1 Introduction

- Many applications such as face recognition, text mining, and microarray gene expression data analysis, involve data with a large number of features. Analysis of such data is challenging due to the curse-of-dimensionality, which states that an enormous number of samples are required to perform accurate predictions on problems with high dimensionality.

- Dimensionality reduction, which extracts a small number of features by removing irrelevant, redundant, and noisy information, can be an effective solution. The commonly used dimensionality reduction methods include supervised approaches such as linear discriminant analysis (LDA), unsupervised ones such as principal component analysis (PCA).

- When the class label information is available, supervised approaches, such as LDA, are usually more effective than unsupervised ones such as PCA for classification.

- Linear discriminant analysis (LDA) is a classical statistical approach for supervised dimensionality reduction and classification. LDA computes an optimal transformation (projection) by minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination. The optimal transformation in LDA can be readily computed by applying an eigendecomposition on the so-called scatter matrices. It has been used widely in many applications involving high-dimensional data.

2 Overview of Linear Discriminant Analysis

- We are given a data set that consists of $n$ samples $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ denotes the $d$-dimensional input, $y_i \in \{1, 2, \cdots, k\}$ denotes the corresponding class label, $n$ is the sample size, and $k$ is the number of classes. Let $X = [x_1, x_2, \cdots, x_n]$ be the data matrix and $X_j \in \mathbb{R}^{d \times n_j}$ be the data matrix of the $j$-th class, where $n_j$ is the sample size of the $j$-th class, and $\sum_{j=1}^k n_j = n$. Classical LDA computes a linear transformation $G \in \mathbb{R}^{d \times \ell}$ that maps $x_i$ in the $d$-dimensional space to a vector $x_i^\ell$ in the $\ell$-dimensional space as follows:
  
  $$x_i \in \mathbb{R}^d \rightarrow x_i^\ell = G^T x_i \in \mathbb{R}^\ell, \quad \ell < d.$$  

- In LDA, three scatter matrices, called within-class, between-class and total scatter matrices
are defined as follows:

\[ S_w = \frac{1}{n} \sum_{j=1}^{k} \sum_{x \in X_j} (x - c^{(j)})(x - c^{(j)})^T, \]  

(1)

\[ S_b = \frac{1}{n} \sum_{j=1}^{k} n_j (c^{(j)} - c)(c^{(j)} - c)^T, \]  

(2)

\[ S_t = \frac{1}{n} \sum_{i=1}^{n} (x_i - c)(x_i - c)^T, \]  

(3)

where \( c^{(j)} \) is the centroid of the \( j \)-th class, and \( c \) is the global centroid. It can be verified from the definitions that \( S_t = S_b + S_w \). Define three matrices \( H_w, H_b, \) and \( H_t \) as follows:

\[ H_w = \frac{1}{\sqrt{n}} [X_1 - c^{(1)}(e^{(1)})^T, \ldots, X_k - c^{(k)}(e^{(k)})^T], \]  

(4)

\[ H_b = \frac{1}{\sqrt{n}} [\sqrt{n_1}(c^{(1)} - c), \ldots, \sqrt{n_k}(c^{(k)} - c)], \]  

(5)

\[ H_t = \frac{1}{\sqrt{n}} (X - ce^T), \]  

(6)

where \( e^{(j)} \) and \( e \) are vectors of all ones of length \( n_j \) and \( n \), respectively. Then the three scatter matrices, defined in Eqs. (1)-(3), can be expressed as

\[ S_w = H_w H_w^T, \quad S_b = H_b H_b^T, \quad S_t = H_t H_t^T. \]  

(7)

• It follows from the properties of matrix trace that

\[ \text{trace}(S_w) = \frac{1}{n} \sum_{j=1}^{k} \sum_{x \in X_j} \|x - c^{(j)}\|^2, \]  

(8)

\[ \text{trace}(S_b) = \frac{1}{n} \sum_{j=1}^{k} n_j \|c^{(j)} - c\|^2. \]  

(9)

Thus \( \text{trace}(S_w) \) measures the distance between the data points to their corresponding class centroid, and \( \text{trace}(S_b) \) captures the distance between the class centroids to the global centroid.

• In the lower-dimensional space resulting from the linear transformation \( G \), the scatter matrices become

\[ S_w^L = G^T S_w G, \quad S_b^L = G^T S_b G, \quad S_t^L = G^T S_t G. \]  

(10)

• An optimal transformation \( G \) would maximize \( \text{trace}(S_b^L) \) and minimize \( \text{trace}(S_w^L) \) simultaneously, which is equivalent to maximizing \( \text{trace}(S_b^L) \) and minimizing \( \text{trace}(S_t^L) \) simultaneously, since \( S_t^L = S_w^L + S_b^L \). The optimal transformation, \( G^{LDA} \), of LDA is computed by solving the following optimization problem:

\[ G^{LDA} = \arg \max_G \left\{ \text{trace} \left( S_b^L (S_t^L)^{-1} \right) \right\}. \]  

(11)
It is known that the optimal solution to the optimization problem in Eq. (11) can be obtained by solving the following generalized eigenvalue problem:

\[ S_b x = \lambda S_t x, \quad (12) \]

More specifically, the eigenvectors corresponding to the \( k - 1 \) largest eigenvalues form columns of \( G^{LDA} \). When \( S_t \) is nonsingular, it reduces to the following regular eigenvalue problem:

\[ S_t^{-1} S_b x = \lambda x. \quad (13) \]

When \( S_t \) is singular, the classical LDA formulation discussed above can not be applied directly. This is known as the singularity or undersampled problem in LDA.

### 3 Generalizations of LDA

- A common way to deal with the singularity problem is to apply an intermediate dimensionality reduction stage, such as PCA, to reduce the data dimensionality before classical LDA is applied. The algorithm is known as PCA+LDA, or subspace LDA. In this two-stage PCA+LDA algorithm, the discriminant stage is preceded by a dimensionality reduction stage using PCA. The dimensionality, \( p \), of the subspace transformed by PCA is chosen such that the “reduced” total scatter matrix in this subspace is nonsingular, so that classical LDA can be applied. The optimal value of \( p \) is commonly estimated through cross-validation.

- Regularization techniques can also be applied to deal with the singularity problem of LDA. The algorithm is known as regularized LDA, or RLDA in short. The key idea is to add a constant \( \mu > 0 \) to the diagonal elements of \( S_t \) as \( S_t + \mu I_d \), where \( I_d \) is the identity matrix of size \( d \). It is easy to verify that \( S_t + \mu I_d \) is positive definite, hence nonsingular. Cross-validation is commonly applied to estimate the optimal value of \( \mu \). Note that regularization is also the key to Support Vector Machines (SVM).

- The null space LDA (NLDA) was proposed to overcome the singularity problem, where the between-class distance is maximized in the null space of the within-class scatter matrix. The singularity problem is thus avoided implicitly. The efficiency of the algorithm can be improved by first removing the null space of the total scatter matrix. It is based on the observation that the null space of the total scatter matrix is the intersection of the null spaces of the between-class and within-class scatter matrices.