Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

On the Equivalence of Linear Discriminant Analysis and Least Squares

Kibok Lee[1,2] and **Junmo Kim**[1]
[1]Department of Electrical Engineering, KAIST, Daejeon, Korea
[2]Samsung Electronics DMC R&D Center, Suwon, Korea
kibok90@gmail.com, junmo.kim@kaist.ac.kr

Abstract

Linear discriminant analysis (LDA) is a popular dimensionality reduction and classification method that simultaneously maximizes between-class scatter and minimizes within-class scatter. In this paper, we verify the equivalence of LDA and least squares (LS) with a set of dependent variable matrices. The equivalence is in the sense that the LDA solution matrix and the LS solution matrix have the same range. The resulting LS provides an intuitive interpretation in which its solution performs data clustering according to class labels. Further, the fact that LDA and LS have the same range allows us to design a two-stage algorithm that computes the LDA solution given by generalized eigenvalue decomposition (GEVD), much faster than computing the original GEVD. Experimental results demonstrate the equivalence of the LDA solution and the proposed LS solution.

Introduction

In classification, dimensionality reduction has been an important problem in many fields dealing with high dimensional data. The main objective of dimensionality reduction is to discard redundant and noisy features while preserving discriminative information so that the curse of dimensionality can be overcome. Linear discriminant analysis (LDA) is a popular dimensionality reduction and classification method that simultaneously maximizes between-class scatter and minimizes within-class scatter (Bishop 2006) (Fukunaga 1990), thereby keeping discriminative information while reducing indiscriminative information. The original formulation of LDA, known as Fisher's linear discriminant, deals with binary classification. This binary-class LDA is equivalent to a least squares (LS) problem with a particular dependent variable matrix (Bishop 2006). Many real-world applications involve multiclass problems, and LDA can be generalized to a multiclass problem. With the generalized Fisher criterion, LDA finds a subspace that has less than or equal to $c-1$ dimensions, where c is the number of classes (Bishop 2006) (Fukunaga 1990). It has been proposed that the range of the LDA solution matrix for the multiclass cases can be obtained from that of the LS solution matrix with a certain dependent variable matrix[1] (Hastie, Tibshirani, and Buja 1994).

Recently, there has been progress in generalizing the equivalence of LS and LDA to multiclass cases without regularization and multiclass cases with regularization. Ye (2007) extended the equivalence to multiclass cases without regularization by considering LS with one specific choice of dependent variable matrix. Sun (2010) and Zhang (2010) dealt with multiclass cases with regularization, which are of more practical importance. They proposed a two-stage algorithm that consists of an LS stage with a dependent variable matrix and a generalized eigenvalue decomposition (GEVD) stage, where the resulting transformation matrix consists of the eigenvectors of the original LDA problem. Cai et al. (2008) also considered multiclass cases with regularization and proposed an algorithm to obtain the LDA solution subspace by solving LS. Their algorithm suggested a way to find the dependent variable matrix. However, the theory for the equivalence of LDA and LS has remained incomplete in the sense that the necessary and sufficient condition for the equivalence is unknown. Consequently, there has been no simple way to check the validity of a potential candidate dependent variable matrix with some desirable properties; the equivalence needs to be proved once again for the candidate dependent variable matrix. Similarly, a two-stage algorithm was proposed and verified only for a specific choice of dependent variable matrix (Sun, Ceran, and Ye 2010) (Zhang et al. 2010).

In this paper, we complete the theory for the equivalence of LDA and LS by establishing the *necessary and sufficient conditions* for the dependent variable matrices, i.e., we identify the *exact* set of dependent variable matrices such that the corresponding LS is equivalent to the LDA for any data so that one can apply LS to obtain the LDA solution. The equivalence is in the sense that they have the same solution subspace (i.e., the same range). The resulting LS with the set of dependent variable matrices provides an interesting and appealing intuitive interpretation in that the mapping for dimensionality reduction given by the LS solution, or equivalently the LDA solution, performs a sort of data clus-

[1]For example, in a typical LS problem $\min_{\mathbf{W}}\|\mathbf{AW}-\mathbf{B}\|_F$, $\mathbf{A}$ and $\mathbf{B}$ are the independent and dependent variable matrices for the LS, respectively. In this paper, we use $\mathbf{Y}$ for the dependent variable matrix.

tering according to class labels. Furthermore, the resulting necessary and sufficient condition is in a simple form that allows us to easily check the validity of a potential candidate dependent variable matrix. For example, we designed a new dependent variable matrix $\mathbf{Y}_B$ that has several desirable properties, including a low construction cost and almost uncorrelated data vectors after dimensionality reduction; the validity of $\mathbf{Y}_B$ is checked according to our theoretical results. We also generalize the two-stage algorithm proposed by Sun (2010) and Zhang (2010) by enlarging the set of dependent variable matrices for the first stage, and we show that one can use this two-stage algorithm to implement a fast dimensionality reduction by LDA to an arbitrary target dimension.

Background

Notation

Let $\mathbf{x} \in \mathbb{R}^d$ be a data point, and let $\mathbf{l} = \mathbf{e}_k \triangleq [\underbrace{0, \cdots, 0}_{k-1 \text{ zeros}}, 1, \underbrace{0, \cdots, 0}_{c-k \text{ zeros}}]^T \in \mathbb{R}^c$ be its corresponding label vector, where k is the class of the data. Also, let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ be the collection of training data (data matrix), and let $\mathbf{L} = [\mathbf{l}_1, \cdots, \mathbf{l}_n] \in \mathbb{R}^{c \times n}$ be the collection of the corresponding label vectors (label matrix), so that if $\mathbf{x}_i$ is in the k-th class, then $\mathbf{l}_i = \mathbf{e}_k$. We refer to a solution matrix $\mathbf{W}^{(p)}$ or $\mathbf{W} \in \mathbb{R}^{d \times p}$ as a p-dimensional solution, and its subspace $\mathrm{ran}\,(\mathbf{W})$ as the solution subspace. The notations are presented in Table 1.

Table 1: Notations.

