A study on three Linear Discriminant Analysis based methods in Small Sample Size problem

Jun Liu, Songcan Chen *, and Xiaoyang Tan

Department of Computer Science and Engineering

Nanjing University of Aeronautics and Astronautics

Nanjing, 210016, P.R. China

Abstract

In this paper, we make a study on three Linear Discriminant Analysis (LDA) based methods: Regularized Discriminant Analysis (RDA), Discriminant Common Vectors (DCV) and Maximal Margin Criterion (MMC) in the Small Sample Size (SSS) problem. Our contributions are that: 1) we reveal that DCV obtains the same projection subspace as both RDA and wMMC (weighted MMC, a general form of MMC) when RDA's regularization parameter tends to zero and wMMC's weight parameter approaches to $+\infty$, which builds on close relationships among these three LDA based methods; 2) we offer efficient algorithms to perform RDA and wMMC in the Principal Component Analysis transformed space, which makes them feasible and efficient to applications such as face recognition; 3) we formulate the eigenvalue distribution of wMMC. On one hand, the formulated eigenvalue distribution can guide practitioners in choosing wMMC's projection vectors, and on the other hand, the underlying methodology can be employed in analyzing the eigenvalue distribution of matrices such as $AA^T - BB^T$, where the rows of A and B are far larger than their columns; and 4) we compare their classification performance on several benchmarks to get that, when the Mean Standard Variance (MSV) criterion is small, DCV can obtain competitive classification performance to both RDA and wMMC under optimal parameters, but when MSV is large, DCV generally yields lower classification accuracy than RDA and wMMC under optimal parameters.

Key words: Regularized Discriminant Analysis (RDA), Discriminant Common Vectors (DCV), Maximal Margin Criterion (MMC), Small Sample Size (SSS), Eigenvalue Distribution.

1 Introduction

In applications such as face recognition, the sample dimensionality d is typically larger than n, the number of training samples, which leads to the so-called Small Sample Size (SSS) problem and thus inevitably degrades the performance of the designed classifier. To solve this problem, one main category of methods is to perform Dimensionality Reduction (DR) by PCA (Principal Component Analysis, Eigenfaces) [31] and LDA (Linear Discriminant Analysis, Fisherfaces) [1]. The DR methods have been received wide interests in the pattern recognition domain, and a nice guidance on the DR methods can be found in [5]. As an unsupervised method, PCA looks for a subspace where the samples have the minimum reconstruction error. In contrast to PCA, LDA takes the class labels into consideration, and searches for a subspace where the samples from the same class are as compact as possible and meanwhile the samples from the different classes are as far as possible. The relationship between PCA and LDA has been studied by Martinez, with the main results

* Corresponding author, Tel: +86-25-8489-6481 Ext 12106, Fax: +86-25-8489-8069, E-mails: j.liu@nuaa.edu.cn (J. Liu), s.chen@nuaa.edu.cn (S. Chen), x.tan@nuaa.edu.cn(X. Tan).

being: "when the training data set is small, PCA can outperform LDA and, also, that PCA is less sensitive to different training data sets" [20].

Naturally, it is very important to come up with new and effective dimensionality reduction methods, but we believe that a study of the existing methods is also quite important, since it can correct some misunderstandings, guide practitioners in choosing appropriate methods, build on relationships among existing methods, and help invent better methods. In this paper, we make a study on three methods: Regularized Discriminant Analysis (RDA) [6], Discriminant Common Vectors (DCV) [3] and Maximal Margin Criterion (MMC) [12,13] in the SSS problem, whose underlying motivations and contributions are:

- These methods are all LDA based methods that can effectively deal with the SSS problem, but the techniques employed are distinct: RDA is from the viewpoint of regularization, DCV is originated from obtaining the projection subspace in the null space of the within-class scatter matrix, and MMC aims at maximizing the average margin between classes. Therefore, a comparison among them will shed light on these DR methods. Our main argument is that DCV obtains the same projection subspace as both RDA and wMMC (weighted MMC, a general form of MMC) when RDA's regularization parameter tends to zero and wMMC's weight parameter approaches to $+\infty$.
- RDA is a classical and pioneer work that solves the SSS problem in LDA and is widely cited in literature, e.g., [3,10,13], however, it is often criticized for demanding space and time costs in applications such as face recognition and is consequently seldom employed in face recognition. In this paper, we show that RDA can be performed in the PCA transformed space and

propose an efficient RDA algorithm in space and time complexities of $O(dn)$ and $O(dn^2)$, which are far more efficient than the original one's $O(d^2)$ and $O(d^3)$. Furthermore, $O(dn)$ and $O(dn^2)$ are also the corresponding space and time complexities of DCV, which has been proven to be a very efficient DR method [3,15], and thus the newly proposed algorithm makes RDA both feasible and efficient for face recognition.

- MMC is a good DR method that maximizes the average margin between classes and is reported to be very effective for face recognition. However, in [13], Li et al. suggested to utilize the eigenvectors corresponding to zero eigenvalues for DR, which is a misunderstanding due to a lack of formulating the inherent eigenvalue distribution of MMC. We reveal and prove that, when the training samples are independent (it is usually the case with applications such as face recognition), wMMC exactly has $c-1$ (c is the number of classes) positive, $n-c$ negative, and $d-n+1$ zero eigenvalues. The revealed eigenvalue distribution helps verify that the eigenvectors corresponding to zero eigenvalues contain no discriminant information and thus helps correct the aforementioned misunderstanding. Moreover, the underlying methodology is also useful mathematically, e.g., in analyzing the eigenvalue distribution of $AA^T - BB^T$, where A and B have the characteristic that the number of rows are typically larger than that of columns.
- Similar to the original RDA, the space and time complexities of the original MMC [1] are respectively $O(d^2)$ and $O(d^3)$ [12], which are very demanding for applications such as face recognition. In this paper, we show that wMMC can

[1] In [13], Li et al. claimed that they found an efficient algorithm for MMC, however, as pointed in [17], the efficient algorithm proposed in [13] is problematic (refer to Section 3.2 for further discussion).

be performed in the PCA transformed space and offer an efficient wMMC algorithm in space and time complexities of $O(dn)$ and $O(dn^2)$, which are far more efficient than the original MMC's $O(d^2)$ and $O(d^3)$.

- We compare the classification performance of these three methods on several benchmarks to get that, when the Mean Standard Variance (MSV) [15] criterion is small, DCV can obtain competitive classification performance to both RDA and wMMC under optimal parameters, but when MSV is large, DCV generally yields inferior classification performance to RDA and wMMC under optimal parameters.

In what follows, we briefly review these three methods in Section 2, develop efficient algorithms for RDA and wMMC in Section 3, derive wMMC's eigenvalue distribution in Section 4, reveal that DCV obtains the same projection subspace as both RDA and wMMC under certain circumstances in Section 5, report experimental results in Section 6, and draw a conclusion to this paper in Section 7.

2 A Brief Review of These Three Methods

We assume that, the training set is composed of c classes $C_1, \ldots, C_c$, the i-th class has n_i training samples, x_j^i denotes the j-th d-dimensional sample from the i-th class, and $n = \sum_{i=1}^{c} n_i$ is the total number of training samples which are generally independent in applications such as face recognition. The within-class scatter matrix S_w and between-class scatter matrix S_b can be respectively denoted as

$$S_w = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_j^i - m_i)(x_j^i - m_i)^T = H_w H_w^T, \tag{1}$$

$$S_b = \frac{1}{n}\sum_{i=1}^{c} n_i(m_i - m)(m_i - m)^T = H_b H_b^T, \tag{2}$$

where H_w and H_b are

$$H_w = \frac{1}{\sqrt{n}}\left[x_1^1 - m_1, \ldots, x_{n_1}^1 - m_1, \ldots, x_{n_c}^c - m_c\right], \tag{3}$$

$$H_b = \frac{1}{\sqrt{n}}\left[\sqrt{n_1}(m_1 - m), \ldots, \sqrt{n_c}(m_c - m)\right], \tag{4}$$

m_i is the centroid of the i-th class, and m is the centroid of the training set. LDA looks for a projection matrix, W, that maximizes the Fisher's criterion

$$J_{Fisher}(W) = arg\max_{W} |W^T S_b W| / |W^T S_w W|. \tag{5}$$

In applications such as face recognition, both S_w and S_b will be singular, due to $d \gg n$ or the so-called SSS problem, and as a result it is impossible to directly calculate W from (5). Regularized Discriminant Analysis (RDA) [6], Discriminant Common Vectors (DCV) [3] and Maximal Margin Criterion (MMC) [12,13] are three methods that can deal with this problem and will be respectively reviewed in Sections 2.1, 2.2 and 2.3.

