Aprenentatge Automàtic 1

GCED

Lluís A. Belanche

belanche@cs.upc.edu

Soft Computing Research Group
Departament de Ciències de la Computació (Computer Science Department)
Universitat Politècnica de Catalunya - Barcelona Tech

2019-2020

LECTURE 2: Linear Data Visualization

Dimensionality reduction

There are two main tasks (goals) associated to these techniques:

Signal representation The goal is to represent the data accurately in a lower-dimensional space

Signal classification The goal is to enhance the class-discriminatory information in the lower-dimensional space

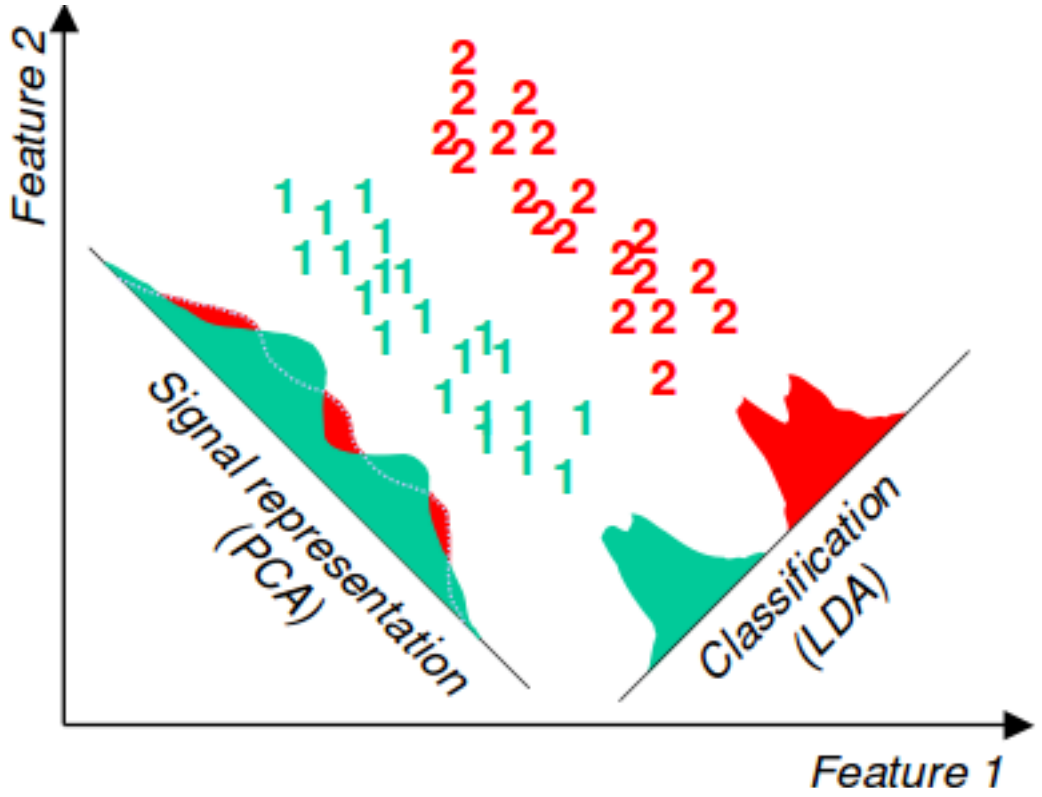
Unfortunately, there is no systematic way to generate non-linear transforms $\longrightarrow$ we will focus on **linear** methods for **feature extraction**:

