When Fisher meets Fukunaga-Koontz: A New Look at Linear Discriminants

Sheng Zhang Terence Sim
School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543
{zhangshe, tsim}@comp.nus.edu.sg

Abstract

The Fisher Linear Discriminant (FLD) is commonly used in pattern recognition. It finds a linear subspace that maximally separates class patterns according to Fisher's Criterion. Several methods of computing the FLD have been proposed in the literature, most of which require the calculation of the so-called scatter matrices. In this paper, we bring a fresh perspective to FLD via the Fukunaga-Koontz Transform (FKT). We do this by decomposing the whole data space into four subspaces, and show where Fisher's Criterion is maximally satisfied. We prove the relationship between FLD and FKT analytically, and propose a method of computing the most discriminative subspace. This method is based on the QR decomposition, which works even when the scatter matrices are singular, or too large to be formed. Our method is general and may be applied to different pattern recognition problems. We validate our method by experimenting on synthetic and real data.

1. Introduction

In recent years, discriminant subspace analysis has been extensively studied in computer vision and pattern recognition. It has been widely used for feature extraction and dimension reduction in face recognition [1] and text classification [5]. One popular method is the Fisher Linear Discriminant (FLD), also known as Linear Discriminant Analysis (LDA) [2, 3]. It tries to find an optimal subspace such that the separability of two classes is maximized. This is achieved by minimizing the within-class distance and maximizing the between-class distance simultaneously. To be more specific, in terms of the between-class scatter matrix $\mathbf{S}_b$ and the within-class scatter matrix $\mathbf{S}_w$, the Fisher's Criterion can be written as

$$J_F(\Phi) = \frac{|\Phi^T \mathbf{S}_b \Phi|}{|\Phi^T \mathbf{S}_w \Phi|} \tag{1}$$

By maximizing the criterion J_F, Fisher Linear Discriminant finds the subspaces in which the classes are most linearly separable. The solution [3] that maximizes J_F is a set of the eigenvectors $\{\phi_i\}$ which must satisfy

$$\mathbf{S}_b \phi = \lambda \mathbf{S}_w \phi. \tag{2}$$

This is called the generalized eigenvalue problem [2, 3]. The discriminant subspace is spanned by the generalized eigenvectors. The discriminability of each eigenvector is measured by the corresponding generalized eigenvalue, i.e., the most discriminant subspace corresponds to the maximal generalized eigenvalue. The generalized eigenvalue problem can be solved by matrix inversion and eigen-decomposition, i.e., applying the eigen-decomposition on $\mathbf{S}_w^{-1}\mathbf{S}_b$. Unfortunately, for many applications with high dimensional data and few training samples, such as face recognition, the scatter matrix $\mathbf{S}_w$ is singular because generally the dimension of sample data is greater than the number of samples. This is known as the undersampled or small sample size problem [3, 2].

Up till now, a great number of methods have been proposed to circumvent the requirement of nonsingularity of $\mathbf{S}_w$, such as Fisherface [1] and LDA/GSVD [5]. In [1], Fisherface first applies PCA [6, 8] to reduce dimension such that $\mathbf{S}_w$ is nonsingular, then followed by LDA. The LDA/GSVD algorithm [5] avoids the inversion of $\mathbf{S}_w$ by the simultaneous diagonalization via the Generalized Singular Value Decomposition (GSVD).

However, these methods are designed to overcome the singularity problem and do not directly relate to the generalized eigenvalue, the essential measure of the discriminability. In this paper, we propose to apply Fukunaga Koontz Transform (FKT) [3] on the LDA problem. Based on the eigenvalue ratio of FKT, we decompose the whole data space into four subspaces. Then our theoretical analyses show the relationship between LDA, FKT and GSVD.

Our work has three main contributions:

1. We present a unifying framework to understand differ-

ent methods, namely, LDA, FKT and GSVD. To be more specific, for the LDA problem, GSVD is equivalent to FKT; and the generalized eigenvalue of LDA is equal to both the eigenvalue ratio of FKT and the square of the generalized singular value of GSVD.

2. The proposed theory is useful for pattern recognition. Our theoretical analyses demonstrate how to choose the best subspaces for maximum discriminability. The experiments on synthetic data and real data validate our theory.

3. In connecting FKT with LDA, we show how FKT, originally meant for 2-class problems can be generalized to handle multiple classes.

The rest of this paper is organized as follows: Section 2 briefly reviews the mathematical background for LDA, GSVD, and FKT. In Section 3, we first analyze the discriminant subspace of LDA based on FKT, then set up the connections between LDA, FKT and GSVD. We apply our theory to the classification problem on synthetic and real data in Section 4, and conclude our paper in Section 5.

2. Mathematical Background

Notations. Let $\mathbf{A} = \{\mathbf{a}_1, ..., \mathbf{a}_N\}$, $\mathbf{a}_i \in \mathbb{R}^D$ denote a data set of given D-dimensional vectors. Each data point belongs to exactly one of C object classes $\{L_1, ..., L_C\}$. The number of vectors in class L_i is denoted by N_i, thus $N = \sum N_i$. Observe that for high-dimensional data, e.g., face images, generally, $C \leq N \ll D$. The between-class scatter matrix $\mathbf{S}_b$, the within-class scatter matrix $\mathbf{S}_w$, and the total scatter matrix $\mathbf{S}_t$ are defined as follows:

$$\mathbf{S}_b = \sum_{i=1}^{C} N_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T = \mathbf{H}_b\mathbf{H}_b^T \quad (3)$$