2.1 Regularized Discriminant Analysis

To solve the singularity problem of S_w, RDA adds a multiple of identity matrix to S_w, as $S_w + \alpha I_d$, for some $\alpha > 0$, where I_d is a $d \times d$ identity matrix. $S_w + \alpha I_d$ is nonsingular now and RDA's projection vector w can be computed by

$$S_b w = \lambda(S_w + \alpha I_d)w, \tag{6}$$

or equivalently

$$(S_w + \alpha I_d)^{-1} S_b w = \lambda w. \tag{7}$$

Generally speaking, RDA employs those projection vectors (w's) corresponding to positive eigenvalues [25]. Further, when α tends to $+\infty$, $S_w + \alpha I_d$ can be regarded as αI_d and w becomes the eigenvector of S_b; and when α tends to zero, w lies in the null space of S_w. As a result, the regularization parameter α tunes the projection vector w between the range space of S_b and the null space of S_w, which contains important discriminant information.

2.2 Discriminant Common Vectors

The idea of DCV is first mentioned by Belhumeur et al. in [1], where they suggested to maximize the between-class scatter subject to the constraint that the within-class scatter is zero, i.e.,

$$W_{opt} = arg \max_{\substack{W^T S_w W = 0 \\ W^T W = I_{c-1}}} |W^T S_b W|, \tag{8}$$

where I_{c-1} is a $(c-1) \times (c-1)$ identity matrix. It is obvious that the obtained projection matrix W_{opt} will make the Fisher's criterion (5) infinite. Following this idea, Chen et al. [4] employed it for face recognition, but the proposed method is not efficient; Huang et al. [10] proposed a PCA plus Null Space (PNS) algorithm to realize this idea; Cevikalp et al. [3] proposed the DCV method and revealed that the samples from the same class are projected to a common vector; and in [15], we pointed out that the existing null space based methods such as PNS and DCV in fact obtain the same projection subspace, and showed that DCV can be implemented by a thin QR decomposition [2] [7],

[2] Suppose $A \in \mathbb{R}^{m \times n}$ has full column rank ($m > n$), then its thin QR decomposition is: $A = QR$, where the column vectors in $Q \in \mathbb{R}^{m \times n}$ span the range space of A and $R \in \mathbb{R}^{n \times n}$ is an upper triangle matrix.

which makes DCV easy be understood and extended to the nonlinear form using the kernel trick [27,28].

2.3 Weighted Maximal Margin Criterion

Weighted Maximal Margin Criterion (wMMC) [3] aims at maximizing the average margin between classes, where the weighted interclass margin between the i-th class and the j-th class can be denoted as

$$d(C_i, C_j) = d(m_i, m_j) - \beta(S(C_i) + S(C_j)), \tag{9}$$

$d(m_i, m_j)$ is defined as the square *Euclidean* distance between m_i and m_j, $S(C_i)$ and $S(C_j)$ are respectively defined as the traces of the scatter matrices of the i-th and j-th classes, and β is a positive weight parameter. After some deduction, the average margin between classes under projection matrix W is

$$J(W) = tr(W^T(S_b - \beta S_w)W). \tag{10}$$

When $\beta = 1$, (10) is the objective function of MMC [12,13], and thus wMMC just becomes MMC. Confining the column vectors in W to be unit vectors, W that maximizes (10) can be calculated through the following eigenvalue equation

$$(S_b - \beta S_w)w = \lambda w. \tag{11}$$

Despite of the simplicities of the eigenvalue equations of RDA's (7) and wMMC's (11), RDA and wMMC both manipulate on matrices of size $d \times d$, and thus

[3] The positive weight parameter β is added here for generality, and the weight parameter was firstly employed in [34] to propose a weighted kernelized MMC. Moreover, when $\beta = 1$, wMMC just reduces to MMC.

they both have space and time complexities of $O(d^2)$ and $O(d^3)$. In applications such as face recognition, d is typically large, e.g., when the image resolution is 100×100, d equals 10000 and it will cost bytes in the order of 10^8 for storage and floating operations (flops) [7] in the order of 10^{12} for computation, which are very expensive. As a result, 1) although RDA is widely cited in the recent LDA based papers on face recognition, e.g., [3,10,13], it is seldom implemented for face recognition; 2) in [12], when carrying out face recognition experiments by MMC, Li et al. resized the original image resolution of 112×92 to 28×23 for computation efficiency; and 3) in [13], Li et al. aimed at developing an efficient algorithm for implementing MMC, but as revealed in [17], their proposed efficient algorithm is problematic (refer to Section 3.2 for discussion). Therefore, efficient algorithms for RDA and wMMC are quite necessary and will be discussed in the next section.

3 Efficient Algorithms for Regularized Discriminant Analysis and Weighted Maximal Margin Criterion

Before proposing the efficient RDA and wMMC algorithms, we first discuss two related studies on efficient algorithms. The first one is the efficient Eigenfaces algorithm, which aims at solving the following eigenvalue equation

$$S_t w = \lambda w, \tag{12}$$

where $S_t = S_b + S_w$ is the total scatter matrix. In applications such as face recognition, it is intractable to directly solve (12), since S_t is a typically large $d \times d$ matrix. As an alternative, a commonly employed practice is to:

1) Formulate (not to calculate) S_t as:

$$S_t = H_t H_t^T, \tag{13}$$

where H_t is computed as

$$H_t = \frac{1}{\sqrt{n}} \left[x_1^1 - m, x_2^1 - m, \ldots, x_{n_1}^1 - m, \ldots, x_{n_c}^c - m \right], \tag{14}$$

2) Take advantage of the existence of the following singular value decomposition

$$H_t = U \Lambda V^T, \tag{15}$$

where U and V are respectively $d \times r$ and $n \times r$ matrices with orthonormal columns, Λ is an $r \times r$ diagonal matrix containing the positive singular values of H_t, r ($\leq n-1$) is the rank of H_t (or equivalently S_t), and V and Λ can be obtained by solving the following singular value decomposition

$$H_t^T H_t = V \Lambda^2 V^T, \tag{16}$$

and 3) Obtain U according to (15) as

$$U = H_t V \Lambda^{-1}, \tag{17}$$

which contains the eigenvectors of S_t corresponding to the positive eigenvalues due to

$$S_t = H_t H_t^T = U \Lambda^2 U^T. \tag{18}$$

Since we do not explicitly compute the $d \times d$ matrix S_t, and the largest matrix we manipulate on is the $d \times n$ matrix H_t, the space and time complexities of the aforementioned efficient Eigenfaces is $O(dn)$ and $O(dn^2)$, which are far less than $O(d^2)$ and $O(d^3)$ of directly solving (12).

The second related study is the Efficient Quadratic Regularization (EQR) method [8] for gene expression arrays. EQR aims at obtaining

$$\hat{\beta} = (X^T X + \lambda I_p)^{-1} X^T y, \tag{19}$$

where $X \in \mathbb{R}^{n \times p}$ is a gene expression array consisting of n samples and p genes, $p \gg n$ [4], and y contains the descriptions for the n genes. (19) is very expensive to compute, since p typically varies between 1,000 and 20,000 [8]. Hastie and Tibshirani converted the solution of (19) to

$$\hat{\beta} = V(XX^T + \lambda I_n)^{-1} R^T y, \tag{20}$$

where $X = UDV^T = RV^T$ is the singular value decomposition of X. Since $(XX^T + \lambda I_n)$ is of size $n \times n$, far smaller than $p \times p$, the size of $(X^T X + \lambda I_p)$, and thus EQR can be much efficiently computed by converting (19) to (20).

A common characteristic of the techniques employed in deriving efficient Eigenfaces and EQR algorithms lies in that, a primal problem is converted to a corresponding dual problem. This technique of switching from a primal to a dual formulation is nicely depicted in the representer theorem [27] and is widely employed in designing learning algorithms such as support vector machine classifiers [28]. The efficient RDA and wMMC algorithms to be proposed in Sections 3.1 and 3.2 will also employ such a technique. Moreover, due to such conversion, the proposed efficient RDA and wMMC methods employ two stages, namely, PCA and a further analysis in the PCA transformed space. And Campbell and Atchley gave a geometrical analysis of such two stages in [2].

[4] Unlike face recognition, the number of samples is far larger than the dimensionality of samples in gene expression arrays.

For convenience of discussion, we let $\tilde{U}$[5] denote a $d \times (d-r)$ matrix whose column vectors are the orthonormal eigenvectors of S_t corresponding to the zero eigenvalues. By U defined in (15), we let $S_w^{'} = U^T S_w U$, $S_b^{'} = U^T S_b U$, $S_t^{'} = U^T S_t U$, and Q and $\tilde{Q}$ respectively be the orthonormal eigenvectors of $S_w^{'}$ corresponding to zero and positive eigenvalues. Furthermore, since the column vectors in $\begin{bmatrix} U & \tilde{U} \end{bmatrix}$ constitute a set of orthonormal bases for the space R^d, then any w in R^d can be written as [7]:

$$w = Up + \tilde{U}\tilde{p}, \tag{21}$$

where p and $\tilde{p}$ are respectively r and $d-r$ dimensional column vectors.

