Least Squares Linear Discriminant Analysis

Jieping Ye JIEPING.YE@ASU.EDU
Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287 USA

Abstract

Linear Discriminant Analysis (LDA) is a well-known method for dimensionality reduction and classification. LDA in the binary-class case has been shown to be equivalent to linear regression with the class label as the output. This implies that LDA for binary-class classifications can be formulated as a least squares problem. Previous studies have shown certain relationship between multivariate linear regression and LDA for the multi-class case. Many of these studies show that multivariate linear regression with a specific class indicator matrix as the output can be applied as a preprocessing step for LDA. However, directly casting LDA as a least squares problem is challenging for the multi-class case. In this paper, a novel formulation for multivariate linear regression is proposed. The equivalence relationship between the proposed least squares formulation and LDA for multi-class classifications is rigorously established under a mild condition, which is shown empirically to hold in many applications involving high-dimensional data. Several LDA extensions based on the equivalence relationship are discussed.

1. Introduction

Linear Discriminant Analysis (LDA) is a well-known method for dimensionality reduction and classification that projects high-dimensional data onto a low-dimensional space where the data achieves maximum class separability (Duda et al., 2000; Fukunaga, 1990; Hastie et al., 2001). The derived features in LDA are linear combinations of the original features, where the coefficients are from the transformation matrix. The optimal projection or transformation in classical LDA is obtained by minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination. It has been applied successfully in many applications (Belhumeour et al., 1997; Swets & Weng, 1996; Dudoit et al., 2002) including face recognition and microarray gene expression data analysis. The optimal transformation is readily computed by solving a generalized eigenvalue problem. The original LDA formulation, known as the Fisher Linear Discriminant Analysis (FLDA) (Fisher, 1936) deals with binary-class classifications. The key idea in FLDA is to look for a direction that separates the class means well (when projected onto that direction) while achieving a small variance around these means.

FLDA bears strong connections to linear regression with the class label as the output. It has been shown (Duda et al., 2000; Mika, 2002) that FLDA is equivalent to a least squares problem. Many real-world applications deal with multi-class classifications, and LDA is generally used to find a subspace with $k-1$ dimensions for multi-class problems, where k is the number of classes in the training dataset (Fukunaga, 1990; Hastie et al., 2001). However, directly casting LDA as a least squares problem is challenging for multi-class problems (Duda et al., 2000; Hastie et al., 2001; Zhang & Riedel, 2005).

Multivariate linear regression with a specific class indicator matrix has been considered in (Hastie et al., 2001) for multi-class classifications. It follows the general framework of linear regression with multiple outputs. As pointed out in (Pages 83–84, Hastie et al. (2001)), this approach has a serious problem when the number of classes $k \geq 3$, especially prevalent when k is large. More specifically, classes can be masked by others due to the rigid nature of the regression model, which is not the case for LDA (Hastie et al., 2001). This multivariate linear regression model is in general different from LDA. However, there is a close connection between multivariate linear regression and LDA. More specifically, it can be shown (Hastie et al., 1994; Hastie et al., 2001) that LDA applied to the trans-

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007. Copyright 2007 by the author(s)/owner(s).

formed space by multivariate linear regression with a specific class indicator matrix as the output is identical to LDA applied to the original space. In this case, multivariate linear regression is applied as a preprocessing step for LDA. One natural question is whether LDA in the multi-class case can be directly formulated as a least squares problem.

In this paper, we propose a novel formulation for multivariate linear regression based on a new class indicator matrix. We establish the equivalence relationship between the proposed least squares formulation and LDA under a mild condition (see Section 5), which holds in many applications involving high-dimensional and undersampled data. We call the proposed LDA formulation **L**east **S**quares **L**inear **D**iscriminant **A**nalysis (or LS-LDA in short). We have conducted experiments using a collection of high-dimensional datasets from various data sources, including text documents, face images, and microarray gene expression data. Experimental results are consistent with the presented theoretical analysis.

2. Overview of Linear Discriminant Analysis

Given a dataset that consists of n samples $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$, and $y_i \in \{1, 2, \cdots, k\}$ denotes the class label of the i-th sample, n is the sample size, d is the data dimensionality, and k is the number of classes. Let the data matrix $X = [x_1, x_2, \cdots, x_n]$ be partitioned into k classes as $X = [X_1, \cdots, X_k]$, where $X_j \in \mathbb{R}^{d \times n_j}$, n_j is the size of the j-th class X_j, and $\sum_{j=1}^k n_j = n$. Classical LDA computes a linear transformation $G \in \mathbb{R}^{d \times \ell}$ that maps x_i in the d-dimensional space to a vector x_i^L in the ℓ-dimensional space as follows: $x_i \in \mathbb{R}^d \rightarrow x_i^L = G^T x_i \in \mathbb{R}^\ell \, (\ell < d)$. In discriminant analysis (Fukunaga, 1990), three scatter matrices, called *within-class*, *between-class* and *total* scatter matrices are defined as follows:

$$S_w = \frac{1}{n} \sum_{j=1}^k \sum_{x \in X_j} (x - c^{(j)})(x - c^{(j)})^T, \quad (1)$$

$$S_b = \frac{1}{n} \sum_{j=1}^k n_j (c^{(j)} - c)(c^{(j)} - c)^T, \quad (2)$$

$$S_t = \frac{1}{n} \sum_{i=1}^n (x_i - c)(x_i - c)^T, \quad (3)$$

where $c^{(j)}$ is the *centroid* of the j-th class, and c is the *global centroid*. It follows from the definition that $S_t = S_b + S_w$. Furthermore, trace(S_w) measures the within-class cohesion, while trace(S_b) measures the between-class separation. In the lower-dimensional space resulting from the linear transformation G, the scatter matrices become

$$S_w^L = G^T S_w G, \quad S_b^L = G^T S_b G, \quad S_t^L = G^T S_t G. \quad (4)$$

An optimal transformation G would maximize trace(S_b^L) and minimize trace(S_w^L) simultaneously, which is equivalent to maximizing trace(S_b^L) and minimizing trace(S_t^L) simultaneously, since $S_t^L = S_w^L + S_b^L$.