$$\mathbf{S}_w = \sum_{i=1}^{C} \sum_{\mathbf{a}\in L_i} (\mathbf{a} - \mathbf{m}_i)(\mathbf{a} - \mathbf{m}_i)^T = \mathbf{H}_w\mathbf{H}_w^T \quad (4)$$

$$\mathbf{S}_t = \sum_{i=1}^{N} (\mathbf{a}_i - \mathbf{m})(\mathbf{a}_i - \mathbf{m})^T = \mathbf{H}_t\mathbf{H}_t^T \quad (5)$$

$$\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w \quad (6)$$

Here $\mathbf{m}_i$ denotes the class mean and $\mathbf{m}$ is the global mean of $\mathbf{A}$. The matrices $\mathbf{H}_b \in \mathbb{R}^{D\times C}$, $\mathbf{H}_w \in \mathbb{R}^{D\times N}$, and $\mathbf{H}_t \in \mathbb{R}^{D\times N}$ are the *precursor* matrices of the between-class scatter matrix, the within-class scatter matrix and the total scatter matrix respectively. Let us denote the rank of each scatter matrix: $r_w = Rank(\mathbf{S}_w)$, $r_b = Rank(\mathbf{S}_b)$, and $r_t = Rank(\mathbf{S}_t)$. Note that for high-dimensional data ($D \gg N$), $r_b \leq C - 1$, $r_w \leq N - 1$ and $r_t \leq N - 1$.

2.1. Linear Discriminant Analysis

Given the data matrix $\mathbf{A}$, which can be divided into C classes, we try to find a linear transformation matrix $\Phi \in \mathbb{R}^{D\times d}$, where $d < D$. It maps high dimensional data to a low dimensional space. From the perspective of pattern classification, LDA aims to find the optimal transformation Φ such that the projected data are well separated.

Usually, two types of criteria are used to measure the separability of classes [3]. One type gives the upper bound on the Bayes error, e.g., Bhattacharyya distance. The other type is based on scatter matrices. As shown in Equation 1, Fisher's criterion belongs to the latter type. The solution of the criterion is the generalized eigenvector and eigenvalue (See Equation 2). However, as we mentioned, if $\mathbf{S}_w$ is nonsingular it can be solved by the generalized eigendecomposition: $\mathbf{S}_w^{-1}\mathbf{S}_b\phi = \lambda\phi$. Otherwise, $\mathbf{S}_w$ is singular and we have to circumvent the nonsingularity requirement via LDA/GSVD [5], for example.

2.2. Generalized SVD

The Generalized Singular Value Decomposition (GSVD) was developed by Van Loan et al. [4]. We will briefly review the mechanism of GSVD using LDA as an example.

Howland et. al. [5] extended the applicability of LDA to the case when $\mathbf{S}_w$ is singular. This is done by using the simultaneous diagonalization of the scatter matrices via GSVD [4]. First, to reduce computation load, $\mathbf{H}_b$ and $\mathbf{H}_w$ are used instead of $\mathbf{S}_b$ and $\mathbf{S}_w$. Then, based on GSVD there exist orthogonal matrices $\mathbf{Y} \in \mathbb{R}^{C\times C}$, $\mathbf{Z} \in \mathbb{R}^{N\times N}$, and a nonsingular matrix $\mathbf{X} \in \mathbb{R}^{D\times D}$ such that:

$$\mathbf{Y}^T\mathbf{H}_b^T\mathbf{X} = [\Sigma_b, 0], \quad (7)$$

$$\mathbf{Z}^T\mathbf{H}_w^T\mathbf{X} = [\Sigma_w, 0]. \quad (8)$$

where

$$\Sigma_b = \begin{bmatrix} \mathbf{I}_b & & \\ & \mathbf{D}_b & \\ & & \mathbf{O}_b \end{bmatrix}, \Sigma_w = \begin{bmatrix} \mathbf{O}_w & & \\ & \mathbf{D}_w & \\ & & \mathbf{I}_w \end{bmatrix}.$$

Since $\mathbf{Y}$, $\mathbf{Z}$ are orthogonal matrices and $\mathbf{X}$ is nonsingular, from Equations 7, 8 we obtain

$$\mathbf{H}_b^T = \mathbf{Y}\,[\Sigma_b, 0]\,\mathbf{X}^{-1}, \quad (9)$$

$$\mathbf{H}_w^T = \mathbf{Z}\,[\Sigma_w, 0]\,\mathbf{X}^{-1}. \quad (10)$$

The matrices $\mathbf{I}_b \in \mathbb{R}^{(r_t-r_w)\times(r_t-r_w)}$ and $\mathbf{I}_w \in \mathbb{R}^{(r_t-r_b)\times(r_t-r_b)}$ are identity matrices, and $\mathbf{O}_b \in \mathbb{R}^{(C-r_b)\times(r_t-r_b)}$ and $\mathbf{O}_w \in \mathbb{R}^{(N-r_w)\times(r_t-r_w)}$ are rectangle, zero matrices which may have no rows or no columns, and $\mathbf{D}_b$=diag$(\alpha_{r_t-r_w+1}, ..., \alpha_{r_b})$ and $\mathbf{D}_w$=diag$(\beta_{r_t-r_w+1}, ..., \beta_{r_b})$ satisfy $1 > \alpha_{r_t-r_w+1} \geq ... \geq$

$\alpha_{r_b} > 0, 0 < \beta_{r_t - r_w + 1} \leq ... \leq \beta_{r_b} < 1$, and $\alpha_i^2 + \beta_i^2 = 1$. Thus,

$$\Sigma_b^T \Sigma_b + \Sigma_w^T \Sigma_w = \mathbf{I}, \tag{11}$$

where $\mathbf{I} \in \mathbb{R}^{r_t \times r_t}$ is an identity matrix. The columns of $\mathbf{X}$, the generalized singular vectors for the matrix pair $(\mathbf{H}_b, \mathbf{H}_w)$, can be used as the discriminant feature subspace based on GSVD.

