Available online at www.sciencedirect.com
SCIENCE DIRECT
Pattern Recognition 39 (2006) 1002–1006
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Rapid and brief communication

Why direct LDA is not equivalent to LDA

Hui Gao*, James W. Davis

Computer Vision Laboratory, Department of Computer Science and Engineering, The Ohio State University, 395 Dreese Lab, 2015 Neil Avenue, Columbus, OH 43210, USA

Received 26 August 2005; accepted 25 November 2005

Abstract

In this paper, we present counterarguments against the direct LDA algorithm (D-LDA), which was previously claimed to be equivalent to Linear Discriminant Analysis (LDA). We show from Bayesian decision theory that D-LDA is actually a special case of LDA by directly taking the linear space of class means as the LDA solution. The pooled covariance estimate is completely ignored. Furthermore, we demonstrate that D-LDA is not equivalent to traditional subspace-based LDA in dealing with the Small Sample Size problem. As a result, D-LDA may impose a significant performance limitation in general applications.

Keywords: Linear discriminant analysis; Direct LDA; Small sample size problem

1. Introduction

Recently, an algorithm called direct Linear Discriminant Analysis (D-LDA) has received considerable interest in Pattern Recognition and Computer Vision. It was first proposed in Ref. [1] to deal with the small sample size (SSS) problem in face recognition and has been followed with several extensions, e.g., fractional direct LDA [2], kernel based direct LDA [3], and regularized direct discriminant analysis [4].

The key idea in this method is that the null space of the between-class scatter matrix S_b contains no useful information for recognition and is discarded by diagonalization. The within-class scatter matrix S_w is then projected into the linear subspace of S_b and factorized using eigenanalysis to obtain the solution. It was claimed in Ref. [1] that

(1) D-LDA gives the "exact solution for Fisher's criterion".
(2) D-LDA is equivalent to subspace-based LDA (e.g., PCA + LDA) in dealing with the SSS problem.

However, we observe that these claims of D-LDA are flawed in theory. Although the null components of S_b do not influence the projection of S_b in the feature space, they do influence the projection of S_w and hence should not be discarded. Since all "direct" approaches share the same idea (e.g. Refs. [1–4]), we focus on the original work of D-LDA [1] to simplify the discussion. Similar arguments can be made to any of the extensions.

Our analysis originates from the viewpoint of Bayesian decision theory. It is well-known [5] that Fisher's LDA (ratio of S_b and S_w in the projection space) is equivalent to a classification problem of c Gaussians with equal covariance when the model parameters are estimated in the maximum-likelihood (ML) fashion. The solution requires a *minimum* of $c-1$ linear features (assuming input dimension $D \gg c$) to form a sufficient statistic. However in D-LDA, because the null space of S_b is first discarded, its solution is constrained to be in the linear space of S_b (no matter the form of S_w), which is *maximally* $c-1$ dimensional. Hence, the complete $c-1$ dimensional linear space of S_b must be kept as the D-LDA solution in order for it to *possibly* be a sufficient statistic. Due to the fact of ignoring S_w, D-LDA is a special case of LDA.

We additionally point out one missing assumption in the linear algebra derivation of D-LDA given in Ref. [1]. When any singular matrix (S_b or S_w) is involved in the generalized eigenvector and eigenvalue problem, the

* Corresponding author. Tel.: +1 614 247 6095; fax: +1 614 292 2911.
E-mail address: gaoh@cse.ohio-state.edu (H. Gao).

0031-3203/$30.00
doi:10.1016/j.patcog.2005.11.016

diagonalization should start from the non-singular matrix. Since S_b has a maximal rank of $c-1$ as a $D \times D$ matrix ($c \ll D$), it is often singular by the nature of the problem. It should not be diagonalized first. Although the SSS problem further results in a singular or badly scaled S_w, it is due to the lack of examples, which should be carefully handled and *not ignored* (as in D-LDA). Lastly, we show that D-LDA is not equivalent to subspace-based LDA (e.g., PCA + LDA) in dealing with the SSS problem. The claim of D-LDA as a "unified PCA + LDA" [1] is not valid.

In the remainder of this paper, we first review the theory of LDA in Section 2. Then we describe the D-LDA algorithm and prove it as a special case of LDA in Section 3. Experiments are presented in Section 4 with conclusions given in Section 5.

2. Linear discriminant analysis

There are two different perspectives of LDA. One is Fisher's LDA, which is defined by maximizing the ratio of the between-class and within-class scatter matrices (S_b and S_w) in a linear feature space. The other comes from Bayesian decision theory with LDA as a straightforward application of c Gaussians with equal covariance.

