NONLINEAR DISCRIMINANT ANALYSIS USING KERNEL FUNCTIONS AND THE GENERALIZED SINGULAR VALUE DECOMPOSITION*

CHEONG HEE PARK† AND HAESUN PARK‡

Abstract. Linear Discriminant Analysis (LDA) has been widely used for linear dimension reduction. However, LDA has some limitations that one of the scatter matrices is required to be nonsingular and the nonlinearly clustered structure is not easily captured. In order to overcome the problems caused by the singularity of the scatter matrices, a generalization of LDA based on the generalized singular value decomposition (GSVD) has been developed recently. In this paper, we propose a nonlinear discriminant analysis based on the kernel method and the generalized singular value decomposition. The GSVD is applied to solve the generalized eigenvalue problem which is formulated in the feature space defined by a nonlinear mapping through kernel functions. Our GSVD-based kernel discriminant analysis is theoretically compared with other kernel-based nonlinear discriminant analysis algorithms. The experimental results show that our method is an effective nonlinear dimension reduction method.

Key words. Dimension reduction, Generalized singular value decomposition, Kernel functions, Linear Discriminant Analysis, Nonlinear Discriminant Analysis

AMS subject classifications. 15A09, 68T10, 62H30, 65F15, 15A18

1. Introduction. In Linear Discriminant Analysis (LDA), a linear transformation is found which maximizes the between-class scatter and minimizes the within-class scatter [7, 9]. Although LDA is conceptually simple and has been used in many application areas, it has some limitations: it requires at least one of the scatter matrices to be nonsingular and it can not easily capture a nonlinearly clustered structure.

One common situation where all of the scatter matrices are singular is when the number of data points is smaller than the dimension of the data space, and this situation is often referred to as an undersampled problem. Numerous methods have been proposed to overcome this difficulty [8, 4, 19]. Recently, a method called LDA/GSVD has been proposed, which is a generalization of LDA based on the generalized singular value decomposition (GSVD) [10]. It overcomes the problem caused by the singularity of the scatter matrices in undersampled problems by applying the GSVD to solve a generalized eigenvalue problem.

In order to make LDA applicable to nonlinearly structured data, kernel-based methods have been applied. The main idea of kernel-based methods is to map the input data to a feature space by a nonlinear mapping where inner products in the feature space can be computed by a kernel function without knowing the nonlinear mapping explicitly. Kernel Principal Component Analysis (Kernel PCA) [18], Kernel Fisher Discriminant Analysis (KFD) [12] and nonlinear Discriminant Analysis [1, 16, 2] are nonlinear extensions of the well known PCA, Fisher Discriminant Analysis, Linear Discriminant Analysis based on the kernel method, respectively. However, PCA or Kernel PCA may not be appropriate as a dimension reduction method for clustered data, since the purpose of these methods is an optimal lower dimensional representation rather than discrimination. KFD [12] and the method proposed in [2]

*This material is based upon work supported by the National Science Foundation Grants CCR-0204109 and ACI-0305543. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

†Dept. of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 (chpark@cs.umn.edu).

‡Dept. of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 (hpark@cs.umn.edu) and The National Science Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230. The work of Haesun Park has been performed while at the NSF and was partly supported by IR/D from the National Science Foundation (NSF).

have been developed to handle the data that consists of two classes only. In Generalized Discriminant Analysis (GDA) [1], centering in the feature space is performed by shifting each vector by the global average, and then the kernel matrix is computed. Although this kernel matrix is assumed to be nonsingular, centering in the feature space makes the kernel matrix singular, even when the kernel function is symmetric positive definite. This makes the theoretical development of GDA [1] break down. In addition, when the input space is mapped to a feature space through a kernel function, the dimension of the feature space often becomes much larger than that of the original data space, and as a result, the scatter matrices become singular.

Towards a general nonlinear discriminant analysis, we propose a kernel-based nonlinear extension of LDA using the GSVD. We also show the relationships of our GSVD-based kernel discriminant analysis with other kernel-based nonlinear discriminant analysis algorithms. After reviewing the linear dimension reduction method LDA/GSVD in Section 2, we present the new Kernel Discriminant Analysis, KDA/GSVD, in Section 3. The relationships of KDA/GSVD with other kernel-based methods are discussed in Section 4. Experimental results are given in Section 5 and we conclude with discussions.

2. Linear Discriminant Analysis. Throughout the paper, a data set of n data vectors in an m-dimensional space is denoted as

$$A = [a_1, \cdots, a_n] = [A_1, \cdots, A_r] \in R^{m \times n} \tag{2.1}$$

where the data is clustered to r classes and each block $A_i \in R^{m \times n_i}$ has n_i data vectors. Let N_i $(1 \leq i \leq r)$ be the set of column indices that belong to the class i. The between-class scatter matrix S_b and the within-class scatter matrix S_w are defined as

$$S_b = \sum_{i=1}^{r} n_i (c_i - c)(c_i - c)^T \quad \text{and} \quad S_w = \sum_{i=1}^{r} \sum_{j \in N_i} (a_j - c_i)(a_j - c_i)^T \tag{2.2}$$

where

$$c_i = \frac{1}{n_i} \sum_{j \in N_i} a_j \quad \text{and} \quad c = \frac{1}{n} \sum_{j=1}^{n} a_j \tag{2.3}$$

are the centroid of the class i and the global centroid, respectively. The separability of classes in a data set can be measured by using the traces of these scatter matrices.

The goal of Linear Discriminant Analysis (LDA) is to find a transformation matrix $G \in R^{m \times l}$ for some integer l with $l \leq m$ that defines a linear transformation

$$G^T : a \in R^{m \times 1} \rightarrow y = G^T a \in R^{l \times 1}$$

and preserves the cluster structure by maximizing the between-class scatter and minimizing the within-class scatter. In the transformed space by G^T, the between-class scatter matrix $\tilde{S}_b$ and the within-class scatter matrix $\tilde{S}_w$ become

$$\tilde{S}_b = G^T S_b G \quad \text{and} \quad \tilde{S}_w = G^T S_w G,$$

respectively. A commonly used criterion in LDA for finding an optimal clustered structure preserving transformation G^T is

$$\max_G \text{trace}((G^T S_w G)^{-1} (G^T S_b G)). \tag{2.4}$$

It is well known [9] that this criterion is satisfied when $l = r - 1$ where r is the number of the classes in the data, and the columns of $G \in R^{m\times(r-1)}$ are the eigenvectors corresponding to the $r-1$ largest eigenvalues for the eigenvalue problem

$$S_w^{-1} S_b x = \lambda x. \tag{2.5}$$

However, as in many applications such as information retrieval [11] and face recognition [4], when the number of data items is smaller than the dimension of data space, S_w becomes singular. Recently, a method which applies the GSVD to solve the generalized eigenvalue problem

$$S_b x = \lambda S_w x \tag{2.6}$$

has been developed [10].