2.3. Fukunaga Koontz Transform

The FKT was designed for the 2-class recognition problem. Given data matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ from two classes, the autocorrelation matrices $\mathbf{S}_1 = \mathbf{A}_1\mathbf{A}_1^T$ and $\mathbf{S}_2 = \mathbf{A}_2\mathbf{A}_2^T$ are positive semi-definite (p.s.d.) and symmetric. The sum of these two matrices is still p.s.d. and symmetric and can be factorized in the form

$$\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2 = [\mathbf{U}, \mathbf{U}_\perp] \begin{bmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{U}^T \\ \mathbf{U}_\perp^T \end{bmatrix} \tag{12}$$

Without loss of generality, $\mathbf{S}$ may be singular and r=Rank(**S**)$< D$, thus $\mathbf{D}$ = diag$\{\lambda_1, ..., \lambda_r\}$, $\lambda_1 \geq ... \geq \lambda_r > 0$. $\mathbf{U} \in \mathbb{R}^{D \times r}$ is the set of eigenvectors corresponding to nonzero eigenvalues and $\mathbf{U}_\perp \in \mathbb{R}^{D \times (D-r)}$ is the orthogonal complement of $\mathbf{U}$. Now we can whiten $\mathbf{S}$ by a transformation operator $\mathbf{P} = \mathbf{U}\mathbf{D}^{-1/2}$. The sum of the two matrices $\mathbf{S}_1, \mathbf{S}_2$ becomes

$$\mathbf{P}^T\mathbf{S}\mathbf{P} = \mathbf{P}^T(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{P} = \widetilde{\mathbf{S}}_1 + \widetilde{\mathbf{S}}_2 = \mathbf{I} \tag{13}$$

where $\widetilde{\mathbf{S}}_1 = \mathbf{P}^T\mathbf{S}_1\mathbf{P}$ and $\widetilde{\mathbf{S}}_2 = \mathbf{P}^T\mathbf{S}_2\mathbf{P}$, $\mathbf{I} \in \mathbb{R}^{r \times r}$ is an identity matrix. Suppose an eigenvector of $\widetilde{\mathbf{S}}_1$ is $\mathbf{v}$ with eigenvalue λ_1, that is: $\widetilde{\mathbf{S}}_1\mathbf{v} = \lambda_1\mathbf{v}$. Since $\widetilde{\mathbf{S}}_1 = \mathbf{I} - \widetilde{\mathbf{S}}_2$, we can rewrite it as:

$$(\mathbf{I} - \widetilde{\mathbf{S}}_2)\mathbf{v} = \lambda_1\mathbf{v} \tag{14}$$

$$\widetilde{\mathbf{S}}_2\mathbf{v} = (1 - \lambda_1)\mathbf{v} \tag{15}$$

This means that $\widetilde{\mathbf{S}}_2$ has the same eigenvector as $\widetilde{\mathbf{S}}_1$ but the corresponding eigenvalue is $\lambda_2 = 1 - \lambda_1$. Consequently, the dominant eigenvector of $\widetilde{\mathbf{S}}_1$ is the weakest eigenvector of $\widetilde{\mathbf{S}}_2$, and vice versa.

3. Theory

In this section, we first employ FKT on $\mathbf{S}_b$ and $\mathbf{S}_w$[1], which results in decomposing the whole data space $\mathbf{S}_t$ into four subspaces. Then we explain the relationship between LDA, FKT and GSVD based on the generalized eigenvalue. This gives insight into the different discriminant subspace analyses.

[1]In this paper, we focus on linear discriminant subspace analysis, but our approach can be easily extended to nonlinear discriminant subspace analysis. For example, based on kernel method, we can apply FKT on kernelized $\mathbf{S}_b$ and kernelized $\mathbf{S}_w$.

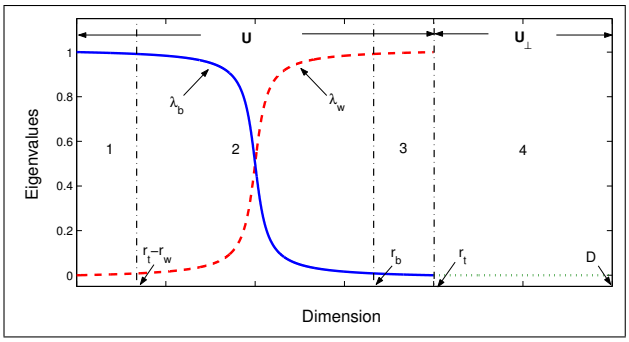

Figure 1. The whole data space is decomposed into four subspaces via FKT. In $\mathbf{U}_\perp$, the null space of $\mathbf{S}_t$, there is no discriminant information. In $\mathbf{U}$, the sum of $\lambda_b + \lambda_w$ is equal to 1. Note that we represent all possible subspaces, but in practice, some of these subspaces may not exist.

