SIAM J. MATRIX ANAL. APPL.
Vol. 32, No. 3, pp. 820–844
© 2011 Society for Industrial and Applied Mathematics

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

CHARACTERIZATION OF ALL SOLUTIONS FOR UNDERSAMPLED UNCORRELATED LINEAR DISCRIMINANT ANALYSIS PROBLEMS*

DELIN CHU†, SIONG THYE GOH†, AND Y. S. HUNG‡

Abstract. In this paper the uncorrelated linear discriminant analysis (ULDA) for undersampled problems is studied. The main contributions of the present work include the following: (i) all solutions of the optimization problem used for establishing the ULDA are parameterized explicitly; (ii) the optimal solutions among all solutions of the corresponding optimization problem are characterized in terms of both the ratio of between-class distance to within-class distance and the maximum likelihood classification, and it is proved that these optimal solutions are exactly the solutions of the corresponding optimization problem with minimum Frobenius norm, also minimum nuclear norm; these properties provide a good mathematical justification for preferring the minimum-norm transformation over other possible solutions as the optimal transformation in ULDA; (iii) explicit necessary and sufficient conditions are provided to ensure that these minimal solutions lead to a larger ratio of between-class distance to within-class distance, thereby achieving larger discrimination in the reduced subspace than that in the original data space, and our numerical experiments show that these necessary and sufficient conditions hold true generally. Furthermore, a new and fast ULDA algorithm is developed, which is eigendecomposition-free and SVD-free, and its effectiveness is demonstrated by some real-world data sets.

Key words. data dimensionality reduction, uncorrelated linear discriminant analysis, QR factorization

AMS subject classifications. 15A09, 68T10, 62H30, 65F15, 15A18, 15A23, 68T05

DOI. 10.1137/100792007

1. Introduction.

Linear discriminant analysis (LDA) is a powerful technique for data dimensionality reduction [1], [2], [3], [4], [6], [8], [10], [11], [12], [14], [16], [19], [23], [24], [27], [28], [29], [30], [33], [34], [35], [36], [37], [38], [39], [40], [41]. It seeks an optimal linear transformation of the data to a low-dimensional subspace. Preferably the reduced dimension is as small as possible, and in the reduced subspace the data features can be modeled with maximal discriminative power. LDA has found many important applications, for example, in pattern recognition [10], [24], [36], face recognition [25], [32], text classification [38], information retrieval [30], [34], and microarray data analysis [20], [21]. A major disadvantage of the classical LDA is that the so-called total scatter matrix must be nonsingular. But, in many applications such as those mentioned above, the total scatter matrix is singular since usually the number of the data samples is smaller than the data dimension. This is known as the undersampled problem [36], also commonly called the small sample size problem. As a result, the classical LDA cannot be applied directly to undersampled problems. To apply LDA to undersampled problems, many extensions of the classical LDA have been proposed recently. These extensions include uncorrelated LDA (ULDA) [13], [15], [25], [26], orthogonal LDA (OLDA) [13], the regularized LDA [17], [37], null space–based LDA (NLDA) [22], [28], GSVD-based LDA (LDA/GSVD) [14], [16], [18], Bayes optimal LDA [5], and least squares LDA [9]. However, all these extended LDA compute the optimal linear transformations by computing

*Received by the editors April 12, 2010; accepted for publication (in revised form) by L. De Lathauwer January 28, 2011; published electronically August 23, 2011.
http://www.siam.org/journals/simax/32-3/79200.html

†Department of Mathematics, Faculty of Science, National University of Singapore, Block S17, 10 Lower Kent Ridge Road, Singapore 119076 (matchudl@nus.edu.sg, g0700501@nus.edu.sg). The work of these authors was supported by NUS research grant R-146-000-140-112.

‡Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong (yshung@hkueee.hku.hk).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

some eigendecompositions/singular value decompositions (SVD), which are computationally expensive. Hence, it is important to develop new and fast algorithms for these extended LDA; preferably the new algorithms are eigendecomposition-free and SVD-free.

ULDA has been studied in [13], [15], [25], [26], and its effectiveness has been demonstrated by many numerical experiments. The feature vectors transformed by ULDA are mutually uncorrelated. This is highly desirable for feature extraction in many applications in order to contain minimum redundancy. The optimal transformation of ULDA in [13] is a solution of an optimization problem. However, this optimization problem has so many different solutions. It is not clear yet how a particular solution should be selected as the optimal transformation of ULDA in [13]. It is necessary to find a mathematical criterion for selecting a particular solution from all solutions of the related optimization problem as the optimal transformation of ULDA.

In this paper we focus on the ULDA for the undersampled problems. The main contributions of the present work include the following:

(i) All solutions of the optimization problem used for establishing the ULDA are parameterized explicitly.

(ii) The optimal solutions among all solutions of the corresponding optimization problem are characterized in terms of both the ratio of between-class distance to within-class distance and the maximum likelihood classification; it has been proved that these optimal solutions are exactly the solutions of the corresponding optimization problem with minimum Frobenius norm, also exactly the solutions with minimum nuclear norm. Hence, these minimal solutions can be considered to be optimal candidates for the optimal transformations in ULDA. These properties provide a mathematical criterion for the selection of the optimal transformations in ULDA.

(iii) Explicit necessary and sufficient conditions are provided to ensure that these minimal solutions lead to a larger ratio of between-class distance to within-class distance, thereby achieving larger discrimination in the reduced subspace than that in the original data space, and our numerical experiments show that these necessary and sufficient conditions hold true generally.

Along with the above mathematical findings, a new and fast ULDA algorithm is also developed, which is eigendecomposition-free and SVD-free. Real-world data sets show that the new algorithm has improved performance over the fast ULDA algorithm in [7].

2. Uncorrelated LDA. Consider a data matrix $A \in \mathbf{R}^{m\times n}$ with $m \gg n$ representing a set of n m-dimensional data points. Assume that a class label is available for every data point and that A is partitioned into k classes as

$$A = [\, a_1 \quad a_2 \quad \cdots \quad a_n \,] = [\, \mathcal{A}_1 \quad \mathcal{A}_2 \quad \cdots \quad \mathcal{A}_k \,],$$

where

$$\mathcal{A}_i \in \mathbf{R}^{m\times n_i}, \qquad i = 1, \ldots, k,$$

and

$$\sum_{i=1}^{k} n_i = n.$$

Further, let

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$e = [1 \quad \cdots \quad 1]^T \in \mathbf{R}^{n \times 1},$$
$$e_i = [1 \quad \cdots \quad 1]^T \in \mathbf{R}^{n_i \times 1}, \qquad i = 1, \ldots, k,$$

and denote the set of column indices that belong to the class i by $\mathcal{N}_i$. The centroid $c^{(i)}$ and the global centroid are given by

$$c^{(i)} = \frac{1}{n_i} \mathcal{A}_i e_i, \qquad i = 1, \ldots, k,$$

and

$$c = \frac{1}{n} A e,$$

respectively. Then the between-class scatter matrix S_b, the within-class scatter matrix S_w, and the total scatter matrix S_t are defined as

$$S_b = \sum_{i=1}^{k} \sum_{j \in \mathcal{N}_i} (c^{(i)} - c)(c^{(i)} - c)^T = \sum_{i=1}^{k} n_i (c^{(i)} - c)(c^{(i)} - c)^T,$$
$$S_w = \sum_{i=1}^{k} \sum_{j \in \mathcal{N}_i} (a_j - c^{(i)})(a_j - c^{(i)})^T,$$
$$S_t = \sum_{j=1}^{n} (a_j - c)(a_j - c)^T.$$

It is well known [18] that $S_t = S_b + S_w$. Let

$$H_b = [\sqrt{n_1}(c^{(1)} - c) \ \cdots \ \sqrt{n_k}(c^{(k)} - c)] \in \mathbf{R}^{m \times k},$$
$$H_w = [\mathcal{A}_1 - c^{(1)} e_1^T \ \cdots \ \mathcal{A}_k - c^{(k)} e_k^T] \in \mathbf{R}^{m \times n},$$
$$H_w = [\mathcal{A}_1 - c^{(1)} e_1^T \ \cdots \ \mathcal{A}_k - c^{(k)} e_k^T] \in \mathbf{R}^{m \times n},$$

The scatter matrices S_b, S_w, and S_t can be expressed as

$$S_b = H_b H_b^T, \qquad S_w = H_w H_w^T, \qquad S_t = H_t H_t^T, \tag{1}$$

since

$$\mathrm{Trace}(S_b) = \sum_{i=1}^{k} \sum_{j \in \mathcal{N}_i} \| c^{(i)} - c \|_2^2,$$

and

$$\mathrm{Trace}(S_w) = \sum_{i=1}^{k} \sum_{j \in \mathcal{N}_i} \| a_j - c^{(i)} \|_2^2.$$

Obviously, $\mathrm{Trace}(S_b)$ measures the distance between classes, while $\mathrm{Trace}(S_w)$ measures the closeness of the data within the classes over all k classes. Note that when the between-class relationship is remote, i.e., the centroids of the classes are remote,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Trace(S_b) will have a large value, whereas when data within each class are located tightly around their own class centroid, Trace(S_w) will have a small value. Hence, the cluster quality can be measured using Trace(S_b) and Trace(S_w).

In the lower-dimensional space mapped upon using the linear transformation $G^T \in \mathbf{R}^{l\times m}$, the between-class, within-class, and total scatter matrices are of the forms

$$S_b^G = G^T S_b G, \qquad S_w^G = G^T S_w G, \qquad S_t^G = G^T S_t G.$$

Ideally, the optimal transformation G^T would maximize Trace(S_b^L) and minimize Trace(S_w^L) simultaneously and equivalently maximize Trace(S_b^L) and minimize Trace(S_t^L) simultaneously, which leads to the optimization in classical LDA for determining the optimal linear transformation G^T, namely, the classical Fisher criterion:

$$G = \arg\max_G \{\mathrm{Trace}((S_t^G)^{-1} S_b^G)\}. \tag{2}$$

In the classical LDA [36], the above optimization problem is solved by computing all the eigenpairs

$$S_b x = \lambda S_t x, \qquad \lambda \neq 0.$$

Thus, the solution G can be characterized explicitly through the eigendecomposition of the matrix $S_t^{-1} S_b$ if S_t is nonsingular. It is easy to know that rank(S_b) $\leq k-1$, and so the reduced dimension by the classical LDA is at most $k-1$.

