Feature Extraction via Generalized Uncorrelated Linear Discriminant Analysis

Jieping Ye JIEPING@CS.UMN.EDU
Department of Computer Science & Engineering, University of Minnesota

Ravi Janardan JANARDAN@CS.UMN.EDU
Department of Computer Science & Engineering, University of Minnesota

Qi Li QILI@CIS.UDEL.EDU
Department of Computer Information & Science, University of Delaware

Haesun Park HPARK@CS.UMN.EDU
National Science Foundation & Department of Computer Science & Engineering, University of Minnesota

Abstract

Feature extraction is important in many applications, such as text and image retrieval, because of high dimensionality. Uncorrelated Linear Discriminant Analysis (ULDA) was recently proposed for feature extraction. The extracted features via ULDA were shown to be statistically uncorrelated, which is desirable for many applications. In this paper, we will first propose the ULDA/QR algorithm to simplify the previous implementation of ULDA. Then we propose the ULDA/GSVD algorithm, based on a novel optimization criterion, to address the singularity problem. It is applicable for undersampled problem, where the data dimension is much larger than the data size, such as text and image retrieval. The novel criterion used in ULDA/GSVD is the perturbed version of the one from ULDA/QR, while surprisingly, the solution to ULDA/GSVD is shown to be independent of the amount of perturbation applied. We did extensive experiments on text and face image data to show the effectiveness of ULDA/GSVD and compare with other popular feature extraction algorithms.

1. Introduction

Feature extraction is important for many applications, such as text and image retrieval, because of the so-called *curse of dimensionality* (Fukunaga, 1990). Many methods have been proposed for feature extraction, such as Principal Component Analysis (PCA) (Jolliffe, 1986), Linear Discriminant Analysis (LDA) (Fukunaga, 1990), etc. LDA aims to find optimal discriminant features by maximizing the ratio of the between-class distance to the within-class distance of a given data set under supervised learning conditions. Its simplest implementation, so-called *classical LDA*, applies an eigen-decomposition on the scatter matrices, and fails when the scatter matrices are singular, as is the case for undersampled data. This is known as the *singularity problem* or *undersampled problem*. Many schemes have been proposed to address the singularity problem in classical LDA, such as regularized LDA (Friedman, 1989), subspace LDA (Swets & Weng, 1996), etc.

Uncorrelated features[1], are desirable in many applications, because they contain minimum redundancy. Motivated by extracting feature vectors having uncorrelated attributes, *uncorrelated LDA* (ULDA), was recently proposed in (Jin et al., 2001a; Jin et al., 2001b). However, the proposed algorithm in (Jin et al., 2001a) involves d generalized eigenvalue problems, if there exist d optimal discriminant vectors. It is computationally expensive for large and high-dimensional dataset. Like classical LDA, it does not address the singularity problem either. We thus call it *classical ULDA*. More details can be found in Section 3.

Classical LDA and classical ULDA were introduced from different perspectives. But it has been found that there is an intrinsic relationship between classical

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

[1]Two variable x and y are said to be uncorrelated, if their covariance is zero, i.e., $\mathrm{cov}(x, y) = 0$

LDA and classical ULDA (Jin et al., 2001b). More precisely, under the assumption that the eigenvalue problem in classical LDA has no multiple eigenvalues, (Jin et al., 2001b) showed that classical ULDA is equivalent to classical LDA. In this paper, we will further show that the equivalence between these two is still held without any assumption. Based on this equivalence, ULDA/QR is proposed to simplify the ULDA implementation in (Jin et al., 2001a), where ULDA/QR stands for ULDA based on QR-decomposition.

Classical LDA and classical ULDA do not address the singularity problem, hence it is difficult to apply them to undersampled data. Such high-dimensional, undersampled problems frequently occur in many applications including information retrieval (Howland et al., 2003), face recognition (Swets & Weng, 1996) and microarray analysis (Dudoit et al., 2002). Several schemes have been proposed to address the singularity problem in classical LDA in the past, including the subspace based method (Swets & Weng, 1996), regularization (Friedman, 1989), etc. The subspace-based method applies the Karhunen-Loeve (KL) expansion, also known as Principal Component Analysis (PCA) (Jolliffe, 1986), before LDA. Its limitation is that some useful information may be lost in the KL expansion. Regularized LDA overcomes the singularity problem by increasing the magnitude of the diagonal elements of the scatter matrices (usually by adding a scaled identity matrix). The difficulty in using regularized LDA for feature extraction is the choice of the amount of perturbation. A small perturbation is desirable to preserve the original matrix structure, while a large perturbation is more effective in dealing with the singularity problem.

There is much less work in addressing the singularity problem in classical ULDA than those on classical LDA. To the best of our knowledge, (Jin et al., 2001a) is the only result known, where a subspace based method was applied (PCA is applied to the between-class scatter matrix). The algorithm is named *subspace ULDA*.

