Linear Discriminant Analysis for Data with Subcluster Structure

Haesun Park, Jaegul Choo
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
{hpark, joyfull}@cc.gatech.edu

Barry L. Drake
SEAL/AMDD
Georgia Tech Research Institute
Smyrna, GA 30080, USA
barry.drake@gtri.gatech.edu

Jinwoo Kang
Center for Signal & Image Processing
Georgia Institute of Technology
Atlanta, GA 30332, USA
jinu@ece.gatech.edu

Abstract

Linear discriminant analysis (LDA) is a widely-used feature extraction method in classification. However, the original LDA has limitations due to the assumption of a unimodal structure for each cluster, which is satisfied in many applications such as facial image data when variations such as angle and illumination can significantly influence the images of the same person. In this paper, we propose a novel method, hierarchical LDA(h-LDA), which takes into account hierarchical subcluster structures in the data sets. Our experiments show that regularized h-LDA produces better accuracy than LDA, PCA, and tensorFaces.

1 Introduction

Linear discriminant analysis (LDA) has been one of the most widely-used dimension reduction methods for classification problems. However, LDA assumes that each class is modeled as a unimodal Gaussian that can be fully described only with the first and the second order statistics, i.e., mean and covariance. In reality, there is no guarantee that the data conforms to such assumptions. When the data is significantly dependent on other factors than the cluster label of interest, the data corresponding to a particular label cannot be simply modeled as a unimodal Gaussian with a single mean and a covariance.

In order to circumvent such problems, one may apply several variants such as nonparametric discriminant analysis (NDA) [2], subclass discriminant analysis (SDA) [6], or regularized LDA [1]. NDA uses a nonparametric version of the between-cluster scatter matrix to relax the unimodal Gaussian assumption. SDA applies the multi-modal Gaussian model directly by replacing each cluster centroid with subcluster centroids in the definition of the between-cluster scatter matrix. Although regularized LDA was originally introduced in order to avoid singularity of the within-cluster scatter matrix for undersampled cases, regularization also plays a role of controlling overfitting by adding a small identity matrix to the within-cluster scatter matrix.

On the other hand, we can utilize the available information other than the cluster label. In face recognition, various information such as angles, illuminations, and/or pose of images may be incorporated so that recognition performance can be enhanced [5].

In this paper, we propose a novel method called hierarchical LDA (h-LDA) that formulates a new feature extraction method based on the hierarchical structure with depth-2 in the data. In h-LDA, clusters are considered to have several subclusters determined by other factors than the cluster label of interest, and such subclusters do not need to be necessarily gathered closely as in the unimodal Gaussian model. Based on this motivation, h-LDA maintains the control against overfitting issues that LDA has.

The rest of this paper is organized as follows. In Section 2, the classical LDA is briefly reviewed, and our new h-LDA is presented in Section 3. In section 4, a regularized version of h-LDA is introduced. Experimental results are reported in Section 6, and finally conclusions are given in Section 7.

2 Linear Discriminant Analysis

In LDA, an optimal dimension-reduced representation of data is obtained by a linear transformation that maximizes the *conceptual* ratio of the between-cluster scatter (variance) versus the within-cluster scatter of the data. In this section, we present an overview of the basic ideas.

Given a data matrix $A \in \mathbb{R}^{m \times n}$, where n columns a_i, $i = 1, \ldots, n$, of A represent n data items in an m dimensional space, let us assume that it is partitioned

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

into p clusters as

$$A = [a_1\, a_2 \cdots a_n] = [A_1 \quad A_2 \quad \cdots \quad A_p],$$

where $A_i \in \mathbb{R}^{m\times n_i}$ and $\sum_{i=1}^{p} n_i = n$.

Let $\mathcal{N}_i$ denote the set of column indices that belong to cluster i, n_i the size of $\mathcal{N}_i$, a_k the data point represented in the k-th column vector of A, $c^{(i)}$ the centroid of the i-th cluster, and c the global centroid. In face recognition, A_i corresponds to the i-th person's image data set. The scatter matrix within the i-th cluster $S_w^{(i)}$, the within-cluster scatter matrix S_w, the between-cluster scatter matrix S_b, and the total (or mixture) scatter matrix S_t, are defined, respectively, as

$$\begin{aligned}
S_w^{(i)} &= \sum_{k\in\mathcal{N}_i} (a_k - c^{(i)})(a_k - c^{(i)})^T, \\
S_w &= \sum_{i=1}^{p} S_w^{(i)}, \\
S_b &= \sum_{i=1}^{p}\sum_{k\in\mathcal{N}_i} (c^{(i)} - c)(c^{(i)} - c)^T, \text{ and} \qquad (1) \\
S_t &= \sum_{i=1}^{p}\sum_{k\in\mathcal{N}_i} (a_k - c)(a_k - c)^T \\
&= S_w + S_b. \qquad (2)
\end{aligned}$$

LDA finds the optimal linear transformation matrix,

$$G^T : x \in \mathbb{R}^{m\times 1} \rightarrow y \in \mathbb{R}^{l\times 1},$$

that maximizes

$$J_1(G) = trace((G^T S_w G)^{-1}(G^T S_b G)), \qquad (3)$$

which is the ratio of the within-cluster radius and the between-cluster distance in the reduced dimensional space.

