Linear Discriminant Analysis

Instructor: Junghye Lee

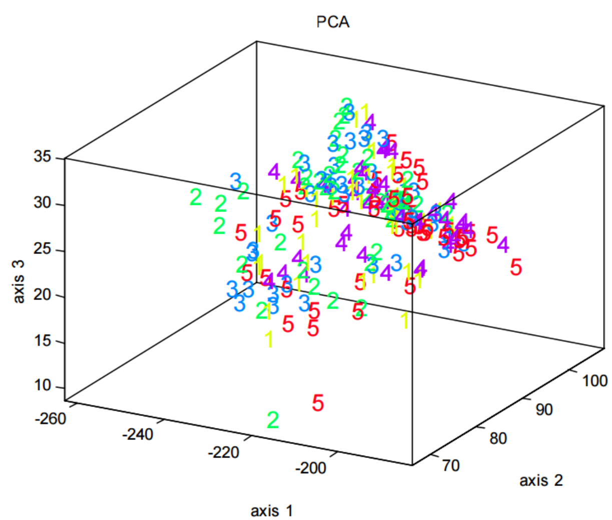
Department of Industrial Engineering

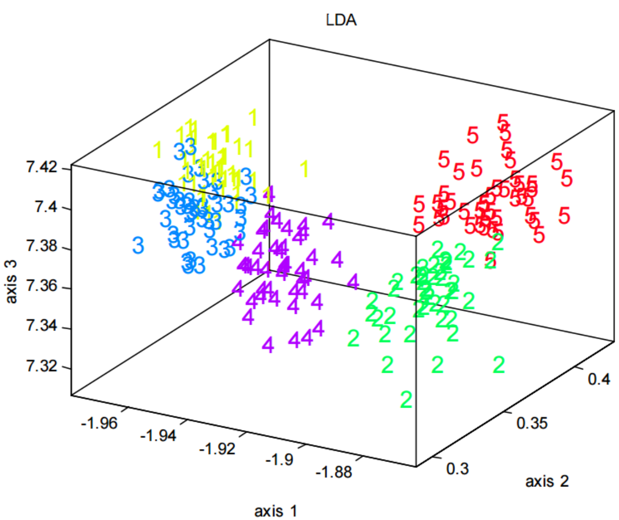
junghyelee@unist.ac.kr

Contents

Introduction

- Dimensionality reduction is a crucial concept in machine learning and data classification.
- The most famous example of dimensionality reduction is Principal Component Analysis (PCA):
  - Is an unsupervised method, so it doesn't include label information.
  - Searches for the directions the data have the largest variance.
  - There are difficulty issues with the number of principal components to choose.





Introduction

- Discriminant analysis methods can be good candidates to address such problems.
  - These methods are supervised, so they include label information.
  - The goal is to find directions on which the data is best separable.
- One of the very well-known discriminant analysis method is the Linear Discriminant Analysis (LDA).

Linear Discriminant Analysis - 2 classes

- The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible
  - Assume we have a *D*-dimensional *N* samples $\{\boldsymbol{x}^{(\mathbf{1})}, \boldsymbol{x}^{(\mathbf{2})}, \dots, \boldsymbol{x}^{(\boldsymbol{N})}\}$, N_1 of which belong to class ω_1, and N_2 to class ω_2
  - We seek to obtain a scalar vector $\boldsymbol{y}$ by projecting the samples $\boldsymbol{x}$ onto a line

$$y_j = \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x}_{\boldsymbol{j}} \text{ where } j = 1, \dots, N$$

or

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}$$

What is the dimension of x, X and $\boldsymbol{w}$?

Illustration

Sample #	AGE	gender	WAIST	BP_HIGH	BP_LWST	BLDS	Diabetes
1	44	1	86	120	80	75	0
2	56	0	84	110	70	151	0
3	38	1	78	103	61	82	1
4	60	1	88	130	77	153	1
5	28	1	92	128	77	101	0
6	42	1	94	134	85	91	0

parameters

$\boxed{\boldsymbol{w}^T}\boldsymbol{x}_1 =$

Projection to ***w***

scalar

$$(w_1, w_2, w_3, w_4, w_5, w_6) \times \begin{pmatrix} 44 \\ 1 \\ 86 \\ 120 \\ 80 \\ 75 \end{pmatrix} = \boxed{w_1 \cdot 44 + w_2 \cdot 1 + w_3 \cdot 86 + w_4 \cdot 120 + w_5 \cdot 80 + w_6 \cdot 75} = y_1$$

