Journal of Machine Learning Research 6 (2005) 483–502 Submitted 9/04; Revised 1/05; Published 4/05

Characterization of a Family of Algorithms for Generalized Discriminant Analysis on Undersampled Problems

Jieping Ye JIEPING@CS.UMN.EDU
Department of Computer Science
University of Minnesota
Minneapolis, MN 55455, USA

Editor: Bin Yu

Abstract

A generalized discriminant analysis based on a new optimization criterion is presented. The criterion extends the optimization criteria of the classical Linear Discriminant Analysis (LDA) when the scatter matrices are singular. An efficient algorithm for the new optimization problem is presented.

The solutions to the proposed criterion form a family of algorithms for generalized LDA, which can be characterized in a closed form. We study two specific algorithms, namely Uncorrelated LDA (ULDA) and Orthogonal LDA (OLDA). ULDA was previously proposed for feature extraction and dimension reduction, whereas OLDA is a novel algorithm proposed in this paper. The features in the reduced space of ULDA are uncorrelated, while the discriminant vectors of OLDA are orthogonal to each other. We have conducted a comparative study on a variety of real-world data sets to evaluate ULDA and OLDA in terms of classification accuracy.

Keywords: dimension reduction, linear discriminant analysis, uncorrelated LDA, orthogonal LDA, singular value decomposition

1. Introduction

Many machine learning and data mining problems involve data in very high-dimensional spaces. We consider dimension reduction of high-dimensional, undersampled data, where the data dimension is much larger than the sample size. The high-dimensional, undersampled problems frequently occur in many applications including information retrieval (Berry et al., 1995; Deerwester et al., 1990), face recognition (Belhumeur et al., 1997; Swets and Weng, 1996; Turk and Pentland, 1991) and microarray data analysis (Dudoit et al., 2002).

Linear Discriminant Analysis (LDA) is a classical statistical approach for feature extraction and dimension reduction (Duda et al., 2000; Fukunaga, 1990; Hastie et al., 2001). LDA computes the optimal transformation (projection), which minimizes the within-class distance (of the data set) and maximizes the between-class distance simultaneously, thus achieving maximum discrimination. The optimal transformation can be readily computed by applying an eigen-decomposition on the scatter matrices of the given training data set. However classical LDA requires the total scatter matrix to be nonsingular. In many applications such as information retrieval, face recognition, and microarray data analysis, all scatter matrices in question can be singular since the data points are from a very high-dimensional space and in general the sample size does not exceed this dimension. This is known as the *singularity* or *undersampled* problems (Krzanowski et al., 1995).

In recent years, many approaches have been brought to bear on such high-dimensional, undersampled problems, including PCA+LDA (Belhumeur et al., 1997; Swets and Weng, 1996; Zhao et al., 1999), Regularized LDA (Friedman, 1989), Penalized LDA (Hastie et al., 1995), Pseudo-inverse LDA (Fukunaga, 1990; Raudys and Duin, 1998; Skurichina and Duin, 1996, 1999), and LDA/GSVD (Howland et al., 2003; Ye et al., 2004b). More details will be given in Section 2.

1.1 Contribution

In this paper, we present a new optimization criterion for discriminant analysis, which is applicable to undersampled problems. A detailed mathematical derivation for the proposed optimization problem is presented in Section 3.

The solutions to the proposed criterion characterize a family of algorithms for generalized LDA. Among the family of algorithms, we study two specific ones in detail, namely Uncorrelated LDA (ULDA) and Orthogonal LDA (OLDA). ULDA was developed in the past for feature extraction and dimension reduction, whereas OLDA is a novel LDA based algorithm proposed in this paper.

ULDA was recently proposed for extracting feature vectors with uncorrelated attributes (Jin et al., 2001a,b). A more recent work (Ye et al., 2004a) showed 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 is nonsingular. Based on this equivalence, an efficient algorithm was presented in (Ye et al., 2004a) for computing the optimal discriminant vectors of ULDA. Interestingly, the solution in (Ye et al., 2004a) is a special case of the solutions to the proposed criterion in this paper (See Section 4).

OLDA is a novel dimension reduction algorithm proposed in this paper. The key property of OLDA is that the discriminant vectors of OLDA are orthogonal to each other, i.e., the transformation matrix of OLDA is orthogonal. There has been some early development on LDA based algorithms with orthogonal transformations. The algorithm is known as Foley-Sammon LDA (FSLDA). FSLDA was first proposed by Foley and Sammon for two-class problems (Foley and Sammon, 1975). It was then extended to the multi-class problems by Duchene and Leclercq (Duchene and Leclerq, 1988). The OLDA algorithm proposed in this paper provides an alternative, but simple and efficient way for computing orthogonal transformations in the framework of LDA.

We have conducted a comparative study on a variety of real-world data sets, including text documents, face images, and gene expression data to evaluate ULDA and OLDA, and compare with Regularized LDA (RLDA). Results have shown that OLDA is competitive with ULDA and RLDA in terms of classification accuracy.

The main contributions of this paper include:

- A generalization of the classical discriminant analysis to small sample size data using a new criterion, where the nonsingularity of the scatter matrices is not required;
- Mathematical derivation of the solutions to the new optimization criterion, based on the simultaneous diagonalization of the scatter matrices;
- Characterization of a family of algorithms for generalized LDA based on the proposed criterion and derivation of two specific algorithms, namely ULDA and OLDA.

1.2 Organization

The rest of the paper is organized as follows: We review classical LDA and several extensions in Section 2. A generalization of classical LDA using the new criterion is presented in Section 3. Two specific solutions to the proposed criterion, namely ULDA and OLDA, are discussed in Section 4. Experimental results are presented in Section 5. Finally, concluding discussions and future directions are presented in Section 6.

1.3 Notation

For convenience, we present in Table 1 the important notations used in the rest of the paper.

Notation	Description	Notation	Description
n	sample size	m	number of variables (dimensions)
k	number of classes	A	data matrix
A_i	data matrix of the i-th class	n_i	size of the i-th class
$c^{(i)}$	centroid of the i-th class	c	global centroid of the training set
S_b	between-class scatter matrix	S_w	within-class scatter matrix
S_t	total scatter matrix	G	transformation matrix
q	rank of the matrix S_b	t	rank of the matrix S_t

Table 1: Important notations used in the paper

2. Classical Discriminant Analysis

Given a data matrix $A \in \mathbb{R}^{m\times n}$, classical linear discriminant analysis computes a linear transformation $G \in \mathbb{R}^{m\times \ell}$ that maps each column a_i of A in the m-dimensional space to a vector y_i in the ℓ-dimensional space:

$$G : a_i \in \mathbb{R}^m \rightarrow y_i = G^T a_i \in \mathbb{R}^\ell \, (\ell < m).$$

Assume the original data is already clustered and ordering is imposed on the samples based on cluster membership. The goal of classical LDA is to find a transformation G such that the cluster structure of the original high-dimensional space is preserved in the reduced-dimensional space. Let the data matrix A be partitioned into k classes as $A = [A_1, \cdots, A_k]$, where $A_i \in \mathbb{R}^{m\times n_i}$, and $\sum_{i=1}^k n_i = n$.

