Null Space versus Orthogonal Linear Discriminant Analysis

Jieping Ye JIEPING.YE@ASU.EDU
Center for Evolutionary Functional Genomics, and Department of Computer Science & Engineering, Arizona State University, Tempe, AZ 85287, USA

Tao Xiong TXIONG@ECE.UMN.EDU
Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA

Abstract

Dimensionality reduction is an important pre-processing step for many applications. Linear Discriminant Analysis (LDA) is one of the well known methods for supervised dimensionality reduction. However, the classical LDA formulation requires the nonsingularity of scatter matrices involved. For undersampled problems, where the data dimension is much larger than the sample size, all scatter matrices are singular and classical LDA fails. Many extensions, including null space based LDA (NLDA), orthogonal LDA (OLDA), etc, have been proposed in the past to overcome this problem. In this paper, we present a computational and theoretical analysis of NLDA and OLDA. Our main result shows that under a mild condition which holds in many applications involving high-dimensional data, NLDA is equivalent to OLDA. We have performed extensive experiments on various types of data and results are consistent with our theoretical analysis. The presented analysis and experimental results provide further insight into several LDA based algorithms.

1. Introduction

Dimensionality reduction is important in many applications of data mining, machine learning, and bioinformatics, due to the so-called *curse of dimensionality* (Duda et al., 2000; Fukunaga, 1990; Hastie et al., 2001). Many methods have been proposed for dimensionality reduction, including Principal Component Analysis (PCA) (Jolliffe, 1986), Linear Discriminant Analysis (LDA) (Fukunaga, 1990), etc. LDA aims to find optimal discriminant vectors (transformation) by maximizing the ratio of the between-class distance to the within-class distance, thus achieving maximum class discrimination. It has been applied successfully in many applications including face recognition (Belhumeour et al., 1997; Swets & Weng, 1996; Turk & Pentland, 1991), document classification (Ye et al., 2004a), and microarray data analysis (Dudoit et al., 2002; Ye et al., 2004b). However classical LDA requires the total scatter matrix to be nonsingular. In many applications such as face recognition, document classification, and microarray data analysis, all scatter matrices in question can be singular since the sample size is much smaller than the data dimension. This is known as the *singularity* or *undersampled problem* (Krzanowski et al., 1995). In recent years, many approaches have been proposed to overcome the singularity problem, including null space based LDA (NLDA) (Chen et al., 2000; Huang et al., 2002), orthogonal LDA (OLDA) (Ye, 2005), subspace LDA (Belhumeour et al., 1997; Swets & Weng, 1996), regularized LDA (Friedman, 1989), penalized LDA (Hastie et al., 1995), etc.

In (Chen et al., 2000), Chen *et al.* proposed the null space based LDA (NLDA), where the between-class scatter is maximized in the null space of the within-class scatter matrix. The singularity problem is thus implicitly avoided. Similar idea has been mentioned briefly in (Belhumeour et al., 1997). Huang *et al.* improved the efficiency of the algorithm by first removing the null space of the total scatter matrix (Huang et al., 2002). It is based on the observation that the null space of the total scatter matrix is the intersection of the null space of the between-class scatter matrix and the null space of the within-class scatter matrix. In orthogonal LDA (OLDA) (Ye, 2005), a set of orthogonal discriminant vectors is computed, based on a new

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

optimization criterion. The optimal transformation is computed through the simultaneous diagonalization of scatter matrices, while the singularity problem is overcome implicitly. Note that discriminant analysis with orthogonal transformation has been studied by Foley and Sammon (Foley & Sammon, 1975) and by Duchene and Leclercq (Duchene & Leclerq, 1988).

Both the NLDA algorithm (Huang et al., 2002) and the OLDA algorithm (Ye, 2005) result in orthogonal transformations. However, they applied different schemes in deriving the optimal transformation. NLDA computes an orthogonal transformation in the null space of the within-class scatter matrix, while OLDA computes an orthogonal transformation through the simultaneous diagonalization of scatter matrices. Surprisingly, we show in Section 4 that NLDA is equivalent to OLDA, under a mild condition C1 that the rank of the total scatter matrix equals to the sum of the rank of the between-class scatter matrix and the rank of the within-class scatter matrix.

We have conducted extensive experiments using nine high-dimensional datasets from various data sources, including text documents, face images, and gene expression data. (details on these datasets can be found in Section 5.) Experimental results show that Condition C1 holds for most datasets (eight out of the nine datasets). NLDA and OLDA achieve the same performance, in all cases when condition C1 holds. For cases where condition C1 does not hold, OLDA outperforms NLDA, as OLDA has a larger number of reduced dimensions than NLDA. These empirical results are consistent with our theoretical analysis.