PCA Principal Components Analysis

FDA Fisher's Discriminant Analysis

ICA Independent Components Analysis

Dimensionality reduction



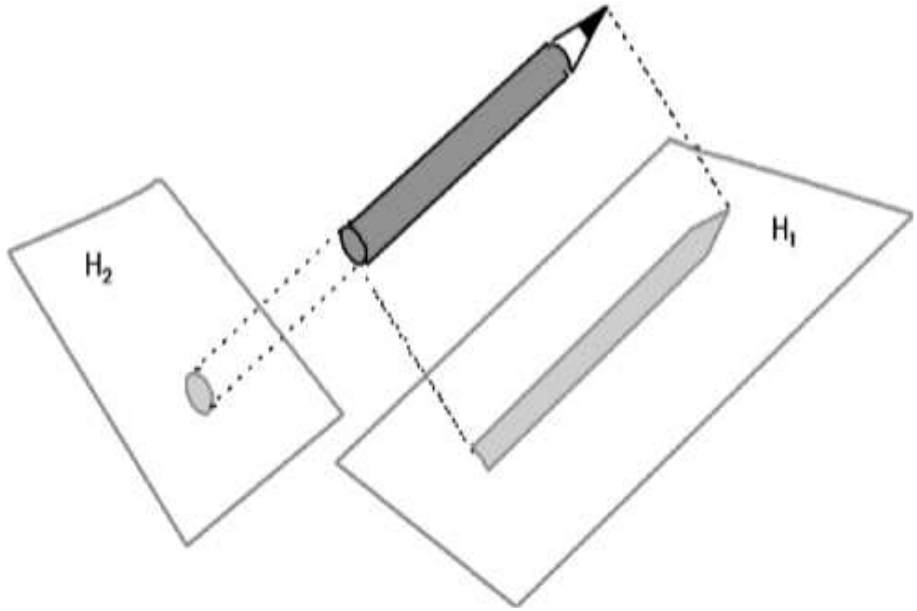
PCA theory

PCA is a technique for:

1. dimensionality reduction
2. lossy data compression
3. feature extraction
4. data visualization

Idea: orthogonal projection of the data onto a lower dimensional linear space, such that the variance of projected data is maximized.

PCA theory



We aim at projecting the data onto a new space so that a maximum of information is preserved. **In PCA, information is variance**.

PCA theory

We depart from a data sample generated by a stochastic mechanism $X = (X_1, \ldots, X_d)^\top$ of mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma = (\sigma_{ij}^2)$.

Let us recall that:

- $\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])^\top = \boldsymbol{\mu}$ and $\mathbb{E}[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^\top] = \Sigma$.
- $\Sigma_{d\times d}$ is symmetric (this implies its eigenvalues are real)
- $\Sigma_{d\times d}$ is p.d. (this implies its eigenvalues are positive)
- $CoVar[X_i, X_j] = \sigma_{ij}^2$ and $Var[X_i] = \sigma_{ii}^2 = \sigma_i^2$

PCA theory

Another useful statistical matrix is the **correlation** matrix $R = (r_{ij})$:

$$r_{ij} = \frac{CoVar[X_i, X_j]}{\sqrt{Var[X_i]}\sqrt{Var[X_j]}} = \frac{\sigma_{ij}^2}{\sigma_i \sigma_j}$$

If the variables are centered, i.e., $\mathbb{E}[X] = 0$, then

$$r_{ij} = \frac{\mathbb{E}[X_i X_j]}{\sqrt{\mathbb{E}[X_i^2]}\sqrt{\mathbb{E}[X_j^2]}}, \text{ which holds because:}$$

1. $CoVar[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$

2. $Var[X_i] = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2$

PCA theory

We consider the problem of finding a new set of variables $Z = (Z_1, \ldots, Z_d)^\top$ s.t. they are decorrelated and their variances decrease (Z_1 holding the greatest variance). We then take Z_j to be a linear combination of the X:

$$Z_j = \sum_{i=1}^{d} a_{ij} X_i = \boldsymbol{a}_j^\top X$$

where the $\boldsymbol{a}_j^\top = (a_{1j} \ldots, a_{dj})^\top$ are the combination coefficients. We also impose the normalization condition $||\boldsymbol{a}_j||^2 = \boldsymbol{a}_j^\top \boldsymbol{a}_j = 1$. This sets up an **orthogonal transformation**.

We choose $\boldsymbol{a}_1$ s.t. it maximizes the variance of Z_1, subject to $\boldsymbol{a}_1^\top \boldsymbol{a}_1 = 1$.

$$Var(Z_1) = Var(\boldsymbol{a}_1^\top X) = \boldsymbol{a}_1^\top Var(X) \boldsymbol{a}_1 = \boldsymbol{a}_1^\top \Sigma \boldsymbol{a}_1$$

PCA theory

Lagrange multipliers

Procedure to maximize a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ subject to a restriction $g(x_1, \ldots, x_d) = c$.

The stationary solution points of f fulfill:

$$\frac{\partial f}{\partial x_i} - \lambda \frac{\partial g}{\partial x_i} = 0, \ i = 1, \ldots, d$$

for some $\lambda \in \mathbb{R}$. We must solve these d equations and that of the restriction. A usual way is by forming the Lagrangian:

$$L(\boldsymbol{x}) := f(\boldsymbol{x}) - \lambda(g(\boldsymbol{x}) - c), \ \boldsymbol{x} = (x_1, \ldots, x_d)^\top$$

PCA theory

In our case,

$$L(\boldsymbol{a}_1) = \boldsymbol{a}_1^\top \Sigma \boldsymbol{a}_1 - \lambda(\boldsymbol{a}_1^\top \boldsymbol{a}_1 - 1)$$

and so

$$\frac{\partial L}{\partial \boldsymbol{a}_1} = 2\Sigma \boldsymbol{a}_1 - 2\lambda \boldsymbol{a}_1 = 0$$

Therefore $\Sigma \boldsymbol{a}_1 = \lambda \boldsymbol{a}_1$ or $(\Sigma - \lambda I_d)\boldsymbol{a}_1 = 0$.