In discriminant analysis (Fukunaga, 1990), three scatter matrices, called *within-class*, *between-class* and *total* scatter matrices are defined as follows:

$$\begin{aligned}
S_w &= \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in A_i}(x-c^{(i)})(x-c^{(i)})^T, \\
S_b &= \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in A_i}(c^{(i)}-c)(c^{(i)}-c)^T = \frac{1}{n}\sum_{i=1}^{k} n_i (c^{(i)}-c)(c^{(i)}-c)^T, \\
S_t &= \frac{1}{n}\sum_{j=1}^{n}(a_j-c)(a_j-c)^T, \qquad (1)
\end{aligned}$$

where the *centroid* $c^{(i)}$ of the i-th class is defined as $c^{(i)} = \frac{1}{n_i} A_i e^{(i)}$ with

$$e^{(i)} = (1, 1, \cdots, 1)^T \in \mathbb{R}^{n_i},$$

and the *global centroid* c is defined as $c = \frac{1}{n} Ae$ with

$$e = (1, 1, \cdots, 1)^T \in \mathbb{R}^n.$$

It is easy to verify that $S_t = S_b + S_w$.

Define the matrices

$$\begin{aligned} H_w &= \frac{1}{\sqrt{n}}[A_1 - c^{(1)}(e^{(1)})^T, \cdots, A_k - c^{(k)}(e^{(k)})^T], \\ H_b &= \frac{1}{\sqrt{n}}[\sqrt{n_1}(c^{(1)} - c), \cdots, \sqrt{n_k}(c^{(k)} - c)], \\ H_t &= \frac{1}{\sqrt{n}}(A - ce^T). \end{aligned} \tag{2}$$

Then S_w, S_b, and S_t can be expressed as

$$S_w = H_w H_w^T, \quad S_b = H_b H_b^T, \quad S_t = H_t H_t^T.$$

The *traces* of the two scatter matrices S_w and S_b can be computed as follows:

$$\begin{aligned} \text{trace}(S_w) &= \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in A_i}(x - c^{(i)})^T(x - c^{(i)}) = \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in A_i}||x - c^{(i)}||^2 \\ \text{trace}(S_b) &= \frac{1}{n}\sum_{i=1}^{k} n_i (c^{(i)} - c)^T(c^{(i)} - c) = \frac{1}{n}\sum_{i=1}^{k} n_i ||c^{(i)} - c||^2. \end{aligned} \tag{3}$$

Hence, $\text{trace}(S_w)$ measures the within-class cohesion, while $\text{trace}(S_b)$ measures the between-class separation.

In the lower-dimensional space resulting from the linear transformation G, the scatter matrices become

$$S_w^L = G^T S_w G, \quad S_b^L = G^T S_b G, \quad S_t^L = G^T S_t G. \tag{4}$$

An optimal transformation G would maximize $\text{trace}(S_b^L)$ and minimize $\text{trace}(S_w^L)$ simultaneously, which is equivalent to maximizing $\text{trace}(S_b^L)$ and minimizing $\text{trace}(S_t^L)$ simultaneously, since $S_t^L = S_w^L + S_b^L$. A common optimization in classical discriminant analysis (Fukunaga, 1990) is

$$G = \arg\max_G \left\{ \text{trace}((S_t^L)^{-1} S_b^L) \right\}. \tag{5}$$

The optimization problem in Eq. (5) is equivalent to finding all the eigenvectors that satisfy $S_b x = \lambda S_t x$, for $\lambda \neq 0$ (Fukunaga, 1990). The solution can be obtained by applying an eigen-decomposition on the matrix $S_t^{-1} S_b$, if S_t is nonsingular. There are at most $k-1$ eigenvectors corresponding to nonzero eigenvalues, since the rank of the matrix S_b is bounded from above by $k-1$. Therefore, the reduced dimension by classical LDA is at most $k-1$. A stable way to solve this eigen-decomposition problem is to apply Singular Value Decomposition (SVD) (Golub and Loan, 1996) on the scatter matrices. Details can be found in (Swets and Weng, 1996).

Assuming normal distribution for each class with the common covariance matrix, classification based on maximum likelihood estimation results in a nearest class centroid rule, where the distance is measured in terms of the within-class Mahalanobis distance (Hastie et al., 2001). Assuming equal prior for all classes for simplicity, a test point h is classified as class j if

$$(h-c^{(j)})^T S_w^{-1}(h-c^{(j)}) \tag{6}$$

is minimized over $j = 1,\cdots,k$. It was shown in (Hastie et al., 1995) that

$$\arg\min_j\{(h-c^{(j)})^T S_w^{-1}(h-c^{(j)})\} = \arg\min_j\{||G^T(h-c^{(j)})||^2\}, \tag{7}$$

where G is the optimal transformation solving the optimization problem in Eq. (5). Thus, classical LDA is equivalent to maximum likelihood classification assuming normal distribution for each class with the common covariance matrix. When the dimension m is much larger than the number of classes k, classification using the reduced representation, i.e., classification based on $\arg\min_j\{G^T(h-c^{(j)})\}$ may give considerable savings (Hastie et al., 1995).

Although relying on heavy assumptions which are not true in many applications, LDA has been proven to be effective. This is mainly due to the fact that a simple, linear model is more robust against noise, and most likely will not overfit. Generalization of LDA by fitting Gaussian mixtures to each class has been studied in (Hastie and Tibshirani, 1996).

Note that classical discriminant analysis requires the total scatter matrix S_t to be nonsingular, which may not hold for undersampled data. Several extensions, including two-stage PCA+LDA, Regularized LDA, Penalized LDA, Pseudo-inverse LDA, and LDA/GSVD were proposed in the past to deal with the singularity problems as follows.

A common way to deal with the singularity problems is to apply an intermediate dimension reduction stage such as PCA to reduce the dimension of the original data before classical LDA is applied. The algorithm is known as PCA+LDA (Belhumeur et al., 1997; Swets and Weng, 1996; Zhao et al., 1999). In this two-stage PCA+LDA algorithm, the discriminant stage is preceded by a dimension reduction stage using PCA. The dimension of the subspace transformed by PCA is chosen such as the "reduced" total scatter matrix in the subspace is nonsingular, so that classical LDA can be applied. A limitation of this approach is that the optimal value of the reduced dimension for PCA is difficult to determine. Moreover, the PCA stage may lose some useful information for discrimination.

A simple way to deal with the singularity of S_t is to apply the idea of regularization, by adding some constant values to the diagonal elements of S_t, as $S_t + \mu I_m$, for some $\mu > 0$, where I_m is an identity matrix. It is easy to verify that $S_t + \mu I_m$ is positive definite, hence nonsingular. This approach is called Regularized LDA, or RLDA in short (Friedman, 1989). Regularization is a key in the theory of splines (Wahba, 1998) and is used widely in machine learning, such as Support Vector Machines (SVM) (Vapnik, 1998). It is evident that when $\mu \to \infty$, we lose the information on S_t, while very small values of μ may not be sufficiently effective. Cross-validation is commonly applied for estimating the optimal μ. More recent studies on RLDA can be found in (Dai and Yuen, 2003; Krzanowski et al., 1995).