2. Classical Linear Discriminant Analysis

Given a data matrix $A = (a_{ij}) \in \mathbb{R}^{m\times n}$, where each column corresponds to a data point and each row corresponds to a particular feature, we consider finding a linear transformation $G \in \mathbb{R}^{m\times \ell}$ $(\ell < m)$ that maps each column a_i, for $1 \le i \le n$, of A in the m-dimensional space to a vector y_i in the ℓ-dimensional space as follows:

$$G : a_i \in \mathbb{R}^m \to y_i = G^T a_i \in \mathbb{R}^\ell.$$

In classical LDA, the transformation matrix G is computed so that the class structure is preserved.

Assume that there are k classes in the dataset. Suppose c_i, S_i, n_i are the centroid, covariance matrix, and sample size of the i-th class, respectively, and c is the global centroid. 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): $S_w = \sum_{i=1}^k n_i S_i$, $S_b = \sum_{i=1}^k n_i (c_i - c)(c_i - c)^T$, and $S_t = \sum_{i=1}^k (A_i - ce^T)(A_i - ce^T)^T$. The covariance matrix S_i of the i-th class can be decomposed as

$$S_i = \frac{1}{n_i}(A_i - c_i e^T)(A_i - c_i e^T)^T,$$

where A_i is the data matrix of the i-th class, and e is the vector of all ones with an appropriate length. That is, each column of $A_i - c_i e^T$ corresponds to a data point from the i-th class subtracted by its centroid c_i. It is easy to verify that $S_t = S_b + S_w$.

Define the matrices

$$H_w = [(A_1 - c_1 e^T), \cdots, (A_k - c_k e^T)], \quad (1)$$
$$H_b = [\sqrt{n_1}(c_1 - c), \cdots, \sqrt{n_k}(c_k - c)], \quad (2)$$
$$H_t = A - ce^T. \quad (3)$$

Then the scatter matrices S_w, S_b, and S_t can be expressed as $S_w = H_w H_w^T$, $S_b = H_b H_b^T$, and $S_t = H_t H_t^T$. The *traces* of the three scatter matrices can be computed as follows: $\mathrm{trace}(S_w) = \sum_{i=1}^k ||A_i - c_i e^T||_F^2$, $\mathrm{trace}(S_b) = \sum_{i=1}^k n_i ||c_i - c||^2$, $\mathrm{trace}(S_t) = \sum_{i=1}^k ||A_i - ce^T||_F^2$, where $||\cdot||_F$ denotes the Frobenius norm (Golub & Van Loan, 1996). Hence, $\mathrm{trace}(S_w)$ measures the within-class cohesion, $\mathrm{trace}(S_b)$ measures the between-class separation, and $\mathrm{trace}(S_t)$ measures the variance of the dataset.

In the lower-dimensional space resulting from the linear transformation G, the scatter matrices become $S_w^L = G^T S_w G$, $S_b^L = G^T S_b G$, and $S_t^L = G^T S_t G$. An optimal transformation G would maximize $\mathrm{trace}(S_b^L)$ and minimize $\mathrm{trace}(S_w^L)$. Classical LDA aims to compute the optimal G by solving the following optimization problem:

$$G = \arg \max_{G\in\mathbb{R}^{m\times\ell}: G^T S_w G = I_\ell} \mathrm{trace}\left(\left(G^T S_w G\right)^{-1} G^T S_b G\right). \quad (4)$$

Other optimization criteria, including those based on the determinant could also be used instead (Duda et al., 2000; Fukunaga, 1990). The solution to the optimization problem in Eq. (4) 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. The columns of G form the discriminant vectors of classical LDA. Since the rank of the between-class scatter matrix is bounded from above by $k-1$, there are at most $k-1$ discriminant vectors in classical LDA. A stable way to solve this eigenvalue problem is to apply SVD on the scatter matrices. Details can be found in (Swets & Weng, 1996).

Classical LDA does not handle singular scatter matrices, which limits its applicability to low-dimensional data. Several methods, including null space based LDA, orthogonal LDA, subspace LDA, etc, were proposed in the past to deal with this problem as discussed in Section 1.

2.1. Orthogonal LDA

Orthogonal LDA (OLDA) was proposed in (Ye, 2005) as an extension of classical LDA. The discriminant vectors in OLDA are orthogonal to each other. Furthermore, OLDA is applicable, even when all scatter matrices are singular, thus overcoming the singularity problem. It has been applied successfully in many applications, including document classification, face recognition, and gene expression data classification. The optimal transformation in OLDA can be computed by solving the following optimization problem:

$$G = \text{argmax}_{G \in \mathbb{R}^{m \times \ell}: G^T G = I_\ell} \text{trace}\left((S_t^L)^+ S_b^L\right), \quad (5)$$

where $S_t^L = G^T S_t G$, $S_b^L = G^T S_b G$, and $(S_t^L)^+$ denotes the pseudo-inverse (Golub & Van Loan, 1996) of S_t^L. The orthogonality condition is imposed in the constraint. The computation of the optimal transformation of OLDA is based on the simultaneous diagonalization of the three scatter matrices (Ye, 2005).

