CSI 436/536
Introduction to Machine Learning

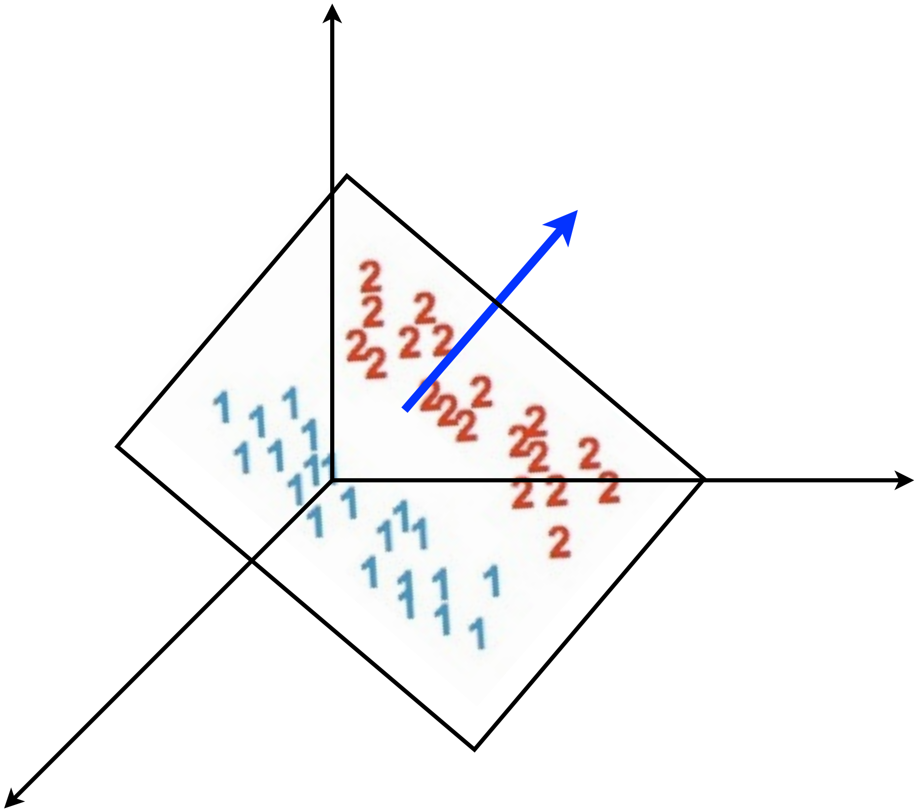
classification: LDA

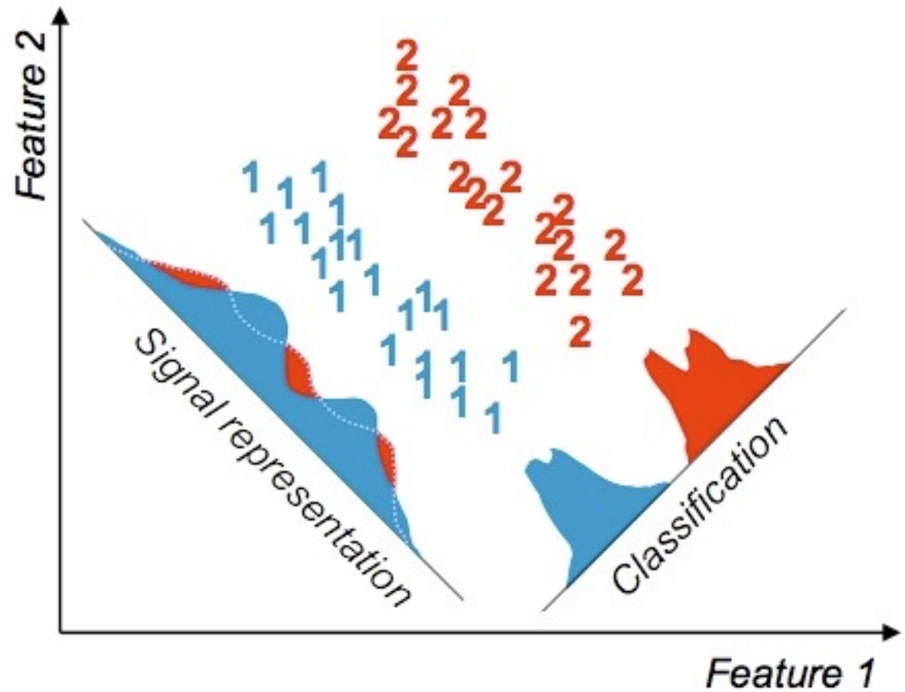
Professor Siwei Lyu
Computer Science
University at Albany, State University of New York

