Journal of Machine Learning Research 11 (2010) 2199-2228 Submitted 11/09; Revised 5/10; Published 8/10

Regularized Discriminant Analysis, Ridge Regression and Beyond

Zhihua Zhang ZHZHANG@ZJU.EDU.CN
Guang Dai GUANG.GDAI@GMAIL.COM
Congfu Xu XUCONGFU@ZJU.EDU.CN
College of Computer Science & Technology
Zhejiang University
Hangzhou, Zhejiang 310027, China

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Editor: Inderjit Dhillon

Abstract

Fisher linear discriminant analysis (FDA) 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 squares procedures. In this paper we address these issues within the framework of regularized estimation. Our approach leads to a flexible and efficient implementation of FDA as well as KDA. We also uncover a general relationship between regularized discriminant analysis and ridge regression. This relationship yields variations on conventional FDA based on the pseudoinverse and a direct equivalence to an ordinary least squares estimator.

Keywords: Fisher discriminant analysis, reproducing kernel, generalized eigenproblems, ridge regression, singular value decomposition, eigenvalue decomposition

1. Introduction

In this paper we are concerned with Fisher linear discriminant analysis (FDA), an enduring classification method in multivariate analysis and machine learning. It is well known that the FDA formulation reduces to the solution of a generalized eigenproblem (Golub and Van Loan, 1996) that involves the between-class scatter matrix and total scatter matrix of the data vectors. To solve the generalized eigenproblem, FDA 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; that is, there are a small number of samples but a very large number of variables. There are two main variants of FDA in the literature that aim to deal with this issue: the *pseudoinverse* method and the *regularization* method (Hastie et al., 2001; Webb, 2002).

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 Kittler and Young (1973). Recently, Howland et al. (2003) used this approach to introduce the

generalized singular value decomposition (GSVD) (Paige and Saunders, 1981) into the FDA solution by using special representations of the pooled scatter matrix and between-class scatter matrix. A similar general approach has been used in the development of efficient approximation algorithms for FDA (Cheng et al., 1992; Ye et al., 2004). However, the challenge of developing an efficient general implementation methodology for FDA still remains.

In the binary classification problem, FDA is equivalent to a least mean squared error procedure (Duda et al., 2001). It is of great interest to obtain a similar relationship in multi-class problems. A significant literature has emerged to address this issue (Hastie et al., 2001; Park and Park, 2005b; Ye, 2007). However, the results obtained by these authors are subject to restrictive conditions. The problem of finding a general theoretical link between FDA 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 FDA (RFDA) problem. Our algorithm is more efficient than the GSVD-based algorithm (Howland et al., 2003), especially in the setting of "small n but large p" problems. More importantly, our algorithm leads us to an equivalence between RFDA and a ridge estimator for multivariate linear regression (Hoerl and Kennard, 1970). This equivalence is derived in a general setting and it is fully consistent with the established result in the binary problem (Duda et al., 2001).

Our algorithm is also appropriate for the pseudoinverse variant of FDA. Indeed, we establish an equivalence between the pseudoinverse variant and an ordinary least squares (OLS) estimation problem. Thus, we are able to resolve the open problem concerning the relationship between the multi-class FDA and multivariate linear estimation problems.

FDA relies on the assumption of linearity of the data manifold. In recent years, kernel methods (Shawe-Taylor and Cristianini, 2004) have aimed at removing such linearity assumptions. The kernel technology can circumvent the linearity assumption of FDA, because it works by nonlinearly mapping vectors in the input space to a higher-dimensional feature space and then implementing traditional versions of FDA in the feature space. Many different approaches have been proposed to extend FDA to kernel spaces in the existing literature (Baudat and Anouar, 2000; Mika et al., 2000; Roth and Steinhage, 2000).

The KDA method in Mika et al. (2000) was developed for binary problems only, and it was based on using the relationship between KDA and the least mean squared error procedure. A more general method, known as generalized discriminant analysis (GDA) (Baudat and Anouar, 2000), requires that the kernel matrix be nonsingular. Unfortunately, centering in the feature space will violate this requirement. Park and Park (2005a) argued that this might break down the theoretical justification for GDA and proposed their GSVD method to avoid this requirement for nonsingularity.

KDA methods have been successfully deployed in many practical problems. The approach to FDA that we present in the current paper not only handles the nonsingularity issue but also extends naturally to KDA, both in its regularization and pseudoinverse forms. We will see that our regularized KDA is different from the existing regularization methods for KDA (see, e.g., Park and Park, 2005a), where as we discuss later, there is a problem with inconsistency of solutions. Our methods for regularized KDA derive directly from the corresponding methods for regularized FDA and avoid the inconsistency problem.

Finally, we extend our approach for FDA as well as KDA to a certain family of generalized eigenvalue problems.

The paper is organized as follows. Section 2 reviews FDA and KDA, and Section 3 presents our KDA formulations. In Sections 4 and 5 we propose two new algorithms for FDA and KDA, respectively. An equivalence between FDA and multivariate linear regression problems is presented in Section 6. We conduct empirical comparisons in Section 7. We extend the approach to a certain family of generalized eigenproblems in Section 8 and conclude in Section 9.

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; that is, $V_i \cap V_j = \varnothing$ 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 FDA and KDA solutions to this multi-class classification problem. We begin by presenting our notation.

2.1 Preliminaries

Throughout this paper, $\mathbf{I}_m$ denotes the $m \times m$ identity matrix, $\mathbf{1}_m$ the $m \times 1$ vector of ones, $\mathbf{0}$ the zero vector or matrix with appropriate size, and $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ the $n \times n$ 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 $\mathbf{A}$, $\mathrm{tr}(\mathbf{A})$ be the trace of $\mathbf{A}$, $\mathrm{rk}(\mathbf{A})$ be the rank of $\mathbf{A}$ and $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}'\mathbf{A})}$ be the Frobenius norm of $\mathbf{A}$. For an $m \times q$ real matrix $\mathbf{A}$, $\mathcal{R}(\mathbf{A})$ and $\mathcal{N}(\mathbf{A})$ denote its range and null spaces; that is, $\mathcal{R}(\mathbf{A}) = \{\mathbf{A}\mathbf{x} | \mathbf{x} \in \mathbb{R}^q\}$ and $\mathcal{N}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^q | \mathbf{A}\mathbf{x} = \mathbf{0}\}$.

For a matrix $\mathbf{A} \in \mathbb{R}^{m \times q}$ with $m \geq q$, we always express the (reduced) singular value decomposition (SVD) of $\mathbf{A}$ as $\mathbf{A} = \mathbf{U}\Gamma\mathbf{V}'$ where $\mathbf{U} \in \mathbb{R}^{m \times q}$ is a matrix with orthonormal columns (that is, $\mathbf{U}'\mathbf{U} = \mathbf{I}_q$), $\mathbf{V} \in \mathbb{R}^{q \times q}$ is orthogonal (i.e., $\mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{I}_q$), and $\Gamma = \mathrm{diag}(\gamma_1,\ldots,\gamma_q)$ 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 ($\leq \min\{m,q\}$) and denote $\mathrm{rk}(\mathbf{A}) = r$. The condensed SVD of $\mathbf{A}$ is then $\mathbf{A} = \mathbf{U}_A\Gamma_A\mathbf{V}_A'$ where $\mathbf{U}_A \in \mathbb{R}^{m \times r}$ and $\mathbf{V}_A \in \mathbb{R}^{q \times r}$ are matrices with orthonormal columns (i.e., $\mathbf{U}_A'\mathbf{U}_A = \mathbf{I}_r$ and $\mathbf{V}_A'\mathbf{V}_A = \mathbf{I}_r$), and $\Gamma_A = \mathrm{diag}(\gamma_1,\ldots,\gamma_r)$ satisfies $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_r > 0$.

Given two matrices Σ_1 and $\Sigma_2 \in \mathbb{R}^{m \times m}$, we refer to $(\Lambda, \mathbf{B})$ where $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_q)$ and $\mathbf{B} = [\mathbf{b}_1,\ldots,\mathbf{b}_q]$ as q eigenpairs of the matrix pencil (Σ_1, Σ_2) if $\Sigma_1\mathbf{B} = \Sigma_2\mathbf{B}\Lambda$; namely,

$$\Sigma_1\mathbf{b}_i = \lambda_i\Sigma_2\mathbf{b}_i, \quad \text{for } i = 1,\ldots,q.$$

The problem of finding eigenpairs of (Σ_1, Σ_2) is known as a *generalized eigenproblem*. In this paper, we especially consider the problem with the nonzero λ_i for $i = 1,\ldots,q$ and refer to $(\Lambda, \mathbf{B})$ as the nonzero eigenpairs of (Σ_1, Σ_2). If Σ_2 is nonsingular, $(\Lambda, \mathbf{B})$ is also referred to as the (nonzero) eigenpairs of $\Sigma_2^{-1}\Sigma_1$ because the generalized eigenproblem is equivalent to the eigenproblem:

$$\Sigma_2^{-1}\Sigma_1\mathbf{B} = \mathbf{B}\Lambda.$$

In the case that Σ_2 is singular, one typically resorts to a pseudoinverse eigenproblem:

$$\Sigma_2^{+}\Sigma_1\mathbf{B} = \mathbf{B}\Lambda.$$

Fortunately, we are able to establish a connection between the solutions of the generalized eigenproblem and its corresponding pseudoinverse eigenproblem. In particular, we have the following theorem, the proof of which is given in Appendix A.

Theorem 1 *Let Σ_1 and Σ_2 be two $m\times m$ real matrices. Assume $\mathcal{R}(\Sigma_1) \subseteq \mathcal{R}(\Sigma_2)$. Then, if $(\Lambda, \mathbf{B})$ are the nonzero eigenpairs of $\Sigma_2^+\Sigma_1$, we have that $(\Lambda, \mathbf{B})$ are the nonzero eigenpairs of the matrix pencil (Σ_1, Σ_2). Conversely, if $(\Lambda, \mathbf{B})$ are the nonzero eigenpairs of the matrix pencil (Σ_1, Σ_2), then $(\Lambda, \Sigma_2^+\Sigma_2\mathbf{B})$ are the nonzero eigenpairs of $\Sigma_2^+\Sigma_1$.*

As we see from Appendix A, a necessary and sufficient condition for $\mathcal{R}(\Sigma_1) \subseteq \mathcal{R}(\Sigma_2)$ is

$$\Sigma_2\Sigma_2^+\Sigma_1 = \Sigma_1.$$

Since $\mathbb{R}^m$ is equal to the direct sum of $\mathcal{R}(\Sigma_1)$ (or $\mathcal{R}(\Sigma_2)$) and $\mathcal{N}(\Sigma_1')$ (or $\mathcal{N}(\Sigma_2')$), we obtain that $\mathcal{R}(\Sigma_1) \subseteq \mathcal{R}(\Sigma_2)$ if and only if $\mathcal{N}(\Sigma_2') \subseteq \mathcal{N}(\Sigma_1')$. Furthermore, if both Σ_1 and Σ_2 are symmetric, then $\mathcal{N}(\Sigma_2) \subseteq \mathcal{N}(\Sigma_1)$ is equivalent to $\mathcal{R}(\Sigma_1) \subseteq \mathcal{R}(\Sigma_2)$.

2.2 Fisher Linear Discriminant Analysis

Let $\mathbf{m} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ be the sample mean, and let $\mathbf{m}_j = \frac{1}{n_j}\sum_{i\in V_j} \mathbf{x}_i$ be 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 between-class scatter matrix $\mathbf{S}_b = \sum_{j=1}^c n_j(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})'$. Conventional FDA solves the following generalized eigenproblem:

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

where $q \leq \min\{p,\ c-1\}$ and where we refer 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, FDA is equivalent to finding a solution to

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

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

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

Here $\mathbf{A} = [\mathbf{a}_1,\ldots,\mathbf{a}_q]$ $(n\times q)$ and $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_q)$ $(q\times q)$. If $\mathbf{S}_t$ is nonsingular, we obtain

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

Thus, the $(\lambda_j, \mathbf{a}_j)$ are the eigenpairs 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, $\mathbf{S}_t^{-1}\mathbf{S}_b$ is not well defined or cannot be computed accurately.

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$ $(n \times p)$, $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_c]'$ $(c \times p)$, $\Pi = \mathrm{diag}(n_1, \ldots, n_c)$ $(c \times c)$, $\Pi^{\frac{1}{2}} = \mathrm{diag}(\sqrt{n}_1, \ldots, \sqrt{n}_c)$, $\pi = (n_1, \ldots, n_c)'$, $\sqrt{\pi} = (\sqrt{n}_1, \ldots, \sqrt{n}_c)'$ and $\mathbf{H}_\pi = \mathbf{I}_c - \frac{1}{n}\sqrt{\pi}\sqrt{\pi}'$. It then follows that $\mathbf{1}_n'\mathbf{E} = \mathbf{1}_c'\Pi = \pi'$, $\mathbf{E}\mathbf{1}_c = \mathbf{1}_n$, $\mathbf{1}_c'\pi = n$, $\mathbf{E}'\mathbf{E} = \Pi$, $\Pi^{-1}\pi = \mathbf{1}_c$, and

$$\mathbf{M} = \Pi^{-1}\mathbf{E}'\mathbf{X}.$$

In addition, we have

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

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

Based on these results and the idempotency of $\mathbf{H}$, $\mathbf{S}_t$ can be written as

$$\mathbf{S}_t = \mathbf{X}'\mathbf{H}\mathbf{H}\mathbf{X} = \mathbf{X}'\mathbf{H}\mathbf{X}, \tag{3}$$

and we have

$$\begin{aligned}
\mathbf{S}_b &= \mathbf{M}'\left[\Pi - \frac{1}{n}\pi\pi'\right]\mathbf{M} \\
&= \mathbf{M}'\left[\Pi^{\frac{1}{2}} - \frac{1}{n}\pi\sqrt{\pi}'\right]\left[\Pi^{\frac{1}{2}} - \frac{1}{n}\sqrt{\pi}\pi'\right]\mathbf{M} \\
&= \mathbf{X}'\mathbf{E}\Pi^{-1}\Pi^{\frac{1}{2}}\mathbf{H}_\pi\mathbf{H}_\pi\Pi^{\frac{1}{2}}\Pi^{-1}\mathbf{E}'\mathbf{X} \\
&= \mathbf{X}'\mathbf{H}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{H}\mathbf{X}.
\end{aligned} \tag{4}$$

Given these representations of $\mathbf{S}_t$ and $\mathbf{S}_b$, the problem in (2) can be solved by using the GSVD method (Van Loan, 1976; Paige and Saunders, 1981; Golub and Van Loan, 1996; Howland et al., 2003).