Symbol	Description
$\dim(\cdot)$	Dimension of the given subspace
$\mathrm{null}\,(\cdot)$	Nullspace of the given matrix
$\mathrm{ran}\,(\cdot)$	Range space of the given matrix
$\mathrm{rank}\,(\cdot)$	Rank of the given matrix, $\mathrm{rank}\,(\cdot) = \dim(\mathrm{ran}\,(\cdot))$
$\mathrm{tr}\,(\cdot)$	Trace; summation of the diagonals of the given matrix
$\mathbf{1}, \mathbf{0}$	One/zero vector; subscript indicates its dimension
$\mathbf{I}, \mathbf{O}$	Identity/zero matrix; subscript indicates its dimension
n, n_k	The number of all data/data in the k-th class
d, p	Original/projected data dimension
c	The number of classes
γ	Regularization parameter
$\mathbf{W}_{(\cdot)}$	Transformation matrix; subscript indicates the method
$\mathbf{X}$	Data matrix
$\mathbf{L}_{(\cdot)}, \mathbf{Y}_{(\cdot)}$	Dependent variable matrix
$\hat{\mathbf{Y}}$	Transformed data matrix
$\mathbf{m}$	Total mean
$\mathbf{m}_k, \mathbf{M}$	k-th class mean / class mean matrix
$\mathbf{S}_{(\cdot)}$	Scatter matrix
$\mathbf{C}_{(\cdot)}$	Centering matrix; subscript indicates its dimension

For some $\mathbf{1}$, $\mathbf{0}$, $\mathbf{I}$, $\mathbf{O}$, and $\mathbf{C}$, we added a subscript to clarify their dimension. In a GEVD $\mathbf{AW} = \mathbf{BW\Lambda}$, we only consider the positive eigenvalues, and the diagonals of the eigenvalue matrix are sorted in descending order.

Linear Discriminant Analysis

LDA is a popular dimensionality reduction and classification method that simultaneously maximizes between-class scatter and minimizes within-class scatter (Bishop 2006) (Fukunaga 1990). The LDA solution $\mathbf{W} \in \mathbb{R}^{d \times p}$ maps the data matrix $\mathbf{X}$ onto the transformed data matrix $\hat{\mathbf{Y}} = \mathbf{W}^T\mathbf{X}$.

Let $\mathbf{m} \in \mathbb{R}^d$ be the mean vector, and let $\mathbf{m}_k \in \mathbb{R}^d$ be the k-th class mean vector. With $\mathbf{L}_B \triangleq (\mathbf{L}\mathbf{L}^T)^{-\frac{1}{2}}\mathbf{L}$, the scatter matrices, which are in $\mathbb{R}^{d \times d}$, can be defined as follows: $\mathbf{S}_T \triangleq \sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T = \mathbf{X}_C\mathbf{X}_C^T = \mathbf{X}\mathbf{C}\mathbf{X}^T$ and $\mathbf{S}_B \triangleq \sum_{k=1}^{c} n_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T = \mathbf{X}\mathbf{C}_B\mathbf{X}^T$ where $\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, $\mathbf{X}_C = \mathbf{X}\mathbf{C}$, and $\mathbf{C}_B = \mathbf{C}\mathbf{L}_B^T\mathbf{L}_B\mathbf{C}$. The scatter matrices $\mathbf{S}_T$ and $\mathbf{S}_B$ are referred to as the total scatter matrix and between-class scatter matrix, respectively. Note that $\mathbf{C}$ and $\mathbf{C}_B$ are symmetric, idempotent (the square of the matrix is the matrix itself), and centered (the mean of column vectors is the zero vector). The matrix $\mathbf{C}$ is often referred to as the centering matrix since multiplying this matrix has the same effect as subtracting the mean vector from every column of the target matrix.

The Fisher criterion can be defined in several ways, and the following is a well-known form (Fukunaga 1990):

$$J_{LDA}(\mathbf{W}) \triangleq \mathrm{tr}\left((\mathbf{W}^T\mathbf{S}_T\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_B\mathbf{W})\right) \quad (1)$$

with an assumption that $\mathrm{rank}\,(\mathbf{S}_T) = d$. Note that total scatter is used instead of within-class scatter, since it generates the same solution, and it is easier to deal with (Fukunaga 1990). One way to obtain an LDA solution is to solve the generalized eigenvalue problem $\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_T\mathbf{w}$. In that case, the LDA solution consists of the p eigenvectors corresponding to the p largest eigenvalues and the value of J_{LDA} is the sum of the eigenvalues. Since it has at most $\mathrm{rank}\,(\mathbf{S}_B)$-positive eigenvalues, $p = \mathrm{rank}\,(\mathbf{S}_B)$ is enough to maximize J_{LDA}. Note that $\mathrm{rank}\,(\mathbf{S}_B) \leq c - 1$ where the equality holds in most cases; for instance, it holds when class mean vectors $\{\mathbf{m}_k\}_{k=1}^c$ are linearly independent.

On the other hand, if $\mathrm{rank}\,(\mathbf{S}_T) < d$, (1) is ill-posed. In that case, there exist vectors in $\mathrm{null}\,(\mathbf{S}_T)$ such that they are geometrically the worst solution: if one takes these vectors as projectors, all data are projected onto zero. In terms of the Fisher criterion, theses vectors make the Fisher criterion zero divided by zero since $\mathbf{w}^T\mathbf{S}_B\mathbf{w} = 0$ and $\mathbf{w}^T\mathbf{S}_T\mathbf{w} = 0$. To deal with the problem, one can add a regularization matrix into the total scatter matrix so that the problem becomes well-posed, which is referred to as regularized LDA (RLDA) (Friedman 1989). To represent the original LDA simultaneously, we define the regularized total scatter matrix as

$$\mathbf{S}_{T,\gamma} \triangleq \mathbf{S}_T + \gamma\mathbf{I},$$

where $\gamma \geq 0$ and $\mathrm{rank}\,(\mathbf{S}_{T,\gamma}) = d$ so that $\mathbf{S}_{T,\gamma}$ is always invertible. Note that it can be separated in two cases: either a regularized case, $\gamma > 0$, or the original LDA case, $\gamma = 0$ and $\mathrm{rank}\,(\mathbf{S}_T) = d$.

$$J_{RLDA}(\mathbf{W}) \triangleq \mathrm{tr}\left((\mathbf{W}^T\mathbf{S}_{T,\gamma}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_B\mathbf{W})\right).$$

Least Squares

LS finds a transformation from an independent variable matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ to a dependent variable matrix $\mathbf{B} \in \mathbb{R}^{n \times p}$ by minimizing the sum of squared error. Adding a positive regularization term makes the solution unique: An regularized LS (RLS) problem $\min_{\mathbf{W}} \|\mathbf{AW} - \mathbf{B}\|_F^2 + \gamma\|\mathbf{W}\|_F^2$ has a unique solution $\mathbf{W} = (\mathbf{A}^T\mathbf{A} + \gamma\mathbf{I})^{-1}\mathbf{A}^T\mathbf{B}$.