3. Null Space LDA

Chen *et al.* proposed the null space based LDA (NLDA) for dimensionality reduction, where the between-class scatter is maximized in the null space of the within-class scatter matrix (Chen et al., 2000). The basic idea behind this algorithm is that the null space of S_w may contain significant discriminant information if the projection of S_b is not zero in that direction (Chen et al., 2000; Lu et al., 2003). The singularity problem is thus overcome implicitly. The optimal transformation of NLDA can be computed by solving the following optimization problem:

$$G = \text{argmax}_{G^T S_w G = 0} \text{trace}(G^T S_b G). \quad (6)$$

The computation of the optimal G involves the computation of the null space of S_w, which may be large for high-dimensional data. Indeed, the dimension of the null space of S_w is at least $m + k - n$, where m is the data dimension, k is the number of classes, and n is the sample size. In (Chen et al., 2000), a pixel grouping method is used to extract geometric features and reduce the dimension of samples, and then NLDA is applied in the new feature space.

Huang *et al.* improved the efficiency of the algorithm by first removing the null space of the total scatter matrix S_t (Huang et al., 2002). It is based on the observation that the null space of S_t is the intersection of the null space of S_b and the null space of S_w, as stated in the following lemma (Huang et al., 2002):

Lemma 3.1. *Let $N(S_t)$, $N(S_b)$, and $N(S_w)$ denote the null spaces of S_t, S_b, and S_w, respectively. Then $N(S_t) = N(S_b) \cap N(S_w)$.*

Proof. Let y belong to the null space of S_t, that is $S_t y = 0$. Since $S_t = S_b + S_w$, we have $0 = y^T S_t y = y^T S_b y + y^T S_w y$. It follows that $y^T S_b y = y^T S_w y = 0$, and $S_b y = S_w y = 0$, since both S_b and S_t are positive semi-definite. Hence, $y \in N(S_b) \cap N(S_w)$.

Oh the other hand, if $y \in N(S_b) \cap N(S_w)$, then $y^T S_b y = y^T S_w y = 0$. It follows that $y^T S_t y = y^T S_b y + y^T S_w y = 0$, and $S_t y = 0$. Hence, $y \in N(S_t)$. □

We can efficiently remove the null space of S_t as follows. Let $H_t = U \Sigma V^T$ be the SVD of H_t, where H_t is defined in Eq. (3), 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 with the diagonal entries sorted in the nonincreasing order, and $t = \text{rank}(S_t)$. Then

$$S_t = H_t H_t^T = U \Sigma \Sigma^T U^T = U \begin{pmatrix} \Sigma_t^2 & 0 \\ 0 & 0 \end{pmatrix} U^T. \quad (7)$$

Let $U = (U_1, U_2)$ be a partition of U with $U_1 \in \mathbb{R}^{m \times t}$ and $U_2 \in \mathbb{R}^{m \times (m-t)}$. Then the null space of S_t can be removed by projecting the data onto the subspace spanned by the columns of U_1. Let $\tilde{S}_b$, $\tilde{S}_w$, and $\tilde{S}_t$ be the scatter matrices after the removal of the null space of S_t. That is, $\tilde{S}_b = U_1^T S_b U_1$, $\tilde{S}_w = U_1^T S_w U_1$, and $\tilde{S}_t = U_1^T S_t U_1$. Note that only U_1 is involved for the projection. We can thus apply the reduced SVD computation (Golub & Van Loan, 1996) on H_t with the time complexity of $O(mn^2)$, instead of $O(m^2 n)$. When the data dimension m is much larger than the sample size n, this leads to a big reduction in terms of computational cost.

With the computed U_1, the optimal transformation of NLDA is given by $G = U_1 N$, where N solves the following optimization problem:

$$N = \text{argmax}_{N^T \tilde{S}_w N = 0} \text{trace}(N^T \tilde{S}_b N). \quad (8)$$

That is, the columns of N lie in the null space of $\tilde{S}_w$, while maximizing $\text{trace}(N^T \tilde{S}_b N)$.

