Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence

Linear Discriminant Analysis: New Formulations and Overfit Analysis

Dijun Luo, Chris Ding, Heng Huang
The University of Texas at Arlington, Arlington, Texas, USA
dijun.luo@gmail.com, chqding@uta.edu, heng@uta.edu

Abstract

In this paper, we will present a unified view for LDA. We will (1) emphasize that standard LDA solutions are not unique, (2) propose several new LDA formulations: St-orthonormal LDA, Sw-orthonormal LDA and orthogonal LDA which have unique solutions, and (3) show that with St-orthonormal LDA and Sw-orthonormal LDA formulations, solutions to all four major LDA objective functions are identical. Furthermore, we perform an indepth analysis to show that the LDA sometimes performs poorly due to over-fitting, *i.e.*, it picks up PCA dimensions with small eigenvalues. From this analysis, we propose a stable LDA which uses PCA first to reduce to a small PCA subspace and do LDA in the subspace.

Introduction

Linear discriminant analysis (LDA)(Fisher 1936) is widely used classification method, especially in applications where the data dimension is large, such as in computer vision(Turk and Pentland 1991; Belhumeur, Hespanha, and Kriengman 1997) where data objects are images with typically 100^2 dimensions. Since it is invented in late 1940's, there is a large number of studies on LDA methodology, among them Fukunaga's book(Fukunaga 1990) is the most authoritative. Since 1990, there are many developments, such as uncorrelated LDA(Jin et al. 2001), orthogonal LDA, (Ye and Xiong 2006), null-space LDA(Chen et al. 2000), and a host of other methods such as generalized SVD (Park and Howland 2004) for LDA, 2DLDA (Ye et al. 2004; Luo, Ding, and Huang 2009) , etc.

For the simple formulation of LDA of Eq. (1), it is a bit surprising to have this large number of varieties. In this paper, we undertake a different route. Instead of developing newer methods, we ask a few fundamental questions about LDA.

Given the fact that there are so many LDA varieties, a natural question is: is the LDA solution unique? A related question: is the LDA solution global solution? Consulting on Fukunaga's book and other books(Duda, Hart, and Stork 2000; Hastie, Tibshirani, and Friedman 2001), and reading recent papers, these questions were not addressed (or not emphasized at least), to the best of our knowledge.

Our investigation of this neglected area of LDA uncover a large number of new results regarding uniqueness, normalization, global solutions. We also investigate the LDA overfitting problem. Our experiments on real life datasets show that LDA often overfits by incorporating PCA (principal component analysis) dimensions with small eigenvalues which causes poor performance. Our results suggest several new approaches to improve the performance.

Outline of New Results

We first introduce the LDA formulation and outline the new results.

Classic LDA. In LDA, the optimal subspace $G = (\mathbf{g}_1, \cdots, \mathbf{g}_k)$ is obtained by optimizing

$$\max_G J_1(G) = \mathrm{Tr}\frac{G^T S_b G}{G^T S_w G} \tag{1}$$

where the between-class (S_b) and within-class (S_b) scatter matrices are defined as

$$S_b = \sum_k n_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T, \tag{2}$$

$$S_w = \sum_k \sum_{i \in C_k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T, \tag{3}$$

where $\mathbf{m}_k$ is the mean of class C_k and $\mathbf{m}$ is the global total mean. The total covariance matrix is $S_t = S_b + S_w$. The central idea is to separate different classes as much as possible (maximize the between-class scatter S_b) while condense each class as much as possible (minimize the within-class scatter S_w).

Traditional Solution In both the book and most (if not all) previous papers, the solution of G for LDA is considered to be given the k eigenvectors of $S_w^{-1} S_b$

$$S_w^{-1} S_b G_0 = G_0 \Lambda_0,\ \Lambda_0 = (\lambda_1^0, \cdots, \lambda_k^0), \tag{4}$$

associated with the largest eigenvalues. The dimension k of the subspace is set to $k = C - 1$ where C is the number of classes. We call this *traditional* solution.