We address the singularity problem of ULDA, in the second part of this paper, by introducing a novel optimization criterion that combines the key ingredients of ULDA/QR and regularized LDA. More specifically, the novel criterion is the perturbed version of the criterion used in ULDA/QR. Based on this criterion, in addition to Generalized Singular Value Decomposition (GSVD) tool (Golub & Van Loan, 1996), we propose a novel feature extraction algorithm, called ULDA/GSVD, where ULDA/GSVD stands for ULDA based on GSVD. ULDA/GSVD solves the singularity problem directly, thus avoiding the information loss in the subspace method. The difference between ULDA/GSVD and the traditional regularized LDA is that the optimal discriminant feature vectors via ULDA/GSVD are independent of the amount of perturbation (a quite surprising but firmly provable result), thus avoiding the limitation in regularized LDA. The details are given in Section 5.

With K-Nearest-Neighbor (K-NN) as classifier, we evaluate the effectiveness of ULDA/GSVD and compare with several other popular feature extraction algorithms, including OCM (Orthogonal Centroid Method) (Park et al., 2003), PCA (Jolliffe, 1986), and subspace ULDA (Jin et al., 2001a), on text and face image datasets. We observe that the accuracies on text datasets achieved by ULDA/GSVD are usually distinctly higher than all other tested algorithms. The result on face image datasets is also appealing. Even though the subspace ULDA is found to be quite competitive to ULDA/GSVD via 1-NN classifier, the latter one shows its extreme stability to different numbers of nearest neighbors in K-NN.

The rest of the paper is organized as follows: Sections 2 and 3 give brief reviews on classical LDA and classical ULDA respectively. Sections 4 and 5 propose the ULDA/QR and ULDA/GSVD algorithms, respectively. Experiments are presented in Section 6. We conclude in Section 7.

2. Classical Linear Discriminant Analysis

For convenience, Table 1 lists the important notations used in this paper.

Table 1. Notations

Notations	Descriptions
A	data matrix
n	number of training data points
N	dimension of the training data
ℓ	reduced dimension
k	number of the classes
S_b	between-class scatter matrix
S_w	within-class scatter matrix
S_t	total scatter matrix
G	transformation matrix
S_i	covariance matrix of the i-th class
m_i	mean of data from the i-th class
P_i	*a priori* probability of the i-th class
m	total mean of the training data
K	number of nearest neighbors in K-NN

Given a data matrix $A = (a_{ij}) \in R^{N \times n}$, where each column corresponds to a data point and each row corresponds to a particular feature, we consider finding a linear transformation $G \in R^{N \times \ell}$ $(\ell < N)$ that maps each column a_i, for $1 \leq i \leq n$, of A in the N-dimensional space to a vector y_i in the ℓ-dimensional space as follows: $G : a_i \in R^N \rightarrow y_i = G^T a_i \in R^\ell$. The resulting data matrix $Z = G^T A \in R^{\ell \times n}$ contains ℓ rows, i.e. there are ℓ features for each data point in the reduced (transformed) space. It is also clear that the features in the reduced space are linear combinations of the features in the original high dimensional space, where the coefficients of the linear combinations depend on the transformation matrix G.

A common way to compute the transformation matrix G is through classical LDA. Classical LDA computes the optimal G such that the class structure is preserved. More details are given below.

Assume there are k classes in the dataset. Suppose m_i, S_i, P_i are the mean vector, covariance matrix, and *a priori* probability of the i-th class, respectively, and m is the total mean. 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 follows (Fukunaga, 1990):

$$\begin{aligned} S_w &= \sum_{i=1}^{k} P_i S_i, \\ S_b &= \sum_{i=1}^{k} P_i (m_i - m)(m_i - m)^T, \\ S_t &= S_b + S_w. \end{aligned}$$

For the covariance matrix S_i of the ith class, we can decompose it as $S_i = X_i X_i^T$, where each column in X_i corresponds to a data point from the ith class subtracted by its mean m_i.

Define the matrices

$$H_w = [\sqrt{P_1} X_1, \cdots, \sqrt{P_k} X_k], \quad (1)$$

$$H_b = [\sqrt{P_1}(m_1 - m), \cdots, \sqrt{P_k}(m_k - m)]. \quad (2)$$

Then the scatter matrices S_w and S_b can be expressed as $S_w = H_w H_w^T$, $S_b = H_b H_b^T$. The *traces* of the two scatter matrices can be computed as follows: trace $(S_w) = \sum_{i=1}^{k} P_i ||X_i||_F^2$, and trace $(S_b) = \sum_{i=1}^{k} P_i ||m_i - m||^2$, where $|| \cdot ||_F$ denotes the Frobenius norm (Golub & Van Loan, 1996). Hence, trace (S_w) measures the closeness of the vectors within classes, while trace (S_b) measures the separation between classes.

