization criterion in terms of these scatter matrices, the problem can be expressed as a generalized eigenvalue problem.

As we explain in the next section, the current discriminant analysis approach can be applied only in the case where $m \leq n$, i.e., when the number of terms does not exceed the number of documents. By recasting the generalized eigenvalue problem in terms of a related generalized singular value problem, we circumvent this restriction on the relative dimensions of A, thus extending the applicability to any data matrix. At the same time, we improve the numerical properties of the approach by working with the data matrix directly rather than forming the scatter matrices explicitly. Our algorithm follows the generalized singular value decomposition (GSVD) [2, 5, 16] as formulated by Paige and Saunders [11]. For a data matrix with k clusters, we can limit our computation to the *generalized right singular vectors* that correspond to the $k-1$ largest generalized singular values. In this way, our algorithm remains computationally simple while achieving its goal of preserving cluster structure. Experimental results demonstrating its effectiveness are described in section 5 of the paper.

2. Dimension reduction based on discriminant analysis. Given a term-document matrix $A \in \mathbb{R}^{m\times n}$, the general problem we consider is to find a linear transformation $G^T \in \mathbb{R}^{l\times m}$ that maps each column a_i, $1 \leq i \leq n$, of A in the m-dimensional space to a column y_i in the l-dimensional space:

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

Rather than looking for the mapping that achieves this explicitly, one may rephrase this as an approximation problem where the given matrix A is decomposed into two matrices B and Y as

$$A \approx BY, \tag{2}$$

where both $B \in \mathbb{R}^{m\times l}$ with $\text{rank}(B) = l$ and $Y \in \mathbb{R}^{l\times n}$ with $\text{rank}(Y) = l$ are to be found. Note that what we need ultimately is the lower-dimensional representation Y of the matrix A, where B and Y are both unknown. In [8, 12], methods that determine the matrix B have been presented. In those methods, after B is determined, the matrix Y is computed, for example, by solving the least squares problem [2]

$$\min_{B,Y} \|BY - A\|_F, \tag{3}$$

where B and A are given. The method we present here computes the matrix G^T directly from A without reformulating the problem as a matrix approximation problem as in (2).

Now our goal is to find a linear transformation such that the cluster structure existing in the full-dimensional space is preserved in the reduced-dimensional space, assuming that the given data are already clustered. For this purpose, first we need to formulate a measure of cluster quality. To have high cluster quality, a specific clustering result must have a tight within-cluster relationship while the between-cluster relationship has to be remote. To quantify this, in discriminant analysis [4, 15], within-cluster, between-cluster, and mixture scatter matrices are defined. For simplicity of discussion, we will assume that the given data matrix $A \in \mathbb{R}^{m\times n}$ is partitioned into k clusters as

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

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

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

and the global centroid is

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

Then the within-cluster scatter matrix S_w is defined as

$$S_w = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})(a_j - c^{(i)})^T,$$

and the between-cluster scatter matrix S_b is defined as

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

Finally, the mixture scatter matrix is defined as

$$S_m = \sum_{j=1}^{n} (a_j - c)(a_j - c)^T.$$

It is easy to show [7] that the scatter matrices have the relationship

$$S_m = S_w + S_b. \tag{4}$$

Writing $a_j - c = a_j - c^{(i)} + c^{(i)} - c$ for $j \in N_i$, we have

$$S_m = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)} + c^{(i)} - c)(a_j - c^{(i)} + c^{(i)} - c)^T \tag{5}$$

$$= \sum_{i=1}^{k} \sum_{j \in N_i} [(a_j - c^{(i)})(a_j - c^{(i)})^T + (c^{(i)} - c)(c^{(i)} - c)^T] \tag{6}$$

$$+ \sum_{i=1}^{k} \sum_{j \in N_i} [(a_j - c^{(i)})(c^{(i)} - c)^T + (c^{(i)} - c)(a_j - c^{(i)})^T]. \tag{7}$$

This gives the relation (4), since each inner sum in (7) is zero.