This amounts to finding an eigenvalue λ of Σ with eigenvector $\boldsymbol{a}_1$.

PCA theory

We can arrange the eigenvalues of Σ as $\lambda_{(1)} > \lambda_{(2)} > \ldots \lambda_{(d)} > 0$.

Which one do we choose?

$$Var(Z_1) = Var(\boldsymbol{a}_1^\top X) = \boldsymbol{a}_1^\top \Sigma \boldsymbol{a}_1 = \boldsymbol{a}_1^\top \lambda \boldsymbol{a}_1 = \lambda(\boldsymbol{a}_1^\top \boldsymbol{a}_1) = \lambda$$

So we must choose $\lambda = \lambda_{(1)}$.

Therefore $\boldsymbol{a}_1$ is the eigenvector associated to $\lambda_{(1)}$.

PCA theory

To find $Z_2 = \boldsymbol{a}_2^\top X$, we must work a little harder ... we must impose both $\boldsymbol{a}_2^\top \boldsymbol{a}_2 = 1$ and Z_2 to be uncorrelated with Z_1.

$$CoVar[Z_2, Z_1] = CoVar[\boldsymbol{a}_2^\top X, \boldsymbol{a}_1^\top X] = \mathbb{E}[\boldsymbol{a}_2^\top (X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^\top \boldsymbol{a}_1]$$
$$= \mathbb{E}[\boldsymbol{a}_2^\top]\mathbb{E}[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^\top]\mathbb{E}[\boldsymbol{a}_1] = \boldsymbol{a}_2^\top \boldsymbol{\Sigma} \boldsymbol{a}_1$$

which should be equal to zero. But $\boldsymbol{\Sigma}\boldsymbol{a}_1 = \lambda_{(1)}\boldsymbol{a}_1$, so we end up in

$$\boldsymbol{a}_2^\top \lambda_{(1)} \boldsymbol{a}_1 = 0$$

which is equivalent to require that $\boldsymbol{a}_2$ and $\boldsymbol{a}_1$ are orthogonal.

PCA theory

So we build the Lagrangian:

$$L(\boldsymbol{a}_2) = \boldsymbol{a}_2^\top \Sigma \boldsymbol{a}_2 - \lambda(\boldsymbol{a}_2^\top \boldsymbol{a}_2 - 1) - \delta(\boldsymbol{a}_2^\top \boldsymbol{a}_1)$$

and so

$$\frac{\partial L}{\partial \boldsymbol{a}_2} = 2\Sigma \boldsymbol{a}_2 - 2\lambda \boldsymbol{a}_2 - \delta \boldsymbol{a}_1 = 0$$

If we left-multiply by $\boldsymbol{a}_1^\top$:

$$2\boldsymbol{a}_1^\top \Sigma \boldsymbol{a}_2 - \delta = 0$$

PCA theory

Since Σ is symmetric, $\boldsymbol{a}_1^\top \Sigma \boldsymbol{a}_2 = \boldsymbol{a}_2^\top \Sigma \boldsymbol{a}_1$ and we obtain $\delta = 0$.

Therefore $\Sigma \boldsymbol{a}_2 = \lambda \boldsymbol{a}_2$ or $(\Sigma - \lambda I_d)\boldsymbol{a}_2 = 0$

This now we know how to do it ... we choose $\lambda = \lambda_{(2)}$.

Therefore $\boldsymbol{a}_2$ is the eigenvector associated to $\lambda_{(2)}$.

This way we would be choosing the remaining $\boldsymbol{a}_j, j = 3, \ldots, d$.

(If we ever encounter two equal eigenvalues, nothing is wrong provided we choose orthogonal eigenvectors)

PCA theory

Summary (1)

1. Let us call $A = [\boldsymbol{a}_{(1)}; \ldots; \boldsymbol{a}_{(d)}]$ the matrix of eigenvectors, arranged by columns. We have $Z = A^\top X$.

   This transformation can also be used to project new points!

2. Let us call Δ the covariance matrix of the Z. It is clear that $\Delta = diag(\lambda_{(1)}, \ldots, \lambda_{(d)})$. Therefore the new variables are uncorrelated.

3. It can also be proven that $\Delta = A^\top \Sigma A$ (or, equivalently, $\Sigma = A \Delta A^\top$), since A is orthogonal and $AA^\top = I_d$.