In the lower-dimensional space resulting from the linear transformation G, the within-class and between-class matrices become

$$\begin{aligned} S_w^L &= (G^T H_w)(G^T H_w)^T = G^T S_w G, \\ S_b^L &= (G^T H_b)(G^T H_b)^T = G^T S_b G. \end{aligned}$$

An optimal transformation G would maximize trace (S_b^L) and minimize trace (S_w^L). Classical LDA aims to compute the optimal G, such that

$$G = \arg\max_G \text{trace}\left(\left(G^T S_w G \right)^{-1} G^T S_b G \right). \quad (3)$$

The solution can be obtained by solving an eigenvalue problem on $S_w^{-1} S_b$ (Fukunaga, 1990), provided that the within-class scatter matrix S_w is nonsingular. Since the rank of the between-class scatter matrix is bounded from above by $k - 1$, there are at most $k - 1$ discriminant vectors by classical LDA.

Classical LDA does not handle singular scatter matrices, which limits its applicability to undersampled problems. Several methods, including subspace LDA (Swets & Weng, 1996), and regularized LDA (Friedman, 1989), were proposed in the past to deal with the singularity problem.

In subspace LDA, an intermediate dimension reduction algorithm, such as PCA, is applied to reduce the dimension of the original data, before classical LDA is applied. A limitation of this approach is that the optimal value of the reduced dimension for the intermediate dimension reduction algorithm is difficult to determine. In regularized LDA, a constant μ is added to the diagonal elements of S_w, as $S_w + \mu I_N$, for some $\mu > 0$, where I_N is an identity matrix. It is easy to check that $S_w + \mu I_N$ is positive definite, hence nonsingular. A limitation of this approach is that the optimal value of the parameter μ is difficult to determine. More details can be found in (Krzanowski et al., 1995).

3. Uncorrelated Linear Discriminant Analysis

ULDA aims to find the optimal discriminant vectors that are S_t-orthogonal[2]. Specifically, suppose r vectors $\phi_1, \phi_2, \cdots, \phi_r$ are obtained, then the $(r+1)$th vector ϕ_{r+1} is the one that maximizes the Fisher criterion function $f(\phi) = \frac{\phi^T S_b \phi}{\phi^T S_w \phi}$, 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 $U_1 = I_N$,

[2] Two vectors x and y are S_t-orthogonal, if $x^T S_t y = 0$.

$U_j = I_N - S_t D_j^T (D_j S_t S_w^{-1} S_t D_j^T)^{-1} D_j S_t S_w^{-1} (j > 1)$, $D_j = [\phi_1, \cdots, \phi_{j-1}]^T (j > 1)$, and I_N is the identity matrix.

It was shown that the feature vectors transformed by ULDA are mutually uncorrelated. This is a desirable property for feature extraction. More details on the role of uncorrelated attributes can be found in (Jin et al., 2001a). The limitation of the ULDA algorithm in (Jin et al., 2001a) lies in the expensive computation of the d generalized eigenvalue problems, where d is number of optimal discriminant vectors by ULDA.

4. The ULDA/QR algorithm

In this section, we show the equivalence between classical ULDA and a variant of classical LDA, regardless of the distribution of the eigenvalues of $S_w^{-1} S_b$. This result enhances the one in (Jin et al., 2001b) where the equivalence between these two is based on the assumption that there are no multiple eigenvalues for $S_w^{-1} S_b$ (note that both results assume that the within-class scatter matrix S_w is nonsingular). Based on the equivalence, we propose ULDA/QR to simplify the ULDA implementation in (Jin et al., 2001a).

Consider a variant of classical LDA in Eq. (3) as follows:

$$G = \arg \min_{G^T S_t G = I_\ell} F(G), \tag{4}$$

where $F(G) = \text{trace}\left(\left(G^T S_w G \right)^{-1} G^T S_b G \right)$.

From linear algebra, there exists a nonsingular matrix X such that

$$\begin{aligned} X^T S_w X &= I_N, & (5) \\ X^T S_b X &= \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_N), & (6) \end{aligned}$$

where $\lambda_1 \geq \cdots \geq \lambda_N$. An efficient algorithm to compute the matrix X will be given later in this section. It can be shown that the matrix consisting of the first q columns of X (with normalization) solves the optimization problem in Eq. (4), where q is the rank of the matrix S_b, as stated in the following theorem:

Theorem 4.1. *Let the matrix X be defined as in Eq. (5) and Eq. (6), and $q = rank(S_b)$. Let $G^* = [\tilde{x}_1, \cdots, \tilde{x}_q]$, where $\tilde{x}_i = \frac{1}{\sqrt{1+\lambda_i}} x_i$, x_i is the i-th column of the matrix X, and λ_i's are defined in Eq. (6). Then G^* solves the optimization problem in Eq. (4).*