Let W be the matrix so that the columns of W span the null space of $\tilde{S}_w$. Then $N = WM$, for some matrix M, which is to be determined next. Since the

constraint in Eq. (8) is satisfied with $N = WM$ for any M, the optimal M can be computed by maximizing

$$\text{trace}(M^T W^T \tilde{S}_b W M).$$

By imposing the orthogonality constraint on M (Huang et al., 2002), the optimal M is given by the eigenvectors of $W^T \tilde{S}_b W$ corresponding to the nonzero eigenvalues. With the computed U_1, W, and M above, the optimal transformation of NLDA is given by

$$G = U_1 W M.$$

In (Huang et al., 2002), the matrix W is computed via the eigen-decomposition of $\tilde{S}_w$. More specifically, let

$$\tilde{S}_w = [W, \tilde{W}] \begin{pmatrix} 0 & 0 \\ 0 & \Delta_w \end{pmatrix} [W, \tilde{W}]^T$$

be its eigen-decomposition, where $[W, \tilde{W}]$ is orthogonal and Δ_w is diagonal with positive diagonal entries. Then, W forms the null space of $\tilde{S}_w$.

4. Relationship between NLDA and OLDA

Both the NLDA algorithm from Section 3 and the OLDA algorithm from Section 2.1 result in orthogonal transformations. Our empirical results show that they often lead to similar performance, especially for high-dimensional data, which implies there is an intrinsic relationship between these two algorithms. In this section, we examine the relationship between NLDA and OLDA in more detail. More specifically, we show that NLDA is equivalent to OLDA, under a mild condition

$$C1 : \text{rank}(S_t) = \text{rank}(S_b) + \text{rank}(S_w),$$

which holds in many applications involving high-dimensional data (see Section 5). It is easy to verify from the definition of scatter matrices that $\text{rank}(S_t) \leq \text{rank}(S_b) + \text{rank}(S_w)$.

From Eq. (7), the null space of S_t, that is, U_2, can be removed, as follows:

$$\tilde{S}_t = U_1^T S_t U_1 = U_1^T S_b U_1 + U_1^T S_w U_1 = \tilde{S}_w + \tilde{S}_b \in \mathbb{R}^{t \times t}.$$

Since the null space of S_t is the intersection of the null space of S_b and the null spaces of S_w, the following equalities hold: $\text{rank}(\tilde{S}_t) = \text{rank}(S_t) = t$, $\text{rank}(\tilde{S}_b) = \text{rank}(S_b)$, and $\text{rank}(\tilde{S}_w) = \text{rank}(S_w)$. Thus condition C1 is equivalent to

$$\text{rank}(\tilde{S}_t) = \text{rank}(\tilde{S}_b) + \text{rank}(\tilde{S}_w).$$

The null space of $\tilde{S}_b$ and the null space of $\tilde{S}_w$ are critical in our theoretical analysis. We will first study the relationship between these two null spaces in the following lemma.

Lemma 4.1. *Let $\tilde{S}_t$, $\tilde{S}_b$, and $\tilde{S}_w$ be defined as above. Let $\{w_1, \cdots, w_r\}$ forms an orthonormal basis for the null space of $\tilde{S}_w$, and let $\{b_1, \cdots, b_s\}$ forms an orthonormal basis for the null space of $\tilde{S}_b$. Then, $\{w_1, \cdots, w_r, b_1, \cdots, b_s\}$ are linearly independent.*

Proof. Prove by contradiction. Assume there exist α_i's and β_j's, not all zeros, such that

$$\sum_{i=1}^{r} \alpha_i w_i + \sum_{j=1}^{s} \beta_j b_j = 0.$$

It follows that

$$\begin{aligned} 0 &= (\sum_{i=1}^{r} \alpha_i w_i + \sum_{j=1}^{s} \beta_j b_j)^T \tilde{S}_w (\sum_{i=1}^{r} \alpha_i w_i + \sum_{j=1}^{s} \beta_j b_j) \\ &= (\sum_{j=1}^{s} \beta_j b_j)^T \tilde{S}_w (\sum_{j=1}^{s} \beta_j b_j), \end{aligned}$$

since w_i's lie in the null space of $\tilde{S}_w$. Hence,

$$(\sum_{j=1}^{s} \beta_j b_j)^T \tilde{S}_t (\sum_{j=1}^{s} \beta_j b_j) = 0,$$

as

$$(\sum_{j=1}^{s} \beta_j b_j)^T \tilde{S}_b (\sum_{j=1}^{s} \beta_j b_j) = 0.$$

Since $\tilde{S}_t$ is nonsingular, we have $\sum_{j=1}^{s} \beta_j b_j = 0$. Thus $\beta_j = 0$, for all j, since $\{b_1, \cdots, b_s\}$ forms an orthonormal basis for the null space of $\tilde{S}_b$. Similarly, we have $\alpha_i = 0$, for all i. This contradicts our original assumption that not all of the α_i's and the β_j's are zero, Thus, $\{w_1, \cdots, w_r, b_1, \cdots, b_s\}$ are linearly independent. □

Next, we show how to compute the optimal transformation of NLDA using these two null spaces. Recall that in NLDA, the null space of S_t may be removed first. In the following, we work on the reduced scatter matrices $\tilde{S}_w$, $\tilde{S}_b$, and $\tilde{S}_t$ directly as in Lemma 4.1. The main result is summarized in the following theorem.