The Penalized LDA (PLDA) is more general than Regularized LDA. PLDA penalizes the within-class scatter matrix as $S_w + \Omega$, for some penalty matrix Ω. Ω is symmetric and positive semidefinite. The penalties are designed to produce smoothness in the discriminant functions. Details on PLDA and the choices of penalties for different applications refer to (Hastie et al., 1995).

Pseudo-inverse is commonly applied to deal with the singularity problems, which is equivalent to approximating the solution using a least-squares solution method. The use of pseudo-inverse in discriminant analysis has been studied in the past. The *Pseudo Fisher Linear Discriminant* (PFLDA) (Fukunaga, 1990; Raudys and Duin, 1998; Skurichina and Duin, 1996, 1999) is based on the pseudo-inverse of the scatter matrices. The generalization error of PFLDA was studied in (Skurichina and Duin, 1996), when the size and dimension of the training data vary. Pseudo-inverses of the scatter matrices were also studied in (Krzanowski et al., 1995). Experiments in (Krzanowski et al., 1995) showed that the pseudo-inverse based methods are competitive with RLDA and PCA+LDA.

The LDA/GSVD algorithm (Howland et al., 2003; Ye et al., 2004b) is a more recent approach. The main technique applied is the Generalized Singular Value Decomposition (GSVD) (Golub and Loan, 1996). The criterion F_0 used in (Ye et al., 2004b) is:

$$F_0(G) = \text{trace}\left((S_b^L)^+ S_w^L\right), \tag{8}$$

where $(S_b^L)^+$ denotes the pseudo-inverse of the between-class scatter matrix. The definition of pseudo-inverse, as well as its computation via SVD, can be found in Appendix A.

LDA/GSVD aims to find the optimal transformation G that minimizes $F_0(G)$, subject to the constraint that $\text{rank}(G^T H_b) = q$, where q is the rank of S_b. The above constraint is enforced to preserve the dimension of the spaces spanned by the centroids in the original and transformed spaces. The optimal solution can be obtained by applying the GSVD. One limitation of this method is the high computational cost of GSVD, especially for large and high-dimensional data sets.

An overview of LDA on undersampled problems can be found in (Krzanowski et al., 1995).

The current paper focuses on linear discriminant analysis, which applies linear decision boundary. Discriminant analysis can also be studied in the non-linear fashion, so-called kernel discriminant analysis, by using the kernel trick (Schökopf and Smola, 2002). It is desirable if the data has weak linear separability. The interested readers can find more details on kernel discriminant analysis in (Baudat and Anouar, 2000; Hand, 1982; Lu et al., 2003; Schökopf and Smola, 2002).

3. Generalization of Discriminant Analysis

Classical discriminant analysis solves an eigen-decomposition problem when S_t is nonsingular. For undersampled problems, S_t is singular, since the sample size n may be smaller than its dimension m. In this section, we define a new criterion F_1, where the nonsingularity of S_t is not required.

The new criterion F_1 is a natural extension of the classical one in Eq. (5), where the inverse of a matrix is replaced by the pseudo-inverse (Golub and Loan, 1996). While the inverse of a matrix may not exist, the pseudo-inverse of any matrix is well defined. Moreover, when the matrix is invertible, its pseudo-inverse coincides with its inverse.

The new criterion F_1 is defined as

$$F_1(G) = \text{trace}\left((S_t^L)^+ S_b^L\right). \tag{9}$$

The optimal transformation matrix G is computed so that $F_1(G)$ is maximized. Note that in the following, the matrix G in $F_1(G)$ may be omitted if it is clear from the content.

In the rest of this section, we show how to solve the above maximization problem. It is based on the simultaneous diagonalization of the three scatter matrices. Details are given below.

3.1 Simultaneous Diagonalization of Scatter Matrices

In this section, we take a closer look at the relationship among three scatter matrices S_b, S_w, and S_t, and show how to diagonalize them simultaneously.

Let $H_t = U\Sigma V^T$ be the SVD of H_t, where H_t is defined in Eq. (2), U and V are orthogonal, $\Sigma = \begin{pmatrix} \Sigma_t & 0 \\ 0 & 0 \end{pmatrix}$, $\Sigma_t \in \mathbb{R}^{t\times t}$ is diagonal, and $t = \text{rank}(S_t)$. Then

$$S_t = H_t H_t^T = U\Sigma V^T V\Sigma^T U^T = U\Sigma\Sigma^T U^T = U \begin{pmatrix} \Sigma_t^2 & 0 \\ 0 & 0 \end{pmatrix} U^T. \tag{10}$$

Let $U = (U_1, U_2)$ be a partition of U, such that $U_1 \in \mathbb{R}^{m\times t}$ and $U_2 \in \mathbb{R}^{m\times(m-t)}$. Since $S_t = S_b + S_w$, we have

$$\begin{aligned} \begin{pmatrix} \Sigma_t^2 & 0 \\ 0 & 0 \end{pmatrix} &= U^T(S_b + S_w)U \\ &= \begin{pmatrix} U_1^T \\ U_2^T \end{pmatrix} S_b(U_1, U_2) + \begin{pmatrix} U_1^T \\ U_2^T \end{pmatrix} S_w(U_1, U_2) \\ &= \begin{pmatrix} U_1^T S_b U_1 & U_1^T S_b U_2 \\ U_2^T S_b U_1 & U_2^T S_b U_2 \end{pmatrix} + \begin{pmatrix} U_1^T S_w U_1 & U_1^T S_w U_2 \\ U_2^T S_w U_1 & U_2^T S_w U_2 \end{pmatrix}. \end{aligned} \tag{11}$$

It follows that $U_2^T S_b U_2 + U_2^T S_w U_2 = 0$. Therefore, $U_2^T S_b U_2 = 0$ and $U_2^T S_w U_2 = 0$, since both are positive semidefinite. We thus have $U_1^T S_b U_2 = 0$ and $U_1^T S_w U_2 = 0$, since both matrices on the right hand size of Eq. (11) are positive semidefinite. That is,

$$U^T S_b U = \begin{pmatrix} U_1^T S_b U_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad U^T S_w U = \begin{pmatrix} U_1^T S_w U_1 & 0 \\ 0 & 0 \end{pmatrix}. \tag{12}$$

From Eq. (11) and Eq. (12), we have $\Sigma_t^2 = U_1^T S_b U_1 + U_1^T S_w U_1$. It follows that

$$I_t = \Sigma_t^{-1} U_1^T S_b U_1 \Sigma_t^{-1} + \Sigma_t^{-1} U_1^T S_w U_1 \Sigma_t^{-1}. \tag{13}$$

Denote $B = \Sigma_t^{-1} U_1^T H_b$ and let $B = P\tilde{\Sigma}Q^T$ be the SVD of B, where P and Q are orthogonal and $\tilde{\Sigma}$ is diagonal. Then

$$\Sigma_t^{-1} U_1^T S_b U_1 \Sigma_t^{-1} = P\tilde{\Sigma}^2 P^T = P\Sigma_b P^T,$$

where

$$\Sigma_b \equiv \tilde{\Sigma}^2 = \text{diag}(\lambda_1, \cdots, \lambda_t),$$
$$\lambda_1 \geq \cdots \geq \lambda_q > 0 = \lambda_{q+1} = \cdots = \lambda_t,$$

and $q = \text{rank}(S_b)$.

