A Flexible and Efficient Algorithm for Regularized Fisher Discriminant Analysis

Zhihua Zhang[1], Guang Dai[1], and Michael I. Jordan[2]

[1] College of Computer Science and Technology
Zhejiang University
Hangzhou, Zhejiang 310027, China
{zhzhang, daiguang116}@gmail.com
[2] Department of Statistics and Division of Computer Science
University of California, Berkeley
Berkeley, CA 94720 USA
jordan@cs.berkeley.edu

Abstract. Fisher linear discriminant analysis (LDA) and its kernel extension—kernel discriminant analysis (KDA)—are well known methods that consider dimensionality reduction and classification jointly. While widely deployed in practical problems, there are still unresolved issues surrounding their efficient implementation and their relationship with least mean squared error procedures. In this paper we address these issues within the framework of regularized estimation. Our approach leads to a flexible and efficient implementation of LDA as well as KDA. We also uncover a general relationship between regularized discriminant analysis and ridge regression. This relationship yields variations on conventional LDA based on the pseudoinverse and a direct equivalence to an ordinary least squares estimator. Experimental results on a collection of benchmark data sets demonstrate the effectiveness of our approach.

1 Introduction

In this paper we are concerned with the supervised dimensionality reduction problem, an enduring issue in data mining and machine learning. Fisher linear discriminant analysis (LDA) provides a classical example of supervised dimension reduction. LDA estimates an effective dimension reduction space defined by linear transformations by maximizing the ratio of between-class scatter to within-class scatter.

The LDA formulation reduces to the solution of a generalized eigenproblem [6] that involves the pooled between-class scatter matrix and total scatter matrix of the input vectors. To solve the generalized eigenproblem, LDA typically requires the pooled scatter matrix to be nonsingular. This can become problematic when the dimensionality is high, because the scatter matrix is likely to be singular. In applications such as information retrieval, face recognition and microarray analysis, for example, we often meet undersampled problems which are in a "small n but large p" regime; i.e., there are a small number of samples but a very large number of variables. There are two main variants of LDA in the literature that aim to deal with this issue: the *pseudoinverse* method and the *regularization* method [7, 19].

Another important family of methods for dealing with singularity is based on a two-stage process in which two symmetric eigenproblems are solved successively. This approach was pioneered by Kitter and Young [12]. Recently, Howland *et al.* [10] used this approach to introduce the generalized singular value decomposition (GSVD) [14] into the LDA solution by utilizing special representations of the pooled scatter matrix and between-class scatter matrix. A similar general approach has been used in the development of efficient approximate algorithms for LDA [2, 21]. However, the challenge of developing an efficient general implementation methodology for LDA still remains.

It is well known that LDA is equivalent to a least mean squared error procedure in the binary classification problem [3]. It is of great interest to obtain a similar relationship in multi-class problems. A significant literature has emerged to address this issue [7, 16, 20]. However, the results obtained by these authors are subject to restrictive conditions. The problem of finding a general theoretical link between LDA and least mean squares is still open.

In this paper we address the issues within a regularization framework. We propose a novel algorithm for solving the regularized LDA (RLDA) problem. Our algorithm is more efficient than the GSVD-based algorithm [10], especially in the setting of "small n but large p" problems. More importantly, our algorithm leads us to an equivalence between RLDA and a ridge estimator for multivariate linear regression [8]. This equivalence is derived in a general setting and it is fully consistent with the established result in the binary problem [3].

Our algorithm is also appropriate for the pseudoinverse variant of LDA. Indeed, we establish an equivalence between the pseudoinverse variant and an ordinary least squares (OLS) estimation problem. Thus, we believe that we completely solve the open problem concerning the relationship between the multi-class LDA problem and multivariate linear estimation problems.

LDA relies on the assumption of linearity of the data manifold. In recent years, kernel methods [18] have aimed at removing such linearity assumptions. The kernel technology can circumvent the linearity assumption of LDA, because it works by nonlinearly mapping vectors in the input space to a higher-dimensional feature space and then implementing traditional versions of LDA in the feature space. Many different approaches have been proposed to extend LDA to kernel spaces in the existing literature [1, 13, 17].

The KDA method in [13] was developed for binary problems only, and it was solved by using the relationship between KDA and the least mean squared error procedure. A more general method, known as generalized discriminant analysis (GDA) [1], requires that the kernel matrix be nonsingular. Unfortunately, centering in the feature space will violate this requirement. Park and Park [15] argued that this might break down the theoretical justification for GDA and proposed their GSVD method to avoid this requirement for nonsingularity. The approach to LDA that we present in the current paper also handles the nonsingularity issue and extends naturally to KDA, both in its regularization and pseudoinverse forms.

The paper is organized as follows. Section 2 reviews LDA and KDA. In Section 3 we propose a new algorithm for LDA as well as KDA. An equivalence between LDA

and multivariate linear regression problems is presented in Section 4. We conduct the empirical comparisons in Section 5 and conclude in Section 6.

2 Problem Formulation

We are concerned with a multi-class classification problem. Given a set of n p-dimensional data points, $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X} \subset \mathbb{R}^p$, we assume that the $\mathbf{x}_i$ are to be grouped into c disjoint classes and that each $\mathbf{x}_i$ belongs to one and only one class. Let $V = \{1, 2, \ldots, n\}$ denote the index set of the data points $\mathbf{x}_i$ and partition V into c disjoint subsets V_j; i.e., $V_i \cap V_j = \varnothing$ for $i \neq j$ and $\cup_{j=1}^{c} V_j = V$, where the cardinality of V_j is n_j so that $\sum_{j=1}^{c} n_j = n$. We also make use of a matrix representation for the partitions. In particular, we let $\mathbf{E} = [e_{ij}]$ be an $n \times c$ indicator matrix with $e_{ij} = 1$ if input $\mathbf{x}_i$ is in class j and $e_{ij} = 0$ otherwise.

In this section we review LDA and KDA solutions to this multi-class classification problem. We begin by presenting our notation.

2.1 Notation

Throughout this paper, $\mathbf{I}_m$ denotes the $m \times m$ identity matrix, $\mathbf{1}_m$ the $m \times 1$ of ones, $\mathbf{0}$ the zero vector or matrix with appropriate size, and $\mathbf{H}_m = \mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m'$ the $m \times m$ centering matrix. For an $m \times 1$ vector $\mathbf{a} = (a_1, \ldots, a_m)'$, $\mathrm{diag}(\mathbf{a})$ represents the $m \times m$ diagonal matrix with $a_1, \ldots, a_m$ as its diagonal entries. For an $m \times m$ matrix $\mathbf{A} = [a_{ij}]$, we let $\mathbf{A}^+$ be the Moore-Penrose inverse of $\mathbf{A}$, $\mathtt{tr}(\mathbf{A})$ be the trace of $\mathbf{A}$, $\mathrm{rk}(\mathbf{A})$ be the rank of $\mathbf{A}$ and $\|\mathbf{A}\|_F = \sqrt{\mathtt{tr}(\mathbf{A}'\mathbf{A})}$ be the Frobenius norm of $\mathbf{A}$.

For a matrix $\mathbf{A} \in \mathbb{R}^{p \times q}$ with $p \geq q$, we always write the the singular value decomposition (SVD) of $\mathbf{A}$ as $\mathbf{A} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{V}'$ where $\mathbf{U} \in \mathbb{R}^{p \times q}$ and $\mathbf{V} \in \mathbb{R}^{q \times q}$ are orthogonal, and $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_q)$ is arrayed in descending order of $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_q$ (≥ 0). Let the rank of $\mathbf{A}$ be $r \leq \min\{p, q\}$ (denoted $\mathrm{rk}(\mathbf{A}) = r$). The thin SVD [6] of $\mathbf{A}$ is then $\mathbf{A} = \mathbf{U}_A\boldsymbol{\Gamma}_A\mathbf{V}_A'$ where $\mathbf{U}_A \in \mathbb{R}^{p \times r}$ and $\mathbf{V}_A \in \mathbb{R}^{q \times r}$ are orthogonal, and $\boldsymbol{\Gamma}_A = \mathrm{diag}(\gamma_1, \ldots, \gamma_r)$ satisfies $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_r > 0$.