2.1. Fisher's LDA

Let S_b and S_w denote the between-class and within-class scatter matrices

$$S_b = \sum_{i=1}^{c} P(\omega_i)(\Psi_i - \Psi)(\Psi_i - \Psi)^\top, \tag{1}$$

$$S_w = \sum_{i=1}^{c} P(\omega_i) \left[\frac{1}{n_i} \sum_{\gamma \in \omega_i} (\gamma - \Psi_i)(\gamma - \Psi_i)^\top \right], \tag{2}$$

where ω_i denotes the ith class with n_i examples and class mean Ψ_i. $P(\omega_i)$ denotes the prior probability of class ω_i. Fisher's LDA looks for a linear subspace W ($c-1$ components), within which the projections of the different classes are best separated, as defined by maximizing the discriminant criteria

$$J(W) = \frac{|W^\top S_b W|}{|W^\top S_w W|}. \tag{3}$$

Along with the orthonormal constraint of W, this can be solved [6] as a generalized eigenvector and eigenvalue problem

$$S_b w_i = \lambda_i S_w w_i \tag{4}$$

with w_i and λ_i being the ith generalized eigenvector and eigenvalue of S_b with regard to S_w. The LDA solution W contains all the $c-1$ eigenvectors with non-zero eigenvalues (S_b has a maximal rank of $c-1$). For a non-singular S_w, it is equivalent to consider a classic eigenvector problem of $S_w^{-1} S_b$ with

$$S_w^{-1} S_b w_i = \lambda_i w_i. \tag{5}$$

However, this requires at least the same number of examples as input dimensions ($N \geqslant D$), which is seldom the case in applications (the SSS problem). For a singular S_w, Fisher's LDA is under-constrained. Any non-trivial vector w in the null space of S_w which yields distinct projections of the class means perfectly maps the variance within each class to 0 (inf Fisher ratio).

2.2. Bayesian decision theory

As a theoretical framework in statistical pattern recognition, Bayesian decision theory assumes the knowledge of the ground-truth probability distributions of each class. The analytical assumption of LDA is the case of c Gaussians with equal covariance Σ. Let μ_i denote the mean of the ith class. The optimal Bayesian classifier can be formulated as the likelihood ratio test (LRT). The ith LRT ($1 \leqslant i \leqslant c-1$) with regard to reference class ω_0 is

$$\begin{aligned} LRT_{0i}(x) &= \frac{P(x \mid \omega_i)}{P(x \mid \omega_0)} \\ &= e^{(\mu_i - \mu_0)^\top \Sigma^{-1} x - (1/2)(\mu_i^\top \Sigma^{-1} \mu_i - \mu_0^\top \Sigma^{-1} \mu_0)} \\ &\underset{0}{\overset{i}{\gtrless}} \frac{P(\omega_0)}{P(\omega_i)} = \tau_{0i}. \end{aligned} \tag{6}$$

An equivalent log-likelihood ratio test (LLRT) is

$$[\Sigma^{-1}(\mu_i - \mu_0)]^\top x \underset{0}{\overset{i}{\gtrless}} \log(\tau_{0i}) + \tfrac{1}{2}(\mu_i^\top \Sigma^{-1} \mu_i - \mu_0^\top \Sigma^{-1} \mu_0). \tag{7}$$

Let $v_i = \Sigma^{-1}(\mu_i - \mu_0)$. Eq. (7) can be interpreted as an input vector x being first projected into a $c-1$ dimensional linear subspace (v_i) followed by thresholding. The linear space $V = span(v_i)$ is the Bayesian solution of LDA, which contains all the linear transformations that are statistically sufficient for optimal classification. When the ML estimates of the Gaussian parameters ($\hat{\mu}_i$ and $\hat{\Sigma}$) are used, $S_w = \hat{\Sigma}$ and S_b is in the linear subspace $span(\hat{\mu}_i - \hat{\mu}_0)$.

Fisher's LDA is closely related to the LDA formulation in Bayesian decision theory in that its solution $eig(S_w^{-1} S_b)$ is one orthonormal basis of the linear subspace $\hat{V}$ (the estimate of the true Bayesian solution V). It is not unique. Any full rank linear transformation in $\hat{V}$ yields the same Fisher ratio defined in Eq. (3). In this sense, the correct LDA solution is a linear subspace, an element in a Grassmann manifold (set of subspaces).