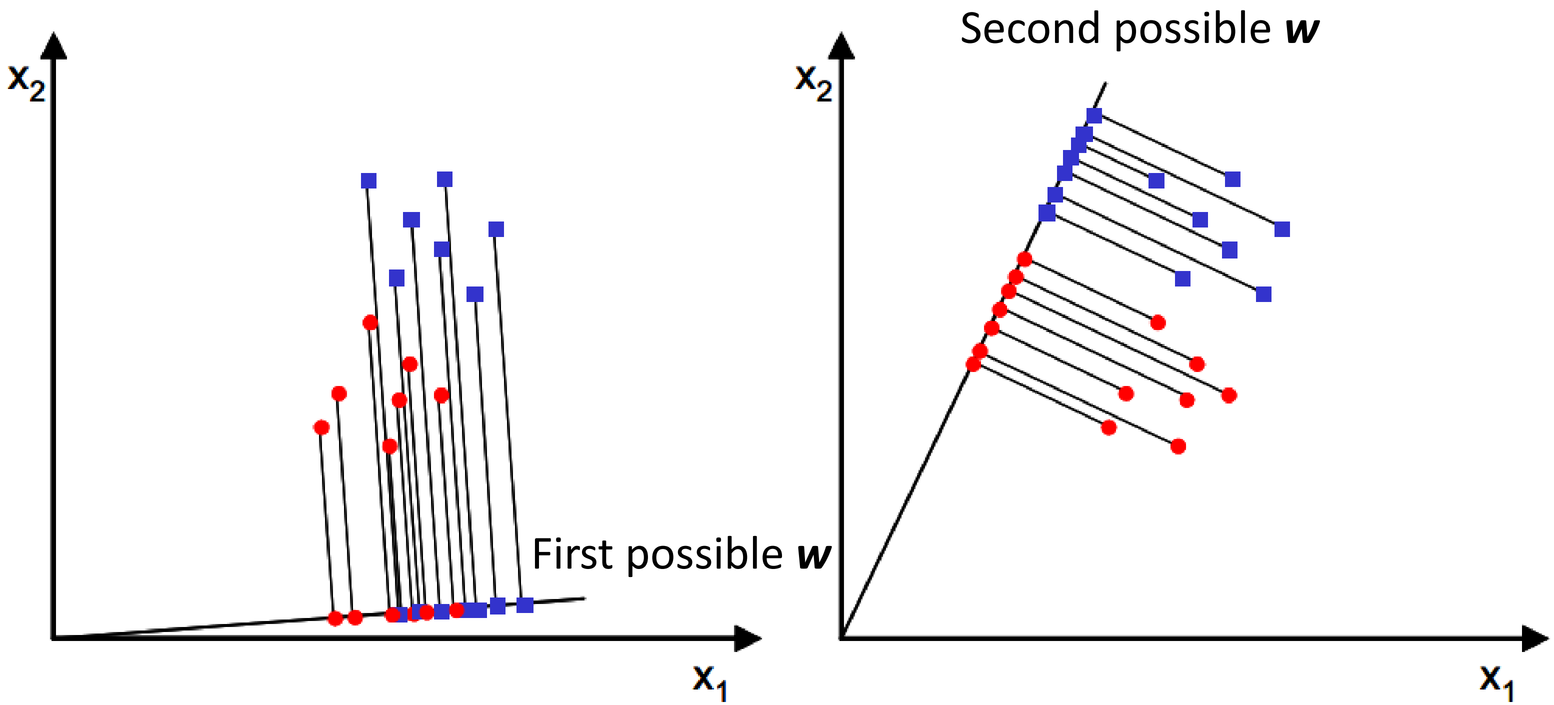
$\boxed{\boldsymbol{w}^T}\boldsymbol{x}_2 =$

$$(w_1, w_2, w_3, w_4, w_5, w_6) \times \begin{pmatrix} 56 \\ 0 \\ 84 \\ 110 \\ 70 \\ 151 \end{pmatrix} = \boxed{w_1 \cdot 56 + w_2 \cdot 0 + w_3 \cdot 84 + w_4 \cdot 110 + w_5 \cdot 70 + w_6 \cdot 151} = y_2$$

$\vdots$

Linear Discriminant Analysis - 2 classes

- Of all the possible lines we would like to select the one that maximizes the separability of the scalars
  - This is illustrated for the two-dimensional case in the following figures.



Which one is better?

Linear Discriminant Analysis – 2 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 $\boldsymbol{x}$ and $\boldsymbol{y}$ feature space is

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

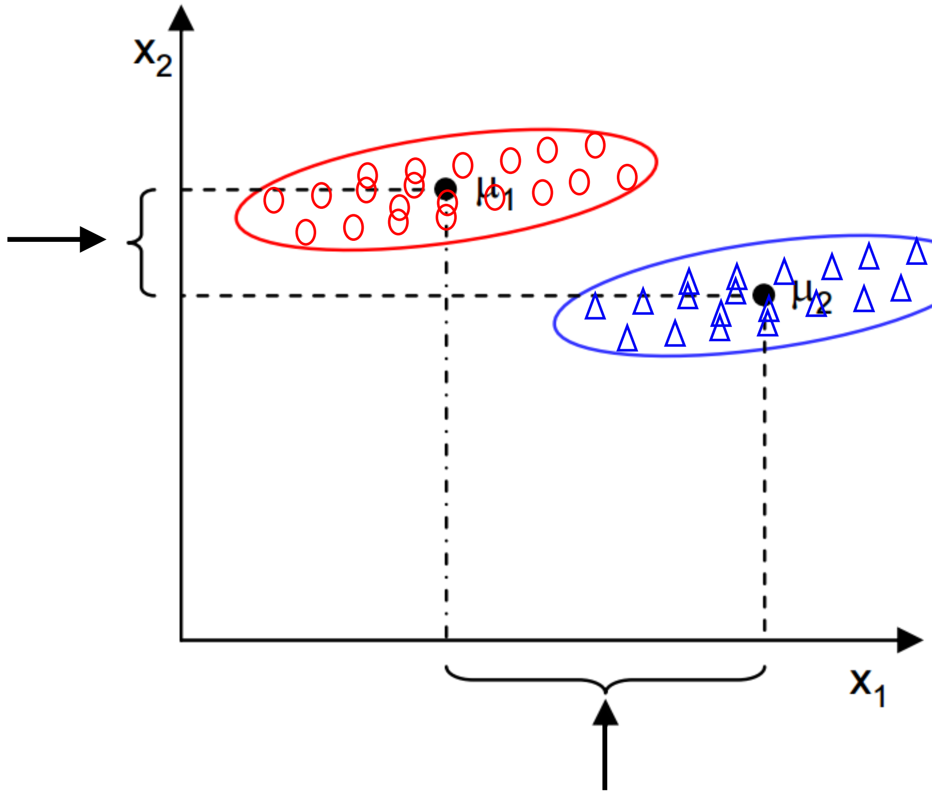
  - ω_i, where $i = 1, 2.$ i.e., $\omega_1 = class\ 1, \omega_2 = class\ 2$
  - We could then choose the distance between the projected means as our objective function