Given two matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, we refer to $(\boldsymbol{\Lambda}, \mathbf{B})$ where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_q]$ as q eigenpairs of the matrix pencil $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ if $\boldsymbol{\Phi}\mathbf{B} = \boldsymbol{\Sigma}\mathbf{B}\boldsymbol{\Lambda}$, namely,

$$\boldsymbol{\Phi}\mathbf{b}_i = \lambda_i\boldsymbol{\Sigma}\mathbf{b}_i, \quad \text{for } i = 1, \ldots, q.$$

The problem of finding eigenpairs of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ is known as a *generalized eigenproblem*.

2.2 Linear Discriminant Analysis

Let $\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ be the sample mean, and $\mathbf{m}_j = \frac{1}{n_j}\sum_{i \in V_j} \mathbf{x}_i$ to the jth class mean for $j = 1, \ldots, c$. We then have the pooled scatter matrix $\mathbf{S}_t = \sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})'$ and the pooled between-class scatter matrix $\mathbf{S}_b = \sum_{j=1}^{c} n_j(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})'$. Conventional LDA solves the generalized eigenproblem as

$$\mathbf{S}_b\mathbf{a}_j = \lambda_j\mathbf{S}_t\mathbf{a}_j, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q > \lambda_{q+1} = 0 \tag{1}$$

where $q \leq \min\{p,\ c-1\}$ and refers to $\mathbf{a}_j$ as the jth discriminant direction. Note that we ignore a multiplier $1/n$ in these scatter matrices for simplicity.

Since $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$ where $\mathbf{S}_w$ is the pooled within-class scatter matrix, LDA is equivalent to finding a solution to

$$\mathbf{S}_b\mathbf{a} = \lambda/(1-\lambda)\mathbf{S}_w\mathbf{a}.$$

We see that LDA involves solving the generalized eigenproblem in (1), which can be expressed in matrix form:

$$\mathbf{S}_b\mathbf{A} = \mathbf{S}_t\mathbf{A}\boldsymbol{\Lambda}, \tag{2}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_q]$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$. If $\mathbf{S}_t$ is nonsingular, it may be shown that

$$\mathbf{S}_t^{-1}\mathbf{S}_b\mathbf{A} = \mathbf{A}\boldsymbol{\Lambda}.$$

Thus, $(\lambda_j, \mathbf{a}_j)$ is eigenpair of $\mathbf{S}_t^{-1}\mathbf{S}_b$ and the eigenvectors corresponding to the largest eigenvalues of $\mathbf{S}_t^{-1}\mathbf{S}_b$ are used for the discriminant directions. Since $\mathrm{rk}(\mathbf{S}_b)$ is at most $c-1$, the projection will be onto a space of dimension at most $c-1$ (i.e., $q \leq c-1$).

In applications such as information retrieval, face recognition and microarray analysis, however, we often meet a "small n but large p" problem. Thus, $\mathbf{S}_t$ is usually ill-conditioned; that is, it is either singular or close to singular. In this case, $\mathbf{S}_t^{-1}\mathbf{S}_b$ cannot be computed accurately.

Let $\boldsymbol{\Pi} = \mathrm{diag}(n_1, \ldots, n_c)$, $\boldsymbol{\Pi}^{\frac{1}{2}} = \mathrm{diag}(\sqrt{n}_1, \ldots, \sqrt{n}_c)$, $\boldsymbol{\pi} = (n_1, \ldots, n_c)'$, $\sqrt{\boldsymbol{\pi}} = (\sqrt{n}_1, \ldots, \sqrt{n}_c)'$ and $\mathbf{H}_\pi = \mathbf{I}_c - \frac{1}{n}\sqrt{\boldsymbol{\pi}}\sqrt{\boldsymbol{\pi}}'$. It follows that $\mathbf{1}_n'\mathbf{E} = \mathbf{1}_c'\boldsymbol{\Pi} = \boldsymbol{\pi}'$, $\mathbf{E}\mathbf{1}_c = \mathbf{1}_n$, $\mathbf{1}_c'\boldsymbol{\pi} = n$, $\mathbf{E}'\mathbf{E} = \boldsymbol{\Pi}$, $\boldsymbol{\Pi}^{-1}\boldsymbol{\pi} = \mathbf{1}_c$, and

$$\mathbf{M} = \boldsymbol{\Pi}^{-1}\mathbf{E}'\mathbf{X}, \tag{3}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$ and $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_c]'$. In addition, we have

$$\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{H}_\pi = \mathbf{H}_n\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}} \tag{4}$$

due to $\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{H}_\pi = \mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}} - \frac{1}{n}\mathbf{1}_n\sqrt{\boldsymbol{\pi}}'$ and $\mathbf{H}_n\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}} = \mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}} - \frac{1}{n}\mathbf{1}_n\sqrt{\boldsymbol{\pi}}'$.

Using these results and the idempotency of $\mathbf{H}_n$, we reexpress $\mathbf{S}_t$ as

$$\mathbf{S}_t = \mathbf{X}'\mathbf{H}_n\mathbf{H}_n\mathbf{X} = \mathbf{X}'\mathbf{H}_n\mathbf{X}.$$

We also have

$$\begin{aligned}
\mathbf{S}_b &= \mathbf{M}'\left[\boldsymbol{\Pi} - \frac{1}{n}\boldsymbol{\pi}\boldsymbol{\pi}'\right]\mathbf{M} \\
&= \mathbf{M}'\left[\boldsymbol{\Pi}^{\frac{1}{2}} - \frac{1}{n}\boldsymbol{\pi}\sqrt{\boldsymbol{\pi}}'\right]\left[\boldsymbol{\Pi}^{\frac{1}{2}} - \frac{1}{n}\sqrt{\boldsymbol{\pi}}\boldsymbol{\pi}'\right]\mathbf{M} \\
&= \mathbf{X}'\mathbf{E}\boldsymbol{\Pi}^{-1}\boldsymbol{\Pi}^{\frac{1}{2}}\mathbf{H}_\pi\mathbf{H}_\pi\boldsymbol{\Pi}^{\frac{1}{2}}\boldsymbol{\Pi}^{-1}\mathbf{E}'\mathbf{X} \\
&= \mathbf{X}'\mathbf{H}_n\mathbf{E}\boldsymbol{\Pi}^{-1}\mathbf{E}'\mathbf{H}_n\mathbf{X}.
\end{aligned}$$

Utilizing the above representations of $\mathbf{S}_t$ and $\mathbf{S}_b$, Howland *et al.* [10] proved that the GSVD method can be used to solve the problem in (2).

There are also two variants of conventional LDA in the literature that aim to handle this problem [19]. The first variant involves replacing $\mathbf{S}_t^{-1}$ by $\mathbf{S}_t^+$ and solving the following eigenproblem:

$$\mathbf{S}_t^+ \mathbf{S}_b \mathbf{A} = \mathbf{A} \boldsymbol{\Lambda}. \tag{5}$$

Note that $\mathbf{S}_t^+$ exists and is unique [6]. Moreover, $\mathbf{S}_t^+$ is equal to $\mathbf{S}_t^{-1}$ whenever $\mathbf{S}_t$ is nonsingular. Thus, we will use (5) when $\mathbf{S}_t$ is either nonsingular or singular.