The method in [10] utilizes representations of the scatter matrices as

$$S_b = H_b H_b^T \quad \text{and} \quad S_w = H_w H_w^T \tag{2.7}$$

where

$$H_b = [\sqrt{n_1}(c_1 - c), \cdots, \sqrt{n_r}(c_r - c)] \in R^{m\times r},$$
$$H_w = [A_1 - c_1 e_1^T, \cdots, A_r - c_r e_r^T] \in R^{m\times n} \quad \text{and} \quad e_i = [1, \cdots, 1]^T \in R^{n_i \times 1}.$$

By applying the GSVD to the pair (H_b^T, H_w^T), we have

$$U^T H_b^T X = [\underbrace{\Sigma_b}_{t} \underbrace{0}_{m-t}] \quad \text{and} \quad V^T H_w^T X = [\underbrace{\Sigma_w}_{t} \underbrace{0}_{m-t}], \quad t = \text{rank}\left(\begin{bmatrix} H_b^T \\ H_w^T \end{bmatrix}\right) \tag{2.8}$$

where U and V are orthogonal and X is nonsingular, $\Sigma_b^T\Sigma_b + \Sigma_w^T\Sigma_w = I_t$ and $\Sigma_b^T\Sigma_b$ and $\Sigma_w^T\Sigma_w$ are diagonal matrices with nonincreasing and nondecreasing diagonal components respectively. Then the simultaneous diagonalizations of S_b and S_w can be obtained as

$$X^T S_b X = \begin{bmatrix} \Sigma_b^T\Sigma_b & 0 \\ 0 & 0 \end{bmatrix} \quad X^T S_w X = \begin{bmatrix} \Sigma_w^T\Sigma_w & 0 \\ 0 & 0 \end{bmatrix}. \tag{2.9}$$

Let us denote the diagonal elements of $\Sigma_b^T\Sigma_b$ and $\Sigma_w^T\Sigma_w$ as η_i and ζ_i, i.e.

$$\Sigma_b^T\Sigma_b = \text{diag}(\eta_1, \cdots, \eta_t), \quad \eta_1 \geq \cdots \geq \eta_t \text{ and}$$
$$\Sigma_w^T\Sigma_w = \text{diag}(\zeta_1, \cdots, \zeta_t), \quad \zeta_1 \leq \cdots \leq \zeta_t$$

and $X = [x_1, \cdots, x_m]$. From Eqs. in (2.9), the column vectors x_i of X satisfy

$$\zeta_i S_b x_i = \eta_i S_w x_i, \ 1 \leq i \leq t. \tag{2.10}$$

For $t + 1 \leq i \leq m$

$$x_i^T S_b x_i = 0 \quad \text{and} \quad x_i^T S_w x_i = 0$$

and they do not convey discriminative information among classes. Hence an optimal dimension reducing transformation G^T can be obtained from the first $r-1$ columns[1] of X as

$$G = [x_1, \cdots, x_{r-1}].$$

The algorithm for LDA/GSVD can be found in [10, 14].

[1]When $l = \text{rank}(S_b)$, $x_i^T S_b x_i = 0$ and $x_i^T S_w x_i = 1$ for $l + 1 \leq i \leq t$. Hence one can consider taking only the first l columns from X. However, in practice the rank of S_b is one less than the number of classes, that is, $r - 1$ in most of real data sets.

3. Nonlinear Discriminant Analysis based on Kernel Functions and the GSVD. In this section, we present a nonlinear extension of LDA based on kernel functions and the GSVD. The main idea of the kernel method is that without knowing the nonlinear feature mapping or the mapped feature space explicitly, we can work on the feature space through kernel functions, as long as the problem formulation depends only on the inner products between data points. This is based on the fact that for any kernel function κ satisfying Mercer's condition [5], there exists a mapping Φ such that

$$< \Phi(a), \Phi(b) >= \kappa(a, b) \tag{3.1}$$

where $<, >$ is an inner product in the feature space transformed by Φ [17, 3]. For a finite data set $\{a_1, \cdots, a_n\}$, a kernel function κ satisfying Mercer's condition can be rephrased as the kernel matrix $K = [\kappa(a_i, a_j)]_{1 \le i,j \le n}$ being positive semi-definite [5]. The polynomial kernel

$$\kappa(x, y) = (\gamma_1(x \cdot y) + \gamma_2)^d, d > 0 \text{ and } \gamma_1, \gamma_2 \in R \tag{3.2}$$

and the Gaussian kernel

$$\kappa(x, y) = \exp(-\|x - y\|^2/\sigma), \sigma \in R \tag{3.3}$$

are two of the most widely used kernel functions. The feature map Φ can be either linear or nonlinear depending on kernel functions used. If the inner product kernel function $\kappa(x, y) = x \cdot y$ is used, the feature map is an identity map. In the kernel methods neither the feature map nor the feature space needs to be formed explicitly due to the relation (3.1) once the kernel function κ is known.

We apply the kernel method to perform LDA in the feature space instead of the original input space. Given a kernel function κ, let Φ be a mapping satisfying (3.1) and define $\mathcal{F} \subset R^N$ to be the feature space from the mapping Φ. As in (2.7), scatter matrices $\mathcal{S}_b$ and $\mathcal{S}_w$ in the feature space $\mathcal{F}$ can be expressed as

$$\mathcal{S}_b = \mathcal{H}_b \mathcal{H}_b^T \quad \text{and} \quad \mathcal{S}_w = \mathcal{H}_w \mathcal{H}_w^T \tag{3.4}$$

where

$$\begin{aligned} &\mathcal{H}_b = [\sqrt{n_1}(\tilde{c}_1 - \tilde{c}), \cdots, \sqrt{n_r}(\tilde{c}_r - \tilde{c})] \in R^{N \times r}, \\ &\mathcal{H}_w = [\Phi(A_1) - \tilde{c}_1 e_1^T, \cdots, \Phi(A_r) - \tilde{c}_r e_r^T] \in R^{N \times n}, \\ &\tilde{c}_i = \frac{1}{n_i} \sum_{j \in N_i} \Phi(a_j), \ \tilde{c} = \frac{1}{n} \sum_{i=1}^{n} \Phi(a_i) \ \text{ and } \ e_i = [1, \cdots, 1]^T \in R^{n_i \times 1}. \end{aligned} \tag{3.5}$$