3.1. LDA/FKT

Generally speaking, for the LDA problem there are more than 2 classes. To handle multiple classes, we replace the autocorrelation matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ with the scatter matrices $\mathbf{S}_b$ and $\mathbf{S}_w$. Since $\mathbf{S}_b$, $\mathbf{S}_w$ and $\mathbf{S}_t$ are p.s.d., symmetric and $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$, we can apply FKT on $\mathbf{S}_b$, $\mathbf{S}_w$ and $\mathbf{S}_t$, which we shall henceforth call LDA/FKT. The whole data space is decomposed into two subspaces: $\mathbf{U}$ and $\mathbf{U}_\perp$ (See Fig. 1). On the one hand, $\mathbf{U}_\perp$ is the set of eigenvectors corresponding to the zero eigenvalues of $\mathbf{S}_t$. It is the intersection of the null spaces of $\mathbf{S}_b$ and $\mathbf{S}_w$, and contains no discriminant information. On the other hand, $\mathbf{U}$ is the set of eigenvectors corresponding to the nonzero eigenvalues of $\mathbf{S}_t$. It contains discriminant information.

Based on FKT, $\widetilde{\mathbf{S}}_b = \mathbf{P}^T\mathbf{S}_b\mathbf{P}$ and $\widetilde{\mathbf{S}}_w = \mathbf{P}^T\mathbf{S}_w\mathbf{P}$ share the same eigenvectors, and the sum of two eigenvalues corresponding to the same eigenvector is equal to 1.

$$\widetilde{\mathbf{S}}_b = \mathbf{V}\Lambda_b\mathbf{V}^T \tag{16}$$

$$\widetilde{\mathbf{S}}_w = \mathbf{V}\Lambda_w\mathbf{V}^T \tag{17}$$

$$\mathbf{I} = \Lambda_b + \Lambda_w \tag{18}$$

Where $\mathbf{V} \in \mathbb{R}^{r_t \times r_t}$ is the orthogonal eigenvector matrix, $\Lambda_b, \Lambda_w \in \mathbb{R}^{r_t \times r_t}$ are diagonal eigenvalue matrices. According to the eigenvalue ratio $\frac{\lambda_b}{\lambda_w}$, $\mathbf{U}$ can be further decomposed into three subspaces. To keep the integrity of the whole data space, we incorporate $\mathbf{U}_\perp$ as the fourth subspace. See Fig. 1.

1. *Subspace 1*: the span of eigenvectors $\{\mathbf{v}_i\}$ corresponding to $\lambda_w = 0$, $\lambda_b = 1$. Since $\frac{\lambda_b}{\lambda_w} = \infty$, in this subspace, the eigenvalue ratio is maximized.

2. *Subspace 2*: the span of eigenvectors $\{\mathbf{v}_i\}$ corresponding to $0 < \lambda_w < 1$ and $0 < \lambda_b < 1$. Since $0 < \frac{\lambda_b}{\lambda_w} < \infty$, the eigenvalue ratio is smaller than that of Subspace 1.

3. *Subspace 3*: the span of eigenvectors $\{\mathbf{v}_i\}$ corresponding to $\lambda_w = 1$ and $\lambda_b = 0$. Since $\frac{\lambda_b}{\lambda_w} = 0$, the eigenvalue ratio is minimal.

4. *Subspace 4*: $\mathbf{U}_\perp$, the span of eigenvectors corresponding to the zero eigenvalues of $\mathbf{S}_t$.

Note that in practice, any of these four subspaces may not exist, depending on the ranks of $\mathbf{S}_b$, $\mathbf{S}_w$ and $\mathbf{S}_t$. For example, if $r_t = r_w + r_b$, then Subspace 2 does not exist. As illustrated in Fig. 1, the null space of $\mathbf{S}_w$ is the union of Subspace 1 and Subspace 4, and the null space of $\mathbf{S}_b$ is the union of Subspace 3 and Subspace 4, if they exist.

3.2. Relationship between LDA, GSVD and FKT

How do these four subspaces help to maximize the Fisher criterion J_F? We explain this in the following Theorem that connects the generalized eigenvalue of J_F to the eigenvalues of FKT. We begin with a Lemma:

Lemma. *For the LDA problem, GSVD is equivalent to FKT, with* $\mathbf{X} = \left[\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}, \mathbf{U}_\perp\right]$, $\Lambda_b = \Sigma_b^T\Sigma_b$ *and* $\Lambda_w = \Sigma_w^T\Sigma_w$. *Where* $\mathbf{X}$, Σ_b *and* Σ_w *are from GSVD (See Equations 7, 8), and* $\mathbf{U}$, $\mathbf{D}$, $\mathbf{V}$, $\mathbf{U}_\perp$, Λ_w *and* Λ_b *are matrices from FKT (See Equations 16, 17, 18).*

Proof. **GSVD $\Longrightarrow$ FKT**
Based on GSVD,

$$\mathbf{S}_b = \mathbf{H}_b\mathbf{H}_b^T = \mathbf{X}^{-T}\begin{bmatrix}\Sigma_b^T\Sigma_b & 0\\ 0 & 0\end{bmatrix}\mathbf{X}^{-1} \quad (19)$$

$$\mathbf{S}_w = \mathbf{H}_w\mathbf{H}_w^T = \mathbf{X}^{-T}\begin{bmatrix}\Sigma_w^T\Sigma_w & 0\\ 0 & 0\end{bmatrix}\mathbf{X}^{-1} \quad (20)$$

Thus,

$$\mathbf{X}^T\left(\mathbf{S}_b + \mathbf{S}_w\right)\mathbf{X} = \begin{bmatrix}\mathbf{I} & 0\\ 0 & 0\end{bmatrix} \quad (21)$$

Since $\Sigma_b^T\Sigma_b + \Sigma_w^T\Sigma_w = \mathbf{I} \in R^{r_t\times r_t}$, if we choose the first r_t columns of $\mathbf{X}$ as $\mathbf{P}$, i.e., $\mathbf{P} = \mathbf{X}_{(D,r_t)}$, then $\mathbf{P}^T(\mathbf{S}_b + \mathbf{S}_w)\mathbf{P} = \mathbf{I}$. This is exactly the outcome of FKT. Meanwhile, we can reach the conclusion that $\Lambda_b = \Sigma_b^T\Sigma_b$ and $\Lambda_w = \Sigma_w^T\Sigma_w$.