Defining the matrices,

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

$$H_b = [\sqrt{n_1}(c^{(1)} - c), \sqrt{n_2}(c^{(2)} - c), \ldots, \sqrt{n_k}(c^{(k)} - c)] \in \mathbb{R}^{m \times k}, \tag{9}$$

and

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

the scatter matrices can be expressed as

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

Note that another way to define H_b is

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

but using the lower-dimensional form in (9) reduces the storage requirements and computational complexity of our algorithm.

Now, trace(S_w), which is

$$\text{trace}(S_w) = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})^T (a_j - c^{(i)}) = \sum_{i=1}^{k} \sum_{j \in N_i} \|a_j - c^{(i)}\|_2^2, \tag{12}$$

provides a measure of the closeness of the columns within the clusters over all k clusters, and trace(S_b), which is

$$\text{trace}(S_b) = \sum_{i=1}^{k} \sum_{j \in N_i} (c^{(i)} - c)^T (c^{(i)} - c) = \sum_{i=1}^{k} \sum_{j \in N_i} \|c^{(i)} - c\|_2^2, \tag{13}$$

provides a measure of the distance between clusters. When items within each cluster are located tightly around their own cluster centroid, then trace(S_w) will have a small value. On the other hand, when the between-cluster relationship is remote, and hence the centroids of the clusters are remote, trace(S_b) will have a large value. Using the values trace(S_w), trace(S_b), and relationship (4), the cluster quality can be measured. In general, when trace(S_b) is large while trace(S_w) is small, or trace(S_m) is large while trace(S_w) is small, we expect the clusters of different classes to be well separated and the items within each cluster to be tightly related, and therefore the cluster quality will be high. There are several measures of cluster quality which involve the three scatter matrices [4, 15], including

$$J_1 = \text{trace}(S_w^{-1} S_b) \tag{14}$$

and

$$J_2 = \text{trace}(S_w^{-1} S_m). \tag{15}$$

Note that both of the above criteria require S_w to be nonsingular or, equivalently, H_w to have full rank. For more measures of cluster quality, their relationships, and their extension to document data, see [6].

In the lower-dimensional space obtained from the linear transformation G^T, the within-cluster, between-cluster, and mixture scatter matrices become

$$S_w^Y = \sum_{i=1}^{k} \sum_{j \in N_i} (G^T a_j - G^T c^{(i)})(G^T a_j - G^T c^{(i)})^T = G^T S_w G,$$

$$S_b^Y = \sum_{i=1}^{k} \sum_{j \in N_i} (G^T c^{(i)} - G^T c)(G^T c^{(i)} - G^T c)^T = G^T S_b G,$$

$$S_m^Y = \sum_{j=1}^{n} (G^T a_j - G^T c)(G^T a_j - G^T c)^T = G^T S_m G,$$

where the superscript Y denotes values in the l-dimensional space. Given k clusters in the full dimension, the linear transformation G^T that best preserves this cluster structure in the reduced dimension would maximize $\text{trace}(S_b^Y)$ and minimize $\text{trace}(S_w^Y)$. We can approximate this simultaneous optimization using measure (14) or (15) by looking for the matrix G that maximizes

$$J_1(G) = \text{trace}((G^T S_w G)^{-1}(G^T S_b G))$$

or

$$J_2(G) = \text{trace}((G^T S_w G)^{-1}(G^T S_m G)).$$

For computational reasons, we will focus our discussion on the criterion of maximizing J_1. Although J_1 is a less obvious choice than the quotient

$$\text{trace}(G^T S_b G)/\text{trace}(G^T S_w G),$$

it is formulated to be invariant under nonsingular linear transformations, a property that will prove useful below.