The classical LDA does not work when S_t is singular, which is the case for undersampled problems. To deal with the singularity of S_t, several generalized optimization criteria for determining the transformation G have been proposed. In particular, the optimization criterion

$$G = \arg \max_{G^T S_t G = I} \mathrm{Trace}((S_t^G)^{(+)} S_b^G) = \arg \max_{G^T S_t G = I} \mathrm{Trace}(S_b^G) \tag{3}$$

is used for ULDA in [13], [15], [25], [26]. ULDA was originally proposed in [25] for extracting feature vectors with uncorrelated attributes. The idea in [25] for computing the optimal discriminant vectors of ULDA is as follows: suppose r optimal discriminant vectors $g_1, \dots, g_r$ are obtained; then the $(r+1)$th vector g_{r+1} is obtained by maximizing the Fisher criterion function

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

subject to the constraints

$$g_{r+1}^T S_t g_i = 0, \qquad i = 1, \dots, r.$$

As a result, the algorithm in [25] computes the jth discriminant vector g_j of ULDA as the eigenvector corresponding to the maximum eigenvalue of the following generalized eigenvalue problem:

$$U_j S_b g_j = \lambda_j S_w g_j,$$

where

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$
\begin{aligned}
&U_1 = I_m,\\
&D_j = [g_1 \ \cdots \ g_{j-1}]^T \qquad (j > 1),\\
&U_j = I_m - S_t D_j^T (D_j S_t S_w^{-1} S_t D_j^T)^{-1} D_j S_t S_w^{-1} \qquad (j > 1).
\end{aligned}
$$

The feature vectors transformed by ULDA are mutually uncorrelated. This is desirable for feature extraction in many applications in order to reduce data redundancy. The main limitations of the algorithm above for ULDA are that it is computationally very expensive for large and high-dimensional data sets, and it is not applicable to undersampled problems.

It was later shown in [13], [15], [26] that classical LDA is equivalent to ULDA in the sense that both classical LDA and ULDA produce the same transformation matrix when the total scatter matrix S_t is nonsingular. The ULDA in [25] was also generalized in [13], [15] for undersampled problems based on simultaneous diagonalization of scatter matrices. Let the SVD of H_t be given by

$$H_t = U\Sigma V^T,$$

where U and V are orthogonal and $\Sigma = \left[\begin{smallmatrix} \Sigma_\gamma & 0 \\ 0 & 0 \end{smallmatrix}\right]$ with $\gamma = \operatorname{rank}(H_t)$ and Σ_γ being diagonal. Then

$$S_t = H_t H_t^T = U \begin{bmatrix} \Sigma_\gamma^2 & 0 \\ 0 & 0 \end{bmatrix} U^T.$$

Let $U = [\, U_1 \quad U_2 \,]$ with $U_1 \in \mathbf{R}^{m\times\gamma}$ and $U_2 \in \mathbf{R}^{m\times(m-\gamma)}$. Since $S_t = S_b + S_w$, we have

$$
\begin{aligned}
&U^T S_b U = \begin{bmatrix} U_1^T S_b U_1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad U^T S_w U = \begin{bmatrix} U_1^T S_w U_1 & 0 \\ 0 & 0 \end{bmatrix},\\
&\Sigma_\gamma^2 = U_1^T S_b U_1 + U_1^T S_w U_1;
\end{aligned}
$$

thus,

$$\Sigma_\gamma^{-1} U_1^T S_b U_1 \Sigma_\gamma^{-1} + \Sigma_\gamma^{-1} U_1^T S_w U_1 \Sigma_\gamma^{-1} = I.$$

Next, let the SVD of $\Sigma_\gamma^{-1} U_1^T H_b$ be

$$\Sigma_\gamma^{-1} U_1^T H_b = P\Lambda Q^T,$$

where P and Q are orthogonal, $\Lambda = \left[\begin{smallmatrix} \Lambda_b & 0 \\ 0 & 0 \end{smallmatrix}\right]$, and $\Lambda_b \in \mathbf{R}^{q\times q}$ is diagonal with $q = \operatorname{rank}(S_b)$. Define

$$X = U \begin{bmatrix} \Sigma_\gamma^{-1} P & 0 \\ 0 & I \end{bmatrix}.$$

Then we have

$$X^T S_b X = \begin{bmatrix} \Lambda_b^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad X^T S_w X = \begin{bmatrix} I - \Lambda_b^2 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad X^T S_t X = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

The above analysis yields that the matrix $X\left[\begin{smallmatrix} I_q \\ 0 \end{smallmatrix}\right]$, i.e., the first q columns of X, is a solution to the optimization problem (3), giving rise to the following ULDA algorithm [13].

Algorithm 1 (ULDA [13]).
Input: data matrix A, class number k;
Output: transformation matrix G

1. Form matrices H_b, H_t;
2. Compute the reduced SVD of H_t without forming V explicitly to get $H_t = U_1[\,\Sigma_\gamma \quad 0\,]V^T$, with $\Sigma_\gamma \in \mathbf{R}^{\gamma\times\gamma}$, $\gamma = \mathrm{rank}(H_t)$;
3. Compute the reduced SVD of $\Sigma_\gamma^{-1}U_1^T H_b$ without forming Q explicitly to get $\Sigma_\gamma^{-1}U_1^T H_b = P_1[\,\Lambda_b \quad 0\,]Q^T$, with $\Lambda_b \in \mathbf{R}^{q\times q}$, $P_1 \in \mathbf{R}^{\gamma\times q}$, $q = \mathrm{rank}(H_b)$;
4. $G = U_1\Sigma_\gamma^{-1}P_1$.

A fast ULDA algorithm is given in [7]. The following is its pseudocode.

Algorithm 2 (Fast ULDA [7]).
Input: data matrix A.
output: transformation matrix G.

1. Compute the economic QR factorization of A as $A = U_1 R$ and partition R into $R = [\,R_1 \quad \ldots \quad R_k\,]$, $R_i \in \mathbf{R}^{n\times n_i}$, $i = 1, \ldots, k$;
2. Compute $\hat{c} = \frac{1}{n}Re \in \mathbf{R}^n$, $\hat{c}^{(i)} = \frac{1}{n_i}R_i e_i \in \mathbf{R}^n$, $i = 1, \ldots, k$, and then form matrices $\hat{H}_b = \left[\sqrt{n_1}(\hat{c}^{(1)} - \hat{c}) \quad \sqrt{n_2}(\hat{c}^{(2)} - \hat{c}) \quad \ldots \quad \sqrt{n_k}(\hat{c}^{(k)} - \hat{c})\right] \in \mathbf{R}^{n\times k}$, $\hat{H}_w = [\,R_1 - \hat{c}^{(1)}e_1^T \quad R_2 - \hat{c}^{(2)}e_2^T \quad \ldots \quad R_k - \hat{c}^{(k)}e_k^T\,] \in \mathbf{R}^{n\times n}$;
3. Compute the complete orthogonal decomposition of $\left[\begin{smallmatrix} \hat{H}_b^T \\ \hat{H}_w^T \end{smallmatrix}\right]$ as $\left[\begin{smallmatrix} \hat{H}_b^T \\ \hat{H}_w^T \end{smallmatrix}\right] = \hat{P}\left[\begin{smallmatrix} \hat{R} & 0 \\ 0 & 0 \end{smallmatrix}\right]\hat{V}^T$ and let $\gamma = \mathrm{rank}(\hat{R})$;
4. Compute the SVD of $\hat{P}(1\!:\!k, 1\!:\!\gamma)$ as $\hat{P}(1\!:\!k, 1\!:\!\gamma) = \hat{U}\,\hat{\mathcal{R}}\,\hat{W}^T$;
5. Compute the first $k-1$ columns of $U_1\hat{V}\left[\begin{smallmatrix} \hat{R}^{-1}\hat{W} & 0 \\ 0 & I \end{smallmatrix}\right]$, and then assign them to G.

Many numerical experiments in [7], [13], [15], [25], [26] have shown that Algorithms 1 and 2 are powerful for data dimensionality reduction. However, Algorithms 1 and 2 have implicitly chosen without any theoretical basis a particular solution from so many different solutions of the optimization problem (3) as the optimal transformation of ULDA. In the next section, we will study the properties of the set of all solutions to the optimization problem (3), with an aim to provide a theoretical justification for selecting the optimal transformation for ULDA among all possible solutions of (3).

3. New results. In this section we will first derive an explicit characterization of all solutions (in Theorem 4) to the optimization problem (3). As a result, we can explore optimal solutions with further properties (in Theorems 5, 7, and 8) among the set of all solutions to the optimization problem (3).

Denote

$$E = \frac{1}{n}ee^T, \qquad E_i = \frac{1}{ni}e_ie_i^T, \qquad i = 1, \ldots, k.$$

The scatter matrices S_t, S_b, and S_w can be written as

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

$$S_t = A(I-E)A^T, \qquad S_b = A\left(\begin{bmatrix} E_1 & & \\ & \ddots & \\ & & E_k \end{bmatrix} - E\right)A^T,$$
$$(4) \qquad S_w = A\left(I - \begin{bmatrix} E_1 & & \\ & \ddots & \\ & & E_k \end{bmatrix}\right)A^T.$$

Note that

$$I-E, \qquad \begin{bmatrix} E_1 & & \\ & \ddots & \\ & & E_k \end{bmatrix} - E, \qquad I - \begin{bmatrix} E_1 & & \\ & \ddots & \\ & & E_k \end{bmatrix}$$

are orthogonal projections in $\mathbf{R}^n$. Let $\mathcal{R}_t$, $\mathcal{R}_b$, and $\mathcal{R}_w$ be the range spaces of the above orthogonal projections, respectively. It can be shown that $\mathcal{R}_t = \mathcal{R}_b + \mathcal{R}_w$ with

$$\dim(\mathcal{R}_t) = n-1, \qquad \dim(\mathcal{R}_b) = k-1, \qquad \dim(\mathcal{R}_w) = n-k.$$

We now devise an orthogonal basis in $\mathbf{R}^n$ containing partitions that span the subspaces $\mathcal{R}_b$ and $\mathcal{R}_w$. Define Householder transformations

$$W_i = I - \left(\begin{bmatrix} 1-\sqrt{n_i} \\ 1 \\ \vdots \\ 1 \end{bmatrix} \Big/ \sqrt{n_i - \sqrt{n_i}}\right)\left(\begin{bmatrix} 1-\sqrt{n_i} \\ 1 \\ \vdots \\ 1 \end{bmatrix} \Big/ \sqrt{n_i - \sqrt{n_i}}\right)^T,$$
$$i = 1, \ldots, k,$$

$$W = I - \left(\begin{bmatrix} \sqrt{n_1}-\sqrt{n} \\ \sqrt{n_2} \\ \vdots \\ \sqrt{n_k} \end{bmatrix} \Big/ \sqrt{n - \sqrt{nn_1}}\right)\left(\begin{bmatrix} \sqrt{n_1}-\sqrt{n} \\ \sqrt{n_2} \\ \vdots \\ \sqrt{n_k} \end{bmatrix} \Big/ \sqrt{n - \sqrt{nn_1}}\right)^T.$$

Matrices W and $W_i (i = 1, \ldots, k)$ are orthogonal satisfying

$$W = W^T, \qquad W_i = W_i^T (i = 1, \ldots, k),$$

$$W^T\left(\begin{bmatrix} \sqrt{n_1} \\ \sqrt{n_2} \\ \vdots \\ \sqrt{n_k} \end{bmatrix} \Big/ \sqrt{n}\right) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad W_i^T(e_i/\sqrt{n_i}) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad i = 1, \ldots, k.$$