FKT $\Longrightarrow$ GSVD
Based on FKT, $\mathbf{P} = \mathbf{U}\mathbf{D}^{-1/2}$

$$\widetilde{\mathbf{S}}_b = \mathbf{P}^T\mathbf{S}_b\mathbf{P} = \mathbf{D}^{-1/2}\mathbf{U}^T\mathbf{H}_b\mathbf{H}_b^T\mathbf{U}\mathbf{D}^{-1/2} \quad (22)$$

$$\widetilde{\mathbf{S}}_b = \mathbf{V}\Lambda_b\mathbf{V}^T \quad (23)$$

Hence,

$$\mathbf{D}^{-1/2}\mathbf{U}^T\mathbf{H}_b\mathbf{H}_b^T\mathbf{U}\mathbf{D}^{-1/2} = \mathbf{V}\Lambda_b\mathbf{V}^T \quad (24)$$

In general, there is no unique decomposition on the above equation because $\mathbf{H}_b\mathbf{H}_b^T = \mathbf{H}_b\mathbf{Y}\mathbf{Y}^T\mathbf{H}_b^T$ for any orthogonal matrix $\mathbf{Y} \in \mathbb{R}^{C\times C}$. That is:

$$\mathbf{D}^{-1/2}\mathbf{U}^T\mathbf{H}_b\mathbf{Y}\mathbf{Y}^T\mathbf{H}_b^T\mathbf{U}\mathbf{D}^{-1/2} = \mathbf{V}\Lambda_b\mathbf{V}^T \quad (25)$$

$$\mathbf{Y}^T\mathbf{H}_b^T\mathbf{U}\mathbf{D}^{-1/2} = \widehat{\Sigma}_b\mathbf{V}^T \quad (26)$$

$$\mathbf{Y}^T\mathbf{H}_b^T\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V} = \widehat{\Sigma}_b \quad (27)$$

where $\widehat{\Sigma}_b \in \mathbb{R}^{C\times r_t}$, and $\Lambda_b = \widehat{\Sigma}_b^T\widehat{\Sigma}_b$. If we define $\mathbf{X} = \left[\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}, \mathbf{U}_\perp\right] \in \mathbb{R}^{D\times D}$. Then,

$$\mathbf{Y}^T\mathbf{H}_b^T\mathbf{X} = \mathbf{Y}^T\mathbf{H}_b^T[\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}, \mathbf{U}_\perp] \quad (28)$$

$$= [\mathbf{Y}^T\mathbf{H}_b^T\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}, 0] \quad (29)$$

$$= [\widehat{\Sigma}_b, 0] \quad (30)$$

Here, $\mathbf{H}_b^T\mathbf{U}_\perp = 0$ and $\mathbf{H}_w^T\mathbf{U}_\perp = 0$ because $\mathbf{U}_\perp$ is the intersection of the null spaces of $\mathbf{S}_b$ and $\mathbf{S}_w$. Similarly, we can obtain $\mathbf{Z}^T\mathbf{H}_w^T\mathbf{X} = [\widehat{\Sigma}_w, 0]$, where $\mathbf{Z} \in \mathbb{R}^{N\times N}$ is an arbitrary orthogonal matrix and $\widehat{\Sigma}_w \in \mathbb{R}^{N\times r_t}$, and $\Lambda_w = \widehat{\Sigma}_w^T\widehat{\Sigma}_w$. Since $\Lambda_b + \Lambda_w = \mathbf{I}$ and $\mathbf{I} \in \mathbb{R}^{r_t\times r_t}$ is an identity matrix, it is easy to check that $\widehat{\Sigma}_b^T\widehat{\Sigma}_b + \widehat{\Sigma}_w^T\widehat{\Sigma}_w = \mathbf{I}$, which satisfies the constraint of GSVD.

Now we have to prove $\mathbf{X}$ is nonsingular.

$$\mathbf{X}\mathbf{X}^T = \left[\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}, \mathbf{U}_\perp\right]\begin{bmatrix}\mathbf{V}^T\mathbf{D}^{-1/2}\mathbf{U}^T\\ \mathbf{U}_\perp^T\end{bmatrix}$$

$$= \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^T + \mathbf{U}_\perp\mathbf{U}_\perp^T$$

$$= [\mathbf{U}, \mathbf{U}_\perp]\begin{bmatrix}\mathbf{D}^{-1} & 0\\ 0 & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{U}^T\\ \mathbf{U}_\perp^T\end{bmatrix} \quad (31)$$

where $\mathbf{V} \in \mathbb{R}^{r\times r}$ and $[\mathbf{U}, \mathbf{U}_\perp]$ are orthogonal matrices. Note that $\mathbf{U}^T\mathbf{U}_\perp = 0$ and $\mathbf{U}_\perp^T\mathbf{U} = 0$. Thus $\mathbf{X}\mathbf{X}^T$ can be eigen-decomposed with positive eigenvalues, which means $\mathbf{X}$ is nonsingular. □

Based on the above lemma, we can investigate the relationship between the eigenvalue ratio of FKT and the generalized eigenvalue λ of the Fisher criterion J_F.

Theorem. *If* λ *is the solution of Equation 2 (the generalized eigenvalue of* $\mathbf{S}_b$ *and* $\mathbf{S}_w$*), and* λ_b *and* λ_w *are the eigenvalues after applying FKT on* $\mathbf{S}_b$ *and* $\mathbf{S}_w$*, then* $\lambda = \frac{\lambda_b}{\lambda_w}$*, where* $\lambda_b + \lambda_w = 1$.