The second variant is referred to as *regularized discriminant analysis* (RDA) [4]. It replaces $\mathbf{S}_t$ by $\mathbf{S}_t + \sigma^2 \mathbf{I}_p$ and solves the following eigenproblem:

$$(\mathbf{S}_t + \sigma^2 \mathbf{I}_p)^{-1} \mathbf{S}_b \mathbf{A} = \mathbf{A} \boldsymbol{\Lambda}. \tag{6}$$

It is a well known result that LDA is equivalent to a least mean squared error procedure in the binary classification problem ($c = 2$) [3]. Recently, similar relationships have been studied for multi-class ($c > 2$) problems [7, 16, 20]. In particular, Park and Park [16] proposed an efficient algorithm for LDA via a least mean squared error procedure in the multi-class problem.

We can see that the solution $\mathbf{A}$ for (5) or (6) is not unique. For example, if $\mathbf{A}$ is the solution, then so is $\mathbf{AD}$ whenever $\mathbf{D}$ is a $q \times q$ nonsingular diagonal matrix. Thus, constraint $\mathbf{A}'(\mathbf{S}_t + \sigma^2 \mathbf{I}_p)\mathbf{A} = \mathbf{I}_q$ is typically imposed in the literature. In this paper we concentrate on the solution of (6) with or without this constraint, and investigate the connection with ridge regression problems in the multi-class setting.

2.3 Kernel Discriminant Analysis

Kernel methods [18] work in a feature space $\mathcal{F}$, which is related to the original input space $\mathcal{X} \subset \mathbb{R}^p$ by a mapping,

$$\boldsymbol{\varphi} : \mathcal{X} \to \mathcal{F}.$$

That is, $\boldsymbol{\varphi}$ is a vector-valued function which gives a vector $\boldsymbol{\varphi}(\mathbf{s})$, called a *feature vector*, corresponding to an input $\mathbf{s} \in \mathcal{X}$. In kernel methods, we are given a reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $K(\mathbf{s}, \mathbf{t}) = \boldsymbol{\varphi}(\mathbf{s})' \boldsymbol{\varphi}(\mathbf{t})$ for $\mathbf{s}, \mathbf{t} \in \mathcal{X}$. The mapping $\boldsymbol{\varphi}(\cdot)$ itself is typically not given explicitly.

In the sequel, we use the tilde notation to denote vectors and matrices in the feature space. For example, the data vectors and mean vectors in the feature space are denoted as $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{m}}_j$. Accordingly, $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n]'$ and $\tilde{\mathbf{M}} = [\tilde{\mathbf{m}}_1, \ldots, \tilde{\mathbf{m}}_n]'$ are the data matrix and mean matrix in the feature space.

Fisher kernel discriminant analysis (KDA) seeks to solve the following generalized eigenproblem:

$$\tilde{\mathbf{S}}_b \tilde{\mathbf{A}} = \tilde{\mathbf{S}}_t \tilde{\mathbf{A}} \tilde{\boldsymbol{\Lambda}}, \tag{7}$$

where $\tilde{\mathbf{S}}_t$ and $\tilde{\mathbf{S}}_b$ are the total scatter matrix and the between-class scatter matrix in $\mathcal{F}$, respectively:

$$\tilde{\mathbf{S}}_t = \sum_{i=1}^{n} (\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}})',$$

$$\tilde{\mathbf{S}}_b = \sum_{j=1}^{c} n_j (\tilde{\mathbf{m}}_j - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_j - \tilde{\mathbf{m}})'.$$

The KDA problem is to solve (7), doing so by working solely with the kernel matrix $\mathbf{K} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'$. This is done by noting [13, 15] that the eigenvectors $\tilde{\mathbf{a}}_j$ are in the space spanned by $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$ and $\tilde{\mathbf{A}}$ can be expressed as

$$\tilde{\mathbf{A}} = \sum_{i=1}^{n} (\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}})\boldsymbol{\beta}_i' = \tilde{\mathbf{X}}'\mathbf{H}_n\mathbf{B},$$

where $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n]$ is an $n \times q$ coefficient matrix. Hence, (7) is equivalent to

$$\tilde{\mathbf{X}}'\mathbf{H}_n\mathbf{E}\boldsymbol{\Pi}^{-1}\mathbf{E}'\mathbf{H}_n\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H}_n\mathbf{B} = \tilde{\mathbf{X}}'\mathbf{H}_n\mathbf{H}_n\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H}_n\mathbf{B}\tilde{\boldsymbol{\Lambda}}.$$

Pre-multiplying the equation by $\mathbf{H}_n\tilde{\mathbf{X}}$, we have a new generalized eigenvalue problem

$$\mathbf{C}\mathbf{E}\boldsymbol{\Pi}^{-1}\mathbf{E}'\mathbf{C}\mathbf{B} = \mathbf{C}\mathbf{C}\mathbf{B}\tilde{\boldsymbol{\Lambda}}, \tag{8}$$

which involves only the kernel matrix $\mathbf{K} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'$. Here $\mathbf{C} = \mathbf{H}_n\mathbf{K}\mathbf{H}_n$ is the centered kernel matrix. Moreover, given a new input vector $\mathbf{x}$, we can compute the projection $\mathbf{z}$ of its feature vector $\tilde{\mathbf{x}}$ onto $\tilde{\mathbf{A}}$ through

$$\begin{aligned}\mathbf{z} = \tilde{\mathbf{A}}'(\tilde{\mathbf{x}} - \tilde{\mathbf{m}}) &= \mathbf{B}'\mathbf{H}_n\tilde{\mathbf{X}}\Big(\tilde{\mathbf{x}} - \frac{1}{n}\tilde{\mathbf{X}}'\mathbf{1}_n\Big) \\ &= \mathbf{B}'\mathbf{H}_n\Big(\mathbf{k}_{\mathbf{x}} - \frac{1}{n}\mathbf{K}\mathbf{1}_n\Big),\end{aligned} \tag{9}$$

where $\mathbf{k}_{\mathbf{x}} = \big(K(\mathbf{x}, \mathbf{x}_1), \ldots, K(\mathbf{x}, \mathbf{x}_n)\big)'$. This shows that the kernel trick can be used for KDA.

The concern then becomes that of solving the problem (8). Although $\mathbf{K}$ can be assumed to be nonsingular, $\mathbf{C}$ is positive semidefinite but not positive definite because the centering matrix $\mathbf{H}_n$ is singular. In fact, the rank of $\mathbf{C}$ is not larger than $n-1$ because the rank of $\mathbf{H}_n$ is $n-1$. In this case, the method devised by [1] cannot be used for the problem (8). Thus, [15] proposed a GSVD-based algorithm to solve problem (8). Running this algorithm requires the complete orthogonal decomposition [6] of matrix $[\mathbf{C}\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}}, \mathbf{C}]'$, which is of size $(n+c)\times n$. Thus, this approach is infeasible for large values of n.

Another treatment is based on the following regularized variant of (8):

$$\mathbf{C}\mathbf{E}\boldsymbol{\Pi}^{-1}\mathbf{E}'\mathbf{C}\mathbf{B} = (\mathbf{C}\mathbf{C} + \sigma^2\mathbf{I}_n)\mathbf{B}\tilde{\boldsymbol{\Lambda}}. \tag{10}$$

This RKDA problem can be equivalently expressed as

$$(\tilde{\mathbf{S}}_t + \sigma^2\mathbf{I}_d)^{-1}\tilde{\mathbf{A}} = \tilde{\mathbf{S}}_b\tilde{\mathbf{A}}\tilde{\boldsymbol{\Lambda}}, \tag{11}$$

where d is the dimension of the feature space. Although d is possibly infinite, we here assume that it is finite but not necessarily known.