Dimension reduction for classification

- In classification problems, we usually do not map raw data x to class labels y, instead, we use transforms of the raw data $\phi(x) : \mathscr{R}^d \mapsto \mathscr{R}^m$ to build classifiers, this is called (classification) ***features***
- Features are very important for effective classification systems and there are two types of get features
  - Lean features ($m \ll d$): reduce the dimensionality of raw data to smaller number of features
    - Reduce the dimension of input data, keeping important information for classification
  - Rich features ($m \gg d$): increase the dimensionality of input data
    - Data are more likely to separate in higher dim space

Is PCA always useful for classification?



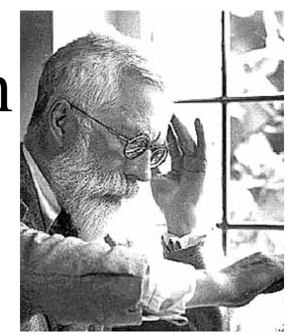


PCA removes the irrelevant dimension that does not affect classification

PCA removes "good" dimension that is important for classification

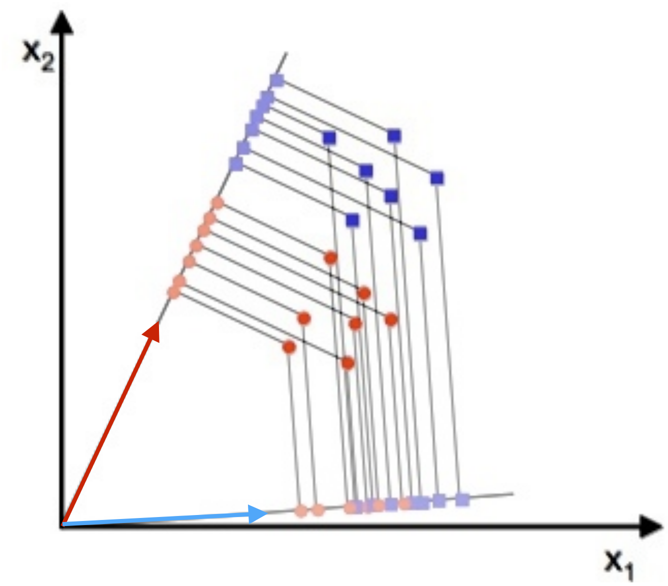
Dimension reduction for classification

- PCA is designed for signal representation but there is no class difference in the definition of PCA
- PCA feature may not be relevant for classification: those discarded PCs may contain important information for classification, even though they have little overall information contributed to represent all the data
- We introduce a new method: *linear discriminant analysis* (LDA), also known as Fisher linear discriminant analysis
  - find low dimensional features for class specific dataset
  - dimensionality is determined by the number of classes
  - LDA is solved as a generalized eigenvalue problem



LDA for binary classification

- we assume data are in two classes with class labels as {-1,+1} and use data matrices $X = [X_+ \ X_-]$, and $N = N_+ + N_-$
- LDA finds a projection direction $v : v^T v = 1$, such that the projection of data of two classes, $v^T X_+$ and $v^T X_-$,
  - The distance between the means of the two classes (**between class scattering**) projections is large
  - The spread of each class (**within class scattering**) is small
  - eg., the red vector in the figure is a better projection direction than the blue vector



