Title
Discriminant subspace analysis: A Fukunaga-Koontz approach

Permalink
https://escholarship.org/uc/item/9k52756v

Journal
IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(10)

ISSN
0162-8828

Authors
Zhang, Sheng
Sim, Terence

Publication Date
2007-10-01

Peer reviewed
Discriminant Subspace Analysis: A Fukunaga-Koontz Approach

Sheng Zhang, Member, IEEE, and Terence Sim, Member, IEEE

Abstract — The Fisher Linear Discriminant (FLD) is commonly used in pattern recognition. It finds a linear subspace that maximally separates class patterns according to the Fisher 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 with different discriminabilities, as measured by eigenvalue ratios. By connecting the eigenvalue ratio with the generalized eigenvalue, we show where the Fisher Criterion is maximally satisfied. We prove the relationship between FLD and FKT analytically and propose a unified framework to understanding some existing work. Furthermore, we extend our theory to the Multiple Discriminant Analysis (MDA). This is done by transforming the data into intraclass and extraclass spaces, followed by maximizing the Bhattacharyya distance. Based on our FKT analysis, we identify the discriminant subspaces of MDA/FKT and propose an efficient algorithm, 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.

Index Terms — Discriminant subspace analysis, Fukunaga-Koontz transform, pattern classification.

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 dimensionality reduction in face recognition [2], [3], [15] and text classification [4]. One popular method is the Fisher Linear Discriminant (FLD), also known as the Linear Discriminant Analysis (LDA) [5], [7]. 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 $S_b$ and the within-class scatter matrix $S_w$, the Fisher Criterion can be written as

$$J_F(\Phi) = \text{trace}\left\{ (\Phi^T S_w \Phi)^{-1} (\Phi^T S_b \Phi) \right\},$$

(1)

where $\Phi$ is a linear transformation matrix. By maximizing the criterion $J_F$, FLD finds the subspaces in which the classes are most linearly separable. The solution [7] that maximizes $J_F$ is a set of the first eigenvectors $\{\phi_i\}$ that must satisfy

$$S_w \phi = \lambda S_b \phi.$$  

(2)

This is called the generalized eigenvalue problem [5], [7]. The discriminant subspace is spanned by the generalized eigenvectors. The discriminability of each eigenvector is measured by the corresponding generalized eigenvalue, for example, the most discriminant subspace corresponds to the largest generalized eigenvalue. Equation (2) can be solved by matrix inversion and eigendecomposition, namely, by applying eigendecomposition on $S_w^{-1} S_b$. Unfortunately, for many applications with high-dimensional data and few training samples, for example, face recognition, the scatter matrix $S_w$ is singular because, generally, the dimension of the data is larger than the number of samples. This is known as the undersampled or small sample size problem [7], [5].

Until now, many methods have been proposed to circumvent the requirement of nonsingularity of $S_w$, such as Fisherface [2], Discriminant Common Vectors [3], Dual Space [19], LDA/GSVD [9], and LDA/QR [20]. In [2], Fisherface first applies PCA [13], [18] to reduce dimension and LDA. The LDA/GSVD algorithm [9] avoids the inversion of $S_w$ by the simultaneous diagonalization via GSVD. In [20], Ye and Li proposed a two-stage LDA method which applies QR decomposition on a small matrix, followed by LDA. Moreover, Ye and Li also showed that both Fisherface and LDA/QR are the approximations of LDA/GSVD.

However, these methods do not directly relate to the generalized eigenvalue $\lambda$, the essential measure of discriminability. In fact, as we will show in Section 4, existing methods result in a suboptimum of the Fisher Criterion because important discriminant information is discarded to make $S_w$ invertible. In our previous work [22], we proposed a better solution by applying the Fukunaga-Koontz Transform (FKT) to the LDA problem. Based on the eigenvalue ratio of FKT, we decomposed the whole data space into four subspaces. This revealed the relationship between LDA, FKT, and GSVD and allowed us to correctly maximize $J_F$ even when $S_w$ is singular.

In this paper, we extend our previous work in two ways: First, we present a unified framework for understanding other LDA-based methods. This provides valuable insights on how to choose the discriminant subspaces of the LDA problem. Second, we propose a new approach to multiple...
discriminant analysis (MDA). This is done by casting the multiclass problem into a two-class one and by maximizing
the Bhattacharyya distance (which is the error bound of the Bayes classifier [5]) rather than the Fisher Criterion. Then, the
discriminant subspace is obtained algebraically via FKT. This
means that our method can find the global optimum directly
(no iteration required), which is not the case in [6]. For
completeness, in this paper, we include details of our previous work [22] as well.

To summarize, our work has three main contributions:

1. We present a unifying framework to understand different
methods, namely, LDA, FKT, and GSVD. To be more specific, we show that, 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. We prove that our approach is useful for general
pattern recognition. Our theoretical analyses dem-
strate how to choose the best subspaces for maximum
discriminability and unify other subspace methods
such as Fisherface, PCA+NULL space, LDA/QR, and
LDA/GSVD.

3. We further propose a new criterion to handle MDA,
derived from the Bhattacharyya distance. Because
the Bhattacharyya distance upper bounds the Bayes
error [5], this new criterion is theoretically superior
to the Fisher Criterion, which is not related to the
Bayes error in general.

The rest of this paper is organized as follows: Section 2
reviews related work, that is, PCA, LDA, Fisherface,
PCA+NULL Space, LDA/QR, and LDA/GSVD. We discuss
FKT in Section 3, where discriminant subspace analysis
based on FKT is also presented. In Section 4, we show how
to unify some LDA-based methods based on our theory.
Moreover, we demonstrate how to handle the multiclass
problem by FKT in Section 5. We apply our theory to the
classification problem on synthetic and real data in Section 6
and conclude our paper in Section 7.

2 RELATED WORK
Notation. Let $A = \{a_1, \ldots, a_N\}, a_i \in \mathbb{R}^d$ denote a data set
of given $D$-dimensional vectors. Each data point belongs to
exactly one of $C$ object classes $\{L_1, \ldots, L_C\}$. The number
of vectors in class $L_i$ is denoted by $N_i$; thus, $N = \sum N_i$.
Observe that, for high-dimensional data, for example, face
images, generally, $C \leq N \ll D$. The between-class scatter
matrix $S_b$, the within-class scatter matrix $S_w$, and the
total scatter matrix $S_t$ are defined as follows:

$$
S_b = \sum_{i=1}^{C} N_i (m_i - \mu)(m_i - \mu)^T = H_b H_b^T,
$$

(3)

$$
S_w = \sum_{i=1}^{C} \sum_{a_j \in L_i} (a_j - m_i)(a_j - m_i)^T = H_w H_w^T,
$$

(4)

$$
S_t = \sum_{i=1}^{C} (a_i - m)(a_i - m)^T = H H^T,
$$

(5)

$$
S_t = S_b + S_w.
$$

(6)