PCA theory

Summary (2)

Interestingly, the total amount of variance remains constant:

$$\sum_{i=1}^{d} Var[Z_i] = \sum_{i=1}^{d} \lambda_{(i)} = Tr(\Delta) = Tr(A^\top \Sigma A)$$

$$= Tr(\Sigma A A^\top) = Tr(\Sigma) = \sum_{i=1}^{d} Var[X_i]$$

And the first m PCs explain a **proportion of variance** equal to:

$$\frac{\sum_{i=1}^{m} \lambda_{(i)}}{\sum_{i=1}^{d} \lambda_{(i)}} \times 100\,\%$$

PCA theory

PCA from the correlation matrix

If we **standardize** the variables X_i prior to analysis, this is equivalent to computing the PCA from the correlation matrix R (instead of the covariance matrix Σ).

$$\tilde{X}_i = \frac{X_i - \mathbb{E}[X_i]}{\sqrt{Var[X_i]}}$$

We get

$$\sum_{i=1}^{d} Var[Z_i] = \sum_{i=1}^{d} Var[\tilde{X}_i] = \sum_{i=1}^{d} 1 = d$$

The results will be different (eigenvectors and eigenvalues differ)

PCA theory

Computations in practice

In practical situations, only an i.i.d data sample $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, generated by $X_1, \ldots, X_d$ is available.

We have unbiased estimates for the vector of means and for the covariance matrix:

$$\widehat{\mu} = \bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$$

PCA theory

Summary of Standard PCA

1. We are given a data set of d-dimensional vectors $X = \{\boldsymbol{x}_i\}$ for $i = 1, ..., n$ which we center around the origin, as $\boldsymbol{x}_i := \boldsymbol{x}_i - \bar{\boldsymbol{x}}$.

2. Compute the (centered) sample covariance matrix of the data S

3. Compute the eigenvalues and eigenvectors of S and sort them $(\lambda_{(j)}, \boldsymbol{u}_{(j)})$; choose the number $m < d$ of PCs to retain.

4. Let $U = [\boldsymbol{u}_{(1)}; \ldots; \boldsymbol{u}_{(m)}]$ the matrix of selected eigenvectors, arranged by columns.

5. The result is the m-dimensional data sample $Z = U^\top X$.

PCA remarks

1. In essence, PCA aligns the transformed axes Z with the directions of maximum variability

2. For Gaussian data X, PCA decorrelates the variables and makes them independent

3. For non-Gaussian data X, PCA simply decorrelates the variables

4. There is no guarantee that these new axes contain good features for discrimination!

5. PCA was introduced by K. Pearson in 1901, and generalized by Loève in 1963; it is also known as the discrete Karhunen-Loève transform in signal processing

FDA theory

FDA is a technique for:

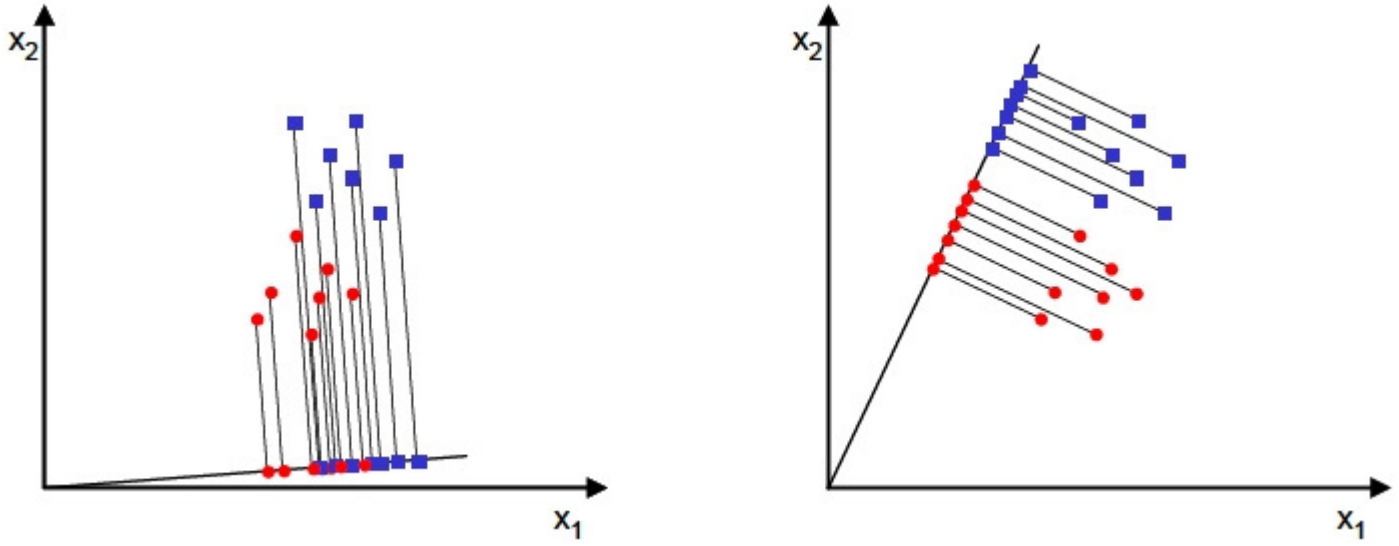
1. dimensionality reduction
2. supervised classification
3. feature extraction
4. data visualization

Idea: projection of the data onto a lower dimensional linear space, such that the separability of projected data is maximized.