New LDA Formulations

We first note that the classic LDA is implicitly defined with constraints:

$$\max_G \mathrm{Tr}\frac{G^T S_b G}{G^T S_w G},\quad s.t.\ \{g_k\}\text{ linearly independent, } ||g_k|| = 1.$$

(5)

Without the constraint the optimal solution would be $G_1 = (g_1, \cdots, g_1)$, because

$$\mathrm{Tr}\frac{G_1^T S_b G_1}{G_1^T S_w G_1} = k\lambda_1 > \mathrm{Tr}\frac{G_0^T S_b G_0}{G_0^T S_w G_0} = \lambda_1 + \cdots + \lambda_k.$$

Now, we introduce several more meaningful constraints.

S_t-orthonormal LDA

First, we consider the S_t-orthonormal LDA:

$$\max_G \mathrm{Tr}\frac{G^T S_b G}{G^T S_w G},\ s.t.\ G^T S_t G = I, \tag{6}$$

This is a meaningful variant, because the projected data

$$y_i = G^T x_i,\ Y = G^T X,$$

are obtained such that the total covariance matrix for Y

$$S_t(Y) = G^T S_t(X) G = I. \tag{7}$$

Thus the projected data Y is not only uncorrelated, they are also properly orthonormal. Transforming data into a unit-covariance is often done in the prepossessing stage in statistical analysis. It often helps the analysis.

S_w-orthonormal LDA

We consider the S_w-orthonormal LDA:

$$\max_G \mathrm{Tr}\frac{G^T S_b G}{G^T S_w G},\ s.t.\ G^T S_w G = I, \tag{8}$$

This is a meaningful variant, because G is obtained such that the total within-class covariance matrix for the projected data $Y = G^T X$

$$S_w(Y) = G^T S_w(X) G = I. \tag{9}$$

This is called sphering the data in statistics (Hastie, Tibshirani, and Friedman 2001). Sphering the within-class covariance matrix is the direct motivation for LDA.

Orthogonal LDA

We will show in Lemma 2 that the classic LDA solution G_0 are not orthogonal, *i.e.*, $G_0^T G_0 \neq I$. In many subspace project, we desire the projection directions are mutually orthonormal. For example, PCA projections are mutually orthonormal. Therefore, we propose the orthogonal LDA as the following:

$$\max_G \mathrm{Tr}\frac{G^T S_b G}{G^T S_w G},\ s.t.\ G^T G = I, \tag{10}$$

Main Results

Our main results are Theorems 1 and 2 below. Let $G_{\text{Lin.ind.}}$ be the solution to the linearly independently constrained LDA of Eq.(5); G_t be the optimal solution to the S_t orthonormal LDA of Eq.(6); G_w be the optimal solution to the S_w orthonormal LDA of Eq.(8); and G_{orth} be the optimal solution to the orthonormal LDA of Eq.(10). Our main results are:

Theorem 1 *(1) All these four solutions are the Global solutions for the four different LDA formulations. (2) All these four different LDA formulations attain the same objective function value:*

$$J_1(G_{Lin.ind.}) = J_1(G_t) = J_1(G_w) = J_1(G_{orth}).$$

The proof is given in later section.

LDA Objective Functions

Besides the J_1 objective function of Eq.(1), there exist three other objective functions as mentioned in Fukunaga's book (Fukunaga 1990)[p.447]. In this paper, we show some interesting results on the different LDA objective functions.

The first and most commonly used LDA objective is J_1 of Eq.(1). The second is determinant based objective:

$$\max_G J_2(G) = \frac{\det G^T S_b G}{\det G^T S_w G}. \tag{11}$$

The third objective is the difference of traces:

$$\max_G J_3(G) = \mathrm{Tr}G^T S_b G - (\mathrm{Tr}G^T S_w G - \mu). \tag{12}$$

The 4th objective is the ratio of traces:

$$\max_G J_4(G) = \frac{\mathrm{Tr}G^T S_b G}{\mathrm{Tr}G^T S_w G}. \tag{13}$$

Fukunaga showed that J_2 is essentially identical to J_1. But he dismissed J_3 and J_4.