Here, $m_i$ denotes the class mean and $\mu$ is the global mean
of $A$. The matrices $H_b \in \mathbb{R}^{D \times C}$, $H_w \in \mathbb{R}^{D \times N}$ and $H \in
\mathbb{R}^{D \times N}$ are the precursor matrices of the between-class
scatter matrix, the within-class scatter matrix, and the total
scatter matrix, respectively,

$$
H_b = \left[ \sqrt{N_1}(m_1 - \mu), \ldots, \sqrt{N_C}(m_C - \mu) \right],
$$

(7)

$$
H_w = [a_1 - m_1 \cdot 1^T, \ldots, a_C - m_C \cdot 1^T],
$$

(8)

$$
H_t = [a_1 - m_1, \ldots, a_C - m_C].
$$

(9)

2.1 Principal Component Analysis (PCA)
Principal Component Analysis (PCA) [13] is one of the well-
known subspace methods for dimensionality reduction. It is
the optimal method for statistical pattern representation in
terms of the mean square error. PCA can be readily computed
by applying the eigendecomposition on the total scatter
matrix, that is, $S_t = U D U^T$. By keeping the eigenvectors
(principal components) corresponding to the largest eigen-
values, we can compute the PCA projection matrix. To solve
the appearance-based face recognition problem, Turk and
Pentland [18] proposed “Eigenface” by using PCA. Note that
PCA is optimal for pattern representation, not necessarily for
classification [5]. LDA [5], however, is another well-known
subspace method designed for pattern classification.

2.2 Linear Discriminant Analysis (LDA)
Given the data matrix $A$, which can be divided into $C$ classes,
we try to find a linear transformation matrix $\Phi \in \mathbb{R}^{D \times d}$,
where $d < D$. This maps a high-dimensional data to a low-
dimensional space. From the perspective of pattern classifi-
cation, LDA aims to find the optimal transformation $\Phi$ such
that the projected data are well separated.

Regarding pattern classification, usually, there are two
types of criteria that are used to measure the separability
of classes [7]. One is a family of criteria that gives the upper
bound on the Bayes error, for example, Bhattacharyya
distance. The other is based on a family of functions of scatter
matrices. As shown in (1), the Fisher Criterion belongs to the
latter one. Moreover, the solution of the criterion is the
generalized eigenvector and eigenvalue of the scatter
matrices (see (2)). However, if $S_w$ is nonsingular, it can be
solved by the generalized eigendecomposition: $S_w^{-1} S_t \lambda = \lambda \Phi$.
Otherwise, $S_w$ is singular and we circumvent this by
methods such as Fisherface [2], PCA+NULL Space [10],
LDA/QR [20], and LDA/GSVD [9].

2.3 Fisherface
To handle face recognition under different lightings,
Belhumeur et al. [2] proposed “Fisherface,” which is an
application of LDA. In the Fisherface method, PCA is
performed first so as to make $S_w$ nonsingular, followed by
LDA. This means that Fisherface = LDA + PCA. However,
there exist at least two problems: 1) During PCA, it is not
clear how many dimensions should be kept so that $S_w$ is
nonsingular and 2) to avoid the singularity of $S_w$, some
directions/eigenvectors (corresponding to the small non-zero eigenvalues) are thrown away in the PCA step, which may contain discriminant information [21].

2.4 PCA + NULL Space

Considering that the null space of \( S_w \) contains discriminant information, Huang et al. [10] first remove the null space of \( S_w \) This is the intersection of null space of \( S_w \) and \( S_o \) and has been proven to be useless for discrimination [10]. It can be done by applying PCA first, followed by computing the principal components of \( S_o \) within the null space of \( S_w \). More precisely, it is realized in three steps:

- **Step 1.** Remove the null space of \( S_o \): Eigendecompose \( S_o = U_o D_o U_o^T \), and \( U \) is the set of eigenvectors corresponding to the non-zero eigenvalues. Let \( S_o' = U S_o U \) and \( S_o'' = U^T S_o U \).

- **Step 2.** Compute the null space of \( S_o'' \): Eigendecompose \( S_o'' \) and let \( Q_{1} \) be the set of eigenvectors corresponding to the zero eigenvalues. Let \( S_o''' = Q_{1} S_o' Q_{1}^T \).

- **Step 3.** Remove the null space of \( S_w''' \) if it exists: Eigendecompose \( S_w''' \) and keep the set of eigenvectors corresponding to the non-zero eigenvalues.

The key difference between PCA+NULL Space and Fisherface is in the first step: PCA+NULL Space only removes the eigenvectors with zero eigenvalues, whereas Fisherface removes eigenvectors corresponding to zero and nonzero eigenvalues.

2.5 LDA/QR

In [20], Ye and Li proposed a two-stage LDA method, namely, LDA/QR. It not only overcomes the singularity problems of LDA but also achieves computational efficiency. This is done by applying QR decomposition on \( H_o \) first, followed by LDA. To be more specific, it is realized in two steps:

- **Step 1.** Apply QR decomposition on \( H_o = QR \), where \( Q \in \mathbb{R}^{D \times r} \) has orthogonal columns that span the space of \( H_o \) and \( R \in \mathbb{R}^{n \times C} \) is an upper triangular matrix. Then, define \( S_o = Q^T S_o Q \) and \( S_w = Q^T S_w Q \).

- **Step 2.** Apply LDA on \( S_o \) and \( S_w \): Keep the set of eigenvectors corresponding to the smallest eigenvalues of \( S_o \) and \( S_w \).

Note that, to reduce computational load, QR decomposition is employed here, whereas, in the Fisherface and PCA+NULL space methods, the subspace is obtained by using eigendecomposition.

2.6 LDA/GSVD

The GSVD was originally defined by Van Loan [14] and then Page and Saunders [16] extended it to handle any two matrices with the same number of columns. We will briefly review the mechanism of GSVD, using LDA as an example.

Howland and Park [9] extended the applicability of LDA to the case when \( S_o \) is singular. This is done by using simultaneous diagonalization of the scatter matrices via the GSVD [8]. First, to reduce computational load, \( H_o \) and \( H_w \) are used instead of \( S_o \) and \( S_w \). Then, based on GSVD, there exist orthogonal matrices \( Y \in \mathbb{R}^{C \times C} \) and \( Z \in \mathbb{R}^{N \times N} \) and a nonsingular matrix \( X \in \mathbb{R}^{d \times d} \) such that

\[
Y^T H_o^T X = [\Sigma_o, 0], \quad (10)
\]

\[
Z^T H_w^T X = [\Sigma_w, 0], \quad (11)
\]

where

\[
\Sigma_o = \begin{bmatrix} I_o & D_o \\ 0 & O_o \end{bmatrix}, \quad \Sigma_w = \begin{bmatrix} O_w & D_w \\ 0 & I_w \end{bmatrix}.
\]

The matrices \( I_o \in \mathbb{R}^{(n-r_o) \times (n-r_o)} \) and \( I_w \in \mathbb{R}^{(n-r_o) \times (n-r_o)} \) are identity matrices, \( O_o \in \mathbb{R}^{(C-r_o) \times (C-r_o)} \) and \( O_w \in \mathbb{R}^{(N-r_o) \times (N-r_o)} \) are rectangular zero matrices that may have no rows or no columns, \( D_o = \text{diag}(\alpha_{r_o}, \ldots, \alpha_{C}) \) and \( D_w = \text{diag}(\beta_{r_o}, \ldots, \beta_{N}) \) satisfy \( 1 > \alpha_{r_o} > \ldots > \alpha_{C} > 0 \), \( 0 < \beta_{r_o} \leq \ldots \leq \beta_{N} < 1 \), and \( \alpha^2 + \beta^2 = 1 \). Thus, \( \Sigma_o^T \Sigma_o + \Sigma_w^T \Sigma_w = I \), where \( I \in \mathbb{R}^{N \times N} \) is an identity matrix. The columns of \( X \), which are the generalized singular vectors for the matrix pair \([H_o, H_w]\), can be used as the discriminant feature subspace based on GSVD.

