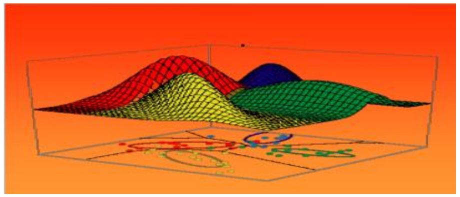


Pattern Recognition (CSE4213)

Faisal Muhammad Shah
Associate Professor,Dept Of CSE, AUST

- Linear Discriminant Analysis (LDA)

Outline

- LDA objective
- Recall … PCA
- Now … LDA
- LDA … Two Classes
  - Counter example
- LDA … C Classes
  - Illustrative Example
- LDA vs PCA Example
- Limitations of LDA

LDA Objective

- The objective of LDA is to perform dimensionality reduction …
  - So what, PCA does this 📪…
- However, we want to preserve as much of the class discriminatory information as possible.
  - OK, that's new, let dwell deeper ☺ …

Recall … PCA

- In PCA, the main idea to re-express the available dataset to extract the relevant information by reducing the redundancy and minimize the noise.

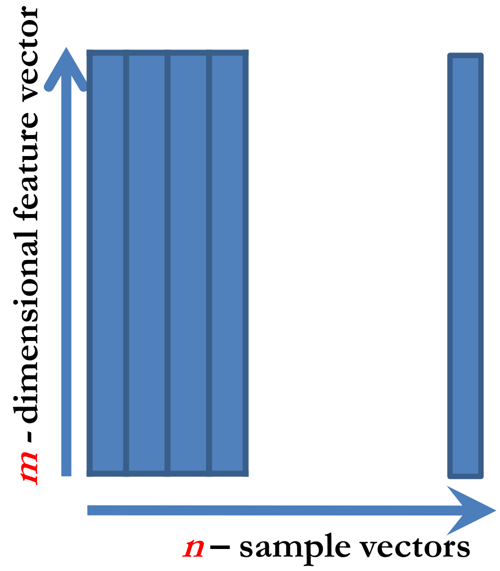
- We didn't care about whether this dataset represent features from one or more classes, i.e. the discrimination power was not taken into consideration while we were talking about PCA.

- In PCA, we had a dataset matrix **X** with dimensions *mxn*, where columns represent different data samples.

- We first started by subtracting the mean to have a zero mean dataset, then we computed the covariance matrix $\mathbf{S_x} = \mathbf{XX^T}$.

- Eigen values and eigen vectors were then computed for $\mathbf{S_x}$. Hence the new basis vectors are those eigen vectors with highest eigen values, where the number of those vectors was our choice.

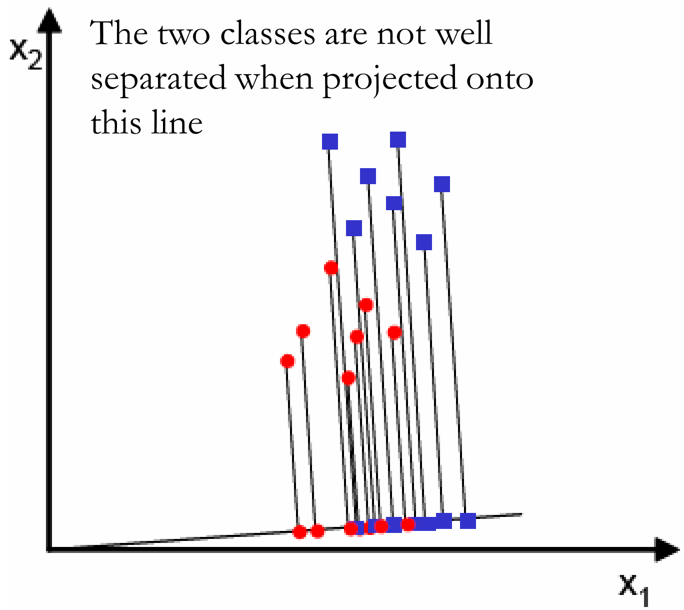
- Thus, using the new basis, we can project the dataset onto a less dimensional space with more powerful data representation.

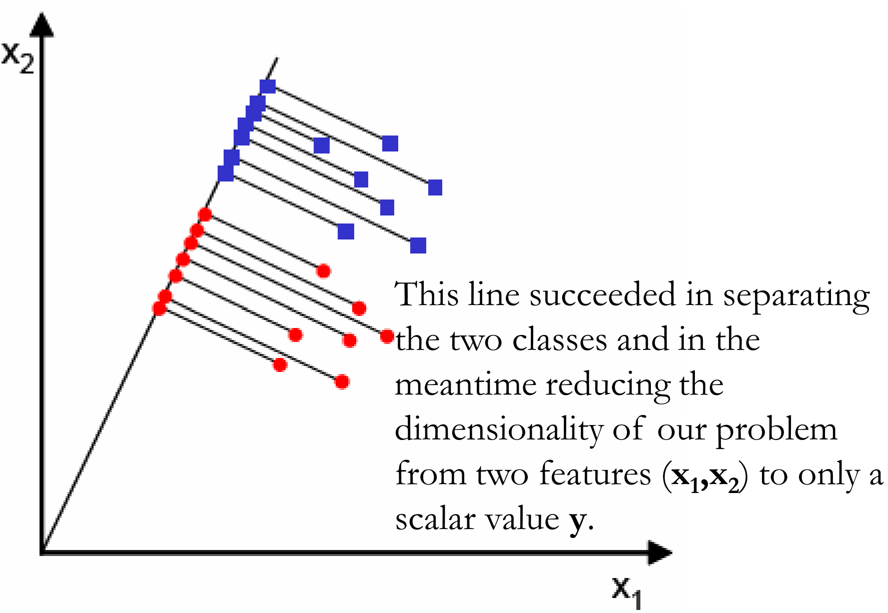


Now … LDA

- Consider a pattern classification problem, where we have C-classes, e.g. seabass, tuna, salmon …
- Each class has $\mathbf{N_i}$ m-dimensional samples, where $i = 1,2, \ldots, C$.
- Hence we have a set of m-dimensional samples $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^{N_i}\}$ belong to class $\boldsymbol{\omega_i}$.
- Stacking these samples from different classes into one big fat matrix **X** such that each column represents one sample.
- We seek to obtain a transformation of **X** to **Y** through projecting the samples in **X** onto a hyperplane with dimension C-1.
- **Let's see what does this mean?**

LDA … Two Classes





Assume we have *m*-dimensional samples $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N\}$, N_1 of which belong to $\boldsymbol{\omega}_1$ and N_2 belong to $\boldsymbol{\omega}_2$.