With regard to the SSS problem, the ML estimate of the common covariance $\hat{\Sigma}$ (or S_w) is singular, whereas $\hat{\Sigma}^{-1}$ is required to describe a Gaussian distribution. There are two approaches to deal with this issue. One is to believe in the data by assuming the estimate $\hat{\Sigma}$ being the true Σ (components

not existing in the current examples should never happen in the future). This approach looks for solutions in the linear subspace of examples (subspace-based LDA). The other method assumes the opposite in that the LDA solution may contain null components (null-space-based LDA). However, due to the lack of evidence from examples (any null component maps $\hat{\Sigma}$ to 0), this approach is under-constrained and typically assumes an identity common covariance in the null space.

3. Direct LDA (D-LDA)

Although D-LDA was previously claimed to be equivalent to LDA [1], we show from Bayesian decision theory that D-LDA is actually a special case of LDA. And with regard to the SSS problem, we show that D-LDA is not equivalent to subspace-based LDA (e.g., PCA + LDA).

3.1. Direct LDA algorithm

D-LDA is based on the idea of "simultaneous diagonalization" of S_b and S_w, which is an alternative approach in linear algebra [7] to solve the generalized eigenvector and eigenvalue problem (Eq. (4)). Unlike traditional methods, which first diagonalize S_w, D-LDA first whitens (diagonalizes and scales) S_b and then diagonalizes S_w (see Algorithm 1). This was claimed in Ref. [1] to overcome the SSS problem, which results in a singular S_w. We now describe the limitations with the D-LDA algorithm.

Algorithm 1. D-LDA Algorithm

1: Diagonalize S_b by eigenanalysis.
Find matrix V such that $V^{\mathrm{T}} S_b V = \Lambda$, where $V^{\mathrm{T}} V = I$ and Λ is diagonal. Only keep components with non-zero eigenvalues (at most $c-1$). Let Y be the new basis and D_b be the diagonal matrix of corresponding non-zero eigenvalues. $Y^{\mathrm{T}} S_b Y = D_b$.
2: Project and diagonalize S_w.
Let $Z = Y D_b^{-1/2}$ (whitening S_b). Factorize $Z^{\mathrm{T}} S_w Z$ using eigenanalysis $U^{\mathrm{T}}(Z^{\mathrm{T}} S_w Z)U = D_w$ where $U^{\mathrm{T}} U = I$ and D_w is diagonal. Keep eigenvectors with smallest eigenvalues.
3: Reconstruct the matrix of feature vectors $W = ZU D_w^{-1/2}$.
For a given input vector x, its projection in the feature space $x^* = W^{\mathrm{T}} x = D_w^{-1/2} U^{\mathrm{T}} Z^{\mathrm{T}} x$.

3.2. Issue 1: Theoretical deficiency

Consider the case of two Gaussians with class means μ_0 and μ_1. Let the common covariance Σ be full rank. The Bayesian LDA solution is a single feature vector (Eq. (7))

$$v = \Sigma^{-1}(\mu_1 - \mu_0). \tag{8}$$

In general, this is not in the linear space of $\mu_1 - \mu_0$ due to the presence of Σ. With ML estimates $S_w = \hat{\Sigma}$ and $S_b = (\hat{\mu}_1 - \hat{\mu}_0)(\hat{\mu}_1 - \hat{\mu}_0)^{\mathrm{T}}$, Fisher's LDA solution (from Eq. (5)) is

$$w = eig(S_w^{-1} S_b) = \alpha \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) \tag{9}$$

with a scalar α as the vector normalization factor [6]. This is theoretically the same as the Bayesian solution (Eq. (8)).

However, if the D-LDA algorithm is followed, the linear subspace of S_b is simply $span(\hat{\mu}_1 - \hat{\mu}_0)$. The projection of S_w into this space therefore is a single number. The D-LDA solution is

$$w = \beta(\hat{\mu}_1 - \hat{\mu}_0) \tag{10}$$

with β being a constant scalar. It is clearly not equivalent to the LDA solution (Eqs. (8) and (9)).

Similar analysis can be extended to a c-class problem. The linear space of S_b has at most $c-1$ components, which results in a $(c-1) \times (c-1)$ dimensional covariance matrix $Z^{\mathrm{T}} S_w Z$ in Step 2 of Algorithm 1. Because each column vector in $W = ZUD^{-1/2}$ is linearly reconstructed from Z (Step 3), the D-LDA solution is limited to be in the linear subspace of S_b. However, from Bayesian decision theory, a minimum of $c-1$ feature vectors are required to form a sufficient statistic. In order for the D-LDA solution to possibly be a sufficient statistic, the entire linear subspace of S_b must be kept as the D-LDA solution. In this sense, D-LDA completely ignores the common covariance estimate $\hat{\Sigma}$ (or S_w) and purely depends on the class means for classification, which is indeed a special case of LDA.