When $S_w = H_w H_w^T$ is assumed to be nonsingular, it is symmetric positive definite. According to results from the symmetric-definite generalized eigenvalue problem [5], there exists a nonsingular matrix $X \in \mathbb{R}^{m\times m}$ such that

$$X^T S_b X = \Lambda = \text{diag}(\lambda_1 \dots \lambda_m) \quad \text{and} \quad X^T S_w X = I_m.$$

Letting x_i denote the ith column of X, we have

$$S_b x_i = \lambda_i S_w x_i, \tag{16}$$

which means that λ_i and x_i are an eigenvalue-eigenvector pair of $S_w^{-1} S_b$, and

$$\text{trace}(S_w^{-1} S_b) = \lambda_1 + \cdots + \lambda_m.$$

Expressing (16) in terms of H_b and H_w and premultiplying by x_i^T, we see that

$$\|H_b^T x_i\|_2^2 = \lambda_i \|H_w^T x_i\|_2^2. \tag{17}$$

Hence $\lambda_i \geq 0$ for $1 \leq i \leq m$.

The definition of H_b in (9) implies that $\text{rank}(H_b) \leq k-1$. Accordingly, $\text{rank}(S_b) \leq k-1$, and only the largest $k-1$ λ_i's can be nonzero. In addition, by using a permutation matrix to order Λ (and likewise X), we can assume that $\lambda_1 \geq \cdots \geq \lambda_{k-1} \geq \lambda_k = \cdots = \lambda_m = 0$.

We have

$$\begin{aligned} J_1(G) &= \text{trace}((S_w^Y)^{-1} S_b^Y) \\ &= \text{trace}((G^T X^{-T} X^{-1} G)^{-1} G^T X^{-T} \Lambda X^{-1} G) \\ &= \text{trace}((\tilde{G}^T \tilde{G})^{-1} \tilde{G}^T \Lambda \tilde{G}), \end{aligned}$$

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. Hence

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

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

$$\begin{aligned}\max_G J_1(G) &= \max_{Q^TQ=I} \operatorname{trace}(Q^T\Lambda Q)\\ &\le \lambda_1 + \cdots + \lambda_{k-1} = \operatorname{trace}(S_w^{-1}S_b).\end{aligned}$$

When $l \ge k-1$, 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 in the sense that $J_1(G) = J_1(GW)$ for any nonsingular matrix $W \in \mathbb{R}^{l\times l}$ since

$$\begin{aligned}J_1(GW) &= \operatorname{trace}((W^TG^TS_wGW)^{-1}(W^TG^TS_bGW))\\ &= \operatorname{trace}(W^{-1}(G^TS_wG)^{-1}W^{-T}W^T(G^TS_bG)W)\\ &= \operatorname{trace}((G^TS_wG)^{-1}(G^TS_bG)WW^{-1}) = J_1(G).\end{aligned}$$

Hence, the maximum $J_1(G)$ is also achieved for

$$G = X \begin{pmatrix} I_l \\ 0 \end{pmatrix}.$$

This means that

$$\operatorname{trace}((S_w^Y)^{-1}S_b^Y) = \operatorname{trace}(S_w^{-1}S_b)$$

whenever $G \in \mathbb{R}^{m\times l}$ consists of l eigenvectors of $S_w^{-1}S_b$ corresponding to the l largest eigenvalues. Therefore, if we choose $l = k-1$, dimension reduction results in no loss of cluster quality as measured by J_1.

Now, a limitation of the criterion $J_1(G)$ in many applications, including text processing in information retrieval, is that the matrix S_w must be nonsingular. For S_w to be nonsingular, we can allow only the case $m \le n$, since S_w is the product of an $m \times n$ matrix, H_w, and an $n \times m$ matrix, H_w^T. In other words, the number of terms cannot exceed the number of documents, which is a severe restriction. We seek a solution which does not impose this restriction, and which can be found without explicitly forming S_b and S_w from H_b and H_w, respectively. Toward that end, we use (17) to express λ_i as α_i^2/β_i^2, and the problem (16) becomes

$$\beta_i^2 H_b H_b^T x_i = \alpha_i^2 H_w H_w^T x_i. \tag{18}$$

(λ_i will be infinite when $\beta_i = 0$, as we discuss later.) This has the form of a problem that can be solved using the GSVD [5, 11, 16], as described in the next section.

