Linear Discriminant Analysis

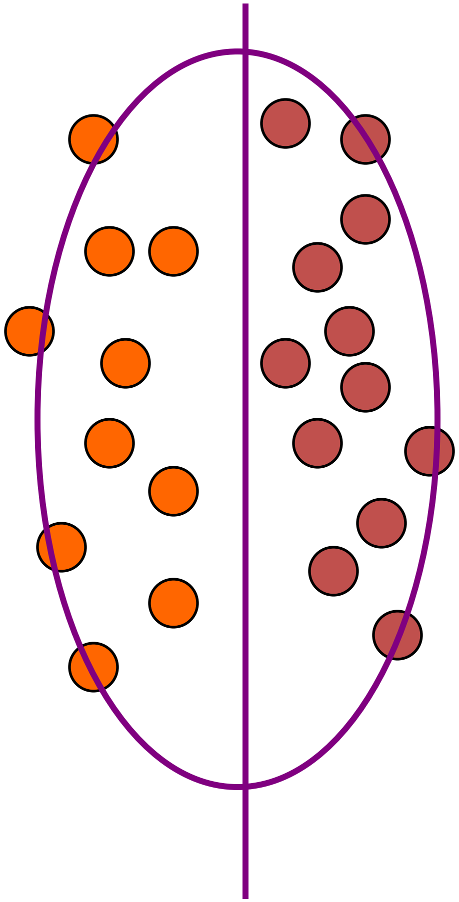
Slides have been prepared by utilizing material from the internet. No authorship is claimed for the same

Linear Discriminant Analysis

- First applied by M. Barnard at the suggestion of R. A. Fisher (1936), Fisher linear discriminant analysis (FLDA):
  - Dimension reduction
    - Finds linear combinations of the features $\mathbf{X}=X_1,\ldots,X_d$ with large ratios of between-groups to within-groups sums of squares - discriminant variables;
  - Classification
    - Predicts the class of an observation $\mathbf{X}$ by the class whose mean vector is closest to $\mathbf{X}$ in terms of the discriminant variables

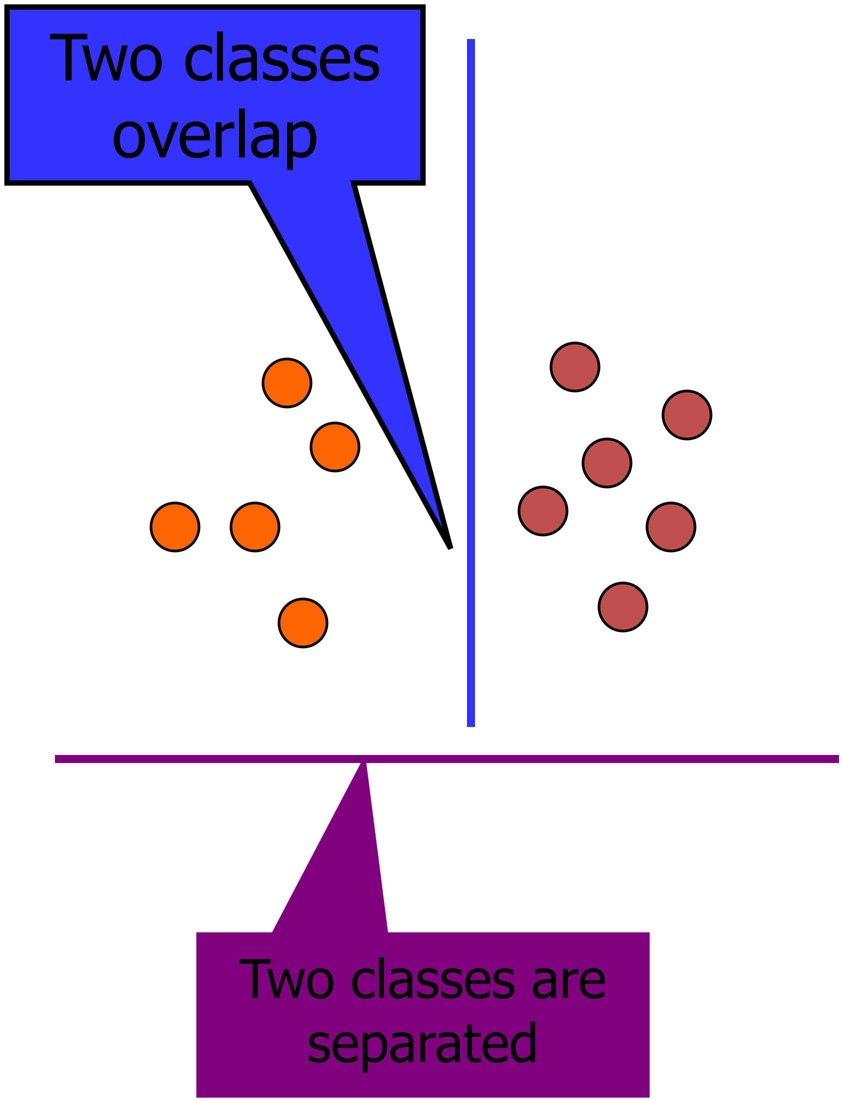
Is PCA a good criterion for classification?