3 Hierarchical LDA (h-LDA)

In many applications, the structure of the data cannot be simply explained by the unimodal Gaussian model of LDA. Relaxing such a simplified assumption, hierarchical LDA (h-LDA) assumes that the data in cluster i, A_i, can be further clustered into q_i subclusters as

$$A_i = [A_{i1} \quad A_{i2} \quad \cdots \quad A_{iq_i}],$$

where $A_{ij} \in \mathbb{R}^{m\times n_{ij}}$ and $\sum_{j=1}^{q_i} n_{ij} = n_i$.

Let $\mathcal{N}_{ij}$ denote the set of column indices that belong to the subcluster j in cluster i, n_{ij} the size of $\mathcal{N}_{ij}$ and $c^{(ij)}$ the centroid of each subcluster. For instance of facial image data, the set of images of a specific person can be further clustered according to angles of view, or illumination conditions. Then, we can define the scatter matrix within subcluster j in cluster i, $S_{w_s}^{(ij)}$, their sum in cluster i, $S_{w_s}^{(i)}$, and the scatter matrix between subclusters in cluster i, $S_{b_s}^{(i)}$, repsectively, as

$$\begin{aligned}
S_{w_s}^{(ij)} &= \sum_{k\in\mathcal{N}_{ij}} (a_k - c^{(ij)})(a_k - c^{(ij)})^T, \\
S_{w_s}^{(i)} &= \sum_{j=1}^{q_i} S_{w_s}^{(ij)}, \text{ and} \\
S_{b_s}^{(i)} &= \sum_{j=1}^{q_i}\sum_{k\in\mathcal{N}_{ij}} (c^{(ij)} - c^{(i)})(c^{(ij)} - c^{(i)})^T.
\end{aligned}$$

Then, the within-subcluster scatter matrix S_{w_s} and the between-subcluster scatter matrix S_{b_s} are defined respectively as

$$S_{w_s} = \sum_{i=1}^{p} S_{w_s}^{(i)} = \sum_{i=1}^{p}\sum_{j=1}^{q_i} S_{w_s}^{(ij)} \text{ and } S_{b_s} = \sum_{i=1}^{p} S_{b_s}^{(i)}.$$

From the identity

$$a_k - c = (a_k - c^{(ij)}) + (c^{(ij)} - c^{(i)}) + (c^{(i)} - c),$$

it can be proved that

$$S_t = S_{w_s} + S_{b_s} + S_b \qquad (4)$$

where the between-cluster scatter matrix S_b is defined as in Eq. (1). Comparing Eq. (2) and Eq. (4), the within-cluster scatter matrix S_w in classical LDA is equivalent to the sum of the within-subcluster scatter matrix S_{w_s} and the between-subcluster scatter matrix S_{b_s} as

$$S_w = S_{w_s} + S_{b_s}. \qquad (5)$$

Now we propose a new within-cluster scatter matrix S_w^h, which is a convex combination of S_{w_s} and S_{b_s} as

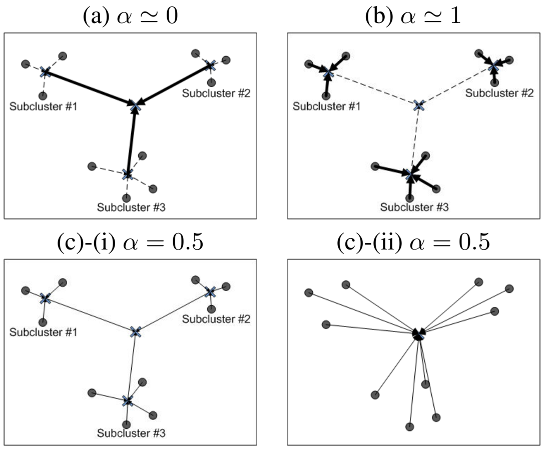
$$S_w^h = \alpha S_{w_s} + (1-\alpha) S_{b_s}, \quad 0 \le \alpha \le 1, \qquad (6)$$

where α determines relative weights between S_{w_s} and S_{b_s}. By replacing S_w with a newly-defined S_w^h, h-LDA finds the solution that maximizes the new criterion

$$L_1(G) = trace((G^T S_w^h G)^{-1}(G^T S_b G)). \qquad (7)$$

Consider the following three cases: $\alpha \simeq 0$, $\alpha \simeq 1$, and $\alpha = 0.5$. When $\alpha \simeq 0$ (see Figure 1(a)), the within-subcluster scatter matrix S_{w_s} is disregarded and the between-subcluster scatter matrix S_{b_s} is emphasized, which can be considered as the original LDA applied after every data point is relocated to its corresponding

