Analyses en composantes principales.

* OBJECTIF :

passer de $X \in \mathbb{R}^{n \times p}$ à $C \in \mathbb{R}^{n \times q}$ avec

$$q < p$$

TEL QUE C SOIT AUSSI PROCHE QUE POSSIBLE DE X

Chaque ~~ligne~~ ligne de X, notée x_i est un vecteur de $\mathbb{R}^p$ qui décrit un individu

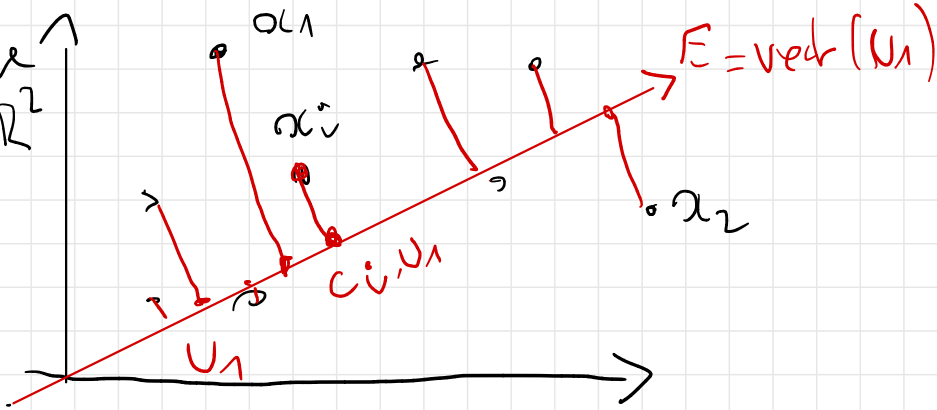
EN PRATIQUE, ON MINIMISE

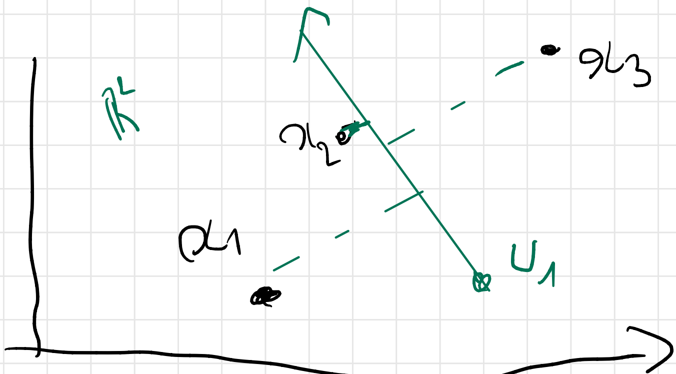
$$J = \sum_{i=1}^{n} \| x_i - U c_i \|^2$$ où c_i est

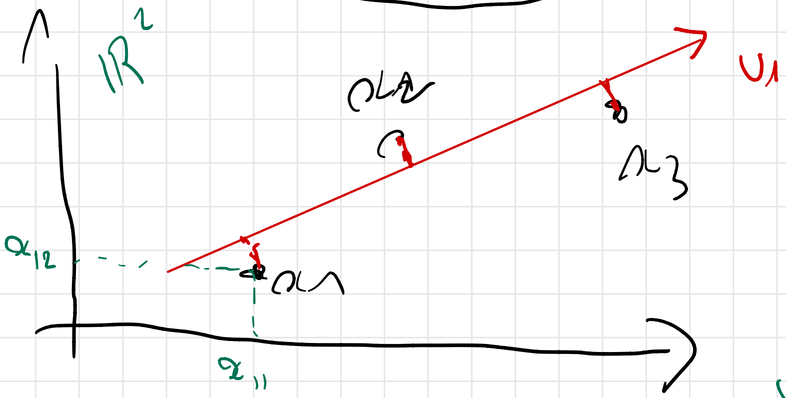
la projection de x_i sur un s.e.v de base $v_1, \dots, v_q$