Proof. Based on GSVD, it is easy to verify that:

$$\mathbf{S}_b = \mathbf{H}_b\mathbf{H}_b^T = \mathbf{X}^{-T}\begin{bmatrix}\Sigma_b^T\Sigma_b & 0\\ 0 & 0\end{bmatrix}\mathbf{X}^{-1} \quad (32)$$

According to our lemma, $\Lambda_b = \Sigma_b^T \Sigma_b$, thus

$$\mathbf{S}_b = \mathbf{X}^{-T} \begin{bmatrix} \Lambda_b & 0 \\ 0 & 0 \end{bmatrix} \mathbf{X}^{-1} \tag{33}$$

Similarly, we can rewrite $\mathbf{S}_w$ but with Λ_w. Since $\mathbf{S}_b\phi = \lambda \mathbf{S}_w\phi$,

$$\mathbf{X}^{-T} \begin{bmatrix} \Lambda_b & 0 \\ 0 & 0 \end{bmatrix} \mathbf{X}^{-1}\phi = \lambda \mathbf{X}^{-T} \begin{bmatrix} \Lambda_w & 0 \\ 0 & 0 \end{bmatrix} \mathbf{X}^{-1}\phi \tag{34}$$

Let $\mathbf{v} = \mathbf{X}^{-1}\phi$, and multiply $\mathbf{X}^T$ on both sides, we can obtain the following.

$$\begin{bmatrix} \Lambda_b & 0 \\ 0 & 0 \end{bmatrix} \mathbf{v} = \lambda \begin{bmatrix} \Lambda_w & 0 \\ 0 & 0 \end{bmatrix} \mathbf{v} \tag{35}$$

If we add $\lambda \begin{bmatrix} \Lambda_b & 0 \\ 0 & 0 \end{bmatrix} \mathbf{v}$ on both sides of Equation 35, then

$$(1+\lambda) \begin{bmatrix} \Lambda_b & 0 \\ 0 & 0 \end{bmatrix} \mathbf{v} = \lambda \begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix} \mathbf{v}. \tag{36}$$

This means that $(1+\lambda)\lambda_b = \lambda$, which can be rewritten as $\lambda_b = \lambda(1-\lambda_b) = \lambda\lambda_w$ because $\lambda_b + \lambda_w = 1$. Now, we can observe that $\lambda = \frac{\lambda_b}{\lambda_w}$. □

Corollary. *If λ is the generalized eigenvalue of $\mathbf{S}_b$ and $\mathbf{S}_w$, and α and β are the solutions of Equations 7, 8 and α/β is the generalized singular value of the matrix pair $(\mathbf{H}_b, \mathbf{H}_w)$, then $\lambda = \frac{\alpha^2}{\beta^2}$, where $\alpha^2 + \beta^2 = 1$.*

Proof. In Lemma we have proved that $\Lambda_b = \Sigma_b^T\Sigma_b$ and $\Lambda_w = \Sigma_w^T\Sigma_w$, that is: $\lambda_b = \alpha^2$ and $\lambda_w = \beta^2$. According to Theorem, we observe that $\lambda = \frac{\lambda_b}{\lambda_w}$. Therefore it is easy to see that $\lambda = \frac{\alpha^2}{\beta^2}$. Note that $\frac{\alpha}{\beta}$ is the generalized singular value of $(\mathbf{H}_b, \mathbf{H}_w)$ by GSVD, and λ is the generalized eigenvalue of $(\mathbf{S}_b, \mathbf{S}_w)$. □

The Corollary suggests how to evaluate the discriminant subspaces of LDA/GSVD. Actually, Howland et. al. in [5] applied the Corollary implicitly, but in this paper we explicitly connect the generalized singular value $\frac{\alpha}{\beta}$ with λ, the measure of discriminability.

Based on our analysis, the eigenvalue ratio $\frac{\lambda_b}{\lambda_w}$ and the square of the generalized singular value $\frac{\alpha^2}{\beta^2}$, both are equal to the generalized eigenvalue λ, the measure of discriminability. Therefore, according to Fig. 1, Subspace 1, with the infinite eigenvalue ratio $\frac{\lambda_b}{\lambda_w}$, is the most discriminant subspace, followed by Subspace 2 and Subspace 3. However, Subspace 4 which contains no discriminant information, can be safely thrown away.

Input: The data matrix $\mathbf{A}$.
Output: Projection matrix Φ_F such that the J_F is maximized.