Let P be the permutation matrix obtained by exchanging the $(\sum_{j=1}^{i-1} n_j + 1)$th column of I_n and the ith column (for $i = 2, \ldots, k$), but otherwise leaving the order of the remaining columns unchanged. It can be verified by a straightforward calculation that

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

$$\left(\begin{bmatrix} W_1 & & \\ & \ddots & \\ & & W_k \end{bmatrix} P \begin{bmatrix} W & \\ & I \end{bmatrix}\right)^T \left(I - \begin{bmatrix} E_1 & & \\ & \ddots & \\ & & E_k \end{bmatrix}\right) \begin{bmatrix} W_1 & & \\ & \ddots & \\ & & W_k \end{bmatrix} P \begin{bmatrix} W & \\ & I \end{bmatrix}$$

$$= \begin{bmatrix} 0 & \\ & I_{n-k} \end{bmatrix} \tag{5}$$

and

$$\left(\begin{bmatrix} W_1 & & \\ & \ddots & \\ & & W_k \end{bmatrix} P \begin{bmatrix} W & \\ & I \end{bmatrix}\right)^T (I - E) \begin{bmatrix} W_1 & & \\ & \ddots & \\ & & W_k \end{bmatrix} P \begin{bmatrix} W & \\ & I \end{bmatrix}$$

$$= \begin{bmatrix} 0_{1\times 1} & & \\ & I_{k-1} & \\ & & 0 \end{bmatrix}. \tag{6}$$

The following lemma is a direct consequence of (4), (5), and (6).

LEMMA 1. *Denote*

$$[\, A_1 \quad A_2 \quad A_3 \,] := A \begin{bmatrix} W_1 & & \\ & \ddots & \\ & & W_k \end{bmatrix} P \begin{bmatrix} W & \\ & I \end{bmatrix},$$

where $A_1 \in \mathbf{R}^{m\times 1}$, $A_2 \in \mathbf{R}^{m\times (k-1)}$, *and* $A_3 \in \mathbf{R}^{m\times(n-k)}$. *Then*

$$S_b = A_2 A_2^T, \qquad S_w = A_3 A_3^T, \qquad S_t = [\, A_2 \quad A_3 \,][\, A_2 \quad A_3 \,]^T. \tag{7}$$

LEMMA 2. *Let* $\mathcal{G}_1 \in \mathbf{R}^{\mu\times\tau}$ *and* $\mathcal{G}_2 \in \mathbf{R}^{\nu\times\tau}$ *satisfy* $\left[\begin{smallmatrix}\mathcal{G}_1\\ \mathcal{G}_2\end{smallmatrix}\right]^T \left[\begin{smallmatrix}\mathcal{G}_1\\ \mathcal{G}_2\end{smallmatrix}\right] = I_\tau$. *Let* $\mathcal{B} \in \mathbf{R}^{\mu\times\mu}$ *be symmetric positive definite. Then*

$$\mathrm{Trace}(\mathcal{G}_1^T \mathcal{B} \mathcal{G}_1) = \mathrm{Trace}(\mathcal{B})$$

if and only if

$$\mathcal{G}_1 \mathcal{G}_1^T = I_\mu.$$

Proof. Since $\left[\begin{smallmatrix}\mathcal{G}_1\\ \mathcal{G}_2\end{smallmatrix}\right]^T \left[\begin{smallmatrix}\mathcal{G}_1\\ \mathcal{G}_2\end{smallmatrix}\right] = I_\tau$, there exist $\tilde{\mathcal{G}}_1 \in \mathbf{R}^{\mu\times(\mu+\nu-\tau)}$ and $\tilde{\mathcal{G}}_2 \in \mathbf{R}^{\nu\times(\mu+\nu-\tau)}$ such that $\left[\begin{smallmatrix}\mathcal{G}_1 & \tilde{\mathcal{G}}_1\\ \mathcal{G}_2 & \tilde{\mathcal{G}}_2\end{smallmatrix}\right]$ is orthogonal, and thus

$$\mathcal{G}_1 \mathcal{G}_1^T + \tilde{\mathcal{G}}_1 \tilde{\mathcal{G}}_1^T = I_\mu. \tag{8}$$

Hence, we have

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\begin{aligned}\mathrm{Trace}(\mathcal{B}) &= \mathrm{Trace}(\mathcal{B}(\mathcal{G}_1\mathcal{G}_1^T + \tilde{\mathcal{G}}_1\tilde{\mathcal{G}}_1^T)) \\ &= \mathrm{Trace}(\mathcal{B}\mathcal{G}_1\mathcal{G}_1^T) + \mathrm{Trace}(\mathcal{B}\tilde{\mathcal{G}}_1\tilde{\mathcal{G}}_1^T) \\ &= \mathrm{Trace}(\mathcal{G}_1^T\mathcal{B}\mathcal{G}_1) + \mathrm{Trace}(\tilde{\mathcal{G}}_1^T\mathcal{B}\tilde{\mathcal{G}}_1^T),\end{aligned}$$

which gives that

$$\begin{aligned}\mathrm{Trace}(\mathcal{G}_1^T\mathcal{B}\mathcal{G}_1) = \mathrm{Trace}(\mathcal{B}) &\Leftrightarrow \mathrm{Trace}(\tilde{\mathcal{G}}_1^T\mathcal{B}\tilde{\mathcal{G}}_1) = 0 \\ &\Leftrightarrow \tilde{\mathcal{G}}_1 = 0 \qquad (\text{since } \mathcal{B} \text{ is symmetric positive definite}) \\ &\Leftrightarrow \mathcal{G}_1\mathcal{G}_1^T = I_\mu \qquad (\text{since (8) holds}).\end{aligned}$$

The following result can be found in [13].

LEMMA 3 (see [13]).

$$\max_G \mathrm{Trace}((S_t^G)^{(+)}S_b^G) = \mathrm{Trace}(S_t^{(+)}S_b).$$

THEOREM 4. *Let the economic QR factorization of the data matrix A be*

$$A = U_1 R, \tag{9}$$

where $U_1 \in \mathbf{R}^{m\times n}$ is column orthogonal and $R \in \mathbf{R}^{n\times n}$. Denote

$$[\,R_1 \quad R_2 \quad R_3\,] = R\begin{bmatrix} W_1 & & \\ & \ddots & \\ & & W_k\end{bmatrix} P \begin{bmatrix} W & \\ & I\end{bmatrix}, \tag{10}$$

where

$$R_1 \in \mathbf{R}^{n\times 1}, \qquad R_2 \in \mathbf{R}^{n\times(k-1)}, \qquad R_3 \in \mathbf{R}^{n\times(n-k)}.$$

Let the economic QR factorization of $[\,R_2 \quad R_3\,]$ with column pivoting be

$$[\,R_2 \quad R_3\,] = Q_1\mathcal{R}, \tag{11}$$

where $Q_1 \in \mathbf{R}^{n\times\gamma}$ is column orthogonal, $\mathcal{R} \in \mathbf{R}^{\gamma\times(n-1)}$, and $\mathrm{rank}(\mathcal{R}) = \mathrm{rank}([\,R_2 \quad R_3\,]) = \gamma$*. Further, let the economic QR factorization of $\mathcal{R}^T$ be*

$$\mathcal{R}^T = P_1^T\Delta^T, \tag{12}$$

where $P_1 \in \mathcal{R}^{\gamma\times(n-1)}$ is row orthogonal and $\Delta \in \mathbf{R}^{\gamma\times\gamma}$ is lower triangular. Denote

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

$$P_1 = [\,P_{11} \quad P_{12}\,], \qquad P_{11} \in \mathbf{R}^{\gamma\times(k-1)}.$$

Finally, let the economic QR factorization of P_{11} *with column pivoting be*

$$P_{11} = V_1\Pi_{11}, \tag{13}$$

where $V_1 \in \mathbf{R}^{\gamma\times q}$ *is column orthogonal,* $\Pi_{11} \in \mathbf{R}^{q\times(k-1)}$*, and* $\mathrm{rank}(\Pi_{11}) = q$*. Then all solutions* $G \in \mathbf{R}^{m\times l}$ *of the optimization problem* (3) *are parameterized by*

$$G = (U_1Q_1\Delta^{-T}[\,V_1 \quad \mathcal{N}_1\,] + \mathcal{N}_2)\mathcal{Z}, \qquad q \le l \le \gamma, \tag{14}$$

where $\mathcal{Z} \in \mathbf{R}^{l\times l}$ *is any orthogonal matrix,* $\mathcal{N}_1 \in \mathbf{R}^{\gamma\times(l-q)}$ *is any column orthogonal matrix satisfying* $\mathcal{N}_1^TV_1 = 0$*, and* $\mathcal{N}_2 \in \mathbf{R}^{m\times l}$ *is any matrix satisfying* $\mathcal{N}_2^TU_1Q_1 = 0$*.*

Proof. Let $U_2 \in \mathbf{R}^{m\times(m-n)}$, $Q_2 \in \mathbf{R}^{n\times(n-\gamma)}$, and $V_2 \in \mathbf{R}^{\gamma\times(\gamma-q)}$ be such that $[\,U_1 \quad U_2\,]$, $[\,Q_1 \quad Q_2\,]$, and $[\,V_1 \quad V_2\,]$ are orthogonal. Denote

$$\begin{aligned}\mathcal{H} &= \left(\begin{bmatrix} U_1 & U_2\end{bmatrix}\begin{bmatrix} Q_1\Delta[\,V_1 \quad V_2\,] & Q_2 & 0\\ 0 & 0 & I\end{bmatrix}\right)^{-T}\\ &= [\,U_1Q_1\Delta^{-T}[\,V_1 \quad V_2\,] \quad U_1Q_2 \quad U_2\,].\end{aligned}$$

In the following we prove Theorem 4 by four arguments, outlined as follows before the full details are given:

- First it is shown in Argument 1 that $\mathcal{H}$ can be used to diagonalize scatter matrices S_t, S_b, and S_w; that is,

$$\left\{\begin{aligned} &\mathcal{H}^TS_t\mathcal{H} = \mathcal{H}^T[\,A_2 \quad A_3\,][\,A_2 \quad A_3\,]^T\mathcal{H} = \begin{bmatrix} I_q & 0 & 0 & 0\\ 0 & I_{\gamma-q} & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix},\\ &\mathcal{H}^TS_b\mathcal{H} = \mathcal{H}^TA_2A_2^T\mathcal{H} = \begin{bmatrix} \Pi_{11}\Pi_{11}^T & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix},\\ &\mathcal{H}^TS_w\mathcal{H} = \mathcal{H}^T(S_t - S_b)\mathcal{H} = \begin{bmatrix} I - \Pi_{11}\Pi_{11}^T & 0 & 0 & 0\\ 0 & I_{\gamma-q} & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}.\end{aligned}\right. \tag{15}$$

- Then it is shown in Argument 2 that

$$\mathrm{Trace}(S_t^{(+)}S_b) = \mathrm{Trace}(\Pi_{11}\Pi_{11}^T). \tag{16}$$

- Next it is shown in Argument 3 using (15) and (16) that G is a solution of the optimization problem (3) if and only if

$$G = \left(U_1Q_1\Delta^{-T}[\,V_1 \quad V_2\mathcal{G}_2\,] + [\,U_1Q_2 \quad U_2\,]\begin{bmatrix} G_{31} & G_{32}\\ G_{41} & G_{42}\end{bmatrix}\right)\mathcal{Z}, \tag{17}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

where $\begin{bmatrix} G_{31} & G_{32} \\ G_{41} & G_{42} \end{bmatrix} = \begin{bmatrix} G_3 \\ G_4 \end{bmatrix} \mathcal{Z}^T$, $G_{31} \in \mathbf{R}^{(n-\gamma)\times q}$, $G_{32} \in \mathbf{R}^{(n-\gamma)\times(l-q)}$, $G_{41} \in \mathbf{R}^{(m-n)\times q}$, and $G_{42} \in \mathbf{R}^{(m-n)\times(l-q)}$.