3 Fukunaga-Koontz Transform and LDA

In this section, we begin by briefly reviewing the Fukunaga-Koontz Transform (FKT). Then, based on the eigenvalue ratio of FKT, we analyze the discriminant subspaces by breaking the whole space into smaller subspaces. Finally, we connect FKT to the Fisher Criterion, which suggests a way to select discriminant subspaces.

3.1 Fukunaga-Koontz Transform

The FKT was designed for the two-class recognition problem. Given the data matrices \( A_1 \) and \( A_2 \) from two classes, the autocorrelation matrices \( S_1 = A_1 A_1^T \) and \( S_2 = A_2 A_2^T \) are positive semidefinite (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

\[
S = S_1 + S_2 = [U, U_\perp] \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U^T \\ U_\perp^T \end{bmatrix}. \quad (12)
\]

Without loss of generality, \( S \) may be singular and \( r = \text{rank}(S) < D \); thus, \( D = \text{diag}(\lambda_1, \ldots, \lambda_r) \), \( \lambda_1 \geq \ldots \geq \lambda_r > 0 \). The set of eigenvectors \( U \in \mathbb{R}^{D \times r} \) corresponds to nonzero eigenvalues and the set \( U_\perp \in \mathbb{R}^{D \times (D-r)} \) is the orthogonal complement of \( U \). Now, we can whiten \( S \) by a transformation operator \( P = UD^{-1/2} \). The sum of the two matrices \( S_1 \) and \( S_2 \) becomes

\[
P^T S P = P^T (S_1 + S_2) P = S_1 + S_2 = I, \quad (13)
\]

where \( S_1 = P^T S_1 P \), \( S_2 = P^T S_2 P \), and \( I \in \mathbb{R}^{N \times r} \) is an identity matrix. Suppose the eigenvector of \( S_1 \) is \( v \) with the eigenvalue \( \lambda_1 \) that is, \( S_1 v = \lambda_1 v \). Since \( S_1 = I - S_2 \), we can rewrite it as

\[
(I - S_2) v = \lambda_1 v, \quad (14)
\]

\[
S_2 v = (1 - \lambda_1) v. \quad (15)
\]

This means that \( S_2 \) has the same eigenvector as \( S_1 \), but the corresponding eigenvalue is \( \lambda_2 = 1 - \lambda_1 \). Consequently, the dominant eigenvector of \( S_1 \) is the weakest eigenvector of \( S_2 \) and vice versa. This suggests that a pattern belonging to Class I ought to yield a large coefficient when projected onto
the dominant eigenvector of $\tilde{S}_1$ and vice versa. The dominant eigenvectors therefore form a subspace in which the two classes are separable. Classification can then be done by, say, picking the nearest neighbor (NN) in this subspace.

Recently, it was proven that, under certain conditions, FKT is the best linear approximation to a quadratic classifier [11]. Interested readers may refer to [7] and [11] for more details.

### 3.2 LDA/FKT

Generally speaking, for the LDA problem, there are more than two classes. To handle the multiclass problem, we replace the autocorrelation matrices $S_1$ and $S_2$ with the scatter matrices $S_b$ and $S_w$. Since $S_b$, $S_w$, and $S_1$ are p.s.d. and symmetric and $S_b = S_1 + S_w$, we can apply FKT on $S_b$, $S_w$, and $S_1$, which is called LDA/FKT hereafter in this paper. The whole data space is decomposed into $U$ and $U_\perp$ (Fig. 1). On one hand, $U_\perp$ is the set of eigenvectors corresponding to the zero eigenvalues of $S_1$. This has been proven to be the intersection of the null spaces of $S_b$ and $S_w$ and contains no discriminant information [10]. On the other hand, $U$ is the set of eigenvectors corresponding to the nonzero eigenvalues of $S_1$. It contains discriminant information.

Based on FKT, $S_b = P^T S_b P$ and $S_w = P^T S_w P$ share the same eigenspace and the sum of two eigenvalues corresponding to the same eigenvector is equal to 1.

1. **Subspace 1.** $\text{span}(S_b) \cap \text{null}(S_w)$, the set of eigenvectors $\{v_i\}$ corresponding to $\lambda_w = 0$ and $\lambda_b = 1$. Since $\frac{\lambda_b}{\lambda_w} = \infty$, in this subspace, the eigenvalue ratio is maximized.

2. **Subspace 2.** $\text{span}(S_b) \setminus \text{span}(S_w)$, the set of eigenvectors $\{v_i\}$ corresponding to $0 < \lambda_w < 1$ and $0 < \lambda_b < 1$. Since $0 < \frac{\lambda_b}{\lambda_w} < \infty$, the eigenvalue ratio is finite and smaller than that of Subspace 1.

3. **Subspace 3.** $\text{null}(S_b) \setminus \text{span}(S_w)$, the set of eigenvectors $\{v_i\}$ corresponding to $\lambda_w = 1$ and $\lambda_b = 0$. Since $\frac{\lambda_b}{\lambda_w} = 0$, the eigenvalue ratio is minimum.

4. **Subspace 4.** $\text{null}(S_b) \cap \text{null}(S_w)$, the set of eigenvectors corresponding to the zero eigenvalues of $S_1$.

Note that, in practice, some of these four subspaces may not exist, depending on the ranks of $S_b$, $S_w$, and $S_1$. As illustrated in Fig. 1, the null space of $S_w$ is the union of Subspace 1 and Subspace 4, whereas the null space of $S_b$ is the union of Subspace 3 and Subspace 4, if they exist. Therefore, from the perspective of FKT, we reach the same conclusion as Huang et al. in [10]. That is, Subspace 4 is the intersection of the null spaces of $S_b$ and $S_w$.

### 3.3 Relationship between FKT, GSVD, and LDA

How do these four subspaces help to maximize the Fisher Criterion $J_F$? We explain this in Theorem 1, which connects the generalized eigenvalue of $J_F$ to the eigenvalues of FKT. We begin with a lemma (see Appendix A for the proof):

**Lemma 1.** For the LDA problem, GSVD is equivalent to FKT, with $X = [UD^{-1/2}V, U_\perp]$, $\lambda_b = \Sigma_b^T \Sigma_b$, and $\lambda_w = \Sigma_w^T \Sigma_w$, where $X$, $\Sigma_b$, and $\Sigma_w$ are from GSVD (10), (11), and $U$, $D$, $V$, $U_\perp$, $\lambda_w$, and $\lambda_b$ are matrices from FKT (16), (17), (18).

Now, based on the above lemma, we can investigate the relationship between the eigenvalue ratio of FKT and the generalized eigenvalue $\lambda$ of the Fisher Criterion $J_F$.

**Theorem 1.** If $\lambda$ is the solution of (2) (the generalized eigenvalue of $S_b$ and $S_w$) and $\lambda_b$ and $\lambda_w$ are the eigenvalues after applying FKT on $S_b$ and $S_w$, then $\lambda = \frac{\lambda_b}{\lambda_w}$, where $\lambda_b + \lambda_w = 1$.

