An Efficient Pseudoinverse Linear Discriminant Analysis method for Face Recognition

Jun Liu, Songcan Chen, Daoqiang Zhang, and Xiaoyang Tan

Department of Computer Science & Engineering,
Nanjing University of Aeronautics & Astronautics, 210016, P.R. China
{j.liu, s.chen, x.tan, dqzhang}@nuaa.edu.cn
WWW home page: http://parnec.nuaa.edu.cn

Abstract. Pseudoinverse Linear Discriminant Analysis (PLDA) is a classical and pioneer method that deals with the Small Sample Size (SSS) problem in LDA when applied to such application as face recognition. However, it is expensive in computation and storage due to manipulating on extremely large $d \times d$ matrices, where d is the dimensionality of the sample image. As a result, although frequently cited in literature, PLDA is hardly compared in terms of classification performance with the newly proposed methods. In this paper, we propose a new feature extraction method named RSw+LDA, which is 1) much more efficient than PLDA in both computation and storage; and 2) theoretically equivalent to PLDA, meaning that it produces the same projection matrix as PLDA. Our experimental results on AR face dataset, a challenging dataset with variations in expression, lighting and occlusion, show that PLDA (or RSw+LDA) can achieve significantly higher classification accuracy than the recently proposed Linear Discriminant Analysis via QR decomposition and Discriminant Common Vectors.

1 Introduction

Linear Discriminant Analysis (LDA) [2] [3] [4] [6] [7] [9] [10] [11] [12] [14] is a popular feature extraction method in pattern recognition. It searches for a set of projection vectors onto which the data points of the same class are close to each other while requiring data points of different classes to be far from each other, in other words, it calculates the projection matrix W that maximizes the Fisher's Linear Discriminant criterion as follows:

$$J_{FLD}(W_{opt}) = \arg\max_{W} |W^T S_b W| / |W^T S_w W| \qquad (1)$$

where S_w and S_b are respectively the within-class scatter matrix and the between-scatter matrix. It has been proven that if S_w is non-singular, then the ratio of (1) is maximized when the column vectors of W are the eigenvectors of $S_w^{-1} S_b$. Unfortunately, in recognition task such as face recognition, S_w is typically singular, due to the fact that the number of the samples is much smaller than the dimension of sample space, i.e., the so-called Small Sample Size (SSS) problem.

Among many methods that address the SSS problem in LDA, Pseudo Linear Discriminant Analysis (PLDA) [6] [9] [10] [11] [12] is a classical and pioneer method that solves the singularity problem by substituting S_w^{-1} with the pseudo inverse S_w^+ [5]. The generalization error of PLDA was studied in [9] [10], when the size and dimension of the training data vary. Pseudo-inverses of scatter matrices were also studied in [6] and the experimental results in [6] showed that the pseudo-inverse based methods are competitive with Fisherfaces [2].

Recently, Discriminant Common Vectors (DCV) [3] [7] and Linear Discriminant Analysis via QR decomposition (LDA/QR) [14] are proposed to solve the SSS problem in LDA. DCV first projects the training samples to the null space of within-class matrix, and then maximizes the between-class matrix in this space. DCV achieves a maximum (infinite) of the criterion in (1), and is reported to yield superior classification accuracy than other LDA based methods, e.g., Fisherfaces. Like DCV, LDA/QR is also a two-stage method which takes the following two steps: 1) Project the training samples to the range space of the between-scatter matrix and 2) Apply LDA in this reduced space. LDA/QR is computationally efficient compared to other LDA based methods and meanwhile can achieve competitive classification accuracy to these methods [14].

There is such an interesting phenomenon in DCV, LDA/QR and other LDA related methods that PLDA was often cited but hardly compared in terms of classification performance. The reason behind this phenomenon can be explained as follows: 1) PLDA is a classical and pioneer method for solving SSS problem incurred in LDA and has been widely studied [6] [9] [10] [11] [12], thus it is often cited; and 2) PLDA manipulates on very large matrices in such application as face recognition, and as a result is expensive in both storage and computation, thus the comparison in terms of classification performance is seldom carried out [3].