3. GSVD. The following theorem introduces the GSVD as was originally defined by Van Loan [16].

THEOREM 1. *Suppose two matrices $K_A \in \mathbb{R}^{m\times n}$ with $m \ge n$ 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*

$$U^TK_AX = \operatorname{diag}(\alpha_1, \ldots, \alpha_n) \quad \text{and} \quad V^TK_BX = \operatorname{diag}(\beta_1, \ldots, \beta_q),$$

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

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 [11] 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 $K_A \in \mathbb{R}^{m\times n}$ and $K_B \in \mathbb{R}^{p\times n}$ are given. Then for*

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

there exist orthogonal matrices

$$U \in \mathbb{R}^{m\times m}, \quad V \in \mathbb{R}^{p\times p}, \quad W \in \mathbb{R}^{t\times t}, \quad \textit{and} \quad Q \in \mathbb{R}^{n\times n}$$

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

$$\underset{m\times t}{\Sigma_A} = \begin{pmatrix} I_A & & \\ & D_A & \\ & & 0_A \end{pmatrix}, \quad \underset{p\times t}{\Sigma_B} = \begin{pmatrix} O_B & & \\ & D_B & \\ & & I_B \end{pmatrix}, \tag{19}$$

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 the values of r and s depend on the data,

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

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

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

satisfy

$$1 > \alpha_{r+1} \geq \cdots \geq \alpha_{r+s} > 0, \quad 0 < \beta_{r+1} \leq \cdots \leq \beta_{r+s} < 1, \tag{20}$$

and

$$\alpha_i^2 + \beta_i^2 = 1 \quad \textit{for } i = r+1, \ldots, r+s.$$

Paige and Saunders gave a constructive proof of Theorem 2, which starts with the complete orthogonal decomposition [5, 2, 10] of K, or

$$P^T K Q = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}, \tag{21}$$

where P and Q are orthogonal and R is nonsingular with the same rank as K. The construction proceeds by exploiting the SVDs of submatrices of P. Partitioning P as

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}, \quad \text{where} \quad P_{11} \in \mathbb{R}^{m\times t} \quad \text{and} \quad P_{21} \in \mathbb{R}^{p\times t},$$

implies $\|P_{11}\|_2 \leq 1$. This means that the singular values of P_{11} do not exceed one, so its SVD can be written as $U^T P_{11} W = \Sigma_A$, where $U \in \mathbb{R}^{m\times m}$ and $W \in \mathbb{R}^{t\times t}$ are orthogonal and Σ_A has the form in (19). Next $P_{21}W$ is decomposed as $P_{21}W = VL$, where $V \in \mathbb{R}^{p\times p}$ is orthogonal and $L = (l_{ij}) \in \mathbb{R}^{p\times t}$ is lower triangular with $l_{ij} = 0$ if $p - i > t - j$ and $l_{ij} \geq 0$ if $p - i = t - j$. This triangularization can be accomplished in the same way as QR decomposition except that columns are annihilated above the diagonal $p - i = t - j$, working from right to left. Then the matrix

$$\begin{pmatrix} \Sigma_A \\ L \end{pmatrix}$$

has orthonormal columns, which implies that $L = \Sigma_B$. These results can be combined with (21) to obtain

$$\begin{pmatrix} K_A \\ K_B \end{pmatrix} Q = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} P_{11}R & 0 \\ P_{21}R & 0 \end{pmatrix} = \begin{pmatrix} U\Sigma_A W^T R & 0 \\ V\Sigma_B W^T R & 0 \end{pmatrix},$$

which completes the proof. In [11], this form of GSVD is related to that of Van Loan by

$$U^T K_A X = (\Sigma_A, 0) \quad \text{and} \quad V^T K_B X = (\Sigma_B, 0), \tag{22}$$

where

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

From the form in (22) 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

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

and

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

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

$$\beta_i^2 K_A^T K_A x_i = \alpha_i^2 K_B^T K_B x_i, \tag{23}$$

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 (23) is satisfied for arbitrary values of α_i and β_i when $t + 1 \leq i \leq n$. Therefore, the columns of X are the generalized right singular vectors for the matrix pair K_A and 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. To determine the number of

generalized singular values of each type, we write explicit expressions for the values of r and s. From (22) and (19), we see that

$$\operatorname{rank}(K_A) = r + s \quad \text{and} \quad \operatorname{rank}(K_B) = t - r.$$