Notations

- mean of positive data $\mu_+ = \frac{1}{n_+}\sum_{i=1}^{n_+} x_i^+ = \frac{1}{n_+}X^+1_{n_+}$

  mean of negative data $\mu_- = \frac{1}{n_-}\sum_{i=1}^{n_-} x_i^- = \frac{1}{n_-}X^-1_{n_-}$,

  mean of all data $\mu = \frac{1}{n}\sum_i x_i = \frac{n_+}{n}\mu_+ + \frac{n_-}{n}\mu_-$

- covariance matrix of positive data

  $S_+ = \frac{1}{n_+}\sum_i (x_i^+ - \mu_+)(x_i^+ - \mu_+)^T = \frac{1}{n_+}\sum_i x_i^+ x_i^{+T} - \mu_+\mu_+^T$covariance

  matrix of negative data

  $S_- = \frac{1}{n_-}\sum_i (x_i^- - \mu_-)(x_i^- - \mu_-)^T = = \frac{1}{n_-}\sum_i x_i^- x_i^{-T} - \mu_-\mu_-^T$

  Covariance of all data

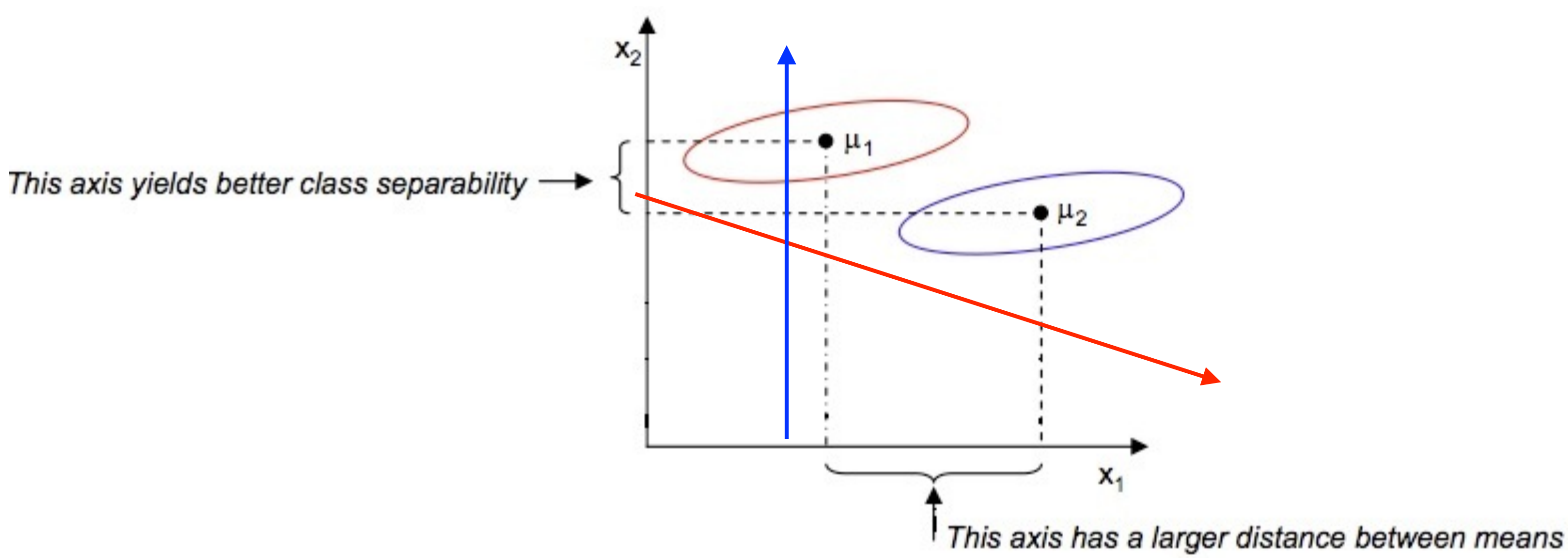
  $S = \frac{1}{n}\sum_i (x_i - \mu)(x_i - \mu)^T = \frac{1}{n}\sum_i x_i x_i^T - \mu\mu^T$