There are also two variants of conventional FDA in the literature that aim to handle the ill-conditioned problem (Webb, 2002). The first variant, the *pseudoinverse method*, involves replacing $\mathbf{S}_t^{-1}$ by $\mathbf{S}_t^+$ and solving the following eigenproblem:

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

Note that $\mathbf{S}_t^+$ exists and is unique (Golub and Van Loan, 1996). 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) (Friedman, 1989). It replaces $\mathbf{S}_t$ by $\mathbf{S}_t + \sigma^2\mathbf{I}_p$ and solves the following eigenproblem:

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

It is a well known result that FDA is equivalent to a least mean squared error procedure in the binary classification problem ($c = 2$) (Duda et al., 2001). Recently, similar relationships have been studied for multi-class ($c > 2$) problems (Hastie et al., 2001; Park and Park, 2005b; Ye, 2007). Moreover, Park and Park (2005b) proposed an efficient algorithm for FDA based on 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 $\mathbf{A}\mathbf{D}$ where $\mathbf{D}$ is an arbitrary $q \times q$ nonsingular diagonal matrix. Thus, the 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 a ridge regression problem in the multi-class setting.

2.3 Kernel Discriminant Analysis

Kernel methods (Shawe-Taylor and Cristianini, 2004) work in a feature space $\mathcal{F}$, which is related to the original input space $\mathcal{X} \subset \mathbb{R}^p$ by a mapping,

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

That is, φ is a vector-valued function which gives a vector $\varphi(\mathbf{s})$, 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 $K(\mathbf{s},\mathbf{t}) = \varphi(\mathbf{s})'\varphi(\mathbf{t})$ for $\mathbf{s},\mathbf{t} \in \mathcal{X}$. The mapping $\varphi(\cdot)$ 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, $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1,\ldots,\tilde{\mathbf{x}}_n]'$ ($n\times g$) and $\tilde{\mathbf{M}} = [\tilde{\mathbf{m}}_1,\ldots,\tilde{\mathbf{m}}_c]'$ ($c\times g$) are the data and mean matrices in the feature space. Here g is the dimension of the feature space. Although g is possibly infinite, we here assume that it is finite but not necessarily known.

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

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

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

$$\begin{aligned}
\tilde{\mathbf{S}}_t &= \sum_{i=1}^{n}(\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}})' = \tilde{\mathbf{X}}'\mathbf{H}\tilde{\mathbf{X}},\\
\tilde{\mathbf{S}}_b &= \sum_{j=1}^{c} n_j(\tilde{\mathbf{m}}_j - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_j - \tilde{\mathbf{m}})' = \tilde{\mathbf{X}}'\mathbf{H}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{H}\tilde{\mathbf{X}}.
\end{aligned}$$

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 that $\tilde{\mathbf{A}}$ can be expressed as

$$\tilde{\mathbf{A}} = \sum_{i=1}^{n}(\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}})\beta_i' + \mathbf{N} = \tilde{\mathbf{X}}'\mathbf{H}\Upsilon + \mathbf{N}, \tag{8}$$

where $\Upsilon = [\beta_1,\ldots,\beta_n]$ ($n\times q$) and $\mathbf{N} \in \mathbb{R}^{g\times q}$ such that $\mathbf{N}'\tilde{\mathbf{X}}'\mathbf{H} = \mathbf{0}$ (Park and Park, 2005a; Mika et al., 2000). It then follows from (7) that

$$\tilde{\mathbf{X}}'\mathbf{H}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{H}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H}\Upsilon = \tilde{\mathbf{X}}'\mathbf{H}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H}\Upsilon\Lambda. \tag{9}$$

This implies that $(\Lambda, \tilde{\mathbf{X}}'\mathbf{H}\Upsilon)$ are also the q eigenpairs of the matrix pencil $(\tilde{\mathbf{S}}_b, \tilde{\mathbf{S}}_t)$. Thus, in the literature the solution of (7) is typically restricted to $\mathcal{R}(\tilde{\mathbf{X}}'\mathbf{H})$; that is, $\mathbf{N} = \mathbf{0}$ is set.

Premultiplying both sides of the Equation (9) by $\mathbf{H}\tilde{\mathbf{X}}$, we have a new generalized eigenvalue problem

$$\mathbf{C}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = \mathbf{C}\mathbf{C}\Upsilon\Lambda, \tag{10}$$

which involves only the kernel matrix $\mathbf{K} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'$ via its centered form $\mathbf{C} = \mathbf{H}\mathbf{K}\mathbf{H}$.

The current concern then becomes that of solving the problem (10). Although $\mathbf{K}$ can be assumed to be nonsingular, $\mathbf{C}$ is positive semidefinite but not positive definite because the centering matrix

$\mathbf{H}$ is singular. In fact, the rank of $\mathbf{C}$ is not larger than $n-1$ because the rank of $\mathbf{H}$ is $n-1$. Thus the GDA method devised by Baudat and Anouar (2000) cannot be used directly for problem (10).

To address this problem Park and Park (2005a) proposed a GSVD-based algorithm to solve (10). Running this algorithm requires the complete orthogonal decomposition (Golub and Van Loan, 1996) of matrix $[\mathbf{CE}\Pi^{-\frac{1}{2}}, \mathbf{C}]'$, which is of size $(n+c)\times n$. This approach is infeasible for large values of n. Thus, Park and Park (2005a) developed an efficient alternative which consists of two SVD procedures but does not involve the complete orthogonal decomposition of an $(n+c)\times n$ matrix. We refer to it as the *SVD-based algorithm*.

Another approach to solving the problem (10) is based on the following regularized version of the problem:

$$\mathbf{CE}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = (\mathbf{CC} + \sigma^2\mathbf{I}_n)\Upsilon\Lambda, \tag{11}$$

which was also studied by Park and Park (2005a). Note that this variant is not a directly regularized form of the original KDA problem in (7).

After having obtained Υ from (10) or (11), for a new input vector $\mathbf{x}$, the projection $\mathbf{z}$ of its feature vector $\tilde{\mathbf{x}}$ onto $\tilde{\mathbf{A}}$ is computed by

$$\mathbf{z} = \Upsilon'\mathbf{H}\tilde{\mathbf{X}}\Big(\tilde{\mathbf{x}} - \frac{1}{n}\tilde{\mathbf{X}}'\mathbf{1}_n\Big) = \Upsilon'\mathbf{H}\Big(\mathbf{k}_\mathbf{x} - \frac{1}{n}\mathbf{K}\mathbf{1}_n\Big), \tag{12}$$

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, and this approach has been widely deployed in practical problems. However, a theoretical justification for using the projection in (12) has been lacking in the literature. We are able to provide such as justification as follows. Recall that if $(\Lambda, \tilde{\mathbf{X}}'\mathbf{H}\Upsilon)$ is the solution of (7), then (Λ,Υ) is the solution of (10). As we will see in Theorem 3, if (Λ,Υ) is the solution of (10), $(\Lambda,\tilde{\mathbf{X}}'\mathbf{H}\Upsilon)$ is the solution of (7). This justifies the projection (12).

Note, however, that if (Λ,Υ) is the solution of (11) this does not imply that $(\Lambda,\tilde{\mathbf{X}}'\mathbf{H}\Upsilon)$ is the solution of (7). This shows that the projection (12) is inconsistent with (11). This is an inconsistency that underlies the regularized KDA methodology of Park and Park (2005a). The new methodology that we propose in the following section surmounts this problem.

3. New Approaches to Kernel Discriminant Analysis

Recall that we are interested in the pseudoinverse and regularization forms of (7), defined respectively by

$$\tilde{\mathbf{S}}_t^+\tilde{\mathbf{S}}_b\tilde{\mathbf{A}} = \tilde{\mathbf{A}}\Lambda \tag{13}$$

and

$$\tilde{\mathbf{S}}_b\tilde{\mathbf{A}} = (\tilde{\mathbf{S}}_t + \sigma^2\mathbf{I}_g)\tilde{\mathbf{A}}\Lambda. \tag{14}$$

We wish to find solutions of (7) and (13) or (14) that are consistent with each other.

Since $\mathcal{R}(\tilde{\mathbf{S}}_t) = \mathcal{R}(\tilde{\mathbf{X}}'\mathbf{H}\mathbf{H}\tilde{\mathbf{X}}) = \mathcal{R}(\tilde{\mathbf{X}}'\mathbf{H})$ and $\mathcal{R}(\tilde{\mathbf{S}}_b) = \mathcal{R}(\tilde{\mathbf{X}}'\mathbf{H}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{H}\tilde{\mathbf{X}}) = \mathcal{R}(\tilde{\mathbf{X}}'\mathbf{H}\mathbf{E})$, we have that $\mathcal{R}(\tilde{\mathbf{S}}_b) \subseteq \mathcal{R}(\tilde{\mathbf{S}}_t)$. As a direct corollary of Theorem 1, we thus obtain a connection between (7) and (13); that is,

Theorem 2 *If $(\Lambda,\tilde{\mathbf{A}})$ (nonzero eigenpairs) is the solution of (13), then $(\Lambda,\tilde{\mathbf{A}})$ is the solution of (7). Conversely, if $(\Lambda,\tilde{\mathbf{A}})$ is the solution of (7), then $(\Lambda,\tilde{\mathbf{S}}_t^+\tilde{\mathbf{S}}_t\tilde{\mathbf{A}})$ is the solution of (13).*

Theorem 2 implies that the solution of (7) can be obtained from (13). To solve (13), we consider the pseudoinverse form of (10), which is

$$\mathbf{C}^+\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = \Upsilon\Lambda \tag{15}$$

due to $\mathbf{C}^+\mathbf{C}^+\mathbf{C} = \mathbf{C}^+$ (see Lemma 11). To obtain the solution of (14), we substitute (8) into (14) and then premultiply by $\mathbf{H}\tilde{\mathbf{X}}$. As a result, we have

$$\mathbf{C}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = (\mathbf{C}\mathbf{C}+\sigma^2\mathbf{C})\Upsilon\Lambda. \tag{16}$$

The following theorem shows that we are able to obtain the solutions of (7), (13) and (14) respectively from the solutions of (10), (15) and (16). That is,

Theorem 3 *Considering the KDA problems, we have:*

(i) *If* (Λ,Υ) *is the solution of (10), then* $\left(\Lambda,\tilde{\mathbf{X}}'\mathbf{H}\Upsilon\right)$ *is the solution of (7).*

(ii) *If* (Λ,Υ) *is the solution of (15), then* $\left(\Lambda,\tilde{\mathbf{X}}'\mathbf{H}\Upsilon\right)$ *is the solution of (13).*

(iii) *If* (Λ,Υ) *is the solution of (16), then* $\left(\Lambda,\tilde{\mathbf{X}}'\mathbf{H}\Upsilon\right)$ *is the solution of (14).*

The proof of this theorem is given in Appendix C. Theorem 3 shows that the solutions of (7), (13) and (14) lie in span$\{\tilde{\mathbf{X}}'\mathbf{H}\}$. Moreover, we see that $\left(\Lambda,\tilde{\mathbf{X}}'\mathbf{H}\Upsilon\right)$ are their solutions. We also note that (16) is different from (11). Theorem 3 provides a relationship between (14) and (16); there is not a similar relationship between (14) and (11).

Finally, as a corollary of Theorems 2 and 3, we have

Corollary 4 *If* (Λ,Υ) *is the solution of (15), then* $\left(\Lambda,\tilde{\mathbf{X}}'\mathbf{H}\Upsilon\right)$ *is the solution of (7). Moreover, if* (Λ,Υ) *is the solution of (10), then* $\left(\Lambda,\tilde{\mathbf{X}}'\mathbf{H}\Upsilon\right)$ *is the solution of (13).*

We see from Corollary 4 that the solution of (7) can be also obtained from (15). We now concentrate our attention on the regularized KDA problem (14). We first handle (14) by using (16) and Theorem 3, and then we present an approach for directly solving (14).

4. SVD-based Algorithms for RDA Problems

It is clear that we can solve the regularized KDA (RKDA) problem in (11) by solving

$$(\mathbf{C}\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{C}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = \Upsilon\Lambda.$$

However, since $\mathbf{C}\mathbf{C}+\sigma^2\mathbf{C}$ is singular, this approach is not appropriate for the RKDA problem in (16). In this section we show how to solve the RKDA problems in both (11) and (16), as well as the regularized FDA (RFDA) problem in (6).

In particular, we exploit the SVD-based algorithm developed for solving (10) by Park and Park (2005a). Our algorithms are summarized as Algorithm 1, which is an SVD-based method for solving RFDA problem (6), and Algorithms 2 and 3, which are SVD-based algorithms for RKDA problems (11) and (16). We defer the derivations to Section 8 in which we consider the SVD-based algorithm for a more general generalized eigenvalue problem. According to Theorem 3 and the following theorem, we immediately have the solution of (14) as $\tilde{\mathbf{A}} = \tilde{\mathbf{X}}'\mathbf{H}\Upsilon$.