Previous studies showed a relation between LDA and LS with $\mathbf{A} = \mathbf{X}_C^T$ and $\mathbf{B} = \mathbf{Y}^T$, where $\mathbf{Y}$ is related to the label matrix $\mathbf{L}$, e.g., $\mathbf{Y} = \mathbf{L}_B$ (Sun, Ceran, and Ye 2010)

or $\mathbf{Y} = \mathbf{L}_B\mathbf{C}$ (Ye 2007) (Zhang et al. 2010). With these choices, the RLS solution $\mathbf{W}_{RLS}$ can be represented as

$$\mathbf{W}_{RLS} = \mathbf{S}_{T,\gamma}^{-1}\mathbf{X}_C\mathbf{Y}^T \tag{2}$$

and the equivalence of RLDA and RLS with these particular choices of $\mathbf{Y}$ has been found in (Ye 2007), (Sun, Ceran, and Ye 2010), and (Zhang et al. 2010) in the sense that the RLDA solution obtained by the generalized eigenvalue problem and this RLS solution have the same range.

Theoretical Analysis

In this section, we show the equivalence of LDA and LS. Using this relationship, the LDA solution subspace can be obtained by solving the LS, which is more computationally efficient than GEVD. For convenience, let $r_B = \operatorname{rank}(\mathbf{S}_B)$.

Range Uniqueness of RLDA Solutions

As is well known, an LDA solution can be obtained by GEVD. However, the solution is not a unique solution but a particular solution. One can verify the fact from the following proposition:

Proposition 1. *For any nonsingular* $\mathbf{\Xi} \in \mathbb{R}^{p\times p}$, $J_{LDA}(\mathbf{W}) = J_{LDA}(\mathbf{W}\mathbf{\Xi})$.

Proof. The proof follows from the properties $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ and $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$. □

Proposition 1 implies that the LDA solution is not unique and that the Fisher criterion value does not depend on the bases but rather on the span of the bases: the Fisher criterion is affine invariant. Consequently, we focus on the subspace in this paper, since the solution subspace $\operatorname{ran}(\mathbf{W})$ matters[2] instead of the solution $\mathbf{W}$.

Let an eigenvector matrix $\mathbf{W} \in \mathbb{R}^{d\times p}$ satisfying $\mathbf{S}_B\mathbf{W} = \mathbf{S}_{T,\gamma}\mathbf{W}\mathbf{\Lambda}$ be the p-dimensional GEVD solution $\mathbf{W}_{GEVD}^{(p)}$. For $p \le r_B$, any p-dimensional RLDA solution can be represented as a linear combination of the first p dominant eigenvectors.

Lemma 1. *For $p \le r_B$, there exist $\mathbf{\Xi}_1 \in \mathbb{R}^{p_1\times p}$ and $\mathbf{\Xi}_2 \in \mathbb{R}^{p_2\times p}$ such that*

$$\mathbf{W}_{RLDA}^{(p)} = \mathbf{W}_{GEVD}^{(p_1+p_2)}\begin{bmatrix}\mathbf{\Xi}_1\\ \mathbf{\Xi}_2\end{bmatrix},$$

where p_1 is the number of eigenvalues larger than the p-th eigenvalue, p_2 is the number of eigenvalues equal to the p-th eigenvalue, $\operatorname{rank}(\mathbf{\Xi}_1) = p_1$, *and* $\operatorname{rank}(\mathbf{\Xi}_2) = p - p_1$.

Since the generalized eigenvalue problem has at most r_B positive eigenvalues, $p \le p_1 + p_2 \le r_B$. In particular, when $p = r_B$, we have an equivalence $\operatorname{ran}\left(\mathbf{W}_{RLDA}^{(r_B)}\right) = \operatorname{ran}\left(\mathbf{W}_{GEVD}^{(r_B)}\right)$, which means that the r_B-dimensional range is uniquely determined by the eigenspace. Further, Lemma 2 shows that the r_B-dimensional RLDA solution subspace is uniquely determined by the data matrix and the label matrix.

Lemma 2. $\operatorname{ran}\left(\mathbf{W}_{RLDA}^{(r_B)}\right) = \operatorname{ran}\left(\mathbf{S}_{T,\gamma}^{-1}\mathbf{X}_C\mathbf{L}^T\right)$.

The equivalence of RLDA and RLS for the case $p = r_B$ is considered in the next section, and then the analysis for the case $p < r_B$ follows in another section.

Solution Subspace Equivalence of RLDA and RLS

In this section, we compare the r_B-dimensional RLDA solution subspace and the RLS solution subspace. From (2) and Lemma 2, one can expect that the range equivalence might hold if $\mathbf{Y}$ has a relation with $\mathbf{L}$. We introduce a lemma about a set of $\mathbf{Y} \in \mathbb{R}^{p_{LS}\times n}$, which turns out to be the exact condition of the equivalence:

Lemma 3.

$$\mathbf{Y} = \mathbf{Z}\mathbf{L} \text{ for some } \mathbf{Z}, \text{ where } \operatorname{rank}(\mathbf{Z}\mathbf{C}_c) = c - 1, \tag{3}$$

or, equivalently,

$$\mathbf{Y} = \mathbf{Z}\mathbf{L} \text{ for some } \mathbf{Z}, \text{ where } \operatorname{rank}\left(\begin{bmatrix}\mathbf{Z}^T & \mathbf{1}\end{bmatrix}\right) = c, \tag{4}$$

if and only if

$$\operatorname{ran}\left(\mathbf{C}_n\mathbf{L}^T\right) = \operatorname{ran}\left(\mathbf{C}_n\mathbf{Y}^T\right).$$

We refer to the condition $\operatorname{rank}(\mathbf{ZC}) = c - 1$ as being "centered rank" $c - 1$, e.g., $\mathbf{Z}$ is of centered rank $c - 1$.

Using a concept of Moore-Penrose pseudoinverse (Penrose 1955), we present a general theorem for an arbitrary data matrix $\mathbf{X}$ with a mild condition that its class mean matrix $\mathbf{M} \triangleq [\mathbf{m}_1, \cdots, \mathbf{m}_c]$ is of centered rank $c - 1$, or equivalently, $r_B = c - 1$, which is true in most cases.