LDA for binary classification

- Between class scattering: squared difference of means on the projection v$(v^T(\mu_+ - \mu_-))^2 = v^T(\mu_+ - \mu_-)(\mu_+ - \mu_-)^T v$

- Within class scattering: variance of the projection of each class: $\frac{1}{n_+}\sum_{i=1}^{n_+}[v^T(x_i^+ - \mu_+)]^2 + \frac{1}{n_-}\sum_{i=1}^{n_+}[v^T(x_i^- - \mu_-)]^2$

- Why we need to consider both types of scatterings



Scattering matrices

- within class scattering matrix:

$$S_w = \frac{n_+}{n} S_+ + \frac{n_-}{n} S_- = \frac{1}{n} \sum_i x_i x_i^T - \frac{n_+}{n} \mu_+ \mu_+^T - \frac{n_-}{n} \mu_- \mu_-^T$$

  - within class scattering of projected data $v^T S_w v$

- between class scattering matrix:

$$S_b = \frac{n_+ n_-}{n^2} (\mu_+ - \mu_-)(\mu_+ - \mu_-)^T \text{[it is a rank one matrix]}$$

  - Between class scattering of projected data $v^T S_b v$

- We want to find v that maximize $v^T S_b v$ but minimize $v^T S_w v$ while respect the constraint $v : v^T v = 1$

  - Multi-objective optimization problem

LDA objective

- Fisher's solution: use the Rayleigh's quotient

$$\max_v J(v) = \frac{v^T S_b v}{v^T S_w v}$$

  - Not using its inverse because between class scattering matrix has rank 1, so $v^T S_b v$ can be zero
  - It becomes an unconstrained optimization problem since any scaling factor in v cancels out

- Solution $\nabla_v \frac{v^T S_b v}{v^T S_w v} = \frac{\nabla_v v^T S_b v}{v^T S_w v} - \frac{v^T S_b v}{(v^T S_w v)^2} \nabla_v v^T S_w v$

$$= \frac{2}{v^T S_w v} S_b v - \frac{2 v^T S_b v}{(v^T S_w v)^2} S_w v = \frac{2}{v^T S_w v} \left(S_b v - \frac{v^T S_b v}{v^T S_w v} S_w v \right) = S_b v - J(v) S_w v$$

maximizing Rayleigh's quotient

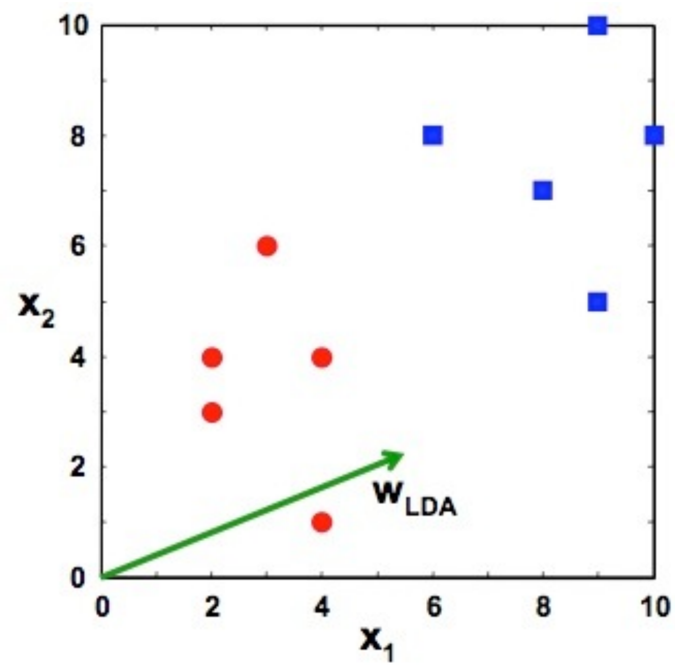
- $S_b v = J(v) S_w v$: solution given by equation $\lambda S_w v = S_b v$, known as the ***generalized eigenvalue*** problem
    - λ is the generalized eigenvalue for pair (S_w, S_b)
    - v is the corresponding generalized eigenvector
  - λ is the optimal value for J(v)
- solve $\lambda S_w v = S_b v$: when S_w is invertible, v is eigenvector of the top eigenvalue for matrix $S_w^{-1} S_b$, with λ being the corresponding eigenvalue
  - There is only one non-zero generalized eigenvalue in this case, because S_b is a rank one matrix