Algorithm 1 SVD-based Algorithm for RFDA problem (6)

procedure RFDA($\mathbf{X}, \mathbf{E}, \pi, \Pi, \sigma^2$)
  Perform the condensed SVD of $\mathbf{HX}$ as $\mathbf{HX} = \mathbf{U}_X \Gamma_X \mathbf{V}_X'$;
  Calculate $\mathbf{F} = (\Gamma_X^2 + \sigma^2 \mathbf{I}_r)^{-\frac{1}{2}} \Gamma_X \mathbf{U}_X' \mathbf{E} \Pi^{-\frac{1}{2}}$ where $r = \mathrm{rk}(\Gamma_X)$;
  Perform the condensed SVD of $\mathbf{F}$ as $\mathbf{F} = \mathbf{U}_F \Gamma_F \mathbf{V}_F'$ and set $q = \mathrm{rk}(\Gamma_F)$;
  Return $\mathbf{A} = \mathbf{V}_X (\Gamma_X^2 + \sigma^2 \mathbf{I}_r)^{-\frac{1}{2}} \mathbf{U}_F$ as the solution of RFDA.
end procedure

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

procedure RKDA($\mathbf{C}, \mathbf{E}, \pi, \Pi, \sigma^2$)
  Perform the condensed SVD of $\mathbf{C}$ as $\mathbf{C} = \mathbf{U}_C \Gamma_C \mathbf{U}_C'$;
  Calculate $\mathbf{F} = (\Gamma_C^2 + \sigma^2 \mathbf{I}_r)^{-\frac{1}{2}} \Gamma_C \mathbf{U}_C' \mathbf{E} \Pi^{-\frac{1}{2}}$ where $r = \mathrm{rk}(\Gamma_C)$;
  Perform the condensed SVD of $\mathbf{F}$ as $\mathbf{F} = \mathbf{U}_F \Gamma_F \mathbf{V}_F'$;
  Let $\Upsilon = \mathbf{U}_C (\Gamma_C^2 + \sigma^2 \mathbf{I}_r)^{-\frac{1}{2}} \mathbf{U}_F$ and set $q = \mathrm{rk}(\Gamma_F)$;
  Calculate $\mathbf{z}$ via (12) as the q-dimensional representation of $\mathbf{x}$.
end procedure

Theorem 5 *Consider Algorithms 1, 2 and 3 for the corresponding RDA problems.*

(i) *If* $\mathbf{A}$ *is obtained from Algorithm 1, then,*

$$\mathbf{A}'(\mathbf{S}_t + \sigma^2 \mathbf{I}_p)\mathbf{A} = \mathbf{I}_q \quad \textit{and} \quad \mathbf{A}'\mathbf{S}_b\mathbf{A} = \Gamma_F^2.$$

(ii) *If* Υ *is obtained from Algorithm 2, then*

$$\Upsilon'(\mathbf{CC} + \sigma^2 \mathbf{I}_n)\Upsilon = \mathbf{I}_q \quad \textit{and} \quad \Upsilon'\mathbf{CE}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = \Gamma_F^2.$$

(iii) *If* Υ *is obtained from Algorithm 3 and* $\tilde{\mathbf{A}} = \tilde{\mathbf{X}}'\mathbf{H}\Upsilon$*, then*

$$\Upsilon'(\mathbf{CC} + \sigma^2 \mathbf{C})\Upsilon = \mathbf{I}_q \quad \textit{and} \quad \Upsilon'\mathbf{CE}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = \Gamma_F^2,$$

and

$$\tilde{\mathbf{A}}'(\tilde{\mathbf{S}}_t + \sigma^2 \mathbf{I}_g)\tilde{\mathbf{A}} = \mathbf{I}_q \quad \textit{and} \quad \tilde{\mathbf{A}}'\tilde{\mathbf{S}}_b\tilde{\mathbf{A}} = \Gamma_F^2.$$

The proof of the theorem is given in Appendix D. According to this theorem, for a new $\mathbf{x}$, the projection $\mathbf{z}$ onto $\tilde{\mathbf{A}}$ is given by

$$\mathbf{z} = \Upsilon'\mathbf{H}\Big(\mathbf{k}_\mathbf{x} - \frac{1}{n}\mathbf{K}\mathbf{1}_n\Big). \tag{17}$$

Note that if $\sigma^2 = 0$, Algorithm 1 degenerates to the SVD-based algorithm for the conventional FDA in (2) (Howland et al., 2003) and Algorithms 2 and 3 become identical.

5. EVD-based Algorithms for RDA

It is desirable to directly find the solution of the RKDA problem (14), rather than obtaining the solution indirectly via (16). However, it is not feasible to devise a SVD-based algorithm for directly solving the RKDA problem in (14). In this section we propose a new approach to solving the RFDA problem (6). We then extend this approach for the solution of the RKDA problem (14).

Algorithm 3 SVD-based Algorithm for RKDA problem (16) as well as for (14)

procedure RKDA($\mathbf{C}, \mathbf{E}, \pi, \Pi, \sigma^2$)
  Perform the condensed SVD of $\mathbf{C}$ as $\mathbf{C} = \mathbf{U}_C \Gamma_C \mathbf{U}_C'$;
  Calculate $\mathbf{F} = (\Gamma_C^2 + \sigma^2 \Gamma_C)^{-\frac{1}{2}} \Gamma_C \mathbf{U}_C' \mathbf{E} \Pi^{-\frac{1}{2}}$ where $r = \mathrm{rk}(\Gamma_C)$;
  Perform the condensed SVD of $\mathbf{F}$ as $\mathbf{F} = \mathbf{U}_F \Gamma_F \mathbf{V}_F'$;
  Let $\Upsilon = \mathbf{U}_C (\Gamma_C^2 + \sigma^2 \Gamma_C)^{-\frac{1}{2}} \mathbf{U}_F$ and set $q = \mathrm{rk}(\Gamma_F)$;
  Calculate $\mathbf{z}$ via (17) as the q-dimensional representation of $\mathbf{x}$.
end procedure

5.1 The Algorithm for RFDA

We reformulate the eigenproblem in (6) as

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

where

$$\mathbf{G} = (\mathbf{X}' \mathbf{H} \mathbf{X} + \sigma^2 \mathbf{I}_p)^{-1} \mathbf{X}' \mathbf{H} \mathbf{E} \Pi^{-\frac{1}{2}} \tag{18}$$

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

$$\mathbf{G} = \mathbf{X}' \mathbf{H} (\mathbf{H} \mathbf{X} \mathbf{X}' \mathbf{H} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{E} \Pi^{-\frac{1}{2}} \tag{19}$$

due to $(\mathbf{X}'\mathbf{H}\mathbf{X} + \sigma^2 \mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{H} = \mathbf{X}'\mathbf{H}(\mathbf{H}\mathbf{X}\mathbf{X}'\mathbf{H} + \sigma^2 \mathbf{I}_n)^{-1}$. If $n < p$, we can use (19) to reduce the computational cost. Moreover, we will see that (19) plays a key role in the development of an efficient algorithm for KDA to be presented shortly.

Let $\mathbf{R} = \Pi^{-\frac{1}{2}} \mathbf{E}' \mathbf{H} \mathbf{X} \mathbf{G}$. Since $\mathbf{G} \Pi^{-\frac{1}{2}} \mathbf{E}' \mathbf{H} \mathbf{X}$ ($p \times p$) and $\mathbf{R}$ ($c \times c$) have the same nonzero eigenvalues (Horn and Johnson, 1985), the λ_j, $j = 1, \ldots, q$, are the nonzero eigenvalues of $\mathbf{R}$. Moreover, if $(\Lambda, \mathbf{V})$ is the nonzero eigenpair of $\mathbf{R}$, $(\Lambda, \mathbf{G}\mathbf{V})$ is the nonzero eigenpair of $\mathbf{G} \Pi^{-\frac{1}{2}} \mathbf{E}' \mathbf{H} \mathbf{X}$. Note that

$$\mathbf{R} = \Pi^{-\frac{1}{2}} \mathbf{E}' \mathbf{H} \mathbf{X} (\mathbf{X}' \mathbf{H} \mathbf{X} + \sigma^2 \mathbf{I}_p)^{-1} \mathbf{X}' \mathbf{H} \mathbf{E} \Pi^{-\frac{1}{2}}. \tag{20}$$

This shows that $\mathbf{R}$ is positive semidefinite.

We use these facts to develop an algorithm for solving the RFDA problem in (6). This is also a two-step process, which is presented in Algorithm 4. The first step computes $(\sigma^2 \mathbf{I}_p + \mathbf{X}'\mathbf{H}\mathbf{X})^{-1}$ (or $(\sigma^2 \mathbf{I}_n + \mathbf{H}\mathbf{X}\mathbf{X}'\mathbf{H})^{-1}$), while the second step is an SVD procedure. Note that the first step can be implemented by computing the condensed SVD of $\mathbf{X}'\mathbf{H}\mathbf{X}$ (or $\mathbf{H}\mathbf{X}\mathbf{X}'\mathbf{H}$). Since $\mathbf{R}$ and $\mathbf{X}'\mathbf{H}\mathbf{X}$ ($\mathbf{H}\mathbf{X}\mathbf{X}'\mathbf{H}$) are positive semidefinite, their SVD are equivalent to the eigenvalue decomposition (EVD). Thus, we refer to this two-step process as an *EVD-based algorithm*, distinguishing it from the SVD-based algorithm.

The first step calculates $\mathbf{G}$ by either (18) or (19). The computational complexity is $O(m^3)$ where $m = \min(n, p)$. The second step forms the condensed SVD of $\mathbf{R}$ for which the computational complexity is $O(c^3)$. If both n and p are large, we recommend an approximate strategy; that is, we first perform the incomplete Cholesky decomposition of $\mathbf{H}\mathbf{X}\mathbf{X}'\mathbf{H}$ (or $\mathbf{X}'\mathbf{H}\mathbf{X}$) and then calculate $(\mathbf{H}\mathbf{X}\mathbf{X}'\mathbf{H} + \sigma^2 \mathbf{I}_n)^{-1}$ (or $(\mathbf{X}'\mathbf{H}\mathbf{X} + \sigma^2 \mathbf{I}_p)^{-1}$) via the Sherman-Morrison-Woodbury formula (Golub and Van Loan, 1996). This strategy makes the first step still efficient.

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

$$\begin{aligned}\mathbf{G} &= (\mathbf{X}'\mathbf{H}\mathbf{X})^{+}\mathbf{X}'\mathbf{H}\mathbf{E}\Pi^{-\frac{1}{2}} \\ &\overset{(or)}{=} \mathbf{X}'\mathbf{H}(\mathbf{H}\mathbf{X}\mathbf{X}'\mathbf{H})^{+}\mathbf{E}\Pi^{-\frac{1}{2}}. \end{aligned} \tag{21}$$

Algorithm 4 EVD-based Algorithm for RFDA problem (6)

1: **procedure** RFDA($\mathbf{X}, \mathbf{E}, \Pi, \sigma^2$)
2: Calculate $\mathbf{G}$ by (18) or (19) and $\mathbf{R}$ by (20);
3: Perform the condensed SVD of $\mathbf{R}$ as $\mathbf{R} = \mathbf{V}_R\Gamma_R\mathbf{V}_R'$;
4: Return $\mathbf{A} = \mathbf{G}\mathbf{V}_R\Gamma_R^{-\frac{1}{2}}$ or $\mathbf{B} = \mathbf{G}\mathbf{V}_R$ as the solution of RFDA problem (6).
5: **end procedure**

Compared with the SVD-based algorithm, the EVD-based algorithm is more efficient, especially for "small n but large p" problems. Using the notation in Algorithms 1 and 4, we have

$$\mathbf{R} = \mathbf{F}'\mathbf{F}$$

by performing some matrix computations. This implies that $\Gamma_R = \Gamma_F^2$. Moreover, it is immediate to obtain the following result.

Theorem 6 *Let* $\mathbf{A}$ *be obtained from Algorithm 4. Then,*

$$\mathbf{A}'(\mathbf{S}_t + \sigma^2\mathbf{I}_p)\mathbf{A} = \mathbf{I}_q \quad \textit{and} \quad \mathbf{A}'\mathbf{S}_b\mathbf{A} = \Gamma_F^2.$$

This theorem shows that Algorithms 1 and 4 are essentially equivalent. As mentioned before, it is not feasible to develop an SVD-based algorithm for solving the RKDA problem (14), which is the kernel extension of RFDA in (6). On the other hand, in the next subsection we will see that Algorithm 4 can be used for solving the RKDA problem (14).

5.2 The Algorithm for RKDA

It follows immediately from (19) that

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

from which, using (20), we calculate $\mathbf{R}$ by

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

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

$$\begin{aligned}\mathbf{z} &= \tilde{\mathbf{A}}'(\tilde{\mathbf{x}} - \tilde{\mathbf{m}}) \\ &= \Gamma_R^{-\frac{1}{2}}\mathbf{V}_R'\Pi^{-\frac{1}{2}}\mathbf{E}'(\mathbf{C} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{H}\tilde{\mathbf{X}}\Big(\tilde{\mathbf{x}} - \frac{1}{n}\tilde{\mathbf{X}}'\mathbf{1}_n\Big) \\ &= \Gamma_R^{-\frac{1}{2}}\mathbf{V}_R'\Pi^{-\frac{1}{2}}\mathbf{E}'(\mathbf{C} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{H}\Big(\mathbf{k}_{\mathbf{x}} - \frac{1}{n}\mathbf{K}\mathbf{1}_n\Big). \end{aligned} \tag{22}$$

This shows that we can calculate $\mathbf{R}$ and $\mathbf{z}$ directly using $\mathbf{K}$ and $\mathbf{k_x}$. We thus obtain an EVD-based algorithm for RKDA, which is given in Algorithm 5. Also, when $\sigma^2 = 0$, we calculate $\mathbf{R}$ by

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

and exploit the EVD-based algorithm to solve the following variant of KDA:

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

We see that the EVD-based algorithm is more efficient than the SVD-based algorithm (i.e., Algorithm 2) for the RKDA problem in (11). Recall that the RKDA problem (14) is not fully equivalent to that in (11). Moreover, we also have an EVD-based algorithm for solving (11), by replacing $\mathbf{C}$ by $\mathbf{C}\mathbf{C}$ in calculating $\mathbf{R}$ and (22) by (17) in calculating $\mathbf{z}$. However, the resulting algorithm is less efficient computationally.