Theorem 1. *Suppose that an arbitrary data matrix $\mathbf{X}$ and label matrix $\mathbf{L}$ has a class mean matrix $\mathbf{M}$ of centered rank $c - 1$, or equivalently,* $\operatorname{rank}(\mathbf{S}_B) = c - 1$. *Then, $\mathbf{Y} = \mathbf{ZL} + \mathbf{\Xi}(\mathbf{I} - \mathbf{X}_C^+\mathbf{X}_C)$ for some $\mathbf{Z}$ of centered rank $c-1$ and some $\mathbf{\Xi} \in \mathbb{R}^{p_{LS}\times n}$ if and only if*

$$\operatorname{ran}\left(\mathbf{W}_{RLDA}^{(r_B)}\right) = \operatorname{ran}(\mathbf{W}_{RLS}).$$

Theorem 1 identifies the exact set of $\mathbf{Y}$ so that the RLDA and RLS are equivalent for a specific $\mathbf{X}$ and $\mathbf{L}$. It makes sense that this set of $\mathbf{Y}$ depends on $\mathbf{X}$. This theorem is of theoretical importance as it delineates the exact boundary for the set of valid $\mathbf{Y}$ for the equivalence. In practice, it is more desirable to have a universal algorithm that works for any data matrix $\mathbf{X}$ rather than a specific $\mathbf{X}$. The following theorem identifies the exact set of $\mathbf{Y}$ so that the equivalence holds for any $\mathbf{X}$.

Theorem 2. *Suppose that $d \ge c - 1$ and $d \ge 2$. For any data matrix $\mathbf{X}$ and label matrix $\mathbf{L}$, if $\mathbf{Y}$ is in the form of $\mathbf{ZL}$ where $\mathbf{Z}$ is of centered rank $c - 1$, then* $\operatorname{ran}\left(\mathbf{W}_{RLDA}^{(r_B)}\right) = \operatorname{ran}(\mathbf{W}_{RLS})$. *Conversely, if* $\operatorname{ran}\left(\mathbf{W}_{RLDA}^{(r_B)}\right) = \operatorname{ran}(\mathbf{W}_{RLS})$ *for any $\mathbf{X}$ and $\mathbf{L}$, then $\mathbf{Y}$ should be in the form of $\mathbf{ZL}$ where $\mathbf{Z}$ is of centered rank $c - 1$.*

[2] $J_{LDA}(\mathbf{W}) = J_{LDA}(\mathbf{W}\mathbf{\Xi})$ alone does not guarantee that the classification performance is the same for arbitrary classifiers unless the classifiers are affine invariant; however, since $\mathbf{W}^T\mathbf{x}$ and $\mathbf{\Xi}^T\mathbf{W}^T\mathbf{x}$ are reproducible from each other, they have the same amount of discriminative information.

Note that Proposition 1 also holds for RLDA. The range equivalence $\mathrm{ran}\left(\mathbf{W}_{RLDA}^{(r_B)}\right) = \mathrm{ran}\left(\mathbf{W}_{RLS}\right)$ in Theorem 1 and 2 implies that an RLS solution maximizes J_{RLDA}, which means that the RLS solution is an RLDA solution.

Now, we focus on the dependent variable matrix $\mathbf{Y}$. For checking the validity of $\mathbf{Y}$, (4) is useful: an $\mathbf{Y}$ in the form of $\mathbf{ZL}$ is valid if $\mathrm{rank}\left(\begin{bmatrix}\mathbf{Z}^T & \mathbf{1}\end{bmatrix}\right) = c$ holds. Let us consider this condition in more details. This condition can be separated in two cases: $\mathrm{rank}\left(\mathbf{Z}\right) = c$ or $\mathrm{rank}\left(\mathbf{Z}\right) = c-1$. Hereafter, we refer to such $\mathbf{Z}$s as type 1 and type 2 matrix, and $\mathbf{W}$s as type 1 and type 2 solution, respectively. First, consider the type 1 case. As illustrated in Figure 1, for a matrix $\mathbf{Z}$ whose columns are linearly independent, the RLS clusters k-th class data around $\mathbf{z}_k = \mathbf{Ze}_k$, the k-th column vector of $\mathbf{Z}$. That is, designing class vectors to be linearly independent is sufficient to make the RLS solution equivalent to the RLDA solution. Interestingly, it can be shown that any type

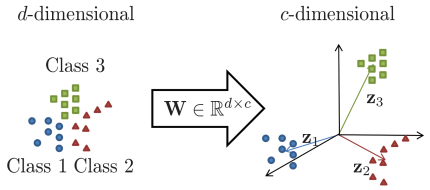

Figure 1: An example of transformation by $\mathbf{W} \in \mathbb{R}^{d \times p_{LS}}$ when $p_{LS} = c = 3$ and $\mathbf{Y} = \mathbf{ZL} \in \mathbb{R}^{c \times n}$ where $\mathbf{Z} = \begin{bmatrix}\mathbf{z}_1 & \mathbf{z}_2 & \mathbf{z}_3\end{bmatrix}$ has linearly independent columns.

1 solution is also a type 2 solution: Let $\mathbf{W}_{t1}$ be a type 1 solution with $\mathbf{Y}_{t1} = \mathbf{Z}_{t1}\mathbf{L}$ where $\mathbf{Z}_{t1} = [\mathbf{z}_1, \ldots, \mathbf{z}_c] \in \mathbb{R}^{p_{t1} \times c}$. For any vector $\mathbf{a} \in \mathbb{R}^{p_{t1}}$, with $\mathbf{Z} = \mathbf{Z}_{t1} - \mathbf{a}\mathbf{1}^T$, we can get the identical type 1 solution: using (2),

$$\begin{aligned}
&\mathbf{S}_{T,\gamma}^{-1}\mathbf{X}_C\mathbf{L}^T(\mathbf{Z}_{t1} - \mathbf{a}\mathbf{1}^T)^T \\
&= \mathbf{S}_{T,\gamma}^{-1}\mathbf{X}(\mathbf{C}\mathbf{L}^T\mathbf{Z}_{t1}^T - \mathbf{C}\mathbf{L}^T\mathbf{1}\mathbf{a}^T) \\
&= \mathbf{S}_{T,\gamma}^{-1}\mathbf{X}\mathbf{C}\mathbf{L}^T\mathbf{Z}_{t1}^T = \mathbf{W}_{t1},
\end{aligned}$$