numerical example

- compute LDA projection for 2D data set
  - X+ = {(4,1),(2,4),(2,3),(3,6),(4,4)}
  - X- = {(9,10),(6,8),(9,5),(8,7),(10,8)}
- class statistics
  - means
    - μ+ = (3.00, 3.60), μ- = (8.40, 7.60)
  - covariances
    - $S+ = \begin{pmatrix} 4.00 & -2.00 \\ -2.00 & 13.00 \end{pmatrix}$, $S- = \begin{pmatrix} 9.20 & -0.20 \\ -0.20 & 13.20 \end{pmatrix}$
- within and between class scattering matrices
  - $Sb = \begin{pmatrix} 29.16 & 21.60 \\ 21.60 & 16.00 \end{pmatrix}$, $Sw = \begin{pmatrix} 13.20 & -2.20 \\ -2.20 & 26.40 \end{pmatrix}$
- solving LDA (generalized eigenvalue problem)

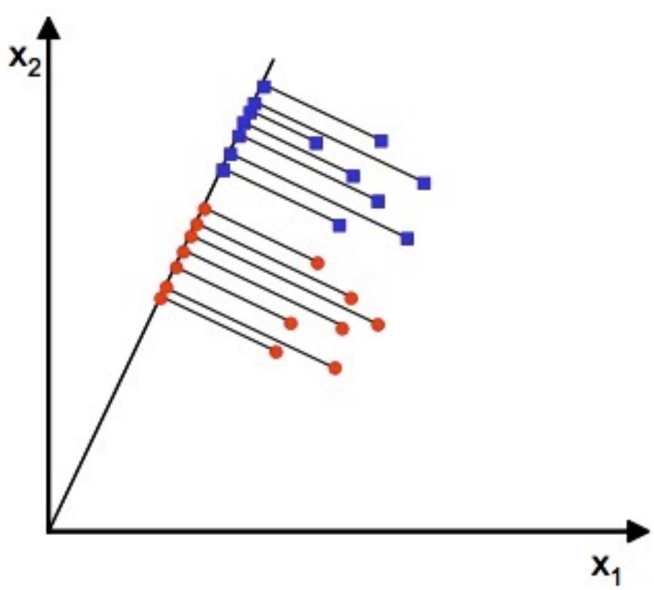
$$S_W^{-1} S_B v = \lambda v \Rightarrow \left| S_W^{-1} S_B - \lambda I \right| = 0 \Rightarrow \begin{vmatrix} 11.89 - \lambda & 8.81 \\ 5.08 & 3.76 - \lambda \end{vmatrix} = 0 \Rightarrow \lambda = 15.65$$