It follows from Eq. (13) that

$$I_t = \Sigma_b + P^T \Sigma_t^{-1} U_1^T S_w U_1 \Sigma_t^{-1} P.$$

Hence

$$P^T \Sigma_t^{-1} U_1^T S_w U_1 \Sigma_t^{-1} P = I_t - \Sigma_b \equiv \Sigma_w$$

is also diagonal.

Combining all these together, we have

$$X^T S_b X = \begin{pmatrix} \Sigma_b & 0 \\ 0 & 0 \end{pmatrix} \equiv D_b, \quad X^T S_w X = \begin{pmatrix} \Sigma_w & 0 \\ 0 & 0 \end{pmatrix} \equiv D_w, \quad X^T S_t X = \begin{pmatrix} I_t & 0 \\ 0 & 0 \end{pmatrix} \equiv D_t, \quad (14)$$

where

$$X = U \begin{pmatrix} \Sigma_t^{-1} P & 0 \\ 0 & I \end{pmatrix}. \quad (15)$$

In summary, the matrix X in Eq. (15) simultaneously diagonalizes S_b, S_w, and S_t.

3.2 Maximization of the F_1 Criterion

In this section, we derive the generalized discriminant analysis by maximizing the F_1 criterion defined in Eq. (9). The main technique applied is the simultaneous diagonalization of scatter matrices from last section. We show in this section that the solutions to the proposed criterion F_1 can be characterized as $G = X_q M$, where X_q is the matrix consisting of the first q columns of X, defined in Eq. (15), $q = \text{rank}(S_b)$, and $M \in \mathbb{R}^{q \times q}$ is an arbitrary nonsingular matrix.

We first present two lemmas. The proof of Lemma 3.1 is straightforward from standard linear algebra and a generalization of Lemma 3.2 can be found in (Edelman et al., 1998).

Lemma 3.1 *For any matrix $A \in \mathbb{R}^{m \times n}$, the following equality holds: $(A^T A)^+ = A^+ (A^+)^T$.*

Lemma 3.2 *Let $A \in \mathbb{R}^{m \times m}$ be symmetric and positive semidefinite and let x_i be the eigenvector of A corresponding to the i-th largest eigenvalue λ_i. Then, for any $M \in \mathbb{R}^{m \times s} (s \leq m)$ with orthonormal columns, the following inequality holds,*

$$\textit{trace}\left(M^T A M\right) \leq \lambda_1 + \cdots + \lambda_s,$$

where the equality holds if $M = [x_1, \cdots, x_s]Q$, for any orthogonal matrix $Q \in \mathbb{R}^{s \times s}$.

The main result of this section is summarized in the following theorem.

Theorem 3.1 *Let X be the matrix defined in Eq. (15) and X_q be the matrix consisting of the first q columns of X, where $q = \textit{rank}(S_b)$. Then $G = X_q M$, for any nonsingular M, maximizes F_1 defined in Eq. (9).*

Proof By the simultaneous diagonalization of the three scatter matrices in Eq. (14), we have

$$\begin{aligned} S_b^L &= G^T S_b G = G^T (X^{-1})^T (X^T S_b X) X^{-1} G = \tilde{G}^T D_b \tilde{G}, \\ S_t^L &= G^T S_t G = G^T (X^{-1})^T (X^T S_t X) X^{-1} G = \tilde{G}^T D_t \tilde{G}, \end{aligned} \quad (16)$$

where $\tilde{G} = X^{-1} G$.

Let $\tilde{G} = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}$ be a partition of $\tilde{G}$ so that $G_1 \in \mathbb{R}^{t \times \ell}$ and $G_2 \in \mathbb{R}^{(m-t) \times \ell}$. It follows that

$$S_b^L = \tilde{G}^T D_b \tilde{G} = G_1^T \Sigma_b G_1, \quad S_t^L = \tilde{G}^T D_t \tilde{G} = G_1^T G_1.$$

Hence

$$F_1 = \text{trace}\left((G_1^T G_1)^+ (G_1^T \Sigma_b G_1)\right) = \text{trace}\left((G_1 G_1^+)^T \Sigma_b (G_1 G_1^+)\right),$$

where the second equality follows from Lemma 3.1.

Recall that $\Sigma_b = \mathrm{diag}(\lambda_1, \cdots, \lambda_t)$, where $\lambda_1 \geq \cdots \geq \lambda_q > 0 = \lambda_{q+1} = \cdots = \lambda_t$. Let $G_1 = R \begin{pmatrix} \Sigma_\delta & 0 \\ 0 & 0 \end{pmatrix} S^T$ be the SVD of G_1, where R and S are orthogonal, Σ_δ is diagonal, and $\delta = \mathrm{rank}(G_1)$. Then $G_1^+ = S \begin{pmatrix} \Sigma_\delta^{-1} & 0 \\ 0 & 0 \end{pmatrix} R^T$, and $G_1 G_1^+ = R \begin{pmatrix} I_\delta & 0 \\ 0 & 0 \end{pmatrix} R^T$. It follows that

$$\begin{aligned} F_1 &= \mathrm{trace}\left((G_1 G_1^+)^T \Sigma_b (G_1 G_1^+)\right) = \mathrm{trace}\left(R \begin{pmatrix} I_\delta & 0 \\ 0 & 0 \end{pmatrix} R^T \Sigma_b R \begin{pmatrix} I_\delta & 0 \\ 0 & 0 \end{pmatrix} R^T \right) \\ &= \mathrm{trace}\left(\begin{pmatrix} I_\delta & 0 \\ 0 & 0 \end{pmatrix} R^T \Sigma_b R \begin{pmatrix} I_\delta & 0 \\ 0 & 0 \end{pmatrix} \right) = \mathrm{trace}\left(R_\delta^T \Sigma_b R_\delta\right) \leq \lambda_1 + \cdots + \lambda_q. \end{aligned}$$

where R_δ is the matrix consisting of the first δ columns of R, and the last inequality follows from Lemma 3.2. By Lemma 3.2 again, the above inequality becomes equality, if $R_\delta = \begin{pmatrix} W \\ 0 \end{pmatrix}$, for any orthogonal $W \in \mathbb{R}^{q \times q}$, $\delta = q$, and $\ell = q$. Under this choice of R_δ,

$$G_1 = R_q \Sigma_q S^T = \begin{pmatrix} W \Sigma_q S^T \\ 0 \end{pmatrix}.$$

We observe that the maximization of F_1 is independent of G_2, and simply set it to zero. Therefore, the maximum of F_1 is attained when

$$\tilde{G} = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} = \begin{pmatrix} W \Sigma_q S^T \\ 0 \end{pmatrix}.$$

Note that the orthogonal matrices W and S, and the diagonal matrix Σ_q are arbitrary. Hence, $M = W \Sigma_q S^T$ is an arbitrary nonsingular matrix. It follows that $G = X\tilde{G} = X_q M$, for any nonsingular M, maximizes F_1. This completes the proof of the theorem. ■

Remark 1 *Note that it is in general not true that $F_1(H) = F_1(HM)$, for any nonsingular M. However, Theorem 3.1 implies that for $H = X_q$, we have $F_1(H) = F_1(HM)$, for any nonsingular M.*

4. Uncorrelated LDA Versus Orthogonal LDA

From last section, $G = X_q M$, for any nonsingular M maximizes the F_1 criterion. A natural question is: How to choose the best M? In this section, we consider two specific choices of M, which lead to two distinct algorithms: Uncorrelated LDA and Orthogonal LDA.