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

$$S_b = H_x H_b^T = X^{-T} \begin{bmatrix} \Sigma_b^T \Sigma_b & 0 \\ 0 & 0 \end{bmatrix} X^{-1}.$$
Corollary 1. \( \Lambda_b = \Sigma_b^T \Sigma_b \) and \( \Lambda_w = \Sigma_w^T \Sigma_w \), thus,

\[
S_b = X^{-T} [\Lambda_b \ 0 \ 0] X^{-1}.
\]

Similiarly,

\[
S_w = X^{-T} [\Lambda_w \ 0 \ 0] X^{-1}.
\]

Since \( S_{b\phi} = \lambda S_{w\phi} \),

\[
X^{-T} [\Lambda_b \ 0 \ 0] X^{-1} = \lambda X^{-T} [\Lambda_w \ 0 \ 0] X^{-1}.
\]

Letting \( v = X^{-1} \phi \), and multiplying \( X^T \) on both sides, we obtain the following:

\[
\Lambda_b [\Lambda_b \ 0 \ 0] v = \lambda \Lambda_w [\Lambda_w \ 0 \ 0] v.
\]

If we add \( \Lambda_b \ 0 \ 0 \) on both sides of the above equation, then

\[
(1 + \lambda) \Lambda_b \ 0 \ 0 = \lambda I \ 0 \ 0 v.
\]

This means that \( (1 + \lambda) \lambda_b = \lambda_c \), 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 \( \lambda = \frac{\lambda_c}{\lambda_w} \). \( \square \)

Corollary 1. If \( \lambda \) is the generalized eigenvalue of \( S_b \) and \( S_w \), \( \alpha \) and \( \beta \) are the solutions of (10) and (11), and \( \alpha/\beta \) is the generalized singular value of the matrix pair \( (H_b, H_w) \), then \( \lambda = \frac{\alpha^2}{\beta^2} \), where \( \alpha^2 + \beta^2 = 1 \).

Proof. In Lemma 1, we have proven 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 1, we observe that \( \lambda = \frac{\alpha^2}{\beta^2} \). Therefore, it is easy to see that \( \lambda = \frac{\alpha^2}{\beta^2} \). Note that \( \frac{\alpha^2}{\beta^2} \) is the generalized singular value of \( (H_b, H_w) \) by GSVD and \( \lambda \) is the generalized eigenvalue of \( (S_b, S_w) \). \( \square \)

The corollary suggests how to evaluate discriminant subspaces of LDA/GSVD. Actually, Howland and Park in [9] applied the corollary implicitly, but, in this paper, we explicitly connect the generalized singular value \( \frac{\alpha^2}{\beta^2} \) with the generalized eigenvalue \( \lambda \), the measure of discriminability.

Based on our analysis, the eigenvalue ratio \( \frac{\lambda_b}{\lambda_w} \) and the square of the generalized singular value \( \frac{\alpha^2}{\beta^2} \) are both equal to the generalized eigenvalue \( \lambda \), the measure of discriminability. According to Fig. 1, Subspace 1, with the infinite eigenvalue ratio \( \frac{\lambda_1}{\lambda_3} \), is the most discriminant subspace, followed by Subspace 2 and Subspace 3. However, Subspace 4 contains no discriminant information and can be safely thrown away. Therefore, the eigenvalue ratio \( \frac{\lambda_1}{\lambda_3} \) or the generalized singular value \( \frac{\alpha^2}{\beta^2} \) suggests how to choose the most discriminant subspaces.

### 3.4 Algorithm for LDA/FKT

Although we proved that FKT is equivalent to GSVD on the LDA problem, as we will see in Table 1, LDA/GSVD is computationally expensive. Since Subspace 4 contains no discriminant information, we may compute the Subspaces 1, 2, and 3 of LDA/FKT based on QR decomposition. Moreover, we use smaller matrices \( H_b \) and \( H_w \), because matrices \( S_b \), \( S_w \), and \( S_b \) may be too large to be formed. Our LDA/FKT algorithm is shown in Fig. 2.

Now, we analyze the computational complexity of the algorithm as follows:

1. **Time complexity.** Line 2 takes \( O(DN^2) \) time to compute the QR decomposition on \( H_b \). To multiply two matrices, Line 3 takes \( O(rDN) \) time. Line 4 takes \( O(r_1DC) \) time, and Line 5 takes \( O(r_2C) \) time. Line 6 takes \( O(r_1^2i) \) time to invert \( S_b \), multiply the matrices,

### Table 1

Comparison between Different Methods: \( N \) is the Number of Training Samples, \( D \) is the Dimension, and \( C \) is the Number of Classes

<table>
<thead>
<tr>
<th>Method</th>
<th>Time complexity</th>
<th>Space complexity</th>
<th>Discriminant subspaces</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>( O(DN^2) )</td>
<td>( O(DN) )</td>
<td>2 and 3</td>
<td>Optimal for pattern representation</td>
</tr>
<tr>
<td>Face</td>
<td>( O(DN^2) )</td>
<td>( O(DN) )</td>
<td>1</td>
<td>Sub-optimal for Fisher Criterion</td>
</tr>
<tr>
<td>LDA/GSVD</td>
<td>( O(DNC) )</td>
<td>( O(DC) )</td>
<td>1 and 2</td>
<td></td>
</tr>
<tr>
<td>LDA/FKT</td>
<td>( O((N+C)^2D) )</td>
<td>( O(DN) )</td>
<td>1, 2, 3 and 4</td>
<td>Optimal for the Fisher Criterion</td>
</tr>
<tr>
<td>MDA/FKT</td>
<td>( O(DN^2) )</td>
<td>( O(DN) )</td>
<td>1, 2 and 3</td>
<td>Optimal for Bhattacharyya distance</td>
</tr>
</tbody>
</table>

For discriminant subspaces, please refer to Fig. 1.
and perform eigendecomposition on the $r_1 \times r_1$ matrix $S_w^{-1}S_b$. Since $r_1 < N$, $C \ll D$, the most intensive step is Line 2, which takes $O(DN^2)$ time to compute the QR decomposition. Thus, the time complexity is $O(DN^2)$.

2. **Space complexity.** Lines 2 and 4 involve matrices $H_c$ and $H_b$. Because of the size of the matrix, $H_c$ requires $O(DN)$ space in memory, and $H_b$ requires $O(DC)$. Lines 3, 5, and 6 only involve $E \in \mathbb{R}^{n \times N}$, $Z \in \mathbb{R}^{D \times C}$, and $S_1, S_2 \in \mathbb{R}^{n \times n}$, which are all small matrices. Therefore, the space complexity is $O(DN)$.

### 4 Comparison

Although Fisherface, PCA+NULL, LDA/GSVD, and LDA/QR were all proposed independently and appear to be different algorithms, in this section, we explain how FKT provides insights into these methods.

#### 4.1 Fisherface: Subspaces 2 and 3

In the Fisherface method, PCA is performed first so as to make $S_b$ nonsingular. This is done by throwing away Subspaces 1 and 4 (Fig. 1). As a result, Subspace 1, the most discriminant subspace in terms of the Fisher Criterion, is discarded. Therefore, Fisherface operates only in Subspaces 2 and 3 and is suboptimal.