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

Linear Discriminant Analysis – 2 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

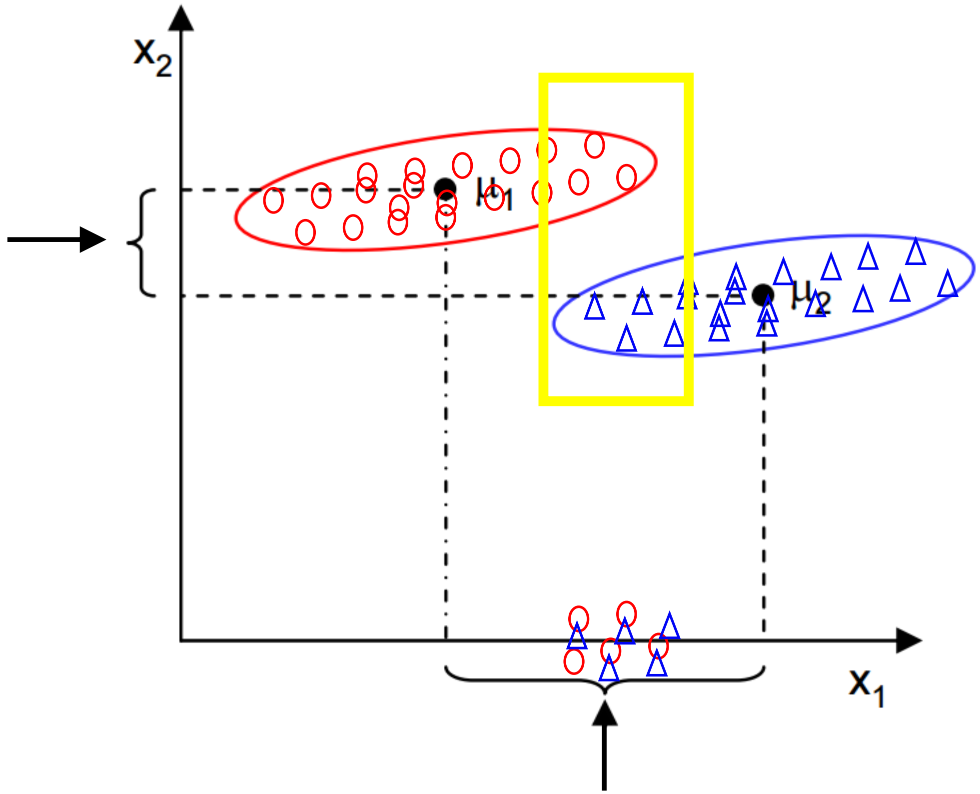


This axis has a larger distance between means

Linear Discriminant Analysis – 2 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

Linear Discriminant Analysis – 2 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 scatter***
  - For each class we define the scatter, an equivalent of the variance, as

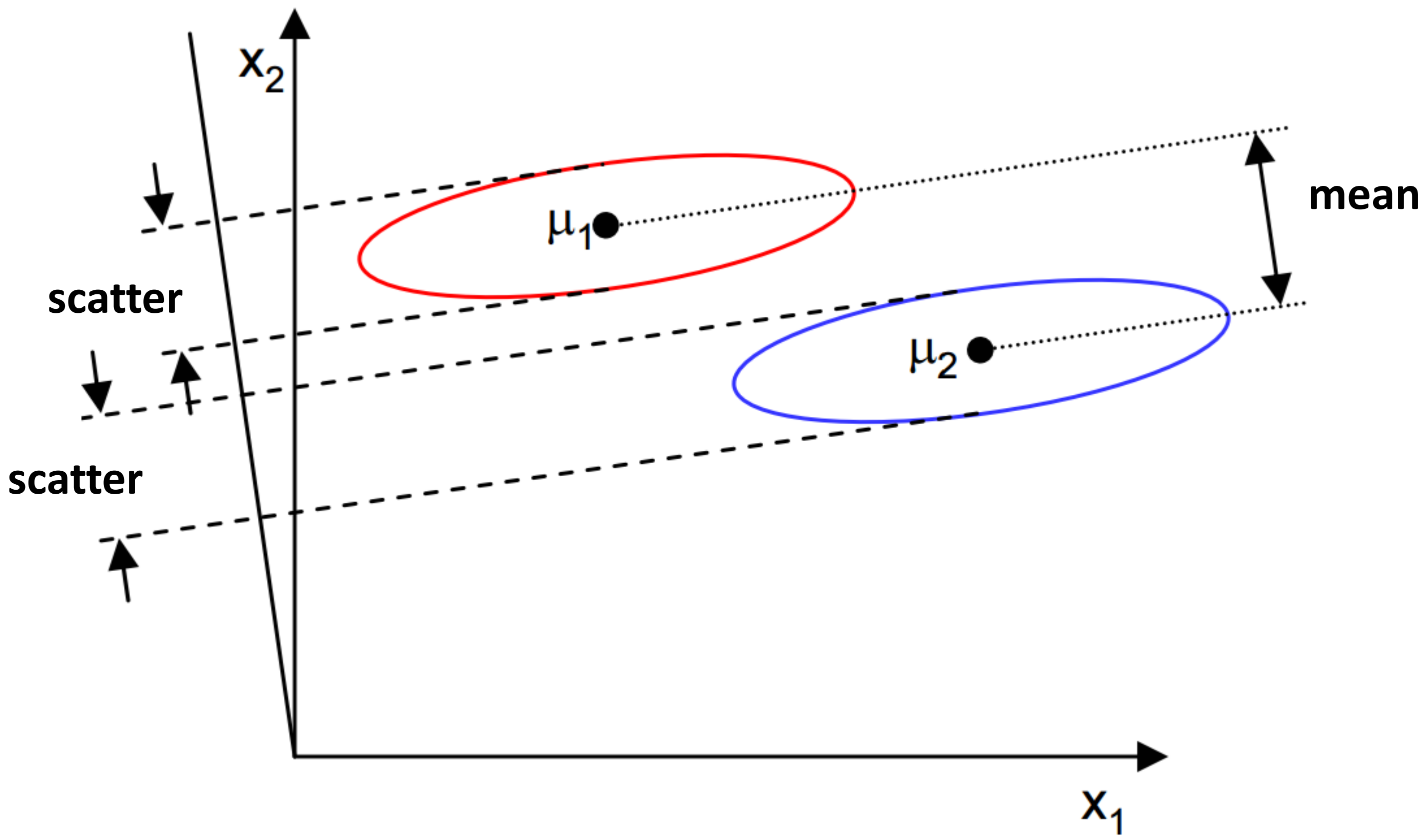
$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2$$

