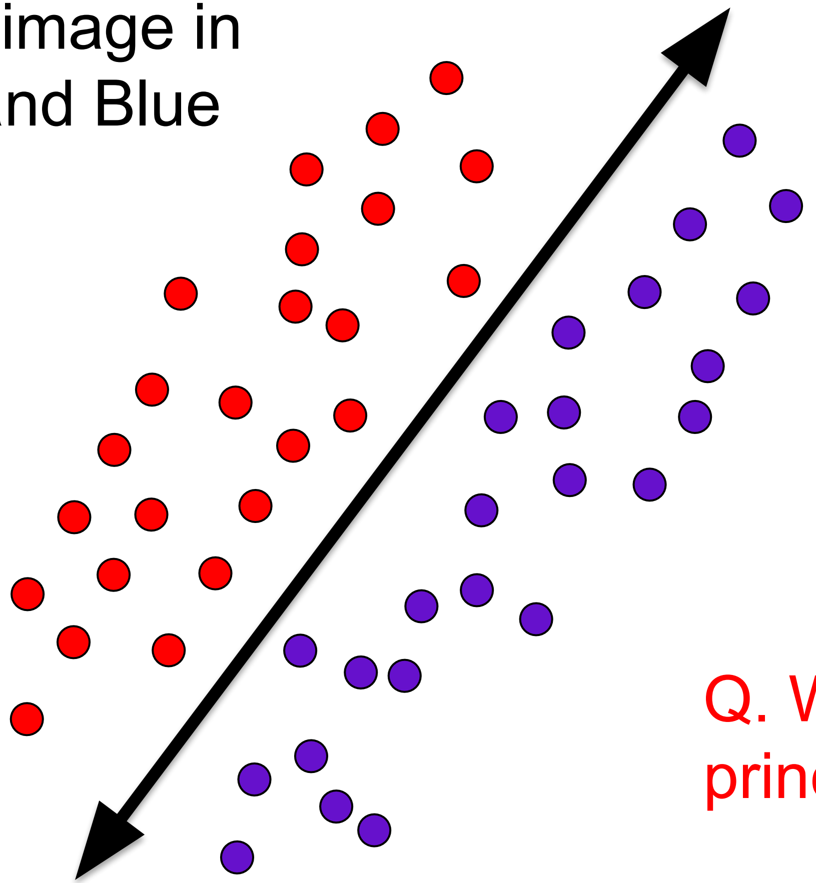
Recitation 8

LDA

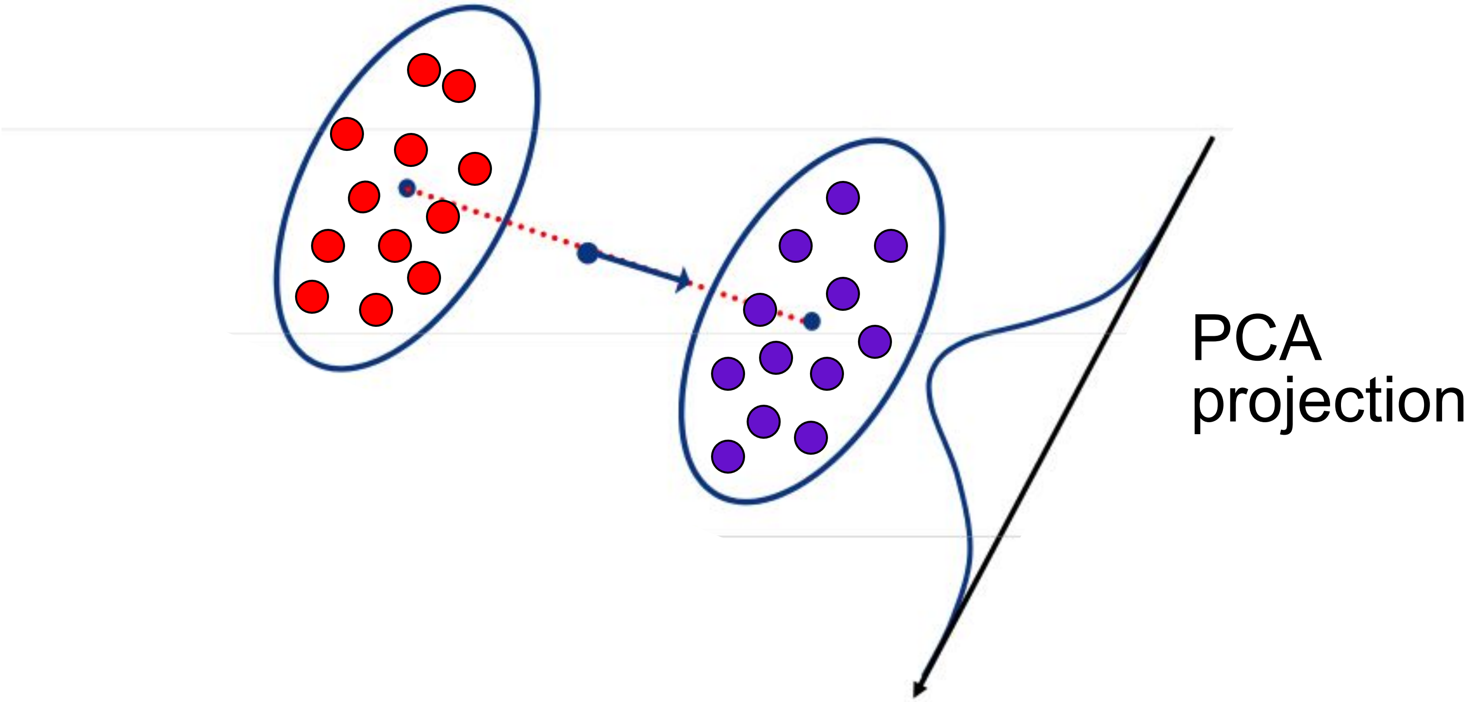
Let’s say that we this hypothetical 2-dimensional feature space.

Here I am showing each image in this feature space. Red and Blue are the two classes.