3 RR-SVD Algorithms for RDA

In this section, we propose a novel approach to solving the RLDA problem in (6). We then extend this approach for the solution of the RKDA problem in (11).

3.1 The Algorithm for RLDA

We reformulate the eigenproblem in (6) as

$$\mathbf{G}\boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{E}'\mathbf{H}_n\mathbf{X}\mathbf{A} = \mathbf{A}\boldsymbol{\Lambda}, \tag{12}$$

where

$$\begin{aligned}\mathbf{G} &= (\mathbf{X}'\mathbf{H}_n\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{H}_n\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}} \\ &= (\mathbf{X}'\mathbf{H}_n\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{M}'\boldsymbol{\Pi}^{\frac{1}{2}}\mathbf{H}_\pi\end{aligned} \tag{13}$$

due to (3) and (4). We also have

$$\mathbf{G} = \mathbf{X}'\mathbf{H}_n(\mathbf{H}_n\mathbf{X}\mathbf{X}'\mathbf{H}_n + \sigma^2\mathbf{I}_n)^{-1}\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}} \tag{14}$$

due to $(\mathbf{X}'\mathbf{H}_n\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{H}_n = \mathbf{X}'\mathbf{H}_n(\mathbf{H}_n\mathbf{X}\mathbf{X}'\mathbf{H}_n + \sigma^2\mathbf{I}_n)^{-1}$. This implies that if $n < p$, we may wish to use (14) to reduce the computational cost. More importantly, we will see that (14) plays a key role in the development of an efficient algorithm for KDA to be presented shortly.

Let $\mathbf{R} = \boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{E}'\mathbf{H}_n\mathbf{X}\mathbf{G}$. Since $\mathbf{G}\boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{E}'\mathbf{H}_n\mathbf{X}$ ($p\times p$) and $\mathbf{R}$ ($c\times c$) have the same nonzero eigenvalues [9], the λ_j, $j = 1, \ldots, q$, are the nonzero eigenvalues of $\mathbf{R}$. Moreover, if $(\boldsymbol{\Lambda}, \mathbf{V})$ is the nonzero eigenpair of $\mathbf{R}$, $(\boldsymbol{\Lambda}, \mathbf{G}\mathbf{V})$ is the nonzero eigenpair of $\mathbf{G}\boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{E}'\mathbf{H}_n\mathbf{X}$. Note that

$$\begin{aligned}\mathbf{R} &= \boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{E}'\mathbf{H}_n\mathbf{X}(\mathbf{X}'\mathbf{H}_n\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{H}_n\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}} \\ &= \mathbf{H}_\pi\boldsymbol{\Pi}^{\frac{1}{2}}\mathbf{M}\mathbf{G}.\end{aligned} \tag{15}$$

This shows that $\mathbf{R}$ is positive semidefinite. Thus, its SVD is equivalent to the eigenvalue decomposition.

We thus develop an algorithm for solving the RLDA problem in (6). This is a two-stage process, which is presented in Algorithm 1. We will prove that the first stage is equivalent to the solution to a ridge regression (RR) problem in Section 4. Thus, we refer to this two-stage process as an RR-SVD algorithm. It is easily obtained that

$$\mathbf{A}'(\mathbf{S}_t + \sigma^2\mathbf{I}_p)\mathbf{A} = \boldsymbol{\Gamma}_R \quad \text{and} \quad \mathbf{A}'\mathbf{S}_b\mathbf{A} = \boldsymbol{\Gamma}_R^2.$$

This implies that $\mathbf{A}\boldsymbol{\Gamma}_R^{-\frac{1}{2}}$ is also a solution of problem (6) such that $\boldsymbol{\Gamma}_R^{-\frac{1}{2}}\mathbf{A}'(\mathbf{S}_t + \sigma^2\mathbf{I}_p)\mathbf{A}\boldsymbol{\Gamma}_R^{-\frac{1}{2}} = \mathbf{I}_q$.

The first stage calculates $\mathbf{G}$ by either (13) or (14). The computational complexity is $O(m^3)$ where $m = \min(n, p)$. The second stage makes use of the thin SVD of $\mathbf{R}$ and the computational complexity is $O(c^3)$. If both n and p are large, we recommend to use the incomplete Cholesky decomposition of $\mathbf{H}_n\mathbf{X}\mathbf{X}'\mathbf{H}_n$ (or $\mathbf{X}'\mathbf{H}_n\mathbf{X}$) and then the Sherman-Morrison-Woodbury formula [6] for calculating $(\mathbf{H}_n\mathbf{X}\mathbf{X}'\mathbf{H}_n + \sigma^2\mathbf{I}_n)^{-1}$ (or $(\mathbf{X}'\mathbf{H}_n\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}$). Compared with the GSVD-based algorithm [15], the RR-SVD algorithm is more efficient for a "small n but large p" problem.

When $\sigma^2 = 0$, we can solve the problem in (5) by simply adjusting the first stage in the RR-SVD algorithm. In particular, we calculate $\mathbf{G}$ by

$$\mathbf{G} = (\mathbf{X}'\mathbf{H}_n\mathbf{X})^+\mathbf{M}'\boldsymbol{\Pi}^{\frac{1}{2}}\mathbf{H}_\pi \overset{(or)}{=} \mathbf{X}'\mathbf{H}_n(\mathbf{H}_n\mathbf{X}\mathbf{X}'\mathbf{H}_n)^+\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}}. \tag{16}$$

Algorithm 1 RR-SVD Algorithm for RLDA problem (6)

1: **procedure** RLDA($\mathbf{X}, \mathbf{E}, \boldsymbol{\Pi}, \sigma^2$)
2: Calculate $\mathbf{G}$ by (13) or (14) and $\mathbf{R}$ by (15);
3: Perform the thin SVD of $\mathbf{R}$ as $\mathbf{R} = \mathbf{V}_R \boldsymbol{\Gamma}_R \mathbf{V}_R'$;
4: Return $\mathbf{A} = \mathbf{G}\mathbf{V}_R$ or $\mathbf{G}\mathbf{V}_R \boldsymbol{\Gamma}_R^{-\frac{1}{2}}$ as the solution of RLDA problem (6).
5: **end procedure**

3.2 The Algorithm for RKDA

We now apply Algorithm 1 to the RKDA problem in (11), which is the kernel extension of RLDA in (6).

It immediately follows from (14) that

$$\tilde{\mathbf{G}} = \tilde{\mathbf{X}}'\mathbf{H}_n(\mathbf{H}_n\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H}_n + \sigma^2\mathbf{I}_n)^{-1}\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}}$$

from which, using (15), we can calculate $\tilde{\mathbf{R}}$ by

$$\tilde{\mathbf{R}} = \boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{E}'\mathbf{C}(\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}}.$$

Moreover, given a new input vector $\mathbf{x}$, we can compute the projection $\mathbf{z}$ of its feature vector $\tilde{\mathbf{x}}$ onto $\tilde{\mathbf{A}}$ through

$$\begin{aligned}\mathbf{z} = \tilde{\mathbf{A}}'(\tilde{\mathbf{x}} - \tilde{\mathbf{m}}) &= \tilde{\mathbf{V}}_{\tilde{\mathbf{R}}}'\boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{E}'(\mathbf{C} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{H}_n\tilde{\mathbf{X}}\Big(\tilde{\mathbf{x}} - \frac{1}{n}\tilde{\mathbf{X}}'\mathbf{1}_n\Big) \\ &= \tilde{\mathbf{V}}_{\tilde{\mathbf{R}}}'\boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{E}'(\mathbf{C} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{H}_n\Big(\mathbf{k}_{\mathbf{x}} - \frac{1}{n}\mathbf{K}\mathbf{1}_n\Big). \end{aligned} \tag{17}$$