Algorithm 5 EVD-based Algorithm for RKDA problem (14)

1: **procedure** RKDA($\mathbf{K}, \mathbf{E}, \mathbf{k_x}, \Pi, \sigma^2$)
2: Calculate $\mathbf{R} = \Pi^{-\frac{1}{2}}\mathbf{E}'\mathbf{C}(\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{E}\Pi^{-\frac{1}{2}}$;
3: Perform the condensed SVD of $\mathbf{R}$ as $\mathbf{R} = \mathbf{V}_R\Gamma_R\mathbf{V}_R'$;
4: Calculate $\mathbf{z}$ by (22);
5: Return $\mathbf{z}$ as the q-dimensional representation of $\mathbf{x}$.
6: **end procedure**

Let us investigate the relationship between the solutions of (14) from Algorithms 3 and 5. First, let Υ be obtained from Algorithm 3. It follows from (16) that $(\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{C}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = \mathbf{C}\Upsilon\Lambda$, that is,

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

Thus, $(\Lambda, \Pi^{-\frac{1}{2}}\mathbf{E}'\mathbf{C}\Upsilon\Gamma_F^{-1})$ is the nonzero eigenpair of $\mathbf{R}$. Finally, we have $\Lambda = \Gamma_F^2 = \Gamma_R$. In addition, it follows from Theorem 5 that

$$\Gamma_F^{-1}\Upsilon'\mathbf{C}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon\Gamma_F^{-1} = \mathbf{I}_q.$$

Moreover, we have

$$\begin{aligned}\tilde{\mathbf{G}}\Pi^{-\frac{1}{2}}\mathbf{E}'\mathbf{C}\Upsilon\Gamma_R^{-2} &= \tilde{\mathbf{X}}'\mathbf{H}(\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon\Lambda^{-1} \\ &= \tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+\mathbf{C}(\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon\Lambda^{-1} \\ &= \tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+\mathbf{C}\Upsilon = \tilde{\mathbf{X}}'\mathbf{H}\Upsilon \qquad (23)\end{aligned}$$

because $\tilde{\mathbf{X}}'\mathbf{H} = \tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+\mathbf{C}$. This implies that $\tilde{\mathbf{A}} = \tilde{\mathbf{X}}'\mathbf{H}\Upsilon$ obtained from Algorithm 3 is equivalent to that obtained from Algorithm 5.

On the other hand, let $\mathbf{R} = \mathbf{V}_R\Gamma_R\mathbf{V}_R'$ be the condensed SVD of $\mathbf{R}$. Then

$$\mathbf{C}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = (\mathbf{C}^2+\sigma^2\mathbf{C})\Upsilon\Gamma_R,$$

where $\Upsilon = (\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{E}\Pi^{-\frac{1}{2}}\mathbf{V}\Gamma_R^{-\frac{1}{2}}$. Moreover, it is easily checked that

$$\Upsilon'(\mathbf{C}^2+\sigma^2\mathbf{C})\Upsilon = \mathbf{I}_q \quad \text{and} \quad \Upsilon'\mathbf{C}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = \Gamma_R.$$

This implies that (Γ_R, Υ) is the solution of (16). Again, using (23), we conclude that the solution of (14) from Algorithm 5 is equivalent to the one from Algorithm 3.

In summary, Algorithms 5 and 3 yield equivalent solutions for (14). However, Algorithm 5 is more efficient than Algorithm 3.

6. Relationships Between RFDA and Ridge Regression

It is a well known result that FDA (or KDA) is equivalent to a least mean squared error procedure in the binary classification problem ($c = 2$) (Duda et al., 2001; Mika et al., 2000). Recently, relationships between FDA and a least mean squared error procedure in multi-class ($c > 2$) problems have been discussed by Hastie et al. (2001), Park and Park (2005b), and Ye (2007).

Motivated by this line of work, we investigate a possible equivalency between RFDA and ridge regression (Hoerl and Kennard, 1970). We then go on to consider a similar relationship between RKDA and the corresponding ridge regression problem.

Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]' = \mathbf{E}\Pi^{-\frac{1}{2}}\mathbf{H}_\pi$. 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:

$$\min_{\mathbf{w}_0, \mathbf{W}} L(\mathbf{w}_0, \mathbf{W}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{1}_n\mathbf{w}_0' - \mathbf{X}\mathbf{W}\|_F^2 + \frac{\sigma^2}{2}\mathrm{tr}(\mathbf{W}'\mathbf{W}). \tag{24}$$

We focus on the solution for $\mathbf{W}$:

$$\mathbf{W} = (\mathbf{X}'\mathbf{H}\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{M}'\Pi^{\frac{1}{2}}\mathbf{H}_\pi = (\mathbf{X}'\mathbf{H}\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{H}\mathbf{E}\Pi^{-\frac{1}{2}}. \tag{25}$$

The derivation is given in Appendix E. It is then seen from (18) 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:

$$\min_{\mathbf{w}_0, \mathbf{W}} L(\mathbf{w}_0, \mathbf{W}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{1}_n\mathbf{w}_0' - \mathbf{X}\mathbf{W}\|_F^2. \tag{26}$$

In this case, if $\mathbf{X}'\mathbf{H}\mathbf{X}$ is singular, a standard treatment uses $(\mathbf{X}'\mathbf{H}\mathbf{X})^+$ in (25). Such a $\mathbf{W}$ is identical with $\mathbf{G}$ in (21).

In summary, we have obtained a relationship between the ridge estimation problem in (24) and the RFDA problem in (6).

Theorem 7 *Let* $\mathbf{W}$ *be the solution of the ridge estimation problem in (24) (resp. the OLS estimation problem in (26)) and* $\mathbf{A}$ *be defined in Algorithm 4 for the solution of the RFDA problem in (6) (resp. the FDA problem in (5)). Then*

$$\mathbf{A} = \mathbf{W}\mathbf{V}_R\Gamma_R^{-\frac{1}{2}},$$

where $\mathbf{V}_R$ *and* Γ_R *are defined in Algorithm 4.*

This theorem provides an important connection between $\mathbf{A}$ and $\mathbf{W}$. Indeed $\mathbf{B} = \mathbf{A}\Gamma_R^{\frac{1}{2}}$ is also a solution of the RFDA problem (6). However, $\mathbf{B}$ satisfies the condition $\mathbf{B}'(\mathbf{S}_t + \sigma^2\mathbf{I}_p)\mathbf{B} = \Gamma_R$, rather than $\mathbf{B}'(\mathbf{S}_t + \sigma^2\mathbf{I}_p)\mathbf{B} = \mathbf{I}_q$. Thus, with this $\mathbf{B}$, we obtain the following result, which is the principal theoretical result of this paper.

Theorem 8 *Under the conditions in Theorem 7, we have*

$$\mathbf{B}\mathbf{B}' = \mathbf{W}\mathbf{W}'.$$

Moreover, we have

$$(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{B}\mathbf{B}'(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{W}\mathbf{W}'(\mathbf{x}_i - \mathbf{x}_j)$$

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

The proof of this theorem is given in Appendix F. Theorem 8 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 the multi-class FDA and multivariate linear estimators are same. Since Theorem 8 holds in general cases, we obtain a complete solution to the open problem concerning the relationship between multi-class FDA problems and multivariate linear estimators.

Similar results have been obtained by Park and Park (2005b); Ye (2007), but under restrictive conditions which arise from a different definition of the label scoring matrix $\mathbf{Y}$ than ours. The choice that of label scoring matrix that we have presented has also been used by Ye (2007), but Ye (2007) attempted to establish a connection between the solution $\mathbf{W}$ and $\mathbf{A}$ as given in Algorithm 1.

It is also worth noting that Zhang and Dai (2009) discussed a connection between the label scoring matrix $\mathbf{Y}$ and the optimal scoring procedure in Hastie et al. (1994). Moreover, Zhang and Jordan (2008) exploited this label scoring matrix in spectral clustering.

Our theorem also goes through immediately in the kernel setting. In particular, for the RKDA problem defined by (11), the corresponding ridge estimator is

$$\min_{\mathbf{w}_0, \mathbf{\Phi} \in \mathbb{R}^{n \times c}} L(\mathbf{w}_0, \Phi) \triangleq \frac{1}{2}\|\mathbf{Y} - \mathbf{1}_n\mathbf{w}_0' - \mathbf{K}\mathbf{H}\Phi\|_F^2 + \frac{\sigma^2}{2}\mathrm{tr}(\Phi'\Phi). \tag{27}$$

The ridge estimation problem corresponding to our RKDA in (14) is given by

$$\min_{\mathbf{w}_0, \tilde{\mathbf{W}} \in \mathbb{R}^{g \times c}} L(\mathbf{w}_0, \tilde{\mathbf{W}}) \triangleq \frac{1}{2}\|\mathbf{Y} - \mathbf{1}_n\mathbf{w}_0' - \tilde{\mathbf{X}}\tilde{\mathbf{W}}\|_F^2 + \frac{\sigma^2}{2}\mathrm{tr}(\tilde{\mathbf{W}}'\tilde{\mathbf{W}}),$$

while the estimation problem for the RKDA in (16) is

$$\min_{\mathbf{w}_0, \mathbf{\Phi} \in \mathbb{R}^{n \times c}} L(\mathbf{w}_0, \Phi) \triangleq \frac{1}{2}\|\mathbf{Y} - \mathbf{1}_n\mathbf{w}_0' - \mathbf{K}\mathbf{H}\Phi\|_F^2 + \frac{\sigma^2}{2}\mathrm{tr}(\Phi'\mathbf{C}\Phi), \tag{28}$$

which is no longer a conventional ridge regression problem. In fact, this problem can be regarded a multi-class extension of the least squares SVM (LS-SVM) (Suykens and Vandewalle, 1999; Suykens et al., 2002); (see, e.g., Van Gestel et al., 2002; Pelckmans et al., 2005). Our work thus provides the relationship between RKDA and the LS-SVM.

Note that when $\sigma^2 = 0$, the problems in (27) and (28) are identical. Moreover, the solution of the problems is given by

$$\Phi = (\mathbf{H}\mathbf{K}\mathbf{H}\mathbf{K}\mathbf{H})^+\mathbf{H}\mathbf{K}\mathbf{H}\mathbf{E}\Pi^{-\frac{1}{2}} = \mathbf{C}^+\mathbf{E}\Pi^{-\frac{1}{2}}.$$

In this case, the RKDA methods in (11) and (16) are also the same. As we see from Section 3, its corresponding pseudoinverse form is given by (15). Let the condensed SVD of $\mathbf{R} = \Pi^{-\frac{1}{2}}\mathbf{E}'\mathbf{C}\mathbf{C}^{+}\mathbf{E}\Pi^{-\frac{1}{2}}$ be $\mathbf{R} = \mathbf{V}_R\Gamma_R\mathbf{V}_R'$. Then $(\Gamma_R, \Phi\mathbf{V}_R\Gamma_R^{-\frac{1}{2}})$ is the solution of (15). This implies that in the case of $\sigma^2 = 0$, there is still the connection between the least squares kernel-based SVM and RKDA shown in Theorem 7.

7. Experimental Study

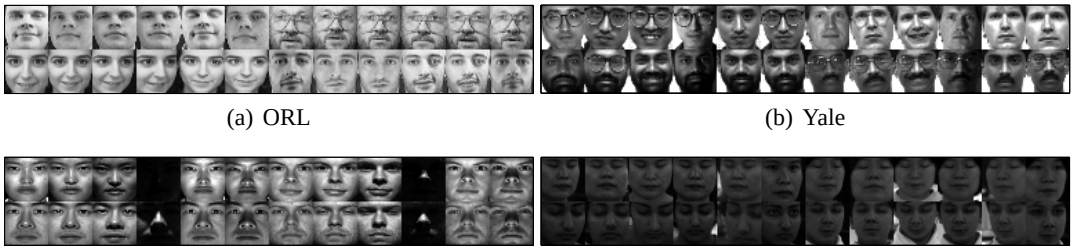
To evaluate the performance of the proposed algorithms for FDA and KDA, we conducted experimental comparisons with other closely related algorithms for FDA and KDA on several real-world data sets. In particular, the comparison was implemented on four face data sets, two handwritten digits data sets, the "letters" data set, and the WebKB data set. All algorithms were implemented in Matlab on a PC configured with an Intel Dual Core 2.0GHz CPU and 2.06GB of memory.

7.1 Setup

The four face data sets are the ORL face database, the Yale face database, the Yale face database B with extension, and the CMU PIE face database, respectively.

- The ORL face database contains 400 facial images of 40 subjects with 10 different images for each subject. This database was developed at the Olivetti Research Laboratories in Cambridge, U.K. The images were taken at different times with variations in facial details (glasses/no glasses), facial expressions (open/closed eyes, smiling/nonsimling), and facial poses (tilted and rotated up to 20 degrees). There is also variation in the scale of up to about 10%. The spatial resolution of the images is 112×92, with 256 gray levels.
- The Yale face images for each subject were captured under different facial expressions or configurations (e.g., center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink).
- The Yale face database with extension includes the Yale face database B (Georghiades et al., 2001) and the extended Yale Face Database B (Lee et al., 2005). The Yale face database B contains 5760 face images of 10 subjects with 576 different images for each subject, and the extended Yale Face Database B contains 16128 face images of 28 subjects, with each subject having 576 different images. The facial images for each subject were captured under 9 poses and 64 illumination conditions. For the sake of simplicity, a subset called the YaleB&E was collected from two databases; it contains the 2414 face images of 38 subjects.
- The CMU PIE face database contains 41,368 face images of 68 subjects. The facial images for each subject were captured under 13 different poses, 43 different illumination conditions, and with 4 different expressions. In our experiments, we considered only the five near-frontal poses under different illuminations and expressions. For simplicity, we collected a subset of the PIE face database, containing the 6800 face images of 68 subjects with 100 different images for each subject.

