A Comparative Study on Vector-based and Matrix-based Linear Discriminant Analysis

Bo Yang
Department of Information Engineering, Hunan Institute of Science and Technology, Yueyang, China
Email: yb mengshen@163.com

Ying-yong Bu
College of Mechanical and Electrical Engineering, Central South University, Changsha, China
Email: byy29@yahoo.com.cn

Abstract—Recently a kind of matrix-based discriminant feature extraction approach called 2D LDA have been drawn much attention by researchers. 2D LDA can avoid the singularity problem and has low computational costs and has been experimentally reported that 2D LDA outperforms traditional LDA. In this paper, we compare 2D LDA with LDA in viewpoint of the discriminant power and find that 2D LDA as a kind of special LDA has no stronger discriminant power than LDA. So, why 2D LDA outperforms LDA in some cases? Through theoretical analysis, we find it is mainly because of the difference of stability under nonsingular linear transformation and linear operation power between 2D LDA and LDA. In experimental parts, the results of experiments give enough proof on our claims and show in some cases the performance of 2D LDA will be possible superior to that of LDA and in other cases the performance of LDA will be possible superior to that of 2D LDA.

Index Terms—Feature Extraction, LDA, 2D LDA

I. INTRODUCTION

Feature extraction is an important research field in pattern recognition, through which we can delete useless information and reduce the dimensionality of data effectively. Many feature extraction methods such as principal component analysis (PCA), linear discriminant analysis (LDA), independent component analysis (ICA), locality preserving projection (LPP), etc have been widely researched in pattern recognition fields.

Among the above mentioned methods, LDA is a kind of supervised feature extraction method which shows good performance to classification tasks. Its main idea is to find the projective vectors which have the largest between-class distance and the shortest within-class distance. However, LDA will fail when the small sample size problem occurs. To deal with this problem, some effective approaches have been proposed such as PCA+LDA[1], Nullspace LDA[2-4], Regularized LDA[5-8], etc. The main idea of PCA+LDA method is reducing dimensionality of samples using PCA firstly in order to generate a full-ranked within-class matrix in reduced dimensional space, and then using LDA in this transformed space. The main idea of Nullspace LDA is searching Null space of within-class matrix firstly, and then extracting the discriminant information from between-class matrix in this Null space. The main idea of regularized LDA is generating a new full-ranked within-class matrix by adding a minor perturbation diagonal matrix to original within-class matrix.

The above feature extraction methods all are vector-based. Recently some matrix-based feature extraction methods have been proposed in image recognition research field such as two-dimensional principle component analysis (2DPCA)[9-10], two-dimensional linear discriminant analysis (2D LDA)[11-15], two-dimensional locality preserving projection (2D LPP)[16-17], etc. Because 2D LDA works in low dimensional space it can avoid the small sample size problem effectively and can achieve higher computational efficiency than LDA. Besides, the 2D LDA based algorithms have been experimentally reported even superior to traditional LDA based algorithms.

Are 2D LDA based algorithms always superior to traditional LDA based algorithms? If the answer is negative so why 2D LDA based algorithms are superior to traditional LDA based algorithms sometimes? Recently, there are some researchers who have made theoretical comparison between 2D LDA and LDA and tried to answer the above two questions. Zheng[18] compared 2D LDA with LDA from the statistical point of view. He indicated that 2D LDA is a kind of feature extraction method which loses covariance information and will be confronted with the “Hetertoscedastic Problem” more seriously than LDA. As for the second question, they think it is mainly because that 2D LDA has more training “row/column samples” to be used which means 2D LDA might be more stable from the bias estimation point of view. Besides, Liang[19] compared 2D LDA with LDA in viewpoint of discriminant power. He indicated that 2D LDA is a kind of special LDA. So the discriminant power of 2D LDA is not stronger than that of LDA when considering the same dimensionality. As for the second question, they think the reason is the training samples size is too small. When the training samples size is large enough, the LDA based algorithms will always superior than 2D LDA based algorithms.
However, the key theorem in Liang’s paper is not correct. In this paper, we continue to compare 2DLD A with LDA and try to answer the above two questions in view of discriminant power in different view. Firstly, we compare 2DLD A with LDA in view of discriminant power again by using a different criteria in contrast to Liang’s paper[19] and indicate that the discriminant power of 2DLD A is not stronger than that of LDA. Then we discuss the stability of 2DLD A and LDA in view of nonsingular linear transformation and the difference of linear operation power between 2DLD A and LDA. Through our theoretical and experimental analysis, we find that because of the difference of linear operation power and stability between 2DLD A and LDA the 2DLD A based algorithms are superior to the LDA based algorithms in some cases.