Furthermore, the projection matrix yielded by DCV resides in the null space of S_w, the projection matrix of LDA/QR is in the range space of S_b, and as will be revealed in section 4, the projection matrix of PLDA resides in the range space of S_w. That is to say that, although these methods are all LDA based methods, they are quite different in the techniques employed. As a result, a comparison among them is meaningful.

In this paper, we propose an efficient two-step method called RSw+LDA for feature extraction. The novelty of the proposed method lies in that: 1) it is much more efficient than PLDA in both computation and storage; and 2) it is theoretically equivalent to PLDA, meaning that it produces the same projection matrix as PLDA.

In the following, we will review PLDA in section 2, give our RSw+LDA method in section 3, and reveal its equivalence relationship with PLDA in section 4. In section 5, experiments on AR face dataset are carried out to verify the effectiveness of PLDA (or RSw+LDA). And finally, we conclude this paper in section 6.

2 Pseudo Linear Discriminant Analysis (PLDA)

Let the training set composed of C classes, and x_{ij} be a d-dimensional column vector which denotes the j-th sample from the i-th class. The within-class and between-class scatter matrices can be defined as:

$$S_w = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{N_i} (x_j^i - m_i)(x_j^i - m_i)^T = H_w H_w^T \tag{2}$$

$$S_b = \frac{1}{N} \sum_{i=1}^{C} N_i (m_i - \bar{m})(m_i - \bar{m})^T = H_b H_b^T \tag{3}$$

where m_i, H_w and H_b are respectively defined as:

$$m_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_j^i \tag{4}$$

$$H_w = \frac{1}{\sqrt{N}} [x_1^1 - m_1, \ldots, x_{N_1}^1 - m_1, \ldots, x_{N_C}^C - m_C] \tag{5}$$

$$H_b = \frac{1}{\sqrt{N}} [\sqrt{N_1}(m_1 - \bar{m}), \ldots, \sqrt{N_C}(m_C - \bar{m})] \tag{6}$$

$\bar{m}$ is the mean sample of the total training set, N_i is the number of training samples for the i-th class and N $(=N_1 + N_2 + \ldots + N_C)$ is the total number of training samples from these C classes.

To calculate pseudo inverse S_w^+, PLDA performs the Singular Value Decomposition (SVD) [5] of S_w as:

$$S_w = Q_1 \Lambda Q_1^T \tag{7}$$

where $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_k)$ contains the positive eigenvalues of S_w, k (generally equals N-C) is rank of S_w and Q_1 consists of the eigenvectors of S_w corresponding to the k positive eigenvalues. According to [5], the pseudo inverse of S_w is:

$$S_w^+ = Q_1 \Lambda^{-1} Q_1^T \tag{8}$$

Then PLDA calculates the eigenvectors of $S_w^+ S_b$ corresponding to positive eigenvalues as the projection vectors.

Although simple in form, PLDA is expensive in both storage and computation. An analysis is given as follows: 1) the SVD in (7) can be calculated indirectly [3] through applying SVD to $H_w^T H_w$ first in $O(dN^2)$ floating point operations (flops) [5], and the space complexity is $O(dN)$; 2) calculating the eigenvalues and corresponding eigenvectors of $S_w^+ S_b$ is expensive, since it costs $O(d^3)$ flops in computation and $O(d^2)$ in storage. In total, the space and time complexity for PLDA is $O(d^2)$ and $O(d^3)$ respectively. In such applications as face recognition, the sample dimensionality d is typically large, e.g., for 100×100 face image, d equals 10000. As a result, it will cost several hundred Mega Bytes (MB) to store the matrix $S_w^+ S_b$, and flops in the order of 10^{12} to calculate its eigenvalues and eigenvectors.