This shows that we can calculate $\tilde{\mathbf{R}}$ and $\mathbf{z}$ directly using $\mathbf{K}$ and $\mathbf{k}_{\mathbf{x}}$. We thus obtain a RR-SVD algorithm for RKDA, which is given in Algorithm 2. Also, when $\sigma^2 = 0$, we can calculate $\tilde{\mathbf{R}}$ by

$$\tilde{\mathbf{R}} = \boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{E}'\mathbf{C}\mathbf{C}^{+}\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}}$$

and exploit the RR-SVD algorithm to solve the following variant of KDA:

$$\tilde{\mathbf{S}}_t^{+}\tilde{\mathbf{S}}_b\tilde{\mathbf{A}} = \tilde{\mathbf{A}}\tilde{\boldsymbol{\Lambda}}.$$

We see that the RR-SVD algorithm is more efficient than the GSVD-based algorithm [15] for the RKDA problem in (11). Recall that problem (11) is not equivalent to that in (10). Moreover, it is not feasible to develop a GSVD-based algorithm for solving problem (11). However, we also have an RR-SVD algorithm for solving (10), by replacing $\mathbf{C}$ by $\mathbf{C}\mathbf{C}$ in calculating $\tilde{\mathbf{R}}$ and (17) by (9) in calculating $\mathbf{z}$. The resulting algorithm may be less computationally efficient, however, because it involves more matrix computations.

4 Relationships Between RFDA and Ridge Regression

It is a well known result that LDA (or KDA) is equivalent to a least mean squared error procedure in the binary classification problem ($c = 2$) [3, 13]. Recently, relationships

Algorithm 2 RR-SVD Algorithm for RKDA problem (11)

procedure RKDA($\mathbf{K}, \mathbf{E}, \mathbf{k}_\mathbf{x}, \boldsymbol{\Pi}, \sigma^2$)
  Calculate $\tilde{\mathbf{R}} = \boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{E}'\mathbf{C}(\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}}$;
  Perform the thin SVD of $\tilde{\mathbf{R}}$ as $\tilde{\mathbf{R}} = \tilde{\mathbf{V}}_{\tilde{R}}\tilde{\boldsymbol{\Gamma}}_{\tilde{R}}\tilde{\mathbf{V}}'_{\tilde{R}}$;
  Calculate $\mathbf{z}$ by (17);
  Return $\mathbf{z}$ as the q-dimensional representation of $\mathbf{x}$.
end procedure

between LDA and a least mean squared error procedure in multi-class ($c > 2$) problems were discussed by [7, 16, 20].

Motivated by this line of work, we investigate a possible equivalency between RLDA and ridge regression [8]. We then go on to consider a similar relationship between RKDA and the corresponding ridge regression problem.

Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]' = \mathbf{E}\boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{H}_\pi$. That is, $\mathbf{y}_i = (y_{i1}, \ldots, y_{ic})$ is defined by

$$y_{ij} = \begin{cases} \frac{n-n_j}{n\sqrt{n_j}} & \text{if } i \in V_j, \\ -\frac{\sqrt{n_j}}{n} & \text{otherwise.} \end{cases}$$

Regarding $\{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, n\}$ as the training samples, we fit the following multivariate linear function:

$$\mathbf{f}(\mathbf{x}) = \mathbf{w}_0 + \mathbf{W}'\mathbf{x}$$

where $\mathbf{w}_0 \in \mathbb{R}^c$ and $\mathbf{W} \in \mathbb{R}^{p\times c}$. We now find ridge estimates of $\mathbf{w}_0$ and $\mathbf{W}$. In particular, we consider the following minimization problem:

$$\min_{\mathbf{w}_0,\mathbf{W}} L(\mathbf{w}_0, \mathbf{W}) = \frac{1}{2}\|\mathbf{Y}-\mathbf{1}_n\mathbf{w}_0'-\mathbf{X}\mathbf{W}\|_F^2 + \frac{\sigma^2}{2}\mathsf{tr}(\mathbf{W}'\mathbf{W}). \tag{18}$$

The solution $\mathbf{W}^*$ for (18) is

$$\mathbf{W}^* = (\mathbf{X}'\mathbf{H}_n\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{M}'\boldsymbol{\Pi}^{\frac{1}{2}}\mathbf{H}_\pi. \tag{19}$$

The derivation is given in Appendix A. It then follows from (13) that $\mathbf{W}^* = \mathbf{G}$. Moreover, when $\sigma^2 = 0$, $\mathbf{W}^*$ reduces to the ordinary least squares (OLS) estimate of $\mathbf{W}$, which is the solution of the following minimization problem:

$$\min_{\mathbf{w}_0,\mathbf{W}} L(\mathbf{w}_0, \mathbf{W}) = \frac{1}{2}\|\mathbf{Y}-\mathbf{1}_n\mathbf{w}_0'-\mathbf{X}\mathbf{W}\|_F^2. \tag{20}$$

In this case, if $\mathbf{X}'\mathbf{H}_n\mathbf{X}$ is singular, a standard treatment is to use the Moore-Penrose inverse $(\mathbf{X}'\mathbf{H}_n\mathbf{X})^+$ in (19). Such a $\mathbf{W}^*$ is identical with $\mathbf{G}$ in (16).

Consequently, we have found a relationship between the ridge estimation problem in (18) and the RLDA problem in (6). This is summarized in the following theorem.

Theorem 1. *Let $\mathbf{W}^*$ be the solutions of the ridge estimation problem in (18)* (resp. the OLS estimation problem in (20)) *and $\mathbf{A}$ be defined in Algorithm 1 for the solution of the RLDA problem in (6)* (resp. the LDA problem in (5)). *Then,*

$$\mathbf{A} = \mathbf{W}^*\mathbf{V}_R$$

where the columns of $\mathbf{V}_R$ *are the eigenvectors of* $\mathbf{R}$ *associated with its* q *nonzero eigenvalues.*

Theorem 1 provides a connection between $\mathbf{A}$ and $\mathbf{W}^*$. Recall that the eigenvector matrix $\mathbf{V}_R$ ($c \times q$) is orthogonal. This leads us to the following main result of this paper.

Theorem 2. *Under the conditions in Theorem 1, we have*

$$\mathbf{A}\mathbf{A}' = \mathbf{W}^*(\mathbf{W}^*)'.$$

Accordingly, we have

$$(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{A}\mathbf{A}'(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{W}^*(\mathbf{W}^*)'(\mathbf{x}_i - \mathbf{x}_j)$$

for any $\mathbf{x}_i$ *and* $\mathbf{x}_j \in \mathbb{R}^p$.

The proof of this theorem is given in Appendix A. Theorem 2 shows that when applying a distance-based classifier such as the K-nearest neighbor (KNN) in the reduced dimensional space, the classification results obtained by multi-class LDA and multivariate linear estimators are the same. Theorem 2 holds in general. Thus we obtain a complete solution to the open problem concerning the relationship between multi-class LDA problems and multivariate linear estimators.

Similar results have been obtained by [16, 20], but under restrictive conditions. The key difference between our work and that of [16] revolves around a different definition for the label scores $\mathbf{Y}$. Ye [20] used the same definition of $\mathbf{Y}$ as ours, but they aimed to establish a connection of the solution $\mathbf{W}^*$ with a matrix $\mathbf{A}$ defined differently from ours.