3.1 RDA in the PCA Transformed Space

Proposition 1 *Regularized Discriminant Analysis in Small Sample Size problem can be performed in the Principal Component Analysis transformed space.*

Proof: Since column vectors in $\tilde{U}$ are the eigenvectors of S_t corresponding to zero eigenvalues, we have $S_t\tilde{U} = 0$, $S_w\tilde{U} = 0$ and $S_b\tilde{U} = 0$. Substituting (21) into (6), we get:

$$S_b U p = \lambda S_w U p + \lambda\alpha(Up + \tilde{U}\tilde{p}). \tag{22}$$

Pre-multiplying U^T to both sides of (22), we get

$$S_b^{'} p = \lambda(S_w^{'} + \alpha I_{n-1})p. \tag{23}$$

Pre-multiplying $\tilde{U}^T$ to both sides of (22), we get

$$0 = \lambda\tilde{p}. \tag{24}$$

[5] Just like S_t, there is no need to compute $\tilde{U}$, and only its formulation is utilized.

Since in RDA, we are interested in those eigenvectors corresponding to positive eigenvalues, then from (24), we have

$$\tilde{p} = 0. \tag{25}$$

Further, substituting (25) into (21), we have

$$w = Up. \tag{26}$$

From (26), we can clearly observe that RDA's projection vector w is composed of two parts: 1) U, which is the projection matrix of PCA; and 2) p, which is the solution to (23), or the projection vector of RDA in the PCA transformed space. As a result, this proposition is proved □.

From Proposition 1, it is easy to observe that the w's for RDA can be calculated in terms of the following three steps: 1) calculate U; 2) compute p's by (23) and 3) obtain w's from (26). The time complexity can be analyzed as: 1) U can be obtained in $O(dn^2)$; 2) $S_b^{'} = U^T S_b U = (U^T H_b)(U^T H_b)^T$ and $S_w^{'} = U^T S_w U = (U^T H_w)(U^T H_w)^T$ can be respectively computed in $O(dnc)$ and $O(dn^2)$; 3) p's can be obtained from (23) in $O(n^3)$ and 4) w's can be got from (26) in $O(dnc)$, thus the time complexity for the newly proposed RDA algorithm is $O(dn^2)$. Besides, it is easy to get that its space complexity is $O(dn)$, since all the matrices to be computed are not larger than $d \times n$. Furthermore, $O(dn)$ and $O(dn^2)$ are just the space and time complexities of DCV which has been proven to be an efficient DR algorithm [3,15], thus the newly proposed RDA algorithm is not only feasible but also efficient to applications such as face recognition.

3.2 wMMC in the PCA Transformed Space

To overcome the high space and time complexities of directly solving eigenvalue equation (11), Li et al. tried to develop an efficient MMC algorithm in [13]. To this end, they looked for a $d \times r$ transformation matrix P that simultaneously diagonalizes S_b and S_t as

$$P^T S_b P = \tilde{\Lambda}, \tag{27}$$

$$P^T S_t P = I_r. \tag{28}$$

Generally speaking, the transformation matrix P that satisfies both (27) and (28) is unique, and can be computed as $P = U\Lambda^{-1}\Psi$, where U and Λ^2 contain the eigenvectors and eigenvalues of S_t, and Ψ contains the eigenvectors of $\Lambda^{-1}S_b^{'}\Lambda^{-1}$. Li et al. then argued that P and $2\tilde{\Lambda} - I_r$ are respectively the eigenvectors and eigenvalues of $S_b - S_w$. However, as pointed out in [17], their argument is problematic, since, although from (27) and (28) we can get

$$P^T(S_b - S_w)P = 2\tilde{\Lambda} - I_r, \tag{29}$$

we can not assure that P and $2\tilde{\Lambda} - I_r$ are the eigenvectors and eigenvalues of $S_b - S_w$. As a result, MMC lacks an efficient algorithm.

Proposition 2 *Weighted Maximal Margin Criterion in Small Sample Size problem can be performed in the Principal Component Analysis transformed space.*

Proof: Keeping in mind $S_w\tilde{U} = 0$ and $S_b\tilde{U} = 0$, and substituting (21) into (11), we get

$$(S_b - \beta S_w)Up = \lambda(Up + \tilde{U}\tilde{p}). \tag{30}$$

Pre-multiplying U^T to both sides of (30), we get

$$(S_b^{'} - \beta S_w^{'})p = \lambda p. \tag{31}$$

Pre-multiplying $\tilde{U}^T$ to both sides of (30), we get

$$0 = \lambda \tilde{p}. \tag{32}$$

Generally speaking, since the objective of wMMC is to maximize (10), we are only interested in eigenvectors corresponding to positive eigenvalues[6]. Therefore, from (32) we have

$$\tilde{p} = 0. \tag{33}$$

Further, substituting (33) into (21), we have

$$w = Up. \tag{34}$$

From (34), we can clearly observe that wMMC's projection vector w is composed of two parts: 1) U, which is the projection matrix of PCA; and 2) p, which is the solution to (31), or the projection vector of wMMC in the PCA transformed space. As a result, wMMC in SSS problem can be performed in the PCA transformed space, which ends the proof of this proposition $\Box$.

From Proposition 2, it is easy to observe that the w's for wMMC can be calculated in terms of the following three steps: 1) calculate U; 2) calculate p's by (31); and 3) obtain w's from (34). Following similar analysis as in

[6] Li et al. [13] argued that the eigenvectors corresponding to zero eigenvalues can still be utilized, however, as will be revealed in Section 4, in applications such as face recognition, the training samples are usually independent and the eigenvectors corresponding to zero eigenvalues contain no discriminant information and should not be utilized.

Section 3.1, it is easy to get that the space and time complexities of the newly proposed efficient wMMC algorithm are respectively $O(dn)$ and $O(dn^2)$, much more efficient than $O(d^2)$ and $O(d^3)$ of the original MMC algorithm, when d is typically larger than n in applications such as face recognition.

4 Eigenvalue Distribution of wMMC

In this section, we reveal the eigenvalue distribution of wMMC to facilitate choosing its projection vectors. Since $S_b - \beta S_w$ is a very large $d \times d$ matrix, it is expensive and intractable to explicitly compute all of its eigenvalues. $S_b - \beta S_w$ can be generalized to such matrix form as $AA^T - BB^T$, where both A and B are rank deficient and have much more rows than columns. Our methodology for revealing the eigenvalue distribution of $AA^T - BB^T$ is that: 1) we make full advantage of the fact that both A and B are rank deficient, and obtain a $d \times r$ ($r = rank([A \quad B])$) matrix U, which spans the range space of matrix $[A \quad B]$ and can be efficiently computed by tools such as Gram-Schmidt Decomposition and Singular Value Decomposition [7]; 2) we denote (note, not calculate) $\tilde{U}$ as the orthogonal complement of U; and 3) we let $P = \begin{bmatrix} U & \tilde{U} \end{bmatrix}$ and employ *Sylvester's Law of Intertia* [7] to get that $AA^T - BB^T$ has at least $d - r$ zero eigenvalues and at most r non-zero eigenvalues. Since r is far less than d, it is not expensive to further obtain the signs of the at most r non-zero eigenvalues in the U transformed space.

In the discussion in this section, we assume that the n training samples are independent (i.e., $r = n - 1$), which is usually the case with applications such as face recognition. Moreover, considering the facts that $rank(S_w^{'}) \leq n - c$, $rank(S_b^{'}) \leq c - 1$, and $rank(S_w^{'}) + rank(S_b^{'}) \geq rank(S_t^{'}) = n - 1$, we have

$rank(S_w^{'}) = n-c$, $rank(S_b^{'}) = c-1$, Q and $\tilde{Q}$ are respectively $(n-1)\times(c-1)$ and $(n-1)\times(n-c)$ matrices. In the following, we first give two lemmas and then formulate the eigenvalue distribution of wMMC.

Lemma 1 ***(Sylvester's Law of Intertia)*** *[7] Let A be any symmetric matrix, P be any non-degenerate matrix and $B = P^T AP$. Then, $\pi(A) = \pi(B)$, $\nu(A) = \nu(B)$, $\delta(A) = \delta(B)$, where $\pi(.)$, $\nu(.)$ and $\delta(.)$ respectively denote the number of positive, negative and zero eigenvalues of given matrix.*

Lemma 2 *Let B be any $k \times k$ symmetric matrix, C be any $k\times(d-k)$ matrix, D be any $(d-k)\times(d-k)$ symmetric matrix and*

$$A = \begin{bmatrix} B & C \\ C^T & D \end{bmatrix}. \tag{35}$$

If B is positive definite, then $\pi(A) \geq k$; and if B is negative definite, then $\nu(A) \geq k$.

Proof: Let O be a $(d-k)\times k$ zero matrix and P be a matrix defined as

$$P = \begin{bmatrix} I_k & -B^{-1}C \\ O & I_{d-k} \end{bmatrix}, \tag{36}$$

we have

$$P^T AP = \begin{bmatrix} B & O^T \\ O & D - C^T B^{-1} C \end{bmatrix}. \tag{37}$$

When B is positive definite, it is clear that $\pi(P^T AP) \geq k$; and similarly, when B is negative definite, $\nu(P^T AP) \geq k$. Furthermore, since P defined in (36) is a non-degenerate matrix, then employing Lemma 1 this lemma is proved □.