The notations $\Phi(A_i)$ are used to denote $\Phi([a_j, \cdots, a_k]) = [\Phi(a_j), \cdots, \Phi(a_k)]$. Then the LDA in $\mathcal{F}$ finds a transformation matrix

$$\mathcal{G} = [\varphi_1, \cdots, \varphi_{r-1}] \in R^{N \times (r-1)}$$

where the columns of $\mathcal{G}$ are the generalized eigenvectors corresponding to the $r - 1$ largest eigenvalues of

$$\mathcal{S}_b \varphi = \lambda \mathcal{S}_w \varphi. \tag{3.6}$$

Now we show how to solve the problem (3.6) without knowing the explicit representation of the mapping Φ and the feature space $\mathcal{F}$, therefore without forming $\mathcal{S}_b$ and $\mathcal{S}_w$ explicitly. Note that any vector $\varphi \in R^{N \times 1}$ can be represented as

$$\varphi = \varphi_1 + \varphi_2$$

where $\varphi_1 \in \text{span}\{\Phi(A)\}$ and $\varphi_2 \in \text{span}\{\Phi(A)\}^\perp$, and $\mathcal{S}_b\varphi_2 = 0$ and $\mathcal{S}_w\varphi_2 = 0$ for any $\varphi_2 \in \text{span}\{\Phi(A)\}^\perp$. Therefore, for any vector φ satisfying (3.6),

$$\mathcal{S}_b\varphi_1 = \mathcal{S}_b(\varphi_1 + \varphi_2) = \lambda S_w(\varphi_1 + \varphi_2) = \lambda S_w\varphi_1.$$

Hence we can restrict the solution space for (3.6) to $\text{span}\{\Phi(A)\}$. One may refer to [13] for an alternative explanation.

Let φ be represented as a linear combination of $\Phi(a_i)$, $i = 1, \cdots, n$,

$$\varphi = \sum_{i=1}^{n} \alpha_i \Phi(a_i) \quad \text{and} \quad \alpha = [\alpha_1, \cdots, \alpha_n]^T. \tag{3.7}$$

The following theorem gives a formula by which $\mathcal{S}_b$ can be expressed through the kernel function.

THEOREM 3.1. *Let*

$$\mathcal{K}_b = [\, b_{ij} \,]_{(1\le i\le n, 1\le j\le r)}, \quad b_{ij} = \sqrt{n_j}\left(\frac{1}{n_j}\sum_{s\in N_j}\kappa(a_i, a_s) - \frac{1}{n}\sum_{s=1}^{n}\kappa(a_i, a_s)\right). \tag{3.8}$$

Then

$$\mathcal{H}_b^T\varphi = \mathcal{K}_b^T\alpha. \tag{3.9}$$

Proof. From (3.5) and (3.7),

$$\begin{aligned}
&\mathcal{H}_b^T\varphi && (3.10)\\
&= \begin{bmatrix} \sqrt{n_1}(\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ \sqrt{n_r}(\tilde{c}_r - \tilde{c})^T \end{bmatrix} \left(\sum_{i=1}^{n}\alpha_i\Phi(a_i)\right)\\
&= \begin{bmatrix} \sqrt{n_1}(\frac{1}{n_1}\sum_{s\in N_1}\Phi(a_s) - \frac{1}{n}\sum_{s=1}^{n}\Phi(a_s))^T \\ \vdots \\ \sqrt{n_r}(\frac{1}{n_r}\sum_{s\in N_r}\Phi(a_s) - \frac{1}{n}\sum_{s=1}^{n}\Phi(a_s))^T \end{bmatrix} \begin{bmatrix} \Phi(a_1), & \cdots, & \Phi(a_n) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}\\
&= \mathcal{K}_b^T\alpha. \qquad \square && (3.11)
\end{aligned}$$

Similarly to Theorem 3.1 , we can obtain

$$\mathcal{H}_w^T\varphi = \mathcal{K}_w^T\alpha \tag{3.12}$$

where

$$\begin{aligned}
&\mathcal{K}_w = [\, w_{ij} \,]_{(1\le i\le n, 1\le j\le n)}, && (3.13)\\
&w_{ij} = \kappa(a_i, a_j) - \frac{1}{n_l}\sum_{s\in N_l}\kappa(a_i, a_s) \text{ when } a_j \text{ belongs to the class } l.
\end{aligned}$$

THEOREM 3.2. *The generalized eigenvalue problem*

$$\mathcal{S}_b\varphi = \lambda\mathcal{S}_w\varphi$$

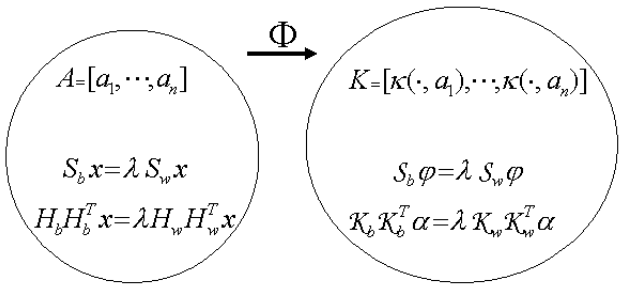

FIG. 3.1. *The problem formulations of LDA in the original data space and the feature space defined by a nonlinear mapping* Φ *through a kernel function,* $< \Phi(a_i), \Phi(a_j) >= \kappa(a_i, a_j)$. *In the kernel matrix* K, $\kappa(\cdot, a_i)$ *denotes a column vector* $[\kappa(a_1, a_i), \cdots, \kappa(a_n, a_i)]^T$.

is equivalent to

$$\mathcal{K}_b \mathcal{K}_b^T \alpha = \lambda \mathcal{K}_w \mathcal{K}_w^T \alpha \tag{3.14}$$

where $\varphi = \sum_{i=1}^n \alpha_i \Phi(a_i)$ *and* $\alpha = [\alpha_1, \cdots, \alpha_n]^T$.