4.1 Uncorrelated LDA

The simplest choice of M is the identity matrix, i.e., $M = I_q$. That is, $G = X_q$. It follows that $X_q^T S_t X_q = I_q$, i.e., the columns of the transformation G are S_t-orthogonal. Recall that two vectors x and y are S_t-orthogonal, if $x^T S_t y = 0$. The solution corresponds to the Uncorrelated LDA, originally proposed by Jin et al. (Jin et al., 2001a,b). The pseudo-code for ULDA is given in **Algorithm 1**.

Algorithm 1: Uncorrelated LDA
Input: data matrix A
Output: transformation matrix G
1. Form three matrices H_b, H_w, and H_t as in Eq. (2);
2. Compute reduced SVD of H_t as $H_t = U_1 \Sigma_t V_1^T$;
3. $B \leftarrow \Sigma_t^{-1} U_1^T H_b$;
4. Compute SVD of B as $B = P \Sigma Q^T$; $q \leftarrow \text{rank}(B)$;
5. $X \leftarrow U_1 \Sigma_t^{-1} P$;
6. $G \leftarrow X_q$;

ULDA was originally proposed to compute the optimal discriminant vectors that are S_t-orthogonal. Specifically, suppose r vectors $\phi_1, \phi_2, \cdots, \phi_r$ are obtained, then the $(r+1)$-th vector ϕ_{r+1} of ULDA is the one that maximizes the Fisher criterion function

$$f(\phi) = \frac{\phi^T S_b \phi}{\phi^T S_w \phi}, \tag{17}$$

subject to the constraints:

$$\phi_{r+1}^T S_t \phi_i = 0, \quad i = 1, \cdots, r.$$

The algorithm in (Jin et al., 2001a) finds ϕ_i successively as follows: The j-th discriminant vector ϕ_j of ULDA is the eigenvector corresponding to the maximum eigenvalue of the following generalized eigenvalue problem:

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

where

$$\begin{aligned} U_1 &= I_m, \\ U_j &= I_m - S_t D_j^T (D_j S_t S_w^{-1} S_t D_j^T)^{-1} D_j S_t S_w^{-1} (j > 1), \\ D_j &= [\phi_1, \cdots, \phi_{j-1}]^T (j > 1), \end{aligned}$$

and I_m is the identity matrix.

A key property of ULDA is that the features in the reduced space are uncorrelated to each other, as stated in the following proposition.

Proposition 4.1 *Let the transformation matrix for ULDA be $G = [g_1, \cdots, g_d]$, for some $d > 0$. The original feature vector A is transformed into $Z = G^T A$, where the i-th feature component of Z is $Z_i = g_i^T A$. Assume that g_i and g_j are S_t-orthogonal to each other, i.e., $g_i^T S_t g_j = 0$, for $i \neq j$. Then the correlation between Z_i and Z_j is 0, for $i \neq j$. That is, Z_i and Z_j are uncorrelated to each other.*

Proof The covariance between Z_i and Z_j can be computed as

$$\text{Cov}(Z_i, Z_j) = E(Z_i - EZ_i)(Z_j - EZ_j) = g_i^T \{E(A - EA)(A - EA)^T\} g_j = g_i^T S_t g_j. \tag{18}$$

Hence, their correlation coefficient is

$$\text{Cor}(Z_i, Z_j) = \frac{g_i^T S_t g_j}{\sqrt{g_i^T S_t g_i} \sqrt{g_j^T S_t g_j}}. \tag{19}$$

Since $g_i^T S_t g_j = 0$, for $i \neq j$, we have $\mathrm{Cor}(Z_i, Z_j) = 0$, for $i \neq j$. This completes the proof of the proposition. ■

In (Ye et al., 2004a), an efficient algorithm for ULDA was proposed, based on the following optimization problem:

$$G = \arg\max_G \{\mathrm{trac}((S_t^L + \mu I_\ell)^{-1} S_b^L)\}. \tag{20}$$

Note that $S_t^L + \mu I_\ell$ is always nonsingular for $\mu > 0$, since S_t^L is positive semidefinite. One key result in (Ye et al., 2004a) shows that the optimal transformation G solving the optimization problem in Eq. (20) is independent of μ.

Interestingly, it can be shown that $G = X_q$ solves the optimization problem in Eq. (20) as stated in the following proposition. Detailed proof follows the one in (Ye et al., 2004a) and is thus omitted.

Proposition 4.2 *Let $G = X_q$, where X_q is the matrix consisting of the first q columns of X, and X is defined in Eq. (15). Then G solves the optimization problem in Eq. (20).*

4.1.1 RELATIONSHIP BETWEEN ULDA AND THE EIGEN-DECOMPOSITION OF $S_t^+ S_b$

In this section, we study the relationship between ULDA and the eigen-decomposition of $S_t^+ S_b$. More specifically, we show that the discriminant vectors of ULDA are eigenvectors of $S_t^+ S_b$ corresponding to nonzero eigenvalues. Recall that classical LDA computes the optimal discriminant vectors by solving an eigenvalue problem on $S_t^{-1} S_b$, assuming S_t is nonsingular (See Section 2). This equivalence result shows that ULDA is a natural extension of classical LDA by replacing inverse with pseudo-inverse, when dealing with singular S_t.

From Eq. (14), we have $X^T S_t X = D_t$, where

$$X = U \begin{pmatrix} \Sigma_t^{-1} P & 0 \\ 0 & 0 \end{pmatrix}, \text{ and } D_t = \begin{pmatrix} I_t & 0 \\ 0 & 0 \end{pmatrix}.$$

Note that P is orthogonal. It follows that

$$S_t = X^{-T} D_t X^{-1} = U \begin{pmatrix} \Sigma_t P & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} I_t & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P^T \Sigma_t & 0 \\ 0 & 0 \end{pmatrix} U^T = U \begin{pmatrix} \Sigma_t^2 & 0 \\ 0 & 0 \end{pmatrix} U^T,$$

Hence,

$$S_t^+ = U \begin{pmatrix} \Sigma_t^{-2} & 0 \\ 0 & 0 \end{pmatrix} U^T.$$

It is easy to verify that

$$X D_t X^T = U \begin{pmatrix} \Sigma_t^{-2} & 0 \\ 0 & 0 \end{pmatrix} U^T.$$

It follows that

$$S_t^+ = X D_t X^T, \tag{21}$$

and

$$S_t^+ S_b = \left(X D_t X^T\right) \left(X^{-1} D_b X^{-1}\right) = X D_t D_b X^{-1}.$$

Therefore, the columns of X_q form the eigenvectors of $S_t^+ S_b$ corresponding to nonzero eigenvalues, since $D_t D_b$ is diagonal with q nonzero diagonal entries.