Proof. It is clear that the constraint in Eq. (4) is satisfied for $G = G^*$. Next we only need to show that the minimum of $F(G)$ is obtained at G^*. By Eq. (5) and Eq. (6), we have

$$G^T S_w G = G^T X^{-T} (X^T S_w X) X^{-1} G = \tilde{G} \tilde{G}^T,$$

$$G^T S_b G = G^T X^{-T} (X^T S_b X) X^{-1} G = \tilde{G} \Lambda \tilde{G}^T,$$

where $\tilde{G} = \left(X^{-1} G \right)^T$. Hence, $F(G) = \text{trace}\left(\left(\tilde{G} \tilde{G}^T \right)^{-1} \left(\tilde{G} \Lambda \tilde{G}^T \right) \right)$. Let $\tilde{G}^T = QR$ be the QR-decomposition of $\tilde{G}^T \in R^{N \times \ell}$ (note that $\tilde{G}^T$ has full column rank), where $Q \in R^{N \times \ell}$ has orthonormal columns and R is nonsingular. Using the fact that $\text{trace}(AB) = \text{trace}(BA)$, for any matrices A and B, we have

$$\begin{aligned} F(G) &= \text{trace}\left(\left(R^T R \right)^{-1} \left(R^T Q^T \Lambda Q R \right) \right) \\ &= \text{trace}\left(Q^T \Lambda Q \right) \leq \lambda_1 + \cdots + \lambda_q, \end{aligned}$$

where the inequality becomes equality for

$$Q = \begin{pmatrix} I_\ell \\ 0 \end{pmatrix} \text{ or } G = X \begin{pmatrix} I_\ell \\ 0 \end{pmatrix} R,$$

when the reduced dimension $\ell = q$. Note that R is an arbitrary upper triangular and nonsingular matrix. Hence, G^* corresponds to the case when R is set to be $R = \text{diag}\left(\frac{1}{\sqrt{1+\lambda_1}}, \cdots, \frac{1}{\sqrt{1+\lambda_q}} \right)$. □

We are now ready to present our main result for this section:

Theorem 4.2. *Let $\tilde{x}_i$'s be defined as in Theorem 4.1. Then $\{\tilde{x}_i\}_{i=1}^q$ forms an optimal discriminant vectors for ULDA, where $q = rank(S_b)$.*

Proof. Proof by induction. It is trivial to check that $\tilde{x}_1 = \arg\max_\phi f(\phi)$, i.e. $\phi_1 = \tilde{x}_1$. Next assume $\phi_i = \tilde{x}_i$, for $i = 1, \cdots, r$. We show in the following that $\phi_{r+1} = \tilde{x}_{r+1}$.

By the definition, $\phi_{r+1} = \arg\max_\phi f(\phi)$, subject to $\phi_{r+1}^T S_t \phi_i = 0$, for $i = 1, \cdots, r$. Let $\phi_{r+1} = \sum_{i=1}^N \gamma_i \tilde{x}_i$, since $\{\tilde{x}_i\}_{i=1}^N$ forms a base for R^N. By the constraints $\phi_{r+1}^T S_t \phi_i = 0$, for $i = 1, \cdots, r$, we have $\gamma_i = 0$, for $i = 1, \cdots, r$, hence $\phi_{r+1} = \sum_{i=r+1}^N \gamma_i \tilde{x}_i$. It follows from Eq. (5) and Eq. (6) that

$$\begin{aligned} f(\phi_{r+1}) &= \frac{\left(\sum_{i=r+1}^N \gamma_i \tilde{x}_i^T \right) S_b \left(\sum_{i=r+1}^N \gamma_i \tilde{x}_i \right)}{\left(\sum_{i=r+1}^N \gamma_i \tilde{x}_i^T \right) S_w \left(\sum_{i=r+1}^N \gamma_i \tilde{x}_i \right)} \\ &= \frac{\sum_{i=r+1}^N \gamma_i^2 \lambda_i}{\sum_{i=r+1}^N \gamma_i^2} \leq \frac{\sum_{i=r+1}^N \gamma_i^2 \lambda_{r+1}}{\sum_{i=r+1}^m \gamma_i^2} \\ &= \lambda_{r+1}, \end{aligned}$$

Algorithm 1: The ULDA/QR Algorithm

Input: Data matrix A.

Output: Discriminant vectors $\tilde{x}_i$'s of ULDA.

1. Construct matrices H_w and H_b as in Eq. (1) and Eq. (2).
2. Compute QR-decomposition on H_w^T as $H_w^T = QR$, where $Q \in R^{n\times N}$, $R \in R^{N\times N}$.
3. Form the matrix $Y \leftarrow H_b^T R^{-1}$.
4. Compute SVD on Y as $Y = U\Sigma V^T$, where $U \in R^{n\times q}$, $\Sigma = \mathrm{diag}(\sigma_1, \cdots, \sigma_q) \in R^{q\times q}$, $V \in R^{N\times q}$, $\sigma_1 \geq \cdots \geq \sigma_q$, and $q = \mathrm{rank}(H_b)$.
5. $[x_1, \cdots, x_q] \leftarrow R^{-1}V$.
6. $\lambda_i \leftarrow \sigma_i^2$, for $i = 1, \cdots, q$.
7. $\tilde{x}_i \leftarrow \frac{1}{\sqrt{1+\lambda_i}} x_i$, for $i = 1, \cdots, q$.

where the inequality becomes equality if $\gamma_i = 0$, for $i = r+2, \cdots, N$. Hence $\tilde{x}_{r+1}$ can be chosen as the $(r+1)$-th discriminant vector of ULDA, i.e. $\phi_{r+1} = \tilde{x}_{r+1}$. □

An efficient algorithm for computing $\{\tilde{x}_i\}_{i=1}^q$ through QR-decomposition is given in **Algorithm 1**.