Figure 1. Example of h-LDA and the parameter α. All data points in each figure belong to one cluster.


(a) $\alpha \simeq 0$

(b) $\alpha \simeq 1$

(c)-(i) $\alpha = 0.5$

(c)-(ii) $\alpha = 0.5$

subcluster centroid. When $\alpha \simeq 1$ (see Figure 1(b)), h-LDA minimizes only the within-subcluster radii, disregarding the distances between subclusters within each cluster. When $\alpha = 0.5$, the within-subcluster scatter matrix S_{w_s} and the between-subcluster scatter matrix S_{b_s} are equally weighted so that h-LDA becomes equivalent to LDA by Eq. (5), which shows the equivalence of the within-cluster scatter matrices between Figure 1(c)-(i) and 1(c)-(ii). Hence, h-LDA can be viewed as a generalization of LDA, and the parameter α can be chosen by parameter optimization schemes such as cross-validation in order to attain maximum classification performance. Considering the motivation of h-LDA, attention should be paid to the case of $0.5 < \alpha \simeq 1$ since this can mitigate the unimodal Gaussian assumption weakness of the classical LDA, which can produce a transformation that projects the points in one cluster onto essentially one point in the reduced dimensional space.

4 h-LDA with Regularization (h-RLDA)

Both h-LDA and LDA take into account only the estimates of the first and the second order statistics of the data, and the quality of these estimates relies on the number of data items. As the number of data items increases, the estimators have smaller variances or deviations from the true underlying statistics. In this sense, the potential problem in h-LDA is that the estimates of S_{w_s} and S_{b_s} may not be as confident as that of S_b due to further splitting of the data into subclusters. In order to compensate such a problem, we suggest introducing a regularization term to S_w^h in Eqs. (6) and in (7) as

$$S_w^h + \gamma I = \alpha S_{w_s} + (1-\alpha) S_{b_s} + \gamma I, \qquad (8)$$

$$L_1(G, \gamma) = trace((G^T(S_w^h + \gamma I)G)^{-1}(G^T S_b G)). \qquad (9)$$

which enables us to avoid the complete dependency on the small sample size of subclusters by controlling the value of $\gamma > 0$. The criterion to maximize Eq. (9) is a regularized form of h-LDA, which we call h-RLDA.

5 Experiments

5.1 Experimental Setup

Our implementation for h-RLDA is based on the generalized SVD framework with efficient QR decomposition by using Eq. (8) as the within-cluster scatter matrix. For more details, see Algorithm 3. LDA/QR-regGSVD in [4]. We have applied h-RLDA to a face recognition problem using Shimon Edelman's face database[1]. This data set contains 27 persons' images where the images vary depending on several factors such as angles of view, illuminations, and facial expressions. We resized the original 512×352 pixel images to 64×44 pixel images, and converted it into a 1-dimensional array by stacking up the columns. Each image was given a set of labels that contain person id, angle of view, illumination, and facial expression. Throughout our experiments, the person id is our target label to estimate, and each of other labels was used as the factor that forms subclasses within each person's images. The proposed h-RLDA is compared to PCA and LDA/GSVD [3], and tensorFaces [5]. The first two methods use only person id, but tensorFaces utilizes other subcluster information than person id as in h-RLDA.

The details of our experimental procedure are as follows: First, the training samples used to build up the dimension reducing matrix are determined depending on particular label set. For instance, among all the available angles in the data, $\{0°, \pm 17°, \pm 34°\}$, the images from the selected angles $\{0°, \pm 34°\}$ are used as a training set and the rest are reserved as a test set to measure the recognition accuracy. Next, the dimension reducing matrix made with such training samples is applied to all of the data, and we perform K-nearest neighbor classification on the test set, where $K = 1$. As a performance measure, we present recognition accuracies.

For h-RLDA, the parameters α $(0 \leq \alpha \leq 1)$ and γ in Eq. (9) were optimized using k-fold cross-validation with step size of 0.1 for α, and $2^{-i}, i = 1, 2, \cdots, 30$ for γ, respectively. In the case of multiple pairs of values of α and γ produced the best cross-validation accuracy, we chose the smallest value for γ and then the largest for α.