We seek to obtain a scalar **y** by projecting the samples **x** onto a line (C-1 space, C = 2).

$$y = w^T x \quad where \quad x = \begin{bmatrix} x_1 \\ . \\ . \\ x_m \end{bmatrix} \quad and \quad w = \begin{bmatrix} w_1 \\ . \\ . \\ w_m \end{bmatrix}$$

Of all the possible lines we would like to select the one that maximizes the separability of the scalars.

- Mean

$m_1 = w^T \mu_1$

$m_2 = w^T \mu_0$

dimension of mean after projection = d*1

- Covariance

$\Sigma_1 = w^T \Sigma_1 w$ $\quad [\, w^T \cdot w = w^2 \,]$

$\Sigma_2 = w^T \Sigma_0 w$

dimension of covariance after projection = d*d

- Goal 1: Maximize the distance of projected mean

$\Rightarrow (w^T \mu_1 - w^T \mu_0)^T (w^T \mu_1 - w^T \mu_0)$

$\Rightarrow (\mu_1 - \mu_0)^T w w^T (\mu_1 - \mu_0)$

$\Rightarrow w^T \underbrace{(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T}_{} w$

S_B = between class covariance

- $\boxed{\max \ w^T S_B w}$ objective fun^c

Goal 2: For each class minimize the variance:

$\Rightarrow w^T \Sigma_1 w + w^T \Sigma_0 w$

$\Rightarrow w^T \underbrace{(\Sigma_1 + \Sigma_0)}_{} w$

S_W = within class covariance

- $\boxed{\text{Min} \quad w^T S_W w}$ objective func

- Now, from fisher linear discriminant,

$$\boxed{\max \frac{w^T S_B w}{w^T S_W w}} \text{ - objective func}$$

- $\boxed{\begin{array}{l} \max \quad w^T S_B w \\ \text{subject to} \quad w^T S_W w = 1 \end{array}}$ → objective func with constraint

From Lagrange,

$$L(w, \lambda) = w^T S_B w - \lambda(w^T S_W w - 1)$$

Now, partial derivative,

$$\frac{\partial L}{\partial w} = 2 S_B w - 2\lambda S_W w = 0$$

$$\Rightarrow S_B w = \lambda S_W w$$

Multiply both side by S_W^{-1}

$\Rightarrow S_W^{-1} S_B w = \lambda w$

Here, w is a eigenvector of $S_W^{-1} S_B$.

LDA … Two Classes

- In order to find a good projection vector, we need to define a measure of separation between the projections.
- The mean vector of each class in **x** and **y** feature space is:

$$\mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x \quad and \quad \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i} w^T x$$

$$= w^T \frac{1}{N_i}\sum_{x\in\omega_i} x = w^T \mu_i$$

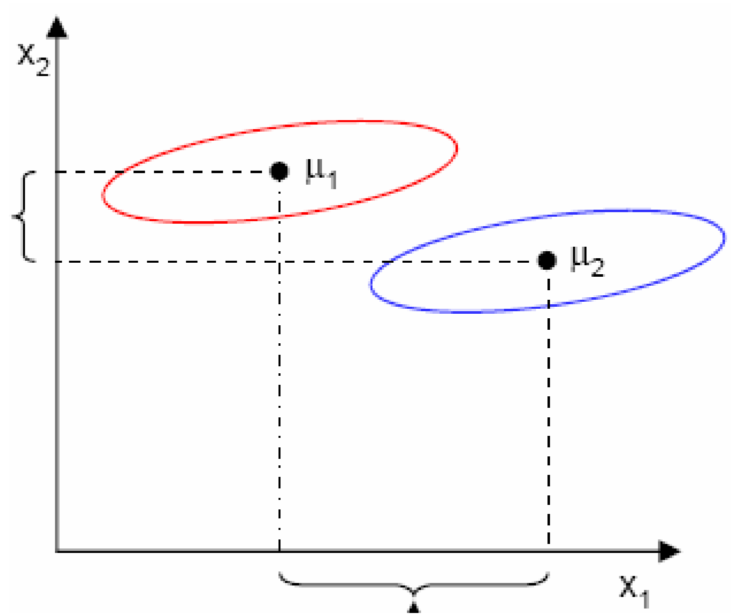
- We could then choose the distance between the projected means as our objective function

$$J(w) = \left|\tilde{\mu}_1 - \tilde{\mu}_2\right| = \left|w^T\mu_1 - w^T\mu_2\right| = \left|w^T(\mu_1 - \mu_2)\right|$$

LDA … Two Classes

- However, the distance between the projected means is not a very good measure since it does not take into account the standard deviation within the classes.

This axis yields better class separability



This axis has a larger distance between means

LDA … Two Classes

- The solution proposed by Fisher is to maximize a function that represents the difference between the means, normalized by a measure of the within-class variability, or the so-called *scatter*.