- Data variation determines the projection direction
- What’s missing?
  - Class information



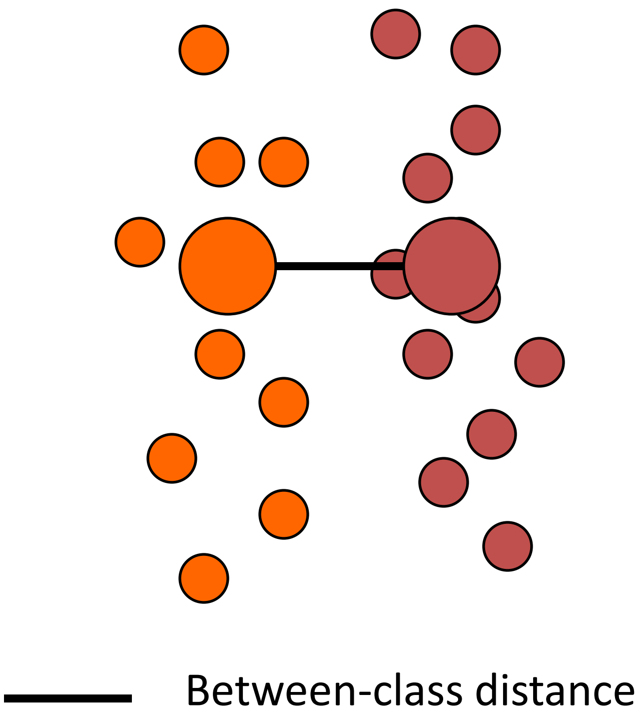
What is a good projection?

- Similarly, what is a good criterion?
  - Separating different classes



What class information may be useful?

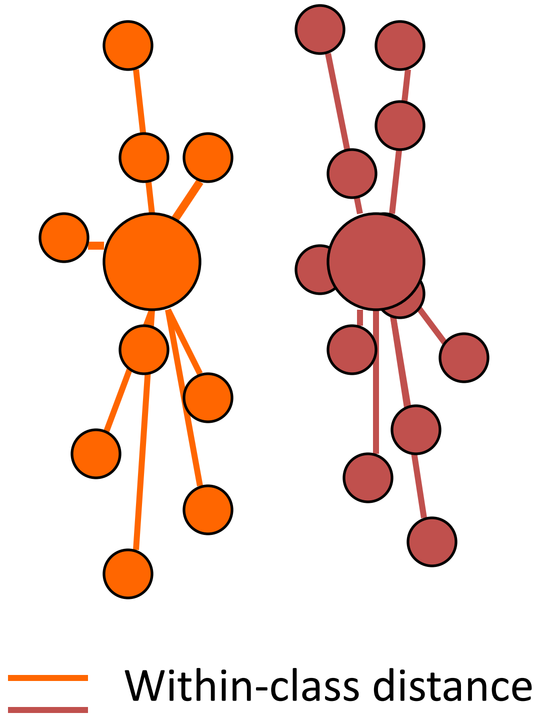
- Between-class distance
  - Distance between the centroids of different classes



What class information may be useful?