5. The ULDA/GSVD algorithm

In the last section, a variant of classical LDA is presented in Eq. (4). It was shown that the solution to the optimization problem in Eq. (4) forms optimal discriminant vectors for classical ULDA. Thus, it provides an efficient way to compute optimal discriminant vectors for ULDA. However, the algorithm assumes the non-singularity of S_w, which limits its applicability to applications involving high-dimensional data. In (Jin et al., 2001a), a subspace based method is presented to solve the singularity problem, where the ULDA algorithm is preceded by PCA. However, the PCA stage may lose some useful information. In this section, we propose a new feature extraction algorithm, called ULDA/GSVD. The new criterion underlying ULDA/GSVD is motivated by the criterion in Eq. (4) and the perturbation in regularized LDA. The new optimization problem for ULDA/GSVD is defined as follows,

$$G_\mu = \arg \max_{G^T S_t G = I_\ell} F_\mu(G), \qquad (7)$$

where $F_\mu(G) = \mathrm{trace}\left((G^T S_w G + \mu I_\ell)^{-1} G^T S_b G\right)$.

Recall that a limitation of regularized LDA is that the optimal value of the perturbation μ is difficult to determine. A key difference between ULDA/GSVD and regularized LDA is that the optimal solution to ULDA/GSVD is independent of the perturbation applied, i.e., $G_{\mu_1} = G_{\mu_2}$, for any $\mu_1, \mu_2 > 0$. The main technique applied here for computing G_μ, for for any $\mu > 0$, is the Generalized Singular Value Decomposition (GSVD). A simple algorithm to compute GSVD can be found in (Howland et al., 2003), where the algorithm is based on (Paige & Saunders, 1981).

We need the following two lemmas to compute G_μ, for any $\mu > 0$.

Lemma 5.1. *Let S_w, S_b, and S_t be defined as in Section 2, and $t = rank(S_t)$. Then there exists a nonsingular matrix $X \in R^{N\times N}$, such that*

$$X^T S_b X = D_1 = diag(\alpha_1^2, \cdots, \alpha_t^2, 0 \cdots, 0),$$

$$X^T S_w X = D_2 = diag(\beta_1^2, \cdots, \beta_t^2, 0 \cdots, 0),$$

where $1 \geq \alpha_1 \geq \cdots \geq \alpha_q > 0 = \alpha_{q+1} = \cdots = \alpha_t$, $0 \leq \beta_1 \leq \cdots \leq \beta_t \leq 1$, $D_1 + D_2 = \begin{pmatrix} I_t & 0 \\ 0 & 0 \end{pmatrix}$, and $q = rank(S_b)$.

Proof. Let $K = \begin{bmatrix} H_b^T \\ H_w^T \end{bmatrix}$, which is an $(n+k)\times N$ matrix. By the generalized singular value decomposition (Paige & Saunders, 1981), there exist orthogonal matrices $U \in R^{k\times k}$, $V \in R^{n\times n}$, and a nonsingular matrix $X \in R^{N\times N}$, such that

$$\begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}^T KX = \begin{bmatrix} \Sigma_1 & 0 \\ \Sigma_2 & 0 \end{bmatrix}, \qquad (8)$$

where $\Sigma_1^T\Sigma_1 = \mathrm{diag}(\alpha_1^2, \cdots, \alpha_t^2)$, $\Sigma_2^T\Sigma_2 = \mathrm{diag}(\beta_1^2, \cdots, \beta_t^2)$, $1 \geq \alpha_1 \geq \cdots \geq \alpha_q > 0 = \alpha_{q+1} = \cdots = \alpha_t$, $0 \leq \beta_1 \leq \cdots \leq \beta_t \leq 1$, $\alpha_i^2 + \beta_i^2 = 1$, for $i = 1, \cdots, t$, and $q = \mathrm{rank}(H_b) = \mathrm{rank}(S_b)$.