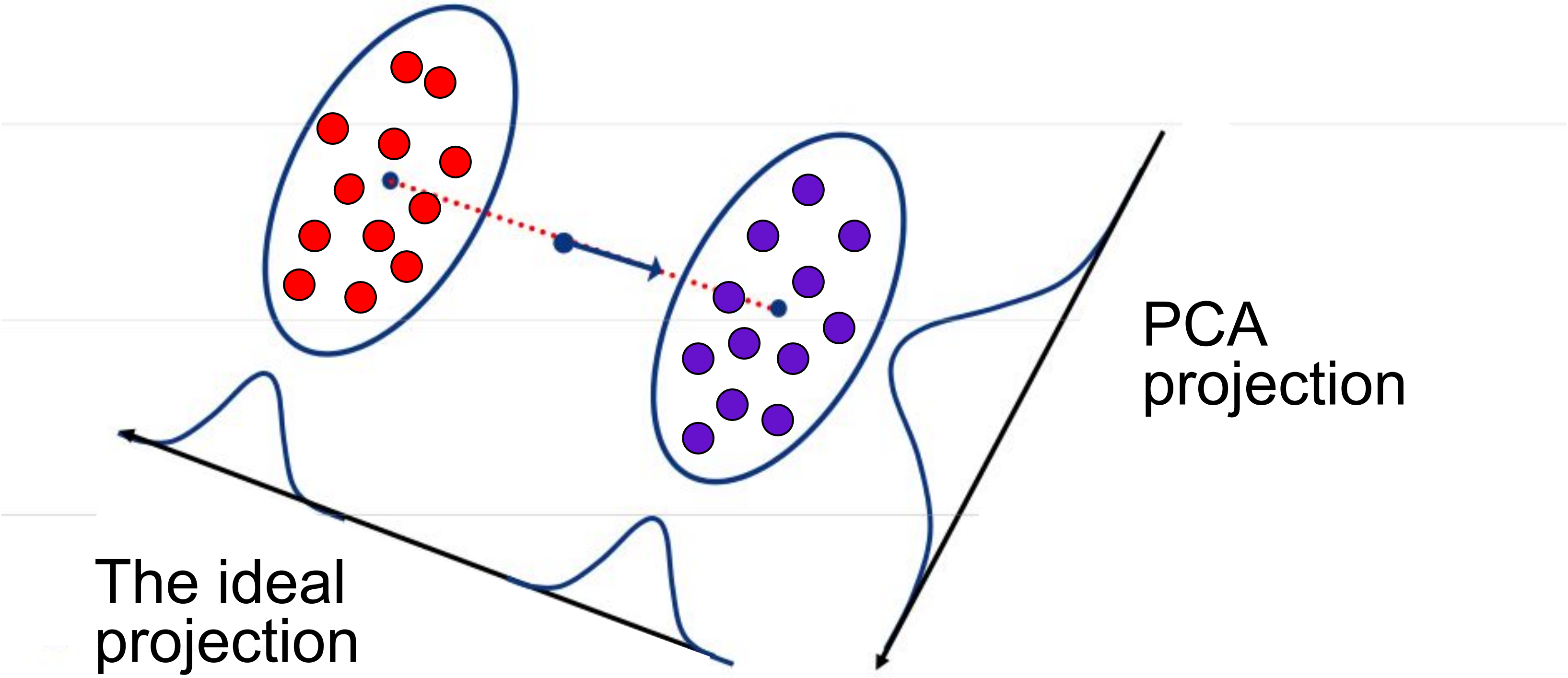


Q. Which direction will is the first principle component?

PCA can project the data such that it will become harder to separate the two classes