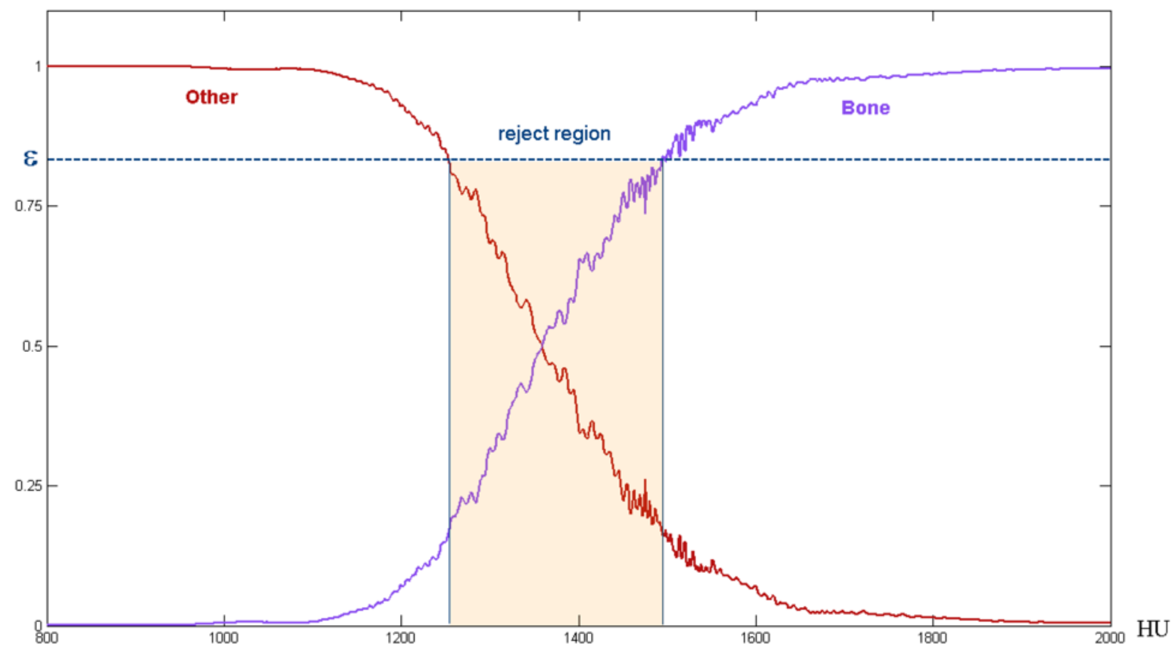
$$\begin{bmatrix} 11.89 & 8.81 \\ 5.08 & 3.76 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 15.65 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \Rightarrow \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0.91 \\ 0.39 \end{bmatrix}$$



choice of optimal threshold in LDA

- LDA only provides a 1D projection direction for the two classes of data, classifier is found as a threshold: pick a threshold on the projection line classification based on which side a datum is on
- the cross-over point to the CDFs of the two classes



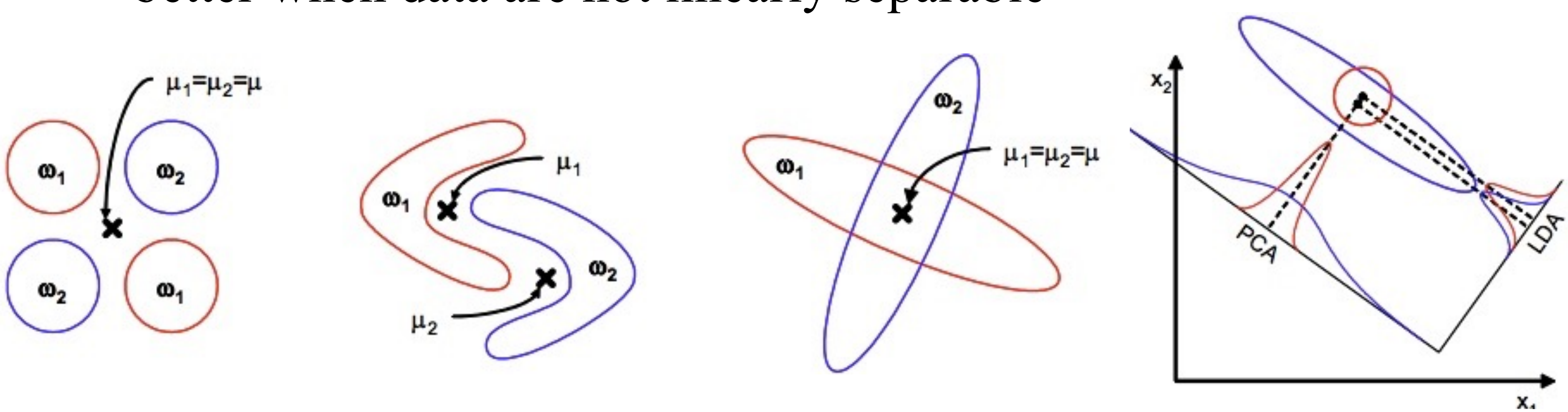


Fisher faces

- Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman, Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, IEEE TPAMI (1997)
- use LDA to obtain faces with glasses and without glasses