In all of the experiments, each whole image was cropped and further resized to have a spatial resolution of 32×32 with 256 gray levels. Figure 1 shows some samples from the four data sets, where four subjects are randomly chosen from each data set and each subject has six sample images.



(a) ORL (b) Yale

(c) YaleB&E (d) PIE

Figure 1: Some sample images randomly chosen from the four data sets, where four subjects from each data set and each subject with six sample images.

The two handwritten digits data sets are the USPS data set and the Binary Alphadigits (BA) data set, respectively.

- The USPS data set was derived from the well-known United States Postal Service (USPS) set of handwritten digits, and contains 2000 images of 10 digits, each digit with 200 images. The spatial resolution of the images in the USPS data set is 16×16, with 256 gray levels.

- The Binary Alphadigits (BA) data set was collected from a binary 20×16 digits database of "0" through "9" and capital "A" through "Z," and thus contains 1404 images of 36 subjects, each subject with 39 image.

The "letters" data set can be obtained from Statlog(`http://www.liacc.up.pt/ML/`) and it consists of images of the letters "A," "B," "C," "D" and "E" with 789, 766, 736, 805 and 768 cases respectively.

Finally, the WebKB data set contains web pages gathered from computer science departments in several universities (Craven et al., 1998). The pages can be divided into seven categories. In our experiments, we used the four most populous categories, namely, student, faculty, course, and project, resulting in a total of 4192 pages. Based on information gain, 300 features were selected.

Table 1 summarizes these benchmark data sets. In our experiments, each data set was randomly partitioned into two disjoint subsets as the 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.

The hyperparameters involved in the following methods were selected by cross-validation. After having obtained the q-dimensional representations $\mathbf{z}_i$ of the $\mathbf{x}_i$ from each method, we used a simple nearest neighbor classifier to evaluate the classification accuracy.

7.2 Comparison of FDA Methods

In the linear setting, we compared Algorithm 4 with Algorithm 1 (RFDA/SVD-based), the FDA/GSVD (Howland and Park, 2004) and FDA/MSE methods. Here the FDA/MSE method was derived by Park and Park (2005b) from the relationship between FDA and the minimum squared error solution. We refer

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%
USPS	10	256	2000	10%
BA	36	320	1404	50%
Letters	5	16	3864	10%
WebKB	4	300	4192	10%

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

to Algorithm 4 working with **A** as the RFDA/EVD-based method. As we have shown, when **B** is used, Algorithm 4 provides the solution of ridge regression. We thus refer to the algorithm working with **B** as RFDA/RR. Similar notation also applies to the kernel setting in the next subsection.

Empirically, the performance of the RFDA/EVD-based method is fully identical to that of the RFDA/SVD-based method. This bears out the theoretical analysis in Theorem 6. Thus, we only report the classification accuracies of the RFDA/RR method for Algorithm 4.

Table 2 presents an overall comparison of the methods on all of the data sets and Figure 2 presents the comparative classification results on the four face data sets. It is seen that the RFDA methods have better classification accuracy overall than other methods throughout a range of choices of the number of discriminant variates. Particularly striking is the performance of the RFDA methods when the number of discriminant variates q is small.

From Figure 2, we see that the performance of RFDA/RR method is a little better than that of RFDA/SVD-based method. This implies that RFDA/RR outperforms RFDA/EVD-based method; that is, the performance using the transformation matrix **B** is better than that using the transformation matrix **A** in Algorithm 4. Therefore, the constraint $\mathbf{A}'(\mathbf{S}_t + \sigma^2\mathbf{I}_p)\mathbf{A} = \mathbf{I}_q$ is not necessarily the best choice for RFDA. This also shows that the ridge regression method given in Section 6 is effective and efficient.

We also compared the computational time of the different methods on the four face data sets. Figure 3 plots the results as a function of the training percentage n/k on the four face data sets. We see that our method has an overall favorable computational complexity in comparison with the other methods on the four face data sets. As the training percentage n/k increases, our method yields more efficient performance.

Note that when the training percentage n/k on the YaleB&E and PIE data sets increases, the singularity problem of the within-class scatter matrix $\mathbf{S}_w$, that is, the small sample size problem, tends to disappear. Figures 3 (c) and (d) show that the FDA/MSE method becomes more efficient in this case, and the corresponding computational time becomes flat with respect to the increase of the training percentage n/k. On the other hand, Figures 3 (c) and (d) also reveal that the computational time of the FDA/SVD-based method significantly increases as the size of training data increases.

Data Set	FDA/GSVD		FDA/MSE		RFDA/SVD-based		RFDA/RR	
	acc ($\pm std$)	*time*	*acc* ($\pm std$)	*time*	*acc* ($\pm std$)	*time*	*acc* ($\pm std$)	*time*
ORL	91.54 ($\pm$1.98)	1.952	91.58 ($\pm$2.00)	0.293	93.17 ($\pm$1.94)	0.347	94.04 ($\pm$1.95)	0.079
Yale	78.56 ($\pm$2.29)	1.281	78.44 ($\pm$2.47)	0.047	79.22 ($\pm$4.19)	0.072	79.56 ($\pm$3.75)	0.014
YaleB&E	59.54 ($\pm$11.8)	43.18	65.34 ($\pm$9.23)	9.967	89.86 ($\pm$1.15)	9.177	90.20 ($\pm$1.09)	1.479
PIE	77.26 ($\pm$1.05)	89.85	77.26 ($\pm$1.05)	23.10	90.40 ($\pm$0.65)	83.88	91.14 ($\pm$0.63)	2.726
USPS	43.02 ($\pm$1.86)	0.392	42.95 ($\pm$1.82)	0.229	82.16 ($\pm$1.07)	0.273	83.49 ($\pm$1.39)	0.035
Letters	91.23 ($\pm$0.98)	0.017	91.23 ($\pm$0.98)	0.011	91.68 ($\pm$0.89)	0.013	91.89 ($\pm$0.65)	0.021
WebKB	67.45 ($\pm$2.29)	0.853	67.45 ($\pm$2.29)	0.595	83.40 ($\pm$0.61)	0.748	83.39 ($\pm$0.63)	0.073
BA	36.40 ($\pm$2.40)	1.586	36.40 ($\pm$2.40)	0.784	68.51 ($\pm$1.91)	0.981	68.85 ($\pm$1.35)	0.157

Table 2: Experimental results for the four methods on different data sets in the linear setting: *acc*$-$ the best classification accuracy percentage; *std*$-$ the corresponding standard deviation; *time*$-$ the corresponding computational time (*s*).

7.3 Comparison of RKDA Methods

In the kernel setting, we compared Algorithms 2, 3 and RKDA/RR (Algorithm 5 working with **B**). We also implemented the KDA/GSVD method (Park and Park, 2005a) as a baseline. 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 among training data points. This setting was empirically found to be effective in real-world applications.

Table 3 summarizes the different evaluation criteria on all the data sets. Figures 4 and 5 further illustrate these results. As we see, our two RKDA methods yield better accuracy than the KDA/GSVD method and Algorithm 2. Moreover, RKDA/RR is more efficient computationally than the other methods, especially as the size of training data increases. It should be mentioned here that the data sets in our experiments range from small-sample to large-sample problems. Thus, Table 3 also confirms that the RKDA method based on (14) is more effective and efficient than the method based on (11).

Finally, Figure 6 presents the performance of the four regularized methods with respect to different regularization parameters σ on the four face data sets. From this figure, it can be seen that the regularized parameter σ plays an important role in our RFDA/RR and RKDA/RR methods. Similar results are obtained for the other regularized FDA or KDA methods compared here.

8. Beyond FDA

In this section we extend our results to a more general setting. We first apply the SVD-based algorithm to a family of generalized eigenvector problems, and then propose an efficient algorithm for penalized kernel canonical correlation analysis (KCCA) (Akaho, 2001; Van Gestel et al., 2001; Bach and Jordan, 2002).

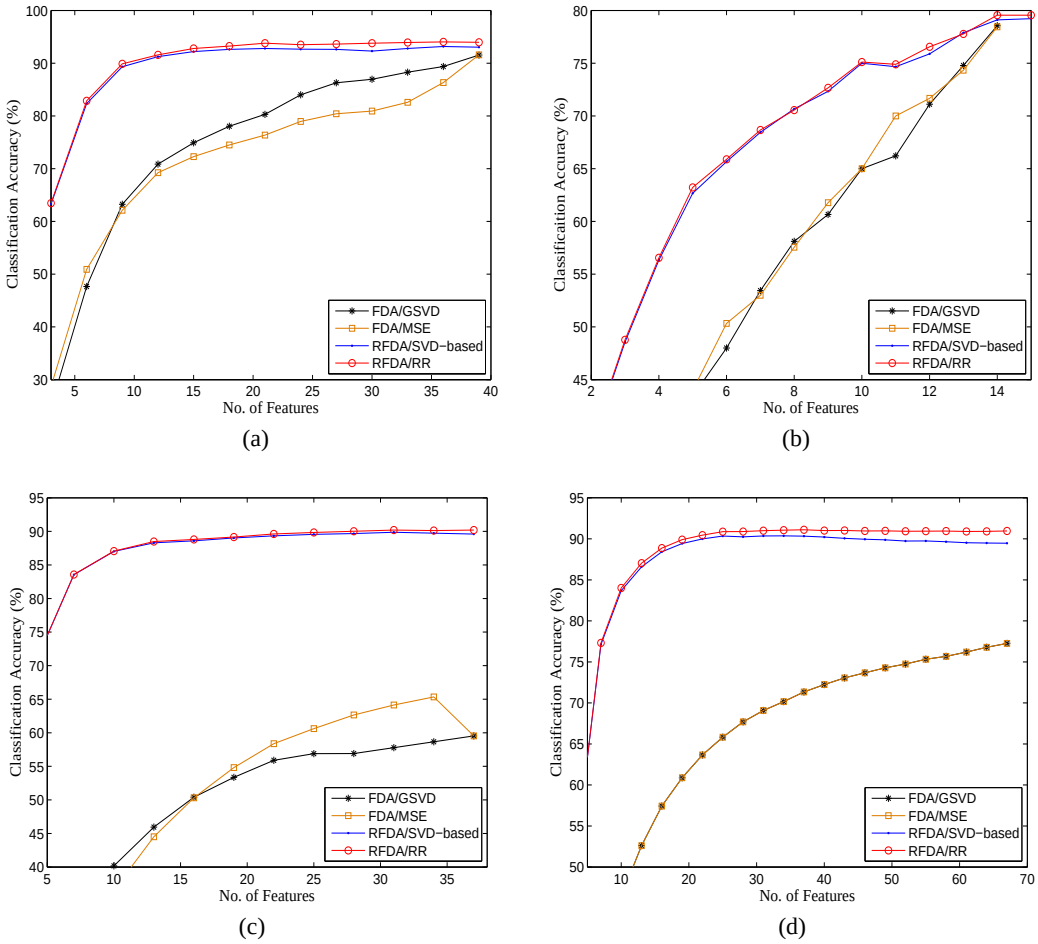


Figure 2: Comparison of the classification accuracies for FDA methods on the four face data sets: (a) ORL; (b) Yale; (c) YaleB&E; (d) PIE.

8.1 A Family of Generalized Eigenvector Problems

Assume that $\mathbf{Q}$ is a $p \times p$ semidefinite positive matrix. Here and later, we define

$$f(\mathbf{Q}) = \sum_{j=0}^{k} b_j \mathbf{Q}^j = b_0 \mathbf{I}_p + b_1 \mathbf{Q} + b_2 \mathbf{Q}^2 + \cdots + b_k \mathbf{Q}^k$$

where $b_0, \ldots, b_k$ are nonnegative real scalars for some positive integer k. We assume that there is at least one b_j such that $b_j > 0$. Let $\mathbf{Q} = \mathbf{V}\Gamma\mathbf{V}'$ be the SVD of $\mathbf{Q}$. We have

$$f(\mathbf{Q}) = \mathbf{V}(b_0 \mathbf{I}_p + b_1 \Gamma + b_2 \Gamma^2 + \cdots + b_k \Gamma^k)\mathbf{V}'.$$

This implies that $f(\mathbf{Q})$ is also semidefinite positive. Moreover, we have $\mathrm{rk}(\mathbf{Q}) \leq \mathrm{rk}(f(\mathbf{Q}))$. In fact, we have $\mathrm{rk}(\mathbf{Q}) = \mathrm{rk}(f(\mathbf{Q}))$ if $b_0 = 0$. However, $f(\mathbf{Q})$ is nonsingular whenever $b_0 > 0$.