FDA theory

Fisher's idea is to regard **dot product** as the projection y of some $\boldsymbol{x} \in \mathbb{R}^p$ from classes ω_1 or ω_2, via a projection vector $\boldsymbol{w}$:

$$y = \boldsymbol{w}^\top \boldsymbol{x} \in \mathbb{R}$$



FDA theory

In order to find a good projection vector, we need to define a measure of separation between the projections:

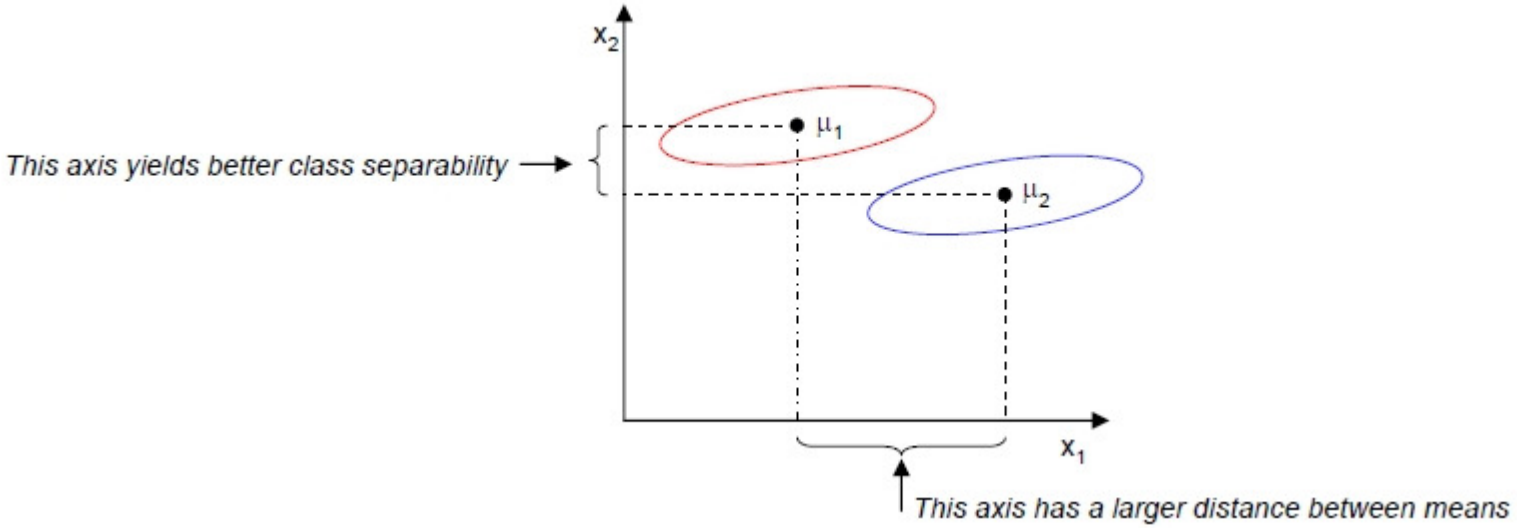
$$\boldsymbol{m}_k = \frac{1}{n_k} \sum_{i \in \omega_k} \boldsymbol{x}_i, \qquad k = 1, 2$$

where $n_1 + n_2 = n$ is the number of examples on every class. We then choose to maximize the (squared) distance between the projected means:

$$(\mu_2 - \mu_1)^2 = (\boldsymbol{w}^\top \boldsymbol{m}_2 - \boldsymbol{w}^\top \boldsymbol{m}_1)^2 = (\boldsymbol{w}^\top (\boldsymbol{m}_2 - \boldsymbol{m}_1))^2$$