1. Compute $\mathbf{H}_b$ and $\mathbf{H}_t$ from data matrix $\mathbf{A}$:
$$\mathbf{H}_b = \left[\sqrt{N_1}(\mathbf{m}_1 - \mathbf{m}), ..., \sqrt{N_C}(\mathbf{m}_C - \mathbf{m})\right],$$
$$\mathbf{H}_t = [\mathbf{a}_1 - \mathbf{m}, ..., \mathbf{a}_N - \mathbf{m}].$$
2. Apply QR decomposition on $\mathbf{H}_t = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{R}^{D \times r_t}, \mathbf{R} \in \mathbb{R}^{r_t \times N}$ and $r_t = Rank(\mathbf{H}_t)$.
3. Let $\widetilde{\mathbf{S}}_t = \mathbf{R}\mathbf{R}^T$, since $\widetilde{\mathbf{S}}_t = \mathbf{Q}^T\mathbf{S}_t\mathbf{Q} = \mathbf{Q}^T\mathbf{H}_t\mathbf{H}_t^T\mathbf{Q} = \mathbf{R}\mathbf{R}^T$.
4. Let $\mathbf{Z} = \mathbf{Q}^T\mathbf{H}_b$.
5. Let $\widetilde{\mathbf{S}}_b = \mathbf{Z}\mathbf{Z}^T$, since $\widetilde{\mathbf{S}}_b = \mathbf{Q}^T\mathbf{S}_b\mathbf{Q} = \mathbf{Q}^T\mathbf{H}_b\mathbf{H}_b^T\mathbf{Q} = \mathbf{Z}\mathbf{Z}^T$.
6. Compute the eigenvectors $\{\mathbf{v}_i\}$ and eigenvalues $\{\lambda_i\}$ of $\widetilde{\mathbf{S}}_t^{-1}\widetilde{\mathbf{S}}_b$.
7. Sort the eigenvectors $\mathbf{v}_i$ according to λ_i in decreasing order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > \lambda_{k+1} = \cdots = \lambda_{r_t} = 0$.
8. The final projection matrix $\Phi_F = \mathbf{QV}$, where $\mathbf{V} = \{\mathbf{v}_i\}$. Note that we can choose the subspaces based on the columns of $\mathbf{V}$.

Figure 2. Algorithm: Apply QR decomposition to compute LDA/FKT.

3.3 Algorithm

Although we proved that FKT is equivalent to GSVD on the LDA problem, LDA/GSVD is computationally expensive. The running time of LDA/GSVD is $O(D(N+C)^2)$, where D is the dimensionality, N is the number of training samples and C is the number of classes. In this paper, we propose an efficient way to compute the Subspaces 1, 2 and 3 of LDA/FKT based on QR decomposition because Subspace 4 contains no discriminant information. Moreover, we use smaller matrices $\mathbf{H}_b$ and $\mathbf{H}_t$ because matrices $\mathbf{S}_b$, $\mathbf{S}_w$ and $\mathbf{S}_t$ may be too large to be formed. Our LDA/FKT algorithm is shown in Fig. 2. The running time is $O(DN^2)$.

4. Experiments

4.1. A Toy Problem

First, we experiment on synthetic data: three single Gaussian classes with the same covariance matrix and different means. In 3D space, the three classes share the same covariance matrix: $diag([1,1,0])$, and each class has 10 points. They have different means: $[0,0,0]^T$, $[0,1,2]^T$, and $[0,1,0]^T$ (See Fig. 3(a)).

Table 1. Eigenvalue and eigenvalue ratio of the toy problem.

LDA/FKT λ_b/λ_w	LDA/GSVD α^2/β^2	LDA λ
$1/0 = \infty$	$(1.6709 \times 10^{16})^2 \rightarrow \infty$	∞
$0.29286/0.70714$ $= 0.4141$	0.64354^2 $= 0.4141$	0.4141
$0/1 = 0$	$(1.0616 \times 10^{-16})^2 \rightarrow 0$	0

As shown in Fig. 4, LDA/FKT decomposes the whole data space into three subspaces: 1, 2 and 3. Here Subspace 4 does not exist because the number of samples (30) is larger than the dimension (3). Note that each subspace consists of only one eigenvector. In terms of the eigenvalue ratio, Subspace 1 is the most discriminative subspace, followed by Subspaces 2 and 3. But if we use Fisherface (PCA followed by LDA), we cannot obtain Subspace 1 because Fisherface restricts to the Subspaces 2 and 3 where $\lambda_w \neq 0$ so that $\mathbf{S}_w$ is invertible. In doing so, Fisherface is discarding the most discriminative information (Subspace 1). To compare LDA/FKT with Fisherface, we project the original 3D data to Subspace 1 and Subspace 2, and we also project to 2D space via Fisherface. As illustrated in Fig. 3(b) and 3(c), the 2D projection of FKT is more separable than the projection of Fisherface. This is consistent with our theory.

Let us recall the Corollary: the eigenvalue ratio of LDA/FKT is equal to the square of the generalized singular value of LDA/GSVD. To check this, we also apply GSVD on $\mathbf{H}_b$ and $\mathbf{H}_w$, which is known as LDA/GSVD [5]. Then we obtain the generalized singular value α/β. From Table 1, we see that $\alpha^2/\beta^2 \simeq \lambda_b/\lambda_w$[2]. Thus we show the validity of our theory experimentally. Moreover, this means we can compute the generalized eigenvalue λ of LDA, although it cannot be obtained directly when $\mathbf{S}_w$ is singular.

4.2. Face Recognition

We now apply LDA/FKT on a real problem: face recognition. We perform experiments on the CMU-PIE [7] face dataset. The CMU-PIE database consists of over $400,000$ images of 68 subjects. Each subject was recorded across 13 different poses, under 43 different illuminations (with and without background lighting), and with 4 different expression. Here we use a subset of PIE for our experiments. We choose 67 subjects [3], and each subject has 24 frontal face images with background lighting. All of these face images are aligned based on eyes coordinates, and cropped to 70×80. Fig. 5 shows a sample of PIE face images used in our experiments. It is easy to see that the major challenge on this data set is to do face recognition under different illuminations.