#### 4.2 PCA + NULL Space: Subspace 1

Considering the discriminant information contained in the null space of $S_w$, PCA+NULL Space first removes the null space of $S_b$, which is Subspace 4 (Fig. 1). Now, only Subspaces 1, 2, and 3 are left. Second, within Subspace 1, the principal components of $S_b$ are computed. Thus, only Subspace 1, the most discriminant feature space, is used.

Other null space methods have also been reported in the literature, such as Direct LDA [21] and NULL Space [3]. The criterion used in these methods is a modified version of the Fisher Criterion, namely,

$$
\Phi_{opt} = \arg \max_{\Phi} \| \Phi^T S_b \Phi \| \quad s.t. \| \Phi^T S_w \Phi \| = 0. \tag{25}
$$

Equation (25) shows that $\Phi_{opt}$ is the set of eigenvectors associated with the zero eigenvalues of $S_w$ and the maximum eigenvalues of $S_b$. Based on the eigenvalue ratio from Fig. 1, this is Subspace 1. Thus, the PCA+NULL, Direct LDA, and NULL space methods all operate only in Subspace 1. However, as we will show in our experiments in Section 6.1, using Subspace 1 alone is sometimes not sufficient for good discrimination because Subspaces 2 and 3 may be necessary. In the worst case, when Subspace 1 does not exist, these null space methods will fail.

#### 4.3 LDA/QR: Subspaces 1 and 2

To circumvent the nonsingularity requirement of $S_w$ and reduce the computation, a two-stage strategy is used in LDA/QR [20]. The eigenspace (corresponding to nonzero eigenvalues) of $S_b$ is computed by applying QR on $H_c$. In fact, this is Subspace 1 (Fig. 1) because the eigenvalues of $S_b$ associated with Subspaces 3 and 4 are all zero, which are thrown away by the QR decomposition. Then, the eigenvectors corresponding to the smallest eigenvalues of $S_w^{-1}S_b$ are computed, equivalently, computing the eigenvectors corresponding to the largest $\frac{1}{n}$. Note that $S_w^{-1}S_b$, rather than $S_w^{-1}S_w$ is eigendecomposed because $S_w^{-1}$ may still be singular ($S_w$ is zero within Subspace 1). Therefore, as Fig. 1 illustrated, Subspaces 1 and 2 are preserved by LDA/QR. This means that LDA/QR operates in Subspaces 1 and 2.

#### 4.4 LDA/GSVD: Subspaces 1, 2, 3, and 4

Both LDA/GSVD and LDA/FTK simultaneously diagonalize two matrices, but, so far, nobody has investigated the relationship between these two methods. In this paper, one of our contributions is the proof that LDA/GSVD and LDA/FTK are equivalent (see Appendix A for the proof). More specifically, from the perspective of FKT, the $Y$ and $Z$ in LDA/GSVD (see (10) and (11)) are just arbitrary rotation matrices. The discriminant subspace of LDA/GSVD $X$ is equal to $[UD^{-1/2}V, U_{1:}]$, where $U_{1:}$ is Subspace 4, and $U$ is the union of Subspaces 1, 2, and 3. This means $X$ contains Subspaces 1, 2, 3, and 4 (see Fig. 1). Therefore, the subspaces obtained by LDA/GSVD are exactly those obtained by LDA/FTK. However, LDA/GSVD is computationally expensive (Table 1). In Fig. 2, we presented an efficient algorithm to compute LDA/FTK. This is achieved by using QR decomposition on $S_b$ to obtain Subspaces 1, 2, and 3. We do not have to compute for Subspace 4, since it contains no discriminant information.

### 5 Multiple Discriminant Analysis (MDA)

#### 5.1 MDA/FTK

From the perspective of the Bayes Classifier, LDA is optimal only for two Gaussian distributions with equal covariance matrices [5], [7], and the Fisher Criterion has been extended to handle multiple Gaussian distributions or classes with unequal covariance matrices. This suggests that, for multiple Gaussian distributions or classes with unequal covariance matrices, LDA-based methods are not optimal with respect to the Bayes Classifier. The worst case occurs when all classes have the same mean. In this case, $S_b = 0$ and all LDA-based methods will fail. Subspaces 1 and 2 do not exist and we are left with only Subspaces 3 and 4, which are less discriminative. To handle these problems, we cast the multiclass problem into a binary pattern classification problem by introducing $\Delta = a_i - a_j$ and defining the intraclass space $\Omega_I = \{(a_i - a_j) \mid L(a_i) = L(a_j)\}$, as well as the extraclass space $\Omega_E = \{(a_i - a_j) \mid L(a_i) \neq L(a_j)\}$, where $L(a_j)$ is the class label of $a_i$. This idea has been used by other researchers, for example, Moghaddam in [15]. The statistics of $\Omega_I$ and $\Omega_E$ are defined as follows:

$$
\mu_i = m_E = 0, \tag{26}
$$

$$
\Sigma_I = H_cH_c^T = \frac{1}{N_I} \sum_{L(a_i) = L(a_j)} (a_i - a_j)(a_i - a_j)^T, \tag{27}
$$

$$
\Sigma_E = H_cH_c^T = \frac{1}{N_E} \sum_{L(a_i) \neq L(a_j)} (a_i - a_j)(a_i - a_j)^T. \tag{28}
$$

Here, $N_I = \frac{1}{2} \sum n_i(n_i - 1)$ is the number of samples in $\Omega_I$ and $N_E = \sum_{i \neq j} n_i n_j$ is the number of samples in $\Omega_E$. For example, if every class has the same number of training samples, $n_i = n$ for $i = 1, \ldots, C$, then $N_I = \frac{1}{2} n(n - 1)$ and $N_E = \frac{1}{2} N(N - n)$. Note that, usually, $\text{rank}(\Sigma_I)$ and $\text{rank}(\Sigma_E)$
are both greater than \( C - 1 \), where \( C \) is the number of classes. \( \mathbf{H}_I \) and \( \mathbf{H}_E \) are the precursor matrices of \( \Sigma_I \) and \( \Sigma_E \) given by

\[
\mathbf{H}_I = \frac{1}{\sqrt{N_I}}[\ldots \mathbf{a}_i - \mathbf{a}_j \ldots], \quad \forall i > j
\]

such that \( L(\mathbf{a}_i) = L(\mathbf{a}_j) \), \( \beta = \frac{\lambda^2}{\sqrt{\mathbf{H} \Sigma \mathbf{H}^T}} \), \( \Sigma_i \) is the total scatter matrix defined in (5) and \( \Sigma_E \) or \( \Sigma_I \) is the total number of samples, \( N_I \) is the number of training samples per class, then \( N = C = 134, N_I = 67, \) and \( N_E = 8,844 \). \( \Sigma_E = 5,600 \times 5,600 \) and \( \Sigma_I = 5,600 \times 8,844 \) in size.