Hence, the number of infinite generalized singular values is

$$r = \operatorname{rank}\begin{pmatrix} K_A \\ K_B \end{pmatrix} - \operatorname{rank}(K_B)$$

and the number of finite and nonzero generalized singular values is

$$s = \operatorname{rank}(K_A) + \operatorname{rank}(K_B) - \operatorname{rank}\begin{pmatrix} K_A \\ K_B \end{pmatrix}.$$

4. Application of the GSVD to dimension reduction. Recall that for the $m \times n$ term-document matrix A, when $m \leq n$ and the scatter matrix S_w is nonsingular, a criterion such as maximization of J_1 can be applied. However, one drawback of this criterion is that both $S_w = H_w H_w^T$ and $S_b = H_b H_b^T$ must be explicitly formed. Forming these cross-product matrices can cause a loss of information [5, p. 239, Example 5.3.2], but by using the GSVD, which works directly with H_w and H_b, we can avoid a potential numerical problem.

Applying the GSVD to the nonsingular case, we include in G those 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 (20), 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 term-document matrix A. We then solve for a very limited portion of the GSVD of the matrix pair H_b^T and H_w^T. This solution is accomplished by following the construction in the proof of Theorem 2. The major steps are limited to the complete orthogonal decomposition of $K = (H_b, H_w)^T$, which produces orthogonal matrices P and Q and a nonsingular matrix R, followed by the SVD of a leading principal submatrix of P. The steps are summarized in Algorithm LDA/GSVD, where LDA stands for linear discriminant analysis.

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 (18), 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 in terms of the simultaneous optimization

$$\max_G \operatorname{trace}(G^T S_b G) \quad \text{and} \quad \min_G \operatorname{trace}(G^T S_w G) \tag{24}$$

that criterion J_1 is approximating.

When $x_i \in \operatorname{null}(S_w) \cap \operatorname{null}(S_b)$, (18) is satisfied for arbitrary values of α_i and β_i. As explained in section 3, this will be the case for the rightmost $m-t$ columns of X. To determine whether these columns should be included in G, consider

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

where g_j represents a column of G. Adding the column x_i to G has no effect on these traces, since $x_i^T S_w x_i = 0$ and $x_i^T S_b x_i = 0$, and therefore does not contribute to

Algorithm 1 LDA/GSVD.

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, and it also computes the $(k-1)$-dimensional representation Y of A.

1. Compute $H_b \in \mathbb{R}^{m\times k}$ and $H_w \in \mathbb{R}^{m\times n}$ from A according to (9) and (8), respectively.
2. Compute the complete orthogonal decomposition of $K = (H_b, H_w)^T \in \mathbb{R}^{(k+n)\times m}$, which is
$$P^T K Q = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}.$$
3. Let $t = \text{rank}(K)$.
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. $Y = G^T A$.

either maximization or minimization in (24). For this reason, we do not include these columns of X in our solution.

When $x_i \in \text{null}(S_w) - \text{null}(S_b)$, then $\beta_i = 0$. As discussed in section 3, 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 $\text{trace}(G^T S_b G)$ while leaving $\text{trace}(G^T S_w G)$ unchanged. We conclude that, even when S_w is singular, the rule regarding which columns of X to include in G should remain the same as for the nonsingular case. Our experiments show that Algorithm LDA/GSVD works very well when S_w is singular, thus extending its applicability beyond that of the original discriminant analysis.

In terms of the matrix pair H_b^T and H_w^T, the columns of X correspond to the generalized singular values as follows. The first
$$r = \text{rank}\begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix} - \text{rank}(H_w^T)$$
columns correspond to infinite values and the next
$$s = \text{rank}(H_b^T) + \text{rank}(H_w^T) - \text{rank}\begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix}$$
columns correspond to finite and nonzero values. The following
$$t - r - s = \text{rank}\begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix} - \text{rank}(H_b^T)$$
columns correspond to zero values and the last
$$m - t = m - \text{rank}\begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix}$$

Algorithm 2 Centroid.

