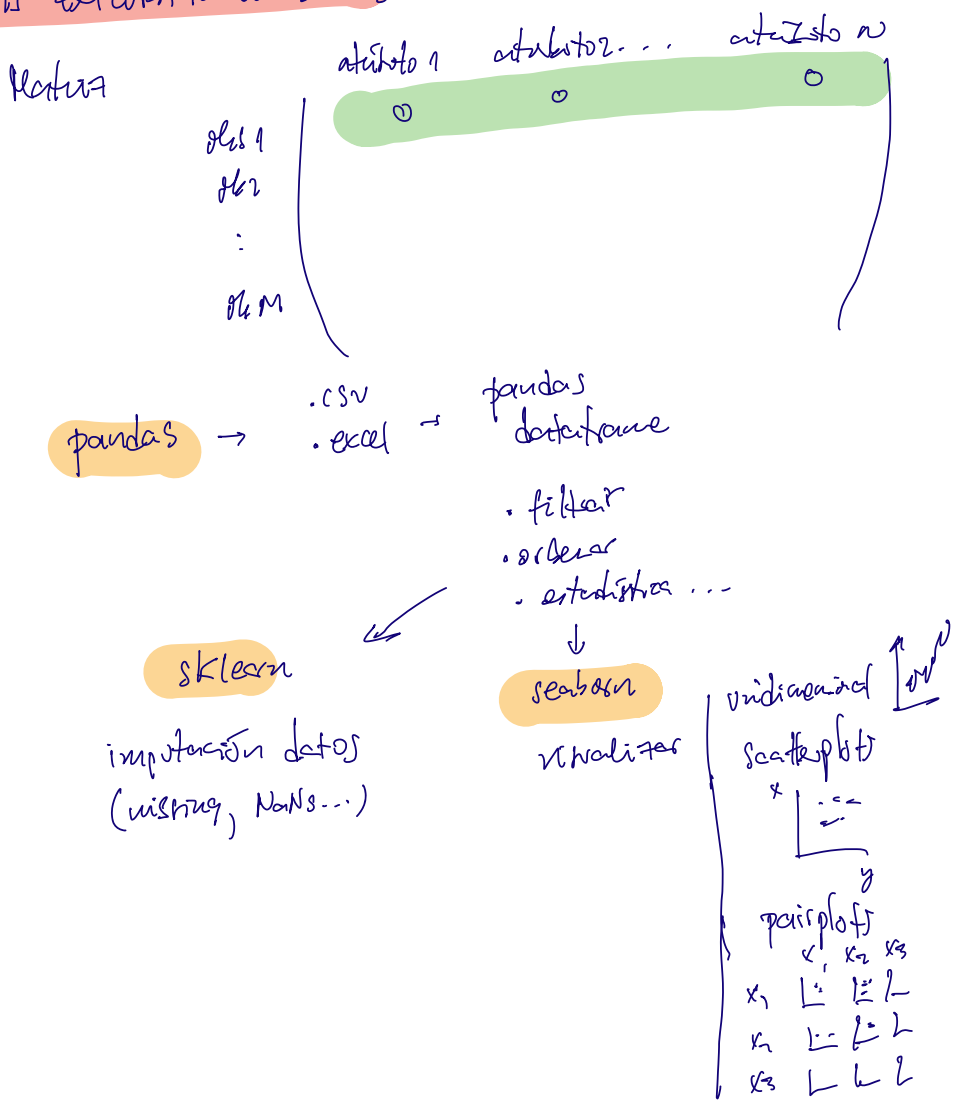
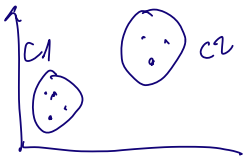
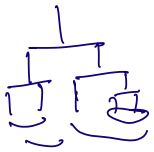
1. Intro AI

2. ANÁLISIS EXPLORATORIO DATOS



3. CLUSTERING



- → K-MEANS → simple
- → AGGLOMERATIVE → dendrograma 
- → GAUSSIAN MIXTURE MODELS → How many clusters?

Índices para selección modelos

BIC, AIC

selección > clustering

mpg, iris...