- For each class we define the scatter, an equivalent of the variance, as;

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} \left(y - \tilde{\mu}_i\right)^2$$

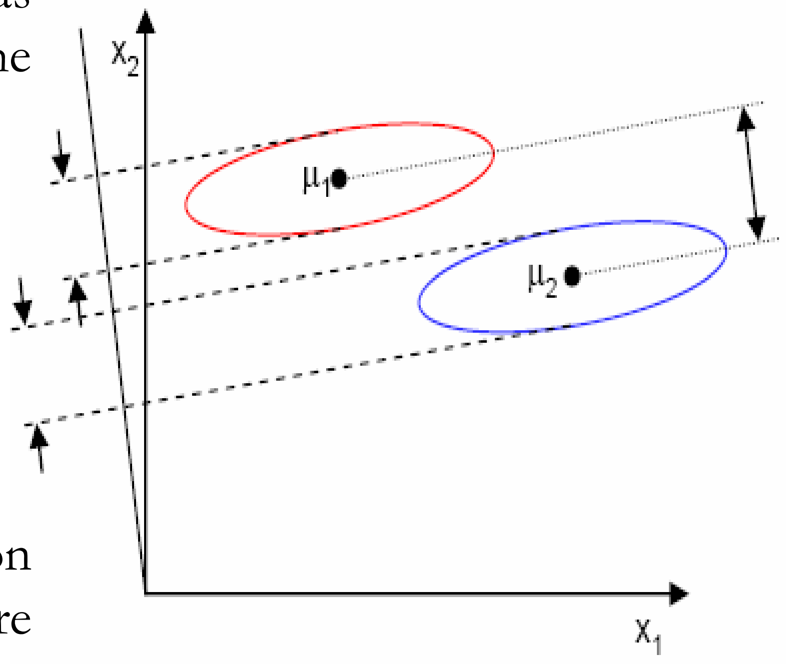
- $\tilde{s}_i^2$ measures the variability within class ω_i after projecting it on the y-space.

- Thus $\tilde{s}_1^2 + \tilde{s}_2^2$ measures the variability within the two classes at hand after projection, hence it is called *within-class scatter* of the projected samples.

LDA … Two Classes

- The Fisher linear discriminant is defined as the linear function $\mathbf{w}^{\mathbf{T}}\mathbf{x}$ that maximizes the criterion function:

$$J(w) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$



- Therefore, we will be looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible

LDA … Two Classes

- In order to find the optimum projection $\mathbf{w}^*$, we need to express *J(w)* as an explicit function of **w**.

- We will define a measure of the scatter in multivariate feature space **x** which are denoted as *scatter matrices*;

$$S_i = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$

$$S_w = S_1 + S_2$$

- Where $\mathbf{S_i}$ is the covariance matrix of class $\boldsymbol{\omega_i}$, and $\mathbf{S_w}$ is called the ***w****ithin-class scatter matrix*.

LDA … Two Classes

- Now, the scatter of the projection **y** can then be expressed as a function of the scatter matrix in feature space **x**.

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2 = \sum_{x \in \omega_i} \left(w^T x - w^T \mu_i\right)^2$$

$$= \sum_{x \in \omega_i} w^T (x - \mu_i)(x - \mu_i)^T w$$

$$= w^T S_i w$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^T S_1 w + w^T S_2 w = w^T (S_1 + S_2) w = w^T S_W w = \tilde{S}_W$$

Where $\tilde{S}_W$ is the within-class scatter matrix of the projected samples **y**.

LDA … Two Classes

- Similarly, the difference between the projected means (in y-space) can be expressed in terms of the means in the original feature space (x-space).

$$
\begin{aligned}
\left(\tilde{\mu}_1 - \tilde{\mu}_2\right)^2 &= \left(w^T \mu_1 - w^T \mu_2\right)^2 \\
&= w^T \underbrace{\left(\mu_1 - \mu_2\right)\left(\mu_1 - \mu_2\right)^T}_{S_B} w \\
&= w^T S_B w = \tilde{S}_B
\end{aligned}
$$

- The matrix $\mathbf{S_B}$ is called the ***b****etween-class scatter* of the original samples/feature vectors, while $\tilde{S}_B$ is the between-class scatter of the projected samples **y**.
- Since $\mathbf{S_B}$ is the outer product of two vectors, its rank is at most one.

LDA … Two Classes

- We can finally express the Fisher criterion in terms of $\mathbf{S_W}$ and $\mathbf{S_B}$ as:

$$J(w) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{w^T S_B w}{w^T S_W w}$$

- Hence *J(w)* is a measure of the difference between class means (encoded in the between-class scatter matrix) normalized by a measure of the within-class scatter matrix.

LDA … Two Classes

- To find the maximum of $J(w)$, we differentiate and equate to zero.

$$\frac{d}{dw}J(w)=\frac{d}{dw}\left(\frac{w^T S_B w}{w^T S_W w}\right)=0$$

$$\Rightarrow \left(w^T S_W w\right)\frac{d}{dw}\left(w^T S_B w\right)-\left(w^T S_B w\right)\frac{d}{dw}\left(w^T S_W w\right)=0$$

$$\Rightarrow \left(w^T S_W w\right)2S_B w-\left(w^T S_B w\right)2S_W w=0$$

$$Dividing\ \ by\ \ 2w^T S_W w:$$

$$\Rightarrow \left(\frac{w^T S_W w}{w^T S_W w}\right)S_B w-\left(\frac{w^T S_B w}{w^T S_W w}\right)S_W w=0$$