    - where the quantity $(\tilde{s}_1^2 + \tilde{s}_2^2)$ is called the within-class scatter of the projected examples
  - The Fisher linear discriminant is defined as the linear function $\boldsymbol{w^T x}$ that maximizes the criterion function

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

Linear Discriminant Analysis – 2 classes

– Therefore, we will be **looking for a projection (*w*)** where **examples from the same class are projected very close to each other (*intra-class*)** and, at the same time, **the projected means between classes (*inter-class*)** are as farther apart as possible.



Linear Discriminant Analysis – 2 classes

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

– We define a measure of the scatter in multivariate feature space $\boldsymbol{x}$, which are scatter matrices

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

$$\boldsymbol{S_1} + \boldsymbol{S_2} = \boldsymbol{S_W}$$

- where S_W is called the within-class scatter matrix

Linear Discriminant Analysis – 2 classes

– The scatter of the projection $\boldsymbol{y}$ can then be expressed as a function of the scatter matrix in feature space $\boldsymbol{x}$

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

Linear Discriminant Analysis – 2 classes

– We can finally express the Fisher criterion in terms of $\boldsymbol{S_W}$ and $\boldsymbol{S_B}$ as

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

– To find the maximum of $\boldsymbol{J}(\boldsymbol{w})$ we derive and equate to zero

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

$$[\boldsymbol{w^T S_W w}]\frac{d[\boldsymbol{w^T S_B w}]}{d\boldsymbol{w}} - [\boldsymbol{w^T S_B w}]\frac{d[\boldsymbol{w^T S_W w}]}{d\boldsymbol{w}} = 0$$

$$[\boldsymbol{w^T S_W w}]2\boldsymbol{S_B w} - [\boldsymbol{w^T S_B w}]2\boldsymbol{S_W w} = 0$$

Linear Discriminant Analysis – 2 classes

Eigenvalue decomposition

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

Generalized eigenvalue decomposition

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{B}\mathbf{v}$$

$$\det(\mathbf{A} - \lambda\mathbf{B}) = 0$$

– Dividing by $\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}$

$$\left[\frac{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}}\right]\boldsymbol{S}_B\boldsymbol{w} - \left[\frac{\boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}}\right]\boldsymbol{S}_W\boldsymbol{w} = 0$$

$$\boldsymbol{S}_B\boldsymbol{w} - \underbrace{\boldsymbol{J}}_{\lambda}\boldsymbol{S}_W\boldsymbol{w} = 0$$

$$\boldsymbol{S}_W^{-1}\boldsymbol{S}_B\boldsymbol{w} - \boldsymbol{J}\boldsymbol{w} = 0$$

to be full rank (invertible)

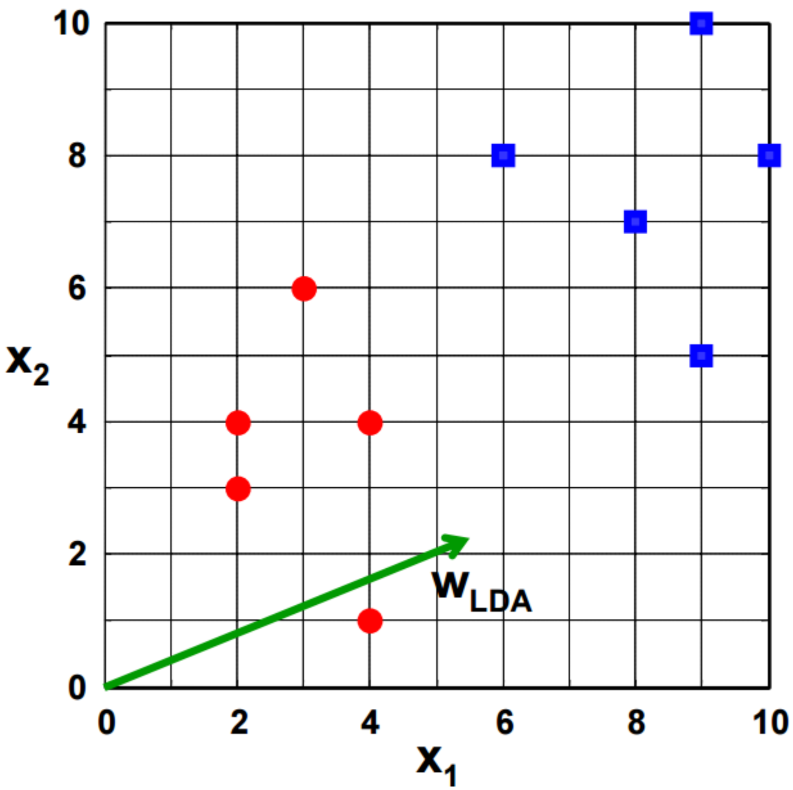
– Solving the **generalized eigenvalue problem** ($\boldsymbol{S}_W^{-1}\boldsymbol{S}_B\boldsymbol{w} = \boldsymbol{J}\boldsymbol{w}$) yields

$$\boldsymbol{w}^* = \operatorname{argmax}_{\boldsymbol{w}}\left\{\frac{\boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}}\right\} = \boldsymbol{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