5 Experimental Study

To evaluate the performance of the proposed algorithm for LDA and KDA, we conducted experimental comparisons with other related algorithms for LDA and KDA on several real-world data sets. In particular, the comparison was implemented on four face datasets [3], two gene datasets, the USPS dataset, the "letters" dataset and the WebKB dataset. Table 1 summarizes the benchmark datasets we used. All algorithms were implemented in Matlab on a PC configured with an Intel Dual Core 2.53GHz CPU and 2.06GB of memory. Matlab code to implement the algorithms can be obtained from the first author.

In our experiments, each dataset was randomly partitioned into disjoint training and test data sets, according to the percentage n/k listed in the last column of Table 1. Ten random partitions were obtained for each data set, and several evaluation criteria were reported, including average classification accuracy rate, standard deviation and average computational time.

In the linear setting, we compared our method to the LDA/GSVD method [11] and the LDA/MSE method [16]. In the kernel setting, we compared our method to the KDA/GSVD method [15] and the KDA/MSE, i.e., the kernel-based extensions

[3] The YaleB(+E) dataset was collected from the YaleB database and its extension.

Table 1. Summary of the benchmark datasets: $c-$the number of classes; $p-$the dimension of the input vector; $k-$the size of the dataset; $n-$the number of the training data.

Data set	c	p	k	n/k
ORL	40	1024	400	40%
Yale	15	1024	165	50%
YaleB(+E)	38	1024	2414	30%
PIE	68	1024	6800	20%
11_Tumors	11	12533	174	31%
14_Tumors	25	15009	305	41%
USPS	10	256	2007	10%
Letters	3	17	2341	5%
WebKB	4	300	4192	10%

of the two linear methods in above. All of the hyperparameters (such as σ^2) were selected by cross-validation [5]. In the kernel setting, the RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\theta^2)$ was employed, and θ was set to the mean Euclidean distance between training data points. After having obtained the q-dimensional representations $\mathbf{z}_i$ of the $\mathbf{x}_i$ from each dimensionality reduction method, we used a simple nearest neighbor classifier to evaluate the classification accuracy.

Figure 1 presents comparative classification results on the four face datasets. We implemented our Algorithm 1 with both $\mathbf{A}$ and $\mathbf{A}\boldsymbol{\Gamma}_R^{-\frac{1}{2}}$. We found that the result with $\mathbf{A}$ was slightly better than that with $\mathbf{A}\boldsymbol{\Gamma}_R^{-\frac{1}{2}}$. Moreover, a similar result was found for kernel leaning. The results reported here were based on the setting with $\mathbf{A}$. From Figure 1, it is clear that in the linear and kernel settings, our method has better classification accuracy than that of the LDA/GSVD and LDA/MSE methods over a range of choices of number of discriminant covariates. Moreover, an appealing characteristic of our method is its effectiveness when q (i.e., the number of discriminant covariates) is small.

We also compared the computational time of the different methods in the linear and kernel settings on the four face datasets. Figure 2 shows the comparisons with respect to the training percentage n/k on the four face datasets. We can also see that our method has an overall low computational time in comparison with the other methods on the four face datasets. As the training percentage n/k increases, our method yields more efficient performance.

Finally, Tables 2 and 3 summarize the different evaluation criteria on all the data sets. As these results show, our method yields accurate and computationally efficient performance in both the linear and kernel settings. Additionally, it should be mentioned here that the data sets in our experiments range over small sample and large sample problems.

6 Conclusion

In this paper we have provided an appealing solution to an open problem concerning the relationship between multi-class LDA problems and multivariate linear estimators,

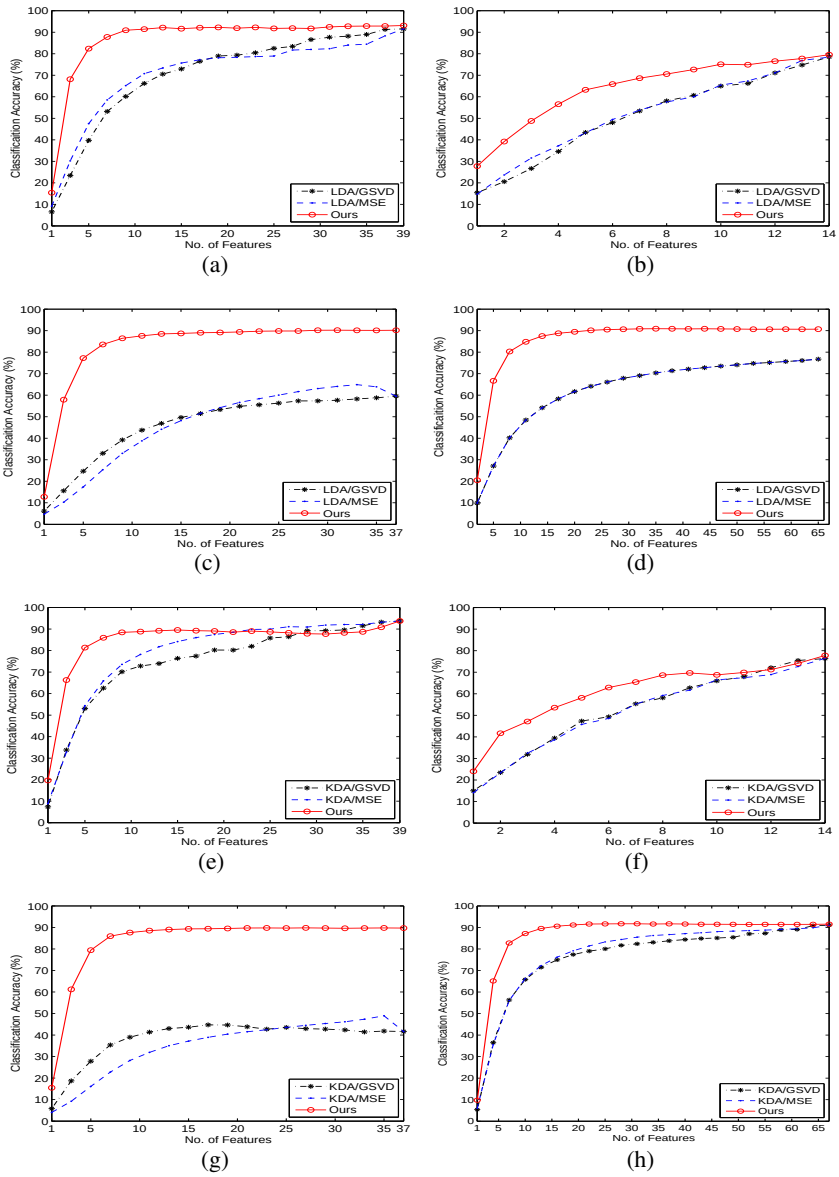

Fig. 1. Comparison of the three different methods on the four face datasets, where (a)∼(d) denote the results in the linear setting and (e)∼(h) denote the results in the kernel setting: (a) ORL−linear; (b) Yale−linear; (c) YaleB(+E)−linear; (d) PIE−linear; (e) ORL−kernel; (f) Yale−kernel; (g) YaleB(+E)−kernel; (h) PIE−kernel. Here "No. of features" is equal to q.

both in the linear setting and the kernel setting. Our theory has yielded efficient and effective algorithms for LDA and KDA within both the regularization and pseudoinverse frameworks. The favorable performance of our algorithms has been demonstrated em-