The optimal transformation, G^{LDA}, of LDA is computed by solving the following optimization problem (Duda et al., 2000; Fukunaga, 1990):

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

The optimal G^{LDA} consists of the top eigenvectors of $S_t^{-1} S_b$ corresponding to the nonzero eigenvalues (Fukunaga, 1990), provided that the total scatter matrix S_t is nonsingular. In the following discussion, we consider the more general case when S_t may be singular. G^{LDA} consists of the eigenvectors of $S_t^+ S_b$ corresponding to the nonzero eigenvalues. Here S_t^+ denotes the pseudo-inverse of S_t (Golub & Van Loan, 1996). Note that when S_t is nonsingular, S_t^+ equals S_t^{-1}.

The above LDA formulation is an extension of the original Fisher Linear Discriminant Analysis (FLDA) (Fisher, 1936), which deals with binary-class problems, i.e., $k = 2$. The optimal transformation, G^F, of FLDA is of rank one and is given by (Duda et al., 2000)

$$G^F = S_t^+ (c^{(1)} - c^{(2)}). \quad (6)$$

Note that G^F is invariant of scaling. That is, αG^F, for any $\alpha \neq 0$ is also a solution to FLDA.

3. Linear Regression versus Fisher LDA

Given a dataset of two classes, $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, and $y_i \in \{-1, 1\}$, the linear regression model with the class label as the output has the following form: $f(x) = x^T w + b$, where $w \in \mathbb{R}^d$ is the weight vector, and b is the bias of the linear model. A popular approach for estimating w and b is the least squares, in which the following objective function is minimized:

$$L(w, b) = \frac{1}{2} ||X^T w + be - y||^2 = \frac{1}{2} \sum_{i=1}^n ||f(x_i) - y_i||^2, \quad (7)$$

where $X = [x_1, x_2, \cdots, x_n]$ is the data matrix, e is the vectors of all ones, and y is the vector of class labels. Assume that both $\{x_i\}$ and $\{y_i\}$ have been centered, i.e., $\sum_{i=1}^n x_i = 0$ and $\sum_{i=1}^n y_i = 0$. It follows that

$$y_i \in \{-2n_2/n, 2n_1/n\},$$

where n_1 and n_2 denote the number of samples from the negative and positive classes, respectively. The bias term b becomes zero and we look for a linear model $f(x) = x^T w$ by minimizing

$$L(w) = \frac{1}{2}||X^T w - y||^2. \tag{8}$$

The optimal w is given by (Duda et al., 2000; Golub & Van Loan, 1996; Hastie et al., 2001) $w = \left(XX^T\right)^+ Xy$. Note that $XX^T = nS_t$ (data matrix X has been centered) and $Xy = \frac{2n_1n_2}{n}(c^{(1)} - c^{(2)})$. It follows that

$$w = \frac{2n_1n_2}{n^2} S_t^+ (c^{(1)} - c^{(2)}) = \frac{2n_1n_2}{n^2} G^F,$$

where G^F is the optimal solution to FLDA in Eq. (6). Hence linear regression with the class label as the output is equivalent to Fisher LDA, as the projection in FLDA is invariant of scaling. More details on this equivalence relationship can be found at (Duda et al., 2000; Mika, 2002).

4. Multivariate Linear Regression with a Class Indicator Matrix

In the multiclass case, we are given a dataset that consists of n samples $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$, and $y_i \in \{1, 2, \cdots, k\}$ denotes the class label of the i-th sample, and $k > 2$. It is common to apply linear regression of a class membership indicator matrix $Y \in \mathbb{R}^{n\times k}$, which applies a vector valued class code for each of the samples (Hastie et al., 2001). There are several well-known indicator matrices in the literature. Denote $Y_1 = (Y_1(ij))_{ij} \in \mathbb{R}^{n\times k}$ and $Y_2 = (Y_2(ij))_{ij} \in \mathbb{R}^{n\times k}$ as the class indicator matrices as follows: $Y_1(ij) = 1$, if $y_i = j$, and $Y_1(ij) = 0$, otherwise; and $Y_2(ij) = 1$, if $y_i = j$, and $Y_2(ij) = -1/(k-1)$, otherwise. The first indicator matrix Y_1 is commonly used in connecting multi-class classification with linear regression (Hastie et al., 2001), while the second indicator matrix has recently been used in extending Support Vector Machines (SVM) to multiclass classification (Lee et al., 2004) and in generalizing the kernel target alignment measure (Guermeur et al., 2004), originally proposed in (Cristianini et al., 2001).