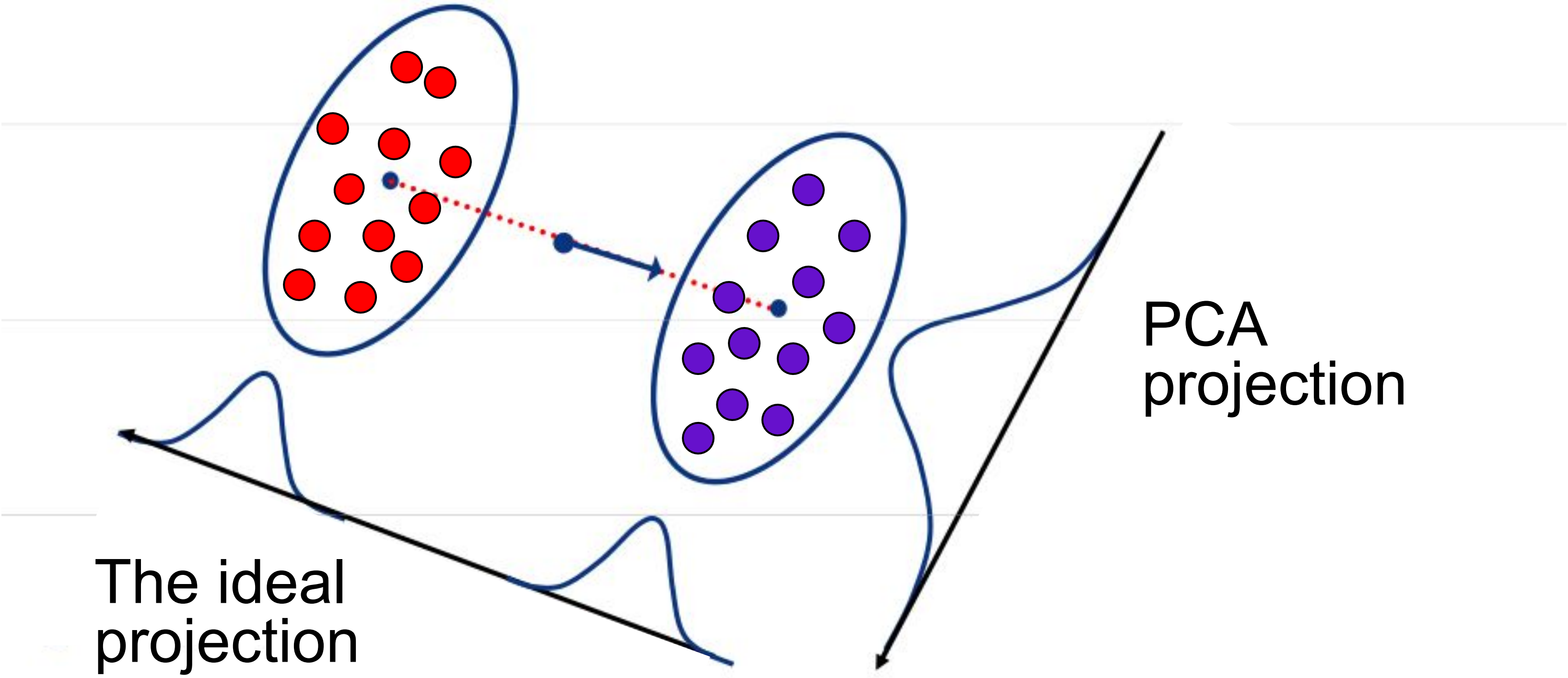


The ideal projection should make it easy to differentiate between images from two classes



Fischer’s Linear Discriminant Analysis (LDA)

- Goal: find the best separation between two classes



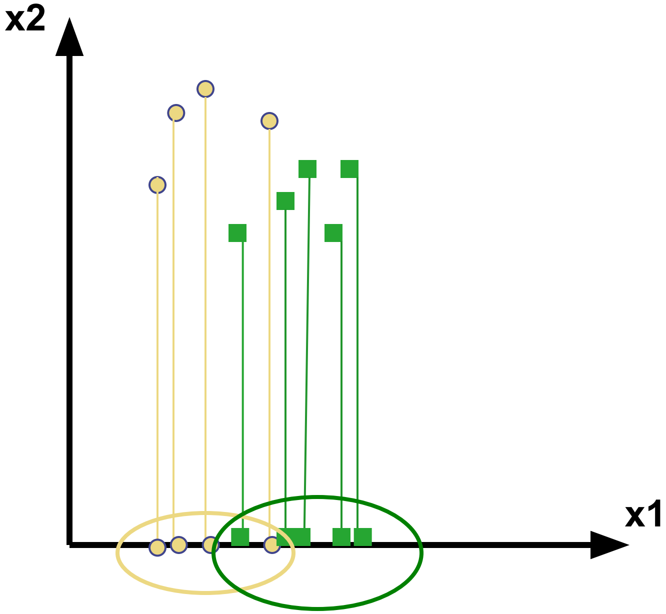
Difference between PCA and LDA

- PCA preserves maximum variance