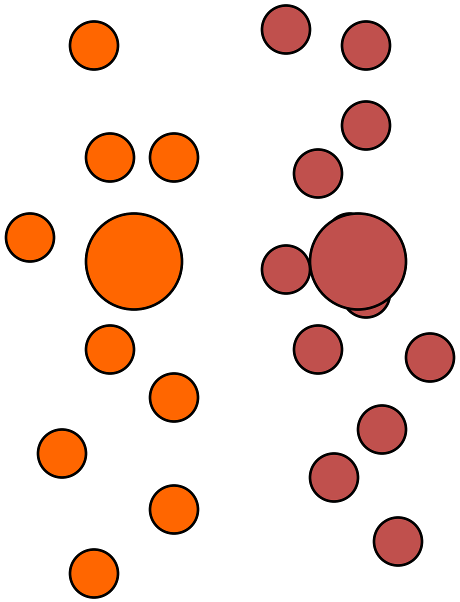
- Between-class distance
  - Distance between the centroids of different classes

- Within-class distance
  - Accumulated distance of an instance to the centroid of its class



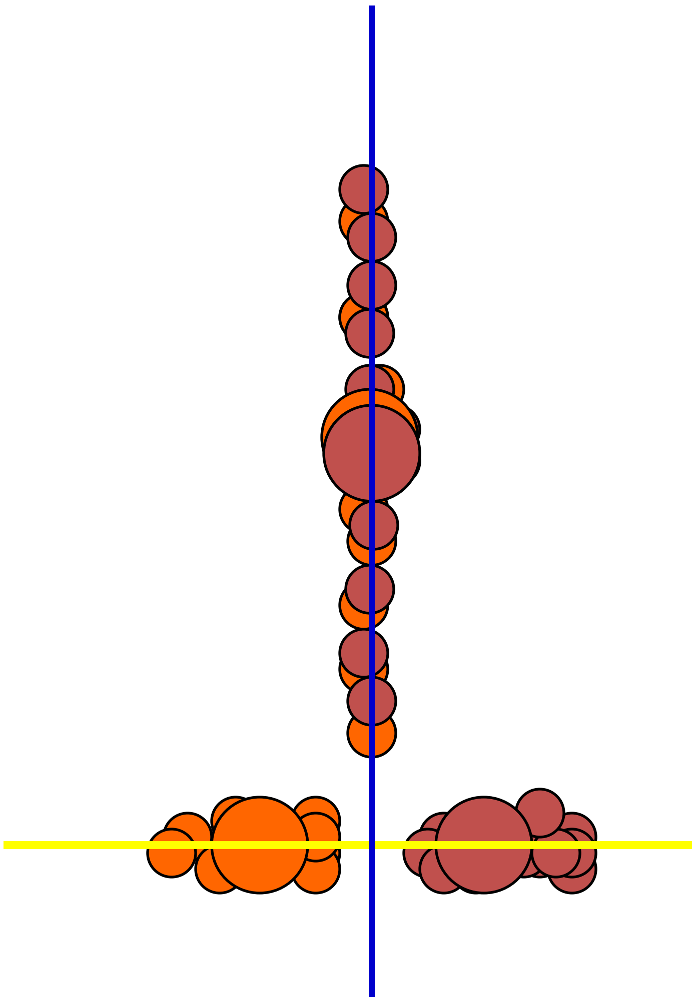
Linear discriminant analysis (LDA)

- Linear discriminant analysis (LDA) finds most discriminant projection by maximizing between-class distance and minimizing within-class distance