In multivariate linear regression (MLR), a k-tuple of separating functions $f(x) = (f_1(x), f_2(x), \cdots, f_k(x))$, for any $x \in \mathbb{R}^d$ is considered. Denote $\tilde{X} = [\tilde{x}_1, \cdots, \tilde{x}_n] \in \mathbb{R}^{d\times n}$, and $\tilde{Y} = \left(\tilde{Y}_{ij}\right) \in \mathbb{R}^{n\times k}$ as the centered data matrix X and the centered indicator matrix Y, respectively. That is, $\tilde{x}_i = x_i - \bar{x}$ and $\tilde{Y}_{ij} = Y_{ij} - \bar{Y}_j$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{Y}_j = \frac{1}{n}\sum_{i=1}^n Y_{ij}$. Then MLR determines the weight vectors, $\{w_j\}_{j=1}^k \in \mathbb{R}^d$, of the k linear models, $f_j(x) = x^T w_j$, for $j = 1, \cdots, k$, via the minimization of the following objective function:

$$L(W) = \frac{1}{2}||\tilde{X}^T W - \tilde{Y}||_F^2 = \frac{1}{2}\sum_{j=1}^k \sum_{i=1}^n ||f_j(\tilde{x}_i) - \tilde{Y}_{ij}||^2, \tag{9}$$

where $W = [w_1, w_2, \cdots, w_k]$ is the weight matrix, and $||\cdot||_F$ denotes the Frobenius norm of a matrix (Golub & Van Loan, 1996). The optimal W is given by (Hastie et al., 2001)

$$W = \left(\tilde{X}\tilde{X}^T\right)^+ \tilde{X}\tilde{Y}, \tag{10}$$

which is determined by $\tilde{X}$ and $\tilde{Y}$.

Both Y_1 and Y_2 defined above, as well as the one in (Park & Park, 2005) could be used to define the centered indicator matrix $\tilde{Y}$. An interesting connection between the linear regression model using Y_1 and LDA can be found in (Page 112, Hastie et al. (2001)). It can be shown that if $X^L = W_1^T \tilde{X}$ is the transformed data by W_1, where $W_1 = \left(\tilde{X}\tilde{X}^T\right)^+ \tilde{X}\tilde{Y}_1$ is the least squares solution in Eq. (10) using the centered indicator matrix $\tilde{Y}_1$, then LDA applied to X^L is identical to LDA applied to the original space. In this case, linear regression is applied as a preprocessing step before the classification, and is in general not equivalent to LDA. The second indicator matrix Y_2 has been used in SVM, and the resulting model using Y_2 is also not equivalent to LDA in general. This is also the case for the indicator matrix in (Park & Park, 2005). A natural question is whether there exists a class indicator matrix $\tilde{Y} \in \mathbb{R}^{n\times k}$, with which the multivariate linear regression is equivalent to LDA. If this is the case, then LDA can be formulated as a least squares problem in the multi-class case.

Note that in multivariate linear regression, each $\tilde{x}_i$ is transformed to $(f_1(\tilde{x}_i), \cdots, f_k(\tilde{x}_i))^T = W^T\tilde{x}_i$, and the centered data matrix $\tilde{X} \in \mathbb{R}^{d\times n}$ is transformed to $W^T\tilde{X} \in \mathbb{R}^{k\times n}$, thus achieving dimensionality reduction if $k < d$. Note that $W = \left(\tilde{X}\tilde{X}^T\right)^+ \tilde{X}\tilde{Y}$. A natural measure for evaluating $\tilde{Y}$ is the class discrimination used in LDA. We thus look for $\tilde{Y}$ which solves the following optimization problem: (The pseudo-inverse is applied to deal with the singular scatter matrix.)

$$\begin{aligned} \max_{\tilde{Y}} \quad & \text{trace}((W^T S_b W)(W^T S_t W)^+) \\ \text{subject to} \quad & W = \left(\tilde{X}\tilde{X}^T\right)^+ \tilde{X}\tilde{Y} \end{aligned} \tag{11}$$

In the following, we construct a specific class indicator matrix Y_3 and show that it solves the optimiza-

tion problem in Eq. (11). More importantly, we show in Section 5 that multivariate linear regression using indicator matrix Y_3 is equivalent to LDA under a mild condition, which has been shown empirically to hold for most high-dimensional and undersampled data. The indicator matrix $Y_3 = (Y_3(ij))_{ij} \in \mathbb{R}^{n\times k}$ is constructed as follows:

$$Y_3(ij) = \begin{cases} \sqrt{\frac{n}{n_j}} - \sqrt{\frac{n_j}{n}} & \text{if } y_i = j, \\ -\sqrt{\frac{n_j}{n}} & \text{otherwise,} \end{cases} \quad (12)$$

where n_j is the sample size of the j-th class, and n is the total sample size. It can be shown that Y_3 defined above has been centered (in terms of rows), and thus $\tilde{Y}_3 = Y_3$.