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, $1 \le i \le k$.
2. Set $C = \begin{pmatrix} c^{(1)} & c^{(2)} & \cdots & c^{(k)} \end{pmatrix}$.
3. Solve $\min_Y \|CY - A\|_F$.

Algorithm 3 Orthogonal Centroid.

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, $1 \le i \le k$.
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, $Y = Q_k^T A$).

columns correspond to the arbitrary values. If S_w is nonsingular, both $r = 0$ and $m-t = 0$, so $s = \text{rank}(H_b^T)$ generalized singular values are finite and nonzero, and the rest are zero. In either case, G should be comprised of the leftmost $r + s = \text{rank}(H_b^T)$ columns of X.

Assuming the centroids are linearly independent, we see from (9) that $\text{rank}(H_b)$ is $k - 1$, so Algorithm LDA/GSVD includes the minimum number of columns in G that are necessary to preserve the cluster structure after dimension reduction. If $\text{rank}(H_b) < k - 1$, then including extra columns in G (some which correspond to the $t - r - s$ zero generalized singular values and, possibly, some which correspond to the arbitrary generalized singular values) will have approximately no effect on cluster preservation.

5. Experimental results. We compare classification results in the full-dimensional space with those in the reduced-dimensional space using Algorithm LDA/GSVD and two other dimension reduction algorithms we have developed, namely, Algorithms Centroid and Orthogonal Centroid [8, 12]. The latter two algorithms assume that the centroids are linearly independent, an assumption for which we have encountered no counterexample in practice. As outlined in Algorithms 2 and 3, centroid and orthogonal centroid solve the same least squares problem (3) for different choices of B. The centroid method chooses the k cluster centroids as the columns of B, whereas orthogonal centroid chooses an orthonormal basis for the cluster centroids.

We employ both a centroid-based classification method and a nearest neighbor classification method [15], which are presented in Algorithms 4 and 5. For the full data matrix A, we apply the classification method with each column of A as the vector q and report the percentage that are misclassified. Likewise, for each dimension reduction method, we apply the classification method to the lower-dimensional representation Y of A. In addition, the quality of classification is assessed by examining traces of the within-class scatter matrix S_w and the between-class scatter matrix S_b.

Two different data types are used to verify the effectiveness of LDA/GSVD. In the first data type, the column dimension of the term-document matrix is higher than the row dimension. This can be dealt with by using the original J_1 criterion, assuming that S_w is nonsingular. In the second data type, the row dimension is higher than the column dimension, so S_w is singular. This means that neither criterion J_1 nor

Algorithm 4 Centroid-Based Classification.

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

- Find the index j such that $sim(q, c^{(i)})$, $1 \leq i \leq k$, 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 5 k Nearest Neighbor (knn) Classification.

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

1. From the similarity measure $sim(q, a_j)$ for $1 \leq j \leq n$, find the k^* nearest neighbors of q. (We use k^* to distinguish the algorithm parameter from the number of clusters.)
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.

J_2 can be applied, but the dimension can be reduced very effectively using our new LDA/GSVD algorithm.

For the first data type, in Test I we use clustered data that are artificially generated by an algorithm adapted from [7, Appendix H]. Table 1 shows the dimensions of the term-document matrix and classification results using the L_2 norm similarity measure. The data consist of 2000 150-dimensional documents with seven clusters. Algorithm LDA/GSVD reduces the dimension from 150 to $k - 1 = 6$, where k is the number of classes. The other methods reduce it to $k = 7$. In Table 1, we also present the results obtained by using the LDA/GSVD algorithm to reduce the dimension to $k - 2 = 5$ and $k = 7$, which are one less than and one greater than the theoretical optimum of $k - 1$, respectively. The results confirm that the theoretical optimum does indeed maximize trace$((S_w^Y)^{-1} S_b^Y)$, and that its value is preserved exactly from the full dimension. In addition, using LDA/GSVD to reduce the dimension to $k - 1$ results in the lowest misclassification rates for both centroid-based and nearest neighbor methods. All three dimension reduction methods produce classification results that are, with one exception, at least as good as the results from the full space. This is remarkable in light of the fact that the row dimension was reduced from 150 to at most 7.