Can we find an efficient way to obtain the Subspaces 1 and 3 of MDA/FKT without \( \mathbf{H}_I \) or \( \mathbf{H}_E \)? Yes. Based on the relationship between \( \Sigma_I, \Sigma_I, \) and \( \Sigma_E \), we devise a method that works with \( \mathbf{H}_I \in \mathbb{R}^{D \times N_I} \) and \( \mathbf{H}_E \in \mathbb{R}^{D \times N_E} \). Throughout the following discussion, we denote \( \Sigma_I = N_I \Sigma_I + N_E \Sigma_E, \) where \( N \) is the total number of samples, \( N_I \) is the number of intraclass samples, and \( N_E \) is the number of extraclass samples.

To efficiently compute the generalized eigenvalues and eigenvectors of \( \Sigma_I, \Sigma_E \), we need the following theorem:

**Theorem 2.** If \( (\lambda, \mathbf{v}) \) is the dominant generalized eigenvalue and eigenvector of the matrix pair \( (\Sigma_I, \Sigma_E) \), then \( (\lambda', \mathbf{v}') \) is the dominant generalized eigenvalue and eigenvector of matrix pair \( (\Sigma_I', \Sigma_E') \), where

\[
\mathbf{v}' = \frac{\mathbf{v}}{\lambda'^N}, \quad \lambda = \frac{\lambda'}{\lambda'^N}.
\]

**Proof.** The generalized eigenvalue equation of matrix pair \( (\Sigma_I, \Sigma_E) \) is \( \Sigma_I \mathbf{v} = \lambda \Sigma_E \mathbf{v} \). Since \( \Sigma_I = \frac{N_I}{C} \Sigma_I \) and \( \Sigma_E = \frac{N_E}{C} \Sigma_E', \) we have

\[
\frac{2N}{N_E} \Sigma_E' \mathbf{v} = \lambda \frac{2N}{N_I} \Sigma_I \mathbf{v}
\]

and, thus,

\[
\Sigma_E' \mathbf{v} = \frac{N_E}{N_I} \lambda \Sigma_I \mathbf{v}.
\]
Input: The data matrix A.
Output: Projection matrix $\mathbf{F}_{MDA}$ such that the $J_{MDA}$ is maximized.

1. Compute $H_t$ and $H_f$ from data matrix A as in Equations (30) and (9).
2. Apply QR decomposition on $H_t = QR$, where $Q \in \mathbb{R}^{t \times t}$, $R \in \mathbb{R}^{t \times N}$ and $r_t = \text{rank}(H_t)$.
3. Let $S_t = RR^T$, since $S_t = Q^T S_t Q = Q^T H_t H_t^T Q = RR^T$.
4. Let $Z = Q^T H_f$.
5. Let $\Sigma_f^t = \frac{N_t}{N} ZZ^T$, since $\Sigma_f^t = Q^T \Sigma_f^t Q = \frac{N_t}{N} Q^T H_f H_f^T Q = \frac{N_t}{N} ZZ^T$.
6. Compute the eigenvectors $\{v_i\}$ and eigenvalues $\{\sigma_i\}$ of $S_t^2 \Sigma_f^t$.
7. Compute the generalized eigenvalues $\{\lambda_i\}$ of $(\Sigma_f^t, \Sigma_E^t)$ using $\lambda_i = \frac{N_t}{N e_i - \sigma_i}$.
8. Sort the eigenvectors $v_i$ according to $\lambda_i + \frac{1}{\lambda_i}$ in decreasing order.
9. The final projection matrix $\mathbf{F}_{MDA} = QV_k$ (the first $k$ columns of $V$), where $V = \{v_i\}$. Note that $k$ could be greater than $C - 1$.

MDA/FKT is comparable to most of LDA-based methods. MDA/FKT, however, is optimal in terms of Bhattacharyya distance, the error bound of the Bayes Classifier, which is not the case for other methods.

### 6 Experiments

Up until now, we have shown that FKT can be used to unify other LDA-based methods. Moreover, we proposed a new approach for MDA. In this section, we evaluate the performance of LDA/FKT and MDA/FKT by using synthetic and real data. The synthetic data has two sets, whereas the real data consists of three sets for digit recognition and face recognition. Table 2 shows the statistics of the data sets in our experiments.

The experimental setting for recognition is described as follows. For PCA, we take the top $C - 1$ principal components, where $C$ is the number of classes. For Fisherface, we apply PCA first and then take 100 principal components, followed by LDA. For MDA/FKT, when performing comparison with other methods, we project MDA/FKT to $C - 1$-dimensional space. With respect to recognition, we employ 1-NN in the low-dimensional space for all methods in these experiments.

#### 6.1 Toy Problems

To evaluate the performance of MDA/FKT, we begin with two toy examples:

- **Toy 1**: Three Gaussian classes; same mean, different covariance matrices. The three classes share the same zero mean in 3D space and each class has 100 points. They have different covariance matrices:

  $$C_1 = [1, 1, 0]^T \times [1, 1, 0] + 0.1[0, 1, 0]^T \times [0, 1, 1],$$

  $$C_2 = [0, 1, 1]^T \times [0, 1, 0] + 0.1[1, 0, 1]^T \times [1, 0, 1],$$

  and

  $$C_3 = [1, 0, 1]^T \times [1, 0, 0] + 0.1[1, 1, 0]^T \times [1, 1, 0]$$

  (Fig. 4a). Note that LDA-based methods will fail here because $S_0 = 0$.

- **Toy 2**: Two classes: Gaussian mixture. We also have two classes in 3-D space. One class contains 50 points and the other contains 75. The first class is generated from a single Gaussian with zero mean and 0.5I covariance. The second class is a Gaussian mixture which consists of three components with different means: $[1, 4, 0]$, $[2\sqrt{3}, -2, 0]$, and $[-2\sqrt{3}, -2, 0]$. Each component has

2. Asymptotically, if PCA can extract more features than LDA, it will perform better.
failure of LDA-based methods. The original 3D data share the same mean. However, MDA/FKT can still work even if all classes share the same mean, whereas all LDA-based methods fail because $S_b = 0$. This results in the failure of LDA-based methods.

In Toy 1, it does not make sense to maximize trace($S_w^{-1}S_b$) because $S_b = 0$. This means that LDA fails when the different classes share the same mean. However, MDA/FKT can still be applied because $\Sigma_I \neq 0$ and $\Sigma_E \neq 0$. As shown in Fig. 4b, by using MDA/FKT, we can obtain a 2D discriminative subspace, which can still approximate the structure of the original 3D data. Even though it is hard to determine the decision boundaries in Fig. 4b, we can discern that the three classes lie principally along three different axes. In Toy 2, LDA/FKT can obtain only a one-dimensional (1D) projection (Fig. 5c) because, for two classes, $\text{rank}(S_b) = 1$. Note that the 1D projections of LDA/FKT overlap significantly, which means it is hard to do classification. However, as shown in Fig. 5b, MDA/FKT can obtain a 2D subspace because the rank of $\Sigma_I$ or $\Sigma_E$ depends on the number of samples, not the number of classes. The larger discriminative subspace of MDA/FKT makes it possible to separate the classes.

Let us summarize these two toy problems: First, MDA/FKT can still work even if all classes share the same mean, whereas all LDA-based methods fail because $S_b = 0$. Second, MDA/FKT can provide larger discriminative subspaces than LDA-based methods because the latter ones are limited by the number of classes.