  - PCA maximizes our ability to reconstruct each image
  - Doesn’t help us find the best projection for classification

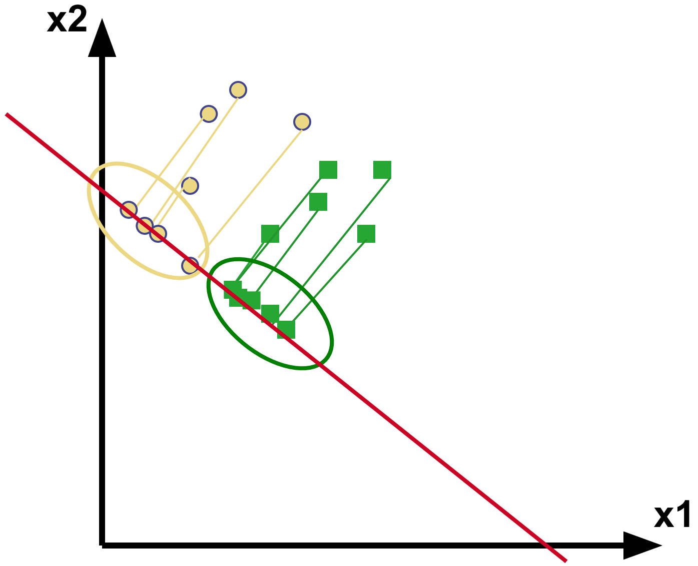
- LDA preserves discrimination (difference between categories)
  - Find projection that **maximizes scatter** between classes and **minimizes scatter** within classes

How LDA reduces dimentionality

- Using two classes as example:

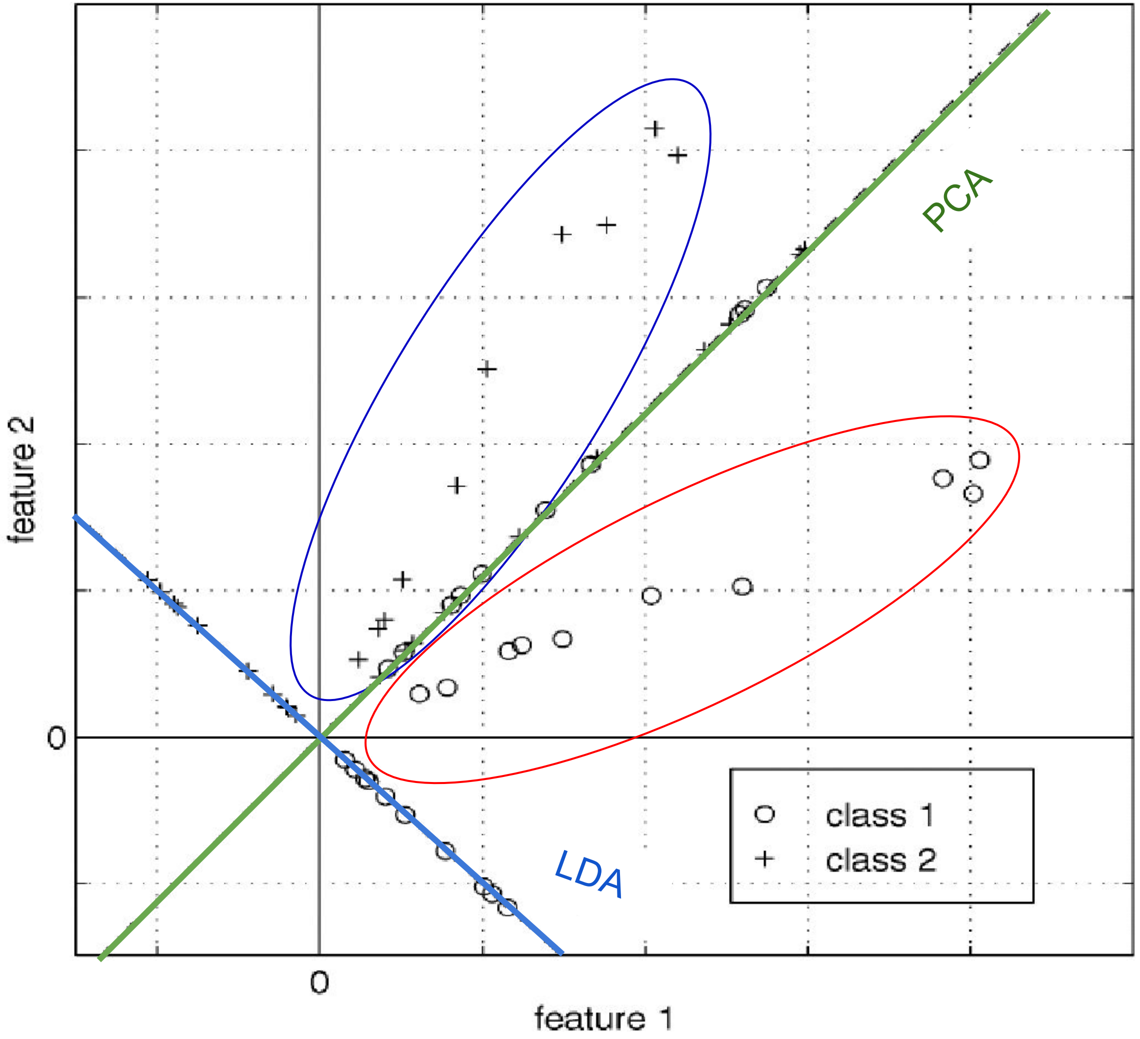


Poor Projection



Good

Basic intuition: PCA vs. LDA



First, let's calculate the per category statistics

- We want to learn a dimension reduction **projection W** such that the projection converts all image features **x** to a lower dimensional space:

$$z = w^T x \qquad z \in \mathbf{R}^m \quad x \in \mathbf{R}^n$$

- First, let's calculate the **per class** means be:

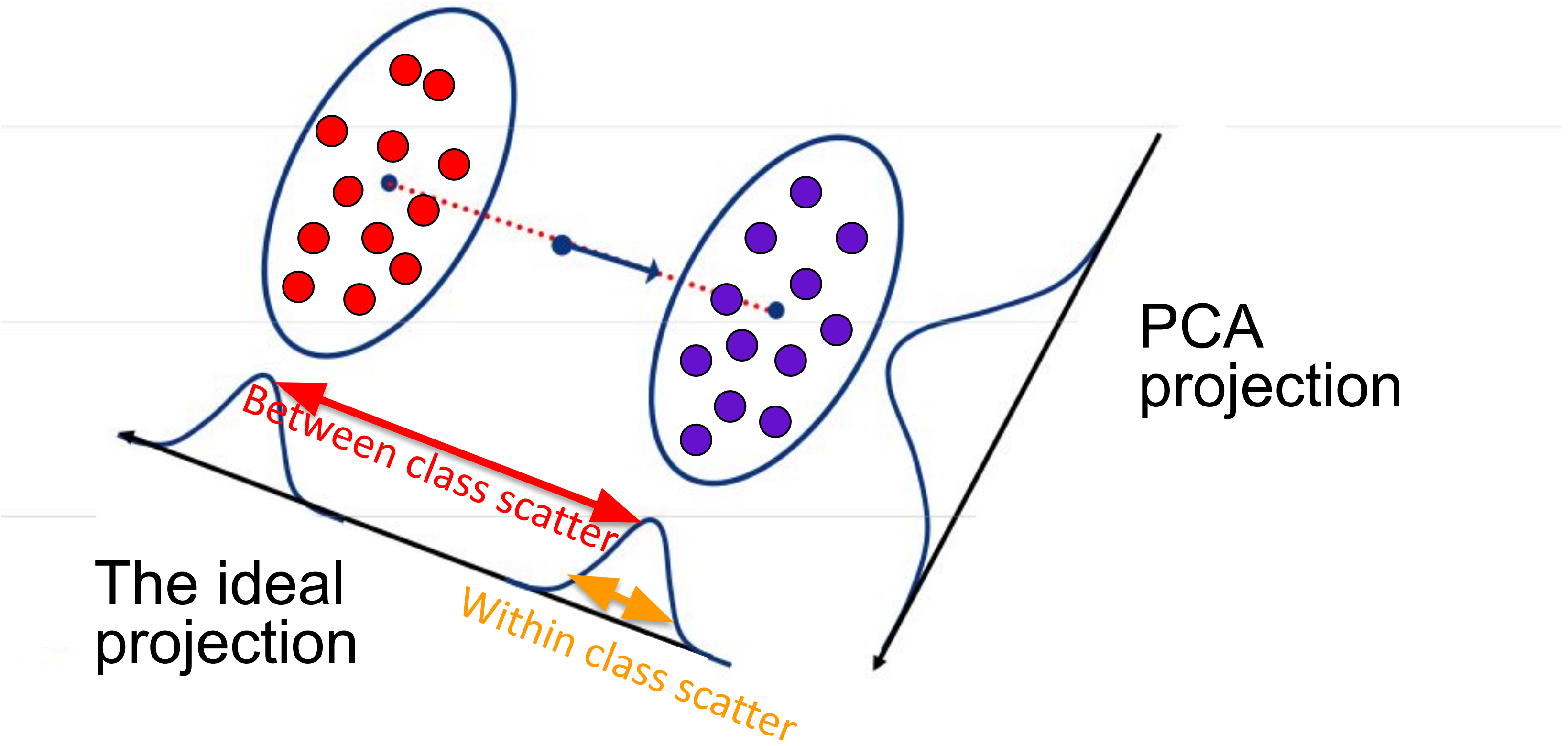
$$\mu_i = E_{X|Y}[X|Y = i]$$

- And the **per class** covariance matrices are:

$$C_i = [(X_i - \mu_i)(X_i - \mu_i)^T|Y = i]$$

Using the per class means and covariance, we want to minimize the following objective:

We want a projection that maximizes: $$J(w) = \max \frac{between\ \ class\ \ scatter}{within\ \ class\ \ scatter}$$