- Finally it is shown in Argument 1 using (17) that G is a solution of the optimization problem (3) if and only if G is of the form (14).

Argument 1. Note that

$$\begin{aligned}
[\, A_1 \quad A_2 \quad A_3 \,] &= A \begin{bmatrix} W_1 & & \\ & \ddots & \\ & & W_k \end{bmatrix} P \begin{bmatrix} W & \\ & I \end{bmatrix} \\
&= [\, U_1 \quad U_2 \,] \begin{bmatrix} R \\ 0 \end{bmatrix} \begin{bmatrix} W_1 & & \\ & \ddots & \\ & & W_k \end{bmatrix} P \begin{bmatrix} W & \\ & I \end{bmatrix} \\
&= [\, U_1 \quad U_2 \,] \begin{bmatrix} R_1 & R_2 & R_3 \\ 0 & 0 & 0 \end{bmatrix};
\end{aligned}$$

thus,

$$[\, A_2 \;\big|\; A_3 \,] = [\, U_1 \quad U_2 \,] \left[\begin{array}{c|c} R_2 & R_3 \\ 0 & 0 \end{array}\right].$$

Consequently, we get

$$\begin{aligned}
[\, A_2 \quad A_3 \,] &= [\, U_1 \quad U_2 \,] \left[\begin{array}{c|c} R_2 & R_3 \\ 0 & 0 \end{array}\right] \\
&= [\, U_1 \quad U_2 \,] \begin{bmatrix} Q_1 & Q_2 & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \mathcal{R} \\ 0 \\ 0 \end{bmatrix} \\
&= [\, U_1 \quad U_2 \,] \begin{bmatrix} Q_1 & Q_2 & 0 \\ 0 & 0 & I \end{bmatrix} \left[\begin{array}{c|c} \Delta P_{11} & \Delta P_{12} \\ 0 & 0 \\ 0 & 0 \end{array}\right] \\
&= [\, U_1 \quad U_2 \,] \begin{bmatrix} Q_1 & Q_2 & 0 \\ 0 & 0 & I \end{bmatrix} \left[\begin{array}{c|c} \Delta \; [\, V_1 \quad V_2 \,] \begin{bmatrix} \Pi_{11} \\ 0 \end{bmatrix} & \Delta \; [\, V_1 \quad V_2 \,] \begin{bmatrix} \Pi_{12} \\ \Pi_{22} \end{bmatrix} \\ 0 & 0 \\ 0 & 0 \end{array}\right] \\
&= [\, U_1 \quad U_2 \,] \begin{bmatrix} Q_1 \Delta [\, V_1 \quad V_2 \,] & Q_2 & 0 \\ 0 & 0 & I \end{bmatrix} \left[\begin{array}{c|c} \Pi_{11} & \Pi_{12} \\ 0 & \Pi_{22} \\ 0 & 0 \\ 0 & 0 \end{array}\right],
\end{aligned}$$

where

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

$$\begin{bmatrix} \Pi_{12} \\ \Pi_{22} \end{bmatrix} = [\, V_1 \quad V_2 \,]^T P_{12}.$$

Note that P_1 is row orthogonal. This yields that

$$\begin{bmatrix} \Pi_{11} & \Pi_{12} \\ 0 & \Pi_{22} \end{bmatrix} \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ 0 & \Pi_{22} \end{bmatrix}^T = [\, V_1 \quad V_2 \,]^T P_1 P_1^T [\, V_1 \quad V_2 \,] = \begin{bmatrix} I_q & 0 \\ 0 & I_{\gamma-q} \end{bmatrix},$$

which together with Lemma 1 gives (15), and thus

$$q = \mathrm{rank}(\Pi_{11}) = \mathrm{rank}(S_b).$$

Argument 2. Now we consider $\mathrm{Trace}(S_t^{(+)} S_b)$. Since Δ is nonsingular, V_1 is column orthogonal, and

$$\begin{cases} S_t = [\, A_2 \quad A_3 \,][\, A_2 \quad A_3 \,]^T \\ \quad = [\, U_1 \quad U_2 \,] \begin{bmatrix} Q_1 & Q_2 & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \Delta\Delta^T & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \left([\, U_1 \quad U_2 \,] \begin{bmatrix} Q_1 & Q_2 & 0 \\ 0 & 0 & I \end{bmatrix} \right)^T, \\ S_b = A_2 A_2^T \\ \quad = [\, U_1 \quad U_2 \,] \begin{bmatrix} Q_1 & Q_2 & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \Delta V_1 \Pi_{11} \Pi_{11}^T V_1^T \Delta^T & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ \qquad \left([\, U_1 \quad U_2 \,] \begin{bmatrix} Q_1 & Q_2 & 0 \\ 0 & 0 & I \end{bmatrix} \right)^T, \end{cases}$$

we obtain

$$\begin{aligned} \mathrm{Trace}(S_t^{(+)} S_b) &= \mathrm{Trace}((\Delta\Delta^T)^{-1} (\Delta V_1 \Pi_{11} \Pi_{11}^T V_1^T \Delta^T)) = \mathrm{Trace}(V_1 \Pi_{11} \Pi_{11}^T V_1^T) \\ &= \mathrm{Trace}(\Pi_{11} \Pi_{11}^T); \end{aligned}$$

i.e., (16) holds true.

Argument 3. For any $G \in \mathbf{R}^{m \times l}$, denote

$$\mathcal{G} = \mathcal{H}^{-1} G = \begin{bmatrix} G_1 \\ G_2 \\ G_3 \\ G_4 \end{bmatrix}, \qquad G_1 \in \mathbf{R}^{q \times l}, \qquad G_2 \in \mathbf{R}^{(\gamma-q) \times l}, \qquad G_3 \in \mathbf{R}^{(n-\gamma) \times l},$$
$$G_4 \in \mathbf{R}^{(m-n) \times l}.$$

It is obvious that

$$\begin{aligned} G^T S_t G &= \mathcal{G}^T (\mathcal{H}^T S_t \mathcal{H}) \mathcal{G} = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}^T \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}, \\ G^T S_b G &= \mathcal{G}^T (\mathcal{H}^T S_b \mathcal{H}) \mathcal{G} = G_1^T \Pi_{11} \Pi_{11}^T G_1. \end{aligned} \tag{18}$$

We have using (16) and (18) that

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

G is a solution of the optimization problem (3)

$$\Leftrightarrow \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}^T \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} = I_l, \quad \mathrm{Trace}(G_1^T \Pi_{11} \Pi_{11}^T G_1) = \mathrm{Trace}(\Pi_{11}\Pi_{11}^T)$$

$$\Leftrightarrow \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}^T \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} = I_l,$$

$G_1 G_1^T = I_q$(by Lemma 2 with $\mathcal{G}_1 := G_1, \mathcal{G}_2 := G_2, \mathcal{B} := \Pi_{11}\Pi_{11}^T, \mu := q, \tau := l$)

$$\Leftrightarrow \begin{cases} q \le l \le \gamma, \\ \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} = \begin{bmatrix} I_q & 0 \\ 0 & \mathcal{G}_2 \end{bmatrix} \mathcal{Z}, \end{cases}$$

$\mathcal{G}_2 \in \mathbf{R}^{(\gamma-q)\times(l-q)}$ is column orthogonal and $\mathcal{Z} \in \mathbf{R}^{l\times l}$ is orthogonal,

$$\Leftrightarrow G = [\, U_1 Q_1 \Delta^{-T} [\, V_1 \quad V_2 \,] \quad U_1 Q_2 \quad U_2 \,] \begin{bmatrix} I_q & 0 \\ 0 & \mathcal{G}_2 \\ G_{31} & G_{32} \\ G_{41} & G_{42} \end{bmatrix} \mathcal{Z}$$

$$= \left(U_1 Q_1 \Delta^{-T} [\, V_1 \quad V_2 \mathcal{G}_2 \,] + [\, U_1 Q_2 \quad U_2 \,] \begin{bmatrix} G_{31} & G_{32} \\ G_{41} & G_{42} \end{bmatrix} \right) Z,$$

where $\left[\begin{smallmatrix} G_{31} & G_{32} \\ G_{41} & G_{42} \end{smallmatrix}\right] = \left[\begin{smallmatrix} G_3 \\ G_4 \end{smallmatrix}\right] \mathcal{Z}^T$, $G_{31} \in \mathbf{R}^{(n-\gamma)\times q}$, $G_{32} \in \mathbf{R}^{(n-\gamma)\times(l-q)}$, $G_{41} \in \mathbf{R}^{(m-n)\times q}$, and $G_{42} \in \mathbf{R}^{(m-n)\times(l-q)}$.