As mentioned in section 2, in a higher quality cluster structure, we will have a smaller value for trace(S_w) and a larger value for trace(S_b). With this in mind, the ratio trace(S_b)/trace(S_w) is another measure of how well trace$(G^T S_b G)$ is maximized while trace$(G^T S_w G)$ is minimized in the reduced space. We observe in Table 1 that the ratio produced by each of the three dimension reduction methods is greater than that of the full-dimensional data. This may explain why, in general, our dimension reduction methods give better classification results than those produced in the full-dimensional space.

TABLE 1
Test I: *Traces and misclassification rates (in %) with L_2 norm similarity.*

Method	Full	Orthogonal centroid	Centroid	LDA/GSVD		
Dim	150×2000	7×2000	7×2000	5×2000	6×2000	7×2000
$\text{trace}(S_w)$	299750	14238	942.3	1.6	2.0	3.0
$\text{trace}(S_b)$	23225	23225	1712	3.4	4.0	4.0
$\frac{\text{trace}(S_b)}{\text{trace}(S_w)}$	0.078	1.63	1.82	2.2	2.0	1.3
$\text{trace}(S_w^{-1} S_b)$	12.3	11.42	11.42	11.0	12.3	12.3
centroid	2.8	2.8	3.2	4.6	2.6	2.6
5nn	20.5	3.3	3.5	5.3	3.0	3.1
15nn	10.2	3.1	3.2	4.6	2.5	2.8
50nn	6.3	3.0	3.4	4.2	2.7	2.8

As proved in our previous work [8], the misclassification rates obtained using the centroid-based classification algorithm in the full space and in the orthogonal centroid-reduced space are identical. It is interesting to observe that the values of $\text{trace}(S_b)$ in these two spaces are also identical, although the motivation for the orthogonal centroid algorithm was not the preservation of $\text{trace}(S_b)$ after dimension reduction. We state this result in the following theorem.

THEOREM 3. *Let $Q_k \in \mathbb{R}^{m \times k}$ be the matrix with orthonormal columns in the reduced QR decomposition of the matrix $C \in \mathbb{R}^{m \times k}$ whose columns are the k centroids (see Algorithm Orthogonal Centroid). Then* $\text{trace}(S_b) = \text{trace}(Q_k^T S_b Q_k) = \text{trace}(S_b^Y)$, *where $Y = Q_k^T A$.*

Proof. There is an orthogonal matrix $\mathrm{Q} \in \mathbb{R}^{m \times m}$ such that

$$C = Q \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where $R \in \mathbb{R}^{k \times k}$ is upper triangular. Partitioning Q as $Q = (Q_k, \hat{Q})$, we have

$$C = (Q_k, \hat{Q}) \begin{pmatrix} R \\ 0 \end{pmatrix} = Q_k R. \tag{25}$$

Premultiplying (25) by $(Q_k, \hat{Q})^T$ gives $Q_k^T C = R$ and $\hat{Q}^T C = 0$. Therefore,

$$\begin{aligned} \text{trace}(S_b) &= \text{trace}(Q^T Q S_b) \\ &= \text{trace}(Q^T S_b Q) \\ &= \text{trace}((Q_k, \hat{Q})^T H_b H_b^T (Q_k, \hat{Q})) \\ &= \text{trace}(Q_k^T H_b H_b^T Q_k) \\ &= \text{trace}(Q_k^T S_b Q_k), \end{aligned}$$

where $H_b = [\sqrt{n_1}(c^{(1)} - c), \sqrt{n_2}(c^{(2)} - c), \ldots, \sqrt{n_k}(c^{(k)} - c)]$ and $\hat{Q}^T H_b = 0$, since $\hat{Q}^T c^{(i)} = 0$ and c is a linear combination of the $c^{(i)}$'s. ∎

In Test II, for the second data type, we use five categories of abstracts from the MEDLINE[1] database. Each category has 40 documents. The total number of terms is 7519 (see Table 2) after preprocessing with stopping and stemming algorithms [9]. For this 7519×200 term-document matrix, the original discriminant analysis breaks down, since S_w is singular. However, our improved LDA/GSVD method circumvents this singularity problem.