Proof. From (3.9) and (3.12),

$$\begin{aligned} \mathcal{S}_b \varphi = \lambda \mathcal{S}_w \varphi \quad &\Leftrightarrow \quad \psi^T \mathcal{H}_b \mathcal{H}_b^T \varphi = \lambda \psi^T \mathcal{H}_w \mathcal{H}_w^T \varphi \\ &\Leftrightarrow \quad \beta^T \mathcal{K}_b \mathcal{K}_b^T \alpha = \lambda \beta^T \mathcal{K}_w \mathcal{K}_w^T \alpha \\ &\Leftrightarrow \quad \mathcal{K}_b \mathcal{K}_b^T \alpha = \lambda \mathcal{K}_w \mathcal{K}_w^T \alpha \end{aligned} \tag{3.15}$$

for any $\psi = \sum_{i=1}^n \beta_i \Phi(a_i)$ and $\beta = [\beta_1, \cdots, \beta_n]^T$. □

Note that $\mathcal{K}_b \mathcal{K}_b^T$ and $\mathcal{K}_w \mathcal{K}_w^T$ can be viewed as the between-class scatter matrix and within-class scatter matrix of the kernel matrix

$$K = [\, \kappa(a_i, a_j) \,]_{(1 \le i \le n, 1 \le j \le n)}$$

when each column in K is considered as a data point in the n-dimensional space. It can be observed by comparing the structures of $\mathcal{K}_b$ and $\mathcal{K}_w$ with those of $\mathcal{H}_b$ and $\mathcal{H}_w$ in (3.5). Figure 3.1 illustrates the corresponding relations in the original data space and the feature space.

Note that $\mathcal{K}_b \mathcal{K}_b^T$ and $\mathcal{K}_w \mathcal{K}_w^T$ are both singular and the classical LDA can not be applied. Now we apply the GSVD to the pair $(\mathcal{K}_b^T, \mathcal{K}_w^T)$ in order to solve (3.14), and as in (2.9) we have

$$\mathcal{X}^T \mathcal{K}_b \mathcal{K}_b^T \mathcal{X} = \begin{bmatrix} \Gamma_b^T \Gamma_b & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathcal{X}^T \mathcal{K}_w \mathcal{K}_w^T \mathcal{X} = \begin{bmatrix} \Gamma_w^T \Gamma_w & 0 \\ 0 & 0 \end{bmatrix} \tag{3.16}$$

where the columns of $\mathcal{X}$ solves (3.14). Let $\mathcal{G}$ be the matrix obtained by the first $r-1$ columns of $\mathcal{X}$ as

$$\mathcal{G} = [\alpha^{(1)}, \cdots, \alpha^{(r-1)}] = \begin{bmatrix} \alpha_{1\,1} & \cdots & \alpha_{1\,r-1} \\ \vdots & & \vdots \\ \alpha_{n\,1} & \cdots & \alpha_{n\,r-1} \end{bmatrix}.$$

Defining

$$\varphi_i = \sum_{j=1}^n \alpha_{j\,i} \Phi(a_j), \quad 1 \le i \le r-1,$$

Algorithm 1 KDA/GSVD

Given a data matrix $A = [a_1, \cdots, a_n] \in R^{m \times n}$ with r classes and a kernel function κ, it computes the $r-1$ dimensional representation of any input vector $z \in R^m$ by applying the GSVD in the feature space defined by the feature mapping Φ such that $\kappa(a_i, a_j) = < \Phi(a_i), \Phi(a_j) >$.

1. Compute $\mathcal{K}_b \in R^{n \times r}$ and $\mathcal{K}_w \in R^{n \times n}$ according to Eqs. (3.8) and (3.13).
2. Apply the GSVD to the pair $(\mathcal{K}_b^T, \mathcal{K}_w^T)$:

$$\mathcal{U}^T \mathcal{K}_b^T \mathcal{X} = [\Gamma_b \quad 0] \quad \text{and} \quad \mathcal{V}^T \mathcal{K}_w^T \mathcal{X} = [\Gamma_w \quad 0].$$

3. Assign the first $r-1$ columns of $\mathcal{X}$ to $\mathcal{G}$:
4. For any input vector $z \in R^{m \times 1}$, a dimension reduced representation is computed as

$$\mathcal{G}^T \begin{bmatrix} \kappa(a_1, z) \\ \vdots \\ \kappa(a_n, z) \end{bmatrix} \in R^{(r-1) \times 1}.$$

by Theorem 3.2, φ_i satisfies

$$\mathcal{S}_b \varphi_i = \lambda_i \mathcal{S}_w \varphi_i$$

and $[\varphi_1, \cdots, \varphi_{r-1}]$ gives a linear transformation by the LDA in the feature space. Hence, for any input vector $z \in R^{m \times 1}$, the dimension reduced representation of z is given by

$$\begin{aligned} &[\varphi_1, \cdots, \varphi_{r-1}]^T \Phi(z) \\ &= \left[\sum_{j=1}^{n} \alpha_{j\,1} \Phi(a_j)^T \Phi(z), \cdots, \sum_{j=1}^{n} \alpha_{j\,r-1} \Phi(a_j)^T \Phi(z) \right] = \mathcal{G}^T \begin{bmatrix} \kappa(a_1, z) \\ \vdots \\ \kappa(a_n, z) \end{bmatrix}. \end{aligned}$$

This method, called KDA/GSVD, is summarized in Algorithm 1.

4. Comparison of Kernel-based Nonlinear Discriminant Analysis Algorithms. In Section 3, we showed that KDA/GSVD finds the solution by solving

$$\mathcal{K}_b \mathcal{K}_b^T \alpha = \lambda \mathcal{K}_w \mathcal{K}_w^T \alpha \tag{4.1}$$

using the GSVD. In this section, we compare our GSVD-based approach with two other methods, the regularization based method [8] and the one based on the minimum squared error function [6, 7], and derive the relationships of KDA/GSVD with other kernel-based nonlinear discriminant analysis algorithms.