What does J(w) look like when we only have 2 classes

The following objective function:

$$J(w) = \frac{between\ \ class\ \ scatter}{within\ \ class\ \ scatter}$$

Can be written as

$$J(w) = \frac{|E_{Z|Y}[Z|Y=1] - E_{Z|Y}[Z|Y=0]|^2}{\mathrm{var}[Z|Y=1] + \mathrm{var}[Z|Y=0]}$$

denom: write out each Z as the reduction of x using w, and we have the square as it's the variance equation

LDA with 2 variables

- **Numerator:** We can write the **between** class scatter as:

$$
\begin{aligned}
|E_{Z|Y}[Z|Y=1] - E_{Z|Y}[Z|Y=0]|^2 &= |w^T(\mu_1 - \mu_0)|^2 \\
&= w^T(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T w
\end{aligned}
$$

- **Each part of Denominator:** Also, the **within** class scatter becomes:

$$
\begin{aligned}
var[Z|Y=i] &= E_{Z|Y}[w^T(x-\mu_i)^2|Y=i] \\
&= E_{Z|Y}[w^T(x-\mu_i)(x-\mu_i)^T w|Y=i] \\
&= w^T C_i w
\end{aligned}
$$

LDA with 2 variables

- We can plug in these scatter values to our objective function:

$$J(w) = \frac{w^T(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T w}{w^T C_1 w + w^T C_0 w}$$

LDA with 2 variables

- We can plug in these scatter values to our objective function:

$$J(w) = \frac{w^T(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T w}{w^T C_1 w + w^T C_0 w}$$
$$= \frac{w^T(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T w}{w^T (C_1 + C_0) w}$$

LDA with 2 variables

- We can plug in these scatter values to our objective function:

$$S_B = (\mu_1 - \mu_0)(\mu_1 - \mu_0)^T$$

Between class scatter

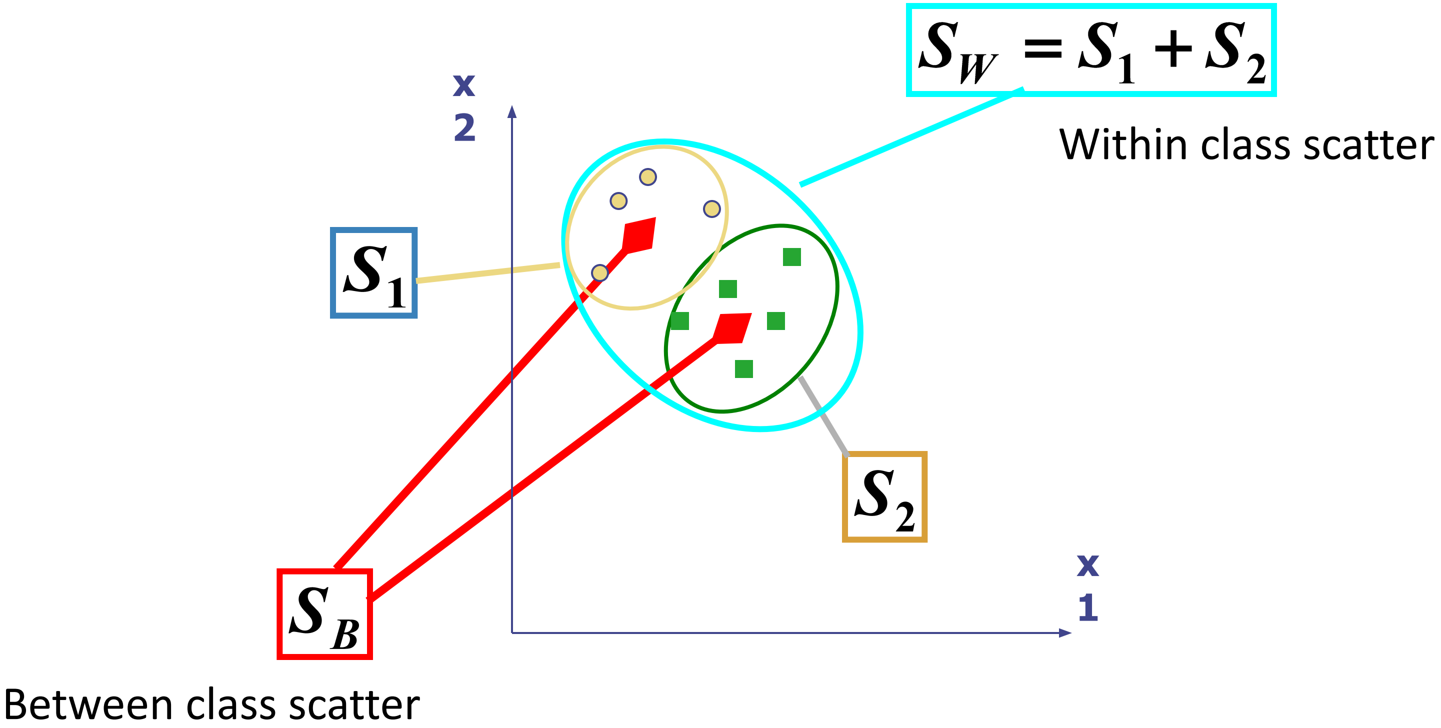
$$J(w) = \frac{w^T(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T w}{w^T C_1 w + w^T C_0 w}$$