$$\Rightarrow S_B w-J(w)S_W w=0$$

$$\Rightarrow S_W^{-1}S_B w-J(w)w=0$$

LDA … Two Classes

- Solving the generalized eigen value problem

$$S_W^{-1} S_B w = \lambda w \qquad where \quad \lambda = J(w) = scalar$$

yields

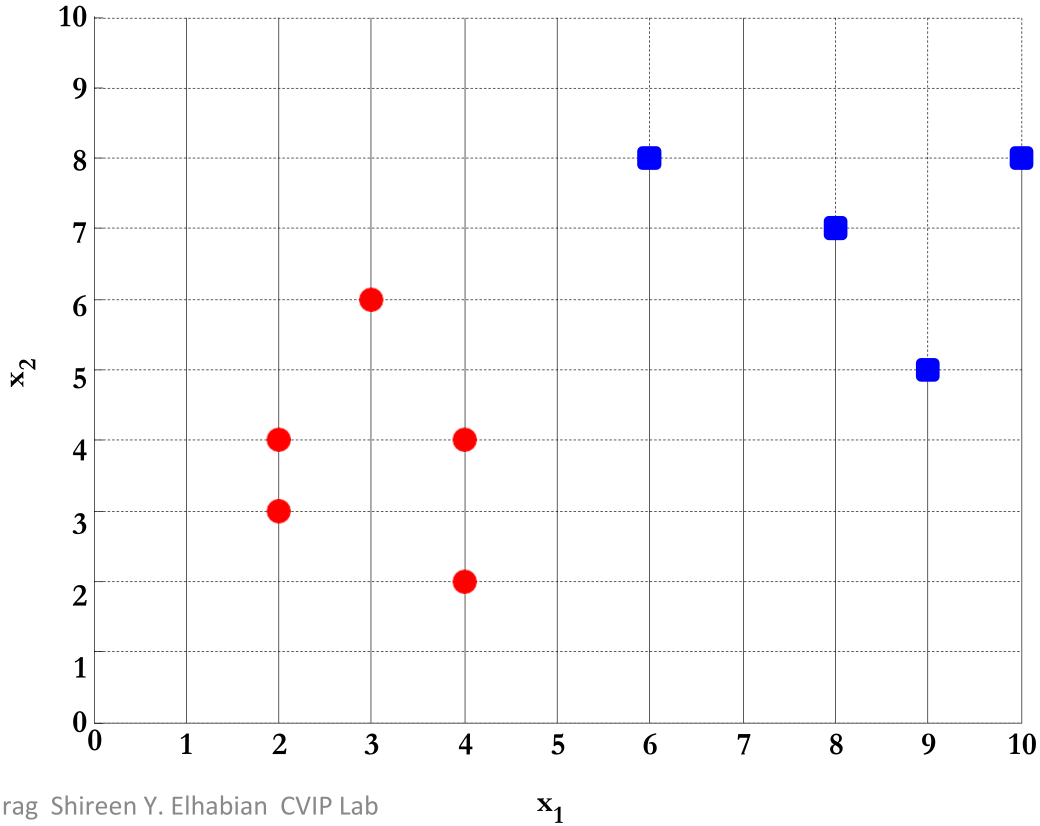
$$w^* = \arg\max_w J(w) = \arg\max_w \left(\frac{w^T S_B w}{w^T S_W w} \right) = S_W^{-1}(\mu_1 - \mu_2)$$

- This is known as Fisher's Linear Discriminant, although it is not a discriminant but rather a specific choice of direction for the projection of the data down to one dimension.

- Using the same notation as PCA, the solution will be the eigen vector(s) of $S_X = S_W^{-1} S_B$

LDA … Two Classes - Example

- Compute the Linear Discriminant projection for the following two-dimensional dataset.
  - Samples for class $\boldsymbol{\omega_1}$: $\mathbf{X_1}=(x_1,x_2)=\{(4,2),(2,4),(2,3),(3,6),(4,4)\}$
  - Sample for class $\boldsymbol{\omega_2}$: $\mathbf{X_2}=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$




```
% samples for class 1
X1 = [4,2;
      2,4;
      2,3;
      3,6;
      4,4];

% samples for class 2
X2 = [9,10;
      6,8;
      9,5;
      8,7;
      10,8];
```

LDA … Two Classes - Example

- The classes mean are :

$$\mu_1 = \frac{1}{N_1}\sum_{x\in\omega_1} x = \frac{1}{5}\left[\begin{pmatrix}4\\2\end{pmatrix}+\begin{pmatrix}2\\4\end{pmatrix}+\begin{pmatrix}2\\3\end{pmatrix}+\begin{pmatrix}3\\6\end{pmatrix}+\begin{pmatrix}4\\4\end{pmatrix}\right]=\begin{pmatrix}3\\3.8\end{pmatrix}$$

$$\mu_2 = \frac{1}{N_2}\sum_{x\in\omega_2} x = \frac{1}{5}\left[\begin{pmatrix}9\\10\end{pmatrix}+\begin{pmatrix}6\\8\end{pmatrix}+\begin{pmatrix}9\\5\end{pmatrix}+\begin{pmatrix}8\\7\end{pmatrix}+\begin{pmatrix}10\\8\end{pmatrix}\right]=\begin{pmatrix}8.4\\7.6\end{pmatrix}$$

```
% class means
Mu1 = mean(X1)';
Mu2 = mean(X2)';
```

LDA … Two Classes - Example

- Covariance matrix of the first class:

$$S_1 = \sum_{x \in \omega_1} (x - \mu_1)(x - \mu_1)^T = \left[\begin{pmatrix} 4 \\ 2 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 2 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2$$

$$+ \left[\begin{pmatrix} 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 3 \\ 6 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 4 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2$$

$$= \begin{pmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{pmatrix}$$

```
% covariance matrix of the first class
S1 = cov(X1);
```

LDA … Two Classes - Example

- Covariance matrix of the second class:

$$S_2 = \sum_{x \in \omega_2} (x - \mu_2)(x - \mu_2)^T = \left[\begin{pmatrix} 9 \\ 10 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 6 \\ 8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2$$

$$+ \left[\begin{pmatrix} 9 \\ 5 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 8 \\ 7 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 10 \\ 8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2$$

$$= \begin{pmatrix} 2.3 & -0.05 \\ -0.05 & 3.3 \end{pmatrix}$$

```
% covariance matrix of the first class
S2 = cov(X2);
```