4.1. A Relationship to Kernel Fisher Discriminant Analysis (KFD). Mika et al. [12] developed a nonlinear extension of Fisher Discriminant Analysis, Kernel Fisher Discriminant Analysis (KFD), using the regularization and kernel methods. KFD is a method for finding a dimension reducing transformation specifically when the data set has only two classes. In a two-class problem, the between-class scatter matrix $\mathcal{S}_b$ is expressed as

$$\mathcal{S}_b = \frac{n_1 n_2}{n} (\tilde{c}_1 - \tilde{c}_2)(\tilde{c}_1 - \tilde{c}_2)^T,$$

where $\tilde{c}_i$, $i = 1, 2$, denote the class centroids. Hence the KFD criterion, maximization of

$$\frac{\varphi^T(\tilde{c}_1 - \tilde{c}_2)(\tilde{c}_1 - \tilde{c}_2)^T\varphi}{\varphi^T\mathcal{S}_w\varphi}, \tag{4.2}$$

is equivalent to maximization of

$$\frac{\varphi^T\mathcal{S}_b\varphi}{\varphi^T\mathcal{S}_w\varphi} = \frac{\alpha^T\mathcal{K}_b\mathcal{K}_b^T\alpha}{\alpha^T\mathcal{K}_w\mathcal{K}_w^T\alpha} \equiv \lambda \tag{4.3}$$

where $\varphi = \sum_{i=1}^n \alpha_i\Phi(a_i)$, $\alpha = [\alpha_1, \cdots, \alpha_n]^T$. Setting the derivative of (4.3) with respect to α as 0 gives the eigenvalue problem

$$\mathcal{K}_b\mathcal{K}_b^T\alpha = \lambda\mathcal{K}_w\mathcal{K}_w^T\alpha. \tag{4.4}$$

The KFD in [12] solves the problem (4.2), that is, the eigenvalue problem (4.4) by the regularization method where a positive diagonal matrix $d\,I$ is added to $\mathcal{K}_w\mathcal{K}_w^T$ to make it nonsingular. However, regularization parameter should be determined experimentally and this procedure can be expensive. The performances by the regularization method and KDA/GSVD are compared in our experiments.

4.2. Using the Minimum Squared Error Function. The Minimum Squared Error (MSE) formulation in a two-class problem (i.e., $r = 2$) seeks a linear discriminant function

$$g(z) = w_0 + w^T z$$

where

$$g(a_i) = w_0 + w^T a_i = \begin{cases} b_1, & \text{if } i \in N_1 \\ b_2, & \text{if } i \in N_2 \end{cases} \tag{4.5}$$

and b_i is the prespecified number for each class. For the data set A given in (2.1), the problem (4.5) can be reformulated as a problem of minimizing the squared error

$$\left\| \begin{bmatrix} 1 & a_1^T \\ \vdots & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \right\|_2^2 \equiv \left\| P \begin{bmatrix} w_0 \\ w \end{bmatrix} - y \right\|_2^2 \tag{4.6}$$

where $y_i = b_1$ if $i \in N_1$ and $y_i = b_2$ if $i \in N_2$. Note that the matrix P is $n \times (m+1)$ and the linear system involved in (4.6) is underdetermined when $n < m + 1$ and overdetermined when $n > m + 1$. In either case, the solution which minimizes the squared error (4.6) can be computed using the pseudoinverse P^+ of P as

$$\begin{bmatrix} w_0 \\ w \end{bmatrix} = P^+ y. \tag{4.7}$$

When $b_1 = n/n_1$ and $b_2 = -n/n_2$, the MSE solution is related with Fisher Discriminant Analysis (FDA) [6, 7]. The vector w in (4.7) is same as the solution of FDA except for a scaling factor and

$$w_0 = -w^T c$$

where c is the global centroid defined in (2.3). For a new data item, it is assigned to the class 1 if

$$w^T z + w_0 = w^T(z - c) > 0, \tag{4.8}$$

otherwise it is assigned to the class 2.

Now (4.6) can be extended to the squared error function in the feature space by a mapping Φ through a kernel function κ. By substituting a_i with $\Phi(a_i)$ and w with $\sum_{j=1}^{n} \alpha_j \Phi(a_j)$ as in (3.7), we obtain

$$\left\| \begin{bmatrix} 1 & \kappa(a_1,a_1) & \cdots & \kappa(a_1,a_n) \\ & \vdots & & \vdots \\ 1 & \kappa(a_n,a_1) & \cdots & \kappa(a_n,a_n) \end{bmatrix} \begin{bmatrix} w_0 \\ \alpha \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \right\|_2^2 \equiv \|\mathcal{P}\theta - y\|_2^2 \qquad (4.9)$$

where $\alpha = [\alpha_1, \cdots, \alpha_n]^T$. The matrix $\mathcal{P}$ in (4.9) is $n \times (n+1)$ and therefore $\mathcal{P}\theta = y$ is an underdetermined linear system. Choosing the solution $\theta = \mathcal{P}^+ y$, a data item z is assigned to the class 1 if

$$\begin{bmatrix} 1 & \kappa(z,a_1) & \cdots & \kappa(z,a_n) \end{bmatrix} \begin{bmatrix} w_0 \\ \alpha \end{bmatrix} > 0 \qquad (4.10)$$

While the MSE solution in (4.6) is related to Fisher Discriminant Analysis (FDA) when the within-class scatter matrix is nonsingular [6, 7], it can be shown that the MSE solution in (4.6) is related to LDA/GSVD in the case of the singular scatter matrix. The relation between the MSE solution of (4.6) and LDA/GSVD for two-class problems and the corresponding relation between the kernel MSE solution of (4.9) and KDA/GSVD are presented in the Appendix.

4.3. Generalized Discriminant Analysis (GDA). In Generalized Discriminant Analysis (GDA) [1], centering of the data in the feature space is performed by shifting each feature vector by the global centroid

$$\tilde{c} = \frac{1}{n} \sum_{i=1}^{n} \Phi(a_i)$$

and then the kernel matrix is computed. This kernel matrix

$$\tilde{K} = [\, k_{i,j} \,]_{1 \le i,j \le n} \quad \text{where} \quad k_{i,j} = (\Phi(a_i) - \tilde{c})^T (\Phi(a_j) - \tilde{c})$$

is assumed to be nonsingular. However, $\tilde{K}$ is always singular since

$$\tilde{K} = \begin{bmatrix} (\Phi(a_1) - \tilde{c})^T \\ \vdots \\ (\Phi(a_n) - \tilde{c})^T \end{bmatrix} \begin{bmatrix} \Phi(a_1) - \tilde{c}, & \cdots, & \Phi(a_n) - \tilde{c} \end{bmatrix}$$

and

$$\text{rank}(\begin{bmatrix} \Phi(a_1) - \tilde{c}, \cdots, \Phi(a_n) - \tilde{c} \end{bmatrix}) \le n - 1.$$

This makes the theoretical development in [1] break down. In numerical experiments, $\tilde{K}$ may only be detected as being very ill-conditioned since $\tilde{K}$ is symmetric positive semidefinite and zero eigenvalues may appear to be extremely small positive numbers due to rounding errors.