FDA theory

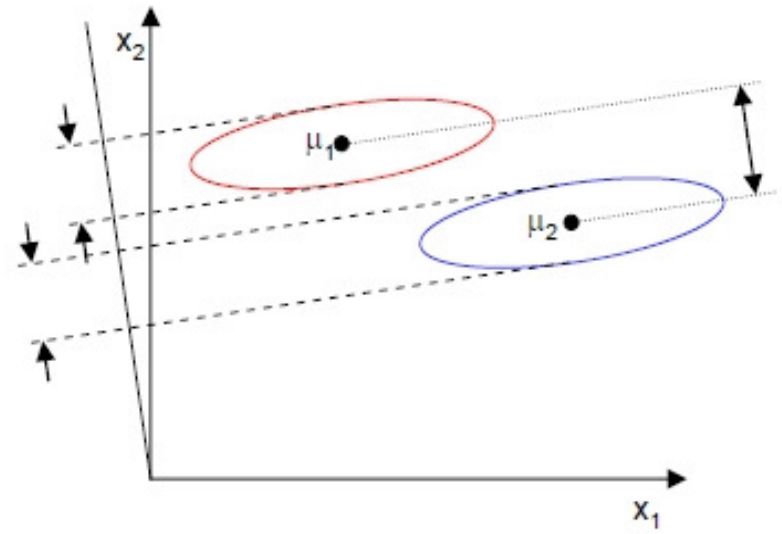
However, the distance between the projected means is not a very good measure since it does not take into account the dispersion (**scatter**) within the classes:



The problem is that the covariance matrices for each class are far from being diagonal ...

FDA theory

We actually want to look for the projection where examples from the same class are projected very close to each other and the projected means are as far apart as possible:



FDA theory

The solution (proposed by R. Fisher) is to maximize a function that represents the difference between the means, normalized by a measure of the within-class scatter:

1. For each class k we define the scatter as:

$$s_k^2 = \sum_{i \in \omega_k} (\boldsymbol{w}^\top \boldsymbol{x}_i - \mu_k)^2, \qquad k = 1, 2$$

2. The total scatter is $s_1^2 + s_2^2$.

3. Fisher's idea was to maximize:

$$J(\boldsymbol{w}) = \frac{(\mu_2 - \mu_1)^2}{s_1^2 + s_2^2}$$

FDA theory

It can be shown that $J(\boldsymbol{w})$ can be rewritten as:

$$J(\boldsymbol{w}) = \frac{(\mu_2 - \mu_1)^2}{s_1^2 + s_2^2} = \frac{\boldsymbol{w}^\top S_B \boldsymbol{w}}{\boldsymbol{w}^\top S_W \boldsymbol{w}}$$

where:

- $S_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\top$
  is the **between-class** scatter matrix (a rank one matrix)

- $S_W = \sum_{i \in \omega_1} (\boldsymbol{x}_i - \boldsymbol{m}_1)(\boldsymbol{x}_i - \boldsymbol{m}_1)^\top + \sum_{i \in \omega_2} (\boldsymbol{x}_i - \boldsymbol{m}_2)(\boldsymbol{x}_i - \boldsymbol{m}_2)^\top$
  is the **within-class** scatter matrix

FDA theory

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

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0$$

Upon solving in $\boldsymbol{w}$ we arrive at:

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

Solving this generalized eigenvalue problem yields:

$$\hat{\boldsymbol{w}} = S_W^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

Known as **Fisher's Linear Discriminant** (1936), although it is not a discriminant but a specific choice for projection down to one dimension.

FDA theory

- FDA generalizes very gracefully for K class problems: the only restriction is that the maximum number of projection directions is $K-1$.

- FDA can also be derived as the Maximum Likelihood result for the case of Gaussian class-conditional densities with equal covariance matrices (in this case it is known as LDA).

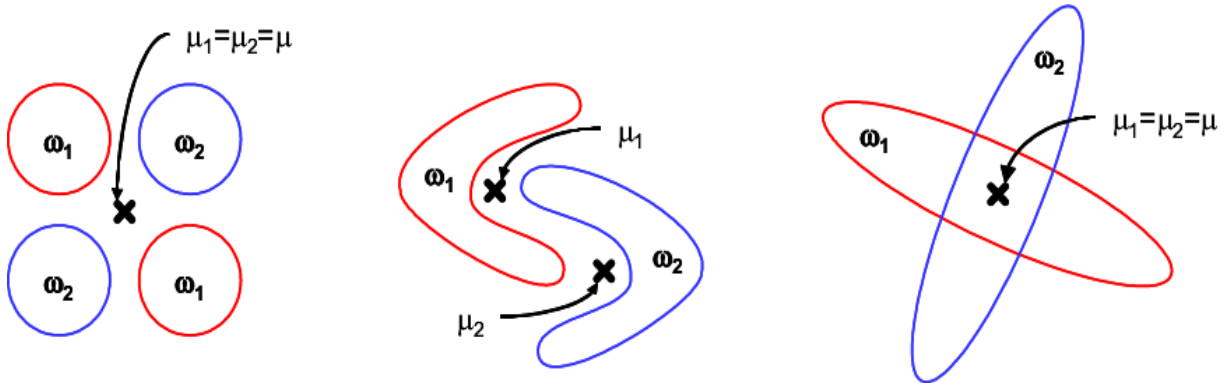
WARNING!