3 RSw+LDA: An alternative way to perform PLDA

To alleviate PLDA's high storage and computation cost, an alternative method, RSw+LDA is proposed in this section. RSw+LDA, which stands for LDA in the Range Space of the within-class scatter matrix (RSw), is a two-stage method that operates as follows:

1) Calculate Q_1 in (7), and project the training samples by Q_1 to the range space of the within-class matrix, where the within-class matrix and between class matrix can respectively be written as:

$$S_w^{'} = Q_1^T S_w Q_1 = \Lambda \tag{9}$$

$$S_b^{'} = Q_1^T S_b Q_1 = (Q_1^T H_b) Q_1^T H_b)^T \tag{10}$$

2) Calculate the eigenvectors of $\Lambda^{-1} S_b^{'}$ corresponding to positive eigenvalues, and put them into a matrix U. Then $W = Q_1 U$ is the projection matrix calculated by RSw+LDA.

Now, we are in a position to analyze the computation and storage costs of RSw+LDA as follows: 1) Similar to PLDA, calculating Q_1 consumes $O(dN^2)$ flops in computation and $O(dN)$ in space; 2) $\Lambda^{-1} S_b^{'}$ is a $k \times k$ matrix, where k generally equals N-C, then it will cost $O(k^2)$ in storage and $O(k^3)$ in computation to calculate the eigenvalues and eigenvectors. Considering the fact that d is typically larger than N in such application as face recognition, the space and computation cost for RSw+LDA is $O(dN)$ and $O(dN^2)$ respectively. We compare the space and time complexity for RSw+LDA and PLDA in Table 1, from which we can observe that RSw+LDA is much more efficient than PLDA in both storage and computation.

Table 1. Space and computation complexity comparison between RSw+LDA and PLDA

Method	Space	Time
PLDA	$O(d^2)$	$O(d^3)$
RSw+LDA	$O(dN)$	$O(dN^2)$

4 Equivalence between RSw+LDA and PLDA

In this section, we try to reveal that RSw+LDA is in fact equivalent to PLDA, meaning that they can obtain the same projection matrix, for which an analysis is given as follows:

PLDA's eigen-equatioin can be formulated as:

$$S_w^{+} S_b w = \lambda w \tag{11}$$

For discussion convenience, let Q_2 be a matrix composed of the *d*-*k* orthonormal eigenvectors of S_w corresponding to its zero eigenvalues. Consequently the column vectors in $[Q_1 \quad Q_2]$ constitute a set of orthonormal basis vectors for the vector space $R^{d\times 1}$. According to the matrix theory [5], w in (11) can be written as:

$$w = [Q_1 \quad Q_2][p_1^T \quad p_2^T]^T = Q_1 p_1 + Q_2 p_2 \tag{12}$$

Now substituting (8) and (12) into (11), we get

$$Q_1 \Lambda^{-1} Q_1^T S_b (Q_1 p_1 + Q_2 p_2) = \lambda (Q_1 p_1 + Q_2 p_2) \tag{13}$$

By the definition of Q_1 and Q_2, we have $Q_1^T Q_1 = I_1$, $Q_2^T Q_2 = I_2$, $Q_1^T Q_2 = O_1$, $Q_2^T Q_1 = O_2$, where I_1 and I_2 are identity matrices, O_1 and O_2 are zero matrices. Hence, pre-multiplying Q_1^T and Q_2^T respectively to both sides of (13) leads to:

$$\Lambda^{-1} Q_1^T S_b (Q_1 p_1 + Q_2 p_2) = \lambda p_1 \tag{14}$$

$$0 = \lambda p_2 \tag{15}$$

As aforesaid, in PLDA, the eigenvectors of $S_w^+ S_b$ corresponding to positive eigenvalues are employed as projection vectors. Consequently $\lambda > 0$, then from (15), we get:

$$p_2 = 0 \tag{16}$$

Substituting (16) into (14) and (12), we get

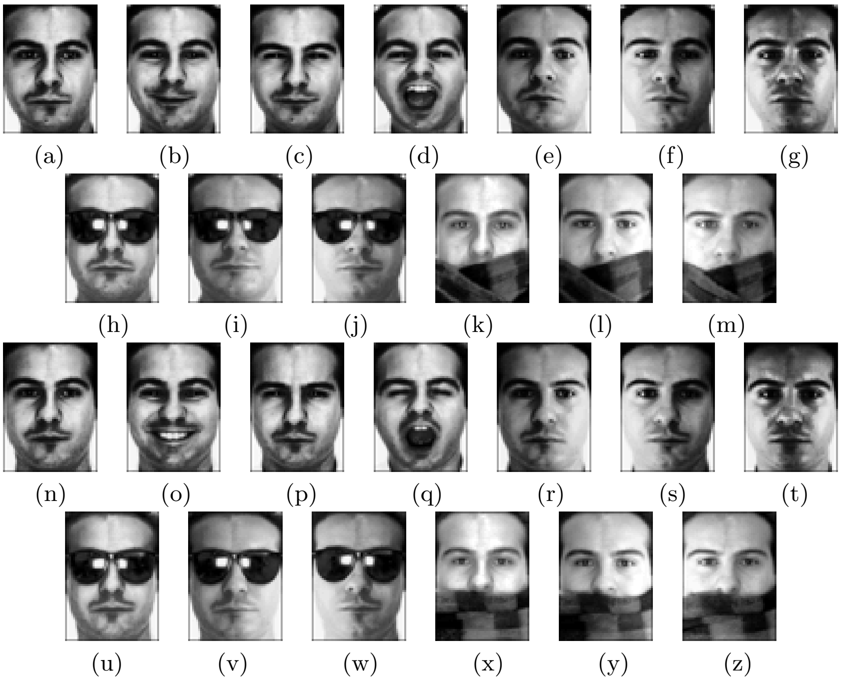
$$\Lambda^{-1} Q_1^T S_b Q_1 p_1 = \lambda p_1 \tag{17}$$

$$w = Q_1 p_1 \tag{18}$$

Note that in RSw+LDA, U calculated in its second stage is indeed the eigenvectors of (17) corresponding to positive eigenvalues, namely the p_1's calculated by PLDA in (17) in fact constitute the matrix U obtained in RSw+LDA. From (18), it is easy to conclude that the projection matrix for PLDA is $Q_1 U$, the same as that of RSw+LDA. As a result, the equivalence relationship between RSw+LDA and PLDA is theoretically verified.

5 Experiments

As mentioned in the introduction, in the recent articles, such as [3] [14], which address the SSS problem in LDA, there is no comparison with PLDA in classification performance yet, which partially attributes to PLDA's demanding computation and storage requirement [3]. Favored by the proposed RSw+LDA, which has been proven to be not only theoretically equivalent to PLDA but also more efficient than PLDA in terms of both computation and storage cost, we can carry out the comparison between PLDA (or RSw+LDA) with other methods. To verify the effectiveness of PLDA, we compare PLDA with the Eigenfaces [13] and the recently proposed Linear Discriminant Analysis via QR decomposition (LDA/QR) [14] and Discriminant Common Vectors (DCV) [3].



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

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

(n) (o) (p) (q) (r) (s) (t)

(u) (v) (w) (x) (y) (z)

Fig. 1. An illustration of 26 images of one subject from AR face database.

We carry out the following experiments on AR [8] face datasets which consists of over 3200 images of frontal images of faces 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. As shown in Fig. 1, for the 13 images, the first one is of neutral expression, the second to the fourth are of facial expression variations, the fifth to the seventh of illumination variations, the eighth to the tenth wearing glasses and the eleventh to the thirteenth wearing scarf. We carry out experiments on a subset of AR face dataset which consists of 2600 images of 100 subjects with each one having 26 images. We make use of the face images from the first session (a-m) for training and those from the second session (n-z) for testing. The face images are preprocessed by Martinez [8] with a resolution of 165×120, and we resize them to 66×48 and rescale the gray level values to [0 1]. When classifying a given unknown sample, it is first projected by the obtained projection matrix, and then it is classified to the same class as its nearest neighbor in the training samples based on Euclidean distance. We report the experimental results in Table 2, from which we can observe that PLDA achieves significantly higher classification accuracy to Eigenfaces (90% information in terms of reconstruction is retained), and such LDA based methods as DCV and LDA/QR in case of great facial variations.