Hence, $H_b^T X = U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix}$, and $H_w^T X = V \begin{bmatrix} \Sigma_2 & 0 \end{bmatrix}$. It follows that

$$X^T S_b X = X^T H_b H_b^T X = \begin{bmatrix} \Sigma_1^T\Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} = D_1,$$

$$X^T S_w X = X^T H_w H_w^T X = \begin{bmatrix} \Sigma_2^T\Sigma_2 & 0 \\ 0 & 0 \end{bmatrix} = D_2,$$

where $D_1 + D_2 = \begin{pmatrix} I_t & 0 \\ 0 & 0 \end{pmatrix}$. □

Lemma 5.2. *Define a trace optimization problem as follows:*

$$G = \arg \max_{G^T G = I_\ell} trace\left(\left(G^T W G\right)^{-1} G^T B G\right), \qquad (9)$$

where $W = diag(w_1, \cdots, w_u) \in R^{u\times u}$ is a diagonal matrix with $0 < w_1 \leq \cdots \leq w_u$, and $B = diag(b_1, \cdots, b_u) \in R^{u\times u}$ is also diagonal with $b_1 \geq \cdots \geq b_q > 0 = b_{q+1} = \cdots = b_u$, i.e. $rank(B) = q$.

Then $G^\star = \begin{pmatrix} I_q \\ 0 \end{pmatrix}$ *solves the optimization problem in Eq. (9) with* $\ell = q$.

Proof. It is clear that the constraint in the optimization in Eq. (9) is satisfied for $G^\star$ with $\ell = q$. Next, we show that $G^\star$ solves the following optimization problem:

$$G = \arg\max_G \operatorname{trace}\left(\left(G^T W G\right)^{-1} G^T B G\right). \qquad (10)$$

The optimization in Eq. (10) corresponds to the trace optimization problem in classical LDA in Eq. (3) with the within-class scatter matrix $S_w = W$ and between-class scatter matrix $S_b = B$. Since the within-class scatter matrix $S_w = W$ is nonsingular, the solution can be obtained by solving the eigenvalue problem on $W^{-1}B$. It is easy to check that $W^{-1}B$ is diagonal and only the first q diagonal entries are nonzero. Hence e_i, for $i = 1, \cdots, q$, is the eigenvector of $W^{-1}B$ corresponding to the i-th largest eigenvalue, where $e_i = (0, \cdots, 1, 0 \cdots, 0)^T$ and the 1 appears at the i-th position. Therefore $G^\star$ solves the optimization in Eq. (10). □

The main result is stated in the following theorem:

Theorem 5.1. *Let the matrix* X *be defined as in Lemma 5.1, and let* $q = rank(S_b)$. *Then* $G^*_\mu = X \begin{pmatrix} I_q \\ 0 \end{pmatrix}$ *solves the optimization problem in Eq. (7) with* $\ell = q$.

Proof. By Lemma 5.1, $X^T S_b X = D_1$, $X^T S_w X = D_2$, where the two diagonal matrices D_1 and D_2 satisfy $D_1 + D_2 = \begin{pmatrix} I_t & 0 \\ 0 & 0 \end{pmatrix}$. It is easy to check that

$$\begin{aligned} (G^*_\mu)^T S_t G^*_\mu &= (I_q, 0)\, X^T (S_b + S_w) X \begin{pmatrix} I_q \\ 0 \end{pmatrix} \\ &= (I_q, 0)\,(D_1 + D_2) \begin{pmatrix} I_q \\ 0 \end{pmatrix} = I_q, \end{aligned}$$

i.e. the constraint in the optimization problem in Eq. (7) is satisfied. Next we show G^*_μ minimizes $F_\mu(G)$.

Since

$$G^T S_b G = G^T (X^{-1})^T (X^T S_b X) X^{-1} G = \tilde{G} D_1 \tilde{G}^T,$$

$$G^T S_w G = G^T (X^{-1})^T (X^T S_w X) X^{-1} G = \tilde{G} D_2 \tilde{G}^T,$$

where $\tilde{G} = (X^{-1}G)^T$, $F_\mu(G)$ can then be rewritten as,

$$F_\mu(G) = \operatorname{trace}\left(\left(\tilde{G} D_2 \tilde{G}^T + \mu I_\ell\right)^{-1} \tilde{G} D_1 \tilde{G}^T\right). \quad (11)$$

Algorithm 2: The ULDA/GSVD Algorithm

Input: Data matrix A
Output: Optimal discriminant vectors $\{\phi_i\}$
1. Form H_b and H_w as in Eq. (2) and Eq. (1).
2. Compute GSVD on the matrix pair (H_b^T, H_w^T) to obtain the matrix X, as in Lemma 5.1.
3. $q \leftarrow \text{rank}(H_b)$.
4. $\phi_i \leftarrow X_i$, for $i = 1, \cdots, q$.

Let $\tilde{G} = \begin{pmatrix} G_1^T & G_2^T \end{pmatrix}$ be a partition of $\tilde{G}$, such that $G_1^T \in R^{\ell \times t}$, $G_2^T \in R^{\ell \times N - t}$.