Interestingly, all four above objective functions are the same under certain reasonable constraints:

Theorem 2 *(1) Under the S_t-orthonormal constraint Eq.(7). The optimal solutions for all four objective functions J_1, J_2, J_3, J_4 of Eqs.(1,11,12,13) are identical. (2) Under the S_w-orthonormal constraint Eq.(9), the optimal solutions for all four objective functions are identical.*

Theorems 1 & 2 are the main results of this paper. They provide a unified view of all LDA objective functions and formulations. (The fact that solutions of J_1 and J_2 are identical is previously known (Fukunaga 1990)[p.447].)

Invariance of LDA

It is comforting that all optimal solutions to various LDA formulations are global solutions, i.e., there is no local optimal solution.

However, global solution may not be unique. Recall the definition of global solution: $\tilde{G}$ is a global solution if $J(G) \leq J(\tilde{G})$ for any G. We could have, however, $\tilde{G}_1 \neq \tilde{G}_2$ and $J(\tilde{G}_1) = J(\tilde{G}_2)$. Thus both $\tilde{G}_1$ and $\tilde{G}_2$ are global solutions.

Sign and Rotational Invariance

Suppose $G = (g_1, g_2, \cdots, g_k)$ is a global optimal solution to LDA objective $J_1(G)$ of Eq.(1). Then it is easy to see that there are 2^k variants

$$G_S = (\pm g_1, \pm g_2, \cdots, \pm g_k) = GS,$$

where $S = \mathrm{diag}(\pm 1, \pm 1, \cdots, \pm 1)$ contains the signs. It is easy to see $J_1(G) = J_1(G_S)$ for any S. This means there are 2^k global solutions.

In fact, $J_1(G)$ has the rotational invariance. Let R be an orthonormal matrix: $RR^T = R^T R = I$. A special case of the rotational transformation is sign transformation $R = S$. The coordinate transformation under the rotation R is: $y_i = R^T x_i$ or $Y = R^T X$. And the scatter matrices S_b, S_w are transformed as

$$S_b(Y) = RS_b(X)R^T,\ \ S_w(Y) = RS_w(X)R^T.$$

It is easy to see that

Proposition 3 *(1) All four LDA objective functions of Eqs.(1,11,12,13) are rotational invariant. (2) All 4 constraints: the linear independent, the S_b-orthonormal, the S_w-orthonormal, the orthonormal constraints are rotational invariant.*

Therefore, LDA solutions can not be unique in the strict sense. If G^* is the global optimal solution for any one of the four LDA objective functions with any one of the four constraints, then G^*R is also a global optimal solution.

However, this non-uniqueness due to a rotational transformation is not a problem in many pattern recognition applications. For examples, if we do KNN classification or K-means clustering, this rotational invariance causes no problem, because KNN and K-means are themselves rotational invariant. We also note that PCA solution has the same rotational invariance.

Generic linear invariance

The $J_1(G)$ objective has a *generic linear invariance* property. Let $A \in \Re^{m\times m}$ be a non-singular matrix, which is more general than rotation R. A defines a linear transformation $y_i = A^T x_i$ or $Y = A^T X$.

Proposition 4 *J_1 is invariant under any nonsingular linear transformation A. J_2 has the same invariance, while J_3, J_4 are not invariant.*

Proof. Under the transformation, $S_b(Y) = A^T S_b(X)A$, and $S_w(Y) = A^T S_w(X)A$. Thus

$$\begin{aligned} J_1(Y) &= \mathrm{Tr}\frac{A^T S_b(X)A}{A^T S_w(X)A} \\ &= \mathrm{Tr}\,[A^T S_w(X)A]^{-1}[A^T S_b(X)A] \\ &= \mathrm{Tr}\,[A^{-1}S_w^{-1}(X)(A^T)^{-1}][A^T S_b(X)A] \\ &= \mathrm{Tr}\,S_w^{-1}(X)S_b(X). \end{aligned} \tag{14}$$

□

Proposition 4 can be stated in a different way: Suppose G^* is an optimal solution. Then $G^{**} = G^*A$ is also an optimal solution.

This generic invariance is important for proving Theorem 1. It was briefly noted in Fukunaga's book(Fukunaga 1990)[p.447] in passing without elaboration.