II. RELATED WORK

Supposing 1D samples are \{x_1, \ldots, x_m\} \ (x_j \in \mathbb{R}^{\text{col}}) , relative 2D samples are \{x_1^{(d)}, \ldots, x_m^{(d)}\} \ (x_j^{(d)} \in \mathbb{R}^{\text{row} \times \text{col}}) . For a C-class classification problem, the 1D between-class scatter matrix \(S_b\) and the 1D within-class scatter matrix \(S_w\) are defined as \(S_b = \sum_{i=1}^{C} m_i (m_i - \mu) (m_i - \mu)^T\) and \(S_w = \sum_{i=1}^{C} \sum_{j \neq i}^{C} (x_j^{(d)} - \mu) (x_j^{(d)} - \mu)^T\), where \(x_j^{(d)}\) is the jth 1D sample of class i, \(m_i\) is the mean vector of 1D samples of class i, \(m_0\) is the mean vector of all 1D samples, and \(m^{(d)}\) is the number of samples of class i. LDA method tries to find the most discriminant projection \(w_{\text{opt}}\) which can be defined as below:

\[
J_1(w) = \max \frac{\text{Tr}(w^T S_b w)}{\text{Tr}(w^T S_w w)}
\]  

(1)

The above criteria can be solved by the generalized eigenvalue problem \(S_b w_i = \lambda_i S_w w_i\). When \(S_w\) is a full rank matrix, it can be rewritten as \(S_w^{-1} S_b w_i = \lambda_i w_i\). Let \(w_i, w_j (i \neq j)\) are the ith and jth best discriminant vector, we have \(w_i^T S_w w_i = 1\ , \ w_j^T S_w w_j = \lambda_i\ , \ w_i^T S_w w_j = 0\ , \ w_j^T S_w w_j = 0\). Supposing eigenvalue \(\lambda_i \geq \cdots \geq \lambda_c > 0\) and \(\lambda_{c+1} = \cdots = \lambda_0 = 0\ , \) when the first r eigenvectors are selected we have \(w_{\text{opt}} = (w_1 \cdots w_r)\). Hence we have \(J_1(w_{\text{opt}}) = \sum_{i=1}^{r} \lambda_i / r\).

Besides, when \(S_w\) is a full-ranked matrix we can use \(\text{Tr}(S_w^{-1} S_b)\) to measure the class separability of 1D samples. Clearly we have

\[
\text{Tr}(S_w^{-1} S_b) = \text{Tr}(w_{\text{opt}}^T S_w w_{\text{opt}}) = (w_{\text{opt}}^T S_w w_{\text{opt}})
\]  

(2)

For C-class classification problem, the 2D between-class scatter matrix \(S_{2b}^{2d}\) , \(S_{2b'}^{2d}\) and the 2D within-class scatter matrix \(S_{2w}^{2d}\) , \(S_{2w'}^{2d}\) are defined as below:

\[
S_{2b}^{2d} = \sum_{i=1}^{C} m_i^{(d)} (m_i^{(d)} - m_{\text{mean}}^{(d)}) (m_i^{(d)} - m_{\text{mean}}^{(d)})^T
\]  

(3)

\[
S_{2r}^{2d} = \sum_{i=1}^{C} m_i^{(d)} (m_i^{(d)} - m_{\text{mean}}^{(d)}) (m_i^{(d)} - m_{\text{mean}}^{(d)})^T
\]  

(4)

\[
S_{2c}^{2d} = \sum_{i=1}^{C} \sum_{j \neq i}^{C} (x_j^{(d)} - m_i^{(d)}) (x_j^{(d)} - m_i^{(d)})^T
\]  

(5)

\[
S_{2w}^{2d} = \sum_{i=1}^{C} \sum_{j \neq i}^{C} (x_j^{(d)} - m_i^{(d)}) (x_j^{(d)} - m_i^{(d)})^T
\]  

(6)

\(x_j^{(d)}\) is the jth 2D sample of class i, \(m_i^{(d)}\) is the mean vector of 2D samples of class i, \(m_{\text{mean}}^{(d)}\) is the mean vector of all 2D samples, and \(L/R\) is the left/ right transformation matrix. 2DLD A method tries to find the most discriminant projection \(L_{\text{opt}}, R_{\text{opt}}\) which can be defined as below:

\[
J_1(L, R) = \max \frac{\text{Tr}(L^T S_{2b}^{2d} L)}{\text{Tr}(L^T S_{2w}^{2d} L)} = \max \frac{\text{Tr}(R^T S_{2b}^{2d} R)}{\text{Tr}(R^T S_{2w}^{2d} R)}
\]  

(7)