1. FDA is able to extract a maximum of $K-1$ projection directions: maybe insufficient for complex data

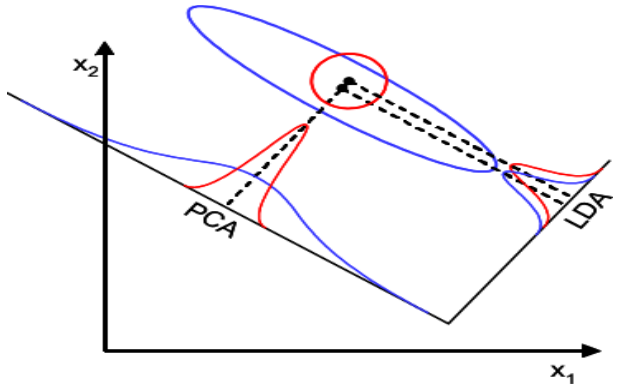
2. PCA is able to extract d projection directions, but how many are necessary is not clear

When will FDA presumably fail?

- If the classes are far from Gaussian, the FDA projections will not be able to preserve any complex structure:



- FDA will also fail when the discriminatory information is not in the mean but rather in the variance of the data (*e.g.*, if $J(\boldsymbol{w}) = 0$):



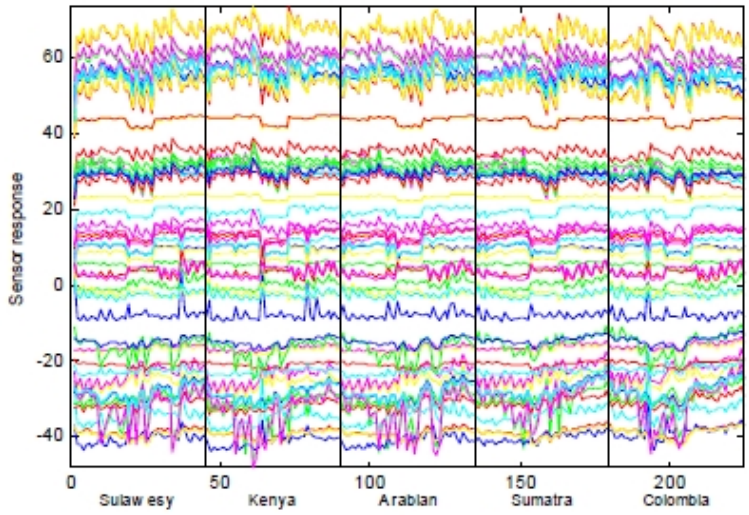
FDA application example

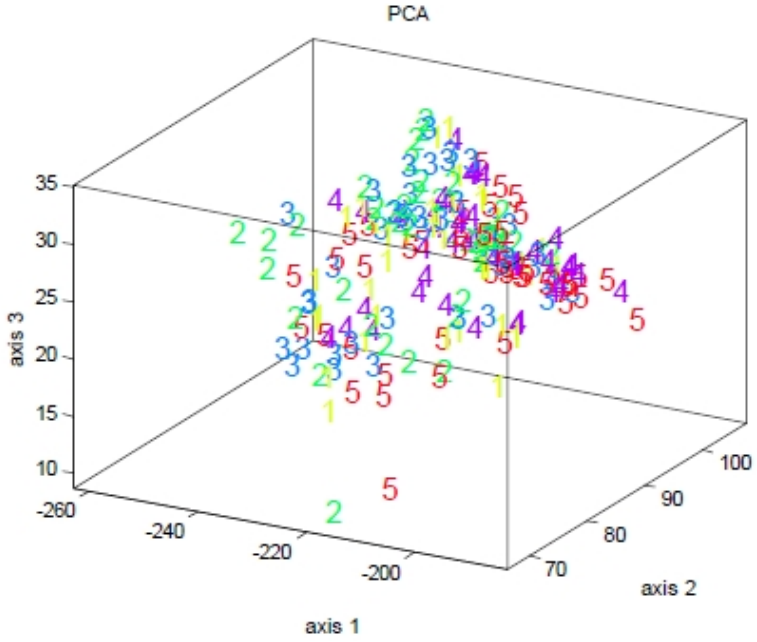
These figures show the performance of PCA and LDA on an odor recognition problem