The generic invariance of J_1 is the source of non-uniqueness of the global solutions to LDA. Fortunately, the four constraints discussed this paper are not generic invariant. Global solutions to the four constrained LDA formulations are *unique* (up to a rotation). This result is the main motivation of emphasizing the 4 new constrained LDA formulations.

Solutions to New LDA Formulations

The S_w-orthonormal LDA solution

The S_w-orthonormal LDA formulation of Eq.(8 can be written as

$$\max_G \mathrm{Tr} G^T S_b G, \;\; s.t.\; G^T S_w G = I, \tag{15}$$

Assuming S_w is non-singular and let $F = S_w^{1/2}G$, this becomes

$$\max_F \mathrm{Tr} F^T S_w^{-\frac{1}{2}} S_b S_w^{-\frac{1}{2}} F, \;\; s.t.\; F^T F = I, \tag{16}$$

Because $S_w^{-1/2} S_b S_w^{-1/2}$ is a positive definite symmetric matrix, solution to the optimization of Eq.(16) are the principal eigenvectors. Therefore, the global solution $F = (f_1, \cdots, f_k)$ are given by the K eigenvectors of (associated with k largest eigenvalues)

$$S_w^{-1/2} S_b S_w^{-1/2} f_k = \lambda_k f_k, \tag{17}$$

This is consistent, because eigenvectors of Eq.(17) automatically satisfies the orthogonality $F^T F = I$. Thus the solution of S_w-orthonormal LDA is

$$G_w = S_w^{-1/2} F. \tag{18}$$

We have automatically $G_w^T S_w G_w = I$. We list some useful relations below:

$$\begin{aligned} S_b G_w &= S_w G_w \Lambda, \\ G_w^T S_b G_w &= \Lambda, \\ \Lambda &= \mathrm{diag}(\lambda_1, \cdots, \lambda_k). \end{aligned} \tag{19}$$

Relations between G_w and G_0, G_t, G_{orth}

Given S_w, we can obtain the traditional solution G_0, the solution G_t to the S_t-orthonormal LDA, and the solution G_{orth} to the orthonormal LDA via the theoretical relations:

Theorem 5 *, G_0, G_t, G_{orth} relate to G_w by the relations*

$$\begin{aligned} &G_0 = S_w[\mathrm{diag}(G_w^T G_w)]^{-\frac{1}{2}}, \\ &G_t = G_w(I+\Lambda)^{-\frac{1}{2}}, \\ &G_{orth} = G_w(G_w^T G_w)^{-\frac{1}{2}}. \end{aligned} \tag{20}$$

Proof of Theorem 1

With the developments in previous sections, we are ready to prove Theorems 1. We first prove the second part of Theorem 1: $J_1(G_0) = J_1(G_\mathrm{t}) = J_1(G_\mathrm{w}) = J_1(G_\mathrm{orth})$. This is obvious now because: (1) By Theorem 5, $G_0, G_\mathrm{t}, G_\mathrm{orth}$ relates to G_w through linear transformations, and (2) By Proposition 4, $J_1(G)$ is invariant w.r.t. these linear transformations.

We now prove the first part of Theorem 1, i.e., $G_0, G_\mathrm{t}, G_\mathrm{orth}, G_\mathrm{w}$ are global optimal solutions.

From §4.1, S_w is the global solution. This fact, together with $J_1(G_0) = J_1(G_\mathrm{t}) = J_1(G_\mathrm{w}) = J_1(G_\mathrm{orth})$, implying that $G_0, G_\mathrm{t}, G_\mathrm{orth}$ are also global optimal solutions.

We prove this by contradiction. Suppose this is not true, i.e., there exists a $G_0' \neq G_0$ and $J_1(G_0') > J_1(G_0)$. Then through the first relation in Theorem 5, we obtain $G_w' = G_0' D^{1/2}$, and $J_1(G_w') = J_1(G_0')$ because $J_1(G)$ is generic invariant. This leads to $J_1(G_w') > J_1(G_w)$ which contradicts to the fact that G_w is global solution.

Proof of Theorem 2

The first part of Theorem 2 becomes obvious because J_3 maximization of Eq.(12) and J_4 maximization of Eq.(13) becomes identical to Eq.(15), which is identical to J_1 maximization of Eq.(8) and J_2 maximization with same constraint. Thus they all have the identical solution S_w.