Define matrices H_w, H_b, and H_t as follows:

$$H_w = \frac{1}{\sqrt{n}}[X_1 - c^{(1)}(e^{(1)})^T, \cdots, X_k - c^{(k)}(e^{(k)})^T], \quad (13)$$

$$H_b = \frac{1}{\sqrt{n}}[\sqrt{n_1}(c^{(1)} - c), \cdots, \sqrt{n_k}(c^{(k)} - c)], \quad (14)$$

$$H_t = \frac{1}{\sqrt{n}}(X - ce^T), \quad (15)$$

where X_j is the data matrix of the j-th class, X is the data matrix, $c^{(j)}$ is the centroid of the j-th class, c is the global centroid, $e^{(j)}$ is the vector of all ones of length n_j and e is the vector of all ones of length n. Then S_w, S_b, and S_t can be expressed as follows:

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

Let $H_t = U\Sigma V^T$ be the Singular Value Decomposition (SVD) (Golub & Van Loan, 1996) of H_t, where H_t is defined in Eq. (15), U and V are orthogonal, $\Sigma = \mathrm{diag}(\Sigma_t, 0)$, $\Sigma_t \in \mathbb{R}^{t\times t}$ is diagonal, and $t = \mathrm{rank}(S_t)$. Then

$$S_t = H_t H_t^T = U\Sigma\Sigma^T U^T = U\mathrm{diag}(\Sigma_t^2, 0)U^T. \quad (16)$$

Let $U = (U_1, U_2)$ be a partition of U, such that $U_1 \in \mathbb{R}^{d\times t}$ and $U_2 \in \mathbb{R}^{d\times(d-t)}$. That is, U_2 lies in the null space of S_t, i.e., $U_2^T S_t U_2 = 0$. Since $S_t = S_b + S_w$, we have $0 = U_2^T S_t U_2 = U_2^T S_b U_2 + U_2^T S_w U_2$. Thus $U_2^T S_b U_2 = 0$, since S_w is positive semi-definite. It follows that

$$U^T S_b U = \begin{pmatrix} U_1^T S_b U_1 & 0 \\ 0 & 0 \end{pmatrix}. \quad (17)$$

Denote

$$B = \Sigma_t^{-1} U_1^T H_b \in \mathbb{R}^{t\times k}, \quad (18)$$

where H_b is defined in Eq. (14) and let

$$B = P\hat{\Sigma}Q^T \quad (19)$$

be the SVD of B, where P and Q are orthogonal and $\hat{\Sigma} \in \mathbb{R}^{t\times k}$ is diagonal. Since $S_b = H_b H_b^T$, we have

$$\Sigma_t^{-1} U_1^T S_b U_1 \Sigma_t^{-1} = BB^T = P\hat{\Sigma}\hat{\Sigma}^T P^T = P\Sigma_b P^T, \quad (20)$$

where

$$\Sigma_b = \hat{\Sigma}\hat{\Sigma}^T = \mathrm{diag}(\alpha_1^2, \cdots, \alpha_t^2), \quad (21)$$

$$\alpha_1^2 \geq \cdots \geq \alpha_q^2 > 0 = \alpha_{q+1}^2 = \cdots = \alpha_t^2, \quad (22)$$

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

It follows from Eqs. (16) and (17) that

$$\begin{aligned} S_t^+ S_b S_t^+ &= U_1 \Sigma_t^{-2} U_1^T S_b U_1 \Sigma_t^{-2} U_1^T \\ &= U_1 \Sigma_t^{-1} \left(P\Sigma_b P^T\right) \Sigma_t^{-1} U_1^T, \end{aligned} \quad (23)$$

where the last equality follows from Eq. (20). We have the following result:

Lemma 4.1. *Let P, U_1, Σ_t, $\tilde{X}$, and $\tilde{Y}$ be defined as above. Then $W^T S_b W = \frac{1}{n^2} F^T \Sigma_b F$, and $W^T S_t W = \frac{1}{n^2} F^T F$, where W is defined in Eq. (10), Σ_b is defined in Eq. (21), and $F = P^T \Sigma_t^{-1} U_1^T (\tilde{X}\tilde{Y})$.*

We are now ready to present the main result of this section, that is $\tilde{Y} = Y_3$, solves the optimization problem in Eq. (11), where Y_3 is defined in Eq. (12), as summarized in the following theorem:

Theorem 4.1. *Let S_b, S_t, Σ_b, W, and $\tilde{Y}$ be defined as above. Then for any $\tilde{Y}$, the following inequality holds: $\mathrm{trace}\left((W^T S_b W)(W^T S_t W)^+\right) \leq \mathrm{trace}(\Sigma_b)$. Furthermore, the equality holds when $\tilde{Y} = Y_3$, where Y_3 is defined in Eq. (12).*

With $\tilde{Y} = Y_3$ as the class indicator matrix, the optimal weight matrix W^{MLR} for multivariate linear regression (MLR) in Eq. (10) becomes

$$W^{MLR} = \left(\tilde{X}\tilde{X}^T\right)^+ \tilde{X}\tilde{Y} = (nS_t)^+ nH_b = S_t^+ H_b. \quad (24)$$

5. Relationship between Multivariate Linear Regression and LDA

Recall that in LDA, the optimal transformation matrix, G^{LDA}, consists of the top eigenvectors of $S_t^+ S_b$ corresponding to the nonzero eigenvalues. In this section, we study the relationship between $W^{MLR} = S_t^+ H_b$ in Eq. (24) and the eigenvectors of $S_t^+ S_b$. From Eqs. (16) and (17), we can decompose matrix $S_t^+ S_b$ as follows:

$$\begin{aligned} S_t^+ S_b &= U \begin{pmatrix} \Sigma_t^{-1} BB^T \Sigma_t & 0 \\ 0 & 0 \end{pmatrix} U^T \\ &= U \begin{pmatrix} \Sigma_t^{-1} P & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_b & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P^T \Sigma_t & 0 \\ 0 & I \end{pmatrix} U^T \end{aligned}$$