Theorem 4.1. *Let U_1, $\tilde{S}_t$, $\tilde{S}_b$, and $\tilde{S}_w$ be defined as above and $t = \text{rank}(\tilde{S}_t)$. Let $R = [W, B]$, where $W = [w_1, \cdots, w_r]$, $B = [b_1, \cdots, b_s]$, and $\{w_1, \cdots, w_r, b_1, \cdots, b_s\}$ are defined as in Lemma 4.1. Assume that the condition C1: $\text{rank}(S_t) = \text{rank}(S_b) + \text{rank}(S_w)$ holds. Then $G = U_1 W M$ solves the optimization problem in Eq. (8), where the matrix M, consisting of the eigenvectors of $W^T \tilde{S}_b W^T$, is orthogonal.*

Proof. From Lemma 4.1, $\{w_1, \cdots, w_r, b_1, \cdots, b_s\} \in \mathbb{R}^t$ is linearly independent. Condition C1 implies that $t = r + s$. Thus $\{w_1, \cdots, w_r, b_1, \cdots, b_s\}$ forms a basis for $\mathbb{R}^t$, that is, $R = [W, B]$ is nonsingular. It follows that

$$\begin{aligned} R^T \tilde{S}_t R &= R^T \tilde{S}_b R + R^T \tilde{S}_w R \\ &= \begin{pmatrix} W^T \tilde{S}_b W & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & B^T \tilde{S}_w B \end{pmatrix}. \end{aligned}$$

Since matrix $R^T \tilde{S}_t R$ has full rank, $W^T \tilde{S}_b W$, the projection of $\tilde{S}_b$ onto the null space of $\tilde{S}_w$, is nonsingular. Let $W^T \tilde{S}_b W = M \Delta_b M^T$ be the eigen-decomposition of $W^T \tilde{S}_b W$, where M is orthogonal and Δ_b is diagonal with positive diagonal entries (note that $W^T \tilde{S}_b W$ is positive definite). Then, from Section 3, the optimal projection G of NLDA is given by $G = U_1 W M$. □

Recall that the matrix M in NLDA is computed so that $\text{trace}(M^T W^T \tilde{S}_b W M)$ is maximized. Since $\text{trace}(QAQ^T) = \text{trace}(A)$ for any orthogonal Q, the solution in NLDA is invariant under an arbitrary orthogonal transformation. Thus $G = U_1 W$ is also a solution to NLDA, since M is orthogonal, as summarized in the following corollary.

Corollary 4.1. *Assume condition C1: rank(S_t) = rank(S_b) + rank(S_w) holds. Let U_1 and W be defined as in Theorem 4.1. Then $G = U_1 W$ solves the optimization problem in Eq. (8). That is, $G = U_1 W$ is an optimal transformation of NLDA.*

Next, we show the equivalence relationship between NLDA and OLDA under the condition C1. The main result is summarized in the following theorem.

Theorem 4.2. *Assume that condition C1: rank(S_t) = rank(S_b) + rank(S_w) holds. Let U_1 and W be defined as in Theorem 4.1. Then, $G = U_1 W$ solves the optimization problem in Eq. (5). That is, under the given assumption, OLDA and NLDA are equivalent.*

Proof. Recall that the optimization involved in OLDA is

$$G = \text{argmax}_{G \in \mathbb{R}^{m \times \ell} : G^T G = I_\ell} \text{trace}\left((S_t^L)^+ S_b^L \right), \quad (9)$$

where $S_t^L = G^T S_t G$ and $S_b^L = G^T S_b G$. From Section 2.1, the maximum number, ℓ, of discriminant vectors is no larger than q, the rank of S_b. Recall that

$$q = \text{rank}(S_b) = \text{rank}(\tilde{S}_b) = \text{rank}(\tilde{S}_t) - \text{rank}(\tilde{S}_w) = r,$$

where r is the dimension of the null space of $\tilde{S}_w$.

Based on the property of the trace of matrices, we have $\text{trace}\left((S_t^L)^+ S_b^L \right) + \text{trace}\left((S_t^L)^+ S_w^L \right) = \text{trace}\left((S_t^L)^+ S_t^L \right) = \text{rank}(S_t^L) \le q = r$, where the second equality follows since $\text{trace}\,(A^+ A) = \text{rank}(A)$ for any square matrix A, and the inequality follows since the rank of $S_t^L \in \mathbb{R}^{\ell \times \ell}$ is at most $\ell \le q$.