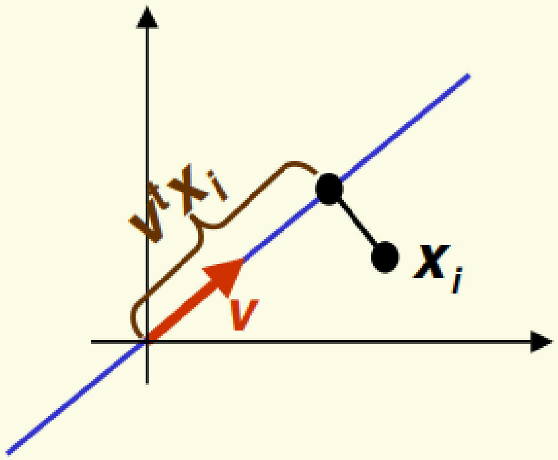
LDA

- Linear discriminant analysis (LDA) finds most discriminant projection by maximizing between-class distance and minimizing within-class distance



Problem Setup

- Suppose we have 2 classes and $\boldsymbol{d}$-dimensional samples $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ where
  - $\boldsymbol{n}_1$ samples come from the first class
  - $\boldsymbol{n}_2$ samples come from the second class
- consider projection on a line
- Let the line direction be given by unit vector $\boldsymbol{v}$



- Scalar $\boldsymbol{v}^t\boldsymbol{x}_i$ is the distance of projection of $\boldsymbol{x}_i$ from the origin
- Thus it $\boldsymbol{v}^t\boldsymbol{x}_i$ is the projection of $\boldsymbol{x}_i$ into a one dimensional subspace

Problem setup

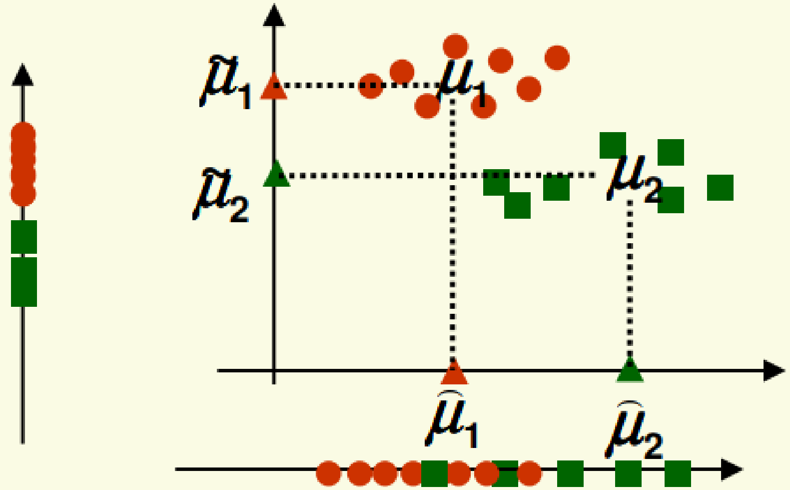
- Thus the projection of sample $\boldsymbol{x_i}$ onto a line in direction $\boldsymbol{v}$ is given by $\boldsymbol{v^t x_i}$
- How to measure separation between projections of different classes?
- Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ be the means of projections of classes 1 and 2
- Let μ_1 and μ_2 be the means of classes 1 and 2
- $\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|$ seems like a good measure

$$\tilde{\mu}_1 = \frac{1}{n_1} \sum_{x_i \in C1}^{n_1} \boldsymbol{v}^t \boldsymbol{x}_i = \boldsymbol{v}^t \left(\frac{1}{n_1} \sum_{x_i \in C1}^{n_1} \boldsymbol{x}_i \right) = \boldsymbol{v}^t \mu_1$$

similarly, $\qquad \tilde{\mu}_2 = \boldsymbol{v}^t \mu_2$

Is it Good Enough ?