Lastly, in the linear algebra derivation given in Ref. [1], a key assumption was missed when using the simultaneous diagonalization method to solve the generalized eigenvector problem. Only when S_b and S_w are both non-singular, the solutions to $eig(S_w^{-1} S_b)$ (equivalent to diagonalizing S_w first) and $eig(S_b^{-1} S_w)$ (equivalent to diagonalizing S_b first) share the same eigenvectors, but reciprocal eigenvalues (compare Eqs. (5) and (11))

$$S_b^{-1} S_w w_i = \frac{1}{\lambda_i} w_i. \tag{11}$$

However, if any singular matrix is involved, diagonalization in this method replaces the inverse matrix with the pseudoinverse, which results in different solutions for Eqs. (5) and (11). To avoid the pseudoinverse, the diagonalization should always start from the non-singular matrix. However, for a c-class problem, the rank of S_b is at most $c-1$, which is not dependent on the sample size and is determined by the nature of the problem to be often singular ($D \gg c$). Hence the matrix S_b should not be first diagonalized as in D-LDA. Only when $D \leqslant c-1$ is D-LDA equivalent to LDA. But there will be no dimensionality reduction in this case.

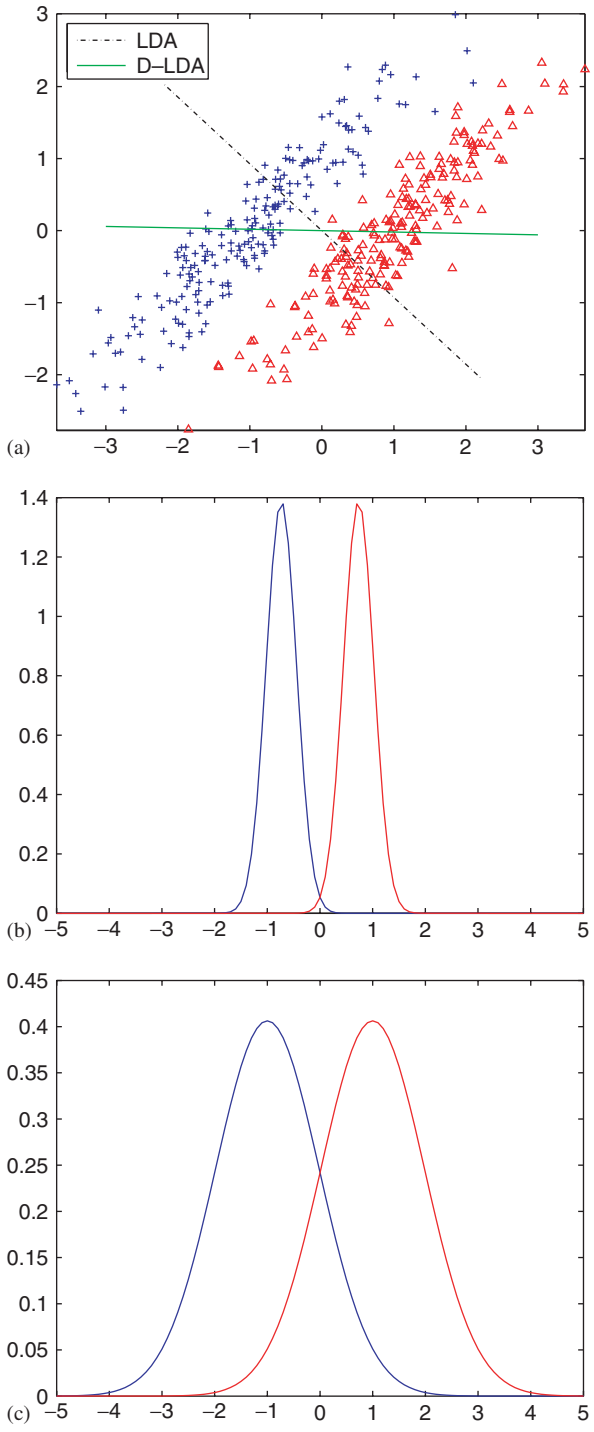

Fig. 1. Simulation results of two Gaussians with 200 samples each. (a) LDA and D-LDA feature vectors. (b) Projections in LDA feature space. (c) Projections in D-LDA feature space.