It follows that $\text{trace}\left((S_t^L)^+ S_b^L \right) \le r$, since $\text{trace}\left((S_t^L)^+ S_w^L \right)$, the trace of the product of two positive semi-definite matrices, is always nonnegative. Next, we show that the maximum is achieved, when $G = U_1 W$.

Recall that W, the null space $\tilde{S}_w$ has dimension r. That is, $W \in \mathbb{R}^{t \times r}$. It follows that $(U_1 W)^T S_t (U_1 W) \in \mathbb{R}^{r \times r}$, and $\text{rank}((U_1 W)^T S_t (U_1 W)) = r$. Furthermore,

$$(U_1 W)^T S_w (U_1 W) = W^T \tilde{S}_w W = 0,$$

as W forms the null space of $\tilde{S}_w$. It follows that,

$$\text{trace}\left(\left((U_1 W)^T S_t (U_1 W) \right)^+ (U_1 W)^T S_w (U_1 W) \right) = 0.$$

Hence,

$$\text{trace}\left(\left((U_1 W)^T S_t (U_1 W) \right)^+ (U_1 W)^T S_b (U_1 W) \right) = r,$$

as

$$\text{trace}\left(\left((U_1 W)^T S_t (U_1 W) \right)^+ (U_1 W)^T S_t (U_1 W) \right) = r,$$

and $S_t = S_b + S_w$. Thus $G = U_1 W$ solves the optimization problem in Eq. (5). That is, OLDA and NLDA are equivalent. □

Theorem 4.2 shows that under the condition C1, OLDA and NLDA are equivalent. Next, we show that condition C1 holds when the data points are linearly independent as summarized below.

Theorem 4.3. *Assume that the condition C2, that the n data points in the data matrix $A \in \mathbb{R}^{m \times n}$ are linearly independent, holds. Then condition C1: rank(S_t) = rank(S_b) + rank(S_w) holds.*

Proof. Since the n columns in A are linearly independent, $H_t = A - ce^T$ is of rank $n-1$. That is, $\text{rank}(S_t) = n-1$. Next we show that $\text{rank}(S_b) = k-1$ and $\text{rank}(S_w) = n-k$. Thus condition C1 holds.

It is easy to verify that $\text{rank}(S_b) \le k-1$ and $\text{rank}(S_w) \le n-k$. We have

$$\begin{aligned} n-1 &= \text{rank}(S_t) \le \text{rank}(S_b) + \text{rank}(S_w) \\ &\le (k-1) + (n-k) = n-1. \end{aligned} \quad (10)$$

All inequalities in Eq. (10) become equalities. That is, $\text{rank}(S_b) = k-1$, $\text{rank}(S_w) = n-k$, and $\text{rank}(S_t) = \text{rank}(S_b) + \text{rank}(S_w)$. Thus, condition C1 holds. □

Table 1. Statistics of our test datasets.

dataset	size (n)	dimensionality (m)	class number (k)
re1	490	3759	5
re0	320	2887	4
tr41	210	7454	7
ORL	400	10304	40
AR	650	8888	50
PIX	300	10000	30
GCM	198	16063	14
colon	62	2000	2
ALLAML4	72	7129	4

Our experimental results in the next section show that for high-dimensional data, the linear independence condition C2, holds in many cases, while condition C1 is satisfied in most cases. This explains why NLDA and OLDA often achieve the same performance in many applications involving high-dimensional data.

5. Experimental Studies

In this section, we perform extensive experimental studies to evaluate the theoretical results presented in this paper. Section 5.1 describes our test datasets. We perform a detailed comparison between NLDA and OLDA in Section 5.2. Results are consistent with our theoretical analysis. The K-Nearest-Neighbor (K-NN) algorithm with $K = 1$ is used as the classifier.

5.1. Datasets

We used nine high-dimensional datasets from various data sources in our experimental studies, including text documents, face images, and gene expression data. re1, re0, and tr41 are three text document datasets, where re1 and re0 are derived from *Reuters-21578* text categorization test collection Distribution 1.0[1], and tr41 is derived from the TREC-5, TREC-6, and TREC-7 collections[2]; ORL[3], AR[4], and PIX[5] are three face image datasets; GCM, colon, and ALLAML4 are three gene expression datasets used in (Ye et al., 2004b). Table 1 summarizes the statistics of our test datasets.

5.2. Comparison of NLDA and OLDA

In this experiment, we did a comparative study of NLDA and OLDA. For all nine datasets, we performed our study by repeated random splittings into training and test sets. The data was partitioned randomly into a training set, where each class consists of two-thirds of the whole class and a test set with each class consisting of one-third of the whole class. The splitting was repeated 10 times and the resulting accuracies of different algorithms are summarized in Table 2. The rank of three scatter matrices, S_b, S_b, and S_w, for each of the splitting is also reported.