Proposition 3 *When the n training samples are independent and β is positive, the matrix $S_b-\beta S_w$ exactly has $c-1$ positive, $n-c$ negative and $d-n+1$ zero eigenvalues.*

Proof: We will prove this proposition in the following three steps:

1) Let $P=\begin{bmatrix}\tilde{U} & U\end{bmatrix}$ which is a non-degenerate matrix, we have

$$P^T(S_b-\beta S_w)P=\begin{bmatrix} O_1 & O_2^T \\ O_2 & U^T(S_b-\beta S_w)U \end{bmatrix}, \tag{38}$$

where O_1 and O_2 are respectively $(d-n+1)\times(d-n+1)$ and $(n-1)\times(d-n+1)$ zero matrices. From Lemma 1, one can easily get $\delta(S_b-\beta S_w)\geq d-n+1$.

2) Since $S_t^{'}$ is positive definite and $[Q \quad \tilde{Q}]$ is an orthonormal matrix, then

$$\begin{bmatrix}Q & \tilde{Q}\end{bmatrix}^T(S_w^{'}+S_b{'})\begin{bmatrix}Q & \tilde{Q}\end{bmatrix}=\begin{bmatrix} Q^TS_b{'}Q & Q^TS_b{'}\tilde{Q} \\ \tilde{Q}^TS_b{'}Q & \tilde{Q}^T(S_w{'}+S_b{'})\tilde{Q}) \end{bmatrix} \tag{39}$$

is positive definite, and so is $Q^TS_b{'}Q$. Let $P=[UQ \quad U\tilde{Q} \quad \tilde{U}]$, we have

$$P^T(S_b-\beta S_w)P=\begin{bmatrix} Q^TS_b{'}Q & C \\ C^T & D \end{bmatrix}, \tag{40}$$

where $C=(UQ)^T(S_b-\beta S_w)\begin{bmatrix}U\tilde{Q} & \tilde{U}\end{bmatrix}$ and $D=\begin{bmatrix}U\tilde{Q} & \tilde{U}\end{bmatrix}^T(S_b-\beta S_w)\begin{bmatrix}U\tilde{Q} & \tilde{U}\end{bmatrix}$. Since P is a non-degenerate matrix and $Q^TS_b{'}Q$ is positive definite, then utilizing Lemmas 1 and 2, we can easily get $\pi(S_b$-$\beta S_w)$ $\geq c$-1.

3) Let R and $\tilde{R}$ respectively be the orthonormal eigenvectors of $S_b^{'}$ corresponding to zero and non-zero eigenvalues, where R and $\tilde{R}$ are respectively

$(n-1) \times (n-c)$ and $(n-1) \times (c-1)$ matrices. Following similar deduction as 2), we can get $\nu(S_b - \beta S_w) \geq n - c$.

Finally, summarizing the results from the above three steps and taking notice of the fact that $\pi(S_b - \beta S_w) + \nu(S_b - \beta S_w) + \delta(S_b - \beta S_w) = d$, this proposition is proved $\Box$.

The revealed eigenvalue distribution of weighted MMC in Proposition 3 can guide us to select wMMC's projection vectors in the case that the training samples are independent, and an analysis is given as follows: on one hand, the $d - n + 1$ column vectors in $\tilde{U}$ are the eigenvectors of $S_b - \beta S_w$ corresponding to zero eigenvalues and the training samples become a common vector when projected by $\tilde{U}$, and thus such $\tilde{U}$ contains no discriminant information; on the other hand, $d - n + 1$ is just the number of $S_b - \beta S_w$'s zero eigenvalues, and thus we can say that, in wMMC, the eigenvectors corresponding to zero eigenvalues contain no discriminant information for classification and we usually only utilize the $c - 1$ eigenvectors corresponding to the positive eigenvalues for DR.

In the end of this section, it should be noted that, when some training samples are dependent (this seldom happens in applications such as face recognition), r should be less than $n - 1$, and the number of zero eigenvalues of $S_b - \beta S_w$ should be over $d - n + 1$. And in this case, although the eigenvectors of $S_b - \beta S_w$ corresponding to zero eigenvalues do not benefit the maximization of wMMC's objection function (10), some of them (excluding the $d - r$ eigenvectors of S_t corresponding to zero eigenvalues) might contain some discriminant information for classification.

5 Relationships among RDA, wMMC and DCV

In this section, we reveal the relationship between DCV and wMMC in Section 5.1 and the relationship between DCV and RDA in Section 5.2, and give two criteria in Section 5.3 for convenience of empirical evaluation of the relationships among these three methods.

5.1 DCV versus wMMC

Proposition 4 *Discriminant Common Vectors obtains the same projection subspace as weighted Maximal Margin Criterion when the latter's weight parameter β approaches to $+\infty$.*

Proof: Since the column vectors in Q are the orthonormal eigenvectors of $S_w^{'}$ corresponding to zero eigenvalues, then $S_w^{'}Q = 0$. Further, since $\begin{bmatrix} Q & \tilde{Q} \end{bmatrix}$ is an orthonormal matrix, thus the p in (31) can be denoted as:

$$p = Qq + \tilde{Q}\tilde{q}. \tag{41}$$

Substituting (41) into (31), we get

$$(S_b^{'} - \beta S_w^{'})(Qq + \tilde{Q}\tilde{q}) = \lambda(Qq + \tilde{Q}\tilde{q}). \tag{42}$$

Pre-multiplying $\tilde{q}^T\tilde{Q}^T$ to both sides of (42), we get

$$\tilde{q}^T\tilde{Q}^T S_b^{'}(Qq + \tilde{Q}\tilde{q}) - \beta\tilde{q}^T\tilde{Q}^T S_w^{'}\tilde{Q}\tilde{q} = \lambda\tilde{q}^T\tilde{q}. \tag{43}$$

Pre-multiplying q^TQ^T to both sides of (42), we get

$$q^TQ^TS_b^{'}(Qq + \tilde{Q}\tilde{q}) = \lambda q^Tq. \tag{44}$$

Adding (43) and (44) and employing (41), we can get

$$p^T S_b^{'} p - \beta \tilde{q}^T \tilde{Q}^T S_w^{'} \tilde{Q} \tilde{q} = \lambda(\tilde{q}^T \tilde{q} + q^T q). \tag{45}$$

If w is a unit vector, then $p^T p = 1$ and $\tilde{q}^T \tilde{q} + q^T q = 1$. Moreover, $p^T S_b^{'} p \leq tr(S_b^{'})$, where $tr(S_b^{'})$ is a positive and finite number for given training samples, and $\tilde{Q}^T S_w^{'} \tilde{Q}$ is a positive definite matrix since $\tilde{Q}$ is composed of the eigenvectors of $S_w^{'}$ corresponding to positive eigenvalues. As a result, when β tends to $+\infty$, λ tends to -∞ so long as $\tilde{q}$ is not a zero vector. Consequently, in order to ensure that $\lambda > 0$, $\tilde{q}$ must be a zero vector. Employing (41), we have $p = Qq$, where q can be calculated through pre-multiplying Q^T to (42) as

$$Q^T S_b^{'} Q q = \lambda q. \tag{46}$$

Put the accordingly obtained $c-1$ q's as column vectors in a matrix V, then V is an orthormal matrix due to the fact that $Q^T S_b^{'} Q$ is positive definite (see the proof of Proposition 3). Thus, when β tends to $+\infty$, the projection matrix for wMMC is UQV, which is just the projection matrix for PNS [10] and has been proven to span the same subspace as DCV in [15]. As a result, DCV obtains the same projection subspace as wMMC when the latter's weight parameter β tends to $+\infty$, and this ends the proof of this proposition □.

5.2 DCV versus RDA

Proposition 5 *Discriminant Common Vectors obtains the same projection subspace as Regularized Discriminant Analysis when the latter's regularization parameter α tends to be zero.*

Proof: Adding $\lambda S_b^{'}$ to both sides of (23) and letting $\tilde{\lambda} = \lambda/(\lambda+1)$, we will get

$$S_b^{'} p = \tilde{\lambda}(S_t^{'} + \alpha I_{n-1})p. \tag{47}$$

It is obvious that, for given pair of λ and $\tilde{\lambda}$, the eigenvectors obtained respectively by (23) and (47) are the same. When α tends to zero, the $c-1$ non-zero eigenvalues for (47) are all ones, and Q makes up of the corresponding $c-1$ eigenvectors, since $S_b^{'} Q = S_t^{'} Q$. The projection matrix obtained by RDA in this case should be UQ, which spans the same subspace as UQV (note that as proven in Proposition 4, V is orthonormal) or DCV. As a result, DCV obtains the same projection subspace as RDA when the latter's regularization parameter α tends to be zero, and this ends the proof □.