- This is known as Fisher's Linear Discriminant (1936), although it is not a discriminant but rather a specific choice of direction for the projection of the data down to one dimension

LDA example

- Compute the Linear Discriminant projection for the following two-dimensional dataset
  - ω_1: $(x_1, x_2) = \{(4{,}1), (2{,}4), (2{,}3), (3{,}6), (4{,}4)\}$
  - ω_2: $(x_1, x_2) = \{(9{,}10), (6{,}8), (9{,}5), (8{,}7), (10{,}8)\}$



LDA example

- Solution (by hand)
  - The class statistics are:

$$\boldsymbol{S_1} = \begin{bmatrix} 4 & -2 \\ -2 & 13.2 \end{bmatrix}; \ \boldsymbol{S_2} = \begin{bmatrix} 9.2 & -0.2 \\ -0.2 & 13.2 \end{bmatrix}$$
$$\boldsymbol{\mu_1} = [3.00 \quad 3.60]^T; \ \ \boldsymbol{\mu_2} = [8.40 \quad 7.60]^T$$

  - The within- and between-class scatter are

$$S_B = \begin{bmatrix} 29.16 & 21.60 \\ 21.60 & 16.00 \end{bmatrix}; \ S_W = \begin{bmatrix} 13.2 & -2.2 \\ -2.2 & 26.4 \end{bmatrix}$$

LDA example

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

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

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

$$\begin{bmatrix} 2.38 - \lambda & 1.76 \\ 1.02 & 0.75 - \lambda \end{bmatrix} = 0$$

$$\lambda = 3.13$$

$$\begin{bmatrix} 2.38 & 1.76 \\ 1.02 & 0.75 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 3.13 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0.91 \\ 0.39 \end{bmatrix}$$

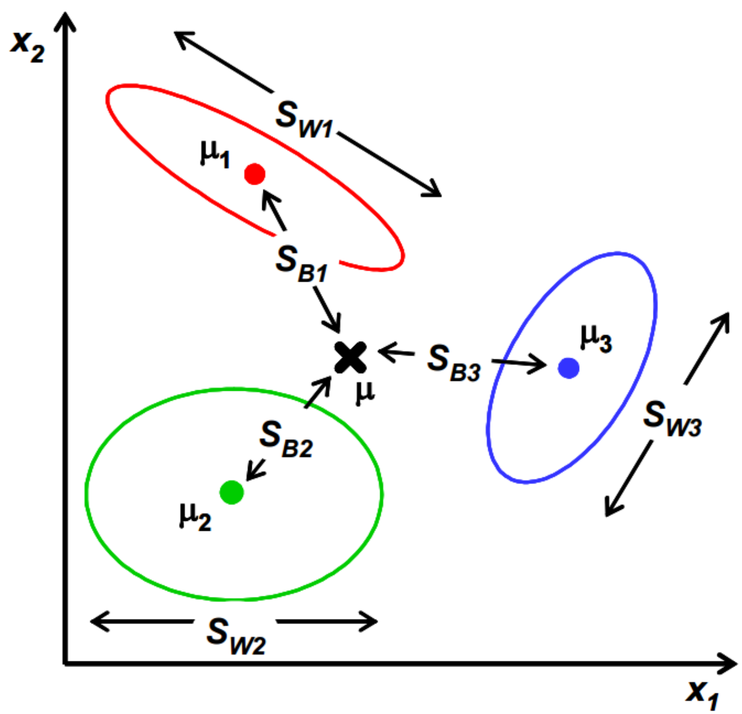
– Or directly by

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

Linear Discriminant Analysis – *C* classes

- Fisher's LDA generalizes very gracefully for *C*-class problems
  - Instead of one projection y, we will now seek (*C-1*) projections $[y_1, y_2, \ldots, y_{C-1}]$ by means of (*C-1*) projection vectors $\boldsymbol{w_i}$, which can be arranged by columns into a projection matrix $\boldsymbol{W} = [\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_{c-1}}]$:

$$y_i = \boldsymbol{w}_i^T \boldsymbol{x} \;\Rightarrow\; \boldsymbol{y} = \boldsymbol{W}^T \boldsymbol{x}$$



Linear Discriminant Analysis – *C* classes

- Derivation
  - The generalization of the within-class scatter is

$$\boldsymbol{S_W} = \sum_{i=1}^{C} \boldsymbol{S_i}$$

where $\boldsymbol{S_i} = \sum_{\boldsymbol{x}\in\omega_i}(\boldsymbol{x} - \boldsymbol{\mu_i})(\boldsymbol{x} - \boldsymbol{\mu_i})^{\boldsymbol{T}}$ and $\boldsymbol{\mu_i} = \frac{1}{N_i}\sum_{\boldsymbol{x}\in\omega_i}\boldsymbol{x}$

  - The generalization of the total scatter is

$$\boldsymbol{S_T} = \frac{1}{N}\sum_{\forall\boldsymbol{x}}(\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^{\boldsymbol{T}}$$

where $\boldsymbol{\mu} = \frac{1}{N}\sum_{\forall\boldsymbol{x}}\boldsymbol{x} = \frac{1}{N}\sum_i\sum_{\boldsymbol{x}\in\omega_i}N_i\boldsymbol{\mu_i}$