The main observations from Table 2 include:

(1) Condition C1: $\text{rank}(S_t) = \text{rank}(S_b) + \text{rank}(S_w)$ is satisfied in all cases except re0. For the re0 dataset, either $\text{rank}(S_t) = \text{rank}(S_b) + \text{rank}(S_w)$ or $\text{rank}(S_t) = \text{rank}(S_b) + \text{rank}(S_w) - 1$ holds, that is, condition C1 is not severely violated. Note that re0 has the smallest number of dimensions among the nine datasets. From the experiments, we may infer that condition C1 is likely to hold for high-dimensional data.

(2) NLDA and OLDA have the same classification accuracy for all cases when condition C1 holds, which confirms the theoretical analysis in Section 4. This explains why NLDA and OLDA often achieve similar performance for high-dimensional data.

(3) The numbers of training data points for the nine high-dimensional data (in the same order as in the table) are 325, 212, 140, 280, 450, 210, 125, 68, and 48, respectively. By examining the rank of S_t in Table 2, we can observe that the training data in six out of nine datasets, including tr41, ORL, AR, GCM, colon, and ALLAML4, are linearly independent. That is, the independence assumption C2 from Theorem 4.3 holds for these datasets. It is clear from the table that for these six datasets, condition C1 holds and NLDA and OLDA achieve the same performance. These are consistent with the theoretical analysis in Section 4.

(4) For the re0 dataset, where condition C1 does not hold, i.e., $\text{rank}(S_t) < \text{rank}(S_b) + \text{rank}(S_w)$, OLDA achieves a higher classification accuracy than NLDA. Recall that the reduced dimension of OLDA equals $\text{rank}(S_b) \equiv q$. The reduced dimension in NLDA equals the dimension of the null space of $\tilde{S}_w$, which equals $\text{rank}(S_t) - \text{rank}(S_w) < \text{rank}(S_b)$. That is, OLDA keeps more dimensions than NLDA. Experimental results in re0 show that these extra dimensions used in OLDA improve its classification performance.

6. Conclusions

In this paper, we present a computational and theoretical analysis of two LDA based algorithms, including null space based LDA and orthogonal LDA.

[1] http://www.research.att.com/∼lewis

[2] http://trec.nist.gov

[3] http://www.uk.research.att.com/facedatabase.html

[4] http://rvl1.ecn.purdue.edu/∼aleix/aleix_face_DB.html

[5] http://peipa.essex.ac.uk/ipa/pix/faces/manchester/test-hard/

Table 2. Comparison of classification accuracy (in percentage) between NLDA and OLDA. Ten different splittings into training and test sets of ratio 2:1 (for each of the k classes) are applied. The rank of three scatter matrices for each splitting is reported.

Data	Method	Ten different splittings into training and test sets of ratio 2:1									
re1	NLDA	92.73	93.33	93.33	93.94	94.55	95.15	96.36	95.15	92.12	93.94
	OLDA	92.73	93.33	93.33	93.94	94.55	95.15	96.36	95.15	92.12	93.94
	S_b	4	4	4	4	4	4	4	4	4	4
	S_w	316	318	319	316	316	320	316	318	317	318
	S_t	320	322	323	320	320	324	320	322	321	322
re0	NLDA	64.81	62.04	64.81	68.52	87.96	70.37	71.30	73.15	87.04	75.93
	OLDA	75.93	75.00	77.78	74.07	87.96	80.56	74.07	78.70	87.04	79.63
	S_b	3	3	3	3	3	3	3	3	3	3
	S_w	205	204	203	203	205	204	201	203	203	205
	S_t	207	206	205	205	208	206	203	205	206	207
tr41	NLDA	97.14	95.71	97.14	98.57	97.14	98.57	100.0	95.71	98.57	95.71
	OLDA	97.14	95.71	97.14	98.57	97.14	98.57	100.0	95.71	98.57	95.71
	S_b	6	6	6	6	6	6	6	6	6	6
	S_w	133	133	133	133	133	133	133	133	133	133
	S_t	139	139	139	139	139	139	139	139	139	139
ORL	NLDA	99.17	96.67	98.33	98.33	95.00	95.83	98.33	97.50	98.33	95.83
	OLDA	99.17	96.67	98.33	98.33	95.00	95.83	98.33	97.50	98.33	95.83
	S_b	39	39	39	39	39	39	39	39	39	39
	S_w	240	240	240	240	240	240	240	240	240	240
	S_t	279	279	279	279	279	279	279	279	279	279
AR	NLDA	96.50	94.50	96.50	94.00	93.50	94.50	93.50	97.00	94.00	96.00
	OLDA	96.50	94.50	96.50	94.00	93.50	94.50	93.50	97.00	94.00	96.00
	S_b	49	49	49	49	49	49	49	49	49	49
	S_b	400	400	400	400	400	400	400	400	400	400
	S_b	449	449	449	449	449	449	449	449	449	449
PIX	NLDA	98.89	97.78	98.89	97.78	98.89	98.89	98.89	97.78	98.89	97.78
	OLDA	98.89	97.78	98.89	97.78	98.89	98.89	98.89	97.78	98.89	97.78
	S_b	29	29	29	29	29	29	29	29	29	29
	S_w	178	179	179	179	178	180	179	179	180	178
	S_t	207	208	208	208	207	209	208	208	209	207
GCM	NLDA	81.54	80.00	81.54	83.08	84.62	87.69	75.38	78.46	84.62	83.08
	OLDA	81.54	80.00	81.54	83.08	84.62	87.69	75.38	78.46	84.62	83.08
	S_b	13	13	13	13	13	13	13	13	13	13
	S_w	111	111	111	111	111	111	111	111	111	111
	S_t	124	124	124	124	124	124	124	124	124	124
colon	NLDA	91.18	94.12	100.0	97.06	91.18	91.18	97.06	94.12	94.12	97.06
	OLDA	91.18	94.12	100.0	97.06	91.18	91.18	97.06	94.12	94.12	97.06
	S_b	1	1	1	1	1	1	1	1	1	1
	S_w	66	66	66	66	66	66	66	66	66	66
	S_t	67	67	67	67	67	67	67	67	67	67
ALLAML4	NLDA	95.83	91.67	95.83	95.83	87.50	95.83	95.83	100.0	91.67	95.83
	OLDA	95.83	91.67	95.83	95.83	87.50	95.83	95.83	100.0	91.67	95.83
	S_b	3	3	3	3	3	3	3	3	3	3
	S_w	44	44	44	44	44	44	44	44	44	44
	S_t	47	47	47	47	47	47	47	47	47	47