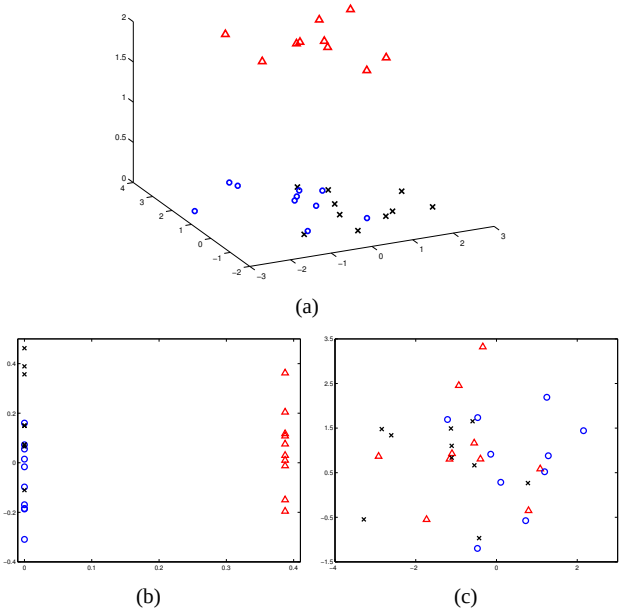

Figure 3. Original 3 classes (triangle, circle and cross) in 3D space (a) and 2D projection by (b) LDA/FKT and (c) Fisherface. Observe that the 2D projection of FKT is more separable than that of Fisherface.

Here, to evaluate the performance of LDA/FKT, we compare it with PCA and Fisherface. We employ 1-NN to do face recognition after projection onto the respective subspaces. For PCA, the projected subspace contains 66 dimensions; for Fisherface, we take the first 100 principal components to make $\mathbf{S}_w$ nonsingular, then followed by LDA. For LDA/FKT, we project data to Subspaces 1, 2.

In face recognition, we usually have an undersampled problem, which is also the reason for the singularity of $\mathbf{S}_w$. To evaluate the performance under such situation, we randomly choose n training samples from each subject, $n = 2, \cdots, 12$, and the remaining images are used for testing. Moreover, for each set of n training samples, we repeat sampling 10 times to compute the mean and standard deviation of classification accuracies. As illustrated in Fig. 6, we observe that the more the training samples, the better the recognition accuracy. To be more specific, for each method, increasing the number of training samples increases the mean recognition rate and decreases the standard deviation.

Fig. 6 also shows that no matter how many training sam-

[2] The difference depends on the machine precision.

[3] In CMU-PIE, there are 68 subjects together, we only choose 67 subjects because one subject has fewer frontal face images.

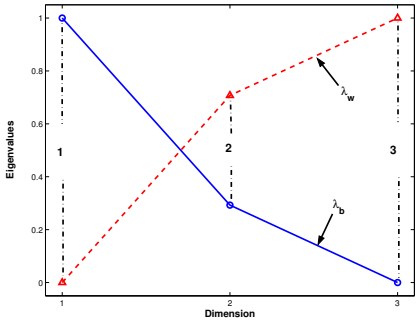

Figure 4. Eigenvalue curve for the toy problem by LDA/FKT. Note that each subspace consists of only one eigenvector.

Figure 5. A sample of face images from PIE dataset. Each subject has 24 frontal face images under different lighting conditions.

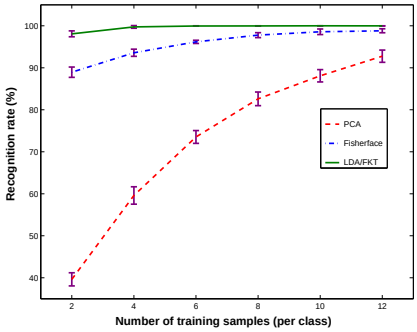

Figure 6. The accuracy curve on PIE with varying training samples. We show the rate and standard deviation from 10 runs.

ples are used, LDA/FKT consistently outperforms PCA and Fisherface. This is easily explained by the subspaces used in each method. LDA/FKT uses Subspaces 1 and 2, which are the most discriminative. Fisherface ignores Subspace 1 and operates in Subspaces 2 and 3. PCA has no notion of discriminative subspaces at all, since it is designed for pattern representation, rather than classification [2, 3].

Note that even when we have only 2 or 4 training samples from each subject, LDA/FKT can still obtain the best recognition accuracy ($\sim 99\%$) with the smallest standard deviation ($< 1\%$). This means LDA/FKT can handle small sample size problem with high and stable performance, which is not the case for PCA or Fisherface.

5 Conclusion

In this paper, we showed how FKT provides valuable insights into LDA by exposing the four subspaces that make up the entire data space. These subspaces differ in discriminability because each has a different value for the Fisher's Criterion. We also proved that the common technique of applying PCA followed by LDA (a.k.a. Fisherface) is not optimal because this discards Subspace 1, which is the most discriminative subspace.

Our result also showed how FKT, originally proposed for 2-class problems, can be generalized to handle multiple classes. This is achieved by replacing the autocorrelation matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ with the scatter matrices $\mathbf{S}_b$ and $\mathbf{S}_w$.

In the near future, we intend to investigate further the insights that FKT reveals about LDA. Another interesting direction to pursue is the extend our theory to nonlinear discriminant analysis. One way is to use the kernel trick employed in SVM, e.g., construct kernelized between-class scatter and within-class scatter matrices. FKT may yet again reveal new insights into kernelized LDA.

References

[1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), 1997.

[2] R. Duda, P. Hart, and D. Stork. *Pattern Classification, 2nd edition.* John Wiley and Sons, 2000.

[3] K. Fukunaga. *Introduction to Statistical Pattern Recognition, 2rd Edition.* Academic Press, 1990.

[4] G. Golub and C. V. Loan. *Matrix Computations, 3rd Edition.* John Hopkins Univ. Press, 1996.

[5] 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, August 2004.

[6] I. Jolliffe. *Principal component analysis*. New York: SpringerVerlag, 1986.

[7] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615 – 1618, December 2003.

[8] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.