avec $U = [v_1 \dots v_q]$

Exemple dans $\mathbb{R}^2$





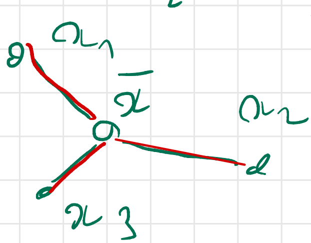


$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ x_3^T \end{bmatrix}$$

$$X = \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} \end{array}$$

Considerons le centre de gravité du nuage X

$$\mu = \bar{x} = \frac{1}{3}(x_1 + x_2 + x_3)$$



$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Notons $\underset{\text{inertie}}{I_T} = \sum_{i=1}^{n} \|x_i - \bar{x}\|^2$

ET CONSIDERONS LA PROJECTION ORTHO.

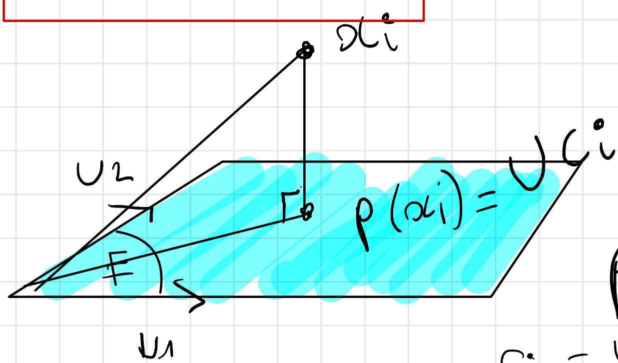
de $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$ sur $E = \text{vect}(u_1, \dots, u_q)$

AVEC $\forall k \ \|u_k\|^2 = 1$ ①

$\forall k \neq l \ \ u_k \perp u_l$ ②

$$\boxed{U = [u_1 \dots u_q]}$$

$$U^T U = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} = \underset{\text{identity}}{I_p}$$



$$p(x_i) = U\underbrace{(U^TU)^{-1}}_{I_p}\underbrace{U^T x_i}_{c_i}$$

$c_i = U^T x_i$

$c_i^T = x_i^T U \quad = \underset{p,q}{U}\ \underset{q,1}{c_i}$

$$\begin{bmatrix} c_1^T \\ c_2^T \\ \vdots \\ c_n^T \end{bmatrix} = \boxed{C = XU}$$

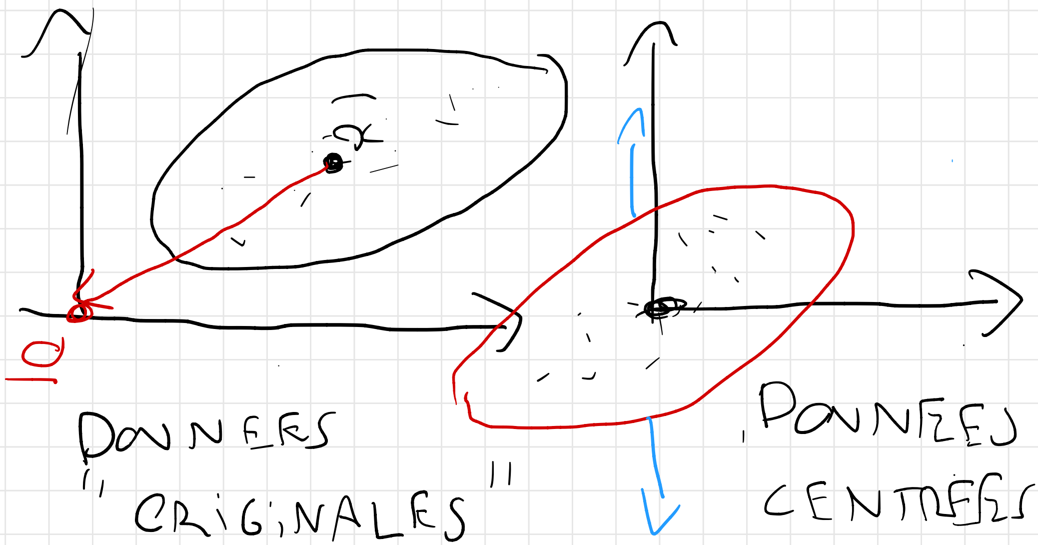
LES DONNÉES X SONT CENTRÉES ET RÉDUITES

$$\forall_{ij}\ X_{ij} \leftarrow \frac{X_{ij} - \overline{X}_j}{\hat{\sigma}_j}$$