Both NLDA and OLDA result in orthogonal transformations. However, they applied different schemes in deriving the optimal transformation. Our theoretical analysis in this paper shows that under a mild Condition C1 which holds in many applications involving high-dimensional data, NLDA is equivalent to OLDA.

We have performed extensive experimental studies on nine high-dimensional datasets. Results show that condition C1 holds for eight out of nine datasets. Results are also consistent with our theoretical analysis. That is, for all cases when condition C1 holds, NLDA and OLDA achieve the same classification performance. We also observe that for other cases with condition C1 violated, OLDA outperforms NLDA, due to the extra number of dimensions used in OLDA.

Our experimental studies have shown that condition C1 holds for most high-dimensional datasets. We plan to carry out theoretical analysis on this property in the future. Some of the theoretical results in (Hall et al., 2005) may be useful for our analysis.

The algorithms in (Yang et al., 2005; Yu & Yang, 2001) are closely related to the null space based algorithm discussed in this paper. The analysis presented in this paper may be useful in understanding why these algorithms perform well in many applications, especially in face recognition. We plan to explore this further in the future.

Acknowledgements

We thank the reviewers for helpful comments. Research of JY is sponsored, in part, by the Center for Evolutionary Functional Genomics of the Biodesign Institute at the Arizona State University.

References

Belhumeour, P., Hespanha, J., & Kriegman, D. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans Pattern Analysis and Machine Intelligence*, *19*, 711–720.

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

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

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification*. Wiley.

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.

Foley, D., & Sammon, J. (1975). An optimal set of discriminant vectors. *IEEE Trans Computers*, *24*, 281–289.

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.

Hall, P., Marron, J., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society series B*, *67*, 427–444.

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

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

Huang, R., Liu, Q., Lu, H., & Ma, S. (2002). Solving the small sample size problem of LDA. *Proc. International Conference on Pattern Recognition* (pp. 29–32).

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.

Lu, J., Plataniotis, K., & Venetsanopoulos, A. (2003). Face recognition using kernel direct discriminant analysis algorithms. *IEEE Trans. Neural Networks*, *14*, 117–126.

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

Turk, M., & Pentland, A. (1991). Face recognition using eigenfaces. *Proc. Computer Vision and Pattern Recognition Conference* (pp. 586–591).

Yang, J., Frangi, A., Yang, J., Zhang, D., & Jin, Z. (2005). KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *27*, 230– 244.

Ye, J. (2005). Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, *6*, 483–502.

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

Ye, J., Li, T., Xiong, T., & Janardan, R. (2004b). Using uncorrelated discriminant analysis for tissue classification with gene expression data. *IEEE/ACM Trans. Computational Biology and Bioinformatics*, *1*, 181–190.

Yu, H., & Yang, J. (2001). A direct LDA algorithm for high-dimensional data with applications to face recognition. *Pattern Recognition*, *34*, 2067–2070.