In solving the generalized eigenvalue problem (4.1), the GSVD is applied to $\mathcal{K}_b^T$ and $\mathcal{K}_w^T$ instead of $\mathcal{K}_b \mathcal{K}_b^T$ and $\mathcal{K}_w \mathcal{K}_w^T$, hence the products in $\mathcal{K}_b \mathcal{K}_b^T$ and $\mathcal{K}_w \mathcal{K}_w^T$ do not need to be computed explicitly. Moreover, the GSVD does not require any parameter optimization such as a regularization parameter. When the regularization method is used for the problem (4.1), the inverse of the matrix $\mathcal{K}_w \mathcal{K}_w^T + d\, I$ should be computed in addition to the eigenvalue decomposition or singular value decomposition which is also required in the GSVD approach or MSE solution. In the next section, experimental comparisons of the performances of KDA/GSVD and other kernel based methods are presented.

TABLE 5.1
The description of datasets.

dataset	Musk	Isolet	Car	Mfeature
no. of classes	2	26	4	10
dim	166	617	6	649
no. of data	6599	7797	1728	2000

5. Experimental Results. We demonstrate that our proposed method KDA/GSVD is an effective nonlinear extension of LDA by comparing the performances of KDA/GSVD and other kernel-based nonlinear discriminant analysis algorithms as well as kernel-based Principal Component Analysis (Kernel PCA) [18].

For the first experiment, data sets were collected from UCI machine learning repository[2]. The detailed description of data sets is shown in Table 5.1. After randomly splitting the data to the training and test sets of equal size, cross-validation is used with the training set in order to determine the optimal value for σ in Gaussian kernel function

$$\kappa(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right). \tag{5.1}$$

In cross-validation, first, the average of pairwise distances in the training data is computed. Then based on the average distance avg_d, an optimal value C in

$$\sigma = C \cdot avg_d \tag{5.2}$$

which gives the highest prediction accuracy is searched. In our experiments, we found that $[0.2, 1.0]$ is a reasonable range for C.

For the generalized eigenvalue problem (4.1), the regularization method is applied for the comparison with KDA/GSVD. A regularization parameter d ($d > 0$) is used to make the matrix $\mathcal{K}_w\mathcal{K}_w^T$ nonsingular and then the eigenvalue problem

$$(\mathcal{K}_w\mathcal{K}_w^T + dI)^{-1}\mathcal{K}_b\mathcal{K}_b^T\alpha = \lambda\alpha$$

is solved. In our experiments, while the regularization parameter d was set as 1, the optimal value σ in the Gaussian kernel function was searched by cross-validation. In Kernel PCA, the reduced dimension was one less than the number of classes as in other methods. After dimension reduction, k-nearest neighbor (k-NN) classifiers were used for the k-values of 1, 15, 29.

Figure 5.1 compares the performances of KDA/GSVD, KFD and Kernel PCA for the Musk data which has two classes. The top figures in Figure 5.1 show the prediction accuracies by 10 cross-validation in the training set of the Musk data, where the x-axis corresponds to the values C in (5.2) which ranged from 0.1 to 1.5 with an interval 0.1. In the second row in Figure 5.1, the prediction accuracies for the various C values in the test data are also shown for the comparison with those obtained by cross-validation. Table 5.2 shows the results in the test sets using the parameters chosen by cross-validation. It also shows the prediction accuracies obtained by the LDA in the original data space. The experimental results demonstrate that the GSVD-based nonlinear discriminant analysis, KDA/GSVD, obtained the competent prediction accuracies over the compared methods, while it does not require any additional parameter optimization as in the regularization method and it can naturally handle the multi-class problems.

[2]http://www.ics.uci.edu/∼mlearn/MLRepository.html

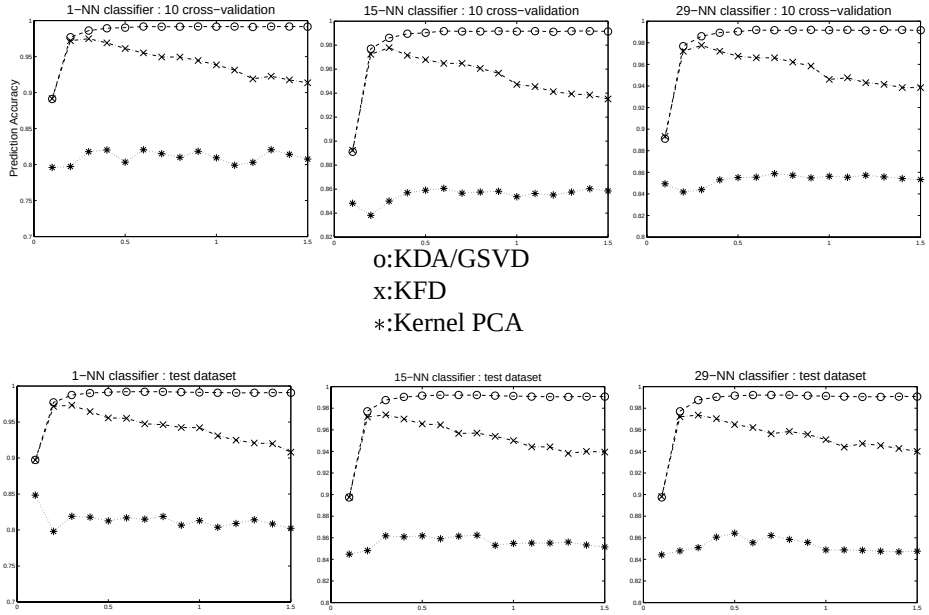

FIG. 5.1. *The Prediction accuracies in the Musk data. The top figures were obtained by the cross-validation with the training data and the bottom figures show the prediction accuracies in the test data. The x-axis corresponds to the values C in $\sigma = C \cdot avg_d$ where avg_d denotes the average of pairwise distances in the training dataset.*

TABLE 5.2
The prediction accuracies (%) in the multi-class problems

Dataset	k-NN	LDA	KDA/GSVD	Regularization	Kernel PCA
Musk	$k = 1$	91.0	**99.2**	97.3	81.7
	15	93.5	**99.2**	97.4	85.9
	29	93.7	**99.2**	97.4	86.2
Isolet	1	94.1	**97.0**	95.8	84.0
	15	94.1	**97.0**	96.1	85.7
	29	93.9	**97.0**	96.2	85.0
Car	1	87.5	94.2	**95.7**	64.9
	15	88.8	94.2	**95.1**	71.3
	29	87.2	94.2	**94.8**	72.8
Mfeature	1	98.2	**98.4**	93.6	87.5
	15	97.9	**98.4**	94.4	85.6
	29	97.8	**98.4**	94.5	82.8