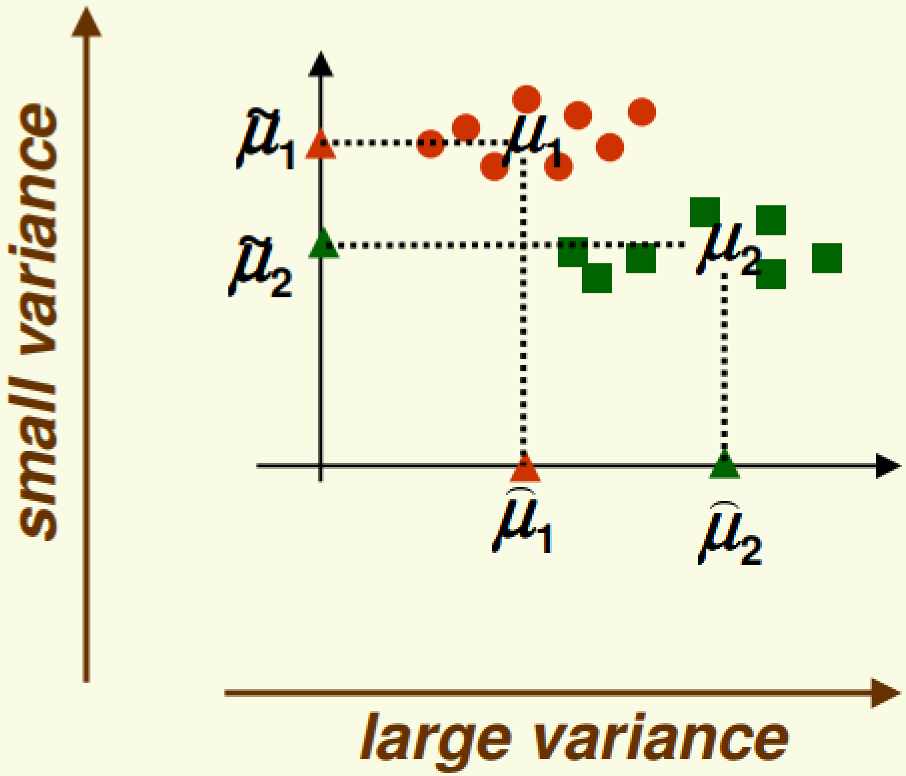
- How good is $|\tilde{\mu}_1 - \tilde{\mu}_2|$ as a measure of separation?
  - The larger $|\tilde{\mu}_1 - \tilde{\mu}_2|$, the better is the expected separation



- the vertical axes is a better line than the horizontal axes to project to for class separability
- however $|\hat{\mu}_1 - \hat{\mu}_2| > |\tilde{\mu}_1 - \tilde{\mu}_2|$

Variance of Classes

- The problem with $|\tilde{\mu}_1 - \tilde{\mu}_2|$ is that it does not consider the variance of the classes

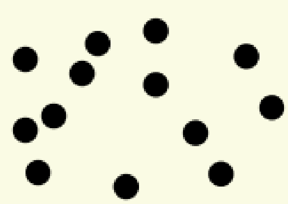
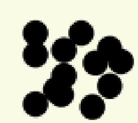


Scatter of Classes

- We need to normalize $|\tilde{\mu}_1 - \tilde{\mu}_2|$ by a factor which is proportional to variance
- Have samples $\boldsymbol{z_1},\ldots,\boldsymbol{z_n}$. Sample mean is $\mu_z = \frac{1}{n}\sum_{i=1}^{n} z_i$
- Define their *scatter* as

$$s = \sum_{i=1}^{n} (z_i - \mu_z)^2$$

- Thus scatter is just sample variance multiplied by $\boldsymbol{n}$
  - scatter measures the same thing as variance, the spread of data around the mean
  - scatter is just on different scale than variance

larger scatter:  *smaller scatter:* 

Projected Scatter

- Fisher Solution: normalize $|\tilde{\mu}_1 - \tilde{\mu}_2|$ by scatter
- Let $\boldsymbol{y_i} = \boldsymbol{v^t x_i}$, i.e. $\boldsymbol{y_i}$'s are the projected samples
- Scatter for projected samples of class 1 is

$$\tilde{s}_1^2 = \sum_{y_i \in Class\ 1} (y_i - \tilde{\mu}_1)^2$$

- Scatter for projected samples of class 2 is

$$\tilde{s}_2^2 = \sum_{y_i \in Class\ 2} (y_i - \tilde{\mu}_2)^2$$

Fisher Discriminant

- We need to normalize by both scatter of class 1 and scatter of class 2
- Thus Fisher linear discriminant is to project on line in the direction ***v*** which maximizes

$$\boldsymbol{J}(\boldsymbol{v}) = \frac{\overbrace{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}^{\text{want projected means are far from each other}}}{\tilde{\mathbf{s}}_1^2 + \tilde{\mathbf{s}}_2^2}$$