where $\mathbf{C}_n\mathbf{L}^T\mathbf{1}_c = \mathbf{C}_n\mathbf{1}_n = \mathbf{0}_n$ is applied. However, choosing an appropriate $\mathbf{a}$, e.g., $\mathbf{a} = \mathbf{z}_1$, we get $\mathrm{rank}\left(\mathbf{Z}_{t1} - \mathbf{a}\mathbf{1}^T\right) = c-1$, hence $\mathbf{Z}_{t1} - \mathbf{a}\mathbf{1}^T$ is type 2 and the corresponding solution is a type 2 solution. Conversely, a type 2 solution with an augmented zero vector turns out to be a type 1 solution: Let $\mathbf{W}_{t2}$ be a type 2 solution with $\mathbf{Y}_{t2} = \mathbf{Z}_{t2}\mathbf{L}$ where $\mathbf{Z}_{t2} \in \mathbb{R}^{p_{t2} \times c}$. Using a type 1 matrix $\begin{bmatrix}\mathbf{Z}_{t2}^T & \mathbf{1}\end{bmatrix}^T$, one can find the relationship:

$$\begin{aligned}
&\mathbf{S}_{T,\gamma}^{-1}\mathbf{X}_C\mathbf{L}^T\begin{bmatrix}\mathbf{Z}_{t2}^T & \mathbf{1}\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{S}_{T,\gamma}^{-1}\mathbf{X}_C\mathbf{L}^T\mathbf{Z}_{t2}^T & \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{W}_{t2} & \mathbf{0}\end{bmatrix}.
\end{aligned}$$

Therefore, type 2 LS generates essentially the same solution as type 1.

Now, we introduce two important and novel type 2 dependent variable matrices, $\mathbf{L}_-$ and $\mathbf{Y}_B$, which can be computed efficiently. With $\mathbf{Z}_- \triangleq \begin{bmatrix}\mathbf{I}_{c-1} & \mathbf{0}\end{bmatrix} \in \mathbb{R}^{(c-1)\times c}$, $\mathbf{L}_- \triangleq \mathbf{Z}_-\mathbf{L}$ is a valid dependent variable matrix. With this choice $\mathbf{Y} = \mathbf{L}_- \in \mathbb{R}^{(c-1)\times n}$, $\mathbf{Y}$ is the sparsest matrix among the cases. Also, we introduce a full-rank factorization of $\mathbf{C}_B = \mathbf{C}\mathbf{L}_B^T\mathbf{L}_B\mathbf{C}$ where $\mathbf{L}_B = (\mathbf{L}\mathbf{L}^T)^{-\frac{1}{2}}\mathbf{L}$ is a scaled label matrix, referred to as $\mathbf{Y}_B$, which satisfies $\mathbf{Y}_B^T\mathbf{Y}_B = \mathbf{C}_B$ and $\mathbf{Y}_B\mathbf{Y}_B^T = \mathbf{I}$. The matrix can be represented as $\mathbf{Y}_B = \mathbf{Z}_B\mathbf{L}$, where $\mathbf{Z}_B = \{z_{Bij}\}$ satisfies

$$z_{Bij} = \begin{cases} \sqrt{\frac{1}{n_i} - \frac{1}{\sum_{h=i}^{c} n_h}} & \text{if } i = j, \\ -\sqrt{\frac{1}{\sum_{j=i+1}^{c} n_j} - \frac{1}{\sum_{j=i}^{c} n_j}} & \text{if } i < j, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where n_k is the number of data in the k-th class. Since $\mathbf{Z}_B$ is nearly triangular, one can easily verify that $\mathrm{rank}\left(\begin{bmatrix}\mathbf{Z}_B^T & \mathbf{1}\end{bmatrix}\right) = c$. Also, $\mathbf{Y}_B^T\mathbf{Y}_B = \mathbf{C}_B$ and $\mathbf{Y}_B\mathbf{Y}_B^T = \mathbf{I}$ are directly derived from (5).

It is often desirable that the transformed data are uncorrelated or that the transformation matrix has orthogonal columns. The dependent variable matrices can be chosen depending on the desired property. Suppose that we want the transformed data to be nearly statistically uncorrelated without additional processing after RLS. With a mild assumption $r_B = c - 1$, which is true in most cases, $\mathbf{Y}_B$ can be the choice: Consider the uncorrelatedness constraint, $\mathbf{W}^T\mathbf{S}_T\mathbf{W} = \mathbf{I}$. Note that LDA with this constraint is referred to as uncorrelated LDA (ULDA) (Ye 2006) (Jin et al. 2001). Choosing $\mathbf{Y} = \mathbf{Y}_B$, we can see that

$$\mathbf{W}^T\mathbf{S}_T\mathbf{W} = (\mathbf{X}_C^T\mathbf{W})^T(\mathbf{X}_C^T\mathbf{W}) \simeq \mathbf{Y}_B\mathbf{Y}_B^T = \mathbf{I}_{c-1}.$$

Therefore, Algorithm 1 returns a transformation that produces nearly statistically uncorrelated transformed data and has the same range as the RLDA solution.

Algorithm 1 Fast approximation of regularized uncorrelated LDA (RULDA) ($p = c - 1$)

1. Compute $\mathbf{Y}_B = \mathbf{Z}_B\mathbf{L}$, where $\mathbf{Z}_B = \{z_{Bij}\}$ in (5).
2. Solve $\mathbf{W} = \mathrm{argmin}_{\mathbf{W}}\|\mathbf{X}_C^T\mathbf{W} - \mathbf{Y}_B^T\|_F^2 + \gamma\|\mathbf{W}\|_F^2$.
3. **return** $\mathbf{W}$

Next, suppose we want to find an orthonormal basis of the solution subspace $\mathrm{ran}\left(\mathbf{W}\right)$, i.e., $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, which is referred to as orthogonal LDA (OLDA) (Ye 2006). In this case, an additional thin QR factorization (Golub and Van Loan 1996) is needed after RLS. Any $\mathbf{Y}$ that satisfies the condition in (3) can be a candidate, but $\mathbf{L}_-$ is preferable due to its sparseness. Note that a thin QR factorization step often improves classification performance of Euclidean distance classifier since the RLS solutions have skewed bases. The fast regularized orthogonal LDA (ROLDA) algorithm is presented in Algorithm 2.