To prove the second second part of Theorem 2, we note the known fact (Fukunaga 1990),

Proposition 6 *Under any orthogonality conditions, the following optimizations are identical*

$$\max_G Tr \frac{G^T S_b G}{G^T S_w G} \quad \Longleftrightarrow \quad \max_G Tr \frac{G^T S_b G}{G^T S_t G}. \tag{21}$$

From this, the S_t-orthonormal LDA can be cast as

$$\max_G \mathrm{Tr} G^T S_b G, \ s.t. \ G^T S_t G = I, \tag{22}$$

Under S_t-orthonormality, $J_3 = 2\mathrm{Tr} G^T S_b G + const$, identical to Eq.(22). Thus J_3 has the same solution G_t. J_4 becomes $J_4 + 1 = \frac{\mathrm{Tr} G^T (S_b + S_w) G}{\mathrm{Tr} G^T S_w G} = \frac{\mathrm{Tr} G^T S_t G}{\mathrm{Tr} G^T S_t G - \mathrm{Tr} G^T S_b G}$. Since $\mathrm{Tr} G^T S_t G$ = constant, $\max_G J_4(G)$ becomes $\max_G \mathrm{Tr} G^T S_b G$, reducing to Eq.(22). Thus the solution to J_4 is G_t, which is also the solution for J_1, J_2.

Overfitting Analysis in LDA

Although elegant and effective in some applications, LDA can also overfit. In this section, we analyze the overfitting problem.

Our results are that LDA often incorporates many dimensions associated with small PCA eigenvalues (we call these dimensions *insignificant PCA dimensions*). These insignificant PCA dimensions are always ignored in PCA. Their inclusion in LDA causes overfit which often shows up in unstable LDA results especially in cross validation.

Theoretically, due to the presences of S_w in the denominator of Eq.(1) or S_t in Eq.(22), the weight of dimensions with small eigenvalues of S_w (S_t) are magnified. This cause the overfitting.

Empirically, we have found (and other studies also implicitly support) that if we use all data points (test data and training data) in computing the LDA subspace, and then do cross validation, the results are particularly good. However, if we compute the LDA subspace using training data only and do cross-validation, the results are much more realistic. This shows the LDA subspace change quite significantly using different portion of the data as training data. This large fluctuations is due to the inclusion of insignificant PCA dimensions.

For this purpose, we use the S_t-orthonormal LDA where the analysis takes a very simple form.

We write the total covariance matrix $S_t = U \Sigma U^T$. We do a 2-stage transform

$$G = G_{\text{PCA}} \Sigma^{-\frac{1}{2}} G_b, \ G_{\text{PCA}} = U$$

This is equivalent to 2 transforms: $y_i = G_{\text{PCA}}^T x_i = U^T x_i$,

$$z_i = G_b^T \Sigma^{-\frac{1}{2}} G_{\text{PCA}}^T x_i = G_b^T \Sigma^{-\frac{1}{2}} U^T x_i = G_b^T \Sigma^{-\frac{1}{2}} \begin{pmatrix} u_1^T x_i \\ u_2^T x_i \\ \cdots \\ u_r^T x_i \end{pmatrix}.$$

Therefore, $G_b^T \Sigma^{-\frac{1}{2}}$ represents net effects of LDA. The r-th column of $G_b^T \Sigma^{-\frac{1}{2}}$ incorporates the r-th PCA dimension $u_r^T x_i$. Therefore we define

Definition. LDA factor is the net effect due to LDA on incorporating r-th PCA dimension, defined to be

$$f_r = \sum_{i=1}^{k} \left[(G_b^T \Sigma^{-\frac{1}{2}})_{ir} \right]^2. \tag{23}$$

We want to show that the overfit of LDA is related to a large value of f_r for these insignificant PCA dimensions.