  - The total scatter is

$$\boldsymbol{S_T} = \boldsymbol{S_B} + \boldsymbol{S_W}$$

  - The generalization of the within-class scatter is

$$\boldsymbol{S_B} = \boxed{\sum_{i=1}^{C} N_i(\boldsymbol{\mu_i} - \boldsymbol{\mu})(\boldsymbol{\mu_i} - \boldsymbol{\mu})^T}$$

different from $\boldsymbol{S_B}$ for 2-class

Linear Discriminant Analysis – *C* classes

– Similarly, we define the mean vector and scatter matrices for the projected samples as

$$\widetilde{\boldsymbol{\mu}}_{\boldsymbol{i}} = \frac{1}{N_i} \sum_{y \in \omega_i} \boldsymbol{y} \qquad \tilde{S}_W = \sum_{i=1}^{C} \sum_{\boldsymbol{y} \in \omega_i} (\boldsymbol{y} - \widetilde{\boldsymbol{\mu}}_{\boldsymbol{i}})^2$$

$$\widetilde{\boldsymbol{\mu}} = \frac{1}{N} \sum_{\forall \boldsymbol{y}} \boldsymbol{y} \qquad \tilde{S}_B = \sum_{i=1}^{C} N_i (\widetilde{\boldsymbol{\mu}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{\mu}})(\widetilde{\boldsymbol{\mu}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{\mu}})^T$$

– From our derivation for the two-class problem, we can write

$$\widetilde{\boldsymbol{S}}_{\boldsymbol{W}} = \boldsymbol{W}^{\boldsymbol{T}} \boldsymbol{S}_{\boldsymbol{W}} \boldsymbol{W}$$
$$\widetilde{\boldsymbol{S}}_{\boldsymbol{B}} = \boldsymbol{W}^{\boldsymbol{T}} \boldsymbol{S}_{\boldsymbol{B}} \boldsymbol{W}$$

W is a matrix!
What is the dimension?

Linear Discriminant Analysis – *C* classes

– Recall that we are looking for a projection that maximizes the ratio of between-class to within-class scatter. Since the projection is no longer a scalar (it has *C-1* dimensions), we then use the determinant of the scatter matrices to obtain a scalar objective function:

$$J(\boldsymbol{W}) = \frac{|\widetilde{\boldsymbol{S}}_{\boldsymbol{B}}|}{|\widetilde{\boldsymbol{S}}_{\boldsymbol{W}}|} = \frac{|\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{S}_{\boldsymbol{B}}\boldsymbol{W}|}{|\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{S}_{\boldsymbol{W}}\boldsymbol{W}|} = \boldsymbol{trace}\left(\left(\widetilde{\boldsymbol{S}}_{\boldsymbol{W}}\right)^{-\boldsymbol{1}}\widetilde{\boldsymbol{S}}_{\boldsymbol{B}}\right)$$

– We will seek the projection matrix $\boldsymbol{W}^*$ that maximizes this ratio.

– It can be shown that the optimal projection matrix $\boldsymbol{W}^*$ is the one whose columns are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem.

$$\boldsymbol{W}^* = [\boldsymbol{w}_{\boldsymbol{1}}^*, \boldsymbol{w}_{\boldsymbol{2}}^*, \cdots, \boldsymbol{w}_{\boldsymbol{C-1}}^*] = \mathrm{argmax}_{\boldsymbol{W}}\left\{\frac{|\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{S}_{\boldsymbol{B}}\boldsymbol{W}|}{|\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{S}_{\boldsymbol{W}}\boldsymbol{W}|}\right\} \Rightarrow (\boldsymbol{S}_{\boldsymbol{B}} - \lambda_i\boldsymbol{S}_{\boldsymbol{W}})\boldsymbol{w}_{\boldsymbol{i}}^* = 0$$

(where $i = 1, \ldots, C - 1$)

Linear Discriminant Analysis – *C* classes

- NOTES
  - $\boldsymbol{S_B}$ is the sum of *C* matrices of rank one or less (i.e., max *C*) and the mean vectors are constrained by

$$\sum_{i=1}^{C} N_i(\boldsymbol{\mu_i} - \boldsymbol{\mu}) = 0$$

We lose one rank.

    - Therefore, $\boldsymbol{S_B}$ will be of rank (*C-1*) or less
    - This means that only (*C-1*) of the eigenvalues λ_i will be non-zero
  - The projections with maximum class separability information are the eigenvectors corresponding to the largest eigenvalues of $\boldsymbol{S_W^{-1} S_B}$
  - LDA can be derived as the Maximum Likelihood method for the case of **normal class-conditional densities with equal covariance matrices**

Decision Boundaries of LDA

- **Okay, I understand how to find the projection for the *C*-class classification.**

- **Then, what is the decision boundary?**

- **Can you guess? Any ideas are welcome!**

Decision Boundaries of LDA

- Once the transformation $\boldsymbol{W}$ is given, the classification is then performed in the transformed space based on some distance metric, such as Euclidean distance and cosine measure:

$$\text{Euclidean distance } d(\boldsymbol{u},\boldsymbol{v}) = \sqrt{\Sigma_j\left(u_j - v_j\right)^2}$$

$$\text{Cosine similarity } d(\boldsymbol{u},\boldsymbol{v}) = 1 - \frac{\Sigma_j u_j v_j}{\sqrt{\Sigma_j u_j^2}\sqrt{\Sigma_j v_j^2}}$$

- Then upon the arrival of the new instance $\boldsymbol{x}_{new}$, it is classified to

$$\mathrm{argmin}_i = d(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_{new}, \widetilde{\boldsymbol{\mu}}_{\boldsymbol{i}}) = d(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_{new}, \boldsymbol{W}^{\mathrm{T}}\boldsymbol{\mu}_{\boldsymbol{i}})$$

  – where $\boldsymbol{W}$ reduces to $\boldsymbol{w}$ in binary classification.