Argument 4. Since $[\, V_1 \quad V_2 \,]$ and $[\, U_1 Q_1 \quad U_1 Q_2 \quad U_2 \,]$ are orthogonal, it holds for any $\mathcal{N}_1 \in \mathbf{R}^{\gamma\times(l-q)}$ and $\mathcal{N}_2 \in \mathbf{R}^{m\times l}$ that

$$\mathcal{N}_1^T V_1 = 0, \quad \mathcal{N}_1 \text{ is column orthogonal} \Leftrightarrow \mathcal{N}_1 = V_2 \mathcal{G}_2, \\ \mathcal{G}_2 \in \mathbf{R}^{(\gamma-q)\times(l-q)} \text{ is column orthogonal,}$$

and

$$\mathcal{N}_2^T U_1 Q_1 = 0$$

$$\Leftrightarrow \begin{cases} \mathcal{N}_2 = [\, U_1 Q_2 \quad U_2 \,] \begin{bmatrix} G_{31} & G_{32} \\ G_{41} & G_{42} \end{bmatrix}, \\ G_{31} \in \mathbf{R}^{(n-\gamma)\times q}, \quad G_{32} \in \mathbf{R}^{(n-\gamma)\times(l-q)}, \quad G_{41} \in \mathbf{R}^{(m-n)\times q}, \quad G_{42} \in \mathbf{R}^{(m-n)\times(l-q)}. \end{cases}$$

Hence, we have that $G \in \mathbf{R}^{m\times l}$ is a solution of the optimization problem (1) if and only if

$$\begin{cases} q \le l \le \gamma, \\ G = (U_1 Q_1 \Delta^{-T} [\, V_1 \quad \mathcal{N}_1 \,] + \mathcal{N}_2) \mathcal{Z}, \end{cases}$$

where $\mathcal{Z} \in \mathbf{R}^{l\times l}$ is orthogonal, $\mathcal{N}_1 \in \mathbf{R}^{\gamma\times(l-q)}$ is column orthogonal satisfying $\mathcal{N}_1^T V_1 = 0$, and $\mathcal{N}_2 \in \mathbf{R}^{m\times l}$ is any matrix satisfying $\mathcal{N}_2^T U_1 Q_1 = 0$.

To have high cluster quality, a specific clustering result must have a tight within-class relationship, while the between-class relationship should be remote. With this objective, the ratio $\mathrm{Trace}(S_b^G)/\mathrm{Trace}(S_w^G)$, that is, the ratio of the between-class distance to within-class distance, is an important measure of how well $\mathrm{Trace}(S_b^G)$ is maximized

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

while $\mathrm{Trace}(S_w^G)$ is minimized in the reduced space [18]. The following result reveals the conditions under which the ratio $\mathrm{Trace}(S_b^G)/\mathrm{Trace}(S_w^G)$ obtained by a solution of the optimization problem (3) is greater than the ratio $\mathrm{Trace}(S_b)/\mathrm{Trace}(S_w)$ of the full-dimension data.

THEOREM 5.

(i) *Let $G \in \mathbf{R}^{m\times l}$ be any arbitrary solution of the optimization problem* (3). *Then*

$$G = \arg\ \max\left\{\frac{\mathrm{Trace}(S_b^{G_l})}{\mathrm{Trace}(S_w^{G_l})}\colon G_l \textit{ is a solution of the optimization problem } (3)\right\} \tag{19}$$

if and only if H

$$l = q, \qquad G = U_1 Q_1 \Delta^{-T} V_1 \mathcal{Z} + \mathcal{N}, \tag{20}$$

where $\mathcal{Z} \in \mathbf{R}^{q\times q}$ is any orthogonal matrix and $\mathcal{N} \in \mathbf{R}^{m\times q}$ is any matrix satisfying $\mathcal{N}^T U_1 Q_1 = 0$.

(ii) *For any solution G in the form* (20) *of the optimization problem* (3),

$$\frac{\mathrm{Trace}(S_b)}{\mathrm{Trace}(S_w)} \le \frac{\mathrm{Trace}(S_b^G)}{\mathrm{Trace}(S_w^G)} \tag{21}$$

if and only if

$$\mathrm{rank}(S_b)\mathrm{Trace}(S_b) \le \mathrm{Trace}(S_t)\mathrm{Trace}(S_t^{(+)} S_b), \tag{22}$$

which, with the notation in Theorem 4, *is equivalent to*

$$\sqrt{q}\|\Delta V_1 \Pi_{11}\|_F \le \|\Delta\|_F \|\Pi_{11}\|_F.$$

Proof. For any solution $G \in \mathbf{R}^{m\times l}$ of the optimization problem (3), Theorem 4 gives that

$$\begin{aligned}
&\mathrm{Trace}(S_t^G) = \mathrm{Trace}(G^T S_t G) = l \ge q,\\
&\mathrm{Trace}(S_b^G) = \mathrm{Trace}(G^T S_b G) = \mathrm{Trace}(S_t^{(+)} S_b),\\
&\mathrm{Trace}(S_w^G) = \mathrm{Trace}(G^T S_w G) = l - \mathrm{Trace}(S_t^{(+)} S_b),\\
&\mathrm{rank}(S_b) = q,\\
&\mathrm{Trace}(S_t^{(+)} S_b) = \mathrm{Trace}(\Pi_{11}\Pi_{11}^T) = \|\Pi_{11}\|_F^2,\\
&\mathrm{Trace}(S_b) = \mathrm{Trace}((\Delta V_1 \Pi_{11})(\Delta V_1 \Pi_{11})^T) = \|\Delta V_1 \Pi_{11}\|_F^2,\\
&\mathrm{Trace}(S_t) = \mathrm{Trace}(\Delta\Delta^T) = \|\Delta\|_F^2,
\end{aligned}$$

and further if $l = q$, then G is of the form (20). Hence, Theorem 5 follows. ∎

It should be pointed out that Theorem 5 holds only for optimization problem (3) but does not hold for the optimization problems, for example, for OLDA [13] and NLDA [22], [28].

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Clearly, Theorem 5 provides explicit necessary and sufficient conditions to ensure that these minimal solutions lead to a larger ratio of between-class distance to within-class distance, thereby achieving larger discrimination in the reduced subspace than that in the original data space.

Noting that

$$\mathrm{Trace}(S_t^{(+)} S_b) \geq \frac{1}{\|S_t\|_2} \mathrm{Trace}(S_b)$$

and

$$\|S_t\|_2 = \|\Delta\|_2^2,$$

we have the following result by using Theorem 5.

COROLLARY 6. *If*

$$\mathrm{rank}(S_b)\|S_t\|_2 \leq \mathrm{Trace}(S_t) \tag{23}$$

or, equivalently,

$$\sqrt{q}\|\Delta\|_2 \leq \|\Delta\|_F,$$

then (21) *holds true.*

It is well known that one important property of the classical LDA is that it is equivalent to maximum likelihood classification, assuming normal distribution for each data class with the common covariance matrix. We will next derive a necessary and sufficient condition under which this property also holds true for the ULDA on undersampled problems. Classification in classical LDA based on the maximum likelihood estimation is based on the Mahalanobis distance as follows: a test data h is classified as class j if

$$j = \arg\min_j (h - c^{(j)})^T S_t^{-1} (h - c^{(j)}).$$

For the undersampled problem, it has been shown in [13] that for the G in Algorithm 1 the following holds for any test data $h \in \mathbf{R}^m$:

$$\arg\min_j \{(h - c^{(j)})^T S_t^{(+)} (h - c^{(j)})\} = \arg\min_j \{\| G^T (h - c^{(j)})\|_2^2\}. \tag{24}$$

The result below is a stronger one which establishes a necessary and sufficient condition, ensuring (24) holds a true for a solution of the optimization problem (3).

THEOREM 7. *With the notation in Theorem* 4, *let $G \in \mathbf{R}^{m\times l}$ be a solution of the optimization problem* (3). *Then* (24) *holds for any test data $h \in \mathbf{R}^m$ if and only if*

$$G = (U_1 Q_1 \Delta^{-T} [\, V_1 \quad \mathcal{N}_1 \,] + [\, 0 \quad \hat{\mathcal{N}}_2 \,])\mathcal{Z}, \tag{25}$$

where $\mathcal{Z} \in \mathbf{R}^{l\times l}$ is any orthogonal matrix, $\mathcal{N}_1 \in \mathbf{R}^{\gamma\times(l-q)}$ is any column orthogonal matrix satisfying $\mathcal{N}_1^T V_1 = 0$, and $\hat{\mathcal{N}}_2 \in \mathbf{R}^{m\times(l-q)}$ is any matrix satisfying $\hat{\mathcal{N}}_2^T U_1 Q_1 = 0$.

Proof. We have from Theorem 4 and its proof that

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

$$q \le l \le \gamma,$$

$$
\begin{aligned}
G &= (U_1 Q_1 \Delta^{-T} [\, V_1 \quad \mathcal{N}_1 \,] + \mathcal{N}_2)\mathcal{Z} \\
&= \left(U_1 Q_1 \Delta^{-T} [\, V_1 \quad V_2 \mathcal{G}_2 \,] + [\, U_1 Q_2 \quad U_2 \,] \begin{bmatrix} G_{31} & G_{32} \\ G_{41} & G_{42} \end{bmatrix} \right) \mathcal{Z},
\end{aligned} \tag{26}
$$

where

$$\mathcal{N}_1 = V_2 \mathcal{G}_2, \qquad \mathcal{N}_2 = [\, U_1 Q_2 \quad U_2 \,] \begin{bmatrix} G_{31} & G_{32} \\ G_{41} & G_{42} \end{bmatrix},$$

$\mathcal{G}_2 \in \mathbf{R}^{(\gamma - q) \times l}$ and $\mathcal{Z} \in \mathbf{R}^{l \times l}$ are any column orthogonal matrix and orthogonal matrix, respectively, $\left[\begin{smallmatrix} G_{31} & G_{32} \\ G_{41} & G_{42} \end{smallmatrix}\right]$ is arbitrary, $G_{31} \in \mathbf{R}^{(n-\gamma) \times q}$, $G_{32} \in \mathbf{R}^{(n-\gamma) \times (l-q)}$, $G_{41} \in \mathbf{R}^{(m-n) \times q}$, $G_{42} \in \mathbf{R}^{(m-n) \times (l-q)}$, and furthermore

$$
\begin{aligned}
S_t^{(+)} &= U_1 Q_1 \Delta^{-T} \Delta^{-1} (U_1 Q_1)^T \\
&= (U_1 Q_1 \Delta^{-T} V_1)(U_1 Q_1 \Delta^{-T} V_1)^T + (U_1 Q_1 \Delta^{-T} V_2)(U_1 Q_1 \Delta^{-T} V_2)^T,
\end{aligned}
$$

and for $j = 1, \ldots, k$,

$$
\begin{cases}
\begin{bmatrix} (U_1 Q_1 \Delta^{-T} V_2)^T \\ (U_1 Q_2)^T \\ U_2^T \end{bmatrix} S_b [\, U_1 Q_1 \Delta^{-T} V_2 \quad U_1 Q_2 \quad U_2 \,] = 0, \\
\begin{bmatrix} (U_1 Q_1 \Delta^{-T} V_2)^T \\ (U_1 Q_2)^T \\ U_2^T \end{bmatrix} c^{(j)} = \begin{bmatrix} (U_1 Q_1 \Delta^{-T} V_2)^T \\ (U_1 Q_2)^T \\ U_2^T \end{bmatrix} c.
\end{cases}
$$