Using Theorem 4, and the S_t-orthonormal LDA of Eq.(22), we construct

$$\begin{aligned} \mathrm{Tr} \frac{G^T S_b G}{G^T S_t G} &= \mathrm{Tr} \frac{(G_b^T \Sigma^{-\frac{1}{2}} U^T) S_b (U \Sigma^{-\frac{1}{2}} G_b)}{(G_b^T \Sigma^{-\frac{1}{2}} U^T) S_t (U \Sigma^{-\frac{1}{2}} G_b)} \\ &= \mathrm{Tr} \frac{G_b^T (\Sigma^{-\frac{1}{2}} U^T S_b U \Sigma^{-\frac{1}{2}}) G_b}{G_b^T G_b}. \end{aligned} \tag{24}$$

Therefore, columns of $G_b = (g_1^b, \cdots, g_k^b)$ are given by the eigenvector of

$$(\Sigma^{-\frac{1}{2}} U^T S_b U \Sigma^{-\frac{1}{2}}) g_k^b = \xi_k^b g_k^b$$

Because $\Sigma^{-\frac{1}{2}} U^T S_b U \Sigma^{-\frac{1}{2}}$ is positive definite symmetric, all eigenvalues $\xi^b \geq 0$, and the eigenvectors $\{g_k^b\}$ are mutually orthogonal, we have $G_b^T G_b = I$. The magnitude of G_b will be show in experiments.

Experiments

MNIST Hand-written Digit Dataset

The MNIST hand-written digits dataset consists of 60,000 training and 10,000 test digits (LeCun et al. 1998), which can be downloaded from "http://yann.lecun.com/exdb/mnist/" with 10 classes. Each image is centered on a 28x28 grid. We randomly pick 20 images for each class, for a total of 200 images for experiment.

Figure 1 shows the LDA classification accuracy (5-fold cross validation) for 5 different LDA solutions at PCA subspace varies from 150 to 10. (Because we do 5-fold cross validation, only 80% data are available for training which limits the maximum PCA-dim to 200*0.8=160.) It is clear that the performance of all LDA solutions are poor near PCA-dim=100 - 150. In Figures 2 and 3, the LDA overfit results are shown. Shown are G_b the LDA Factor defined in Eq.(23) without the eigenvalues $\Sigma^{-\frac{1}{2}}$ and the PCA eigenvalues. From Figure 2, it is clearly the LDA Factor are overwhelmed by the insignificant (small eigenvalue) PCA dimensions (peak near PCA-dim= 140-150). This demonstrates that LDA is clearly overfit at PCA dimension = 150.

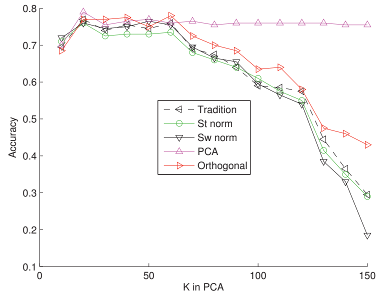

Figure 1: LDA results on MNIST dataset.

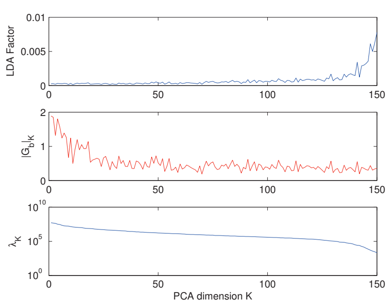

Figure 2: LDA Overfit results on MNIST dataset at PCA-Dim=150.

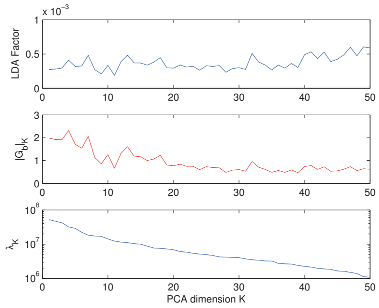

Figure 3: LDA Overfit results on MNIST dataset at PCA-Dim=50.

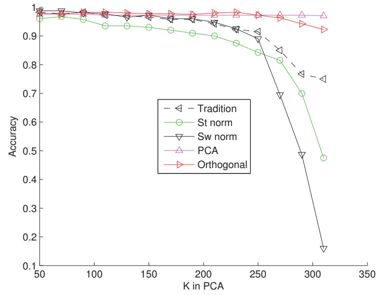

Figure 4: LDA results on ATNT dataset.