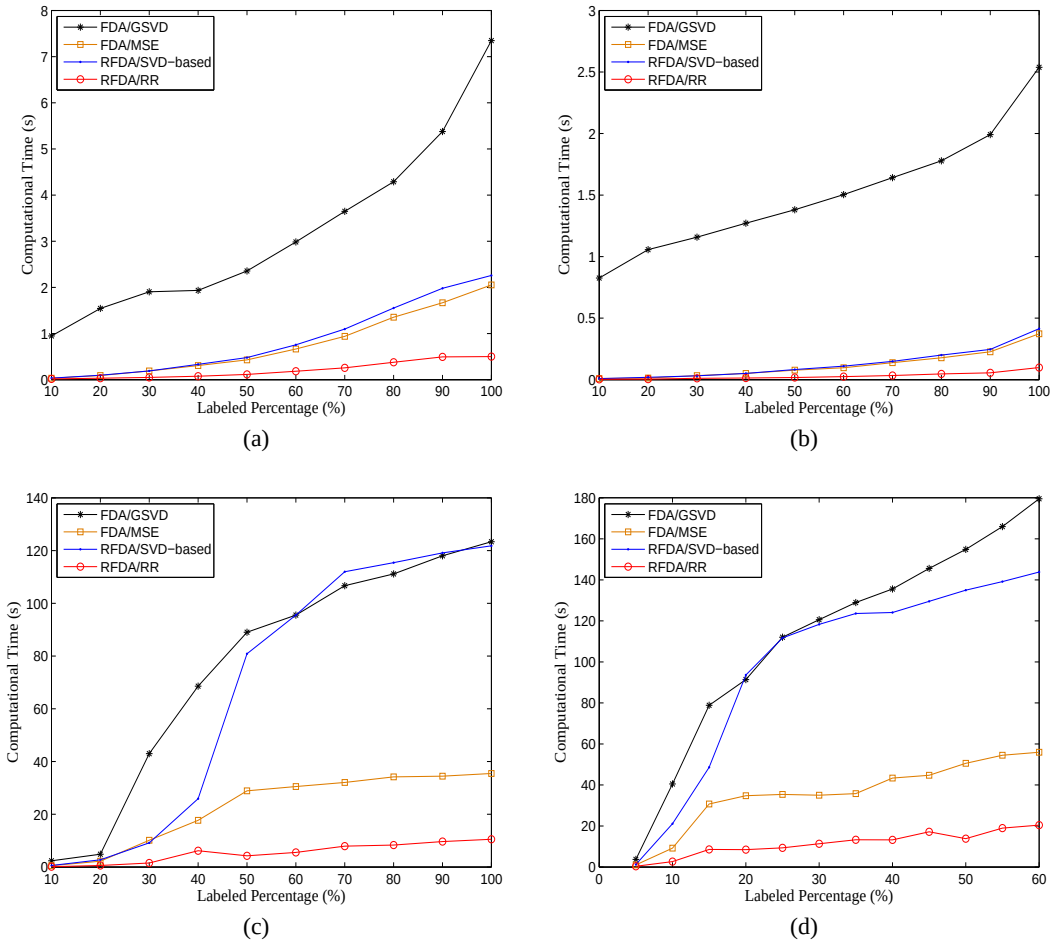


Figure 3: Comparison of the computational times for FDA methods as the training percentage k/n increases on the four face data sets: (a) ORL; (b) Yale; (c) YaleB&E; (d) PIE.

Letting $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} \in \mathbb{R}^{n\times m}$, we consider the following general optimization problem:

$$\max_{\mathbf{A}\in\mathbb{R}^{p\times q}} \operatorname{tr}\big(\mathbf{A}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{A}(\mathbf{A}'f(\mathbf{Q})\mathbf{A})^{-1}\big). \tag{29}$$

where $\mathbf{Q} = (\mathbf{X}'\mathbf{X})^{1/2}$ and $\operatorname{rk}(\mathbf{Q}) = \operatorname{rk}(\mathbf{X}) \geq q$. This problem can be formulated as a generalized eigenproblem as follows:

$$\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{A} = f(\mathbf{Q})\mathbf{A}\Lambda. \tag{30}$$

Thus, we consider the following eigenproblem:

$$\big(f(\mathbf{Q})\big)^{+}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{A} = \mathbf{A}\Lambda. \tag{31}$$

The following theorem shows a relationship between (30) and (31).

Theorem 9 *If $(\Lambda, \mathbf{A})$ (nonzero eigenpairs) is the solution of (31), then $(\Lambda, \mathbf{A})$ is the solution of (30). Conversely, if $(\Lambda, \mathbf{A})$ is the solution of (30), then $(\Lambda, (f(\mathbf{Q}))^{+} f(\mathbf{Q})\mathbf{A})$ is the solution of (31).*

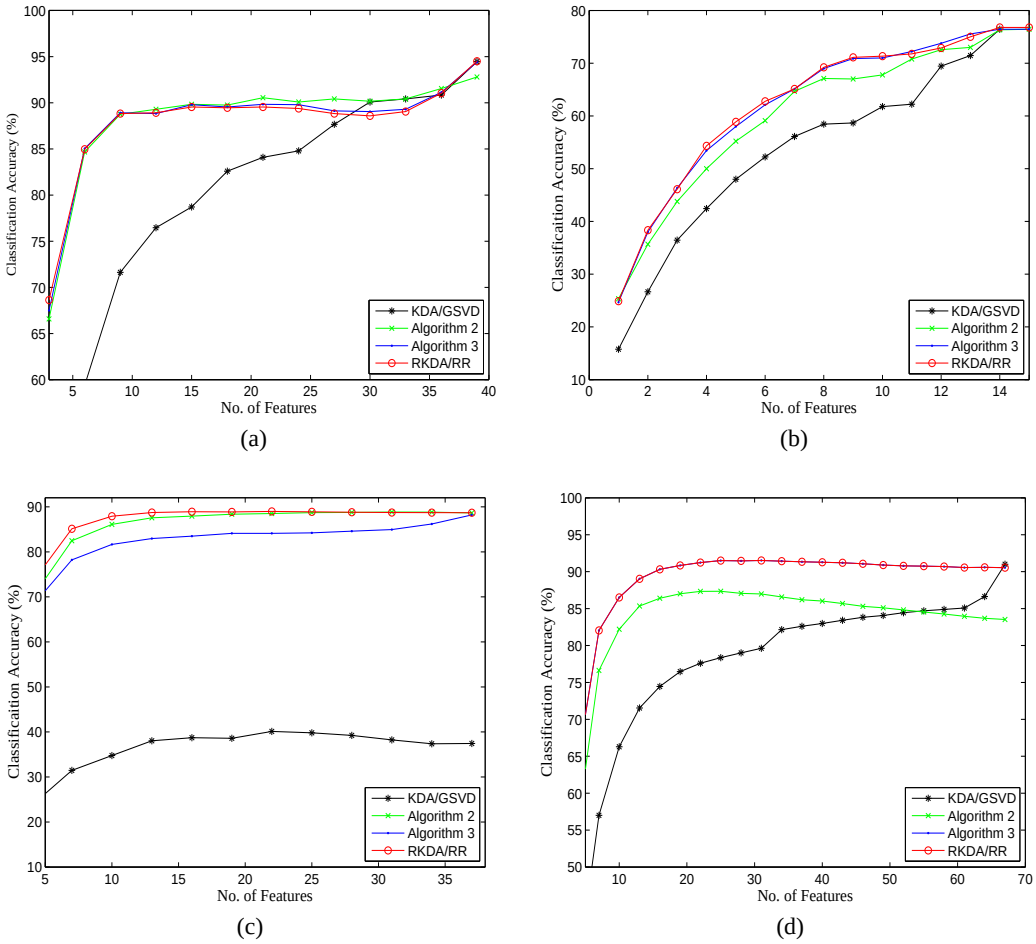


Figure 4: Comparison of the classification accuracies for KDA methods on the four face data sets: (a) ORL; (b) Yale; (c) YaleB&E; (d) PIE.

The theorem obviously holds when $b_0 > 0$, because $f(\mathbf{Q})$ is nonsingular. In the case that $b_0 = 0$, this theorem is a special case of Theorem 1.

Returning to the optimization problem in (29), we have the following theorem.

Theorem 10 *Assume that* $\mathrm{rk}(\mathbf{X}) = r \geq q$. *Let the condensed SVD of* $\mathbf{X}$ *be* $\mathbf{X} = \mathbf{U}_X \Gamma_X \mathbf{V}_X'$ *and the condensed SVD of* $\mathbf{F} = (f(\Gamma_X))^{-\frac{1}{2}} \Gamma_X \mathbf{U}_X' \mathbf{Y}$ *be* $\mathbf{F} = \mathbf{U}_F \Gamma_F \mathbf{V}_F'$. *We have: (i)* $(\Lambda, \mathbf{T})$ *where* $\mathbf{T} = \mathbf{V}_X (f(\Gamma_X))^{-\frac{1}{2}} \mathbf{U}_F$ *and* $\Lambda = \Gamma_F^2$ *are* r *eigenpairs of the pencil* $(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}, f(\mathbf{Q}))$*; (ii) the matrix* $\mathbf{T}_q$ *consisting of the first* q *columns of* $\mathbf{T}$ *is a maximizer of the generalized Rayleigh quotient in* (29).

The proof is given in Appendix G. This theorem shows that we can use the SVD-based algorithm to solve (29). That is, we obtain a derivation of Algorithm 6. Moreover, it is easily seen that the RFDA problems (6), (11), (14) and (16) are special cases of the problem (29) with different settings for the b_j.

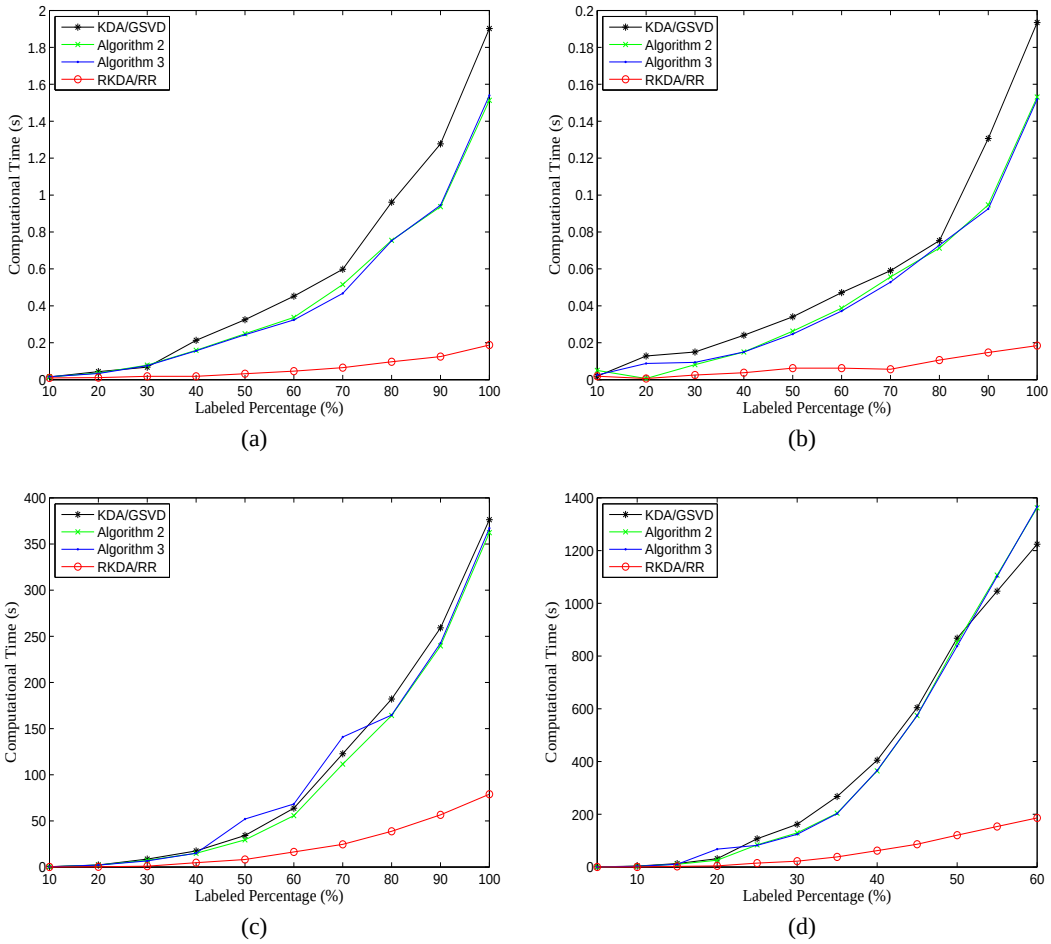


Figure 5: Comparison of the computational times for KDA methods as the training percentage k/n increases on the four face data sets: (a) ORL; (b) Yale; (c) YaleB&E; (d) PIE.