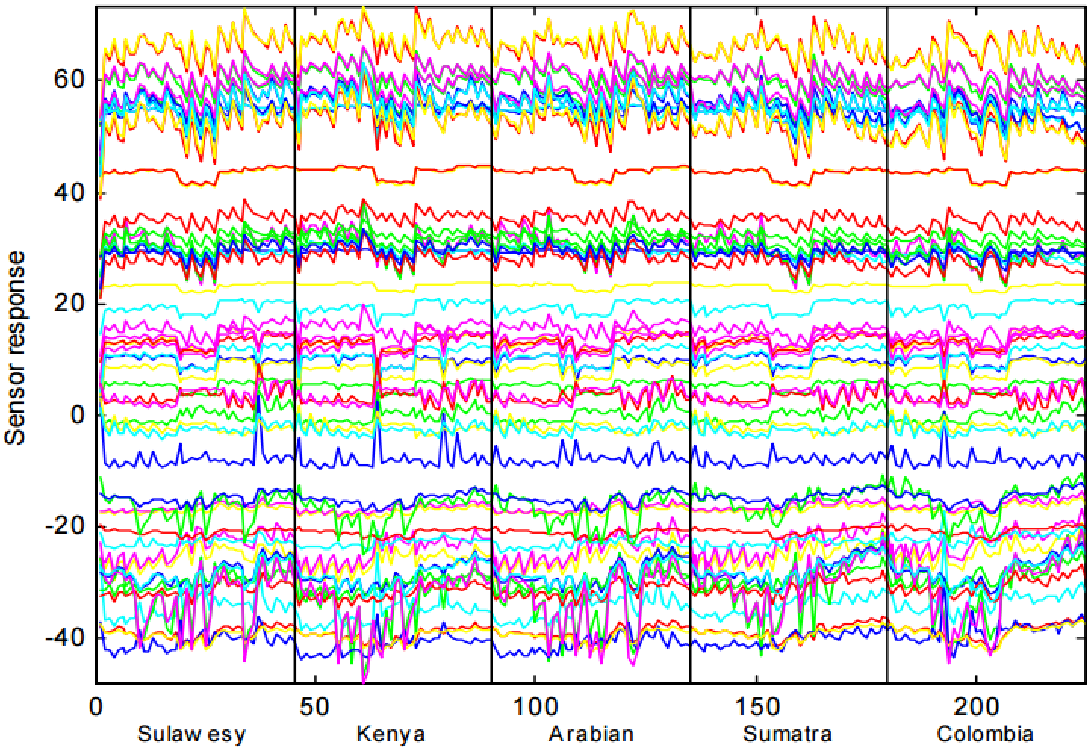
Assumption of LDA

- LDA approaches the problem by assuming that the conditional probability density functions $p(x|y=0)$ and $p(x|y=1)$ are both normally distributed with mean and covariance parameters $(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$ and $(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$, respectively.

- Regarding this, note that we can derive LDA by Bayes Theorem (please see the reference).

LDA vs. PCA Example

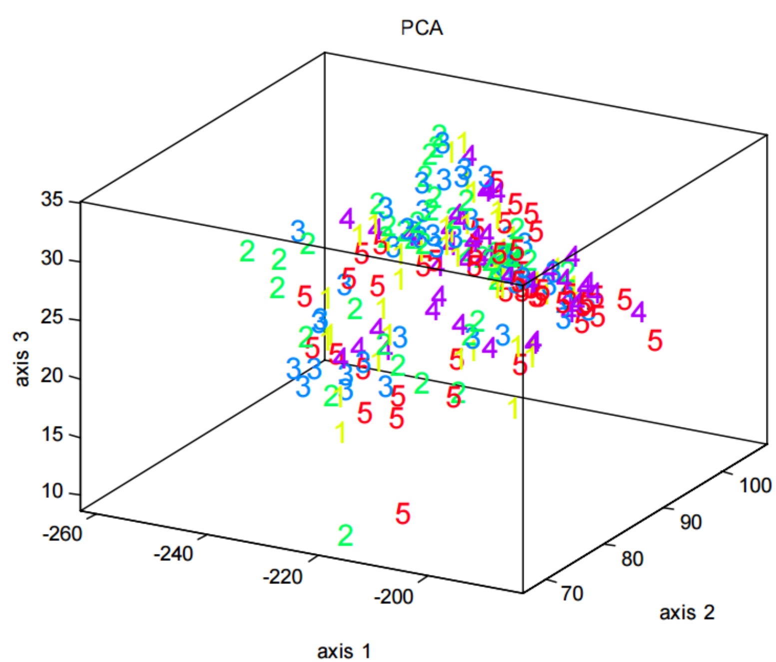
- Coffee discrimination with a gas sensor array
  - Five types of coffee beans were presented to an array of chemical gas sensors
  - For each coffee type, 45 “sniffs” were performed and the response of the gas sensor array was processed in order to obtain a 60-dimensional feature vector