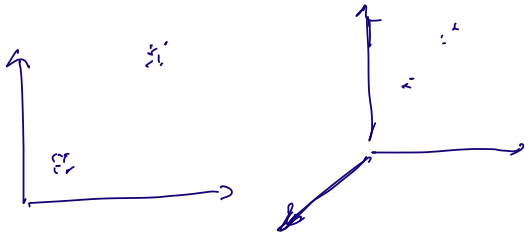
↳

4. REDUCCIÓN DE DIMENSIONALIDAD

↳ ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

Matriz Datos $X_{NOBS \times NFEAT}$ + W class labels

- → IMPUTACIÓN
- → CLUSTERING
- → CLASIFICACIÓN



$X_{NOBS \times NFEAT}$

↓

$X_{NOBS \times d}$

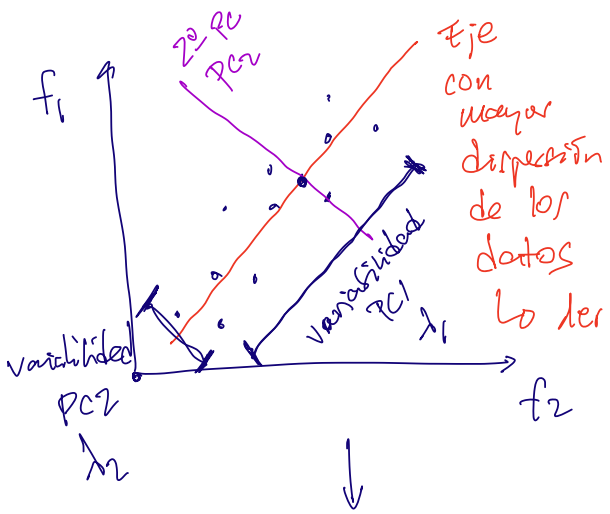
↓

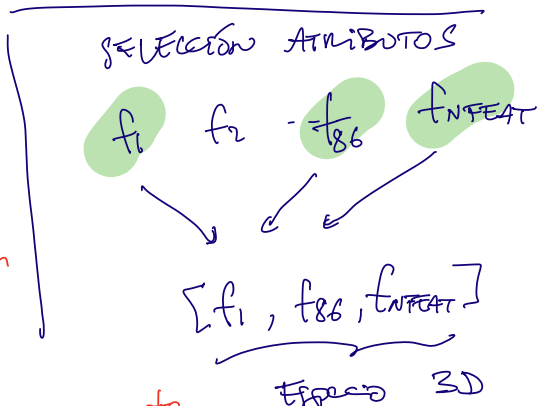
dimensión reducida

$d < NFEAT$

CUANTO MAYOR LA DIMENSIONALIDAD (# FEATURES) → MÁS DATOS SON NECESARIOS PARA IDENTIFICAR PATRONES

Ex: PCA 2D NFEAT=2





NFEAT = 1

$(f_1, f_2) \xrightarrow{PCA} (PC_1, PC_2)$

2d → 2d

Desplazamiento, Rotación } transf. lineal sistema de coord.

$$\begin{pmatrix} PC_1 \\ PC_2 \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

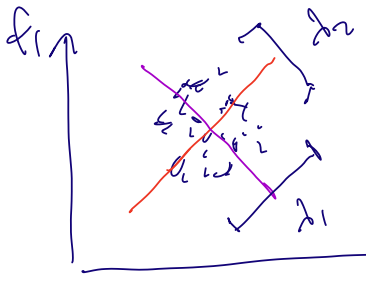
$$PC_1 = \alpha_{11} f_1 + \alpha_{12} f_2$$

↕

La nueva variable es combinación lineal de f_1 y f_2

Cuánta variabilidad explica el $PC_1 = \lambda_1 / (\lambda_1 + \lambda_2)$

$PC_2 = \lambda_2 / (\lambda_1 + \lambda_2)$

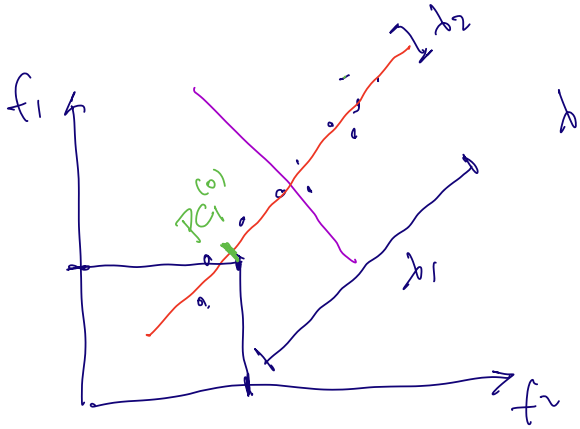


$\lambda_1 \sim \lambda_2 \rightarrow$

$PC_1 \sim 50\%$ variabilidad

$PC_2 \sim 50\%$ variabilidad

NO PUEDO REDUCIR LA DIMENSIONALIDAD PORQUE APENAS HAY CORRELACIÓN ENTRE f_1 y f_2



$\lambda_1 >> \lambda_2$

$PC_1 \sim 97\%$

$PC_2 \sim 3\%$

↓

proyección de los datos en el eje PC_1

$(f_1, f_2) \rightarrow PC_1$

(1d)

Matemáticamente:

① ESCALAR DATOS

$X_{NOBS \times NFEAT}$ ← estandarizar

$$XS = (X - \bar{X}) / \sigma$$

- XS: datos escalados
- $\bar{X}$: media de cada FEATURE centrado
- σ: escalado

(2) matriz covarianza

$$C = \frac{1}{NOBS-1} \cdot \left(\frac{X-\bar{X}}{\sigma}\right)^T \cdot \left(\frac{X-\bar{X}}{\sigma}\right)$$

$(NFEAT \times NOBS) \cdot (NOBS \times NFEAT)$

$NFEAT \times NFEAT$

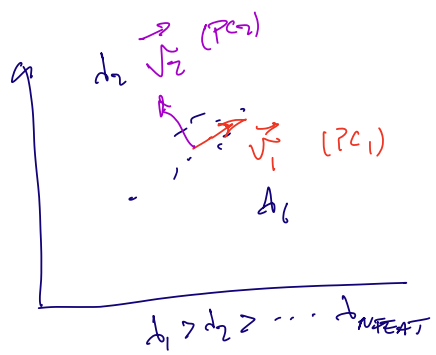
(3) Diagonalizar C

$C \cdot \vec{v} = \lambda \cdot \vec{v}$

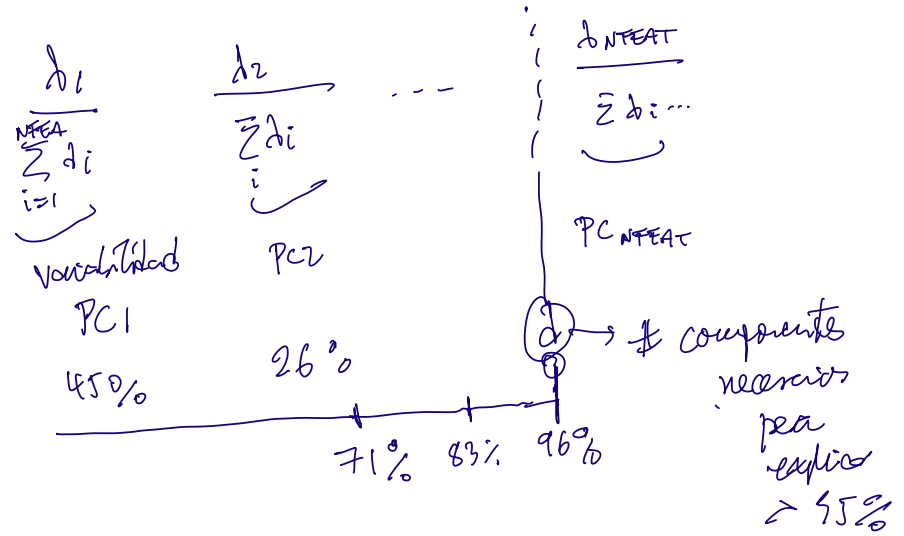
valores propios $\lambda_i \longrightarrow$ variabilidad explicada PC_i

vectores propios $\vec{v}_i \longrightarrow$ dirección PC_i

$i = 1 .. NFEAT$



(4) Cuantos PC_i son necesarios para explicar el 95% variabilidad



$(f_1 \dots f_{NFEAT}) \rightarrow (PC_1 \dots PC_d)$

IMPORTANCIA ESCALAR DATOS:

$$X = \begin{pmatrix} f_1(\text{km}) & f_2(\text{m}) \\ 1 & 33 \\ 0.5 & 89 \\ 0.3 & 124 \\ 1.2 & 96 \\ 5.4 & 33 \\ 1.8 & 430 \end{pmatrix} \rightarrow XS \begin{pmatrix} fs_1 & fs_2 \\ 1 & 1 \end{pmatrix}$$

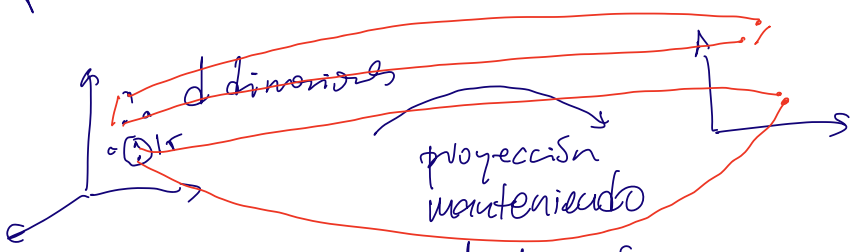
σ_1

rango σ_2

SELECCIÓN ATRIBUTOS

5. TÉCNICAS VISUALIZACIÓN DATOS MULTIDIMENSIONALES

Espacio atributos dimensión d = NºAT



→ MULTIDIMENSIONAL SCALING (MDS)

→ t - SNE