Algorithm 2 Fast ROLDA ($p = r_B$)

1. Solve $\mathbf{W} = \mathrm{argmin}_{\mathbf{W}}\|\mathbf{X}_C^T\mathbf{W} - \mathbf{L}_-^T\|_F^2 + \gamma\|\mathbf{W}\|_F^2$.
2. Find the thin QR factorization $\mathbf{W} = \mathbf{Q}\begin{bmatrix}\mathbf{R}_1 & \mathbf{R}_2\end{bmatrix}$, where $\mathbf{R}_1$ is upper triangular.
3. **return** $\mathbf{Q}$

Two-stage RLDA Method Based on RLS and EVD

Suppose we desire to reduce the dimension to an arbitrary p, where $p \leq r_B$. Although RLS can find $p_{LS} \geq c-1$ vectors such that the solution subspace is optimal, we cannot obtain the optimal p-dimensional solution simply by choosing the first p columns of the solution, unlike the GEVD solution. In this case, we can find the solution in two stages: we first obtain the RLS solution and then perform GEVD with the small scatter matrices transformed by the basis of the RLS solution subspace to obtain the p-dimensional optimal solution instead of GEVD of the large scatter matrices.

Theorem 3. *For $\mathbf{Y}$ in the form of $\mathbf{ZL}$ where $\mathbf{Z}$ is of centered rank $c-1$, let $\mathbf{W}_1 = \mathbf{Q}_1\mathbf{R}_1$ where $\mathbf{Q}_1 \in \mathbb{R}^{d\times r_B}$ be the thin QR factorization of the LS solution*

$$\mathbf{W}_1 = \underset{\mathbf{W}}{\operatorname{argmin}} \|\mathbf{X}_C^T\mathbf{W} - \mathbf{Y}^T\|_F^2 + \gamma\|\mathbf{W}\|_F^2 \in \mathbb{R}^{d\times p_{LS}},$$

and let $\mathbf{W}_2 \in \mathbb{R}^{r_B\times r_B}$ be the eigenvector matrix satisfying

$$(\mathbf{Q}_1^T\mathbf{S}_B\mathbf{Q}_1)\mathbf{W}_2 = (\mathbf{Q}_1^T\mathbf{S}_{T,\gamma}\mathbf{Q}_1)\mathbf{W}_2\mathbf{\Lambda}_2.$$

Then, the two-stage RLDA solution $\mathbf{Q}_1\mathbf{W}_2\begin{bmatrix}\mathbf{I}_p\\ \mathbf{O}\end{bmatrix}$ is the p-dimensional GEVD solution.

To make the transformed total scatter matrix used in the second stage always nonsingular, an additional QR factorization should be conducted before the second stage in Theorem 3. The first stage finds the transformation that reduces the dimension from d to r_B, and the second stage finds the transformation that further reduces the dimension from r_B to p. However, replacing GEVD with EVD, which needs only one transformed scatter matrix, the more efficient algorithm can be conducted, as in Algorithm 3. In the following theorem, an additional constraint $\mathbf{C}\mathbf{Y}^T\mathbf{Y}\mathbf{C} = \mathbf{C}_B$ is sufficient for the application of Algorithm 3. For example, $\mathbf{L}_B$ (Sun, Ceran, and Ye 2010), $\mathbf{L}_B\mathbf{C}$ (Ye 2007) (Zhang et al. 2010), and $\mathbf{Y}_B$, which is proposed in this paper, satisfy this condition.

Theorem 4. *For $\mathbf{Y}$ satisfying $\mathbf{C}\mathbf{Y}^T\mathbf{Y}\mathbf{C} = \mathbf{C}_B$, let $\mathbf{W}_1$ be the LS solution*

$$\mathbf{W}_1 = \underset{\mathbf{W}}{\operatorname{argmin}} \|\mathbf{X}_C^T\mathbf{W} - \mathbf{Y}^T\|_F^2 + \gamma\|\mathbf{W}\|_F^2 \in \mathbb{R}^{d\times p_{LS}},$$

and let $\mathbf{W}_2 \in \mathbb{R}^{p_{LS}\times r_B}$ be the eigenvector matrix of the EVD of $\mathbf{W}_1^T\mathbf{X}_C\mathbf{Y}^T$, which satisfies

$$\left(\mathbf{W}_1^T\mathbf{X}_C\mathbf{Y}^T\right)\mathbf{W}_2 = \mathbf{W}_2\mathbf{\Lambda}_2. \tag{6}$$

Then, the two-stage RLDA solution $\mathbf{W}_1\mathbf{W}_2\begin{bmatrix}\mathbf{I}_p\\ \mathbf{O}\end{bmatrix}$ is the p-dimensional GEVD solution.

Computational Complexity Analysis

We use floating point operations (flops) (Golub and Van Loan 1996) to measure the operation counts. Note that flops do not consider several issues, such as memory allocation.

Table 2 summarizes the dominant term of the computational complexity of the compared methods.

Algorithm 3 Two-stage RLDA ($p \leq r_B$)

1. Solve $\mathbf{W}_1 = \operatorname{argmin}_{\mathbf{W}}\|\mathbf{X}_C^T\mathbf{W} - \mathbf{Y}^T\|_F^2 + \gamma\|\mathbf{W}\|_F^2$ where $\mathbf{Y} \in \{\mathbf{Y}|\mathbf{C}\mathbf{Y}^T\mathbf{Y}\mathbf{C} = \mathbf{C}_B\}$.
2. Find the EVD of $\mathbf{W}_1^T\mathbf{X}_C\mathbf{Y}^T$, $(\mathbf{W}_1^T\mathbf{X}_C\mathbf{Y}^T)\mathbf{W}_2 = \mathbf{W}_2\mathbf{\Lambda}_2$.
3. Select the first p columns of $\mathbf{W}_1\mathbf{W}_2$ and normalize them: for $k = 1\ldots p$, $\mathbf{W}(:,k) = \mathbf{W}_1\mathbf{W}_2(:,k)/\|\mathbf{W}_1\mathbf{W}_2(:,k)\|_2$.
4. **return** $\mathbf{W}$

Table 2: Summary of computational complexity (Golub and Van Loan 1996) (Anderson 1999) (Lehoucq, Sorensen, and Yang 1998) (Stewart 2001). Note that $s = \min(n,d)$, h is the number of iterations, and $l \geq c$ is the number of Lanczos vectors used in the implicitly restarted Lanczos method (Lehoucq, Sorensen, and Yang 1998). In the MATLAB function "eigs", at least $l \geq 2c$ is recommended.