limitations of LDA

- LDA produces at most C-1 feature projections: if the classification error estimates establish that more features are needed, other method must be employed to provide those additional features
- implicit structure of data distributions
- fail when the discriminatory information is not in the mean but rather in the variance of the data
- LDA does not construct a classifier directly, instead, it finds a projection where classification is easier
  - a method directly minimizes classification errors may work better when data are not linearly separable



Generalization to multi-class

- We need to extend the problem formulation to multi-class classification problem (class number = K >2)
- Two important relation
  - $S_b = (\mu_+ - \mu)(\mu_+ - \mu)^T + (\mu_- - \mu)(\mu_- - \mu)^T$, when we have multiple classes of data, we can extend this to
    $$S_b = \sum_{j=1}^{K} (\mu_j - \mu)(\mu_j - \mu)^T$$
  - $S = S_w + S_b$: total data scattering is the sum of within-class scattering and between class scattering
- We can solve the same problem, and the solution is given by

$$S_b v = \lambda S_w v = \lambda (S - S_b) v \Rightarrow S_b v = \frac{\lambda}{\lambda + 1} S v \Rightarrow S_b v = \lambda' S v,$$

solution given by generalized eigenvalue problem

Proof

- Need to show that

$$\frac{n_+}{n}\mu_+\mu_+^T + \frac{n_-}{n}\mu_-\mu_-^T - \frac{n_+n_-}{n^2}(\mu_+ - \mu_-)(\mu_+ - \mu_-)^T = \mu\mu^T$$

- Reformulate the between class scattering matrix

$$\mu_+ - \mu = \mu_+ - \frac{n_+}{n}\mu_+ - \frac{n_-}{n}\mu_- = \frac{n_-}{n}(\mu_+ - \mu_-)$$

$$\mu_- - \mu = \mu_- - \frac{n_+}{n}\mu_+ - \frac{n_-}{n}\mu_- = \frac{n_+}{n}(\mu_- - \mu_+)$$

$$(\mu_+ - \mu)(\mu_+ - \mu)^T = \frac{n_-^2}{n^2}(\mu_+ - \mu_-)(\mu_+ - \mu_-)^T$$

$$(\mu_- - \mu)(\mu_- - \mu)^T = \frac{n_+^2}{n^2}(\mu_+ - \mu_-)(\mu_+ - \mu_-)^T$$