$$= \frac{w^T \boxed{(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T} w}{w^T \boxed{(C_1 + C_0)} w}$$

$$S_W = (C_1 + C_0)$$

Within class scatter

Visualizing S_w and S_B



Linear Discriminant Analysis (LDA)

- Maximizing the ratio

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Linear Discriminant Analysis (LDA)

- Maximizing the ratio

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

- Is equivalent to maximizing the numerator while keeping the denominator constant, i.e.

$$\max_w w^T S_B w \quad \text{subject to} \quad w^T S_W w = K$$

Linear Discriminant Analysis (LDA)

- Maximizing the ratio
$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

- Is equivalent to maximizing the numerator while keeping the denominator constant, i.e.
$$\max_w w^T S_B w \quad \text{subject to} \quad w^T S_W w = K$$

- And can be accomplished using Lagrange multipliers, where we define the Lagrangian as
$$L = w^T S_B w - \lambda\left(w^T S_W w - K\right)$$

- And maximize with respect to both w and λ

Linear Discriminant Analysis (LDA)

- Setting the gradient of $L = w^T(S_B - \lambda S_W)w + \lambda K$ to 0

- Taking the derivative respect to **w** to find the maximum:

$$\nabla_w L = 2(S_B - \lambda S_W)w = 0$$

Linear Discriminant Analysis (LDA)

- Setting the gradient of $L = w^T\left(S_B - \lambda S_W\right)w + \lambda K$ to 0

- Taking the derivative respect to **w** to find the maximum:

$$\nabla_w L = \boxed{2\left(S_B - \lambda S_W\right)w = 0}$$

- This is maximized when $\boxed{S_B w = \lambda S_W w}$

Linear Discriminant Analysis (LDA)

- Setting the gradient of $L = w^T(S_B - \lambda S_W)w + \lambda K$ to 0

- Taking the derivative respect to **w** to find the maximum:

$$\nabla_w L = \boxed{2(S_B - \lambda S_W)w = 0}$$

- This is maximized when $\boxed{S_B w = \lambda S_W w}$

- The solution is easy when S_w has an inverse: $S_W^{-1} = (C_1 + C_0)^{-1}$

Linear Discriminant Analysis (LDA)

$$\boxed{S_B w = \lambda S_W w}$$

If an inverse for S_W exists:

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

We want to find the optimal w.

Q. What does this look like?

Linear Discriminant Analysis (LDA)

$$\boxed{S_B w = \lambda S_W w}$$

If an inverse for S_W exists:

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

The solution is the eigenvector of $S_W^{-1} S_B$ **corresponding to the largest eigenvalue**

LDA with C classes

Same as when C=2. Except S_W and S_B now include all classes.

$$S_W = \sum_i C_i$$

$$S_B = \sum_i \sum_{j \neq i} (\mu_i - \mu_j)^2$$

PCA vs. LDA

- PCA exploits the max scatter of the training images in face space
- LDA attempt to maximise the **between class scatter**, while minimising the **within class scatter**.