So,

$$
\begin{gathered}
(h - c^{(j)})^T S_t^{(+)} (h - c^{(j)}) = \|(U_1 Q_1 \Delta^{-T} V_1)(h - c^{(j)})\|_2^2 + \|(U_1 Q_1 \Delta^{-T} V_2)^T (h - c)\|_2^2, \\
j = 1, \ldots, k,
\end{gathered}
$$

$$\arg\min_j \{(h - c^{(j)})^T S_t^{(+)} (h - c^{(j)})\} = \arg\min_j \|(U_1 Q_1 \Delta^{-T} V_1)^T (h - c^{(j)})\|_2^2, \tag{27}$$

and

$$
\begin{aligned}
G^T (h - c^{(j)}) = \mathcal{Z}^T & \left(\begin{bmatrix} (U_1 Q_1 \Delta^{-T} V_1)^T (h - c^{(j)}) \\ \mathcal{G}_2^T (U_1 Q_1 \Delta^{-T} V_2)^T (h - c) \end{bmatrix} \right. \\
& \left. + \begin{bmatrix} G_{31}^T & G_{41}^T \\ G_{32}^T & G_{42}^T \end{bmatrix} \begin{bmatrix} (U_1 Q_2)^T \\ U_2^T \end{bmatrix} (h - c) \right), \\
& j = 1, \ldots, k.
\end{aligned} \tag{28}
$$

Sufficiency. Let (25) hold. Then $\left[\begin{smallmatrix} G_{31} \\ G_{41} \end{smallmatrix}\right] = 0$, and for any test data $h \in \mathbf{R}^m$ and for $j = 1, \ldots, k$,

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\begin{cases} G^T(h-c^{(j)}) = \mathcal{Z}^T \begin{bmatrix} (U_1Q_1\Delta^{-T}V_1)^T(h-c^{(j)}) \\ \left(\mathcal{G}_2^T(U_1Q_1\Delta^{-T}V_2)^T + [\,G_{32}^T \quad G_{42}^T\,]\begin{bmatrix} (U_1Q_2)^T \\ U_2^T \end{bmatrix}\right)(h-c) \end{bmatrix}, \\ \|G^T(h-c^{(j)})\|_2^2 = \|(U_1Q_1\Delta^{-T}V_1)^T(h-c^{(j)})\|_2^2 \\ \qquad\qquad + \left\|\left(\mathcal{G}_2^T(U_1Q_1\Delta^{-T}V_2)^T + [\,G_{32}^T \quad G_{42}^T\,]\begin{bmatrix} (U_1Q_2)^T \\ U_2^T \end{bmatrix}\right)(h-c)\right\|_2^2, \\ \arg\min_j \|G^T(h-c^{(j)})\|_2^2 = \arg\min_j \|(U_1Q_1\Delta^{-T}V_1)^T(h-c^{(j)})\|_2^2, \end{cases}$$

the above equalities together with (27) give that (24) holds for any test data $h \in \mathbf{R}^m$.

Necessity. Note that (24) holds for any test data $h \in \mathbf{R}^m$, so it also holds true for any test data h of the form

$$h = c + [\,U_1Q_2 \quad U_2\,]x, \qquad x \in \mathbf{R}^{m-\gamma}.$$

For such h, it holds that

$$\arg\min_j \{(h-c^{(j)})^T S_t^{(+)}(h-c^{(j)})\} = \arg\min_j \|(U_1Q_1\Delta^{-T}V_1)^T(c-c^{(j)})\|_2^2,$$

$$G^T(h-c^{(j)}) = \mathcal{Z}^T \begin{bmatrix} (U_1Q_1\Delta^{-T}V_1)^T(c-c^{(j)}) + [\,G_{31}^T \quad G_{41}^T\,]x \\ [\,G_{32}^T \quad G_{42}^T\,]x \end{bmatrix},$$

$$\|G^T(h-c^{(j)})\|_2^2 = \|(U_1Q_1\Delta^{-T}V_1)^T(c-c^{(j)}) + [\,G_{31}^T \quad G_{41}^T\,]x\|_2^2 + \|[\,G_{32}^T \quad G_{42}^T\,]x\|_2^2,$$

and

$$\arg\min_j \|G^T(h-c^{(j)})\|_2^2 = \arg\min_j \|(U_1Q_1\Delta^{-T}V_1)^T(c-c^{(j)}) + [\,G_{31}^T \quad G_{41}^T\,]x\|_2^2.$$

Hence, we obtain by using (24) that

$$[\,G_{31}^T \quad G_{41}^T\,] = 0.$$

Equivalently,

$$\begin{aligned} G &= \left(U_1Q_1\Delta^{-T}[\,V_1 \quad V_2\mathcal{G}_2\,] + [\,U_1Q_2 \quad U_2\,]\begin{bmatrix} 0 & G_{32} \\ 0 & G_{42} \end{bmatrix}\right)\mathcal{Z} \\ &= (U_1Q_1\Delta^{-T}[\,V_2 \quad \mathcal{N}_1\,] + [\,0 \quad \hat{\mathcal{N}}_2\,])\mathcal{Z}, \end{aligned} \tag{29}$$

where $\mathcal{Z}$ is orthogonal, and $\mathcal{N}_1 = V_2\mathcal{G}_2$ is column orthogonal and satisfies that $\mathcal{N}_1^T V_1 = \mathcal{G}_2^T V_2^T V_1 = 0$, $\hat{\mathcal{N}}_2 = [\,U_1Q_2 \quad U_2\,][{}^{G_{32}}_{G_{42}}]$, and $\hat{\mathcal{N}}_2^T U_1Q_1 = 0$. ∎

Note that any solution G of the optimization problem (3) is of the form (26); thus,

$$G = [\,U_1Q_1 \quad U_1Q_2 \quad U_2\,]\begin{bmatrix} \Delta^{-T}V_1 & \Delta^{-T}V_2\mathcal{G}_2 \\ G_{31} & G_{32} \\ G_{41} & G_{42} \end{bmatrix}\mathcal{Z},$$

$$\|G\|_F^2 = \|\Delta^{-T}V_1\|_F^2 + \|\Delta^{-T}V_2\mathcal{G}_2\|_F^2 + \left\|\begin{bmatrix} G_{31} & G_{32} \\ G_{41} & G_{42} \end{bmatrix}\right\|_F^2,$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

and

$$\|G\|_\star = \left\| \begin{bmatrix} \Delta^{-T}V_1 & \Delta^{-T}V_2\mathcal{G}_2 \\ G_{31} & G_{32} \\ G_{41} & G_{42} \end{bmatrix} \right\|_\star \geq \|[\,\Delta^{-T}V_1 \quad \Delta^{-T}V_2\mathcal{G}_2\,]\|_\star \geq \|\Delta^{-T}V_1\|_\star, \tag{30}$$

where $\|\cdot\|_F$ and $\|\cdot\|_\star$ denote the Frobenius norm and nuclear norm, respectively. It is easy to see that the first inequality in (30) becomes an equality if and only if $\left[\begin{smallmatrix} G_{31} & G_{32} \\ G_{41} & G_{42}\end{smallmatrix}\right] = 0$, and the second one becomes an equality if and only if $\Delta^{-T}V_2\mathcal{G}_2$ vanishes (i.e., $l = q$) since $\Delta^{-T}V_2\mathcal{G}_2$ is of full column rank if $l > q$. Therefore, the following result is a direct consequence of Theorems 5 and 7.

Theorem 8. *With the notation in Theorem* 4, *let G be a solution of the optimization problem* (3). *Then the following four statements are equivalent:*

(i) *Equations* (19) *and* (24) *hold simultaneously for any test data $h \in \mathbf{R}^m$.*

(ii)

$$G = U_1Q_1\Delta^{-T}V_1\mathcal{Z}, \qquad \mathcal{Z} \in \mathbf{R}^{q\times q} \textit{ is orthogonal.} \tag{31}$$

(iii) *G is a solution of the optimization problem* (3) *with minimum Frobenius norm; i.e.,*

$$G = \arg\ \min\{\|G_l\|_F\colon G_l \textit{ is a solution of the optimization problem}\,(3)\}.$$

(iv) *G is a solution of the optimization problem* (3) *with minimum nuclear norm;*[1] *i.e.,*

$$G = \arg\ \min\{\|G_l\|_\star\colon G_l \textit{ is a solution of the optimization problem}\,(3)\}.$$

We now summarize the main findings in this section. Theorem 3 provides a characterization of all optimal solutions to the ULDA problem (3). Further properties of the optimal solutions are then explored in Theorems 5, 7, and 8. Theorem 5 shows that all optimal solutions to problem (3) that maximize the ratio of between-class distance to within-class distance are characterized by (20). Theorem 7 shows that all optimal solutions to problem (3) that solve the maximum likelihood classification problem in classical LDA are characterized by (25). Taking the intersection of the above two sets of solutions, i.e., satisfying both (20) and (25), yields (31) which characterizes all optimal solutions to (3) that have both properties of the optimal ratio of between-class distance to within-class distance as well as maximum likelihood classification. Moreover, Theorem 8 shows that the solutions characterized by (31) are exactly those solutions of the optimization problem (3) with minimal Frobenius and/or nuclear norm.

We further note that it is natural to pick the minimum-norm transformation G among all possible solutions to (3), and this is the current practice, perhaps because there is no better reason for other choices of G. With the above remarks, the significance of our results is that we have now provided a good justification for preferring the minimum-norm transformation over other possible solutions.

[1]For any matrix, its nuclear norm is defined as the sum of all its singular values.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

4. Numerical experiments.

4.1. A new ULDA algorithm. The results in section 3 lead to the following new ULDA algorithm, in which the minimal solution $G = U_1 Q_1 \Delta^{-T} V_1$ of the optimization problem (3) is used as the optimal transformation in ULDA.

Algorithm 3 (Proposed ULDA method).
Input: data matrix $A \in \mathbf{R}^{m\times n}$, class number k.
Output: transformation matrix G

1. Compute economic QR factorization (1);
2. Compute R_2 and R_3 by (10);
3. Compute the economic QR factorization (1) with column pivoting and then compute the economic QR factorization (1);
4. Compute the economic QR factorization (13) with column pivoting;
5. Solve the upper triangular linear system of equations $\Delta^T Y_1 = V_1$ and then compute $G = U_1 Q_1 Y_1$.

Obviously, Algorithm 3 is eigendecomposition-free and SVD-free and can be carried out by means of four economic QR factorization with/without pivoting.

In the following we perform extensive experimental studies to evaluate and demonstrate the efficiency of our Algorithm 3. We perform a detailed comparison of Algorithms 1, 2, and 3 in terms of the classification accuracy and the computational time.

First, we estimate the computational complexities of Algorithms 1, 2, and 3 in Table 1, in which

$$q = \mathrm{rank}(S_b) = \mathrm{rank}(H_b) \leq k - 1$$

and

$$\gamma = \mathrm{rank}(S_t) = \mathrm{rank}(H_t) = \mathrm{rank}([\, H_b \quad H_w \,]) \leq n - 1.$$

We consider only the undersampled case, that is, $m > n$.

Table 1 implies that the computational complexities of Algorithms 2 and 3 are much lower than that of Algorithm 1. It can also be seen that Algorithm 3 has a lower computational complexity than Algorithm 2.

Remark 1. In Table 1, the following computational costs for QR factorization and SVD are used [31]:

Computational complexity for QR factorization of $\Theta \in \mathbf{R}^{m\times n}$ with $m \geq n$.