3.3. Issue 2: Relation to subspace-based LDA

It was also claimed in Ref. [1] that D-LDA is equivalent to subspace-based LDA (e.g., PCA + LDA) in dealing with the SSS problem, which results in a singular S_w due to the lack of examples. Although Fisher's LDA can be equivalently defined as

$$\hat{J}(W) = \frac{|W^\top S_t W|}{|W^\top S_w W|} \tag{12}$$

with $S_t = S_b + S_w$ (covariance matrix used in PCA), as employed in Ref. [1] to support the claim of D-LDA as "unified PCA + LDA", the D-LDA algorithm only projects and diagonalizes S_w (Step 2) after extracting the linear space of S_t (PCA) in Step 1. However, this is different from PCA + LDA, which projects both S_b and S_w in the linear space of S_t and simultaneously diagonalize them. Furthermore, it has not been proven that Eqs. (3) and (12) are equivalent under the special case D-LDA algorithm.

4. Experiments

To verify our claim from Bayesian decision theory, we set up a two-class synthetic experiment with $\mu_0 = [-1, 0]^T$, $\mu_0 = [0, 1]^T$, and $\Sigma = [1, 0.92; 0.92, 1]$. The LDA feature vector is $v_{LDA} \approx [13.0, -12.0]^T$ (Eq. (8)), or $[0.74, -0.68]^T$ after normalization Eq. (9). But the normalized D-LDA solution is $v_{D\text{-}LDA} = [1, 0]^T$ (Eq. (10)). In terms of classification, LDA yields the theoretical recognition rate of 99.5%, while D-LDA only gives 84.1%. In a simulation with 200 examples per class, the estimated feature vectors of D-LDA and LDA (along with their Gaussian projections) are shown in Fig. 1. This clearly illustrates the non-equivalence of D-LDA to LDA.

Next we compare D-LDA to a traditional subspace-based LDA method (EFLD, a variation of PCA + LDA which adjusts the number of PCA components [7] for the optimal results) in dealing with the SSS problem in real applications. The same ORL face dataset (40 subjects) in Ref. [1] was employed, where 5 out of 10 images per person were randomly drawn as test images (10 repeats). Although an average test recognition rate of 90.8% was reported in [1] (comparable to 91.4% in our result), significant performance drops (see Table 1) were found for harder versions of the same dataset: "face cropped" (77.1%) and then "intensity normalized" (73.4%). As a comparison, EFLD gave 96.5% for the original data, 88.1% for "cropped", and 85.6% for "normalized", consistently better than D-LDA. This explicitly demonstrates that D-LDA has no performance advantage over subspace-based LDA in dealing with the SSS problem.

Table 1
Classification rate for the ORL dataset. Original data was first tightly "cropped" to the face region (after smoothing and down-sampling) and then "normalized" by subtracting mean and dividing by std of pixel intensities

	Original (%)	Cropped (%)	Normalized (%)
D-LDA	91.4	77.1	73.4
EFLD	96.5	88.1	85.6

5. Conclusion

In this paper, we presented both theoretical and experimental analysis of the shortcomings of D-LDA. Despite its recent popularity, we showed that D-LDA is actually a special case of LDA, which directly takes the linear space of class means as the solution. Furthermore, we demonstrated that D-LDA is not equivalent to traditional subspace-based LDA in dealing with the SSS problem. Though D-LDA may work well in applications with well-separated classes, the method imposes a significant performance limitation in general cases.

References

[1] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data—with application to face recognition, Pattern Recognition 34 (2001) 2067–2070.

[2] J. Lu, K. Plataniotis, A. Venetsanopoulos, Face recognition using LDA-based algorithms, IEEE Trans. Neural Networks 14 (1) (2003) 195–200.

[3] X. Wu, J. Kittler, J. Yang, K. Messer, A new kernel direct discriminant analysis (KDDA) algorithm for face recognition, in: British Machinery and Vision Conference, 2004.

[4] J. Lu, K. Plataniotis, A. Venetsanopoulos, Regularized discriminant analysis for the small sample size problem in face recognition, Pattern Recognition Lett. 24 (2003) 3079–3087.

[5] N. Campbell, Canonical variate analysis—a general model formulation, Austr. J. Statist. 26 (1984) 86–96.

[6] R. Duda, P. Hart, D. Stork, Pattern Classification, Wiley, New York, 2001.

[7] C. Liu, H. Wechsler, Enhanced Fisher linear discriminant models for face recognition, in: Proceedings of International Conference on Pattern Recognition, IEEE, 1998.