

Tema 3: REDUCCIÓ DE DIMENSIONALITAT

↳ ANÀLISI DE COMPONENTS PRINCIPALS (PCA)

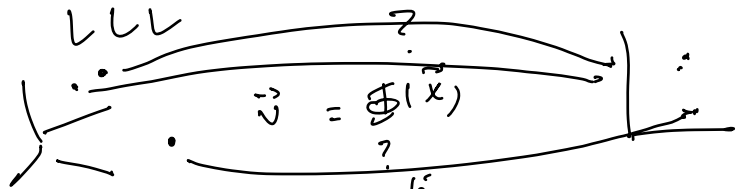
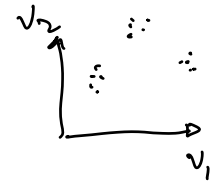
↳ Tècniques de visualització de dades multivariades

ex: dades $NVAR = 18$

$$\vec{x} \in \mathbb{R}^{18}$$

LLLLLL
LLL

projecció
baixa
dimensió



PCA: Principal Component Analysis

Idea: dades observacionals

$NVAR$ variables $1 \dots NVAR$

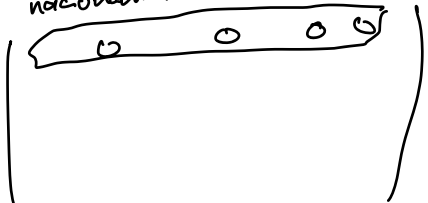


$$\begin{matrix} 1 \\ \vdots \\ NOBS \end{matrix} \begin{pmatrix} & & \\ & X & \\ & & \end{pmatrix}$$



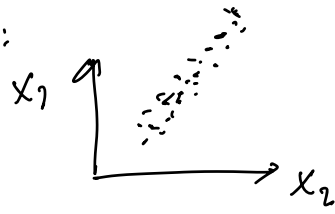
persona 1
persona 2
 \vdots
persona NOBS

nacionalitat \$ edat alçada



estudi

Quint les variables mesurades presenten correlació:

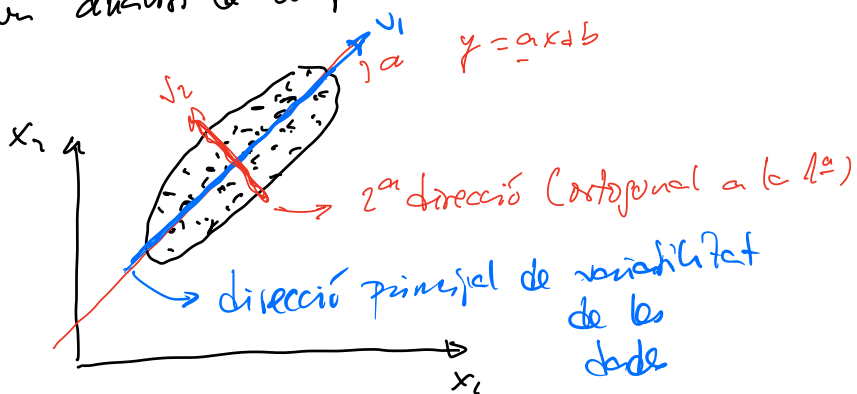


idealment
 varien
 des
 de
 coor
 ones.

	x_1	x_2	\dots	x_{NVAR}
	x_1	x_2	x_3	x_4
	x_1	x_2	x_3	x_4
	x_1	x_2	x_3	x_4
	x_1	x_2	x_3	x_4

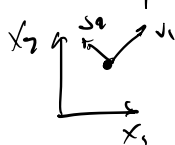
SOLUCIÓ: Aplicar un anàlisi de components principals.

Ex: $NVAR = 2$



transformació d'eixos lineal

$(x_1, x_2) \longrightarrow (v_1, v_2)$
 Rotació
 +
 desplaçament



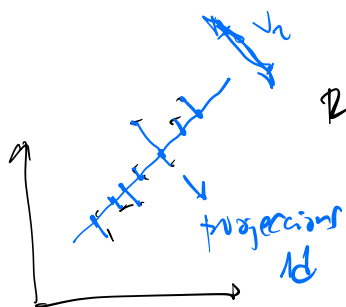
a la nova base:

$v_1 \rightarrow$ molta dispersió/variabilitat
 $v_2 \rightarrow$ baixa dispersió

Reducció dimensional:

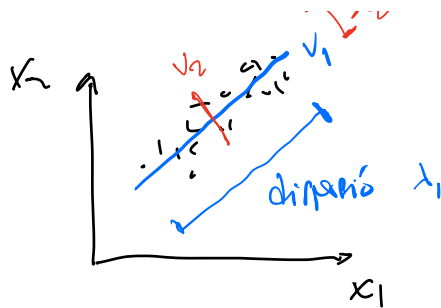
dispersió $v_1 \gg$ dispersió v_2

Dades projectades a v_1
 1-dimensional



Ex. 2D \rightarrow 1D

PCA
 CAL
 ESCALAR
 DADES



dispensió total: $\lambda_1 + \lambda_2$

dispensió relativa eix principal v_1

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} \sim \textcircled{95\%}$$

dispensió relativa v_2

$$\frac{\lambda_2}{\lambda_1 + \lambda_2}$$

$$\{x_1, \dots, x_{NVAR}\}$$

↓ PCA

$$\{v_1, \dots, v_{NVAR}\}$$

$$\{\lambda_1, \lambda_2, \dots, \lambda_{NVAR}\} \quad (\text{ordenats de major a menor})$$

↓ reducció dimensionalitat

Quants λ_i $i=1 \dots NVAR$ calen per explicar un 95% variabilitat?

$$\left| \frac{\lambda_1 + \lambda_2 + \dots + \lambda_p}{\sum_{i=1}^{NVAR} \lambda_i} \sim 0.95 \right|$$

→ p: NAVA DIMENSIONALITATS

p eixos de l'espai PCA

expliquen ~ 95%
variabilitat de les
dades

$$\vec{X} \in \mathbb{R}^{NVAR} \rightarrow \vec{x}_{PCA} \in \mathbb{R}^p$$

$$\text{eix } v_i = \alpha_{i1} x_1 + \alpha_{i2} x_2 + \dots + \alpha_{iNVAR} x_{NVAR}$$

representació
basse
dimensional.

NO ÉS UNA TÈCNICA DE
SELECCIÓ D'ATRIBUTS / VARIABLES!

CÓM? :

X matrix dada

↓

1. ESTANDARITZACIÓ

$$XS = \frac{X - \bar{X}}{\sigma_X} \begin{pmatrix} 1 \\ 1 \\ \vdots \end{pmatrix}$$

2. CALCULAR LA MATRIZ DE CORRELACIÓ

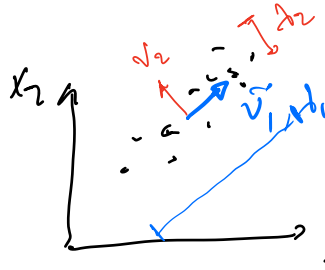
$$C = \text{corr}(XS, XS) \rightarrow \begin{matrix} \text{correlacions} \\ \text{a} \\ \text{parella} \\ \text{de variables} \\ (x_i, x_j) \end{matrix}$$

$NVAR \times NVAR$

3. DIAGONALITZEM C

↳ vectors propis : $\{\vec{v}_1, \vec{v}_2, \dots, \vec{v}_{NVAR}\} \rightarrow \text{DIRECCIÓ}$

↳ valors propis : $\{\lambda_1, \lambda_2, \dots, \lambda_{NVAR}\} \rightarrow \text{FIDELTAT}$



4. ESCOURE DIMENSIONALITAT REDUÏDA P

$$\frac{\sum_{i=1}^P \lambda_i}{\sum_{i=1}^{NVAR} \lambda_i} \sim 0.95$$

5. PROJECTAR DATES A L'ESTAT PCA DE DIMENSIÓ REDUÏDA

$$X_{NOBS \times NVAR} \rightarrow \underbrace{XPCA_{NOBS \times P}}_{\text{NOVES DATES}} /$$

✓
 Reconhecer padrões é mais fácil
 Treinar sistemas de IA.

CONSIDERAÇÕES COMPUTACIONAIS


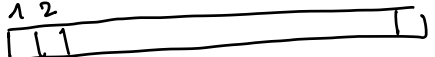
NORMALMENTE:


$N_{OBS} \gg N_{VAR} \rightarrow C_{N_{VAR} \times N_{VAR}} \rightarrow \text{FÁCIL!}$

PENSA, DE VECIADES:

$N_{OBS} \ll N_{VAR}$


Ex: Imagens - base de dados $N_{VAR} = N_{PIXELS}$

I_1  256 pixels \rightarrow 

I_2 

I_3 

\vdots

$I_{N_{OBS}}$ 

1000 Imagens

\times
 $1000 \times 256 \cdot 256$
 \downarrow

$C_{256 \cdot 256 \times 256 \cdot 256}$

\Downarrow

Computar os
 primeiros 1000
 eigenvalues &
 eigenvectors

$$X_{1000 \times 256 \cdot 256}^T =$$

$$X_{256 \cdot 256 \times 1000}$$



$$\tilde{C}_{1000 \times 1000}$$



DIAGONALIZZATA
QUESTA!

(EIGENFACES)