Algorithm 2: Orthogonal LDA

Input: data matrix A

Output: transformation matrix G

1. Compute the matrix X_q as in ULDA (Steps 1–5 of **Algorithm 1**);
2. Compute QR decomposition of X_q as $X_q = \tilde{Q}\tilde{R}$;
3. $G \leftarrow \tilde{Q}$;

4.2 Orthogonal LDA

LDA with orthogonal discriminant vectors is a natural alternative to ULDA. Let $X_q = \tilde{Q}\tilde{R}$ be the QR decomposition of X_q, then we can simply choose $M = \tilde{R}^{-1}$ so that the columns of $G = X_q M = \tilde{Q}$ are orthogonal to each other. The pseudo-code for OLDA is given in **Algorithm 2**.

Note that in the literature of LDA, Foley-Sammon LDA (FSLDA) is also known for its orthogonal discriminant vectors. FSLDA was first proposed by Foley and Sammon for two-class problems (Foley and Sammon, 1975). It was then extended to the multi-class problems by Duchene and Leclercq (Duchene and Leclerq, 1988). Specifically, suppose r vectors $\phi_1, \phi_2, \cdots, \phi_r$ are obtained, then the $(r+1)$-th vector ϕ_{r+1} of FSLDA is the one that maximizes the Fisher criterion function $f(\phi)$ defined in Eq. (17), subject to the constraints: $\phi_{r+1}^T \phi_i = 0$, $i = 1, \cdots, r$.

The algorithm in (Duchene and Leclerq, 1988) finds ϕ_i successively as follows: The j-th discriminant vector ϕ_j of FSLDA is the eigenvector corresponding to the maximum eigenvalue of the following matrix:

$$\left(I_m - S_w^{-1} D_j^T S_j^{-1} D_j\right) S_w^{-1} S_b,$$

where

$$D_j = [\phi_1, \cdots, \phi_{j-1}]^T (j > 1), \text{ and } S_j = D_j S_w^{-1} D_j^T.$$

The above FSLDA algorithm may be expensive for large and high-dimensional data sets. More details on the computation of FSLDA can be found in (Duchene and Leclerq, 1988).

It is worthwhile to point out that both ULDA and FSLDA use the same Fisher criterion function, and the main difference is that the optimal discriminant vectors generated by ULDA are S_t-orthogonal to each other, while the optimal discriminant vectors of FSLDA are orthogonal to each other.

The common point of the proposed OLDA algorithm and the FSLDA algorithm described above is that the transformation matrix has orthogonal columns. However, these two algorithms were derived from distinct perspectives.

4.3 Discussions

As discussed in Section 2, classical LDA is equivalent to maximum likelihood classification assuming normal distribution for each class with the common covariance matrix. Classification in classical LDA based on the maximum likelihood estimation is based on the Mahalanobis distance as follows: a test point h is classified as class j if

$$j = \arg\min_j (h - c^{(j)})^T S_w^{-1} (h - c^{(j)}), \tag{22}$$

which is equivalent to

$$j = \arg\min_j (h - c^{(j)})^T S_t^{-1} (h - c^{(j)}). \tag{23}$$

We show in the following that the classification in ULDA uses the following distance:

$$(h - c^{(j)})^T S_t^{+} (h - c^{(j)}). \tag{24}$$

The main result is summarized in the following theorem.

Theorem 4.1 *Let G be the optimal transformation matrix for ULDA, and let h be any test point. Then*

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

Proof Let X_i be the i-th column of X. From Eq. (21), we have

$$S_t^{+} = X D_t X^T = \sum_{i=1}^{t} X_i X_i^T = GG^T + \sum_{i=q+1}^{t} X_i X_i^T,$$

where G consists of the first q columns of X, and $q = \text{rank}(S_b)$.

Recall from Section 3.1 that X diagonalizes S_b and $X_i^T S_b X_i = 0$, for $i = q+1, \cdots, t$. Hence $H_b X_i = 0$, or $(c^{(j)})^T X_i = c X_i$, for all $j = 1, \cdots, k$. It follows that

$$\begin{aligned} (h - c^{(j)})^T S_t^{+} (h - c^{(j)}) &= (h - c^{(j)})^T GG^T (h - c^{(j)}) + \sum_{i=q+1}^{t} (h - c^{(j)})^T X_i X_i^T (h - c^{(j)}) \\ &= ||G^T (h - c^{(j)})||^2 + \sum_{i=q+1}^{t} (h - c)^T X_i X_i^T (h - c). \end{aligned} \tag{25}$$

The second term on the right hand side of Eq. (25) is independent of class j, hence

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

This completes the proof of the theorem.

■

Theorem 4.1 shows that the classification rule in ULDA is a variant of the one used in classical LDA. ULDA can be considered as an extension of classical LDA for singular scatter matrices. The result does not extend to OLDA. However, with whitened total scatter matrix, that is if S_t is an identity matrix, OLDA is equivalent to ULDA.

Geometrically, both ULDA and OLDA project the data onto the subspace spanned by the centroids. ULDA removes the correlation among the features in the transformed space, which is theoretically sound but may be sensitive to the noise in the data. On the other hand, OLDA applies orthogonal transformation $\tilde{Q}$, by factoring out the $\tilde{R}$ matrix through the QR decomposition of $X_q = \tilde{Q}\tilde{R}$. The removal of $\tilde{R}$ in OLDA may contribute to the noise removal. Our experiments in next section show that OLDA often leads to better performance than ULDA in classification.

Data Set	Size (n)	Dimension (m)	# of classes (k)
tr41	210	7454	7
re0	320	2887	4
PIX	300	10000	30
AR	1638	8888	126
GCM	198	16063	14
ALL	248	12558	6

Table 2: Statistics for our test data sets

5. Experiments

We divide the experiments into three parts. Section 5.1 describes our test data sets. Section 5.2 evaluates ULDA and OLDA in terms of classification accuracy. We study the effect of the matrix M in Section 5.3. Recall that $G = X_q M$, for any nonsingular M maximizes the F_1 criterion.

Both ULDA and OLDA were implemented in MATLAB and the source codes may be accessed at *http://www.cs.umn.edu/∼jieping/UOLDA*.