When \(R\) is a matrix constant and \(R = I^{\text{row} \times \text{col}}\) , it is called Left 2DLD A and its 2DLD A criterion can be written as:

\[
J_1(L, I) = \max \frac{\text{Tr}(L^T S_{2b}^{2d} L)}{\text{Tr}(L^T S_{2w}^{2d} L)}
\]  

(8)

Where \(S_{2b} = \sum_{i=1}^{C} m_i^{(d)} (m_i^{(d)} - m_{\text{mean}}^{(d)})(m_i^{(d)} - m_{\text{mean}}^{(d)})^T\) , \(S_{2w} = \sum_{i=1}^{C} \sum_{j \neq i}^{C} (x_j^{(d)} - m_i^{(d)})(x_j^{(d)} - m_i^{(d)})^T\) . Like 1D methods, Left 2DLD A can be solved in one step by solving the generalized eigenvalue problem \(S_{2b} L = \lambda S_{2w} L\).

When \(L\) is a matrix constant and \(L = I^{\text{row} \times \text{col}}\) , it is called Right 2DLD A and its 2DLD A criterion can be written as:

\[
J_1(I, R) = \max \frac{\text{Tr}(R^T S_{2b}^{2d} R)}{\text{Tr}(R^T S_{2w}^{2d} R)}
\]  

(9)

Where \(S_{2b} = \sum_{i=1}^{C} m_i^{(d)} (m_i^{(d)} - m_{\text{mean}}^{(d)})(m_i^{(d)} - m_{\text{mean}}^{(d)})^T\) , \(S_{2w} = \sum_{i=1}^{C} \sum_{j \neq i}^{C} (x_j^{(d)} - m_i^{(d)})(x_j^{(d)} - m_i^{(d)})^T\) . Like 1D methods, Right 2DLD A can be solved in one step by solving the generalized eigenvalue problem \(S_{2b} R = \lambda S_{2w} R\).

When \(L, R\) both are matrix variables, it is called Bilateral 2DLD A. It is hard to find its global resolution and can only be solved locally by solving Left 2DLD A and Right 2DLD A problem in several times in iterative way.

In Liang’s paper[19], they indicated that 2D methods are a kind of special 1D methods (see[19], the equation (17~21)): 

\[
\text{vec}(L_{\text{opt}}^T x^{(2d)} R_{\text{opt}}) = (R_{\text{opt}}^T \otimes L_{\text{opt}}^T) x
\]  

(10)

Where \text{vec()}\ denotes the \text{vec} operator which convert the matrix into a vector by stacking the columns of the matrix. Hence, (7) can be rewritten as:

© 2011 ACADEMY PUBLISHER
\[ J_2(L, R) = \max \frac{\text{Tr}((R^T \otimes L^T)S_y(R \otimes L))}{\text{Tr}((R^T \otimes L^T)S_y(R \otimes L))} \quad (11) \]

We can use a formation like (2) to measure the class separability of 2D samples:
\[ \text{Tr}((R_{opt}^T \otimes L_{opt}^T)S_y(R_{opt} \otimes L_{opt})) = \text{Tr}((R_{opt}^T \otimes L_{opt}^T)(S_y(R_{opt} \otimes L_{opt}))) \quad (12) \]

(12) means after 2DLDA transformation LDA is used again on transformed samples \( (R_{opt}^T \otimes L_{opt})x \). Let \( A = R_{opt}^T \otimes L_{opt}^T \), we have
\[ \text{Tr}((AS_yA^T)^{-1}AS_yA^T) = \text{Tr}((w_{opt}^T AS_yA^T w_{opt})^{-1}(w_{opt}^T AS_yA^T w_{opt})) \quad (13) \]

Where \( w_{opt} \) are the eigenvectors of the generalized eigenvalue problem \( AS_yA^T w = \gamma AS_yA^T w \). So we have \( \text{Tr}((AS_yA^T)^{-1}AS_yA^T) = \sum \gamma_i \).

### III. THEORETICAL ANALYSIS BETWEEN 2DLDA AND LDA

In this section, we compare LDA with 2DLDA in view of discriminant power. This concept was first introduced by Liang et al [19]. In their paper, Liang indicated that 2DLDA has no stronger discriminant power than LDA. This conclusion is right. However, Theorem 1 as the main theoretical proof on this conclusion in their paper is not correct. We think that mainly because the comparison criteria selected in their paper are not appropriate. So we use another criteria to measure the discriminant power of LDA and 2DLDA. Besides, we analyze the stability of 2DLDA and LDA under nonsingular linear transformation and the linear operation power of 2DLDA and LDA. We also indicate the attributes of 2DLDA and LDA which lead to difference performances.