6.2 Digit Recognition
We perform digit recognition to compare MDA/FKT with LDA-based methods on the MFEAT [12] data set. This consists of handwritten digits (“0”-“9”) (10 classes) with 649-dimensional features. These features comprise six feature sets: Fourier coefficients, profile correlations, Karhunen-Loève coefficients, pixel averages in $2 \times 3$ windows, Zernike moments, and morphological features. For each class, we have 200 patterns; 30 of them are chosen randomly as training samples and the rest for testing. To evaluate the stability of each method, we repeat the sampling 10 times so that we can compute the mean and standard deviation of the recognition accuracy.

As shown in Fig. 6a, the accuracy of LDA/FKT and MDA/FKT is about 95 percent with small standard deviations, which means an accurate and stable performance. For MDA/FKT, we can investigate the relationship between performance and the projected dimension. Fig. 6b shows a plot of accuracy versus projected dimensions. The accuracy reaches the maximum when the projected dimension is around 8, after which it remains flat even if we increase the projected dimension. This suggests that MDA/FKT reaches its best performance around eight dimensions. Another observation is that the projected dimension should not be limited by the number of classes. For example, here, we have $C = 10$ classes, but Fig. 6b illustrates that seven or eight dimensions can give almost the same accuracy as $C - 1 = 9$ projected dimensions.

6.3 Face Recognition
We also perform experiments on real data on two face data sets:

1. PIE face data set [17]. We choose 67 subjects and each subject has 24 frontal face images taken under room lighting. All of these face images are aligned based on eye coordinates and cropped to $70 \times 80$. Fig. 7a shows a sample of PIE face images used in our experiments. The major challenge in this data set is to do face recognition under different illuminations.

2. Banca face data set [1]. This contains 52 subjects and each subject has 120 face images, which are normalized to $51 \times 55$ in size. By using a Webcam and an expensive camera, these subjects were recorded in three different scenarios over a period of three months. Each face image contains illumination, expression, and pose variations because the subjects are required to talk during the recording (Fig. 7b).

Fig. 7 shows a sample of PIE and Banca face images used in our experiments.

For face recognition, usually, we have an undersampled problem, which is also the reason for the singularity of $S_b$. To evaluate the performance under such a situation, we randomly choose $N$ training samples from each subject, $N = 2, \ldots, 12$, and the remaining images are used for testing. For each set of $N$ training samples, we employ cross validation so that we can compute the mean and standard deviation for classification accuracies. We show the mean and standard deviation (in parenthesis) of the recognition rate from 10 runs (see Table 3). Note that the largest number in each row is highlighted in bold.

3. The PIE data set contains 68 subjects altogether. We omitted one because s/he has too few frontal face images.
As shown in Table 3, we observe that the more training samples, the better the recognition accuracy. To be more specific, on both data sets, for each method, increasing the number of training samples increases the mean recognition rate and decreases the standard deviation. Note that, when the training set is small, LDA/FKT significantly outperforms the other methods. For example, for the PIE data set, with two training samples, LDA/FKT achieves about 98 percent accuracy compared with the next highest of 90 percent from PCA+NULL (see the first row of Table 3). Moreover, with four training samples, LDA/FKT achieves about 100 percent compared with 94 percent of the PCA+NULL method (see the second row of Table 3). The standard deviation of LDA/FKT is also significantly smaller than that of the other methods. For the Banca data set, we have a similar observation. With two and four training samples, LDA/FKT achieves about 5 percent higher in accuracy than the next highest (see the seventh and eighth rows of Table 3). This shows that LDA/FKT can handle small-sample-size problems very well. With more training samples, that is, 6-12, LDA/FKT is not the best but falls behind the highest by no more than 1.8 percent (see the bottom four rows of Table 3). One possible reason is that LDA/FKT is optimal as a linear classifier, whereas, for the Banca data set, the face images under expression and pose variations are nonlinearly distributed.

To compare PCA, MDA/FKT, and various LDA-based methods with different training samples, we also visualize the average classification accuracies in Fig. 8.

As shown in Table 3, we observe that the more training samples, the better the recognition accuracy. To be more specific, on both data sets, for each method, increasing the number of training samples increases the mean recognition rate and decreases the standard deviation. Note that, when the training set is small, LDA/FKT significantly outperforms the other methods. For example, for the PIE data set, with two training samples, LDA/FKT achieves about 98 percent accuracy compared with the next highest of 90 percent from PCA+NULL (see the first row of Table 3). Moreover, with four training samples, LDA/FKT achieves about 100 percent compared with 94 percent of the PCA+NULL method (see the second row of Table 3). The standard deviation of LDA/FKT is also significantly smaller than that of the other methods. For the Banca data set, we have a similar observation. With two and four training samples, LDA/FKT achieves about 5 percent higher in accuracy than the next highest (see the seventh and eighth rows of Table 3). This shows that LDA/FKT can handle small-sample-size problems very well. With more training samples, that is, 6-12, LDA/FKT is not the best but falls behind the highest by no more than 1.8 percent (see the bottom four rows of Table 3). One possible reason is that LDA/FKT is optimal as a linear classifier, whereas, for the Banca data set, the face images under expression and pose variations are nonlinearly distributed.

To compare PCA, MDA/FKT, and various LDA-based methods with different training samples, we also visualize the average classification accuracies in Fig. 8.
Note that, since MDA/FKT is not limited by the number of classes (unlike LDA), we may project the data onto a space whose dimension is greater than the number of classes. Fig. 9 shows a plot of accuracy versus projected dimensions for different numbers of training samples for both the PIE and Banca data sets. We observe the following:

1. The more training samples we use, the better the recognition rate. This is consistent with our experiments (in Table 3) and it has also been confirmed by other researchers. The reason is that a larger set of
training data can sample the underlying distribution more accurately than a smaller set.

2. With fewer training samples per class, say, two or four, the highest accuracy is obtained at \( C - 1 \) projected dimensions (\( C = 67 \) for PIE, \( C = 52 \) for Banca).

3. However, with more training samples per class, say, six or more, we can obtain a high classification rate with fewer than \( C - 1 \) projected dimensions. For example, on the PIE data set, with eight or 10 training samples per class, we can obtain 98 percent accuracy by using only 30 projected dimensions (Fig. 9a). The curves remain flat with increasing dimensions. Thus, there is no incentive to use more than 30 dimensions.

Now, let us summarize our experiments on real data sets. First, MDA/FKT is comparable to LDA-based methods with respect to the accuracy. Second, LDA/FKT and MDA/FKT significantly outperform other LDA-based methods for small-sample-size problems.

7 Conclusion

In this paper, we showed how FKT can provide valuable insights into LDA. We derived and proved the relationship between GSVD, FKT, and LDA and then unified different methods, that is, Fisherface, PCA+NULL, LDA/QR, and LDA/GSVD. Our theoretical analyses showed how to choose the discriminant subspaces based on the generalized eigenvalue, the essential measure of separability.

4. We also compared some common LDA methods with LDA/FKT. Most of these methods are suboptimal in terms of the Fisher Criterion. More specifically, Fisherface, PCA+NULL, and LDA/QR all operate in different parts of the discriminative subspaces of LDA/FKT. We showed that LDA/GSVD and LDA/FKT are, in fact, equivalent, but our LDA/FKT is more efficient than LDA/GSVD with respect to computation.