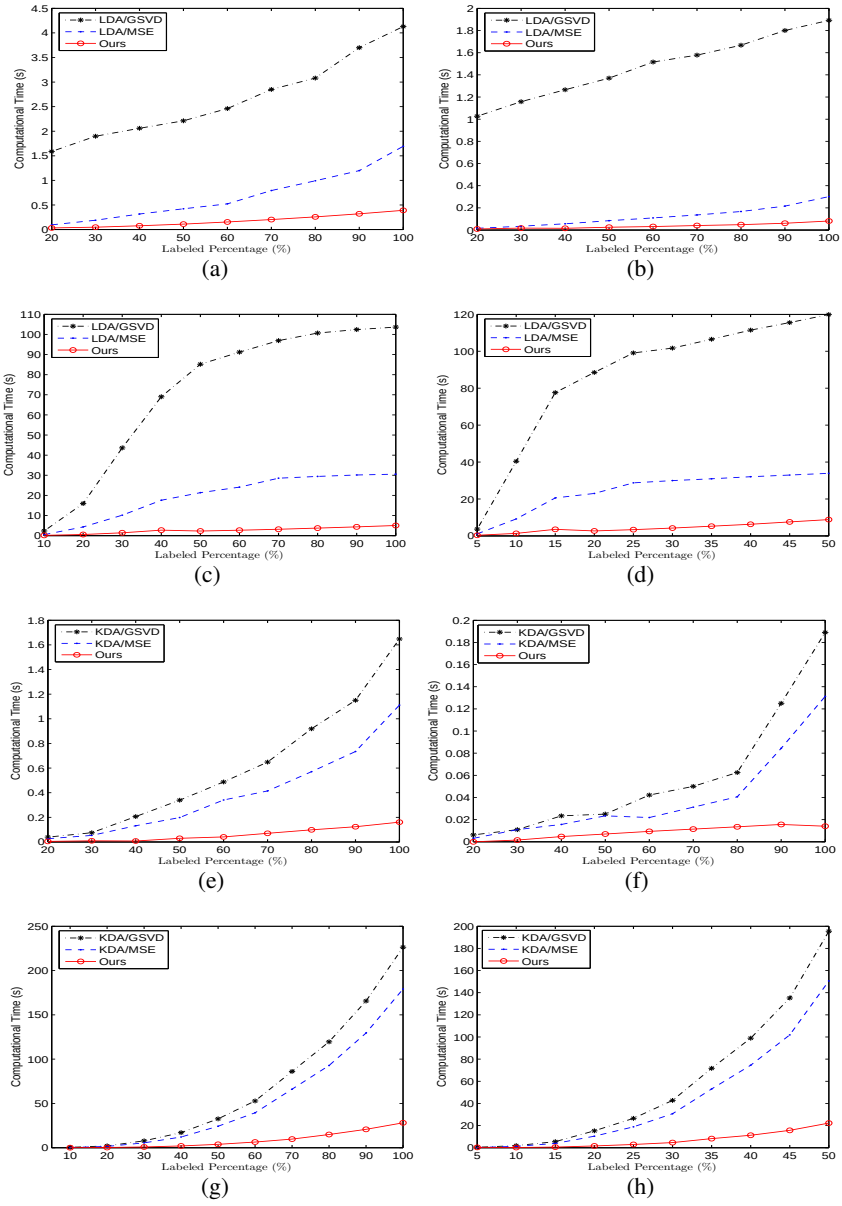

Fig. 2. Comparison of the computational times for the three different methods as the training percentage k/n increases on the four face datasets, where (a)∼(d) denote the results in the linear setting and (e)∼(h) denote the results in the kernel setting: (a) ORL−linear; (b) Yale−linear; (c) YaleB(+E)−linear; (d) PIE−linear; (e) ORL−kernel; (f) Yale−kernel; (g) YaleB(+E)−kernel; (h) PIE−kernel.

Table 2. Experimental results of the three methods on different datasets in the linear setting: $acc-$ the best classification accuracy percentage; $std-$ the corresponding standard deviation; $q-$ the corresponding number of discriminant covariates; $time-$ the corresponding computational time (s).

Dataset	LDA/GSVD			LDA/MSE			Ours		
	$acc\,(\pm std)$	q	$time$	$acc\,(\pm std)$	q	$time$	$acc\,(\pm std)$	q	$time$
ORL	91.54 (±1.98)	39	2.059	91.58 (±2.00)	39	0.316	93.13 (±2.00)	39	0.077
Yale	78.56 (±2.29)	14	1.370	78.44 (±2.47)	14	0.084	79.56 (±3.75)	14	0.025
YaleB(+E)	59.54 (±11.8)	37	43.62	65.19 (±8.36)	34	10.17	90.20 (±1.09)	31	1.422
PIE	77.00 (±0.81)	67	88.51	77.01 (±0.81)	67	23.01	90.91 (±0.55)	45	2.681
11_Tumors	92.35 (±1.51)	10	0.652	90.25 (±2.10)	10	0.877	92.44 (±1.43)	10	0.495
14_Tumors	66.33 (±1.82)	24	2.808	64.94 (±1.69)	24	4.499	66.39 (±1.91)	24	2.035
USPS	52.23 (±3.02)	9	0.979	52.24 (±3.02)	9	0.579	86.84 (±1.31)	9	0.114
Letters	89.16 (±0.81)	2	0.129	89.16 (±0.81)	2	0.021	89.27 (±0.86)	2	0.102
WebKB	65.92 (±1.77)	3	2.348	65.92 (±1.77)	3	1.635	81.76 (±0.87)	3	0.225

Table 3. Experimental results of the three methods on different datasets in the kernel setting: $acc-$ the best classification accuracy percentage; $std-$ the corresponding standard deviation; $q-$ the corresponding number of discriminant covariates; $time-$ the corresponding computational time (s).

Dataset	KDA/GSVD			KDA/MSE			Ours		
	$acc\,(\pm std)$	q	$time$	$acc\,(\pm std)$	q	$time$	$acc\,(\pm std)$	q	$time$
ORL	93.75 (±1.95)	39	0.339	93.75 (±1.89)	39	0.198	93.75 (±1.73)	39	0.031
Yale	76.33 (±2.67)	14	0.025	76.44 (±2.67)	14	0.023	77.78 (±2.87)	14	0.007
YaleB(+E)	44.74 (±24.6)	17	7.751	48.95 (±25.4)	35	5.477	89.80 (±1.02)	27	0.844
PIE	91.04 (±0.49)	67	46.84	91.04 (±0.49)	67	36.44	91.77 (±0.43)	26	8.811
11_Tumors	88.74 (±2.94)	10	0.031	89.16 (±2.36)	10	0.022	89.58 (±2.10)	10	0.011
14_Tumors	59.56 (±2.76)	24	0.170	60.56 (±2.73)	10	0.108	66.33 (±1.51)	24	0.035
USPS	85.13 (±3.57)	9	0.697	85.15 (±3.57)	9	0.493	89.22 (±1.42)	9	0.082
Letters	95.25 (±0.99)	2	0.120	95.24 (±0.99)	2	0.071	99.36 (±1.20)	2	0.024
WebKB	74.79 (±2.78)	3	4.285	74.79 (±2.77)	3	3.255	83.35 (±1.31)	3	0.499

pirically on a collection of benchmark data sets. In future work we plan to extend our algorithms to a broader class of generalized eigenproblems.