5.1 Data Sets

We have three types of data for the evaluation: text documents, including **tr41** and **re0**; face images, including **PIX** and **AR**; and gene expression data, including **GCM** and **ALL**. The important statistics of these data sets are summarized as follows (see also Table 2):

- **tr41** is a text document data set, derived from the TREC-5, TREC-6, and TREC-7 collections (TREC, 1999). It includes 210 documents belonging to 7 different classes. The dimension of this data set is 7454.
- **re0** is another text document data set, derived from *Reuters-21578* text categorization test collection Distribution 1.0 (Lewis, 1999). It includes 320 documents belonging to 4 different classes. The dimension of this data set is 2887.
- **PIX**[1] is a face image data set, which contains 300 face images of 30 persons. The size of PIX images is 512×512. We subsample the images down to a size of $100 \times 100 = 10000$.
- **AR**[2] (Martinez and Benavente, 1998), is a large face image data set. The instance of each face may contain pretty large areas of occlusion, due to the presence of sun glasses and scarves. We use a subset of AR. This subset contains 1638 face images of 126 individuals. Its image size is 768×576. We first crop the image from row 100 to 500, and column 200 to 550, and then subsample the cropped images down to a size of $101 \times 88 = 8888$.
- **GCM** is a gene expression data set consisting of 198 human tumor samples spanning fourteen different cancer types. The data set was first studied in (Ramaswamy and et al., 2001; Yeang and et al., 2001). The breakdown of the sample classes is as follows: 12 *breast* samples, 14

1. http://peipa.essex.ac.uk/ipa/pix/faces/manchester/test-hard/
2. http://rvl1.ecn.purdue.edu/∼aleix/aleix_face_DB.html

prostate samples, 12 *lung* samples, 12 *colorectal* samples, 22 *lymphoma* samples, 11 *bladder* samples, 10 *melanoma* samples, 10 *uterus* samples, 30 *leukemia* samples, 11 *renal* samples, 11 *pancreas* samples, 12 *ovary* samples, 11 *mesothelioma* samples, and 20 *CNS* samples.

- **ALL**[3] is another gene expression data set consisting of six diagnostic groups (Yeoh and et al., 2002). The breakdown of the samples is: 15 samples for *BCR*, 27 samples for *E2A*, 64 samples for *Hyperdip*, 20 samples for *MLL*, 43 samples for *T*, and 79 samples for *TEL*.

5.2 Comparison on Classification Accuracy

In this experiment, we evaluate ULDA and OLDA in terms of classification accuracy. For the **GCM** and **ALL** gene expression data sets, the test sets were provided. In the absence of original test sets, such as the two document data sets and the two face image data sets, we perform our comparative study by repeated random splitting into training and test sets exactly as in (Dudoit et al., 2002). The data were randomly partitioned into a training set consisting of two-thirds of the whole set and a test set consisting of one-third of the whole set. To reduce the variability, the splitting was repeated 50 times and the resulting accuracies were averaged. Note that during each run, dimension reduction is applied to the training set only. For RLDA, the results depend on the choice of the parameter μ. We choose the best μ through cross-validation. The range for μ is between 0.001 and 10.

The results of the three algorithms on the six data sets are presented in Table 3. The main observation from Table 3 is that OLDA is competitive with ULDA and RLDA in all six data sets. We also observe that in most cases, RLDA outperforms ULDA and is competitive with OLDA.

It is interesting to note that OLDA achieves higher accuracies than ULDA for the two face image data sets and two gene expression data sets, while it achieves accuracies close to those of ULDA for the two text document data sets. For the **GCM** gene expression data set, OLDA achieved classification accuracy 3% higher than that of OLDA. This may be related to the effect of the noise removal inherent in OLDA as discussed in Section 4.3.

5.3 Effect of the Matrix M

In this experiment, we study the effect of the matrix M on classification using the **GCM** and **ALL** data sets. Recall that the solution to the proposed criterion is $G = X_q M$, for any nonsingular M. Two specific choices of M were studied, which correspond to ULDA and OLDA. In this experiment, we randomly generated 100 matrices for M and computed the accuracies using the corresponding transformation matrices. Figure 1 shows the histogram of the resulting accuracies on **GCM**, where the x-axis represents the range of resulting accuracies (divided into small intervals), and the y-axis represents the number (count) for each interval. The main observations are:

- None of the accuracies is higher than those of ULDA (73.91%) and OLDA (76.09%). ULDA and OLDA are probably two of the best ones among the family of solutions to the proposed criterion.

- In Figure 1, most of the accuracies are around 55%, which is much lower than those of ULDA and OLDA. Thus, the choice of M does make a big difference. Among the family of solutions to the proposed criterion, most of them perform quite poorly in comparison to ULDA and OLDA.

3. http://www.stjuderesearch.org/data/ALL1/

Data Set	Accuracy		
	ULDA	OLDA	RLDA
tr41	96.69 ± 1.90	96.34 ± 2.10	96.23 ± 2.17
re0	86.26 ± 2.46	86.13 ± 2.58	87.34 ± 2.37
PIX	96.16 ± 2.48	98.00 ± 1.66	96.31 ± 2.20
AR	90.94 ± 0.96	92.77 ± 1.04	91.11 ± 1.02
GCM	73.91	76.09	78.26
ALL	98.82	100.0	98.82

Table 3: Comparison of classification accuracy and standard deviation of three algorithms: ULDA (Uncorrelated LDA), OLDA (Orthogonal LDA), and RLDA (Regularized LDA), on the six data sets. The mean and standard deviation of accuracies from fifty runs are reported for **tr41**, **re0**, **PIX**, and **AR**. Note that for the two gene expression data sets: **GCM** and **ALL**, we use the original test sets. Thus the standard deviation for these two data sets are not reported.

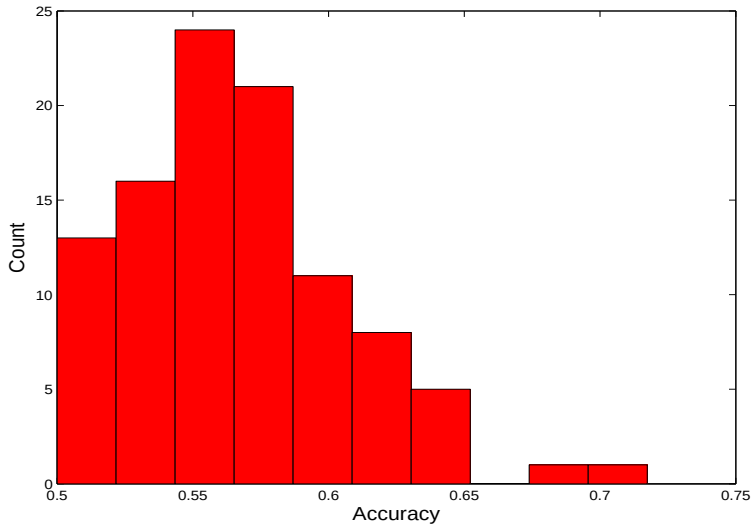

Figure 1: Effect of the matrix M using the **GCM** data set. The corresponding accuracies of ULDA and OLDA are 73.91% and 76.09%, respectively.

The result on **ALL** is shown in Figure 2. We can observe the same trend as in **GCM**, that is, most of the accuracies are much lower than those of ULDA and OLDA.