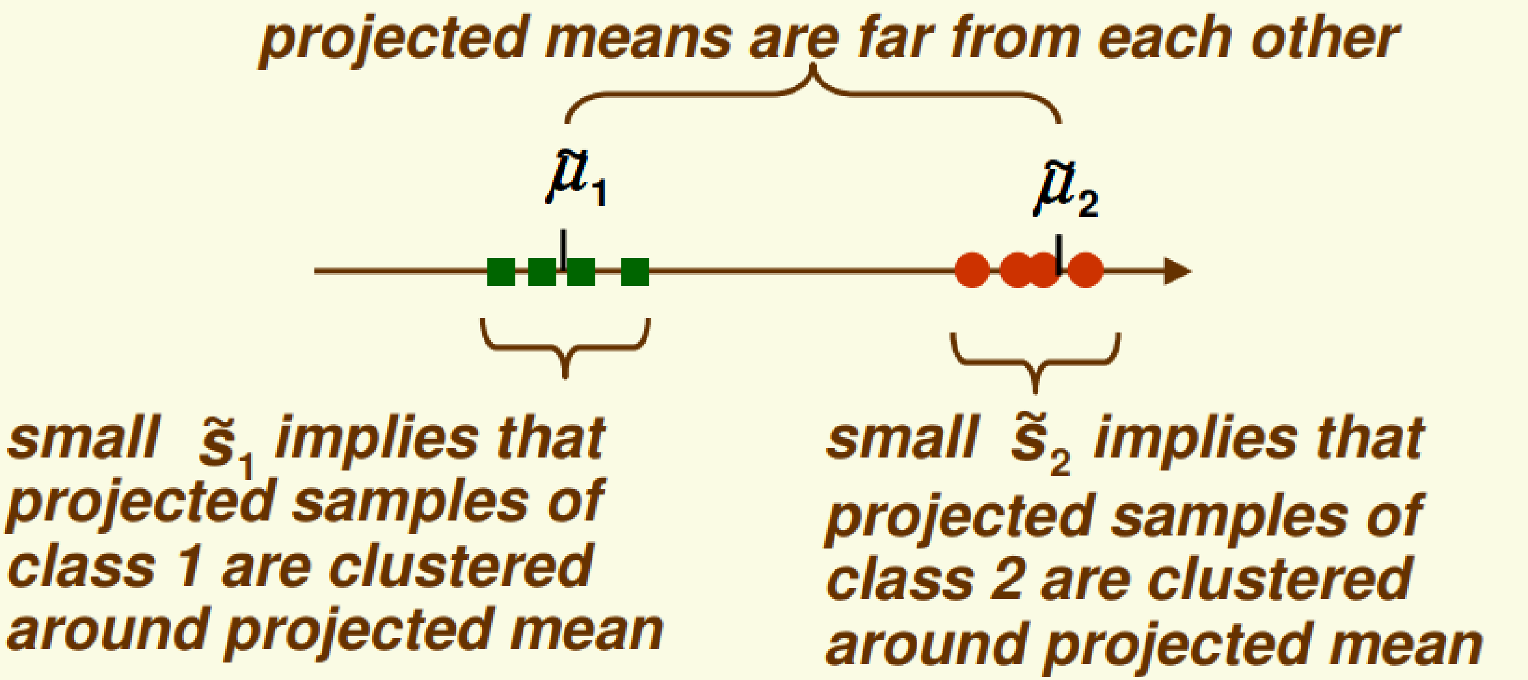
want scatter in class 1 is as small as possible, i.e. samples of class 1 cluster around the projected mean $\tilde{\mu}_1$

want scatter in class 2 is as small as possible, i.e. samples of class 2 cluster around the projected mean $\tilde{\mu}_2$