Although DCV, RDA, and wMMC are originated distinctly, they in fact have close relationships as revealed in Propositions 4 and 5. Furthermore, the revealed relationships can guide us in choosing dimensionality reduction methods in practical application, and an analysis is given as follows:

DCV achieves the maximum (infinite) of the Fisher's criterion (5), however, this does not mean that it can always obtain the optimal classification accuracy or generalization ability in all applications. In other words, extremely high (here infinite) Fisher's criterion value does not definitely yield better generalization ability. In [15], it was argued by us that when the Mean Standard Variance (MSV) criterion (refer to Section 5.3.1 for the definition) that measures the compactness of the training samples from the same classes is relatively small, DCV can achieve better performance than other methods, but on the contrary, when MSV is relatively large, DCV can not assure superior performance. In practical applications, RDA and wMMC may be better choices than DCV, since: 1) as revealed by Propositions 4 and 5, DCV in fact obtains the same projection subspace as RDA and wMMC under certain circumstances, namely, DCV is a special case of both RDA and wMMC; and 2) by adapting the parameters with the specific data, better generalization performance can be obtained by RDA and wMMC. Moreover, we think it is meaningful and important to come up with the data dependent rules for automatically setting the parameters in both RDA and wMMC.

5.3 Two Criteria

5.3.1 Mean Standard Variance

In [15], we gave a Mean Standard Variance (MSV) criterion for explaining the experimental phenomena that DCV works better than other methods on some databases, but not on some others. The MSV criterion measures the compactness of the training samples from the same classes, and is defined as [15]:

$$MSV = \frac{1}{c}\sum_{i=1}^{c} SV_i, \tag{48}$$

where SV_i is the standard variance of the i-th class defined as

$$SV_i = \frac{1}{d}\sum_{k=1}^{d}\sqrt{\frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{jk}^i - m_{ik})^2}, \tag{49}$$

with x_{jk}^i and m_{ik} respectively denoting the k-th element of the d-dimensional samples x_j^i and class mean m_i. From (48) and (49), we can know that MSV can measure the variations among the samples from the same class, and when MSV is small, the training samples from the same class have relatively small variations and vice versa. Moreover, since DCV concentrates the training samples from the same class to a common vector, then it is reasonable that, when MSV is relatively small, DCV can perform well, but when MSV is relatively large, it performs poorly since the projection subspace overfits the training samples [15]. Here, we employ the MSV criterion to get some insights into the relationships among RDA, wMMC and DCV in terms of classification performance.

5.3.2 Subspace Distance

To measure the distance between subspaces S_1 and S_2, where the columns vectors in $P_1 \in R^{d \times k}$ and $P_2 \in R^{d \times k}$ are respectively their orthonormal base vectors, we employ the subspace distance defined in [7] as

$$dist(S_1, S_2) = ||P_1 P_1^T - P_2 P_2^T||_2 = \sqrt{1 - cos^2 \theta}, \tag{50}$$

where $||.||_2$ is the matrix 2-norm, θ is the maximal principal angle between S_1 and S_2, and $cos(\theta)$ equals the minimal singular value of $P_1^T P_2$. Generally speaking, $0 \leq dist(S_1, S_2) \leq 1$, and S_1 is identical to S_2 if and only if $dist(S_1, S_2) = 0$.

6 Experiments

In this section, we carry out experiments to support the arguments made in this paper and to compare the classification performance of these methods. In the following, we describe databases and experimental settings in Section 6.1, and then report experimental results in Section 6.2.

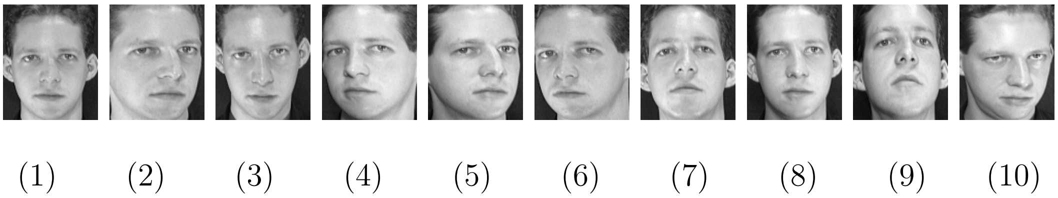


Fig. 1. Illustration of 10 images of one subject from ORL face database.

Table 1

Data partition on ORL face database

Category	Training	Testing
ORL1	1-2	3-10
ORL2	1-3	4-10
ORL3	1-4	5-10
ORL4	1-5	6-10

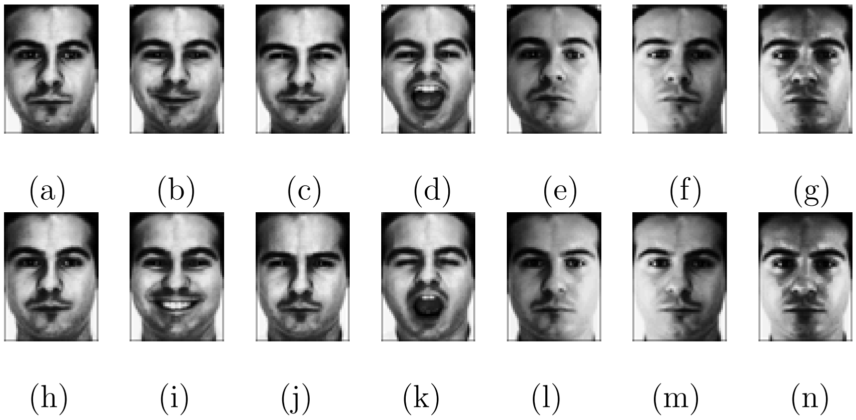
6.1 Database description and experimental setting

6.1.1 Database description

We conduct experiments on the following three benchmarks: ORL [7] face database, AR [21] face database, and the COIL-20 [22] object database.

The ORL face database contains ten different images of each of 40 distinct subjects. All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). The size of each image is 112×92 pixels, with 256 grey levels per pixel, and Fig. 1 illustrates the ten images of one subject, which are numbered between 1 and 10. In our experiments here, the face images are resized to a

[7] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html



(a) (b) (c) (d) (e) (f) (g)

(h) (i) (j) (k) (l) (m) (n)

Fig. 2. Illustration of 14 images of one subject from AR face database.

Table 2

Data partition on AR face database

Category	Training	Testing
AR1	a, h	b-g, i-n
AR2	a, b, h, i	c-g, j-n
AR3	a, e, h, l	b-d, f-g, i-k, m-n
AR4	a-d, h-k	e-g, l-n
AR5	a, e-g, h, l-n	b-d, i-k

resolution of 56 × 46, and the grey level values of all images are rescaled to [0 1]. We carry out four independent experiments ORL1, ORL2, ORL3 and ORL4, where the training samples and testing samples are summarized in Table 1. It is obvious that, on ORLi ($i = 1, 2, 3, 4$), the first $i + 1$ face images of each subject are employed for training while the rest for testing, and thus the experiments on ORL1-ORL4 can test the performance of given method under different number of training samples per subject.

AR consists of over 3200 frontal images of 126 subjects. Each subject has 26 different images which were grabbed in two different sessions separated by two weeks, and 13 images in each session were recorded. In our experiments here, we use a subset of the AR face database provided and preprocessed by Martinez. This subset contains 1400 image faces corresponding to 100 objects (50 men and 50 women) where each subject has 14 non-occluded images with variations in expression and illumination (see Fig. 2 for the 14 images of one subject). Here, the images are resized to 66 × 48 and the gray level values are rescaled to [0 1]. We carry out five independent experiments AR1, AR2, AR3, AR4, AR5, where the training and testing samples are summarized in Table 2. From Fig. 2 and Table 2, we can clearly observe that: 1) on AR1, the training samples from each subject have relatively low variations (both of neural expression), and the testing samples have great variations in expression and illumination; 2) on AR2, the four training samples from each subject have variations in expression, and the testing samples have variations in expression and illumination; 3) on AR3, the four training samples from each subject have variations in expression and illumination, and the testing samples have variations in expression and illumination; 4) on AR4, the eight training samples from each subject have great variations in expression, and the testing samples have great variations in illumination; 5) on AR5, the eight training samples from each subject have great variations in illumination, and the testing samples have great variations in expression. Consequently, these five experiments can test the performance of given method under different kinds of difficulty in the training set and testing set.

Columbia Object Image Library (COIL-20) is a database of gray-scale images of 20 objects. The objects were placed on a motorized turntable against a black background. The turntable was rotated through 360 degrees to vary

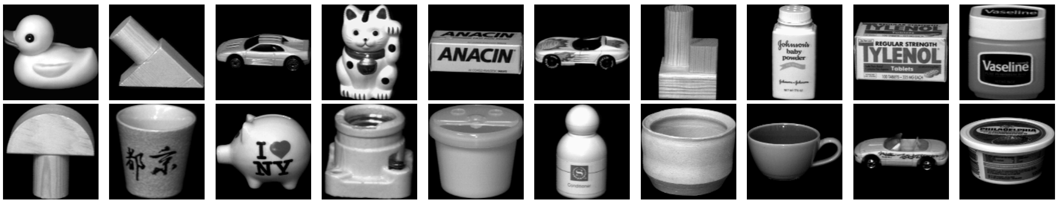


Fig. 3. Illustration of 20 objects in the COIL-20 object database.