where the equalities follow since

$$B = \Sigma_t^{-1} U_1^T H_b = P\hat{\Sigma}Q^T$$

is the SVD of B as in Eq. (19) and $\Sigma_b = \hat{\Sigma}\hat{\Sigma}^T$. Thus, the transformation matrix in LDA is given by $G^{LDA} = U_1\Sigma_t^{-1}P_q$, where P_q consists of the first q columns of P, since only the first q diagonal entries of Σ_b is nonzero. On the other hand,

$$\begin{aligned} S_t^+ H_b &= U \begin{pmatrix} \left(\Sigma_t^2\right)^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^T H_b \\ &= U_1\Sigma_t^{-1}\left(\Sigma_t^{-1}U_1^T H_b\right) \\ &= U_1\Sigma_t^{-1}B \\ &= U_1\Sigma_t^{-1}P\hat{\Sigma}Q^T \\ &= U_1\Sigma_t^{-1}P_q\left[\hat{\Sigma}_q, 0\right]Q^T \\ &= \left[G^{LDA}\Sigma_{bq}^{0.5}, 0\right]Q^T, \end{aligned} \tag{25}$$

where $\hat{\Sigma}_q, \Sigma_{bq} \in \mathbb{R}^{q\times q}$ consists of the first q rows and the first q columns of $\hat{\Sigma}, \Sigma_b$, respectively, the fifth equality follows since only the first q rows and the first q columns of $\hat{\Sigma}$ are nonzero and the last equality follows since $\Sigma_b = \hat{\Sigma}\hat{\Sigma}^T$. It follows that

$$W^{MLR} = \left[G^{LDA}\Sigma_{bq}^{0.5}, 0\right]Q^T,$$

where Q is orthogonal.

The K-Nearest-Neighbor (K-NN) algorithm (Duda et al., 2000) based on the Euclidean distance is commonly applied as the classifier in the dimensionality reduced (transformed) space of LDA. If we apply W^{MLR} for dimensionality reduction before K-NN, the matrix W^{MLR} is invariant of an orthogonal transformation, since any orthogonal transformation preserves all pairwise distance. Thus W^{MLR} is essentially equivalent to $\left[G^{LDA}\Sigma_{bq}^{0.5}, 0\right]$ or $G^{LDA}\Sigma_{bq}^{0.5}$, as the removal of zero columns does not change the pairwise distance either. The essential difference between W^{MLR} and G^{LDA} is thus the diagonal matrix $\Sigma_{bq}^{0.5}$.

Next, we show that matrix Σ_{bq} is an identity matrix of size q, that is, W^{MLR} and G^{LDA} are essentially equivalent, under a mild condition that the rank difference of the three scatter matrices is zero, that is, $\text{rank}(S_b) + \text{rank}(S_w) - \text{rank}(S_t) = 0$, which holds in many applications involving high-dimensional and undersampled data (Ye & Xiong, 2006). The main result is summarized in the following theorem:

Theorem 5.1. *Let $\Sigma_{bq} \in \mathbb{R}^{q\times q}$ consist of the first q rows and the first q columns of Σ_b as defined above, where Σ_b is defined in Eq. (21). Assume that the following equality holds: $rank(S_b)+rank(S_w)-rank(S_t) = 0$. Then $\Sigma_{bq} = I_q$, where I_q is the identity matrix of size q and $q = rank(S_b)$.*

6. Experiments

We performed our experimental studies using nine high-dimensional datasets, including text documents, face images, and gene expression data. DOC1, DOC2, and DOC3 are three text document datasets; ORL, AR, and PIX are three face image datasets; and GCM, ALL, and ALLAML are three gene expression datasets. The statistics of the datasets are summarized in Table 1 (the first column).

To compare LS-LDA and LDA, we use the K-NN algorithm with $K = 1$ as the classifier. For all datasets, we performed our study by repeated random splittings of the whole dataset into training and test sets as in (Dudoit et al., 2002). 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 for the ten splittings are summarized in Table 1. The rank difference of three scatter matrices, i.e., $\text{rank}(S_b) + \text{rank}(S_w) - \text{rank}(S_t)$, as well as the ratio of the largest to the smallest diagonal entries of $\Sigma_{bq}^{0.5}$, for each of the splitting is also reported. Recall from Theorem 5.1 that when the rank difference of the scatter matrices is zero, matrix Σ_{bq} equals to the identity matrix and the ratio of the largest to the smallest diagonal entries of $\Sigma_{bq}^{0.5}$ is 1.

We can observe from Table 1 that the rank difference, $\text{rank}(S_b)+\text{rank}(S_w)-\text{rank}(S_t)$, equals zero in all cases except the DOC2 dataset. For most datasets, the n data points are linearly independent, i.e., $\text{rank}(S_t) = n - 1$. In this case, the k centroids are also linearly independent, i.e., $\text{rank}(S_b) = k - 1$, while in S_w, each data point is subtracted by its class centroid and $\text{rank}(S_w) = n - k$. Hence, the rank difference is zero. Furthermore, LS-LDA and LDA achieve the same classification performance for all cases when the rank difference is zero. The empirical result confirms the theoretical analysis in Section 5. For DOC2, LS-LDA and LDA still achieve the same classification performance, although the rank difference is not zero.