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

TABLE 2
Medline data set for Test II.

Class	Data from MEDLINE	
	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
Test II: *Traces and misclassification rate with L_2 norm similarity.*

Method		Full	Orthogonal centroid	Centroid	LDA/GSVD
Dim		7519×200	5×200	5×200	4×200
Trace values	$\text{trace}(S_w)$	73048	4210	90	0.05
	$\text{trace}(S_b)$	6229	6229	160	3.95
	$\frac{\text{trace}(S_b)}{\text{trace}(S_w)}$	0.09	1.5	1.8	79
Misclassification rate in %	centroid	5	5	2	1
	1nn	40	3	2.5	1

By Algorithm LDA/GSVD the dimension 7519 is dramatically reduced to 4, which is one less than the number of classes. The other methods reduce the dimension to the number of classes, which is 5. Table 3 shows classification results using the L_2 norm similarity measure. As in the results of Test I, LDA/GSVD produces the lowest misclassification rate using both classification methods. Because the J_1 criterion is not defined in this case, we compute the ratio $\text{trace}(S_b)/\text{trace}(S_w)$ as an approximate optimality measure. We observe that the ratio is strikingly higher for the LDA/GSVD reduction than for the other methods, and that, once again, the ratio produced by each of the three dimension reduction methods is greater than that of the full-dimensional data.

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 which the original discriminant analysis cannot handle. In addition, our LDA/GSVD algorithm avoids the numerical problems inherent in explicitly forming the scatter matrices.

In terms of computational complexity, 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 $\mathcal{O}(m^2 t)$ when $m > n$. Therefore, a fast algorithm needs to be developed for step 2.

Finally, we observe that dimension reduction is only a preprocessing stage. Even if this stage is a little expensive, it will be worthwhile if it effectively reduces the cost of the postprocessing involved in classification and document retrieval, which will be the dominating parts computationally.

Acknowledgments. The authors would like to thank Profs. Lars Eldén and Chris Paige for valuable discussions which improved the presentations in this paper. A part of this work was carried out during the summer of 2001, when H. Park was visiting the School of Mathematics, Seoul National University, Seoul, Korea, with the

support of the Brain Korea 21 program. She would like to thank Profs. Hyuck Kim and Dongwoo Sheen, as well as the School of Mathematics at SNU for their kind invitation.

REFERENCES

[1] M. W. Berry, S. T. Dumais, and G. W. O'Brien, *Using linear algebra for intelligent information retrieval*, SIAM Rev., 37 (1995), pp. 573–595.
[2] A. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.
[3] W. B. Frakes and R. Baeza-Yates, *Information Retrieval: Data Structures and Algorithms*, Prentice–Hall, Englewood Cliffs, NJ, 1992.
[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, New York, 1990.
[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[6] P. Howland and H. Park, *Extension of discriminant analysis based on the generalized singular value decomposition*, in Proceedings of the Second SIAM International Conference on Data Mining/Text Mining Workshop, SIAM, Philadelphia, 2002.
[7] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice–Hall, Englewood Cliffs, NJ, 1988.
[8] M. Jeon, H. Park, and J. B. Rosen, *Dimension reduction based on centroids and least squares for efficient processing of text data*, in Proceedings of the First SIAM International Conference on Data Mining, CD-ROM, SIAM, Philadelphia, 2001.
[9] G. Kowalski, *Information Retrieval Systems: Theory and Implementation*, Kluwer Academic, Dordrecht, The Netherlands, 1997.
[10] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, SIAM, Philadelphia, 1995.
[11] C. C. Paige and M. A. Saunders, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.
[12] H. Park, M. Jeon, and J. B. Rosen, *Lower dimensional representation of text data based on centroids and least squares*, BIT, to appear.
[13] G. Salton, *The SMART Retrieval System*, Prentice–Hall, Englewood Cliffs, NJ, 1971.
[14] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw–Hill, New York, 1983.
[15] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, New York, 1999.
[16] C. F. Van Loan, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.