Table 3

Data partition on COIL-20 object database

Category	Training	Testing
COIL1	0, 1, 2, 3	4-71
COIL2	0, 4, 8, 12	1-3, 5-7, 9-11, 13-71
COIL3	0, 8, 16, 24	1-7, 9-15, 17-23, 25-71
COIL4	0, 12, 24, 36	1-11, 13-23, 25-35, 37-71
COIL5	0, 16, 32, 48	1-15, 17-31, 33-47, 49-71

object pose with respect to a fixed camera. Images of the objects were taken at pose intervals of 5 degrees, and each object has 72 (numbered between 0 and 71) images with a resolution of 128×128. Fig. 3 illustrates the 20 objects in the COIL-20 object database. Here, we resize the image resolution to 64×64 and rescale the grey level value to [0 1]. We set the number of training samples per class to 4, and carry out five independent experiments COIL1, COIL2, COIL3, COIL4 and COIL5, where the training and testing samples are summarized in Table 3. These five experiments are employed to test the performance of given method under different kinds of variations in the training set.

6.1.2 Experimental setting

For RDA, we let $\alpha = \lambda_{max}(S_w^{'})e^{t-21}$, where $\lambda_{max}(S_w^{'})$ denotes the maximal eigenvalue of $S_w^{'}$ and t varies from 1 to 21 incremented by 1, and when reporting the best classification performance of RDA under given α, we also report its logarithm value $\ln(\alpha/\lambda_{max}(S_w^{'}))$. Similarly, for MMC, we let $\beta = e^{t-5}$, where t varies from 1 to 21 incremented by 1, and when reporting the best classification performance of wMMC under given β, we also report its logarithm value $\ln(\beta)$.

For methods such as DCV, RDA, wMMC, Fisherfaces and Pseudo-inverse Linear Discriminant Analysis (PLDA) [18,26,30], the number of projection vectors is set to the number of classes minus 1, i.e., $c - 1$. The Fisherfaces here operates as [1], namely, PCA first projects the d-dimensional samples to a dimensionality of $n - c$, and then LDA is applied. Moreover, when the features are extracted by the given method, a nearest neighbor classifier using *Euclidean* distance is utilized for reporting classification accuracy.

6.2 Results

We report results on the eigenvalue distribution of wMMC to support Proposition 3 in Section 6.2.1, results on the subspace distances between the projection subspaces of RDA (wMMC) and DCV to support Propositions 4 and 5 in Section 6.2.2. We report results on the values of MSV on different partitions of ORL, AR and COIL-20 in Section 6.2.3, and the classification performance of DCV, wMMC, RDA, Fisherfaces and PLDA in Section 6.2.4.

6.2.1 Results on the eigenvalue distribution of wMMC

The experimental results on ORL1-ORL4, AR1-AR5, and COIL1-COIL5 consistently show that the n training samples are always independent, and the eigenvalue equation (31) has $c-1$ positive and $n-c$ negative eigenvalues. Further, based on the number of positive and negative eigenvalues and employing (38) and Lemma 1, we can clearly get that S_b-βS_w has $d-n+1$ zero eigenvalues. As a result, Proposition 3 is experimentally verified.

6.2.2 Results on the subspace distance

We choose ORL1, ORL4, AR1, AR5, COIL1 and COIL5 as representatives and report the subspace distances between the projection subspaces of DCV and RDA (wMMC) under different α's (β's) in Fig. 4, from which one can easily observe that when t is less than 4, the subspace distance between the projection subspaces of RDA and DCV is 0; and similarly, when t is greater than 18, the subspace distance between the projection subspaces of wMMC and DCV is 0. Furthermore, the experimental results on ORL2-ORL3, AR2-AR4, and COIL2-COIL4 witness the same phenomena. Therefore, Propositions 4 and 5 are experimentally verified.

By the way, from Fig. 4, it is easy to get that MMC (or setting $\beta = 1$ in wMMC) is not equivalent to DCV (or PNS), since the subspace distance between MMC and DCV is not zero. Thus the argument that MMC is actually equivalent to PNS in SSS problem made in [13] is corrected here.

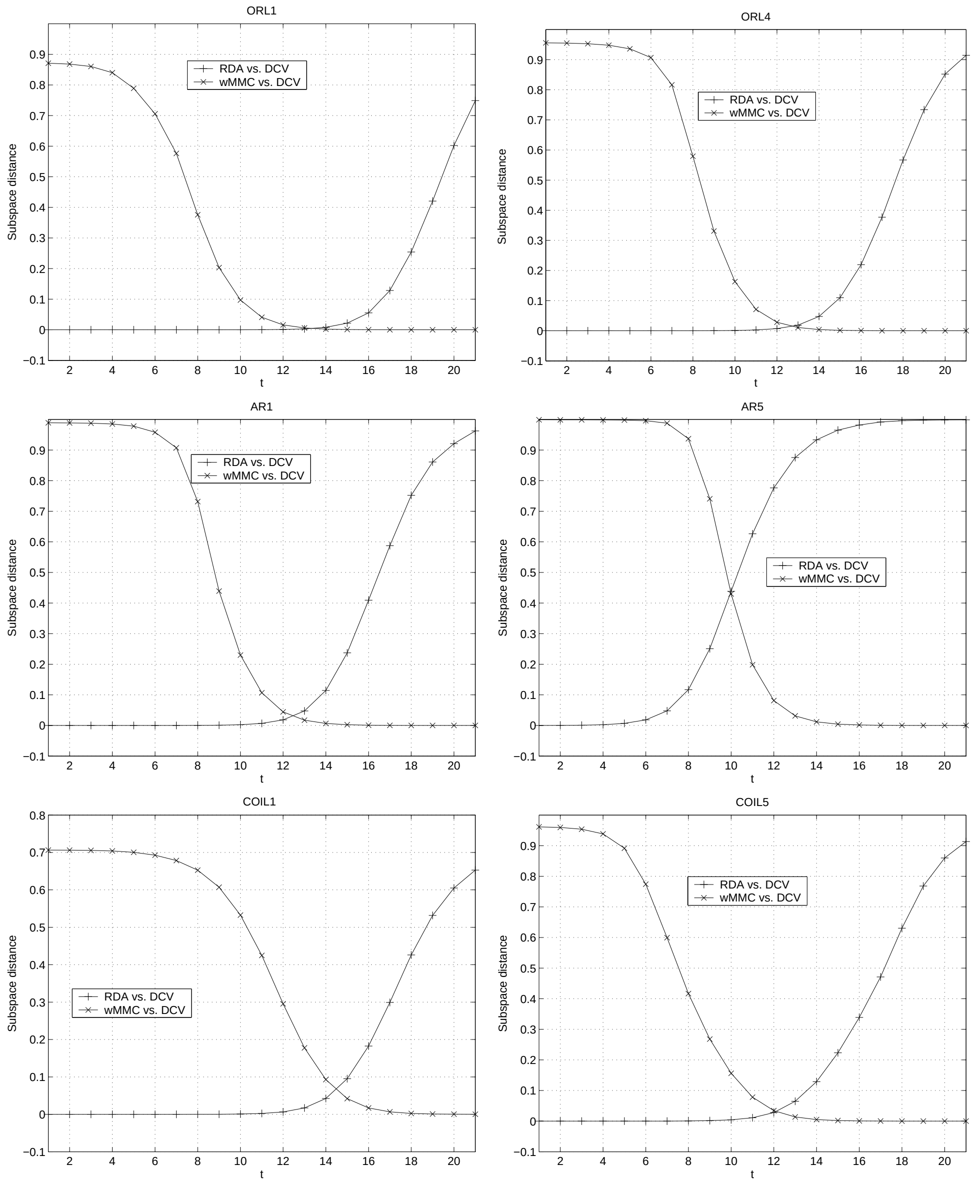


Fig. 4. Substance distance.

6.2.3 MSV criterion

We present the MSV values of each database partition in Tables 4, 5 and 6. From these three tables, we can get that: 1) the MSV values of ORL1-ORL4 are relatively small, which attributes to the fact that ORL is a relatively easy face database with relatively small variations among face images; 2) the MSV value of AR1 is quite small, which is due to the fact that the two training samples

Table 4

Results on ORL

	ORL1	ORL2	ORL3	ORL4
MSV	0.06	0.07	0.08	0.08
Fisherfaces	77.19	81.07	85.83	83.50
PLDA	67.50	77.14	86.67	89.00
DCV	84.06	86.43	91.67	91.50
RDA	85.31 (-2)	88.21 (-1)	92.08 (-3)	92.00 (-4)
wMMC	85.63 (2)	86.43 (1)	92.08 (3)	91.50 (4)

from the same subject are of small variations (both of neutral expression), but the MSV value of AR5 is relatively large, since the eight training samples from the same subject are of great variations in illumination; and 3) with i increasing, COILi generally has larger MSV value, which is due to the fact that the four training samples from the same subject are taken at pose intervals of higher degrees, but COIL4 has smaller MSV value than COIL3, which might attribute to the symmetry characteristic of some objects, e.g., for the object in the second row and first column in Fig. 3, its image numbered 0 is nearer to the one numbered 36 than the ones numbered 8, 16 and 24.