In the next experiment, the purpose is to evaluate the performance of KDA/GSVD for an undersampled problem. The data set was constructed by randomly selecting 500 documents from each of five categories from the MEDLINE data set. The documents were preprocessed with stemming, stop-list and rare term removal and encoded using the term frequency and inverse document frequency [11], resulting in a total of 22095 terms. Equally splitting documents in each category into training and test data sets, each of them has 1250 documents. Support Vector Machine (SVM) classifiers as well as k-nearest neighbors and centroid-based classification method were applied both in the original data space and in the reduced dimensional space. Since the SVM classifier is for binary class problem and our data set has 5 classes, we used a DAG scheme for multi-class classification [15]. Table 5.3 shows the prediction accuracies. After both linear and nonlinear SVMs were applied in the original data

TABLE 5.3
Prediction accuracy (%) on MEDLINE data.

		k-NN			Centroid-	
	dimension	30	45	60	based	SVM
Original data	22095×1250	83.5	84.0	83.8	84.8	89.5
KDA/GSVD	4 ×1250	89.4	89.4	89.4	89.4	89.7

space, the best accuracy was obtained with a linear soft margin SVM. On the other hand, we obtained a competitive result by a linear SVM in the dimension reduced space by KDA/GSVD using the Gaussian kernel. Since the dimension was reduced dramatically from 22095 down to 4 and it was trained with only a linear classifier in the reduced dimension, the training process was much faster than in the full dimension. Even with k-NN and centroid-based classification methods, prediction results that were as good as with SVM were obtained. The high prediction accuracies by k-NN or centroid-based classifiers in the reduced dimensional space by KDA/GSVD show that the difficulty of applying a binary classifier as SVM to multi-class problem can be overcome effectively.

6. Discussion. We have introduced KDA/GSVD which is a nonlinear extension of LDA based on kernel functions and the generalized singular value decomposition. One advantage of KDA/GSVD is that it can be applied regardless of singularity of the scatter matrices both in the original space and in the feature space by a nonlinear mapping. It is also shown that in two-class problem KDA/GSVD is related to the kernel version of the MSE solution. The comparison with other methods in solving the generalized eigenvalue problem demonstrates that KDA/GSVD is an effective dimension reduction method for multi-class problems.

REFERENCES

[1] G. BAUDAT AND F. ANOUAR, *Generalized discriminant analysis using a kernel approach*, Neural computation, 12 (2000), pp. 2385–2404.
[2] S. BILLINGS AND K. LEE, *Nonlinear fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm*, Neural networks, 15(2) (2002), pp. 263–270.
[3] C. BURGES, *A tutorial on support vector machines for pattern recognition*, Data Mining and Knowledge Discovery, 2(2) (1998), pp. 121–167.
[4] L. CHEN, H. LIAO, M. KO, J. LIN, AND G. YU, *A new lda-based face recognition system which can solve the small sample size problem*, pattern recognition, 33 (2000), pp. 1713–1726.
[5] N. CRISTIANINI AND J. SHAWE-TAYLOR, *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge, 2000.
[6] R. DUDA AND P. HART, *Pattern classification and scene analysis*, Wiley Interscience, 1973.
[7] R. DUDA, P. HART, AND D. STORK, *Pattern Classification*, Wiley-interscience, New York, 2001.
[8] J. FRIEDMAN, *Regularized discriminant analysis*, Journal of the American statistical association, 84(405) (1989), pp. 165–175.
[9] K. FUKUNAGA, *Introduction to Statistical Pattern Recognition*, Academic Press, second ed., 1990.
[10] P. HOWLAND, M. JEON, AND H. PARK, *Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition*, SIAM Journal on Matrix Analysis and Applications, 25(1) (2003), pp. 165–179.
[11] G. KOWALSKI, *Information Retrieval System: Theory and Implementation*, Kluwer Academic Publishers, 1997.
[12] S. MIKA, G. RÄTSCH, J. WESTON, B. SCHÖLKOPF, AND K.-R. MÜLLER, *Fisher discriminant analysis with kernels*, in Neural networks for signal processing IX, E. J.Larsen and S.Douglas, eds., IEEE, 1999, pp. 41–48.
[13] S. MIKA, G. RÄTSCH, J. WESTON, B. SCHÖLKOPF, A. SMOLA, AND K.-R. MÜLLER, *Invariant feature extraction and classification in kernel spaces*, Advances in neural information processing systems, 12 (2000), pp. 526–532.
[14] C.H. PARK AND H. PARK, *A comparison of generalized lda algorithms for undersampled problems*. Technical Reports 03-048, Department of Computer Science and Engineering, University of Minnesota, Twin

Cities, 2003.
[15] J. PLATT, N. CRISTIANINI, AND J. SHAWE-TAYLOR, *Large margin dags for multiclass classification*, Advances in Neural Information Processing Systems, 12 (2000), pp. 547–553.
[16] V. ROTH AND V. STEINHAGE, *Nonlinear discriminant analysis using kernel functions*, Advances in neural information processing systems, 12 (2000), pp. 568–574.
[17] B. SCHÖLKOPF, S. MIKA, C. BURGES, P. KNIRSCH, K.-R. MÜLLER, G. RÄTSCH, AND A. SMOLA, *Input space versus feature space in kernel-based methods*, IEEE transactions on neural networks, 10 (September 1999), pp. 1000–1017.
[18] B. SCHÖLKOPF, A. SMOLA, AND K.-R. MÜLLER, *Nonlinear component analysis as a kernel eigenvalue problem*, Neural computation, 10 (1998), pp. 1299–1319.
[19] H. YU AND J. YANG, *A direct lda algorithm for high-dimensional data- with application to face recognition*, pattern recognition, 34 (2001), pp. 2067–2070.

Appendix. In this appendix, a relationship between the MSE solution of (4.6) and the LDA/GSVD solution for two-class problems is first presented. It is an extension of the relation between the MSE and FDA [7, 6] to the case of the singular scatter matrices. From the relation between the MSE solution of (4.6) and the LDA/GSVD, the relationship between the kernel MSE solution of (4.9) and the KDA/GSVD solution is also derived.