Recall that the value of ratio denotes the ratio of the largest to the smallest diagonal entries of the matrix $\Sigma_{bq}^{0.5}$. From Table 1, the value of ratio equals 1 for all cases when the rank difference is zero. This is consistent with the theoretical result in Theorem 5.1. For DOC2, where the rank difference is not zero for several cases, the value of ratio is close to 1 for all cases. That is, matrix $\Sigma_{bq}^{0.5}$ is close to the identity matrix. This explains why LS-LDA and LDA achieve the same classification performance for DOC2, even though the rank difference is not zero.

Table 1. Comparison of classification accuracy (in percentage) between LS-LDA and LDA. Ten different splittings into training and test sets of ratio 2:1 (for each of the k classes) are applied. The rank difference (Diff) of three scatter matrices, i.e., $\text{rank}(S_b)+\text{rank}(S_w)-\text{rank}(S_t)$, for each splitting, as well as the ratio of the largest to the smallest diagonal entries of $\Sigma_{bq}^{0.5}$, is reported. n is the total sample size, d is the data dimensionality, and k is the total number of classes.

Dataset	Method/ratio/Diff	Ten different splittings into training and test sets of ratio 2:1									
DOC1	LS-LDA	93.33	93.33	91.52	95.15	94.55	93.94	93.94	95.15	93.33	93.33
$n = 490$	LDA	93.33	93.33	91.52	95.15	94.55	93.94	93.94	95.15	93.33	93.33
$d = 3759$	ratio	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k = 5$	Diff	0	0	0	0	0	0	0	0	0	0
DOC2	LS-LDA	75.93	67.59	78.70	78.70	80.56	76.85	81.48	85.19	84.26	81.48
$n = 320$	LDA	75.93	67.59	78.70	78.70	80.56	76.85	81.48	85.19	84.26	81.48
$d = 2887$	ratio	1.010	1.023	1.013	1.019	1.019	1.013	1.010	1.000	1.010	1.000
$k = 4$	Diff	1	1	1	1	1	1	1	0	1	0
DOC3	LS-LDA	95.71	1.000	1.000	97.14	97.14	97.14	97.14	1.000	98.57	92.86
$n = 210$	LDA	95.71	1.000	1.000	97.14	97.14	97.14	97.14	1.000	98.57	92.86
$d = 7455$	ratio	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k = 7$	Diff	0	0	0	0	0	0	0	0	0	0
ORL	LS-LDA	90.00	90.00	92.50	97.50	95.00	94.17	92.50	92.50	96.67	94.17
$n = 400$	LDA	90.00	90.00	92.50	97.50	95.00	94.17	92.50	92.50	96.67	94.17
$d = 10304$	ratio	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k = 40$	Diff	0	0	0	0	0	0	0	0	0	0
AR	LS-LDA	96.50	97.50	95.50	94.50	94.00	91.50	94.00	94.50	93.00	92.50
$n = 650$	LDA	96.50	97.50	95.50	94.50	94.00	91.50	94.00	94.50	93.00	92.50
$d = 8888$	ratio	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k = 50$	Diff	0	0	0	0	0	0	0	0	0	0
PIX	LS-LDA	94.44	95.56	92.22	96.67	100.0	98.89	95.56	95.56	95.56	94.44
$n = 300$	LDA	94.44	95.56	92.22	96.67	100.0	98.89	95.56	95.56	95.56	94.44
$d = 10000$	ratio	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k = 300$	Diff	0	0	0	0	0	0	0	0	0	0
GCM	LS-LDA	76.92	81.54	73.85	69.23	78.46	76.92	84.62	78.46	76.92	86.15
$n = 198$	LDA	76.92	81.54	73.85	69.23	78.46	76.92	84.62	78.46	76.92	86.15
$d = 16063$	ratio	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k = 14$	Diff	0	0	0	0	0	0	0	0	0	0
ALL	LS-LDA	97.56	1.000	96.34	97.56	96.34	98.78	98.78	97.56	98.78	97.56
$n = 248$	LDA	97.56	1.000	96.34	97.56	96.34	98.78	98.78	97.56	98.78	97.56
$d = 12559$	ratio	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k = 6$	Diff	0	0	0	0	0	0	0	0	0	0
ALLAML	LS-LDA	100.0	91.67	95.83	83.33	91.67	95.83	91.67	95.83	95.83	83.33
$n = 72$	LDA	100.0	91.67	95.83	83.33	91.67	95.83	91.67	95.83	95.83	83.33
$d = 4106$	ratio	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k = 4$	Diff	0	0	0	0	0	0	0	0	0	0

7. Conclusion and Discussion

In this paper, we analyze the relationship between multi-class LDA and multivariate linear regression. Specifically, we show that under a mild condition, which has been shown empirically to hold for many high-dimensional and undersampled data, multi-class LDA is equivalent to multivariate linear regression with a specific class indicator matrix. That is, under the given condition, multi-class LDA can be formulated as a least squares problem, which extends previous equivalence result for the binary-class case. Our experimental studies on high-dimensional datasets confirm the presented theoretical analysis.

The presented analysis can be extended in several directions. Regularization is commonly applied to stabilize the sample covariance matrix estimation and improve the classification performance of LDA (Friedman, 1989). Regularization using the L_2-norm penalty can also be applied in linear regression, which is known as ridge regression (Hoerl & Kennard, 1970). Based on the equivalence result established in this paper, we obtain the following formulation for regularized LDA:

$$L_2(W,\gamma) = \frac{1}{2}\sum_{j=1}^{k}\left(\frac{1}{n}\sum_{i=1}^{n}||\tilde{x}_i^T w_j - Y_3(ij)||_2^2 + \gamma||w_j||_2^2\right) \tag{26}$$

where $\gamma > 0$ is the regularization parameter, commonly estimated through cross-validation.