6.2.4 Classification performance

The classification accuracies on ORL1-ORL4, AR1-AR5 and COIL1-COIL5 are respectively reported in Tables 4, 5 and 6, where the classification accuracies of RDA and wMMC in these tables are their optimal results, together

Table 5

Results on AR

	AR1	AR2	AR3	AR4	AR5
MSV	0.06	0.08	0.11	0.10	0.17
Fisherfaces	69.25	73.10	81.80	75.67	74.50
PLDA	63.50	70.70	86.00	73.67	78.00
DCV	80.83	79.30	83.00	87.67	78.33
RDA	80.83 (-20)	81.50 (-5)	86.80 (-7)	87.83 (-12)	86.00 (-5)
wMMC	83.92 (4)	84.60 (3)	86.00 (3)	88.00 (7)	85.00 (-4)

Table 6

Results on COIL-20

	COIL1	COIL2	COIL3	COIL4	COIL5
MSV	0.04	0.11	0.14	0.12	0.14
Fisherfaces	58.46	72.35	82.21	82.80	75.15
PLDA	58.02	73.09	81.77	83.53	76.03
DCV	64.19	77.87	86.25	87.79	80.00
RDA	65.00 (-1)	80.07 (0)	89.93 (-1)	89.19 (-1)	84.93 (0)
wMMC	65.81 (-3)	80.22 (-4)	90.44 (0)	90.15 (0)	85.00 (-3)

with $\ln(\alpha/\lambda_{max}(S_w^{'}))$ and $\ln(\beta)$ in the parentheses. Furthermore, we also plot the classification performance of RDA or wMMC under different α's or β's on ORL1, ORL4, AR1, AR5, COIL1 and COIL5 in Fig. 5.

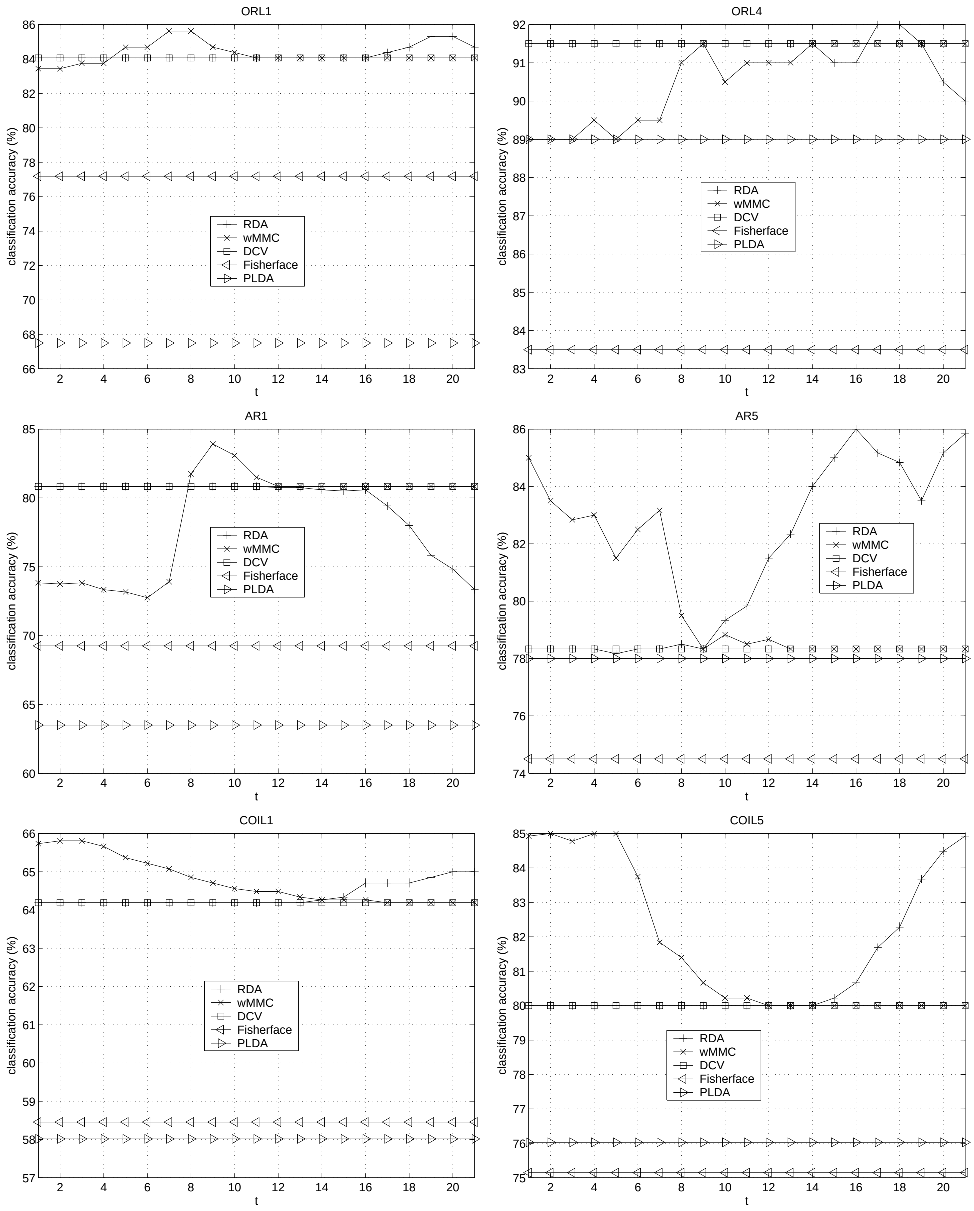


Fig. 5. Classification performance.

Now, we analyze the classification performance as follows:

1) as the results reported in [3], DCV can generally yield superior classification performance to Fisherfaces. The underlying reason might be that, in Fisherfaces, some directions corresponding to the small eigenvalues of S_t are thrown away in the PCA step, and thus applying PCA for dimensionality reduction has the potential to remove dimensions that contain discriminative

information [3].

2) DCV can generally obtain higher classification accuracies than PLDA. The underlying reason might be that, the projection vectors of PLDA are in fact in the range space of S_w (refer to [18] for detail), complement to the projection vectors of DCV (note, the projection vectors of DCV reside in the null space of S_w), and thus DCV fulfills the objective of LDA better than PLDA (note, the Fisher's criterion (5) for DCV is $+\infty$ while that for PLDA is finite).

3) when MSV is relatively small (e.g., on ORL1-ORL2, AR1-AR2 and COIL1), DCV's advantage over Fisherfaces and PLDA is quite obvious, but when MSV is relatively large (e.g., on AR5), DCV just obtains competitive classification performance to Fisherfaces and PLDA. The underlying reason might be that, when MSV is small, the samples from the same class have small variations, and thus it is reasonable that they are concentrated to a common vector; while on the contrary, when MSV is relatively large, overfitting will occur in DCV, since it still concentrates the same class samples with great variations to a common vector, however in this case, it is reasonable and preferable for Fisherfaces and PLDA to acknowledge the great variations among the same class samples and project these samples to different vectors.

4) since DCV is a special case of RDA and wMMC when α and β respectively tend to zero and $+\infty$, then it is natural that the optimal classification accuracies of RDA and wMMC are not inferior to DCV.

5) on one hand, when MSV is relatively small (e.g., on ORL1-ORL4, AR1-AR2 and COIL1), the classification performance of DCV is competitive to the optimal performance of RDA and wMMC. Furthermore, turning to the classification accuracy curves of RDA and wMMC under different α's and

β's on ORL1, ORL4, AR1 and COIL1 (illustrated in Fig. 5), one can easily get that RDA or wMMC under any given parameter α or β obtains either inferior or competitive performance to DCV. On the other hand, when MSV is relatively high (e.g., on AR5 and COIL5), the optimal classification accuracies of RDA and wMMC are higher than those of DCV. Moreover, turning to the classification accuracy curves of RDA and wMMC under different α's and β's on AR5 and COIL5 (illustrated in Fig. 5), one can clearly see that RDA and wMMC under all α's and β's almost always obtain better classification performance than DCV. The underlying reason is as described in 3), namely, when MSV is small, it is intuitively reasonable for DCV to concentrate the same class samples with small variations to a common vector; while when MSV is high, it would be better to employ positive α and finite β that acknowledge the great variations among the samples from the same class in order to prevent overfitting.

6) the results on ORL1-ORL4 shows that, all the methods can generally benefit from more training samples per person. The results on AR1-AR5 show that, when the parameters α and β are well-tuned, RDA and wMMC generally can benefit from more training samples per class, but the performances of Fisherfaces, PLDA and DCV are influenced by the training samples employed. For example, the number of training samples per subject is 4 for AR3 and 8 for AR4, but the classification accuracies of Fisherfaces and PLDA on AR4 are lower than those on AR3, which might be that, AR3 utilizes the samples that have variations in both expression and illumination for training, but AR4 only employs the samples that have variations in expression. The results on COIL-20 further reveal the importance of employing proper training samples, since although the number of training samples per class is 4 for COIL1-COIL5, the

classification accuracies are quite different. Moreover, it is reasonable that the classification accuracies (of all the methods) on COIL3 are higher than those on COIL1 and COIL2, since the four training samples of COIL3 can represent the COIL-20 object library better than the training samples of COIL1 and COIL2. As a result, in the SSS problem, it is very important to employ proper and representative training samples.