Fisher Discriminant

$$J(v) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

- If we find v which makes $J(v)$ large, we are guaranteed that the classes are well separated



Fisher Discriminant

$$J(v) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

- All we need to do now is to express ***J*** explicitly as a function of ***v*** and maximize it
  - straightforward but need linear algebra and Calculus
- Define the separate class scatter matrices $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ for classes 1 and 2. These measure the scatter of original samples $\boldsymbol{x_i}$ (before projection)

$$\boldsymbol{S}_1 = \sum_{x_i \in Class\ 1} (\boldsymbol{x}_i - \mu_1)(\boldsymbol{x}_i - \mu_1)^t$$

$$\boldsymbol{S}_2 = \sum_{x_i \in Class\ 2} (\boldsymbol{x}_i - \mu_2)(\boldsymbol{x}_i - \mu_2)^t$$

Deriving for LDA

- Now define the ***within*** the class scatter matrix

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

- Recall that $\tilde{s}_1^2 = \sum_{y_i \in Class\ 1} (y_i - \tilde{\mu}_1)^2$

- Using $y_i = v^t x_i$ and $\tilde{\mu}_1 = v^t \mu_1$

$$\begin{aligned}
\tilde{s}_1^2 &= \sum_{y_i \in Class\ 1} \left(v^t x_i - v^t \mu_1\right)^2 \\
&= \sum_{y_i \in Class\ 1} \left(v^t (x_i - \mu_1)\right)^t \left(v^t (x_i - \mu_1)\right) \\
&= \sum_{y_i \in Class\ 1} \left((x_i - \mu_1)^t v\right)^t \left((x_i - \mu_1)^t v\right) \\
&= \sum_{y_i \in Class\ 1} v^t (x_i - \mu_1)(x_i - \mu_1)^t v = v^t S_1 v
\end{aligned}$$

Deriving for LDA

- Similarly $\tilde{\mathbf{s}}_2^2 = \mathbf{v}^t \mathbf{S}_2 \mathbf{v}$
- Therefore $\tilde{\mathbf{s}}_1^2 + \tilde{\mathbf{s}}_2^2 = \mathbf{v}^t \mathbf{S}_1 \mathbf{v} + \mathbf{v}^t \mathbf{S}_2 \mathbf{v} = \mathbf{v}^t \mathbf{S}_W \mathbf{v}$
- Define between the class scatter matrix

$$\mathbf{S}_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t$$

- $\mathbf{S}_B$ measures separation between the means of two classes (before projection)
- Let's rewrite the separations of the projected means

$$\begin{aligned}(\tilde{\mu}_1 - \tilde{\mu}_2)^2 &= \left(\mathbf{v}^t \mu_1 - \mathbf{v}^t \mu_2\right)^2 \\ &= \mathbf{v}^t (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t \mathbf{v} \\ &= \mathbf{v}^t \mathbf{S}_B \mathbf{v}\end{aligned}$$

Deriving for LDA

- Thus our objective function can be written:

$$J(v) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{v^t S_B v}{v^t S_W v}$$

- Minimize $J(v)$ by taking the derivative w.r.t. v and setting it to 0

$$\frac{d}{dv} J(v) = \frac{\left(\frac{d}{dv} v^t S_B v\right) v^t S_W v - \left(\frac{d}{dv} v^t S_W v\right) v^t S_B v}{\left(v^t S_W v\right)^2}$$

$$= \frac{(2 S_B v) v^t S_W v - (2 S_W v) v^t S_B v}{\left(v^t S_W v\right)^2} = 0$$