By the constraint that $G^T S_t G = I_\ell$, we have

$$\begin{aligned} I_\ell &= G^T S_t G = G^T (S_w + S_b) G = G^T S_b G + G^T S_w G \\ &= \tilde{G} D_1 \tilde{G}^T + \tilde{G} D_2 \tilde{G}^T = \tilde{G}(D_1 + D_2)\tilde{G}^T = G_1^T G_1. \end{aligned}$$

Hence, $F_\mu(G)$ in Eq. (11) can be rewritten as

$$F_\mu(G) = \operatorname{trace}\left(\left(G_1^T \left(D_2^t + \mu I_\ell\right) G_1\right)^{-1} G_1^T D_1^t G_1\right).$$

where D_1^t and D_2^t are the t-th principal sub-matrices of D_1 and D_2 respectively. It is clear that $F_\mu(G)$ is independent of G_2. Hence we can simplify set $G_2 = 0$. Denote $\Sigma = (D_2^t + \mu I_t)$, which is a nonsingular and diagonal matrix. It follows that $F_\mu(G) = \operatorname{trace}\left(\left(G_1^T \Sigma G_1\right)^{-1} G_1^T D_1^t G_1\right)$. The result then follows from Lemma 5.2, with $W = \Sigma$ and $B = D_1^t$. □

It is clear from Theorem 5.1 that the optimal solution $G^\star_\mu$ to the optimization in Eq. (7) only depends on X, which is determined by H_w and H_b, hence it is independent of μ, as stated in the following theorem:

Theorem 5.2. *Let* G^*_μ, *for* $\mu > 0$, *be as in Theorem 5.1. Then* $G^*_{\mu_1} = G^*_{\mu_2}$, *for any* $\mu_1, \mu_2 > 0$.

The main algorithm is presented in **Algorithm 2**.

Remark 5.1. *Theorem 5.2 implies that the choice of* μ *in Eq. (7) is irrelevant for practical implementation (see* **Algorithm 2***). However, it leads to the theoretical derivation of ULDA/GSVD. The complexity of the* **Algorithm 2** *can be shown to be* $O(n^2 N)$.

6. Experiments

This section consists of two parts. The first part describes our test datasets. The second part compares our ULDA/GSVD algorithm with PCA, OCM, and subspace ULDA, in terms of classification accuracy. The K-Nearest-Neighbor (K-NN) algorithm was used as our classifier. In each test, a dataset is randomly partitioned into training and testing sets with equal

sizes. The final classification accuracy reported is the average over 10 different partitions.

Note that ULDA/QR does not apply for undersampled data, which is the case for all the datasets in this paper. However, when the scatter matrix is non-singular, ULDA/GSVD is equivalent to ULDA/QR.

6.1. Datasets

We have three text document datasets Doc1–3 and three face image datasets Img1–3 as follows: Doc1 is derived from the TREC-5, TREC-6, and TREC-7 collections, available at http://trec.nist.gov; Doc2 and Doc3 are derived from *Reuters-21578* text categorization test collection Distribution 1.0, available at http://www.research.att.com/~lewis; Img1 is the ORL face image dataset available at http://www.uk.research.att.com/facedatabase.html; Img2 is the PIX face image dataset available at http://peipa.essex.ac.uk/ipa/pix/faces/manchester; Img3 is the AR face image dataset available at http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html. For text documents, we use a stop-list to remove common words, and the words are stemmed using Porter's suffix-stripping algorithm (Porter, 1980). Table 2 summarizes the statistics of our test datasets.

6.2. Results and analysis

The accuracy curves of four feature extraction algorithms: OCM, PCA[3], subspace ULDA[4], and ULDA/GSVD, on the six datasets are shown in Fig. 1. The horizontal axis represents the number of nearest neighbors in K-NN algorithm, and the vertical axis represents the classification accuracy/precision.

For the text data, the superiority of ULDA/GSVD over the other three feature extraction algorithms can be easily observed via different K-NNs. The subspace method and OCM are comparable to each other.

As for the images, ULDA/GSVD again outperforms other algorithms. Unlike the behaviors on text data, subspace ULDA becomes quite competitive to ULDA/GSVD and PCA is also quite competitive except on the AR dataset. This is especially true when $K = 1$ nearest neighbor is used in K-NN. Note that the poor performance of OCM and PCA on the AR dataset is probably due to large within-class variation in the AR dataset. Unlike subspace ULDA and ULDA/GSVD, both PCA and OCM omit the within-class information. Another major observation is the stability of ULDA/GSVD to the number of nearest neighbors. Recall that ULDA/GSVD minimizes the within-class distance and maximizes the between-class distance simultaneously with maximum discrimination. Thus, the performance of K-NN algorithm on the reduced space by ULDA/GSVD is expected to be insensitive to the number of neighbors (K).

[3]Extensive experiments showed that using $p = 100$ principal components in PCA gave good overall results.

[4]Extensive experiments showed that using $p = 200$ principal components in the PCA stage of subspace ULDA gave good overall results.