6. Conclusions and Future Directions

In this paper, a new optimization criterion for discriminant analysis is presented. The new criterion extends the optimization criteria of the classical LDA when the scatter matrices are singular. It

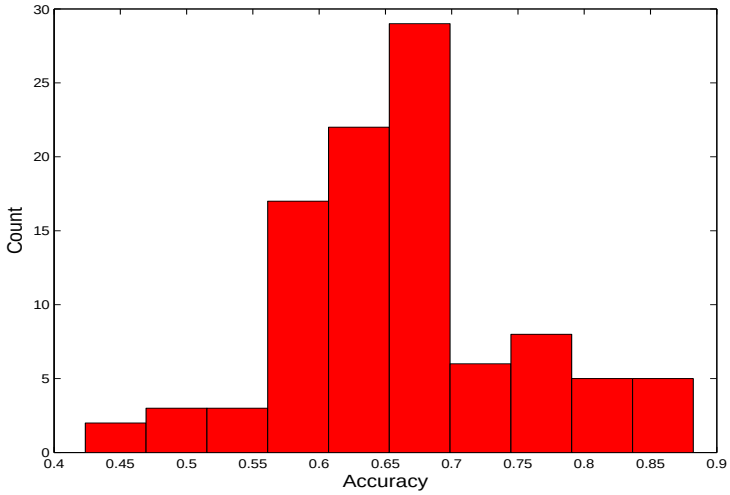

Figure 2: Effect of the matrix M using the **ALL** data set. The corresponding accuracies of ULDA and OLDA are 98.82% and 100.0%, respectively.

is applicable regardless of the relative sizes of the data dimension and sample size, overcoming a limitation of the classical LDA. A detailed mathematical derivation for the proposed optimization problem is presented. It is based on the simultaneous diagonalization of the three scatter matrices.

The solutions to the proposed criterion form a family of algorithms for generalized LDA, which can be characterized in a closed form. Among the family of solutions, we study two specific ones, namely ULDA and OLDA, where ULDA was previously proposed for feature extraction and dimension reduction and OLDA is a novel algorithm. ULDA has the property that the features in the reduced space are uncorrelated, while OLDA has the property that the discriminant vectors obtained are orthogonal to each other. Experiment on a variety of real-world data sets show that OLDA is competitive with ULDA and RLDA in terms of classification accuracy.

In this paper, we focus on two specific algorithms, ULDA and OLDA, for generalized LDA. A promising direction is to find algorithms with sparse transformation matrices. Sparsity has recently received much attention for extending Principal Component Analysis (d'Aspremont et al., 2004; Jolliffe and Uddin, 2003). One of our future work is to incorporate the sparsity criterion in discriminant analysis.

Acknowledgments

Research is sponsored, in part, by the Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014, the content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. Support Fellowships from Guidant Corporation and from the Department of Computer Science & Engineering, at the University of Minnesota, Twin Cities is gratefully acknowledged.

Appendix A.

The pseudo-inverse of a matrix is defined as follows.

Definition 2 *The pseudo-inverse of a matrix A, denoted as A^+, refers to the unique matrix satisfying the following four conditions:*

$$(1)A^+AA^+ = A^+, \quad (2)AA^+A = A, \quad (3)\ (AA^+)^T = AA^+, \quad (4)(A^+A)^T = A^+A.$$

The pseudo-inverse is commonly computed by the SVD as follows (Golub and Loan, 1996). Let $A = U \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} V^T$ be the SVD of A, where U and V are orthogonal and Σ is diagonal with positive diagonal entries. Then, $A^+ = V \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^T$.

References

G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

M. W. Berry, S. T. Dumais, and G. W. O'Brie. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573–595, 1995.

D. Q. Dai and P. C. Yuen. Regularized discriminant analysis and its application to face recognition. *Pattern Recognition*, 36:845–847, 2003.

A. d'Aspremont, L. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. In *Proceedings of the Eighteenth Annual Conference on Advances in Neural Information Processing Systems*, 2004.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Scienc*, 41:391–407, 1990.

L. Duchene and S. Leclerq. An optimal transformation for discriminant and principal component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6):978–983, 1988.

R. O. Duda, P. E. Hart, and D. Stork. *Pattern Classification*. Wiley, 2000.

S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457): 77–87, 2002.

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

D. H. Foley and J. W. Sammon. An optimal set of discriminant vectors. *IEEE Transactions on Computers*, 24(3):281–289, 1975.

J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

K. Fukunaga. *Introduction to Statistical Pattern Classification*. Academic Press, USA, 1990.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.

D. J. Hand. *Kernel discriminant analysis*. Research Studies Press/Wiley, 1982.

T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23: 73–102, 1995.

T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society series B*, 58:158–176, 1996.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer, 2001.

P. Howland, M. Jeon, and H. Park. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25(1):165–179, 2003.

Z. Jin, J. Y. Yang, Z. S. Hu, and Z. Lou. Face recognition based on the uncorrelated discriminant transformation. *Pattern Recognition*, 34:1405–1416, 2001a.

Z. Jin, J. Y. Yang, Z. M. Tang, and Z. S. Hu. A theorem on the uncorrelated optimal discriminant vectors. *Pattern Recognition*, 34(10):2041–2047, 2001b.

I. T. Jolliffe and M. Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12:531–547, 2003.

W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied Statistics*, 44:101–115, 1995.

D. D. Lewis. *Reuters-21578 text categorization test collection distribution 1.0*. http://www.research.att.com/∼lewis, 1999.

J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Face recognition using kernel direct discriminant analysis algorithms. *IEEE Transactions on Neural Networks*, 14(1):117–126, 2003.

A. M. Martinez and R. Benavente. The AR face database. Technical Report No. 24, 1998.

S. Ramaswamy and et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Science*, 98(26):15149–15154, 2001.

S. Raudys and R. P. W. Duin. On expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, 19(5-6):385–392, 1998.

B. Schökopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

M. Skurichina and R. P. W. Duin. Stabilizing classifiers for very small sample size. In *Proc. International Conference on Pattern Recognition*, pages 891–896, 1996.

M. Skurichina and R. P. W. Duin. Regularization of linear classifiers by adding redundant features. *Pattern Analysis and Applications*, 2(1):44–52, 1999.

D. L. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836, 1996.

TREC. *Text Retrieval conference*. http://trec.nist.gov, 1999.

M. A. Turk and A. P. Pentland. Face recognition using Eigenfaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial & Applied Mathematics, 1998.

J. Ye, R. Janardan, Q. Li, and H. Park. Feature extraction via generalized uncorrelated linear discriminant analysis. In *The Twenty-First International Conference on Machine Learning*, pages 895–902, 2004a.

J. Ye, R. Janardan, C. H. Park, and H. Park. An optimization criterion for generalized discriminant analysis on undersampled problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):982–994, 2004b.

C. H. Yeang and et al. Molecular classification of multiple tumor types. *Bioinformatics*, 17(1):1–7, 2001.

E. J. Yeoh and et al. Classification, subtype diswcovery, and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143, 2002.

W. Zhao, R. Chellappa, and P. Phillips. Subspace linear discriminant analysis for face recognition. Technical Report CAR-TR-914. Center for Automation Research, University of Maryland, 1999.