The geometry of the marginal distribution can be exploited through the Laplacian-based regularization (Belkin et al., 2006), which has been applied in regression and SVM. The equivalence relationship presented in this paper results in Laplacian regularized LDA. Furthermore, it naturally leads to semi-supervised dimensionality reduction by combining both label and unlabeled data (Belkin et al., 2006).

Sparsity has recently received much attention for extending Principal Component Analysis (d'Aspremont et al., 2004; Jolliffe & Uddin, 2003; Zou et al., 2006). L_1-norm penalty has been used in regression (Tibshi-

rani, 1996), known as LASSO, and SVM (Zhu et al., 2003) to achieve model sparsity. Sparsity often leads to easy interpretation and good generalization ability of the resulting model. Sparse Fisher Discriminant Analysis has been proposed in (Mika, 2002), for binary-class problems. Based on the equivalence relationship between LDA and regression established in this paper, multi-class sparse LDA can be formulated by minimizing an objective function similar to the one in Eq. (26) by replacing the 2-norm of w_j by the 1-norm as $||w_j||_1$. The optimal w_j can be computed by applying LASSO (Tibshirani, 1996). A solution path can also be obtained through the LARS algorithm (Efron et al., 2004). We plan to study the effectiveness of all these extensions in the future.

Acknowledgments

This research is sponsored by the Center for Evolutionary Functional Genomics of the Biodesign Institute at Arizona State University and by the National Science Foundation Grant IIS-0612069.

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.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, *7*, 2399–2343.

Cristianini, N., Kandola, J., Elisseeff, A., & Shawe-Taylor, J. (2001). On kernel target alignment. *NIPS*.

d'Aspremont, A., Ghaoui, L., Jordan, M., & Lanckriet, G. (2004). A direct formulation for sparse PCA using semidefinite programming. *NIPS*.

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.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics (with discussion)*, *32*, 407–499.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

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

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

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

Guermeur, Y., Lifchitz, A., & Vert, R. (2004). A kernel for protein secondary structure prediction. *Kernel Methods in Computational Biology*, 193–206.

Hastie, T., Tibshirani, R., & Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, *89*, 1255–1270.

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

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

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

Lee, Y., Lin, Y., & Wahba, G. (2004). Multicategory Support Vector Machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, *99*, 67–81.

Mika, S. (2002). *Kernel fisher discriminants*. PhD thesis, University of Technology, Berlin.

Park, C., & Park, H. (2005). A relationship between LDA and the generalized minimum squared error solution. *SIAM Journal on Matrix Analysis and Applications*, *27*, 474–492.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 267–288.

Ye, J., & Xiong, T. (2006). Computational and theoretical analysis of null space and orthogonal linear discriminant analysis. *Journal of Machine Learning Research*, *7*, 1183–1204.

Zhang, P., & Riedel, N. (2005). Discriminant analysis: A unified approach. *ICDM*.

Zhu, J., Rosset, S., Hastie, T., & Tibshirani, R. (2003). 1-norm support vector machines. *NIPS*.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15(2)*, 265–286.

Appendix

A. Proof of Lemma 4.1

Proof. Note that $\tilde{X}\tilde{X}^T = nS_t$. From Eq. (23), we have

$$\begin{aligned} W^T S_b W &= \left(\left(\tilde{X}\tilde{X}^T\right)^+ \tilde{X}\tilde{Y}\right)^T S_b \left(\tilde{X}\tilde{X}^T\right)^+ \tilde{X}\tilde{Y} \\ &= (\tilde{X}\tilde{Y})^T (nS_t)^+ S_b (nS_t)^+ (\tilde{X}\tilde{Y}) \\ &= \frac{1}{n^2}(\tilde{X}\tilde{Y})^T U_1 \Sigma_t^{-1} P \Sigma_b P^T \Sigma_t^{-1} U_1^T (\tilde{X}\tilde{Y}) \\ &= \frac{1}{n^2} F^T \Sigma_b F. \end{aligned}$$

Since $M^+MM^+ = M^+$, for any matrix M (Golub & Van Loan, 1996), and P is orthogonal, i.e., $PP^T = I_t$, we have

$$\begin{aligned} W^T S_t W &= \left(\left(\tilde{X}\tilde{X}^T\right)^+ \tilde{X}\tilde{Y}\right)^T S_t \left(\tilde{X}\tilde{X}^T\right)^+ \tilde{X}\tilde{Y} \\ &= (\tilde{X}\tilde{Y})^T (nS_t)^+ S_t (nS_t)^+ (\tilde{X}\tilde{Y}) \\ &= \frac{1}{n^2}(\tilde{X}\tilde{Y})^T (S_t)^+ (\tilde{X}\tilde{Y}) \\ &= \frac{1}{n^2}(\tilde{X}\tilde{Y})^T U_1 \Sigma_t^{-2} U_1^T (\tilde{X}\tilde{Y}) = \frac{1}{n^2} F^T F. \end{aligned}$$