### k. the discriminant power of 2DLDA and LDA

Liang[19] compared the discriminant power of LDA with 2DLDA using criterion (1) and criterion (11). They tried to prove a theorem that \( J_1(L, R) \leq J_1(w) \) when the dimensionally reduced samples using LDA and 2DLDA are of the same dimensionality. However, this theorem is not right in general case.

Here we give a counterexample about this theorem.

Suppose our vector samples are \( \{x_i\} \) and we use \( J_1(w) \) to reduce sample \( x_i \) to 2-dimensional vector sample. In this case, we have \( J_1(w_{opt}) = (\lambda_i + \lambda_j) / 2 \), where \( \lambda_i, \lambda_j \) are the largest two eigenvalues related to general eigenvalue problem \( S_y w = \lambda S_y w \).

Let’s construct new vector samples \( \{y_i\} \) (\( y_i = (x_i^T, x_i^T) \)) and matrix samples \( \{y_i^{2D}\} \) (\( y_i^{2D} = (x_i, x_i) \)). For \( \{y_i\} \), we also have \( J_1(w_{opt}) = (\lambda_i + \lambda_j) / 2 \) clearly. For \( \{y_i^{2D}\} \), Let \( R = I \). When matrix samples are also reduced to two dimensional samples we have \( J_2(L_{opt}, I) = (\lambda_i + \lambda_j) / 2 = \lambda_i \geq J_1(w_{opt}) \).

In the case of this counterexample, suppose the two eigenvectors related to eigenvalue \( \lambda_i, \lambda_j \) on samples \( \{x_i\} \) are \( w_1, w_2 \), we have the dimension reduced result on \( y_j^{1D} \) is \( \left(\begin{array}{c} w_1^T x_j \\ w_2^T x_j \end{array}\right) \) and the dimension reduced result on \( y_j^{2D} \) is \( \left(\begin{array}{c} w_1^T x_j \ w_1^T x_j \end{array}\right) \). Clearly although \( J_2(L_{opt}, I) \geq J_1(w_{opt}) \) the discriminant power of LDA is also stronger than that of 2DLDA. So we think that this counterexample means the criteria for measuring discriminant power used in their paper are not appropriate.

So we have to choose another discriminant power measurement of LDA and 2DLDA. Here we use (2) and (12) as the measures of the discriminant powers of LDA and 2DLDA. Using these measures, we have the discriminant power of LDA is \( \lambda_i + \lambda_j \) and the discriminant power of 2DLDA is \( \lambda_i \) for the counterexample above.

Now we start our analysis through the below Lemma.

**Lemma 1** [20]. Supposing matrix \( A (A \in R^{m \times n}) \) is a Hermitian matrix, \( \lambda_i \ (1 \leq i \leq n) \) is its ith eigenvalue and \( \lambda_i \geq \cdots \geq \lambda_n \), \( w_1, \cdots, w_n \) is its eigenvectors and \( A w_i = \lambda_i w_i \). Supposing \( x_1, \cdots, x_n \) satisfy \( x_i^T x_j = 0 (i \neq j) \), then
\[
\sum_{i=1}^{n} x_i^T A x_i \leq \sum_{i=1}^{n} \lambda_i = \text{Tr}(A) ;
\]
If and only if \( x_i \propto w_i (1 \leq i \leq n) \), then
\[
\sum_{i=1}^{n} x_i^T A x_i = \sum_{i=1}^{n} \lambda_i = \text{Tr}(A) .
\]

According to Lemma 1, we can prove Lemma 2 as follows.

**Lemma 2.** Suppose marices \( A (A \in R^{m \times n}) \) and \( B (B \in R^{m \times n}, B > 0) \) as Hermitian matrices, \( \lambda_i \ (1 \leq i \leq n) \) as the ith eigenvalue of matrix \( B^{-1}A \) satisfying \( \lambda_i \geq \cdots \geq \lambda_n \), \( w_1, \cdots, w_n \) as its eigenvectors in \( B^{-1}A w_i = \lambda_i w_i \) and suppose \( x_1, \cdots, x_n \) satisfy \( x_i^T B x_j = 0 (i \neq j) \), then
\[
\sum_{i=1}^{n} x_i^T A x_i \leq \sum_{i=1}^{n} \lambda_i = \text{Tr}(B^{-1}A) ;
\]
If and only if \( x_i \propto w_i (1 \leq i \leq n) \), then
\[
\sum_{i=1}^{n} x_i^T A x_i = \sum_{i=1}^{n} \lambda_i = \text{Tr}(B^{-1}A) .
\]

**Proof.** Supposing \( y_i = B^{1/2} x_i \), we have
\[
\sum_{i=1}^{n} y_i^T A y_i = \sum_{i=1}^{n} y_i^T B^{-1/2} A B^{-1/2} y_i.
\]
Because \( x_i^T B x_j = 0, i \neq j \), we have \( y_i^T y_j = 0, i \neq j \).