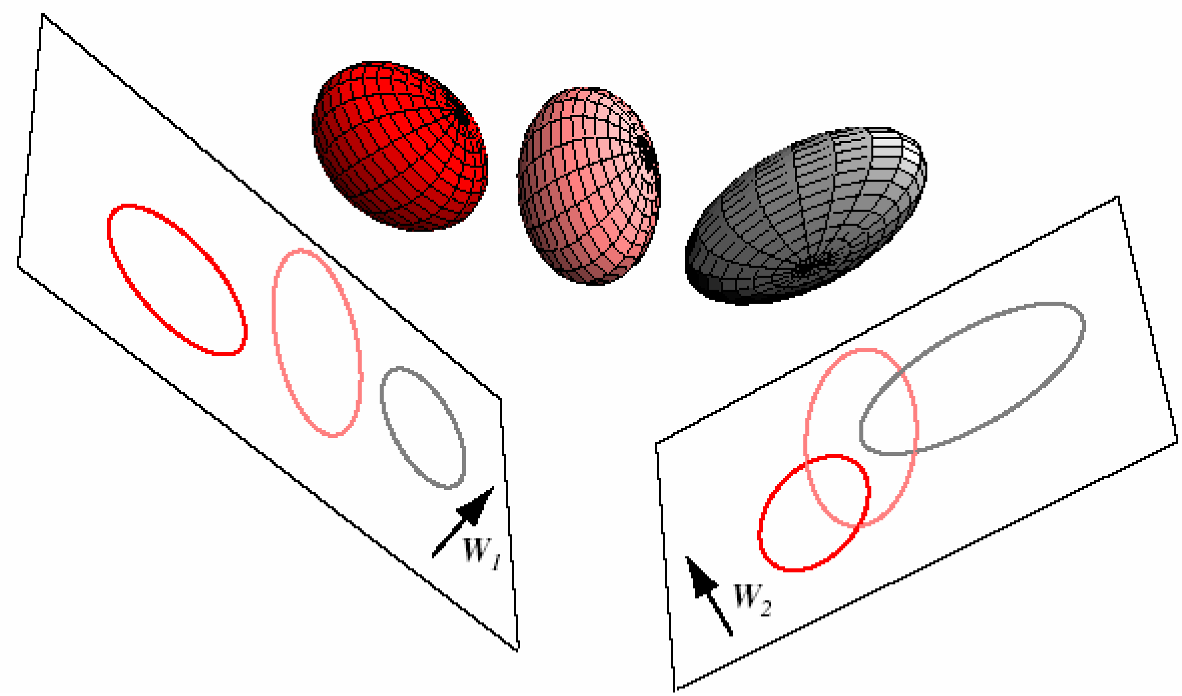
$$\frac{1}{n_+}(\mu_+ - \mu)(\mu_+ - \mu)^T = \frac{n_-^2}{n^2 n_+}(\mu_+ - \mu_-)(\mu_+ - \mu_-)^T$$

$$\frac{1}{n_-}(\mu_- - \mu)(\mu_- - \mu)^T = \frac{n_+^2}{n^2 n_-}(\mu_+ - \mu_-)(\mu_+ - \mu_-)^T$$

$$\frac{n}{n_+}(\mu_+ - \mu)(\mu_+ - \mu)^T + \frac{n}{n_-}(\mu_- - \mu)(\mu_- - \mu)^T = (\mu_+ - \mu_-)(\mu_+ - \mu_-)^T = S_b$$

multi-class LDA

- In multi-class case, the between-class scattering matrix $S_b = \sum_{j=1}^{K} (\mu_j - \mu)(\mu_j - \mu)^T$ has rank K-1
  - The K vectors are linearly dependent with rank K-1 $\sum_{j=1}^{K} \frac{n_k}{n} (\mu_j - \mu) = 0$



Relation with LLSE

- Solution given by $\lambda S_w v = S_b v$, as $S_w = S - S_b$, we can have equivalently $\lambda S v = (\lambda + 1) S_b v$
- Bringing back the definition of S_b, we have $\lambda S v = \dfrac{n_+ n_-}{n^2}(\lambda + 1)(\mu_+ - \mu_-)(\mu_+ - \mu_-)^T v$
- Clearing up all constants, we have $v \propto S^{-1}(\mu_+ - \mu_-)$, this is exactly the same solution we get from the LLSE solution to binary classification
  - The inverse covariance matrix modulated difference of class mean is the optimal direction for 1D linear classification from two different point of views

Comparison with PCA (TLSE)

- If we re-formulate LDA objective function $\max_v \frac{v^T S_b v}{v^T S_w v} = \max_v \frac{v^T (S - S_w) v}{v^T S_w v} = \max_v \frac{v^T S v}{v^T S_w v} - 1$, so equivalently we can solve rewrite it as $\max_v \frac{v^T S v}{v^T S_w v}$
- We can also reformulate PCA (Total LSE)
  - Original formulation: $\max_v v^T S v$ s.t $v^T v = 1$, it is equivalent to $\max_v \frac{v^T S v}{v^T v}$
- Compare the two formulations, we see that PCA denominator has no class specific information, while LDA does