Deriving for LDA

- Need to solve $v^t S_W v (S_B v) - v^t S_B v (S_W v) = 0$

$$\Rightarrow \frac{v^t S_W v (S_B v)}{v^t S_W v} - \frac{v^t S_B v (S_W v)}{v^t S_W v} = 0$$

$$\Rightarrow S_B v - \underbrace{\frac{v^t S_B v (S_W v)}{v^t S_W v}}_{= \lambda} = 0$$

$$\Rightarrow \underbrace{S_B v = \lambda S_W v}_{\textit{generalized eigenvalue problem}}$$

The Final Step

$$S_B v = \lambda S_W v$$

- If S_W has full rank (the inverse exists), can convert this to a standard eigenvalue problem

$$\boxed{S_W^{-1} S_B v = \lambda v}$$

- But $S_B x$ for any vector x, points in the same direction as $\mu_1 - \mu_2$

$$S_B x = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t x = (\mu_1 - \mu_2)\underbrace{\left((\mu_1 - \mu_2)^t x\right)}_{\alpha} = \alpha(\mu_1 - \mu_2)$$

- Thus can solve the eigenvalue problem immediately

$$\boxed{v = S_W^{-1}(\mu_1 - \mu_2)}$$

$$S_W^{-1} S_B \underbrace{\left[S_W^{-1}(\mu_1 - \mu_2)\right]}_{v} = S_W^{-1}\left[\alpha(\mu_1 - \mu_2)\right] = \underbrace{\alpha}_{\lambda}\underbrace{\left[S_W^{-1}(\mu_1 - \mu_2)\right]}_{v}$$

Lets Try a Problem

- **Compute the Linear Discriminant projection for the following two-dimensional dataset**
  - $X1=(x_1,x_2)=\{(4,1),(2,4),(2,3),(3,6),(4,4)\}$
  - $X2=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$

Solution

- **Compute the Linear Discriminant projection for the following two-dimensional dataset**
  - $X1=(x_1,x_2)=\{(4,1),(2,4),(2,3),(3,6),(4,4)\}$
  - $X2=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$
- **SOLUTION (by hand)**
  - The class statistics are:

$$S_1 = \begin{bmatrix} 0.80 & -0.40 \\ -0.40 & 2.60 \end{bmatrix};\ S_2 = \begin{bmatrix} 1.84 & -0.04 \\ -0.04 & 2.64 \end{bmatrix}$$

$$\mu_1 = [3.00 \quad 3.60]; \qquad \mu_2 = [8.40 \quad 7.60]$$

  - The within- and between-class scatter are

$$S_B = \begin{bmatrix} 29.16 & 21.60 \\ 21.60 & 16.00 \end{bmatrix};\ S_W = \begin{bmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{bmatrix}$$

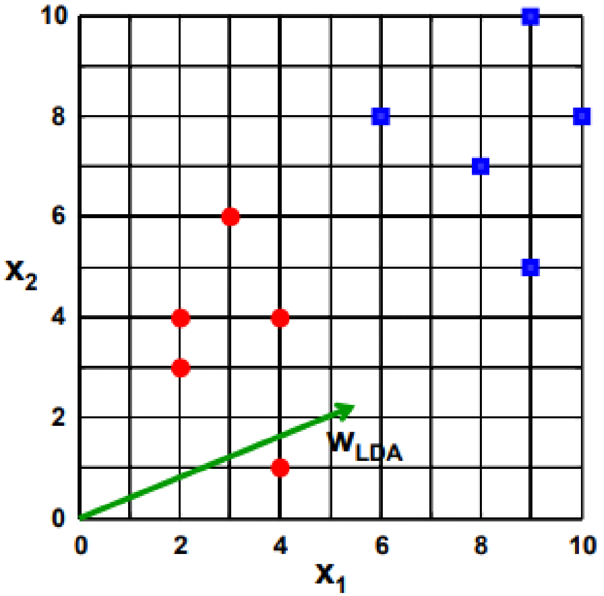
  - The LDA projection is then obtained as the solution of the generalized eigenvalue problem

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

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

  - Or directly by

$$w^* = S_W^{-1}(\mu_1 - \mu_2) = [-0.91 \quad -0.39]^T$$



Questions ?