LDA … Two Classes - Example

- Within-class scatter matrix:

$$S_w = S_1 + S_2 = \begin{pmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{pmatrix} + \begin{pmatrix} 2.3 & -0.05 \\ -0.05 & 3.3 \end{pmatrix}$$

$$= \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix}$$

```
% within-class scatter matrix
Sw = S1 + S2 ;
```

LDA … Two Classes - Example

- Between-class scatter matrix:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

$$= \left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]\left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^T$$

$$= \begin{pmatrix} -5.4 \\ -3.8 \end{pmatrix}\begin{pmatrix} -5.4 & -3.8 \end{pmatrix}$$

$$= \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix}$$

```
% between-class scatter matrix
SB = (Mu1-Mu2)*(Mu1-Mu2)';
```

LDA … Two Classes - Example

- The LDA projection is then obtained as the solution of the generalized eigen value problem $S_W^{-1} S_B w = \lambda w$

$$\Rightarrow \left| S_W^{-1} S_B - \lambda I \right| = 0$$

$$\Rightarrow \left| \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix}^{-1} \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0$$

$$\Rightarrow \left| \begin{pmatrix} 0.3045 & 0.0166 \\ 0.0166 & 0.1827 \end{pmatrix} \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0$$

$$\Rightarrow \left| \begin{pmatrix} 9.2213 - \lambda & 6.489 \\ 4.2339 & 2.9794 - \lambda \end{pmatrix} \right|$$

$$= (9.2213 - \lambda)(2.9794 - \lambda) - 6.489 \times 4.2339 = 0$$

$$\Rightarrow \lambda^2 - 12.2007\lambda = 0 \Rightarrow \lambda(\lambda - 12.2007) = 0$$

$$\Rightarrow \lambda_1 = 0, \lambda_2 = 12.2007$$

LDA … Two Classes - Example

- Hence

$$\begin{pmatrix} 9.2213 & 6.489 \\ 4.2339 & 2.9794 \end{pmatrix} w_1 = \underbrace{0}_{\lambda_1} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

and

$$\begin{pmatrix} 9.2213 & 6.489 \\ 4.2339 & 2.9794 \end{pmatrix} w_2 = \boxed{\underbrace{12.2007}_{\lambda_2}} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

Thus;

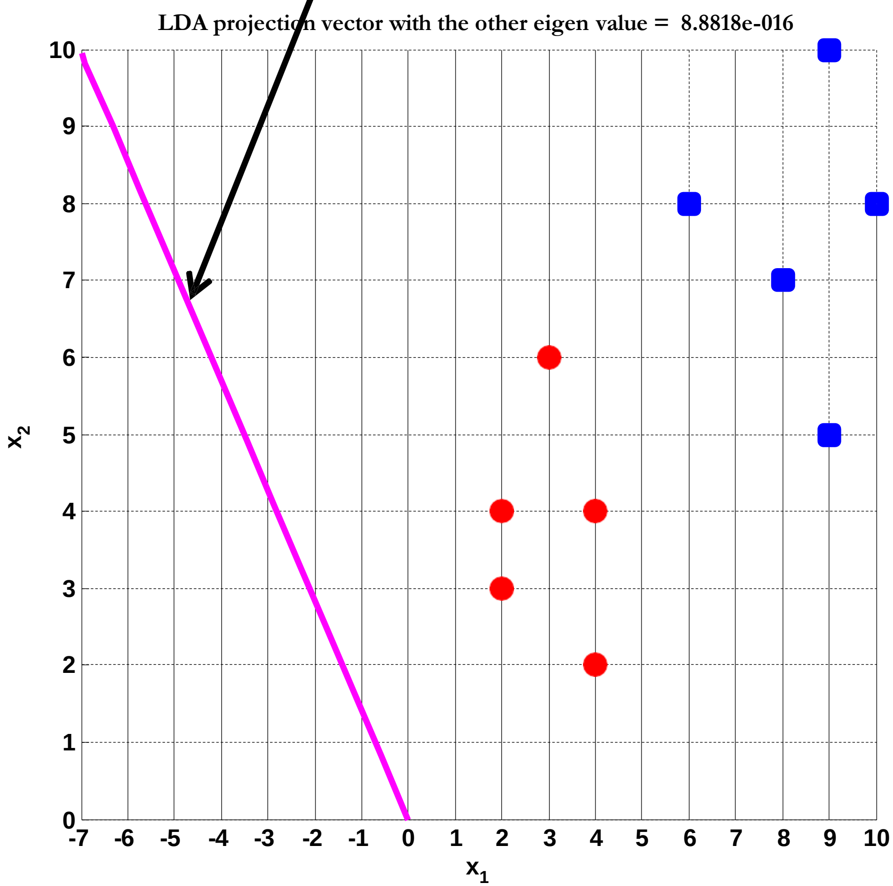
$$w_1 = \begin{pmatrix} -0.5755 \\ 0.8178 \end{pmatrix} \quad and \quad \boxed{w_2 = \begin{pmatrix} 0.9088 \\ 0.4173 \end{pmatrix} = w^*}$$