To partially explain the reason why PLDA performs better than such methods as DCV and LDA/QR on AR face dataset, we give an r criterion that

Table 2. Classification accuracy (%) comparison

PLDA	DCV	LDA/QR	Eigenfaces
78.2	71.4	70.8	56.4

measures the ratio of the variances between within-class and between-class as follows:

$$r = trace(S_b)/trace(S_w) \tag{19}$$

When $r > 1$, the between-class variance is large compared to the within-class variance, vice versa. In our experiment on AR, r equals 0.3823, meaning that the samples within the same class have much larger variance than the ones among difference classes. In this case, the estimate for the class mean m_i is biased due to large within-class variance and limited training samples in each class, and as a result the estimate for the between-class scatter matrix S_b is also biased. However, compared to S_b, the within-class scatter matrix S_w can more reliably be estimated without the explicit calculation of m_i, which can be read from the following equation

$$S_w = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N_i}(x_j^i - m_i)(x_j^i - m_i)^T = \frac{1}{N}\sum_{i=1}^{C}\frac{1}{N_i}\sum_{p=1}^{N_i}\sum_{q=1}^{N_i}(x_p^i - x_q^i)(x_p^i - x_q^i)^T \tag{20}$$

For LDA/QR, it first projects the samples to the range space of the ill-estimated S_b, and thus operates poorly due to the so-called centroid sensitivity problem [14]. For DCV, it discards the information in the range space of within-class matrix and concentrates the samples from the same class to a unique common vector, which is not reasonable when the within-class variance is very large. Furthermore, as reported in [7] when MSV defined in [7] is larger than 0.15, DCV operates poorly, then considering the fact that the MSV value equals 0.1967 here, it is reasonable that DCV operates poorly. For PLDA, or RSw+LDA, it first projects the samples to the range space of S_w, which can be relatively reliably estimated in the case of great within-class variations. Further it acknowledges the variance among the same class samples and projects them to different samples. Thus PLDA can obtain better classification performance compared to DCV and LDA/QR.

6 Conclusion

In this paper, we propose an effective and efficient RSw+LDA method for feature extraction and theoretically verify that RSw+LDA is actually equivalent to PLDA, a classical and pioneer method that addresses SSS problem in LDA. RSw+LDA relaxes the demanding computational and storage requirement in PLDA, and thus makes possible the comparison between PLDA and other methods. We carry out experiments on AR face dataset to compare the classification performance between PLDA and such methods such as Eigenfaces, DCV and

LDA/QR, and the conclusion is that PLDA can significantly outperform these methods on this dataset of great within-class variance. Furthermore, based on the proposed RSw+LDA, we can extend the PLDA method to its nonlinear form utilizing the kernel trick similar to [1], and we are currently exploring on this point to make PLDA deal with data of nonlinear distribution better.

Acknowledgments

We thank the anonymous reviewers for constructive comments and we thank Natural Science Foundations of China and Jiangsu under Grant Nos. 60473035, 60505004, BK2005122, BK2006187 for support.

References

1. G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404, 2000.
2. 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*, 19:711–720, 1997.
3. H. Cevikalp, M. Neamtu, Wilke, and A. Barkana. Discriminative common vectors for face recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(1):4–13, 2005.
4. K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 2nd edition, 1990.
5. G.H. Golub and C.F.V. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 1996.
6. W.J. Krzanowski, P. Jonathan, W.V. McCarthy, and M.R. Thomas. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied Statistics*, 44:101–115, 1995.
7. J. Liu and S. 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.
8. A. M. Martinez and R. Benavente. The ar face database. Technical report, CVC, 1998.
9. 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.
10. M. Skurichina and R. Duin. Stabilizing classifiers for very small sample size. In *International Conference on Pattern Recognition*, pages 891–896, 1996.
11. M. Skurichina and R. Duin. Regularization of linear classifiers by adding redundant features. *Pattern Analysis and Applications*, 2(1):44–52, 1999.
12. Q. Tian, M. Barbero, Z.H. Gu, and S.H. Lee. Image classification by the foley-sammon transform. *Opt. Eng.*, 25(7):834–840, 1986.
13. M. Turk and A Pentland. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3(1):71–96, 1991.
14. J. Ye and Q. Li. A two-stage linear discriminant analysis via qr-decomposition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(6):929–941, 2005.