Noticing that matrix \( B^{-1/2} A B^{-1/2} \) also is a Hermitian matrix. According to Lemma 1, we have
Let $B$ be a full-ranked matrix. After $B$, original 2D samples $B$ are changed into $L$ and $TT$ is invariant when samples $B$ are changed into new samples $L$. According to Lemma 2, the relative generalized eigenvalue problem is

$$\sum_i^n y_i^T B^{-1/2} A B^{-1/2} y_i = \lambda_i y_i,$$

where $\lambda_i$ is an eigenvalue and $y_i$ is an eigenvector.

Notice $y_i = B^T x_i$, we have

$$\sum_i^n x_i y_i^T A x_i = \sum_i^n \lambda_i = \text{trace}(B^{-1} A).$$

This completes the proof.

Let $A = R_{opt} \otimes L_{opt}$. For 2DLDA, its discriminant power value is $\text{trace}(A^T S_A^{-1} A^T S_A A)$ . Its relative generalized eigenvalue problem is $A^T S_A^{-1} A w_i = \lambda_i A^T S_A A w_i$ and its eigenvectors satisfy $w_i^T A^T S_A A w_i = 0 (i \neq j)$.

For LDA, its discriminant power value is $\text{trace}(S_A^{-1} S_\mu)$. Its relative generalized eigenvalue problem is $S_A w_i = \lambda_i S_A w_i$ and its eigenvectors satisfy $w_i^T S_A w_i = 0 (i \neq j)$.

The two basis $\{A w_i\}$ and $\{w_i\}$ both are conjugate orthogonal basis of matrix $S_{\mu}$. According to Lemma 2, when the inverse matrix $S_{\mu}^{-1}$ exists we have

$$\text{trace}(S_{\mu}^{-1} S_\mu) \geq \text{trace}((A^T S_A A)^{-1}(A^T S_A A)).$$

Hence we can draw a conclusion that the discriminant power of LDA is always larger than the discriminant power of 2DLDA.

6. The stability of 2DLDA and LDA under nonsingular linear transformation and the linear operation power of 2DLDA and LDA

Let the matrix $D$ be a full-ranked matrix. After nonsingular linear transformation using $D$, original 1D vector samples $x$ are changed into new samples $Dx$ . Like this, let the matrices $A, B$ be full-ranked matrices. After left/right nonsingular linear transformation using $A$ and $B$, original 2D samples $x$ are changed into new samples $Ax$ and $Bx$. Here we have the below theorems about LDA and 2DLDA.

**Theorem 1.** LDA is invariant under any nonsingular linear transformation on 1D samples.

**Proof.** For transformed samples $Dx$, the relative generalized eigenvalue problem can be written as

$$DS^T D w_i = \lambda_i DS^T w_i.$$

Because $D$ is a nonsingular linear transformation matrix there exists inverse matrix $D^{-1}$. Let $w_i = D^T w_i$ , we have $S_d w_i = \lambda_i S_d w_i$.

So we have $w_i^T x = w_i^T (Dx)$ . It means that the LDA results are invariant under nonsingular linear transformation.

This completes the proof.

From Theorem 1, we can see that a nonsingular linear transformation on samples has no influence on LDA. So the results of classification do not change. Like the proof of Theorem 1, from their relative generalized eigenvalue problem we also have the above theorems on Left/Right 2DLDA.

**Theorem 2.** Left/Right 2DLDA is invariant under any right/left nonsingular linear transformation on 2D samples.

As for Bilateral 2DLDA, supposing the first step of the whole solving process of Bilateral 2DLDA is Right/Left 2D-LDA clearly we have it is invariant under any right/left nonsingular linear transformation according to Theorem 2. However, Left/Right 2DLDA is not invariant under any right/left nonsingular linear transformation any more except that the transformation is unit orthogonal transformation in sense of Euclidian Distance which is proved in the below theorem.

**Theorem 3.** Left/Right 2DLDA is invariant in sense of Euclidian Distance under any right/left unit orthogonal transformation on 2D samples.

**Proof.** To Left 2DLDA, 2D samples $x^{2d}$ are transformed as $x^{2d}B$ after right linear transformation. So the relative generalized eigenvalue problem can be written as

$$S_d, 1 L = \lambda_i S_d, 1 L ;$$

$$= \sum_{j=1}^m (m_j - m_{2d}) B B^T (m_j^{2d} - m_{2d}) \gamma ,$$

and its eigenvectors satisfy $w_i^T S_d, 1 L w_i = 0 (i \neq j)$.