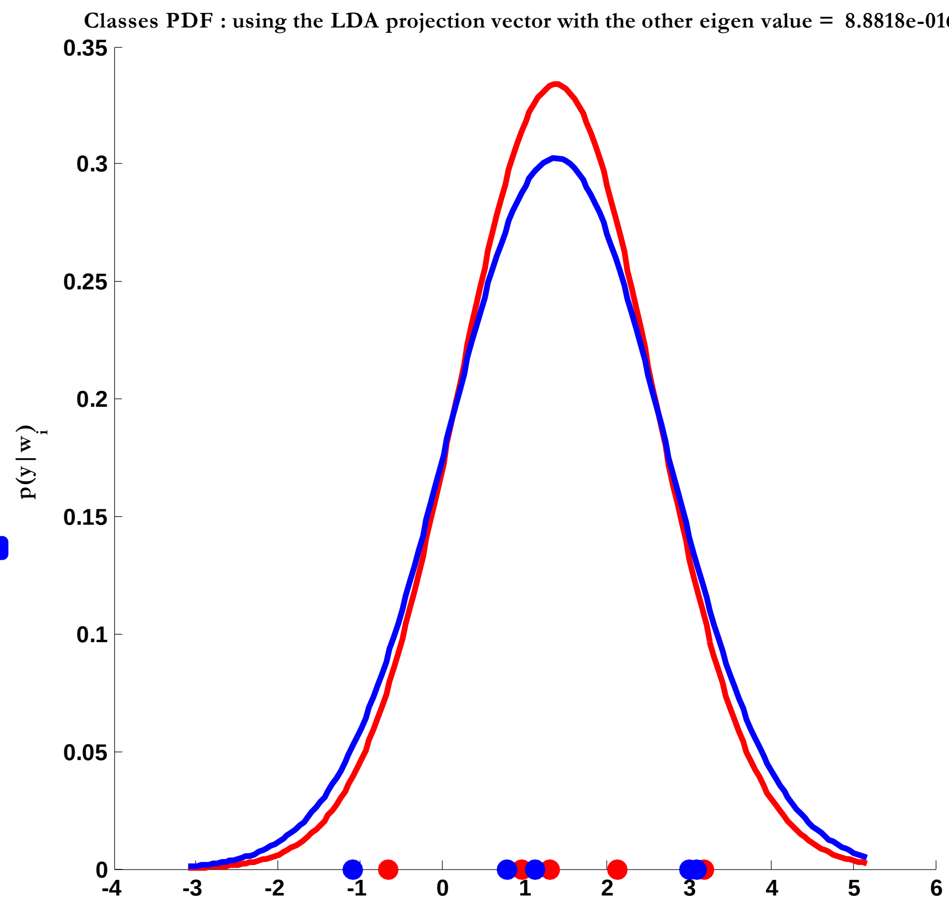
```
% computing the LDA projection
invSw = inv(Sw);

invSw_by_SB = invSw * SB;

% getting the projection vector
[V,D] = eig(invSw_by_SB)

% the projection vector
W = V(:,1);
```