Full QR factorization: $4m^2n + \frac{2}{3}n^3 - 2mn^2$;

Economic QR factorization: $4mn^2 - \frac{4}{3}n^3$;

Full QR factorization with column pivoting: $(4m^2n - 4mn^2 + \frac{4}{3}n^3) + (4mnp - 2p^2(m+n) + \frac{4}{3}p^3)$, $p = \mathrm{rank}(\Theta)$;

Economic QR factorization with column pivoting: $2mn^2 - \frac{2}{3}n^3 + (4mnp - 2p^2(m+n) + \frac{4}{3}p^3)$;

Computational complexity for SVD ($\Theta = U\Sigma V^T$, $U_1 = U(:, 1:n)$) of $\Theta \in \mathbf{R}^{m\times n}$ with $m \geq n$.

Σ: $4mn^2 - \frac{4}{3}n^3$;
Σ, V: $4mn^2 + 8n^3$;
U, Σ: $4m^2n - 8mn^2$;
U_1, Σ: $14mn^2 - 2n^3$.
U, Σ, V: $4m^2n + 8mn^2 + 9n^3$;
U_1, Σ, V: $14mn^2 + 8n^3$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

TABLE 1
Computational complexities of Algorithms 1, 2, *and* 3.

The computational complexity of Algorithm 1.
Step 1: $\mathbf{O}(mn)$,
Step 2: $14mn^2 - 2n^3$,
Step 3: $2m\gamma k + 14\gamma k^2 - 2k^3$,
Step 4: $2m\gamma q$.
The computational complexity of Algorithm 2.
Steps 1 and 2: $4mn^2 + \frac{4}{3}n^3 + \mathbf{O}(n^2)$
Step 3: $4(n+k)^2 n - 4(n+k)n^2 + \frac{4}{3}n^3 + 4(n+k)n\gamma - 2\gamma^2(2n+k) + \frac{4}{3}\gamma^3 + 4n^2\gamma + \frac{2}{3}\gamma^3 - 2n\gamma^2$,
Step 4: $14\gamma k^2 - 2k^3$,
Step 5: $2mn(k-1) + 2n\gamma(k-1) + \gamma^2(k-1)$.
The computational complexity of Algorithm 3.
Steps 1 and 2: $4mn^2 - \frac{4}{3}n^3 + \mathbf{O}(n^2)$.
Step 3: $2n(n-1)^2 - \frac{2}{3}(n-1)^3 + [4n(n-1)\gamma - 2\gamma^2(n+n-1) + \frac{4}{3}\gamma^3] \leq \frac{8}{3}n^3 + 4(n-1)\gamma^2 - \frac{4}{3}\gamma^3 \leq \frac{8}{3}n^3$,
Step 4: $2\gamma(k-1)^2 - \frac{2}{3}(k-1)^3 + [4\gamma(k-1)q - 2q^2(\gamma+k-1) + \frac{4}{3}q^3] \leq 4\gamma k^2 - \frac{4}{3}k^3$,
Step 5: $2mnq + 2n\gamma q + \gamma^2 q$.

4.2. Numerical results. In this subsection we experiment on four face databases, the ORL face database, AR face database, FERET face database, and Palmprint database, to demonstrate the efficiency of Algorithm 3. These face databases are described as follows.

The ORL face database is available at http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. This database consists of 400 different images, 10 for each of 40 distinct subjects. All of the images in the ORL face database were resized to 32×32 pixels.

The AR face database is available at http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html. A subset of the AR database was used in our experiment. This subset includes 1680 color images corresponding to 120 persons' faces (70 men and 50 women). Images feature frontal view faces with different facial expressions, illumination conditions, and occlusions (sunglasses and scarf). The pictures of 120 individuals (65 men and 55 women) were taken in two sessions; 28 face images (each session containing 14) of these 120 individuals were used in our experiment. The face portion of each image was manually cropped and then resized to 50×40 pixels.

The FERET face database is available at http://www.itl.nist.gov/iad/humanid/feret/feret_master.html. This database has become a standard database for testing the state-of-the-art face recognition algorithms. A subset of the FERET database was used in our experiment. This subset includes 1000 images of 200 individuals (each one has 5 images). It is composed of the images whose names are marked with two-character strings: "ba", "bj", "bk", "be", and "bf". This subset involves variations in facial expression, illumination, and pose. In our experiment, the facial portion of each original image was automatically cropped based on the locations of eyes and mouths, and the cropped images were resized to 80×80 pixels and further preprocessed by histogram equalization.

The Palmprint database is available at http://www4.comp.polyu.edu.hk/~biometrics/. This database contains 100 different palms. Six samples from each of these palms were collected in two sessions, where three samples were captured in each session. All images from the Palmprint database were compressed to 64×64 pixels.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

TABLE 2
Data structures.

Data	m (data dimension)	n (training size)	k (number of classes)	(Number of test data)
ORL	1024	200	40	200
AR	2000	840	120	840
FERET	6400	600	200	400
Palmprint	4096	300	100	300

Table 2 summarizes the data structures in our experiments.

For all data sets above, we performed our study by repeated random splitting into training and test sets using the following algorithm: within each class, we randomly reordered the data and then for each class with size n_i, the first $\lceil 0.5n_i \rceil$ data were used as the training data and the others were used as test data, whereby $\lceil \cdot \rceil$ is the ceiling function. The splitting was repeated 10 times.

The experiments were conducted by using the Osprey workstation cluster with 8GB RAM located at the Center for Computational Science and Engineering, National University of Singapore.

We compare the classification accuracies (%) and the ratio of between-class distance to within-class distance of Algorithms 1, 2, and 3, and in the original data space, respectively, in Tables 3 and 4. We also compare the CPU time of Algorithms 1, 2, and 3 in Table 5. The standard deviations of classification accuracies are given in brackets in Table 3.

In the following experiments, we consider different solutions of the optimization problem (3) as the optimal transformations of ULDA and then compare their classification performances:

TABLE 3
Comparison of 1*-NN average classification accuracies (%) for Algorithms* 1*,* 2*, and* 3*, and the original data space, with standard deviations (in brackets).*

Data	Algorithm 1	Algorithm 2	Algorithm 3	The original data space
ORL	94.4000 (1.2247)	94.4000 (1.2247)	94.4000 (1.2247)	89.1500 (1.9200)
AR	95.5357 (0.5195)	95.5357 (0.5195)	95.5357 (0.5195)	85.1667 (1.0108)
FERET	70.8250 (1.7177)	70.8250 (1.7177)	70.8250 (1.7177)	58.9250 (2.7014)
Palmprint	99.3000 (0.5538)	99.3000 (0.5538)	99.3000 (0.5538)	97.6667 (0.9888)

TABLE 4
Comparison of average ratio of between-class distance to within-class distance for Algorithms 1*,* 2*, and* 3*, and the original data space.*

Data	Algorithm 1	Algorithm 2	Algorithm 3	The original data space
ORL	1.2385×10^{28}	1.2096×10^{29}	1.1870×10^{29}	1.4950
AR	1.8020×10^{27}	9.7702×10^{28}	8.2030×10^{28}	1.7305
FERET	7.2614×10^{27}	1.1219×10^{29}	1.0432×10^{29}	2.1525
Palmprint	1.2173×10^{28}	3.0686×10^{29}	3.7712×10^{29}	3.6737

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Table 5
Comparison of average CPU time (seconds) used by Algorithms 1, 2, and 3.

Data	Algorithm 1	Algorithm 2	Algorithm 3
ORL	0.2200	0.0510	0.0400
AR	10.9370	1.9430	1.3770
FERET	9.8440	1.9350	1.5810
Palmprint	1.5870	0.3300	0.2820

- $G = U_1 Q_1 \Delta^{-T} V_1 \in \mathbf{R}^{m\times q}$, i.e., the output G of Algorithm 3;
- Let $l = q + 3i$, with $i = 0, 1, \ldots, 10$, recover V_2 from the QR factorization of V_1; then take

$$G = U_1 Q_1 \Delta^{-T} [\, V_1 \quad V_2 \mathcal{G}_2 \,] \in \mathbf{R}^{m\times l}, \tag{32}$$

where

$$X = \mathrm{rand}(\gamma - q, l - q), \qquad [\mathcal{G}_2, \mathcal{R}] = qr(X, 0).$$

Note that for any $\mathcal{N}_2 \in \mathbf{R}^{m\times l}$ with $\mathcal{N}_2^T U_1 Q_1 = 0$, it holds that

$$\mathcal{N}_2^T S_t \mathcal{N}_2 = 0, \qquad \mathcal{N}_2^T S_b \mathcal{N}_2 = 0, \qquad \mathcal{N}_2^T S_w \mathcal{N}_2 = 0,$$

which means that $\mathcal{N}_2$ does not contain any useful discriminant information. Hence, to remove redundancy as far as possible, G in (32) does not contain such $\mathcal{N}_2$.

Tables 6 and 7 record the average 1-NN classification accuracies achieved and the ratio of between-class distance to within-class distance in the reduced space by different G with different l above over the 10 experiments. The standard deviations of classification accuracies are given in brackets in Table 6.

It is clear from Tables 3–7 that the following hold:

- Algorithms 1, 2, and 3 achieve similar classification accuracies. This is consistent with the fact that the transformation G produced by Algorithms 1, 2, and 3 are theoretically equivalent (since it holds that $q = k - 1$ for 4 data sets AR, ORL, FERET, and Palmprint).
- Algorithms 2 and 3 are much faster than Algorithm 1, and Algorithm 3 is faster than Algorithm 2.
- For 4 data sets AR, ORL, FERET, and Palmprint, it has been verified that

$$\begin{aligned} \mathrm{rank}([\, A_2 \quad A_3 \,]) &= \mathrm{rank}(A_2) + \mathrm{rank}(A_3), \quad \text{i.e.,} \\ \mathrm{rank}(S_t) &= \mathrm{rank}(S_b) + \mathrm{rank}(S_w). \end{aligned} \tag{33}$$

Consequently, we have from the proof of Theorem 5 that

$$\mathrm{Trace}(S_b^G) = \mathrm{Trace}(S_t^{(+)} S_b) = \mathrm{rank}(S_b) = q,$$
$$\mathrm{Trace}(S_w^G) = l - \mathrm{Trace}(S_t^{(+)} S_b) = l - q.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

TABLE 6

Comparison of 1*-NN average classification accuracies (%) for different G with different l, with standard deviations (in brackets).*

	ORL	AR	FERET	Palmprint
$l = q$	94.4000 (1.2247)	95.5357 (0.5195)	70.8250 (1.7177)	99.3000 (0.5538)
$l = q + 3$	94.4500 (1.7146)	95.4881 (0.7865)	71.1000 (1.3124)	99.3000 (0.3266)
$l = q + 6$	94.1000 (1.3191)	95.5476 (0.7664)	70.6500 (1.2452)	99.3333 (0.3350)
$l = q + 9$	93.9000 (1.2855)	95.4048 (0.6326)	70.7250 (1.4276)	99.3667 (0.4534)
$l = q + 12$	93.4500 (1.7212)	95.4167 (0.7112)	70.6000 (1.2945)	99.2667 (0.4269)
$l = q + 15$	93.2500 (1.8848)	95.4048 (0.9528)	70.3500 (1.1675)	99.3000 (0.3786)
$l = q + 18$	92.4000 (1.8364)	95.2619 (0.5954)	70.0500 (1.3096)	99.2667 (0.4014)
$l = q + 21$	92.1500 (1.9551)	95.2024 (0.7716)	70.1000 (1.3730	99.1667 (0.3590)
$l = q + 24$	92.0000 (1.8364)	95.2857 (0.7226)	70.1750 (0.9682)	99.0667 (0.4269)
$l = q + 27$	92.1500 (2.2962)	95.0119 (0.5906)	69.8250 (1.0610)	99.1667 (0.3350)
$l = q + 30$	91.6000 (1.8173)	94.9762 (0.7810)	69.7250 (1.1803)	99.1333 (0.4485)

TABLE 7

The average ratio of between-class distance to within-class distance in the reduced space for different G with different l.