The problem (4.6) can be solved by the normal equations

$$\begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} 1 & a_1^T \\ & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} \frac{n}{n_1} e_{n_1} \\ -\frac{n}{n_2} e_{n_2} \end{bmatrix} \tag{6.1}$$

where e_{n_i} is the $n_i \times 1$ column vector with only 1's as its components. From (6.1), we obtain

$$\begin{cases} nw_0 + (\sum_{1\le j\le n} a_j^T)w = 0 \\ (\sum_{1\le j\le n} a_j)w_0 + (\sum_{1\le j\le n} a_j a_j^T)w = \frac{n}{n_1}\sum_{j\in N_1} a_j - \frac{n}{n_2}\sum_{j\in N_2} a_j. \end{cases} \tag{6.2}$$

The first equation in (6.2) gives

$$w_0 = -c^T w$$

and by substituting it in the second equation and using the alternative expressions[3] of S_b and S_w

$$S_b = \sum_{1\le i\le 2} n_i c_i c_i^T - ncc^T \quad \text{and} \quad S_w = \sum_{1\le j\le n} a_j a_j^T - \sum_{1\le i\le 2} n_i c_i c_i^T, \tag{6.3}$$

we obtain

$$(S_b + S_w)w = n(c_1 - c_2). \tag{6.4}$$

On the other hand, since $\text{rank}(S_b) = 1$ in the two-class case we have

$$\eta_1 > \eta_2 = \cdots = \eta_m = 0 \quad \text{in} \quad \zeta_i S_b x_i = \eta_i S_w x_i$$

by LDA/GSVD. Since $\eta_1 + \zeta_1 = 1$ and $\eta_1 S_w x_1 = \zeta_1 S_b x_1$,

$$\eta_1(S_b + S_w)x_1 = (\eta_1 + \zeta_1)S_b x_1 = S_b x_1 = \frac{n_1 n_2}{n}(c_1 - c_2)(c_1 - c_2)^T x_1. \tag{6.5}$$

[3] In fact, S_b can also be represented as $S_b = \frac{n_1 n_2}{n}(c_1 - c_2)(c_1 - c_2)^T$ which will be used in Eq. (6.5) later.

Denoting

$$\rho = \eta_1 \frac{n^2}{n_1 n_2 (c_1 - c_2)^T x_1}, \tag{6.6}$$

Eq. (6.5) becomes

$$(S_b + S_w)\rho x_1 = n(c_1 - c_2). \tag{6.7}$$

Hence from Eqs. (6.4) and (6.7), we have

$$S_t w = (S_b + S_w)w = (S_b + S_w)\rho x_1 = S_t \rho x_1, \tag{6.8}$$

where

$$S_t = \sum_{i=1}^{n} (a_j - c)(a_j - c)^T = S_b + S_w$$

is the total scatter matrix.

Let the symmetric eigenvalue decomposition (EVD) of S_t be

$$S_t = Y \Sigma_t Y^T = \begin{bmatrix} \underbrace{Y_1}_{s} & \underbrace{Y_2}_{m-s} \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Y_1^T \\ Y_2^T \end{bmatrix}$$

where $s = \text{rank}(S_t)$, Y is orthogonal and Σ_1 is a diagonal matrix with nonincreasing positive diagonal elements. Then from (6.8),

$$Y_1 \Sigma_1 Y_1^T w = Y_1 \Sigma_1 Y_1^T \rho x_1 \quad \text{and} \quad Y_1^T w = Y_1^T \rho x_1.$$

PROPOSITION 6.1. *For any $x \in$ null$(S_b) \cap$ null(S_w), all data items are transformed to one point by the transformation x^T.*

Proof. For any $x \in \text{null}(S_b) \cap \text{null}(S_w)$, $x^T S_b x = x^T S_w x = 0$. Hence

$$0 = x^T S_b x = (x^T H_b)(H_b^T x) = \|x^T H_b\|^2 = \sum_{i=1}^{r} n_i |x^T c_i - x^T c|^2 \quad \text{and} \tag{6.9}$$

$$0 = x^T S_w x = \sum_{j=1}^{n} |x^T a_j - x^T c_i|^2 \quad \text{where } a_j \text{ belongs to the class } i. \tag{6.10}$$

Eqs. (6.9) and (6.10) give

$$\begin{cases} x^T c_i = x^T c \quad \text{for } i = 1, \cdots, r \\ x^T a_j = x^T c_i \quad \text{for all } j \text{ in } N_i \text{ and } i = 1, \cdots, r, \end{cases}$$

and these imply that all data items are transformed to one point by x^T. □

Since

$$\text{span}\{Y_2\} = \text{null}(S_t) \subset \text{null}(S_b) \cap \text{null}(S_w),$$

by Proposition 6.1, $Y_2^T (z - c) = 0$, and we obtain

$$\begin{aligned} w^T z + w_0 &= w^T (z - c) = w^T (Y_1 Y_1^T + Y_2 Y_2^T)(z - c) \\ &= w^T Y_1 Y_1^T (z - c) = \rho x_1^T Y_1 Y_1^T (z - c) \\ &= \rho x_1^T (Y_1 Y_1^T + Y_2 Y_2^T)(z - c) = \rho x_1^T (z - c). \end{aligned} \tag{6.11}$$

Eq. (6.11) gives the relation between the MSE solution and the solution of LDA/GSVD for two-class problem, which holds regardless of the singularity of the scatter matrices.

Now it should be straightforward to derive the corresponding relation between the kernel MSE solution and KDA/GSVD in two-class case. The formulation (4.9) for the kernel MSE solution is obtained by substituting the original data $[a_1, \cdots, a_n]$ in (4.6) with the kernel matrix $K = [\kappa(a_i, a_j)]_{(1 \le i \le n, 1 \le j \le n)}$ where each column of the kernel matrix K can be considered as a data item. On the other hand, as illustrated in Figure 3.1, KDA/GSVD solves the generalized eigenvalue problem

$$\mathcal{K}_b \mathcal{K}_b^T \alpha = \lambda \mathcal{K}_w \mathcal{K}_w^T \alpha.$$

$\mathcal{K}_b \mathcal{K}_b^T$ and $\mathcal{K}_w \mathcal{K}_w^T$ are the scatter matrices of the kernel matrix K when each column of the kernel matrix K is considered as a data item. Hence the relation between the kernel MSE and KDA/GSVD corresponding to (6.11) can be derived by substituting the original data a_i with $[\kappa(a_i, a_1), \cdots, \kappa(a_i, a_n)]^T$ in the proof of the relation between the MSE and LDA/GSVD.