5. We further presented MDA/FKT with the following properties:

a. It is derived from the Bhattacharyya distance, which is the error bound of the Bayes Classifier. This is theoretically superior to the Fisher Criterion, which is based on scatter matrices and which does not relate to the Bayes Classifier.

b. It can provide larger discriminative subspaces; in contrast, LDA-based methods are limited by the number of classes.

c. It works even if \( S_b = 0 \), which is where LDA-based methods fail. Furthermore, for Gaussian mixture pdf, it works better than LDA.

d. It can be realized by an efficient algorithm. This algorithm is comparable to most of LDA-based methods with respect to computation and storage.

6. We experimentally showed the superiority of LDA/FKT and MDA/FKT. In particular, for small-sample-size problems, LDA/FKT and MDA/FKT work significantly better than other methods. In the case of MDA/FKT, we further observed that using a small projected subspace (dimension \( \approx \frac{C}{2} \)) is enough to achieve high accuracy when the training set is sufficiently large.

An interesting future work is to extend our theory to a nonlinear discriminant analysis. One way is to use the kernel trick employed in support vector machines (SVMs), for example, construct kernelized between-class scatter and within-class scatter matrices. FKT may yet again reveal new insights into the kernelized LDA.

Appendix A

proof of lemma 1

Based on GSVD,

\[
S_b = H_bH_b^T = X^{-1} \begin{bmatrix} \Sigma_b & 0 \\ 0 & 0 \end{bmatrix} X^{-1},
\]

\[
S_w = H_wH_w^T = X^{-1} \begin{bmatrix} \Sigma_w & 0 \\ 0 & 0 \end{bmatrix} X^{-1}.
\]

Thus,

\[
X^T (S_b + S_w) X = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.
\]

Since \( \Sigma_b + \Sigma_w = I \), if we choose the first \( r_1 \) columns of \( X \) as \( P \), that is, \( P = X_{(d,r_1)} \), then \( P^T(S_b + S_w)P = I \). This is exactly FKT. Meanwhile, we can obtain that \( \Lambda_b = \Sigma_b^2 \Sigma_b \) and \( \Lambda_w = \Sigma_w^2 \Sigma_w \).

**FTK \rightarrow GSVD.**

Based on FKT \( P = UD^{-1/2} \),

\[
\tilde{S}_b = P^T S_b P = D^{-1/2} U H_b H_b^T U D^{-1/2},
\]

\[
\tilde{S}_b = V \Lambda_b V^T.
\]

Hence,

\[
D^{-1/2} U H_b H_b^T U D^{-1/2} = V \Lambda_b V^T.
\]

In general, there is no unique decomposition on the above equation because \( H_b H_b^T = H_b Y Y^T H_b^T \) for any orthogonal matrix \( Y \in \mathbb{R}^{C \times C} \). That is,

\[
D^{-1/2} U H_b Y Y^T H_b^T U D^{-1/2} = V \Lambda_b V^T,
\]

\[
Y^T H_b U D^{-1/2} = \tilde{S}_b V^T,
\]

\[
Y^T H_b U D^{-1/2} V = \tilde{S}_b,
\]

where \( \tilde{S}_b \in \mathbb{R}^{C \times C} \) and \( \Lambda_b = \tilde{S}_b^2 \tilde{S}_b \). If we define \( X = [UD^{-1/2}V, U_\perp] \in \mathbb{R}^{d \times d} \). Then,
Here, $H^T U_\perp = 0$, and $H^T U = 0$ because $U_\perp$ is the intersection of the null space of $S_t$ and $S_w$. Similarly, we can get $Z^T H^T X = [S_w, 0]$, where $Z \in \mathbb{R}^{r \times r}$ is an arbitrary orthogonal matrix, $S_w \in \mathbb{R}^{r \times r}$, and $A_w = S_w^T S_w$. Since $A_b + A_w = I$ and $I \in \mathbb{R}^{r \times r}$ is an identity matrix, it is easy to check that $S_I^T S_I + S_w^T S_w = I$, which satisfies the constraint of GSVD.

Now, we have to prove $X$ is nonsingular

$$XX^T = [UD^{-1/2}V, U_\perp] \begin{bmatrix} V^T D^{-1/2} U^T \end{bmatrix}$$

$$= [UD^{-1/2} U^T U_\perp] U_\perp^T$$

$$= [U, U_\perp] \begin{bmatrix} D^{-1/2} & 0 \\ 0 & 1 \end{bmatrix} U_\perp^T.$$

Here, $V \in \mathbb{R}^{r \times r}$ and $[U, U_\perp]$ are orthogonal matrices. Note that $U^T U_\perp = 0$ and $U_\perp U^T = 0$. From the above equation, $XX^T$ can be eigendecomposed with positive eigenvalues, which means $X$ is also nonsingular. This completes the proof. \[\Box\]

**APPENDIX B**

**PROOF OF LEMMA 2**

Proof. Since

$$\Sigma_i = \frac{1}{N_I} \sum_{L(a_i)L(a_j) = L(a_i)} (a_i - a_j)(a_i - a_j)^T,$$

$$\Sigma_E = \frac{1}{N_E} \sum_{L(a_i)\neq L(a_j)} (a_i - a_j)(a_i - a_j)^T.$$

Then,

$$N_I \Sigma_I + N_E \Sigma_E = \sum_i \sum_j (a_i - a_j)(a_i - a_j)^T$$

$$= \sum_i \sum_j (a_i a_j^T - a_i a_j^T - a_j a_i^T + a_j a_j^T)$$

$$= 2N \sum_i (a_i a_i^T) - 2 \left( \sum_i a_i \right) \left( \sum_i a_i^T \right)$$

$$= 2N \sum_i (a_i a_i^T) - 2N^2 \mu \mu^T$$

$$= 2N \left( \sum i a_i a_i^T - N \mu \mu^T \right),$$

where $\mu = \frac{1}{N} \sum_i a_i$ is the total mean of the samples.

On the other hand,

$$S_t = \sum_i (a_i - \mu)(a_i - \mu)^T$$

$$= \sum \left( a_i a_i^T - a_i \mu^T - \mu a_i^T + \mu \mu^T \right)$$

$$= \sum a_i a_i^T - N \mu \mu^T.$$

By examining (48) and (49), we can see that

$$2N S_t = N_I \Sigma_I + N_E \Sigma_E.$$

**REFERENCES**


Sheng Zhang received the bachelor’s degree from Zhejiang University, China, in 1998, the Master’s degree from the Institute of Automation, Chinese Academy of Sciences, in 2001, and the PhD degree from the School of Computing, National University of Singapore in 2006. He now works in the Department of Electrical and Computer Engineering, University of California at Santa Barbara (UCSB) as a postdoctoral researcher. His research interests include face recognition, computer vision, statistical pattern recognition, and machine learning. He is a member of the IEEE and the IEEE Computer Society.

Terence Sim received the BS degree in computer science and engineering from the Massachusetts Institute of Technology in 1990, the MS degree in computer science from Stanford University in 1991, and the PhD degree in electrical and computer engineering from Carnegie Mellon University in 2002. He is an assistant professor in the School of Computing, National University of Singapore. His research interests are in biometrics, face recognition, computer vision, digital photography, and music processing. He chairs the Workgroup on Cross-Jurisdictional and Societal Aspects in the Biometrics Technical Committee, Singapore. He is a member of the IEEE.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.