□

B. Proof of Theorem 4.1

Proof. From Lemma 4.1, we have

$$\begin{aligned} \mathrm{tr}\left((W^T S_b W)(W^T S_t W)^+\right) &= \mathrm{tr}\left((F^T \Sigma_b F)(F^T F)^+\right) \\ &= \mathrm{tr}\left((F^T \Sigma_b F) F^+ (F^T)^+\right) \\ &= \mathrm{tr}\left((FF^+)^T \Sigma_b (FF^+)\right), \end{aligned}$$

where the second equality follows since $(F^TF)^+ = F^+(F^T)^+$ (Golub & Van Loan, 1996). Let

$$F = P_1 \mathrm{diag}\left(\Sigma_1, 0\right) Q_1^T$$

be the SVD of F, where P_1 and Q_1 are orthogonal, $\Sigma_1 \in \mathbb{R}^{r\times r}$ is diagonal and nonsingular, and $r = \mathrm{rank}(F)$. It follows that

$$\begin{aligned} FF^+ &= P_1 \mathrm{diag}\left(\Sigma_1, 0\right) Q_1^T Q_1 \mathrm{diag}\left(\Sigma_1^{-1}, 0\right) P_1^T \\ &= P_1 \mathrm{diag}\left(I_r, 0\right) P_1^T \\ &= P_{1r} P_{1r}^T, \end{aligned}$$

where P_{1r} consists of the first r columns of P_1 and thus has orthonormal columns, i.e., $P_{1r}^T P_{1r} = I_r$. It follows that

$$\mathrm{tr}\left((FF^+)^T \Sigma_b (FF^+)\right) = \mathrm{tr}\left(P_{1r}^T \Sigma_b P_{1r}\right) \le \mathrm{tr}(\Sigma_b),$$

where the first equality follows since $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for any two matrices A and B, and $P_{1r}^T P_{1r} = I_r$, and the inequality follows since P_{1r} has orthonormal columns. This completes the proof of the first part of the theorem.

When $\tilde{Y} = Y_3$, we have $\tilde{X}\tilde{Y} = nH_b$. It follows that

$$\begin{aligned} F &= P^T \Sigma_t^{-1} U_1^T \tilde{X}\tilde{Y} = nP^T \Sigma_t^{-1} U_1^T H_b \\ &= nP^T B = n\hat{\Sigma} Q^T. \end{aligned}$$

where B is defined in Eq. (18) and $\hat{\Sigma} \in \mathbb{R}^{t\times k}$ is defined in Eq. (19). Since $\Sigma_b = \hat{\Sigma}\hat{\Sigma}^T$, we have

$$\begin{aligned} W^T S_b W &= \frac{1}{n^2} F^T \Sigma_b F = Q\hat{\Sigma}^T \Sigma_b \hat{\Sigma} Q^T = Q\Sigma_{bk}^2 Q^T \\ W^T S_t W &= \frac{1}{n^2} F^T F = Q\hat{\Sigma}^T \hat{\Sigma} Q^T = Q\Sigma_{bk} Q^T \end{aligned}$$

where Σ_{bk} consists of the first k rows and the first k columns of Σ_b. It follows that

$$\mathrm{tr}\left((W^T S_b W)(W^T S_t W)^+\right) = \mathrm{tr}\left(\Sigma_{bk}\right) = \mathrm{tr}\left(\Sigma_b\right),$$

where the last equality follows since only the first q diagonal entries of Σ_b are nonzero. □

C. Proof of Theorem 5.1

Proof. Let matrix $H \in \mathbb{R}^{d\times d}$ be defined as follows:

$$H = U \begin{pmatrix} \Sigma_t^{-1} P & 0 \\ 0 & I_{d-t} \end{pmatrix}, \tag{27}$$

where U and Σ_t are defined in Eq. (16), and P is defined in Eq. (19). It follows from Eqs. (16)–(21)

$$\begin{aligned} H^T S_b H &= \mathrm{diag}\left(\Sigma_b, 0\right), \\ H^T S_t H &= \mathrm{diag}\left(I_t, 0\right). \end{aligned} \tag{28}$$

Since $S_w = S_t - S_b$, we have

$$H^T S_w H = \mathrm{diag}\left(\Sigma_w, 0\right), \tag{29}$$

for some diagonal matrix $\Sigma_w = I_t - \Sigma_b$. From Eqs. (21), (22), (28) and (29), we have

$$\begin{aligned} H^T S_b H &= \mathrm{diag}(\alpha_1^2, \cdots, \alpha_t^2, 0 \cdots, 0) \\ H^T S_w H &= \mathrm{diag}(\beta_1^2, \cdots, \beta_t^2, 0 \cdots, 0), \end{aligned}$$

where $\alpha_1^2 \ge \cdots \ge \alpha_q^2 > 0 = \alpha_{q+1}^2 = \cdots = \alpha_t^2$, and $\alpha_i^2 + \beta_i^2 = 1$, for all i. Since $\mathrm{rank}(S_b) + \mathrm{rank}(S_w) - \mathrm{rank}(S_t) = 0$, we have

$$t = \mathrm{rank}(H^T S_t H) = \mathrm{rank}(H^T S_b H) + \mathrm{rank}(H^T S_w H).$$

Since $\alpha_i^2 + \beta_i^2 = 1$, at least one of α_i and β_i is nonzero. Thus the following inequality always holds: $\mathrm{rank}(H^T S_b H) + \mathrm{rank}(H^T S_w H) \ge t$. The equality holds only when either α_i or β_i is zero, for all i. That is, $\alpha_i \beta_i = 0$, for all i. Hence, $\alpha_1^2 = \cdots = \alpha_q^2 = 1$, that is, Σ_{bq}, which consists of the first q rows and the first q columns of Σ_b, equals to I_q. □