Algorithm 6 SVD-based Algorithm for Problem (29)

1: **procedure** GEP($\{\mathbf{X}, \mathbf{Y}, \sigma^2\}$)
2: Perform the condensed SVD of $\mathbf{X}$ as $\mathbf{X} = \mathbf{U}_X \Gamma_X \mathbf{V}_X'$.
3: Calculate $\mathbf{F} = (f(\Gamma_X))^{-\frac{1}{2}} \Gamma_X \mathbf{U}_X' \mathbf{Y}$ where $r = \mathrm{rk}(\Gamma_X)$.
4: Perform the condensed SVD of $\mathbf{F}$ as $\mathbf{F} = \mathbf{U}_F \Gamma_F \mathbf{V}_F'$.
5: Let $\mathbf{T} = \mathbf{V}_X (f(\Gamma_X))^{-\frac{1}{2}} \mathbf{U}_F$
6: Return $\mathbf{A} = \mathbf{T}(:, 1:q)$ for $q \leq r$ as a maximizer of Problem (29).
7: **end procedure**

8.2 Penalized KCCA

Given two data matrices $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^{n \times m}$, CCA finds two matrices $\mathbf{A}_x \in \mathbb{R}^{p \times q}$ and $\mathbf{A}_y \in \mathbb{R}^{m \times q}$ of canonical correlation vectors by solving the following optimization problem:

$$\max_{\mathbf{A}_x, \mathbf{A}_y} \mathrm{tr}(\mathbf{A}_x' \mathbf{S}_{xy} \mathbf{A}_y),$$

$$\text{s.t. } \mathbf{A}_x' \mathbf{S}_{xx} \mathbf{A}_x = \mathbf{I}_q \text{ and } \mathbf{A}_y' \mathbf{S}_{yy} \mathbf{A}_y = \mathbf{I}_q,$$

Data Set	KDA/GSVD		Algorithm 2		Algorithm 3		RKDA/RR	
	acc ($\pm std$)	*time*	*acc* ($\pm std$)	*time*	*acc* ($\pm std$)	*time*	*acc* ($\pm std$)	*time*
ORL	94.45 ($\pm$1.63)	0.231	92.79 ($\pm$1.74)	0.159	94.41 ($\pm$2.03)	0.162	94.50 ($\pm$1.64)	0.032
Yale	76.44 ($\pm$3.50)	0.017	76.44 ($\pm$2.71)	0.025	76.44 ($\pm$2.38)	0.025	76.78 ($\pm$3.20)	0.004
YaleB&E	40.34 ($\pm$22.4)	7.898	88.83 ($\pm$0.99)	6.520	88.20 ($\pm$0.88)	6.554	89.06 ($\pm$0.81)	0.818
PIE	91.00 ($\pm$0.36)	48.07	87.33 ($\pm$0.65)	40.71	91.52 ($\pm$0.45)	40.90	91.52 ($\pm$0.45)	5.079
USPS	82.25 ($\pm$1.59)	0.305	83.96 ($\pm$1.10)	0.238	84.92 ($\pm$1.56)	0.234	83.94 ($\pm$0.84)	0.020
Letters	92.74 ($\pm$2.10)	1.162	95.88 ($\pm$0.63)	1.100	94.60 ($\pm$0.99)	1.028	96.05 ($\pm$0.67)	0.134
WebKB	77.27 ($\pm$2.77)	1.684	83.51 ($\pm$0.51)	1.464	83.47 ($\pm$0.49)	1.452	83.47 ($\pm$0.49)	0.156
BA	66.19 ($\pm$1.21)	7.883	68.70 ($\pm$1.76)	6.715	69.86 ($\pm$1.30)	6.626	69.82 ($\pm$1.10)	0.709

Table 3: Experimental results for the five methods on different data sets in the kernel setting: *acc*– the best classification accuracy percentage; *std*– the corresponding standard deviation; *time*– the corresponding computational time (*s*).

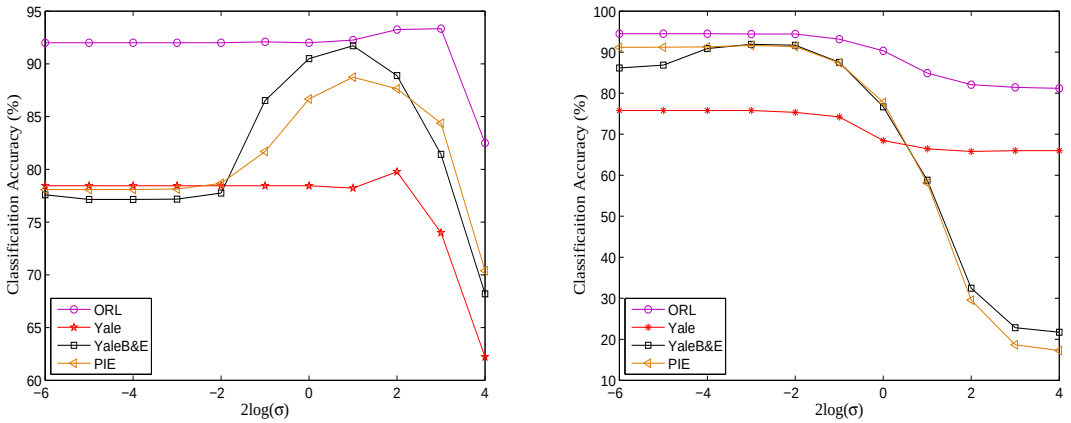


Figure 6: Performance of the RFDA/RR and RKDA/RR methods for different regularization parameters σ, where (a) displays the results of RFDA/RR on different data sets and (b) displays the results of RKDA/RR on different data sets.

where $q \leq \min\{p, m, n-1\}$, $\mathbf{S}_{xx} = \mathbf{X}'\mathbf{H}\mathbf{X}$ and $\mathbf{S}_{yy} = \mathbf{Y}'\mathbf{H}\mathbf{Y}$ are the pooled covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, respectively, and $\mathbf{S}_{xy} = \mathbf{X}'\mathbf{H}\mathbf{Y} = \mathbf{S}'_{yx}$ is the pooled cross-covariance matrix between $\mathbf{x}$ and $\mathbf{y}$.

Consider that either $\mathbf{S}_{xx}$ or $\mathbf{S}_{yy}$ is ill-conditioned. The penalized CCA method (Hastie et al., 1995) solves the following optimization problem

$$\max_{\mathbf{A}_x, \mathbf{A}_y} \ \mathrm{tr}(\mathbf{A}'_x \mathbf{S}_{xy} \mathbf{A}_y),$$

$$\text{s.t. } \mathbf{A}'_x(\mathbf{S}_{xx} + \sigma_x^2 \mathbf{I}_p)\mathbf{A}_x = \mathbf{I}_q \ \text{ and } \ \mathbf{A}'_y(\mathbf{S}_{yy} + \sigma_y^2 \mathbf{I}_m)\mathbf{A}_y = \mathbf{I}_q.$$

This problem can be solved in a two-step process (Mardia et al., 1979). The first step solves the following generalized problem:

$$\mathbf{S}_{yx}(\mathbf{S}_{xx}+\sigma_x^2\mathbf{I}_p)^{-1}\mathbf{S}_{xy}\mathbf{A}_y = (\mathbf{S}_{yy}+\sigma_y^2\mathbf{I}_m)\mathbf{A}_y\Lambda,$$

where Λ is a $q\times q$ diagonal matrix with positive diagonal elements. The second step calculates $\mathbf{A}_x$ by

$$\mathbf{A}_x = (\mathbf{S}_{xx}+\sigma_x^2\mathbf{I}_p)^{-1}\mathbf{S}_{xy}\mathbf{A}_y\Lambda^{-\frac{1}{2}}.$$

Assume that we have kernel functions $K_x(\cdot,\cdot)$: $\mathcal{X}\times\mathcal{X}\to\mathbb{R}$, and $K_y(\cdot,\cdot)$: $\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$. Similar to the linear case, penalized KCCA first solves the following generalized problem:

$$\tilde{\mathbf{S}}_{xy}(\tilde{\mathbf{S}}_{yy}+\sigma_y^2\mathbf{I}_h)^{-1}\tilde{\mathbf{S}}_{yx}\tilde{\mathbf{A}}_x = (\tilde{\mathbf{S}}_{xx}+\sigma_x^2\mathbf{I}_g)\tilde{\mathbf{A}}_x\Lambda, \tag{32}$$

and then calculates $\tilde{\mathbf{A}}_y$ by

$$\tilde{\mathbf{A}}_y = (\tilde{\mathbf{S}}_{yy}+\sigma_y^2\mathbf{I}_h)^{-1}\tilde{\mathbf{S}}_{yx}\tilde{\mathbf{A}}_x\Lambda^{-\frac{1}{2}}.$$

Here g and h are the dimensions of the corresponding feature spaces. We now address the solution to the generalized eigenproblem in (32). Consider

$$\begin{aligned}
&(\tilde{\mathbf{S}}_{xx}+\sigma_x^2\mathbf{I}_g)^{-1}\tilde{\mathbf{S}}_{xy}(\tilde{\mathbf{S}}_{yy}+\sigma_y^2\mathbf{I}_h)^{-1}\tilde{\mathbf{S}}_{yx}\\
&= (\tilde{\mathbf{X}}'\mathbf{H}\tilde{\mathbf{X}}+\sigma_x^2\mathbf{I}_g)^{-1}\tilde{\mathbf{X}}'\mathbf{H}\tilde{\mathbf{Y}}(\tilde{\mathbf{Y}}'\mathbf{H}\tilde{\mathbf{Y}}+\sigma_y^2\mathbf{I}_h)^{-1}\tilde{\mathbf{Y}}'\mathbf{H}\tilde{\mathbf{X}}\\
&= \tilde{\mathbf{X}}'\mathbf{H}(\mathbf{H}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H}+\sigma_x^2\mathbf{I}_n)^{-1}(\mathbf{H}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'\mathbf{H}+\sigma_y^2\mathbf{I}_n)^{-1}\mathbf{H}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'\mathbf{H}\tilde{\mathbf{X}}\\
&= \tilde{\mathbf{X}}'\mathbf{H}(\mathbf{C}_x+\sigma_x^2\mathbf{I}_n)^{-1}(\mathbf{C}_y+\sigma_y^2\mathbf{I}_n)^{-1}\mathbf{C}_y\mathbf{H}\tilde{\mathbf{X}},
\end{aligned}$$

where $\mathbf{C}_x = \mathbf{H}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H}$ and $\mathbf{C}_y = \mathbf{H}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'\mathbf{H}$. Since $\tilde{\mathbf{X}}'\mathbf{H}(\mathbf{C}_x+\sigma_x^2\mathbf{I}_n)^{-1}(\mathbf{C}_y+\sigma_y^2\mathbf{I}_n)^{-1}\mathbf{C}_y\mathbf{H}\tilde{\mathbf{X}}$ and $(\mathbf{C}_x+\sigma_x^2\mathbf{I}_n)^{-1}(\mathbf{C}_y+\sigma_y^2\mathbf{I}_n)^{-1}\mathbf{C}_y\mathbf{C}_x$ have the same nonzero eigenvalues, we let

$$(\mathbf{C}_x+\sigma_x^2\mathbf{I}_n)^{-1}(\mathbf{C}_y+\sigma_y^2\mathbf{I}_n)^{-1}\mathbf{C}_y\mathbf{C}_x\Upsilon = \Upsilon\Lambda$$

where Λ consists of the q largest nonzero eigenvalues of $(\mathbf{C}_x+\sigma_x^2\mathbf{I}_n)^{-1}(\mathbf{C}_y+\sigma_y^2\mathbf{I}_n)^{-1}\mathbf{C}_y\mathbf{C}_x$. We thus define

$$\tilde{\mathbf{A}}_x = \tilde{\mathbf{X}}'\mathbf{H}\Upsilon$$

and

$$\tilde{\mathbf{A}}_y = \tilde{\mathbf{Y}}'\mathbf{H}(\mathbf{C}_y+\sigma_y^2\mathbf{I}_n)^{-1}\mathbf{C}_x\Upsilon\Lambda^{-\frac{1}{2}}$$

as the solution of the KCCA problem. Given $\mathbf{x}\in\mathbb{R}^p$ and $\mathbf{y}\in\mathbb{R}^m$, we can directly calculate their canonical variables by

$$\mathbf{z}_x = \tilde{\mathbf{A}}_x'(\tilde{\mathbf{x}}-\tilde{\mathbf{m}}_x) = \Upsilon'\Big(\mathbf{k}_x-\frac{1}{n}\mathbf{K}_x\mathbf{1}_n\Big),$$

where $\tilde{\mathbf{m}}_x = \frac{1}{n}\sum_{i=1}^n\tilde{\mathbf{x}}_i$, $\mathbf{k}_x = (K_x(\mathbf{x},\mathbf{x}_1),\ldots,K_x(\mathbf{x},\mathbf{x}_n))'$ and $\mathbf{K}_x = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'$, and

$$\mathbf{z}_y = \tilde{\mathbf{A}}_y'(\tilde{\mathbf{y}}-\tilde{\mathbf{m}}_y) = \Lambda^{-\frac{1}{2}}\Upsilon'\mathbf{C}_x(\mathbf{C}_y+\sigma_y^2\mathbf{I}_n)^{-1}\Big(\mathbf{k}_y-\frac{1}{n}\mathbf{K}_y\mathbf{1}_n\Big),$$

where $\tilde{\mathbf{m}}_y = \frac{1}{n}\sum_{i=1}^n\tilde{\mathbf{y}}_i$, $\mathbf{k}_y = (K_y(\mathbf{y},\mathbf{y}_1),\ldots,K_y(\mathbf{y},\mathbf{y}_n))'$ and $\mathbf{K}_y = \tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'$.

As we see, the canonical vectors $\mathbf{z}_x$ and $\mathbf{z}_y$ can be calculated without the explicit use of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$. Moreover, if $\sigma_x^2=0$ or $\sigma_y^2=0$, the algorithm still works by using $\mathbf{C}_x^+$ or $\mathbf{C}_y^+$ instead.

9. Conclusion

In this paper we have provided a solution to an open problem concerning the relationship between multi-class discriminant analysis problems and multivariate regression problems, both in the linear setting and the kernel setting. Our theory has yielded efficient and effective algorithms for FDA and KDA within both the regularization and pseudoinverse paradigms. The favorable performance of our algorithms has been demonstrated empirically on a collection of benchmark data sets. We have also extended our algorithms to a more general family of generalized eigenvalue problems.

Acknowledgments

Zhihua Zhang and Congfu Xu acknowledge support from the 973 Program of China (No. 2010CB327903). Zhihua Zhang acknowledges support from Natural Science Foundations of China (No. 61070239), Doctoral Program of Specialized Research Fund of Chinese Universities, and the Fundamental Research Funds for the Central Universities. Michael Jordan acknowledges support from Google, Intel and Microsoft Research.