- The optimal projection is the one that given maximum $\lambda = J(w)$

LDA - Projection

The projection vector corresponding to the **smallest** eigen value

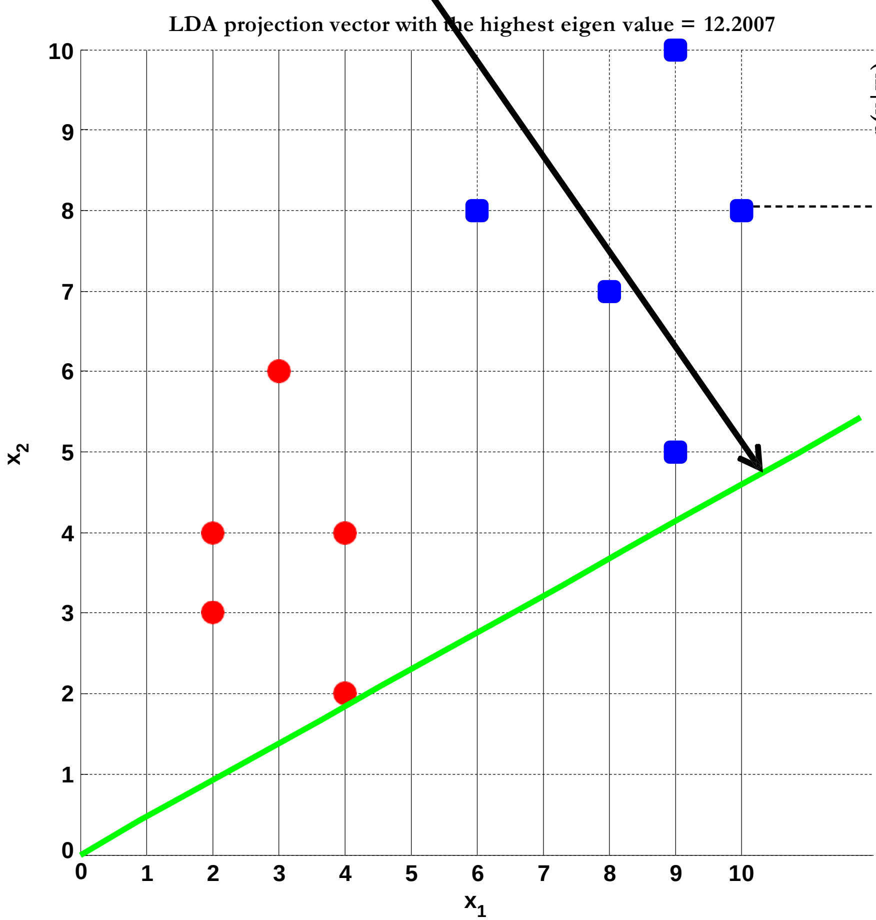


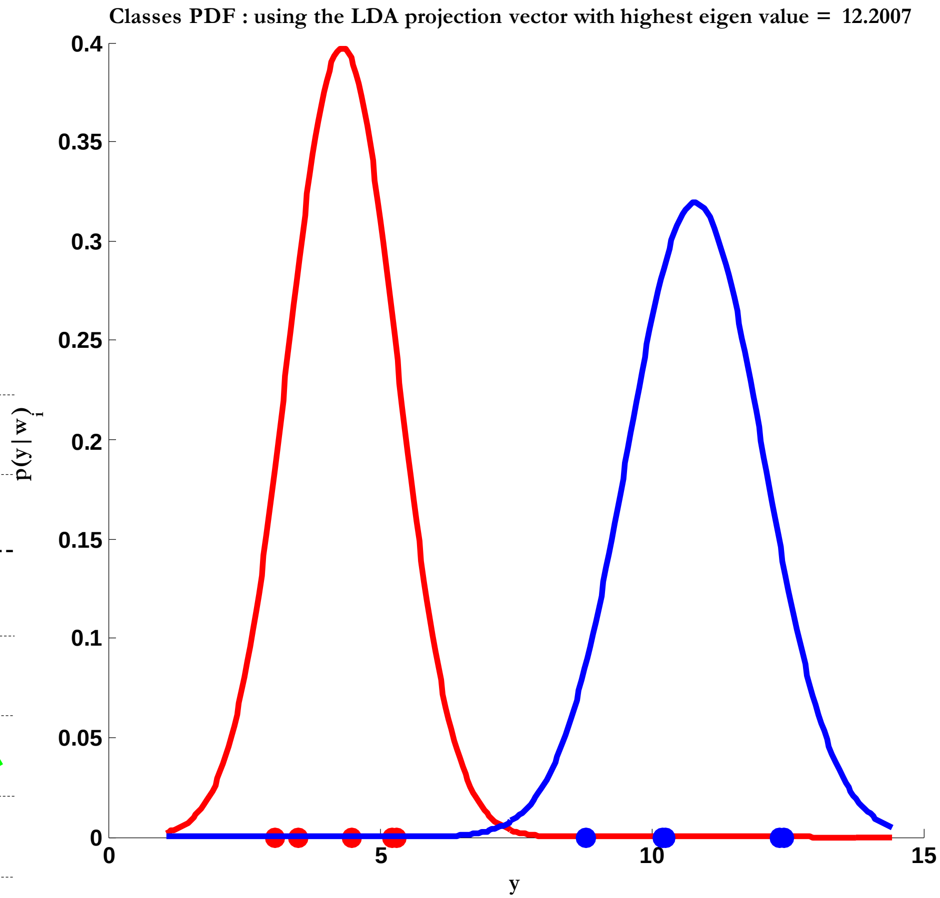


Using this vector leads to **bad separability** between the two classes

LDA - Projection

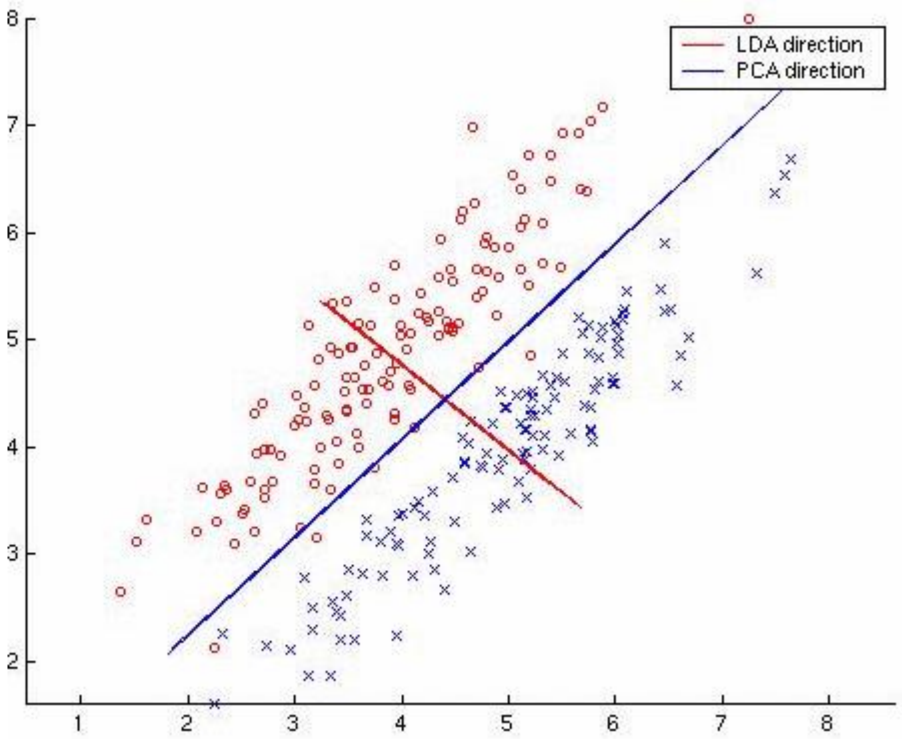
The projection vector corresponding to the **highest** eigen value