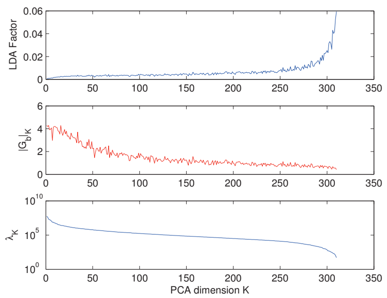

Figure 5: LDA Overfit results on ATNT dataset at PCA-Dim=310.

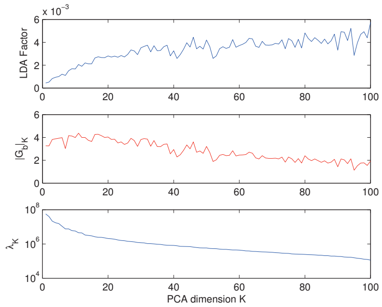

Figure 6: LDA Overfit results on ATNT dataset at PCA-Dim=100.

As the PCA-dimension is reduced towards 20, performance of all LDA solutions increase steadily, because the overfit problem is gradually reduced. From Figure 3, the LDA Factor is no longer dominated by the insignificant PCA dimensions. The best results are achieved at PCA-dim = 20 = 2C (C=10 is the number of classes).

AT&T Face Image Dataset

The AT&T database, which can be downloaded from "http://www.cl.cam.ac.uk/research/dtg/attarchive/", is widely used in computer vision as a benchmark for classification. There are total 400 images for 40 persons. Each image has a size 112×92, which is reduced to size 56×46 before analysis. Figure 4 shows the LDA classification accuracy (5-fold cross validation) for 5 different LDA solutions at PCA subspace varies from 320 to 50. (Because we do 5-fold cross validation, only 80% data are available for training which limits the maximum PCA-dim to $400 \times 0.8 = 320$.) It is clear that the performance of all LDA solutions are poor near PCA-dim=250-320. In Figure 5 the LDA overfit results are shown at PCA dimension = 320. Clearly LDA is overfitting at PCA dimension = 320. This explains the poor performance of LDA solutions. As the PCA-dimension is reduced from 320 towards 50, performances of all LDA solutions increase steadily, because the overfit problem is gradually reduced. At PCA dimension = 100, as shown in Figure 6, the LDA Factor is no longer dominated by insignificant PCA dimensions. This trend is consistent. The best results are achieved at PCA-dim = 50 = 1.2C (C=40 is the number of classes).

YaleB Dataset

The Yale database B (Georghiades, Belhumeur, and Kriegman 2001) contains images of 31 persons (This is a standard subset of of the original 38 persons, but some images of 7 persons were corrupted). We randomly select 10 illumination conditions for a total 310 images. The size of each original image is 192×168, which is reduced to 48×42 for our experiments. Figure 7 shows the LDA classification accuracy (5-fold cross validation) for 5 different LDA solutions at PCA subspace at PCA-dim= $31, 61, 91, 121, 181, 211, 241$. (Because we do 5-fold cross validation, only 80% data are available for training which limits the maximum PCA-dim to $310 \times 0.8 = 248$.) It is clear that the performance of all LDA solutions are poor near PCA-dim=$181 \sim 241$. In Figure 8 the LDA overfit results are shown at PCA dimension = 230. Clearly LDA is overfitting. This explains the poor performance of LDA solutions. As the PCA-dimension is reduced from 320 towards 150, performances of all LDA solutions increase steadily, because

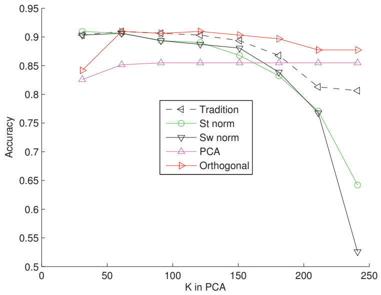

Figure 7: LDA results on YaleB dataset.

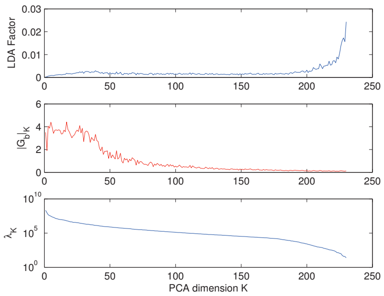

Figure 8: LDA Overfit results on YaleB dataset at PCA-Dim=230.