A Proof of Theorem 1

Proof. The first-order derivatives of $L(\mathbf{w}_0, \mathbf{W})$ with respect to $\mathbf{w}_0$ and $\mathbf{W}$ are given by

$$\frac{\partial L}{\partial \mathbf{w}_0} = n\mathbf{w}_0 + \mathbf{W}'\mathbf{X}'\mathbf{1}_n - \mathbf{Y}'\mathbf{1}_n,$$
$$\frac{\partial L}{\partial \mathbf{W}} = (\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{I}_p)\mathbf{W} + \mathbf{X}'\mathbf{1}_n\mathbf{w}_0' - \mathbf{X}'\mathbf{Y}.$$

Letting $\frac{\partial L}{\partial \mathbf{w}_0} = \mathbf{0}$, $\frac{\partial L}{\partial \mathbf{W}} = \mathbf{0}$ and $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \frac{1}{n}\mathbf{X}'\mathbf{1}_n$, we obtain

$$\begin{cases} \mathbf{w}_0 + \mathbf{W}'\bar{\mathbf{x}} = \mathbf{0} \\ n\bar{\mathbf{x}}\mathbf{w}_0' + (\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{I}_p)\mathbf{W} = \mathbf{M}'\boldsymbol{\Pi}^{\frac{1}{2}}\mathbf{H}_\pi \end{cases}$$

due to $\mathbf{Y}'\mathbf{1}_n = \mathbf{0}$ and $\mathbf{X}'\mathbf{Y} = \mathbf{M}'\boldsymbol{\Pi}^{\frac{1}{2}}\mathbf{H}_\pi$. Further, it follows that $\mathbf{w}_0 = -\mathbf{W}\bar{\mathbf{x}}$, and hence,

$$(\mathbf{X}'\mathbf{H}_n\mathbf{X} + \sigma^2\mathbf{I}_p)\mathbf{W} = \mathbf{M}'\boldsymbol{\Pi}^{\frac{1}{2}}\mathbf{H}_\pi$$

because $\mathbf{X}'\mathbf{X} - n\bar{\mathbf{x}}\bar{\mathbf{x}}' = \mathbf{X}'\mathbf{H}_n\mathbf{X}$. We thus obtain $\mathbf{W}^*$ in (19). It then follows from (13) that $\mathbf{W}^* = \mathbf{G}$. Moreover, when $\sigma^2 = 0$, $\mathbf{W}^*$ reduces to the solution of the minimization problem in (20). In this case, if $\mathbf{X}'\mathbf{H}_n\mathbf{X}$ is singular, a standard treatment is to use the Moore-Penrose inverse $(\mathbf{X}'\mathbf{H}_n\mathbf{X})^+$ in (19). Such a $\mathbf{W}^*$ is identical with $\mathbf{G}$ in (16).

Consequently, we have the relationship between the ridge estimation problem in (18) and the RLDA problem in (6). This is summarized in Theorem 1.

B Proof of Theorem 2

Proof. Since $\mathbf{V}_R$ is an $c\times q$ orthogonal matrix, there exists an $c\times(c-q)$ orthogonal matrix $\mathbf{V}_2$ such that $\mathbf{V} = [\mathbf{V}_R, \mathbf{V}_2]$ is an $c\times c$ orthogonal matrix. Noting that $\mathbf{R} = \mathbf{V}_R\boldsymbol{\Gamma}_R\mathbf{V}_R'$, we have $\mathbf{R}\mathbf{V}_2 = \mathbf{0}$ and $\mathbf{V}_2'\mathbf{R}\mathbf{V}_2 = \mathbf{0}$. Let $\mathbf{Q} = \mathbf{M}'\boldsymbol{\Pi}^{\frac{1}{2}}\mathbf{H}_\pi\mathbf{V}_2$. Then we obtain $\mathbf{Q}'(\mathbf{X}'\mathbf{H}_n\mathbf{X}+\sigma^2\mathbf{I}_p)^{-1}\mathbf{Q} = \mathbf{0}$. This implies $\mathbf{Q} = \mathbf{0}$ because $(\mathbf{X}'\mathbf{H}_n\mathbf{X}+\sigma^2\mathbf{I}_p)^{-1}$ is positive definite. Hence, $\mathbf{W}^*\mathbf{V}_2 = (\mathbf{X}'\mathbf{H}_n\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{Q} = \mathbf{0}$. As a result, we have

$$\begin{aligned} \mathbf{W}^*(\mathbf{W}^*)' &= \mathbf{W}^*\mathbf{V}\mathbf{V}'(\mathbf{W}^*)' \\ &= \mathbf{W}^*\mathbf{V}_R\mathbf{V}_R'(\mathbf{W}^*)' + \mathbf{W}^*\mathbf{V}_2\mathbf{V}_2'(\mathbf{W}^*)' \\ &= \mathbf{A}\mathbf{A}'. \end{aligned}$$

Note that if $\sigma^2 = 0$ and $\mathbf{X}'\mathbf{H}_n\mathbf{X}$ is nonsingular, we still have $\mathbf{W}^*(\mathbf{W}^*)' = \mathbf{A}\mathbf{A}'$. In the case that $\mathbf{X}'\mathbf{H}_n\mathbf{X}$ is singular, we have $\mathbf{Q}'(\mathbf{X}'\mathbf{H}_n\mathbf{X})^+\mathbf{Q} = \mathbf{0}$. Since $(\mathbf{X}'\mathbf{H}_n\mathbf{X})^+$ is positive semidefinite, its square root matrix (denoted $\mathbf{F}$) exists. It thus follows from $\mathbf{Q}'(\mathbf{X}'\mathbf{H}_n\mathbf{X})^+\mathbf{Q} = \mathbf{Q}\mathbf{F}\mathbf{F}\mathbf{Q}' = \mathbf{0}$ that $\mathbf{F}\mathbf{Q}' = \mathbf{0}$. This shows that $\mathbf{W}^*\mathbf{V}_2 = (\mathbf{X}'\mathbf{H}_n\mathbf{X})^+\mathbf{Q} = \mathbf{0}$. Thus, we also obtain $\mathbf{W}^*(\mathbf{W}^*)' = \mathbf{A}\mathbf{A}'$. The proof is complete.

Acknowledgements Zhihua Zhang is supported in part by program for Changjiang Scholars and Innovative Research Team in University (IRT0652, PCSIRT), China.

References

1. G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404, 2000.
2. Y.-Q. Cheng, Y.-M. Zhuang, and J.-Y. Yang. Optimal Fisher discriminant analysis using the rank decomposition. *Pattern Recognition*, 25(1):101–111, 1992.
3. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, New York, second edition, 2001.

4. J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.
5. G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.
6. G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.
7. T. Hastie, R. Tishiran, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.
8. A. E. Hoerl and R. W. Kennard. Ridge regression. *Technometrics*, 12:56–67 and 69–82, 1970.
9. R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985.
10. P. Howland, M. Jeon, and H. Park. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25(1):165–179, 2003.
11. P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):995–1006, 2004.
12. J. Kittler and P. C. Young. A new approach to feature selection based on the Karhunen-Loève expansion. *Pattern Recognition*, 5:335–352, 1973.
13. S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K. R. Müller. Invariant feature extraction and classification in kernel space. In *Advances in Neural Information Processing Systems 12*, volume 12, pages 526–532, 2000.
14. C. C. Paige and M. A. Saunders. Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, 18(3):398–405, 1981.
15. C. H. Park and H. Park. Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 27(1):87–102, 2005.
16. C. H. Park and H. Park. A relationship between linear discriminant analysis and the generalized minimum squared error solution. *SIAM Journal on Matrix Analysis and Applications*, 27(2):474–492, 2005.
17. V. Roth and V. Steinhage. Nonlinear discriminant analysis using kernel functions. In *Advances in Neural Information Processing Systems 12*, volume 12, pages 568–574, 2000.
18. J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
19. A R. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, Hoboken, NJ, 2002.
20. J. Ye. Least squares linear discriminant analysis. In *The Twenty-Fourth International Coference on Machine Learning (ICML)*, 2007.
21. J. Ye, Q. Li, H. Xiong, H. Park, R. Janardan, and V. Kumar. An incremental dimension reduction algorithm via QR decomposition. In *ACM SIGKDD*, pages 364–373, 2004.