Table 2. Statistics for the test datasets

Dataset	Source	Size	Dim	# of classes
Doc1	TREC	210	7454	7
Doc2	Reuters	320	2887	4
Doc3	Reuters	490	3759	5
Img1	ORL	400	10304	40
Img2	PIX	300	10000	30
Img3	AR	1638	8888	126

7. Conclusion

Uncorrelated attributes with minimum redundancy are highly desirable in feature extraction for many applications such as text retrieval, image retrieval, etc. In this paper, we presented a study on uncorrelated Linear Discriminant Analysis (ULDA). This study contains two major contributions. The first one is the theoretical result on the equivalence between classical ULDA and classical LDA, which leads to a fast implementation, ULDA/QR, of ULDA. Then we propose ULDA/GSVD, based on a novel optimization criterion, that can successfully overcome the singularity problem in classical ULDA. The novel criterion used in ULDA/GSVD is the perturbed version of the one from ULDA/QR, while the solution to ULDA/GSVD is shown to be independent of the amount of perturbation applied, thus avoiding the limitation in regularized LDA. The experiments on text and face image data show the superiority of ULDA/GSVD over other competing algorithms including PCA, OCM, and subspace ULDA.

Acknowledgment

Research of J. Ye and R. Janardan 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

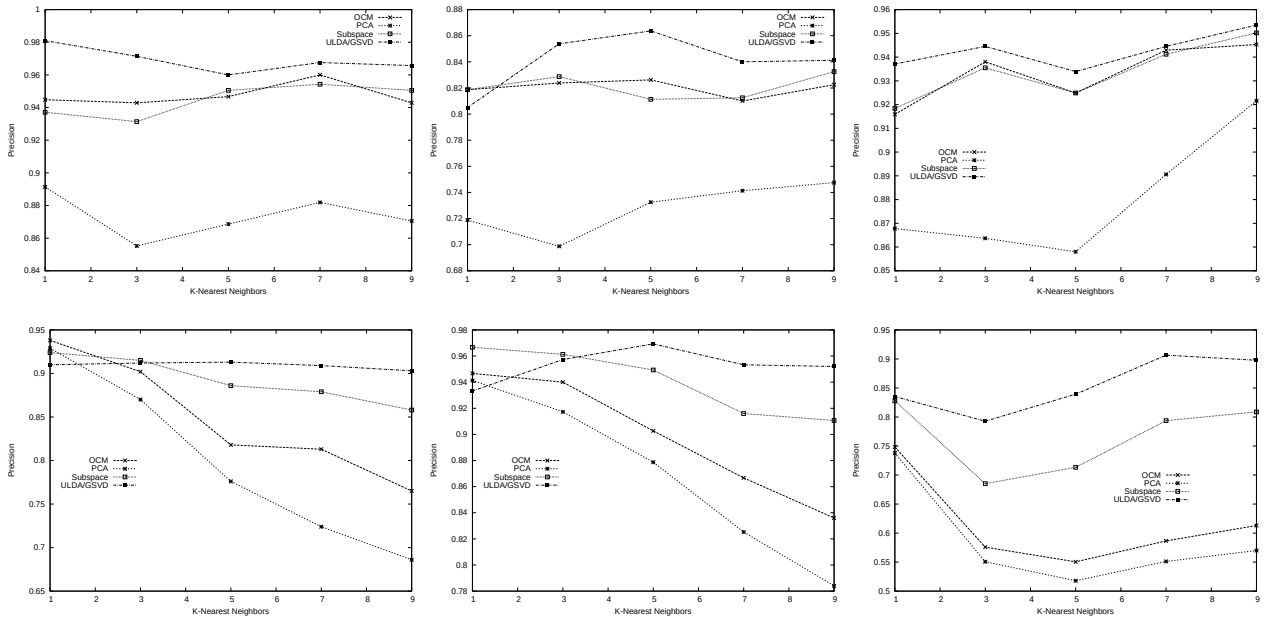

Figure 1. Comparison on Doc1–3 document datasets (from left to right on the top) and Img1–3 face image datasets (from left to right in the bottom)

should be inferred. The work of Haesun Park has been performed while at the NSF and was partly supported by IR/D from the National Science Foundation (NSF). This material is based upon work supported in part by the National Science Foundation Grants CCR-0204109 and ACI-0305543. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

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

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

Fukunaga, K. (1990). *Introduction to statistical pattern classification*. San Diego, California, USA: Academic Press.

Golub, G. H., & Van Loan, C. F. (1996). *Matrix computations*. Baltimore, MD, USA: The Johns Hopkins University Press. Third edition.

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

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

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

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.

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

Paige, C., & Saunders, M. (1981). Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, *18*, 398–405.

Park, H., Jeon, M., & Rosen, J. (2003). Lower dimensional representation of text data based on centroids and least squares. *BIT*, *43*, 1–22.

Porter, M. (1980). An algorithm for suffix stripping program. *Program*, *14*, 130–137.

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