The transformed sample is $L x^{2d} B$. To any two of the transformed samples $L x^{2d} B$ and $L x_j^{2d} B$, the Euclidian distance between them is:

$$D_{ij}^2 = || L (x_i^{2d} - x_j^{2d}) ||_2^2$$

$$= \text{trace}(L (x_i^{2d} - x_j^{2d}) B B^T (x_i^{2d} - x_j^{2d}) \gamma L)$$

$$= \text{trace}(L (x_i^{2d} - x_j^{2d})(x_i^{2d} - x_j^{2d}) \gamma L)$$

$$= || L (x_i^{2d} - x_j^{2d}) ||_2^2$$

So right unit orthogonal transformation has no influence to Left 2DLDA on dimensional reduced results in sense of Euclidian Distance. Like this, for Right 2DLDA we also have the analogous conclusion.

This completes the proof.

As for Bilateral 2DLDA, supposing the first step of the whole solving process of Bilateral 2DLDA is Right/Left 2DLDA we also have it is invariant under any left/right unit orthogonal transformation according to Theorem 3.
However, when unit orthogonal transformation is relaxed to nonsingular linear transformation the results using 2D LDA methods are not invariant anymore. It will lead to the difference in the results of classification after different nonsingular linear transformation on samples.

Besides, 1D methods and 2D methods have the difference of the power of linear operation. Let 1D sample be \( x = [x_{1,1} \cdots x_{1,n} \cdots x_{m,1} \cdots x_{m,n}]^T \) and relative 2D sample be \( x^{2d} = \begin{pmatrix} x_{1,1} \cdots x_{1,n} \\ \vdots \\ x_{m,1} \cdots x_{m,n} \end{pmatrix} \). So the linear operation on 1D sample \( x \) on vector \( w \) is \( w^T x \) and the linear operation of 2D sample \( x^{2d} \) on vectors \( l, r \) is \( l^T x^{2d} r \). Clearly 1D methods have the whole power of linear operation while 2D methods have not. For example, using linear operation on 1D sample \( x \), we can obtain a certain linear combination like \( x_{1,1} + x_{2,2} \). However, we cannot find any vectors \( l, r \) to generate this linear combination on 2D sample \( x^{2d} \).

From above analysis, we can see that the performance of 2D LDA on the samples under different full ranked linear transformation will be different. If the discriminant information of transformed samples were mainly located along column/row direction, the performance of 2D LDA would be satisfying. However, the discriminant information of transformed samples were not mainly located along column/row direction, the performance of 2D LDA would be degenerate.

Besides, because of the deficiency of linear operation power of 2D LDA, 2D LDA can not abstract all discriminant information. However, when the small sample size problem occurs, the whole discriminant information will contain some illusive discriminant information. In this case, 2D LDA can avoid the influence of illusive discriminant information come from different columns/rows effectively and the performance of 2D LDA will be possible to superior to that of LDA.

IV. Experiments Results

In this section, we do comparative experiments on an artificial dataset and ORL face dataset[21]. In our experiments, we choose the nearest-neighbor (NN) classifier. In the experiments on artificial dataset and ORL dataset, because the matrix \( S_u \) is not full-ranked the inverse matrix of \( S_u \) does not exist. Here a kind of regularized LDA is used. In regularized LDA method[22], \( S_u \) is replaced into \( S_u^{-1} = S_u + \lambda I \). Where \( \lambda \) is a regularized parameter, \( I \) is a unit matrix. In our experiments, the regularized parameter \( \lambda \) is fixed as \( \lambda = 0.0001 \). All the algorithms are developed using Matlab 6.5.

\( \lambda \), experimental counterexample about Liang’s theorem

Here we generate vector samples \( \{y_i\}_{i=1}^{232} (y_i \in R^{30}) \) which belong to contain three classes and construct matrix samples \( \{y_i^{2d}\}_{i=1}^{232} (y_i^{2d} \in R^{50}) \) using \( \{y_i\} \). The mean vector \( m_1 \) of class 1 is \( (0 \cdots 0)^T \); The mean vector \( m_2 \) of class 2 is \( (1 \cdots 1)^T \); The mean vector \( m_3 \) of class 3 is \( (2 \cdots 2)^T \); Every dimension is normally distributed and the variance of every dimension is 0.01. The number of samples per class is 50.

We use 1D LDA reducing \( \{y_i\} \) to 2-dimensional vector samples and use 2D LDA reducing \( \{y_i^{2d}\} \) to 1×2-dimensional matrix samples. For matrix-based linear discriminant analysis, here we do Right 2D LDA and Left 2D LDA only one time.