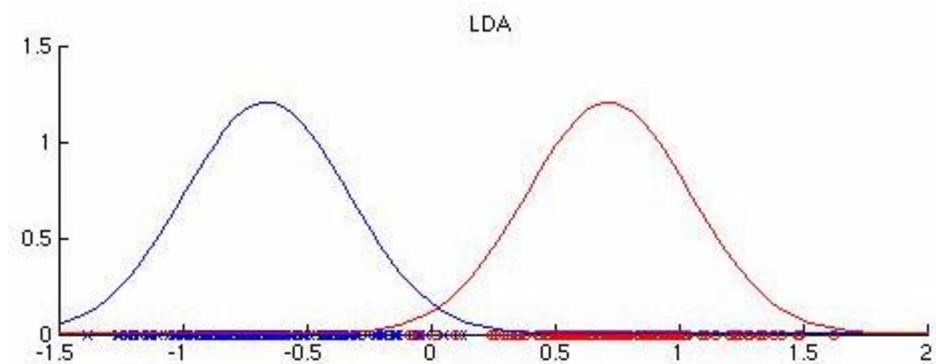


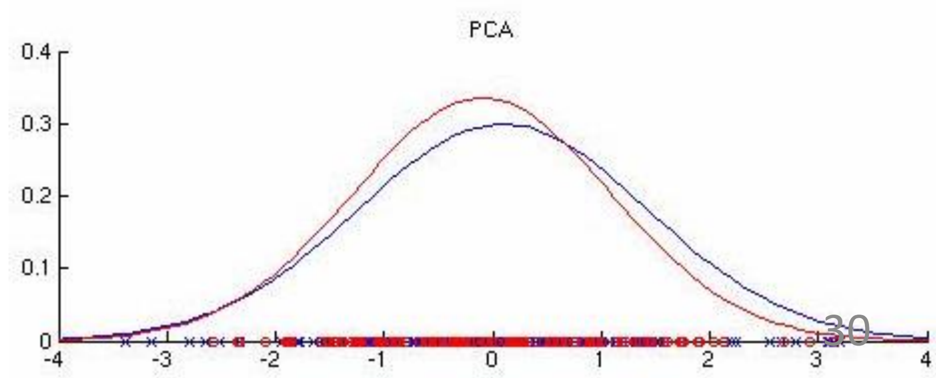


Using this vector leads to **good separability** between the two classes

PCA vs LDA

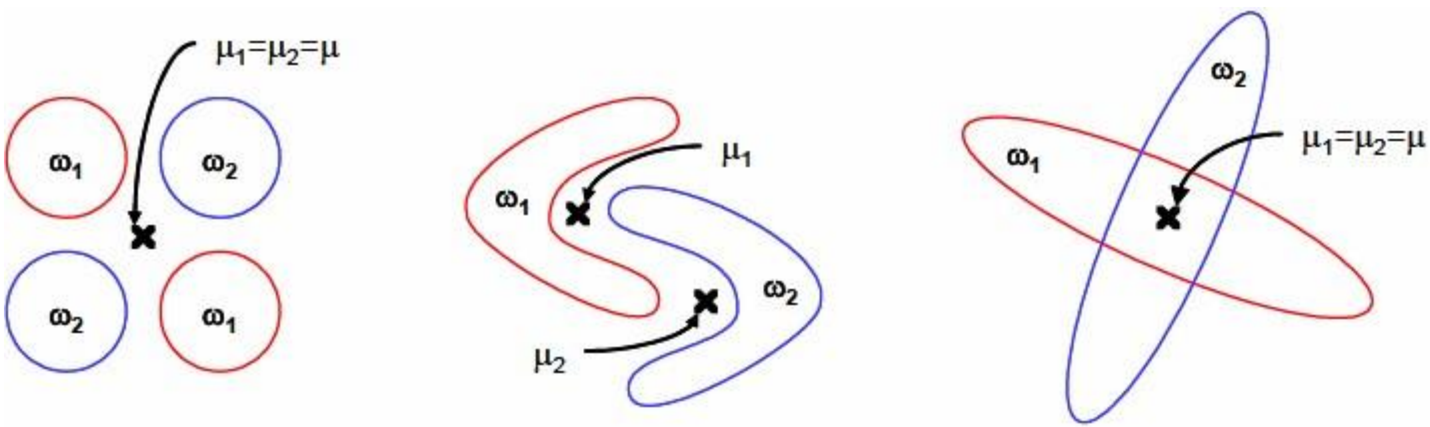






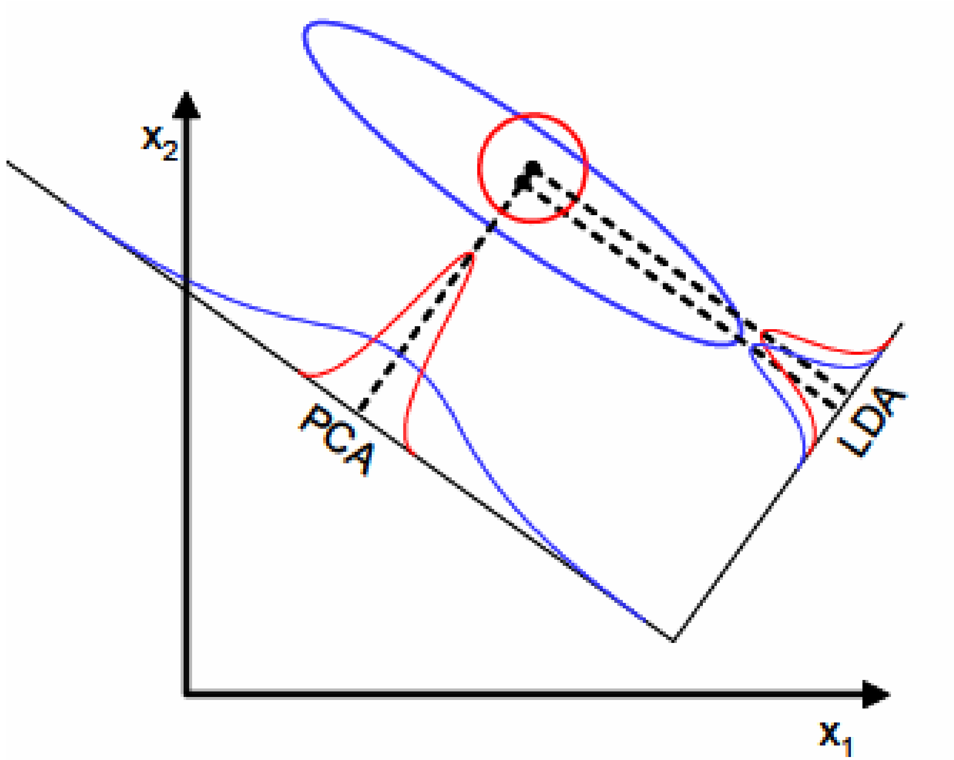
Limitations of LDA

- **LDA produces at most C-1 feature projections**
  - If the classification error estimates establish that more features are needed, some other method must be employed to provide those additional features
- **LDA is a parametric method since it assumes unimodal Gaussian likelihoods**
  - If the distributions are significantly non-Gaussian, the LDA projections will not be able to preserve any complex structure of the data, which may be needed for classification.



Limitations of LDA

- LDA will fail when the discriminatory information is not in the mean but rather in the variance of the data



Thank You