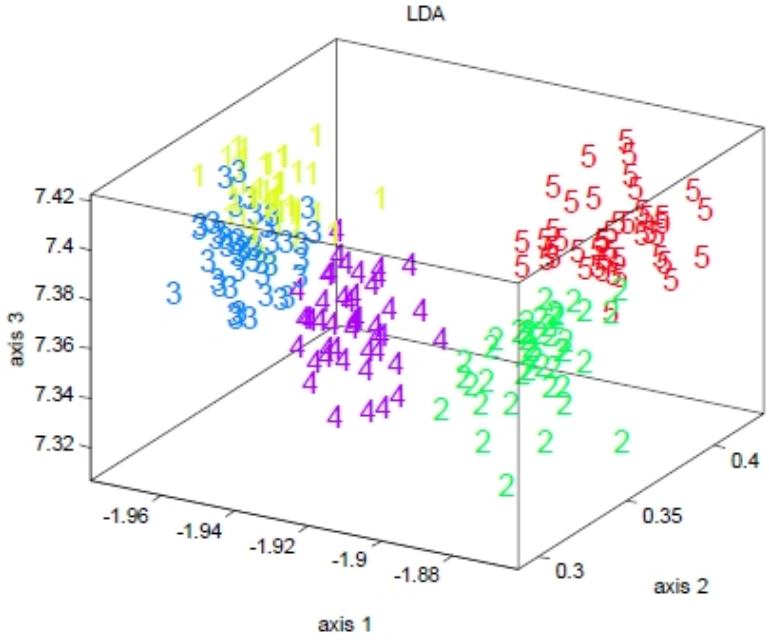
- Five types of coffee beans were presented to an array of chemical gas sensors
- For each coffee type, 45 "sniffs" were performed and the response of the gas sensor array was processed in order to obtain a 60-dimensional feature vector

Results

- 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

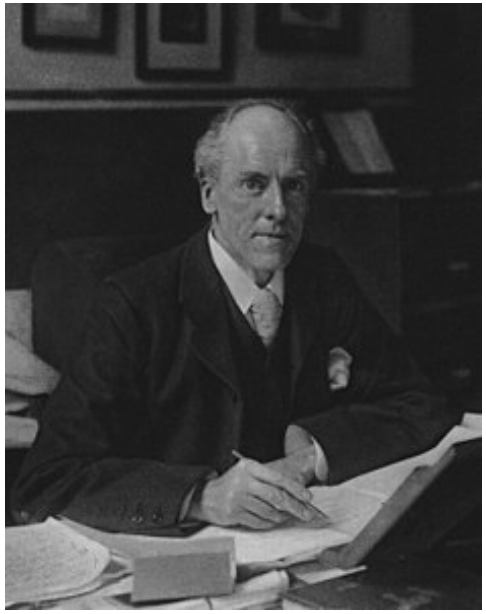




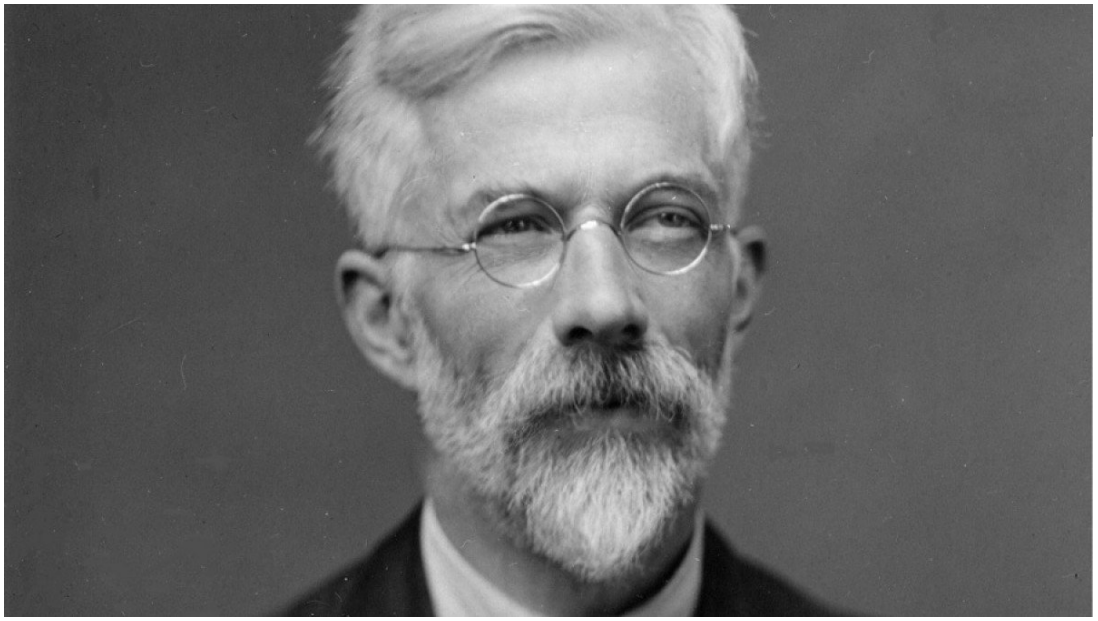


Thanks to R. Gutiérrez-Osuna

Relevant characters for this lecture



Karl Pearson
(1857-1936)



Sir Ronald Fisher
(1890-1962)