	ORL	AR	FERET	Palmprint
$l = q$	1.1870×10^{29}	8.2030×10^{28}	1.0432×10^{29}	3.7712×10^{29}
$l = q + 3$	13.0000	39.667	66.3333	33.0000
$l = q + 6$	6.5000	19.833	33.1667	16.5000
$l = q + 9$	4.3333	13.222	22.1111	11.0000
$l = q + 12$	3.2500	9.9167	16.5833	8.2500
$l = q + 15$	2.6000	7.9333	13.2667	6.6000
$l = q + 18$	2.1667	6.6111	11.0556	5.5000
$l = q + 21$	1.8571	5.6667	9.4762	4.7143
$l = q + 24$	1.6250	4.9583	8.2917	4.1250
$l = q + 27$	1.4444	4.4074	7.3704	3.6667
$l = q + 30$	1.3000	3.9667	6.6333	3.3000

Thus, when $l = q$, it holds that $\mathrm{Trace}(S_w^G) = 0$, which leads to the huge numerical values in Table 4 and the second row of Table 7.

It should be pointed out that the equality $\mathrm{rank}([\, A_2 \quad A_3\,]) = \mathrm{rank}(A_2) + \mathrm{rank}(A_3)$ holds true for almost all A_2 and A_3 with appropriate sizes, so the condition (33) holds for almost all data sets. It is also worthy to note that condition (22) holds true in all our experiments. This gives that the optimal transformation G produced by Algorithm 3 yields a larger ratio of between-class distance to within-class distance, thereby achieving larger discrimination in the reduced subspace than that in the original data space.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

- Although the transformation G with the largest ratio of between-class distance to within-class distance does not always achieve the best classification accuracy, it always achieves at least comparable classification accuracy, and usually the transformation G with a relatively small ratio of between-class distance to within-class distance yields relatively low classification accuracy. Hence, our numerical experiments confirm the well-known fact that the ratio of between-class distance to within-class distance is a very important measure for data cluster quality.
- The minimal solution $G = U_1 Q_1 \Delta^{-T} V_1$ always achieves comparative classification accuracy compared with other solutions G in the form (32). Hence, it is reasonable to select it as the optimal transformation of the ULDA.

4.3. Conclusions. In this paper, all solutions to the optimization problem (3) for establishing ULDA have been characterized explicitly. With such a characterization, all optimal solutions to problem (3) that further maximize the ratio of between-class distance to within-class distance and also solve the maximum likelihood classification problem have been obtained. It turns out that these optimal solutions are exactly the solutions of the optimization problem (3) with minimum Frobenius norm and/or nuclear norm. Hence, it is natural to pick such a minimum-norm transformation G among all possible solutions to optimization problem (3) to be the transformation in ULDA. These properties provide a good mathematical justification for preferring the minimum-norm transformation over other possible solutions as the optimal transformation in ULDA. The explicit characterization of all solutions of the optimization problem (3) has led to Algorithm 3—a new and fast ULDA algorithm. Algorithm 3 is eigendecomposition-free and SVD-free, and its effectiveness has been demonstrated by some real-world data sets.

REFERENCES

[1] L. Zhang, L. Liao, and M. Ng, *Fast algorithms for the generalized Foley–Sammon discriminant analysis*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 1584–1605.

[2] D. Chu and S. T. Goh, *A new and fast implementation for null space based linear discriminant analysis*, Pattern Recognition, 43 (2010), pp. 1373–1379.

[3] D. Chu and S. T. Goh, *A new and fast orthogonal linear discriminant analysis on undersampled problems*, SIAM J. Sci. Comput., 32 (2010), pp. 2274–2297.

[4] D. Tao, X. Li, X. Wu, and S. J. Maybank, *Geometric mean for subspace selection*, IEEE Trans. Pattern Anal. Mach. Intell., 31 (2009), pp. 260–274.

[5] O. C. Hamsici and A. M. Martinez, *Bayes optimality in linear discriminant analysis*, IEEE Trans. Pattern Anal. Mach. Intell., 30 (2008), pp. 647–657.

[6] Y. Guo, T. Hastie, and R. Tibshirani, *Regularized linear discriminant analysis and its application in microarray*, Biostatistics, 8 (2007), pp. 86–100.

[7] H. Park, B. Drake, S. Lee, and C. Park, *Fast Linear Discriminant Analysis Using QR Decomposition and Regularization*, Technical report GT-CSE-07-21, Georgia Institute of Technology, Atlanta, GA, 2007, http://www.cc.gatech.edu/~hpark/papers/reglda_ieee.pdf.

[8] D. Tao, X. Li, X. Wu, and S. J. Maybank, *General averaged divergence analysis*, in Proceedings of the IEEE International Conference on Data Mining, 2007, pp. 302–311.

[9] J. Ye, *Least squares linear discriminant analysis*, in Proceedings of the Twenty-Fourth International Conference on Machine Learning, 2007, pp. 1087–1093.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[11] J. Ye and T. Xiong, *Computational and theoretical analysis of null space and orthogonal linear discriminant analysis*, J. Mach. Learn. Res., 7 (2006), 1183–1204.

[12] P. Hall, J. S. Marron, and A. Neeman, *Geometric representation of high dimension, low sample size data*, J. R. Stat. Soc. Ser. B, 67 (2005), pp. 427–444.

[13] J. Ye, *Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems*, J. Mach. Learn. Res., 6 (2005), pp. 483–502.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[14] P. Howland and H. Park, *Generalizing discriminant analysis using the generalized singular value decomposition*, IEEE Trans. Pattern Anal. Mach. Intell., 26 (2004), pp. 995–1006.

[15] J. Ye, R. Janardan, Q. Li, and H. Park, *Feature extraction via generalized uncorrelated linear discriminant analysis*, in Proceedings of the Twenty-First International Conference on Machine Learning, 2004, pp. 895–902.

[16] J. Ye, R. Janardan, C. H. Park, and H. Park, *An optimization criterion for generalized discriminant analysis on undersampled problems*, IEEE Trans. Pattern Anal. Mach. Intell., 26 (2004), pp. 982–994.

[17] D. Q. Dai and P. C. Yuen, *Regularized discriminant analysis and its application to face recognition*, Pattern Recognition, 36 (2003), pp. 845–847.

[18] P. Howland, M. Jeon, and H. Park, *Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 165–179.

[19] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, *Face recognition using LDA based algorithms*, IEEE Trans. Neural Netw., 14 (2003), pp. 195–200.

[20] P. Baldi and G. W. Hatfield, *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*, Cambridge University Press, Cambridge, UK, 2002.

[21] S. Dudoit, J. Fridlyand, and T. P. Speed, *Comparison of discrimination methods for the classification of tumors using gene expression data*, J. Amer. Statist. Assoc., 97 (2002), pp. 77–87.

[22] R. Huang, Q. Liu, H. Lu, and S. Ma, *Solving the small sample size problem of LDA*, in Proceedings of the International Conference on Pattern Recognition, 2002, pp. 29–32.

[23] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, New York, 2002.

[24] R. Q. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., Wiley, New York, 2001.

[25] Z. Jin, J. Y. Yang, Z. S. Hu, and Z. Lou, *Face recognition based on the uncorrelated discriminant transformation*, Pattern Recognition, 34 (2001), pp. 1405–1416.

[26] Z. Jin, J. Y. Yang, Z. M. Tang, and Z. S. Hu, *A theorem on the uncorrelated optimal discriminant vectors*, Pattern Recognition, 34 (2001), pp. 2041–2047.

[27] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, *Multiclass linear dimension reduction by weighted pairwise Fisher criteria*, IEEE Trans. Pattern Anal. Mach. Intell., 23 (2001), pp. 762–766.

[28] L. Chen, H. M. Liao, M. Ko, J. Lin, and G. Yu, *A new LDA-based face recognition system which can solve the small sample size problem*, Pattern Recognition, 33 (2000), pp. 1713–1726.

[29] R. Lotlikar and R. Kothari, *Fractional-step dimensionality reduction*, IEEE Trans. Pattern Anal. Mach. Intell., 22 (2000), pp. 623–627.

[30] G. Kowalski, *Information Retrieval Systems: Theory and Implementation*, Kluwer Academic Publishers, Norwell, MA, 1997.

[31] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[32] D. L. Swets and J. Weng, *Using discriminant eigenfeatures for image retrieval*, IEEE Trans. Pattern Anal. Mach. Intell., 18 (1996), pp. 831–836.

[33] M. W. Berry, S. T. Dumais, and G. W. O'Brie, *Using linear algebra for intelligent information retrieval*, SIAM Rev., 37 (1995), pp. 573–595.

[34] W. B. Frakes and R. Baeza-Yates, *Information Retrieval: Data Structures and Algorithms*, Prentice–Hall, Englewood Cliffs, NJ, 1992.

[35] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 1992.

[36] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, New York, 1990.

[37] J. H. Friedman, *Regularized discriminant analysis*, J. Amer. Statist. Assoc., 84 (1989), pp. 165–175.

[38] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice–Hall, Englewood Cliffs, NJ, 1988.

[39] L. Duchene and S. Leclerq, *An optimal transformation for discriminant and principal component analysis*, IEEE Trans. Pattern Anal. Mach. Intell., 10 (1988), pp. 978–983.

[40] J. Ye, Q. Li, H. Xiong, H. Park, R. Janardan, and V. Kumar, *IDR/QR: An incremental dimension reduction algorithm via QR decomposition*, IEEE Trans. Knowl. Data Eng., 17 (2005), pp. 1208–1221.

[41] J. Ye and T. Xiong, *Null space versus orthogonal linear discriminant analysis*, in Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

Downloaded 11/20/12 to 147.8.230.103. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.