We do this experiment 10 times and the experimental results are shown in Table 1. From Table 1, we can see that at the \( 1^{st}, 3^{rd}, 5^{th}, 10^{th} \) steps we have \( J_2(J_{opt}, R_{opt}) \geq J_1(w_{opt}) \) in our experiments. Here \( J_1 \) and \( J_2 \) value are calculated by using (1) and (11).

<table>
<thead>
<tr>
<th>Test No</th>
<th>( J_1 ) Value</th>
<th>( J_2 ) Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>208.02</td>
<td>225.41</td>
</tr>
<tr>
<td>2</td>
<td>249.24</td>
<td>243.32</td>
</tr>
<tr>
<td>3</td>
<td>194.83</td>
<td>197.1</td>
</tr>
<tr>
<td>4</td>
<td>206.51</td>
<td>203.42</td>
</tr>
<tr>
<td>5</td>
<td>163.84</td>
<td>179.32</td>
</tr>
<tr>
<td>6</td>
<td>218.94</td>
<td>212.65</td>
</tr>
<tr>
<td>7</td>
<td>260.91</td>
<td>244.91</td>
</tr>
<tr>
<td>8</td>
<td>231.24</td>
<td>224.34</td>
</tr>
<tr>
<td>9</td>
<td>231.05</td>
<td>221.85</td>
</tr>
<tr>
<td>10</td>
<td>240.03</td>
<td>245.66</td>
</tr>
</tbody>
</table>

§. Comparative experiment on artificial dataset

Here we construct a three-class classification task. The 2D sample \( x^{2d} \in R^{50 \times 30} \), the distribution of \( x_{ij} \) in each class is normal distribution. In this case only the diagonal element \( x_{ij} \) contains discriminant information. The mean of \( x_{ij} (i \neq j) \) is 0. The variance of \( x_{ij} (i \neq j) \) is 0.1. The mean values of diagonal element \( x_{ij} \) in class 1, class 2 and class 3 are 0.1, 0.4 and 0.8. The variance value of \( x_{ij} \) is 0.01. This experiment is repeated 10 times and in each time we generate 30 samples per class.

In our experiment, when sample \( x^{2d} \) is reduced to \( y^{2d} \in R^{50 \times 30} \) and \( y^{2d} \in R^{17 \times 30} \) using Left/Right 2D LDA and Bilateral 2D LDA the best classification performance is obtained. In this case the number of dimension is not reduced apparently using 2D LDA. As shown in Table 2, we can find that the decent rate of eigenvalues is low which means the dimensional reduced efficiency using 2D LDA is weak in this case.
Besides, as shown in Fig 1, the best classification performance using LDA is apparently superior to it using 2DLDA.

We do this experiment 10 times and the experimental results are shown in Table 1. From Table 1, we can see that at the 1st, 3rd, 5th, 10th steps we have \( J_2(I_{opt}, R_{opt}) \geq J_1(w_{opt}) \) in our experiments. Here \( J_1 \) and \( J_2 \) value are calculated by using (1) and (11).

Table II.

<table>
<thead>
<tr>
<th></th>
<th>Left</th>
<th>Right</th>
<th>Bilateral</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0045</td>
<td>0.0047</td>
<td>0.0141</td>
</tr>
<tr>
<td>5</td>
<td>0.0036</td>
<td>0.0034</td>
<td>0.0079</td>
</tr>
<tr>
<td>10</td>
<td>0.0026</td>
<td>0.0026</td>
<td>0.0034</td>
</tr>
<tr>
<td>15</td>
<td>0.0018</td>
<td>0.0019</td>
<td>0.0014</td>
</tr>
<tr>
<td>20</td>
<td>0.0014</td>
<td>0.0014</td>
<td>0.0005</td>
</tr>
<tr>
<td>25</td>
<td>0.0008</td>
<td>0.0009</td>
<td>0.0002</td>
</tr>
<tr>
<td>30</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Figure 1. Comparisons of Average Error Rates between using 2DLDA and LDA under different training sample size per class on artificial dataset when discriminant information is along the diagonal direction.

However, when the diagonal elements which contain discriminant information are rearranged to the first column \( x_{ij} \leftrightarrow x_{1j} \) the performance of 2DLDA is improved apparently. In this case, when sample \( x_{ij} \) is reduced to \( y_{ij} \in R^{20} \) using Left/Right/ Bilateral 2DLDA the average classification error rates are 0 and are apparently superior to it using LDA. As shown in Table 3, we can find that the decent speed of eigenvalue is high which means the dimensional reduced efficiency using 2DLDA is high in this case.