Algorithm	Dominant Term (Flops)
GEVD	$nd^2 + \frac{16}{3}d^3 + cd^2$
Iterative GEVD	$nd^2 + \frac{1}{3}d^3 + 2hld^2 + 2h(l^2-c^2)d + O(hl^3)$
LS	$nds + \frac{1}{3}s^3 + 2cnd + 2cs^2$

GEVD and iterative GEVD are implemented by the MATLAB functions "eig" and "eigs", respectively; "eig" directly finds the GEVD fully, while "eigs" iteratively finds the selected number of eigenvalues and eigenvectors in descending order. LS is implemented with Cholesky factorization (Golub and Van Loan 1996) and matrix inversion by forward/backward substitution. Note that the computational complexity of LS is related to $s = \min(n,d)$ because the RLS solution can be represented in terms of the Gram matrix of $\mathbf{X}_C$:

$$\mathbf{W}_{RLS} = \mathbf{X}_C(\mathbf{X}_C^T\mathbf{X}_C + \gamma\mathbf{I})^{-1}\mathbf{Y}^T.$$

Therefore, LS is faster than both GEVD and iterative GEVD when the data are undersampled ($n < d$). In an oversampled case ($n > d$), although the first two dominant terms of iterative GEVD and LS are the same, the rest of the terms of iterative GEVD are larger than those of LS.

Experimental Results

In this section, we verify the range equivalence and the correctness of the two-stage solutions. We also investigate the performance of the proposed algorithms. All experiments were done in MATLAB on a PC with an Intel Core i7-3610QM CPU at 2.30 GHz and with 8 GB RAM.

Data Sets

We used three data sets for our experiment: the extended Yale Face Database B (Georghiades, Belhumeur, and Kriegman 2001), the MNIST database of handwritten digits (LeCun and Cortes 1998), and Isolet (Bache and Lichman 2013). For each data set, we chose two different sizes of training data so that the experiments cover both undersampled and oversampled problems, which are marked as (u) and (o), respectively. We normalized the data so that their norm is one. The data sets satisfy the condition $r_B = c-1$. The data sets are summarized in Table 3.

Table 3: Summary of data sets.

Data Set	# of Training Samples (n)	# of Test Samples	Dimension (d)	# of Classes (c)
Yale B (u)	608	608	896	38
Yale B (o)	1806			
MNIST (u)	600	10000	784	10
MNIST (o)	6000			
Isolet (u)	520	1559	617	26
Isolet (o)	6238			

Compared Algorithms

There are two categories of methods to solve the LDA problem that we discussed in the previous sections: GEVD and LS. All RLDA solutions obtained by GEVD are normalized: $\|\mathbf{w}_k\|_2 = 1$ for $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_p]$. An iterative GEVD (I-GEVD) based on implicitly restarted Arnoldi method (Lehoucq, Sorensen, and Yang 1998) is also compared. On the other hand, we compare two LS algorithms with fixed $\mathbf{Y}$s, aRULDA and ROLDA, and three LS algorithms with $\mathbf{Y} = \mathbf{ZL}$, where $\mathbf{Z}$ is randomly selected, referred to as type 1, 2, and 3, to verify the range equivalence. First, aRULDA is the fast approximation of RULDA implemented in Algorithm 1, and ROLDA is the regularized version of OLDA implemented in Algorithm 2, respectively. The other three LS have different properties: type 1 has $\mathrm{rank}\,(\mathbf{Z}) = c$, type 2 has $\{\mathrm{rank}\,(\mathbf{Z}) = c-1$ and $\mathrm{rank}\,([\mathbf{Z}^T \quad \mathbf{1}]) = c\}$, and type 3 has $\mathrm{rank}\,(\mathbf{Z}) = \mathrm{rank}\,([\mathbf{Z}^T \quad \mathbf{1}]) = c-1$, so that type 3 does not satisfy (3). Note that Theorem 2 guarantees the validity of type 1 and type 2, but not type 3. The three LS algorithms, type 1, 2, and 3, are also used to verify the correctness of the two-stage solution. Finally, 2sRLDA is the two-stage RLDA based on RLS and EVD, which is implemented in Algorithm 3. Both $\mathbf{L}_B$ and $\mathbf{Y}_B$ can be used and there is no difference in terms of performance.

Results

Let $\mathbf{W} = \mathbf{QR}$ be the thin QR decomposition of $\mathbf{W}$. To verify the range equivalence of two categories of methods, $\mathrm{ran}\,(\mathbf{W}_{RLDA}) = \mathrm{ran}\,(\mathbf{W}_{RLS})$ as claimed in Theorem 2, we use the fact that $\mathrm{ran}\,(\mathbf{W}_{RLDA}) = \mathrm{ran}\,(\mathbf{W}_{RLS})$ if and only if $\mathbf{Q}_{RLDA}\mathbf{Q}_{RLDA}^T = \mathbf{Q}_{RLS}\mathbf{Q}_{RLS}^T$, which holds because $\mathbf{QQ}^T$ is an orthogonal projection onto $\mathrm{ran}\,(\mathbf{W})$ and there is an one-to-one correspondence between $\mathbf{QQ}^T$ and $\mathrm{ran}\,(\mathbf{W})$ (Golub and Van Loan 1996). Therefore, we can empirically verify the correctness of Theorem 2 by checking whether $\mathbf{Q}_{RLDA}\mathbf{Q}_{RLDA}^T = \mathbf{Q}_{RLS}\mathbf{Q}_{RLS}^T$ holds. On the other hand, note that eigenvectors can be permuted or reflected, and these operations can be summarized as an orthogonal matrix: any eigenvector matrix can be expressed as $\mathbf{W}_*\mathbf{P}$, where $\mathbf{P}$ is orthogonal. Since $\mathbf{PP}^T = \mathbf{I}$, we compared $\mathbf{WW}^T = \mathbf{W}_*\mathbf{W}_*^T$ to verify the correctness of the two-stage solutions. The regularization parameter is set to be 10^{-4} on extended Yale B and 1 on the others. The regularization parameter selection is not included, which is out of scope of this paper. After the dimensionality reduction, we used a nearest centroid method for classification, which is a Euclidean distance classifier: $\mathbf{l} = \mathbf{e}_{k_*}$ where $k_* = \mathrm{argmin}_k \|\mathbf{x} - \mathbf{m}_k\|_2$.