Appendix A. Proof of Theorem 1

Let $\Sigma_1 = \mathbf{U}_1\Gamma_1\mathbf{V}_1'$ and $\Sigma_2 = \mathbf{U}_2\Gamma_2\mathbf{V}_2'$ be the condensed SVD of Σ_1 and Σ_2. Thus, we have $\mathcal{R}(\Sigma_1) = \mathcal{R}(\mathbf{U}_1)$ and $\mathcal{R}(\Sigma_2) = \mathcal{R}(\mathbf{U}_2)$. Moreover, we have $\Sigma_2^+ = \mathbf{V}_2\Gamma_2^{-1}\mathbf{U}_2'$ and $\Sigma_2\Sigma_2^+ = \mathbf{U}_2\mathbf{U}_2'$. It follows from $\mathcal{R}(\Sigma_1) \subseteq \mathcal{R}(\Sigma_2)$ that $\mathcal{R}(\mathbf{U}_1) \subseteq \mathcal{R}(\mathbf{U}_2)$. This implies that $\mathbf{U}_1$ can be expressed as $\mathbf{U}_1 = \mathbf{U}_2\mathbf{Q}$ where $\mathbf{Q}$ is some matrix of appropriate order. As a result, we have

$$\Sigma_2\Sigma_2^+\Sigma_1 = \mathbf{U}_2\mathbf{U}_2'\mathbf{U}_2\mathbf{Q}\Gamma_1\mathbf{V}_1' = \Sigma_1.$$

It is worth noting that the condition $\Sigma_2\Sigma_2^+\Sigma_1 = \Sigma_1$ is not only necessary but also sufficient for $\mathcal{R}(\Sigma_1) \subseteq \mathcal{R}(\Sigma_2)$.

If $(\Lambda, \mathbf{B})$ are the eigenpairs of $\Sigma_2^+\Sigma_1$, then it is easily seen that $(\Lambda, \mathbf{B})$ are also the eigenpairs of (Σ_1, Σ_2) due to $\Sigma_2\Sigma_2^+\Sigma_1 = \Sigma_1$.

Conversely, suppose $(\Lambda, \mathbf{B})$ are the eigenpairs of (Σ_1, Σ_2). Then we have $\Sigma_2\Sigma_2^+\Sigma_1\mathbf{B} = \Sigma_2\mathbf{B}\Lambda$. This implies that $(\Lambda, \Sigma_2^+\Sigma_2\mathbf{B})$ are the eigenpairs of $\Sigma_2^+\Sigma_1$ due to $\Sigma_2\Sigma_2^+\Sigma_1 = \Sigma_1$ and $\Sigma_2^+\Sigma_2\Sigma_2^+ = \Sigma_2^+$.

Appendix B. Some Properties of Moore-Penrose Inverses

In order to prove Theorem 3, we will need some properties of Moore-Penrose inverses.

Lemma 11 *Let* $\mathbf{C} = \mathbf{H}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H}$, $\tilde{\mathbf{S}}_t = \tilde{\mathbf{X}}'\mathbf{H}\tilde{\mathbf{X}}$ *and* $\tilde{\mathbf{S}}_b = \tilde{\mathbf{X}}'\mathbf{H}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{H}\tilde{\mathbf{X}}$. *Then*

(a) $\mathbf{C}\mathbf{C}^+ = (\mathbf{C}\mathbf{C}^+)' = \mathbf{C}^+\mathbf{C}$, $\mathbf{C}^+\mathbf{C}\mathbf{C} = \mathbf{C}\mathbf{C}^+\mathbf{C} = \mathbf{C}$, $\mathbf{C}^+\mathbf{C}^+\mathbf{C} = \mathbf{C}^+\mathbf{C}\mathbf{C}^+ = \mathbf{C}^+$;

(b) $\tilde{\mathbf{S}}_t^+\tilde{\mathbf{X}}'\mathbf{H} = (\tilde{\mathbf{X}}'\mathbf{H}\mathbf{H}\tilde{\mathbf{X}})^+\tilde{\mathbf{X}}'\mathbf{H} = \tilde{\mathbf{X}}'\mathbf{H}(\mathbf{H}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H})^+ = \tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+$;

(c) $\tilde{\mathbf{X}}'\mathbf{H} = \tilde{\mathbf{S}}_t\tilde{\mathbf{S}}_t^+\tilde{\mathbf{X}}'\mathbf{H} = \tilde{\mathbf{X}}'\mathbf{H}\mathbf{H}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{H}\mathbf{H}\tilde{\mathbf{X}})^+\tilde{\mathbf{X}}'\mathbf{H}$
$= \tilde{\mathbf{X}}'\mathbf{H}(\mathbf{H}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H})^+\mathbf{H}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H} = \tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+\mathbf{C}$;

(d) $\tilde{\mathbf{S}}_t\tilde{\mathbf{S}}_t^+\tilde{\mathbf{S}}_b = \tilde{\mathbf{S}}_b$.

These results can be found in Lütkepohl (1996).

Appendix C. Proof of Theorem 3

First, if (Λ, Υ) is the solution of (10), we have

$$\begin{aligned}\tilde{\mathbf{S}}_b\tilde{\mathbf{X}}'\mathbf{H}\Upsilon &= \tilde{\mathbf{X}}'\mathbf{H}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon \\ &= \tilde{\mathbf{X}}'\mathbf{H}\mathbf{H}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{H}\mathbf{H}\tilde{\mathbf{X}})^+\tilde{\mathbf{X}}'\mathbf{H}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon \\ &= \tilde{\mathbf{S}}_t\tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = \tilde{\mathbf{S}}_t\tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+\mathbf{C}^+\mathbf{C}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon \\ &= \tilde{\mathbf{S}}_t\tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+\mathbf{C}^+\mathbf{C}\mathbf{C}\Upsilon\Lambda = \tilde{\mathbf{S}}_t\tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+\mathbf{C}\Upsilon\Lambda \\ &= \tilde{\mathbf{S}}_t\tilde{\mathbf{X}}'\mathbf{H}\Upsilon\Lambda.\end{aligned}$$

This implies that $\left(\Lambda, \tilde{\mathbf{X}}'\mathbf{H}\Upsilon\right)$ is the solution of (7).

Second, if (Λ, Υ) is the solution of (15), we have

$$\begin{aligned}\tilde{\mathbf{S}}_t^+\tilde{\mathbf{S}}_b\tilde{\mathbf{X}}'\mathbf{H}\Upsilon &= (\tilde{\mathbf{X}}'\mathbf{H}\mathbf{H}\tilde{\mathbf{X}})^+\tilde{\mathbf{X}}'\mathbf{H}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{H}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H}\Upsilon \\ &= \tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon \\ &= \tilde{\mathbf{X}}'\mathbf{H}\Upsilon\Lambda.\end{aligned}$$

This implies that $(\Lambda, \tilde{\mathbf{X}}'\mathbf{H}\Upsilon)$ is the solution of (13).

Finally, it follows from (16) that

$$(\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{C}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = \mathbf{C}\Upsilon\Lambda.$$

In addition, note that $(\tilde{\mathbf{S}}_t+\sigma^2\mathbf{I}_g)^{-1}\tilde{\mathbf{X}}'\mathbf{H} = \tilde{\mathbf{X}}'\mathbf{H}(\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}$. Hence, we have

$$\begin{aligned}(\tilde{\mathbf{S}}_t+\sigma^2\mathbf{I}_g)^{-1}\tilde{\mathbf{S}}_b\tilde{\mathbf{X}}'\mathbf{H}\Upsilon &= (\tilde{\mathbf{S}}_t+\sigma^2\mathbf{I}_g)^{-1}\tilde{\mathbf{X}}'\mathbf{H}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon \\ &= \tilde{\mathbf{X}}'\mathbf{H}(\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon \\ &= \tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+\mathbf{C}(\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon \\ &= \tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+(\mathbf{C}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{C}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = \tilde{\mathbf{X}}'\mathbf{H}\mathbf{C}^+\mathbf{C}\Upsilon \\ &= \tilde{\mathbf{X}}'\mathbf{H}\Upsilon\Lambda.\end{aligned}$$

This completes the proof of part (iii).

Appendix D. Proof of Theorem 5

We prove the final part. As for other parts, their proof can be immediately obtained from Appendix G. In terms of Algorithm 3, we have

$$\begin{aligned}\tilde{\mathbf{A}}'(\tilde{\mathbf{S}}_t+\sigma^2\mathbf{I}_g)\tilde{\mathbf{A}} &= \Upsilon'\mathbf{H}\tilde{\mathbf{X}}(\tilde{\mathbf{S}}_t+\sigma^2\mathbf{I}_g)\tilde{\mathbf{X}}'\mathbf{H}\Upsilon \\ &= \Upsilon'\mathbf{C}(\mathbf{C}+\sigma^2\mathbf{I}_n)\Upsilon = \mathbf{I}_q\end{aligned}$$

and

$$\tilde{\mathbf{A}}'\tilde{\mathbf{S}}_b\tilde{\mathbf{A}} = \Upsilon'\mathbf{C}\mathbf{E}\Pi^{-1}\mathbf{E}'\mathbf{C}\Upsilon = \Gamma_F^2.$$

Appendix E. Derivation of Equation 25

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

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

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

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

because of $\mathbf{X}'\mathbf{X} - n\bar{\mathbf{x}}\bar{\mathbf{x}}' = \mathbf{X}'\mathbf{H}\mathbf{X}$. We thus obtain $\mathbf{W}$ in (25). It then follows from (18) that $\mathbf{W} = \mathbf{G}$. Moreover, when $\sigma^2 = 0$, $\mathbf{W}$ reduces to the solution of the minimization problem in (26). In this case, if $\mathbf{X}'\mathbf{H}\mathbf{X}$ is singular, a standard treatment is to use the Moore-Penrose inverse $(\mathbf{X}'\mathbf{H}\mathbf{X})^+$ in (25). Such a $\mathbf{W}$ is identical with $\mathbf{G}$ in (21).

Appendix F. Proof of Theorem 8

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

$$\begin{aligned} \mathbf{W}\mathbf{W}' &= \mathbf{W}\mathbf{V}\mathbf{V}'\mathbf{W}' \\ &= \mathbf{W}\mathbf{V}_R\mathbf{V}_R'\mathbf{W}' + \mathbf{W}\mathbf{V}_2\mathbf{V}_2'\mathbf{W}' \\ &= \mathbf{B}\mathbf{B}'. \end{aligned}$$

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

Appendix G. Proof of Theorem 10

It is immediate that

$$\begin{aligned} \mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{T} &= \mathbf{V}_X\Gamma_X\mathbf{U}_X'\mathbf{Y}\mathbf{Y}'\mathbf{U}_X\Gamma_X\mathbf{V}_X'\mathbf{V}_X(f(\Gamma_X))^{-\frac{1}{2}}\mathbf{U}_F \\ &= \mathbf{V}_X(f(\Gamma_X))^{\frac{1}{2}}\mathbf{U}_F\Gamma_F\mathbf{V}_F'\mathbf{V}_F\Gamma_F\mathbf{U}_F'\mathbf{U}_F \\ &= \mathbf{V}_X(f(\Gamma_X))^{\frac{1}{2}}\mathbf{U}_F\Gamma_F^2 \end{aligned}$$

and

$$f(\mathbf{Q})\mathbf{T}\Lambda = \mathbf{V}\begin{bmatrix} f(\Gamma_X) & \mathbf{0} \\ \mathbf{0} & b_0\mathbf{I}_{p-r} \end{bmatrix}\mathbf{V}'\mathbf{V}_X(f(\Gamma_X))^{-\frac{1}{2}}\mathbf{U}_F\Gamma_F^2$$
$$= \mathbf{V}_X(f(\Gamma_X))^{\frac{1}{2}}\mathbf{U}_F\Gamma_F^2.$$

where $\mathbf{V} = [\mathbf{V}_X, \mathbf{V}_2]$ such that $\mathbf{V}_X'\mathbf{V}_2 = \mathbf{0}$. In addition, we have

$$\mathbf{T}'f(\mathbf{Q})\mathbf{T} = \mathbf{I}_r \quad \text{and} \quad \mathbf{T}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{T} = \Gamma_F^2.$$

References

S. Akaho. A kernel method for canonical correlation analysis. In *International Meeting of Psychometric Society*, 2001.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404, 2000.

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.

M. Craven, D. Dopasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the World Wide Web. In *The Fifteenth Conference on Artificial Intelligence*, 1998.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, New York, second edition, 2001.

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

A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. 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, 2001.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.

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

T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23(1): 73–102, 1995.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2001.

A. E. Hoerl and R. W. Kennard. Ridge regression. *Technometrics*, 12:56–82, 1970.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985.

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.

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.

J. Kittler and P. C. Young. A new approach to feature selection based on the Karhunen-Loève expansion. *Pattern Recognition*, 5:335–352, 1973.

K. C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

H. Lütkepohl. *Handbook of Matrices*. John Wiley & Sons, New York, 1996.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, New York, 1979.

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.

C. C. Paige and M. A. Saunders. Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, 18(3):398–405, 1981.

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, 2005a.

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, 2005b.

K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor. The differogram: Nonparametric noise variance estimation and its use for model. *Neurocomputing*, 69:100–122, 2005.

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.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293–300, 1999.

J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

C. F. Van Loan. Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis*, 13(3):76–83, 1976.

T. Van Gestel, J. A. K. Suykens, J. De Brabanter, B. De Moor, and J. Vandewalle. Kernel canonical correlation analysis and least squares support vector machines. In *The International Conference on Artificial Neural Networks (ICANN)*, pages 381–386, 2001.

T. Van Gestel, J. A. K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle. Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis. *Neural Computation*, 14:1115–1147, 2002.

A. R. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, Hoboken, NJ, 2002.

J. Ye. Least squares linear discriminant analysis. In *The Twenty-Fourth International Conference on Machine Learning (ICML)*, 2007.

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.

Z. Zhang and G. Dai. Optimal scoring for unsupervised learning. In *Advances in Neural Information Processing Systems 23*, volume 12, pages 2241–2249, 2009.

Z. Zhang and M. I. Jordan. Multiway spectral clustering: A margin-based perspective. *Statistical Science*, 3:383–403, 2008.