From this experiment on artificial dataset, the main comparative conclusions between 2DLDA and LDA have been proven clearly. When discriminant information is not located along the row/column direction, the performance of 2DLDA is not superior to that of LDA because of its limited linear operation power and its smaller discriminant power than LDA. However, when discriminant information is located along the row/column direction the performance of 2DLDA is possible to be superior to that of LDA for the same reason. In this case, the illusive discriminant information from different row/column elements is excluded.

Table III.

<table>
<thead>
<tr>
<th></th>
<th>Left</th>
<th>Right</th>
<th>Bilateral</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0332</td>
<td>0.3112</td>
<td>0.4738</td>
</tr>
<tr>
<td>5</td>
<td>0.0018</td>
<td>0.0018</td>
<td>0.0008</td>
</tr>
<tr>
<td>10</td>
<td>0.0012</td>
<td>0.0011</td>
<td>0.0003</td>
</tr>
<tr>
<td>15</td>
<td>0.0008</td>
<td>0.0008</td>
<td>0.0000</td>
</tr>
<tr>
<td>20</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0000</td>
</tr>
<tr>
<td>25</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0000</td>
</tr>
<tr>
<td>30</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

C. Comparative experiment on ORL dataset

The third experiment is completed on ORL human face dataset. ORL dataset contains forty classes. There are 10 samples per class. All image samples have the resolution of 112 × 96 pixels. For the computational efficiency, here samples are resized to 56 × 48 pixels. We random select the training samples and the left samples treated as test samples.

After samples are rotated to the 45° direction which are illustrated in Fig 2, the experiment is done again. In this case, the samples are enlarged to 74 × 74 pixels in order to keep the original samples not changed and the blank part of every sample is filled with 255(white).

Figure 2. Illustrations of some rotated face images in ORL database.

We do this experiment 10 times. The average misclassification rates using LDA and 2DLDA are shown in Table 4. The values in parentheses denote the standard deviations of error rates. From Table 4, we can see that 2DLDA methods outperform 1DLDA method when the number of training samples is 2/4/6/8 per person. However, when the image samples are rotated the classification results degenerate and 2DLDA methods do not outperform 1DLDA anymore.
TABLE IV.
AVERAGE MISCLASSIFICATION RATES(%) USING 2D LDA ON ORIGINAL SAMPLES AND THE ROTATED SAMPLES

<table>
<thead>
<tr>
<th>No. of training images/person</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>25.36(3.23)</td>
<td>10.06(2.12)</td>
<td>3.67(1.53)</td>
<td>1.89(1.76)</td>
</tr>
<tr>
<td>Right</td>
<td>24.45(4.58)</td>
<td>8.93(3.67)</td>
<td>2.93(3.22)</td>
<td>1.82(2.04)</td>
</tr>
<tr>
<td>Bilateral</td>
<td>23.17(2.82)</td>
<td>6.77(2.56)</td>
<td>4.38(1.17)</td>
<td>0.75(0.87)</td>
</tr>
<tr>
<td>Left (45)</td>
<td>29.10(4.17)</td>
<td>14.49(3.28)</td>
<td>4.62(2.54)</td>
<td>2.80(3.47)</td>
</tr>
<tr>
<td>Right (45)</td>
<td>30.66(3.95)</td>
<td>12.45(1.84)</td>
<td>4.79(1.71)</td>
<td>3.25(1.66)</td>
</tr>
<tr>
<td>Bilateral (45)</td>
<td>30.43(2.89)</td>
<td>14.89(1.96)</td>
<td>5.96(2.23)</td>
<td>2.64(1.52)</td>
</tr>
<tr>
<td>1DLDA</td>
<td>27.54(2.40)</td>
<td>10.62(1.78)</td>
<td>3.74(1.37)</td>
<td>1.23(1.28)</td>
</tr>
</tbody>
</table>

V. Conclusions

In this paper, we discuss the differences between traditional vector-based LDA and matrix-based 2D LDA. It is found that the discriminant power of LDA is always larger than that of 2D LDA. Furthermore, we try to answer the question why 2D LDA outperforms LDA sometimes and we think the main reasons accounting for it are the difference of the stability under nonsingular linear transformation and the power of linear operation between LDA and 2D LDA. Experimental results show that when discriminant information is mainly located along the row/column direction the performance of 2D LDA is superior to that of LDA.

Acknowledgment

This work is supported by National Science Foundation of China under grant No 50875265,50474052.

References


Bo Yang was born in Yueyang, China on 22 October 1974. He is a teacher at Hunan Institute of Science and Technology, China. At the same time, he is working for his Ph.D. in the College of Mechanical and Electrical Engineering, Central South University, China. His current research relates to Sonar signal processing, pattern recognition.

Yingyong Bu is a professor in the College of Mechanical and Electrical Engineering, Central South University, China. His current research to pattern recognition, equipment information management.