Table 4 supports our theoretical results. First, observe that type 1 and type 2, which satisfy (3), have the same range as GEVD, while type 3 does not. Our proposed algorithms, aRULDA and ROLDA, do as well. Next, as we expected, 2sRLDA, type 1, and type 2 make the same solution as GEVD using two-stage method, while type 3 does not. From Table 5, observe that the training time is much less in the LS methods than in the GEVD methods, while the test set accuracy is similar.

Table 4: The value of $\|\mathbf{QQ}^T - \mathbf{Q}_*\mathbf{Q}_*^T\|_2$ and $\|\mathbf{WW}^T - \mathbf{W}_*\mathbf{W}_*^T\|_2$ for verifying the range equivalence and the correctness of the two-stage solutions, respectively.

Data Set	Range Equivalence: $\|\mathbf{QQ}^T - \mathbf{Q}_*\mathbf{Q}_*^T\|_2$					
	I-GEVD	LS		LS		
	RLDA	aRULDA	ROLDA	Type 1	Type 2	Type 3
Yale B (u)	2.4e-12	1.5e-11	4.7e-10	3.2e-09	1.5e-11	1.0e+00
Yale B (o)	1.4e-11	4.6e-13	4.3e-11	1.8e-10	1.9e-11	1.0e+00
MNIST (u)	1.5e-14	6.8e-15	6.3e-14	4.9e-13	4.0e-14	1.0e+00
MNIST (o)	6.5e-14	9.3e-14	5.6e-13	7.6e-12	1.2e-12	1.0e+00
Isolet (u)	3.0e-14	8.0e-15	1.4e-13	1.3e-11	9.2e-14	9.9e-01
Isolet (o)	1.4e-13	4.1e-14	2.3e-12	9.9e-12	6.4e-13	1.0e+00

Data Set	Solution Correctness: $\|\mathbf{WW}^T - \mathbf{W}_*\mathbf{W}_*^T\|_2$				
	I-GEVD	LS-EVD	LS-GEVD		
	RLDA	2sRLDA	Type 1	Type 2	Type 3
Yale B (u)	2.7e-12	2.4e-10	2.4e-09	1.8e-11	1.0e+00
Yale B (o)	2.1e-11	6.4e-13	1.2e-11	1.9e-11	1.1e+00
MNIST (u)	1.6e-14	8.5e-15	5.5e-14	4.5e-14	1.1e+00
MNIST (o)	6.6e-14	1.6e-14	2.3e-13	1.1e-12	1.0e+00
Isolet (u)	3.0e-14	1.3e-14	8.1e-14	9.4e-14	1.0e+00
Isolet (o)	1.3e-13	5.4e-14	2.6e-13	7.4e-13	1.0e+00

Table 5: Test set accuracy and training time.

Data Set	Test Set Accuracy (%)				
	GEVD	I-GEVD	LS		LS-EVD
	RLDA	RLDA	aRULDA	ROLDA	2sRLDA
Yale B (u)	93.59	93.59	92.43	89.64	93.59
Yale B (o)	98.85	98.85	98.68	97.70	98.85
MNIST (u)	85.60	85.60	85.85	84.65	85.60
MNIST (o)	87.86	87.86	87.99	87.43	87.86
Isolet (u)	87.30	87.30	86.08	85.44	87.30
Isolet (o)	94.55	94.55	93.91	93.84	94.55

Data Set	Training Time (msec)				
	GEVD	I-GEVD	LS		LS-EVD
	RLDA	RLDA	aRULDA	ROLDA	2sRLDA
Yale B (u)	373.5	178.7	18.1	17.6	20.3
Yale B (o)	424.2	202.6	61.3	55.8	59.2
MNIST (u)	201.0	68.0	15.9	15.4	16.9
MNIST (o)	311.5	151.9	93.6	105.2	96.2
Isolet (u)	150.4	68.2	11.5	11.1	12.3
Isolet (o)	209.2	135.4	75.9	77.5	93.9

Conclusion

A relation between LDA and LS is discussed in this paper. LDA has the same solution subspace with LS if and only if the dependent variable matrix $\mathbf{Y}$ is in the form of $\mathbf{ZL}$ where $\mathrm{rank}\,(\mathbf{ZC}) = c-1$. That is, the solution subspace, where the solution maximizes between-class scatter while minimizing within-class scatter, is equal to the LS solution subspace, where the solution clusters all data according to the class. In addition, based on the relation, we generalized the two-stage algorithm by enlarging the set of dependent variable matrices.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) 2010-0028680 and 2014-003140.

References

Anderson, E. 1999. *LAPACK Users' Guide*, volume 9. SIAM.

Bache, K., and Lichman, M. 2013. UCI machine learning repository.

Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Springer.

Cai, D.; He, X.; and Han, J. 2008. SRDA: An efficient algorithm for large-scale discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering* 20(1):1–12.

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

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

Georghiades, A.; Belhumeur, P.; and Kriegman, D. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(6):643–660.

Golub, G., and Van Loan, C. 1996. *Matrix Computations*. Johns Hopkins Univ Press, 3rd edition.

Hastie, T.; Tibshirani, R.; and Buja, A. 1994. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association* 89(428):1255–1270.

Jin, Z.; Yang, J.; Hu, Z.; and Lou, Z. 2001. Face recognition based on the uncorrelated discriminant transformation. *Pattern Recognition* 34(7):1405–1416.

LeCun, Y., and Cortes, C. 1998. The MNIST database of handwritten digits.

Lehoucq, R. B.; Sorensen, D. C.; and Yang, C. 1998. *ARPACK Users' Guide: Solution of Large-scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, volume 6. SIAM.

Penrose, R. 1955. A generalized inverse for matrices. In *Proceedings of the Cambridge Philosophical Society*, volume 51, 406–413. Cambridge Univ Press.

Stewart, G. 2001. *Matrix Algorithms: Eigensystems*, volume 2. Society for Industrial Mathematics.

Sun, L.; Ceran, B.; and Ye, J. 2010. A scalable two-stage approach for a class of dimensionality reduction techniques. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 313–322. ACM.

Ye, J. 2006. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research* 6(1):483–502.

Ye, J. 2007. Least squares linear discriminant analysis. In *Proceedings of the 24th International Conference on Machine Learning*, 1087–1093. ACM.

Zhang, Z.; Dai, G.; Xu, C.; and Jordan, M. I. 2010. Regularized discriminant analysis, ridge regression and beyond. *Journal of Machine Learning Research* 99:2199–2228.