[1] ftp://ftp.wisdom.weizmann.ac.il/pub/facebase

Table 1. Comparison data of recognition accuracies(%)

	Training/Test Data	PCA	LDA	tensorFaces	h-RLDA
Data 1	Training : **3 angles of view**$(0^\circ, \pm 34^\circ)$, 4 illuminations, 3 facial expressions Test : **2 other angles of view**$(\pm 17^\circ)$, 4 illuminations, 3 facial expressions	88.73%	96.24%	85.92%	**98.59%**
Data 2	Training : 5 angles of view$(0^\circ, \pm 17^\circ \pm 34^\circ)$, **2 illuminations**, 3 facial expressions Test : 5 angles of view$(0^\circ, \pm 17^\circ \pm 34^\circ)$, **2 other illuminations**, 3 facial expressions	86.47%	97.58%	90.94%	**99.82%**
Data 3	Training : 1 angles of view(0°), 4 illuminations, **2 facial expressions** Test : 1 angles of view(0°), 4 illuminations, **1 other facial expression**	87.15%	98.34%	89.42%	**100%**

Tucker decomposition in Matlab Tensor Toolbox[2] was used for tensorFaces algorithm.

5.2 Results

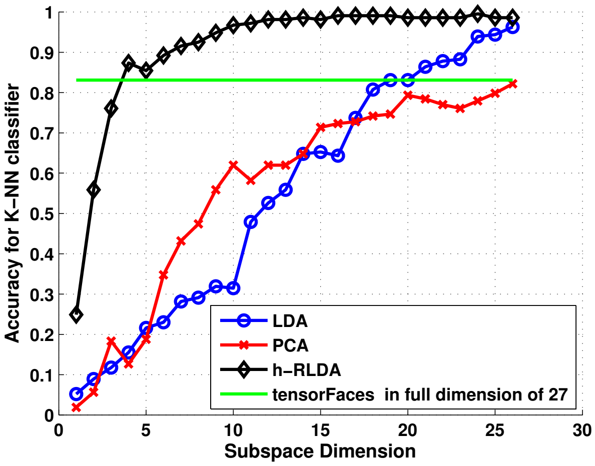
For three different training/test sets, the recognition accuracies of PCA, LDA, tensorFaces, and h-RLDA are shown in Table 1. In all cases, h-RLDA shows consistently better performances than any other methods. Another interesting observation is that the tensorFaces did not outperform PCA as clearly as was reported in [5], and it did not perform as well as LDA although it utilized more information than LDA.

Figure 2 shows the recognition accuracies versus subspace dimensions using Data 1 shown in Table 1. LDA and h-RLDA allow the maximum reduced dimension of $p-1$ where p is the number of different people, and PCA allows the maximum reduced dimension which is the same as the total number of training images unless it exceeds the original dimension. This experiment shows the recognition results when we use only a subset of fully extracted features. Reduced space of dimension d was obtained from the d leading eigenvectors and generalized singular vectors of PCA and LDA/h-RLDA respectively. From Figure 2, we can observe that h-RLDA reaches its maximum performance very fast even with only about 10 features whereas LDA requires almost the full reduced dimensional space of rank $p-1$ to produce its best performance, which means the extracted feature quality of h-RLDA can be much better than that of LDA.

6 Conclusions

To remedy the drawback of LDA that assumes a unimodal Gaussian model in each cluster, hierarchical LDA was introduced by enhancing the within-cluster scatter matrices using additional information available from the data. Combined with regularization and recently proposed regularized LDA algorithm, the idea of hierarchical LDA showed superior performances over other methods such as PCA, LDA, and tensorFaces.

Figure 2. Recognition accuracies versus subspace dimensionality of Data 1 in Table 1



References

[1] J. Friedman. Regularized Discriminant Analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[2] K. Fukunaga and J. M. Mantock. Nonparametric discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:671–678, 1983.

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

[4] H. Park, B. Drake, S. Lee, and C. Park. Fast Linear Discriminant Analysis using QR Decomposition and Regularization. *Technical Report GT-CSE-07-21*, 2007.

[5] M. A. O. Vasilescu and D. Terzopoulos. Multilinear image analysis for facial recognition. In *Proceedings of International Conference on Pattern Recognition (ICPR 2002)*, pages 511–514, Quebec City, Canada, August.

[6] M. Zhu and A. Martinez. Subclass Discriminant Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1274–1286, 2006.

[2] http://csmr.ca.sandia.gov/˜tgkolda/TensorToolbox/