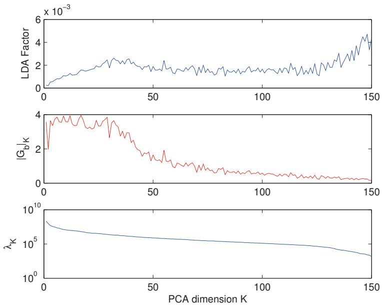

Figure 9: LDA Overfit results on YaleB dataset at PCA-Dim=150.

the overfit problem is gradually reduced. At PCA dimension = 150, as shown in Figure 9, the LDA Factor is less dominated by insignificant PCA dimensions. The best results are achieved at PCA-dim = 31-61 = (1 - 2)C (C=31 is the number of classes).

Which LDA solution is the best? From the LDA performance of Figures 1, 4, and 7, it seems that G_w consistently performs the best, G_0 consistently performs well, G_t performs well 2 out of 3 datasets.

Implication. Our results show that in application of LDA, due to overfitting, we should first perform PCA to reduce data to a suitable subspace and do LDA in the subspace. The exact PCA subspace dimension L should be chosen as small as possible by cross-validation. Our experiments suggest

$$\text{PCA dimension} \quad \approx 2C,$$

where C is the number of classes. In early work (Belhumeur, Hespanha, and Kriengman 1997), PCA dimension is mostly set at rank$(S_w) - C$. This is far larger than $2C$. As shown in Figures 2, 5, and 8, at PCA-dim = rank$(S_w) - C$, overfitting often occur.

Conclusions

In this paper, we have clarified a large number of issues regarding to LDA on whether the solution is (1) unique or not, (2) global not local, (3) rotational invariant or generic invariant. We also show the solutions of S_t-orthonormal LDA and S_w-orthonormal LDA are also the solutions of all four possible LDA objective functions. We systematically analyze the overfitting problem of LDA and show that LDA often incorporate insignificant PCA dimensions. We carry out extensive experiments on 3 widely used datasets to demonstrate the overfitting problem. Overall, our analysis provides a unified and systematic analysis of the LDA classification methodology.

Acknowledgments This research was supported by NSF-CCF 0830780, NSF-CCF 0917274, NSF-DMS 0915228.

References

Belhumeur, P.; Hespanha, J.; and Kriengman, D. 1997. Eigenfaces vs fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence* 19(7):711–720.

Chen, L.; Liao, H.; Lin, J.; Ko, M.; and Yu, G. 2000. A new lda-based face recognition system which can solve the small sample size problem. *Pattern Recognition* 33:17131726.

Duda, R. O.; Hart, P. E.; and Stork, D. G. 2000. *Pattern Classification, 2nd ed.* Wiley.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7.

Fukunaga, K. 1990. Introduction to statistical pattern recognition. *Academic Press Professional, 2nd edition*.

Georghiades, A.; Belhumeur, P.; and Kriegman, D. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence* 23(6):643–660.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2001. *Elements of Statistical Learning*. Springer Verlag.

Jin, Z.; Yang, J.; Hu, Z.; and Lou, Z. 2001. Face recognition based on the uncorrelated discriminant transformation. *Pattern Recognition* 34:14051416.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86(11):2278–2324.

Luo, D.; Ding, C.; and Huang, H. 2009. Symmetric two dimensional linear discriminant analysis (2DLDA). *IEEE Conf. Computer Vision and Pattern Recognition*.

Park, H., and Howland, P. 2004. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE. Trans. on Pattern Analysis and Machine Intelligence* 26:995 – 1006.

Turk, M. A., and Pentland, A. P. 1991. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 586–591.

Ye, J., and Xiong, T. 2006. Computational and theoretical analysis of null space and orthogonal linear discriminant analysis. *J. Machine Learning Research* 7:1183–1204.

Ye, J.; Janardan, R.; Li, Q.; et al. 2004. Two-dimensional linear discriminant analysis. *Advances in Neural Information Processing Systems* 17:1569–1576.