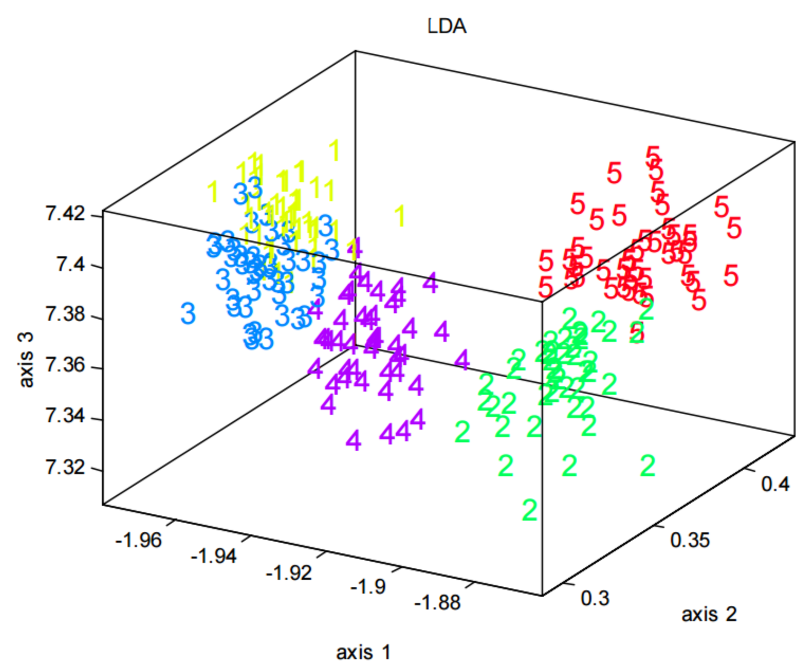


I.V. $\boldsymbol{X} \in \mathbb{R}^{225\times 60}$
D.V. $y_i \in \{1,2,3,4,5\}$
where $i = 1, \ldots, 225$

LDA vs. PCA Example

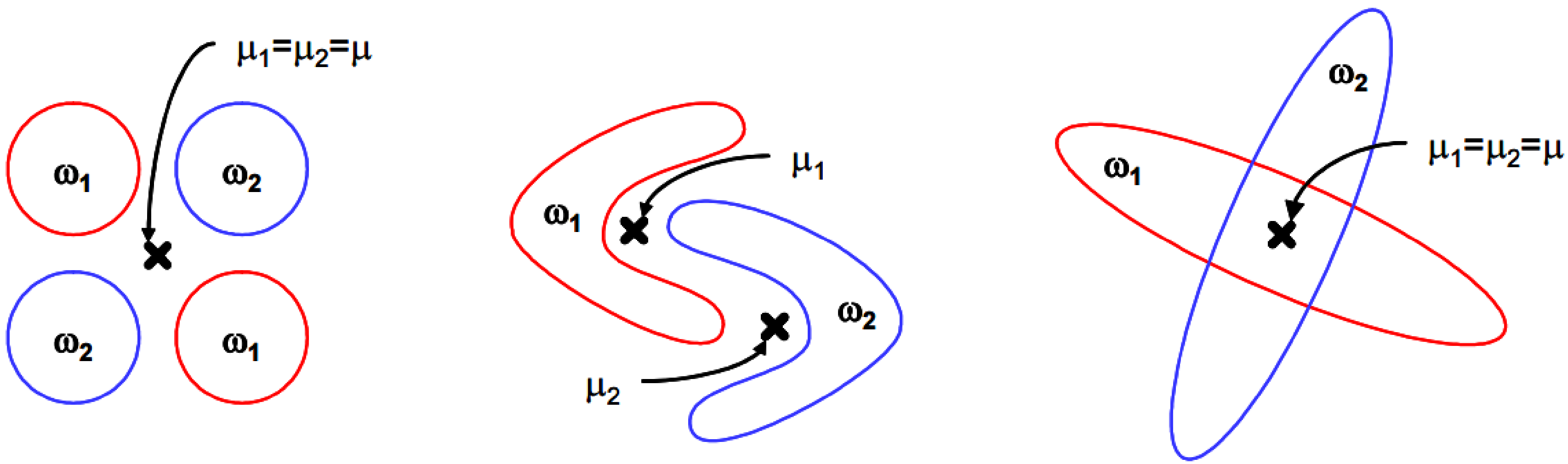
- Results
  - These figures show the performance of PCA and LDA on an odor recognition problem
  - From the 3D scatter plots it is clear that LDA outperforms PCA in terms of class discrimination
  - This is one example where the discriminatory information is not aligned with the direction of maximum variance





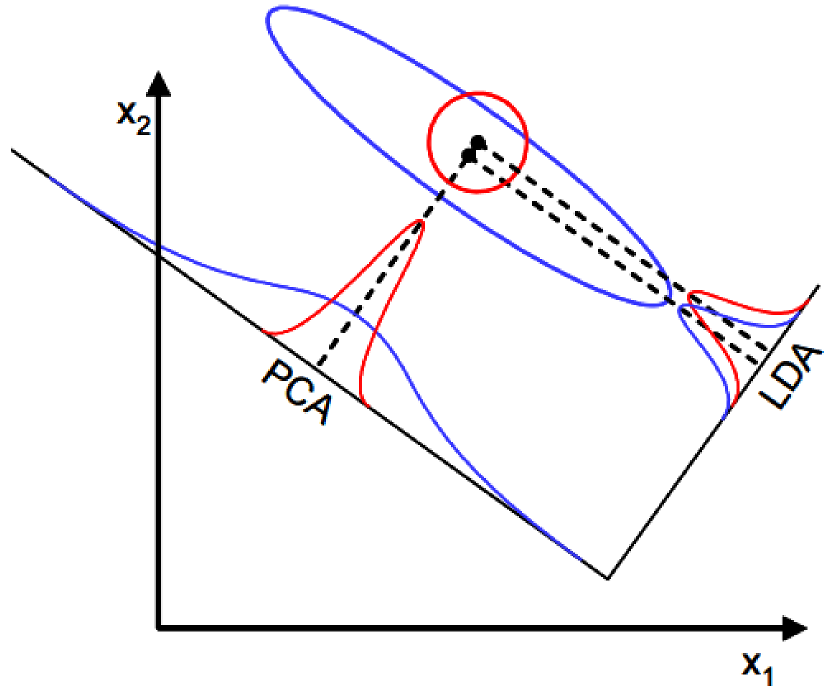
Limitations of LDA

- 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.



In this case, PCA is encouraged.
You should look at your data first.

Questions?