7 Conclusion

In this paper, we first reveal that RDA and wMMC in SSS problem can be performed in the PCA transformed space and propose efficient algorithms for RDA and wMMC to be implemented in space and time complexities of $O(dn)$ and $O(dn^2)$ respectively, which are much more efficient than the original ones' $O(d^2)$ and $O(d^3)$. Therefore, both RDA and wMMC can be applied to areas such as face recognition much more efficiently.

Second, we reveal the eigenvalue distribution for wMMC in the case that the training samples are independent. On one hand, such a revelation facilitates choosing projection vectors in wMMC, and on the other hand, the underlying methodology can be employed to analyze the eigenvalue distribution of matrices such as AA^T-BB^T, where A and B have the characteristic that the number of rows are typically larger than that of columns.

Third, we reveal the relationships among these three powerful DR methods, namely, when α and β respectively tend to 0 and $+\infty$, both RDA and wMMC will obtain the same projection subspace as DCV, or equivalently, DCV is in fact a special case of RDA and wMMC under the aforementioned circumstances.

Finally, we compare the classification performance among the three methods to show that, when the MSV criterion is relatively small, DCV can obtain competitive recognition accuracy as both RDA and wMMC under optimal parameters α and β; but when the MSV criterion is relatively high, the optimal classification accuracies of both RDA and wMMC are generally higher than DCV.

In our viewpoint, it is worthwhile to carry out the following studies:

1) The three methods discussed in this paper are mainly for SSS problem, but we think it interesting to extend them to the large sample size problem. A possible way is to resample the training samples to yield a set of SSS representations of the original question and then to ensemble the results of these SSS representations by these methods. In fact, the recent papers such as [14,19] have witnessed the usefulness and powerfulness of the resampling technique in ensembling the results of SSS methods such as Fisherfaces.

2) It is meaningful to offer the boundary of MSV for employing DCV in real applications and to come up with some criteria for automatically setting α (for RDA) and β (for wMMC) appropriate values in case of high MSV, where the criteria should take both the training samples and the given testing sample into consideration. Moreover, for DCV, a possible countermeasure in the case of high MSV value is to split the same class samples to a set of subclasses to lower the MSV value.

3) It is worthwhile to carry out comparative study between one-dimensional and two-dimensional based LDA methods. In this paper, the three LDA based

methods all treat image samples as one-dimensional vectors. Recently, researchers have proposed an important category of LDA based methods that treat image samples as their native two-dimensional matrices to solve the SSS problem, e.g., the two-dimensional LDA method [33] and the framework of 2D Fisher Discriminant Analysis method [11]. Therefore, it is important to compare the one-dimensional based LDA methods and the two-dimensional based LDA methods both theoretically and experimentally.

4) The pattern recognition community has recently witnessed quite a few extensions on subspace analysis methods for the SSS problem, e.g., improved discriminate analysis [35], subspace evolution analysis [32], generalized null space uncorrelated Fisher discriminant analysis [24], generalized discriminant analysis [9], relevance weighted LDA [29], uncorrelated heteroscedastic LDA [23], etc. Although these methods have achieved successes, we think it necessary to conduct in-depth comparative studies to explore when and why a given method is better for classification, which is beneficial for practitioners to select appropriate methods in real applications. Moreover, in [16], we have shown that it is beneficial for classification by employing an intermediate representation before performing subspace analysis. Therefore, when developing new subspace analysis methods, attention should also be paid to the preprocessing of sample patterns.

Acknowledgments

The authors would like to thank the associate editor and the five anonymous reviewers for their constructive comments that significantly improved this paper, AT&T Lab for providing the ORL face database, Martinez for the AR database, Nene et al. for the COIL-20 object database. This work is supported by the National Natural Science Foundation of China under Grant No. 60473035, the Natural Science Foundation of Jiangsu Province under Grant Nos. BK2005122 and BK2006187, and the Innovation Foundation of NUAA under Grand No. Y0603-042.

References

[1] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[2] N.A. Campbell and W.R. Atchley. The geometry of canonical variate analysis. *Systematic Zoology*, 30(3):268-280, 1981.

[3] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana. Discriminative common vectors for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):4–13, 2005.

[4] L.F. Chen, H.Y.M. Liao, M.T. Ko, J. Lin, and G.J. Yu. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713–1726, 2000.

[5] R. Duin, M. Loog, and T.K. Ho. Editorial: Recent submissions in linear dimensionality reduction and face recognition. *Pattern Recognition Letters*, 27(7):707–708, 2006.

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

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

[8] T. Hastie and R. Tibshirani. Efficient Quadratic Regularization for Expression Arrays. *Biostatistics*, 5(3):329-340, 2004.

[9] P. Howland, J. Wang, and H. Park. Solving the small sample size problem in face recognition using generalized discriminant analysis. *Pattern Recognition*, 39(2):277–287, 2006.

[10] R. Huang, Q. Liu, H. Lu, and S. Ma. Solving the small sample size problem of LDA. In *International Conference on Pattern Recognition*, pages 29–32, 2002.

[11] H. Kong, L. Wang, E. Teoh, J. Wang, and R. Venkateswarlu. A framework of 2d fisher discriminant analysis: Application to face recognition with small number of training samples. In *Computer Vision and Pattern Recognition*, pages 1083–1088, 2005.

[12] H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. In *Advances in Neural Information Processing Systems*, 2003.

[13] H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. *IEEE Transactions on Neural Networks*, 17(1):157–165, 2006.

[14] J. Liu and S.C. Chen. Resampling LDA/QR and PCA+LDA for face recognition. In *The Australian Joint Conference on Artificial Intelligence*, pages 1221–1224, 2005.

[15] J. Liu and S.C. Chen. Discriminant common vecotors versus neighbourhood components analysis and laplacianfaces: A comparative study in small sample size problem. *Image and Vision Computing*, 24(3):249–262, 2006.

[16] J. Liu, S.C. Chen, and X. Tan. Fractional order singular value decomposition representation for face recognition. *Pattern Recognition*, Accepted for publication, 2007.

[17] J. Liu, S.C. Chen, X. Tan, and D. Zhang. Comments on “Efficient and robust maximal margin criterion”. *IEEE Transactions on Neural Networks*, Accepted for publication, 2007.

[18] J. Liu, S.C. Chen, X. Tan, and D. Zhang. Efficient pseudo-inverse linear discriminant analysis and its nonlinear form for face recognition. *International Journal of Pattern Recognition and Artifcial Intelligence*, Accepted for publication, 2007.

[19] X. Lu and A. K. Jain. Resampling for face recognition. In *International Conference on Audio- and Video-based Biometric Person Authentication*, pages 869–877, 2003.

[20] A.M. Martinez. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

[21] A.M. Martinez and R. Benavente. The AR face database. Technical report, CVC, 1998.

[22] S.A. Nene, S.K. Nayar, and H. Murase. Columbia object image library (coil-20). Technical report, Department of Computer Science, Columbia University, 1996.

[23] A. Qin, P. Suganthan, and M. Loog. Uncorrelated heteroscedastic LDA based on the weighted pairwise chernoff criterion. *Pattern Recognition*, 38(4):613–616, 2005.

[24] A. Qin, P. Suganthan, and M. Loog. Generalized null space uncorrelated fisher discriminant analysis for linear dimensionality reduction. *Pattern Recognition*, 39(9):1805–1808, 2006.

[25] S. Raudys. *Statistical and Neural Classifiers*. Springer-Verlag, 2001.

[26] S. Raudys and R.P.W. Duin. On expected classification error of the fish linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, 19(5-6):385–392, 1998.

[27] B. Scholkopf, R. Herbrich, and A.J. Smola. A Generalized Representer Theorem. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 416–426, 2001

[28] J.A.K. Suykens, and J. Vandewalle. Least squares support vector machine classifiers *Neural Processing Letters*, 9(3):293-300, 1999.

[29] E. Tang, P. Suganthan, X. Yao, and A. Qin. Linear dimensionality reduction using relevance weighted lda. *Pattern Recognition*, 38(4):485–493, 2005.

[30] Q. Tian, M. Barbero, Z.H. Gu, and S.H. Lee. Image classification by the foley-sammon transform. *Optical Engineering*, 25(7):834–840, 1986.

[31] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–96, 1991.

[32] H. Wang, Y. Zhou, X. Ge, and J. Yang. Subspace evolution analysis for face representation and recognition. *Pattern Recognition*, 40(1):335–338, 2007.

[33] J. Ye, R. Janardan, and Q. Li. Two-dimensional linear discriminant analysis. In *Advances in Neural Information Processing Systems*, 2004.

[34] W. Zheng, C. Zou, and L. Zhao. Weighted maximum margin discriminant analysis with kernels. *Neurocomputing*, 67:357–362, 2005.

[35] X. Zhuang and D. Dai. Improved discriminate analysis for high-dimensional data and its application to face recognition. *Pattern Recognition*, 40(5):1570–1578, 2007.