AVEC

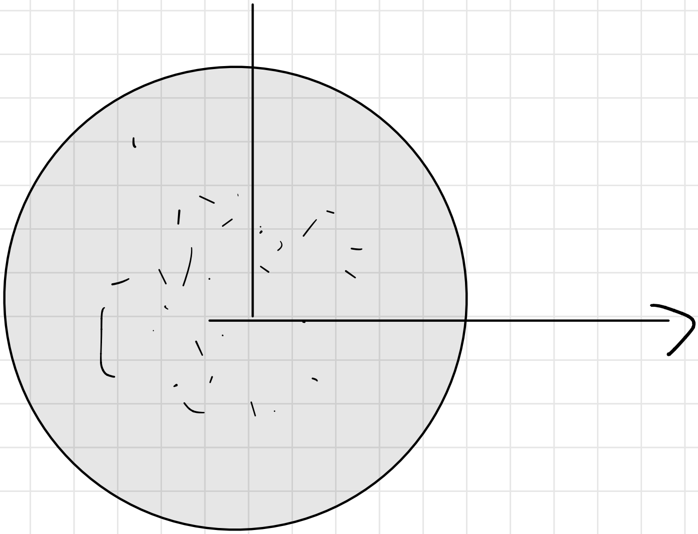
$$\overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$$

$$\hat{\sigma}_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2}$$

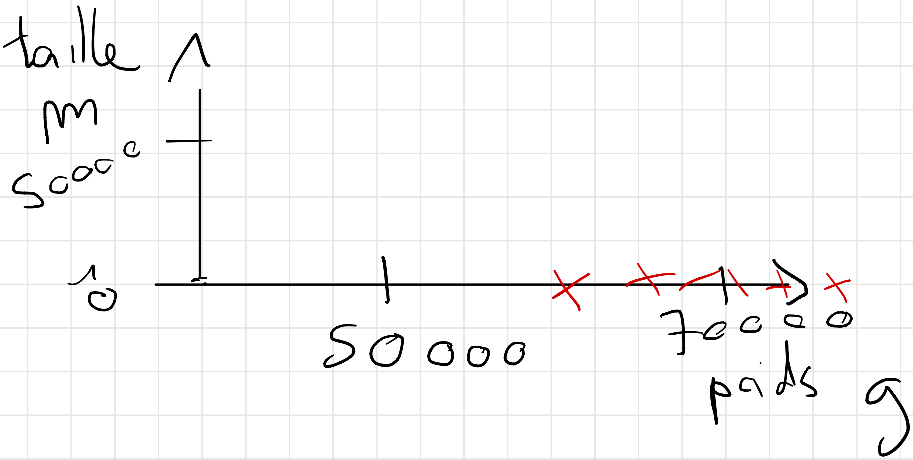


"DONNÉES ORIGINALES"

DONNÉES CENTRÉES



DONNÉES CENTRÉES
RÉDUITES



SI $\bar{x} = 0$ ALORS $\|x_i - \bar{x}\|^2 = \|x_i\|^2$

$$I_T = \sum_{i=1}^{n} \|x_i - \bar{x}\|_2^2$$

$$= \sum_{i=1}^{n} \underbrace{\|x_i\|_2^2}_{\sum_{j=1}^{p} x_{ij}^2} = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2$$

$$= \sum_{i=1}^{n} \|x_i - p(x_i) + p(x_i)\|^2$$

$$= \sum_i \langle x_i - p(x_i) + p(x_i), x_i - p(x_i) + p(x_i) \rangle$$

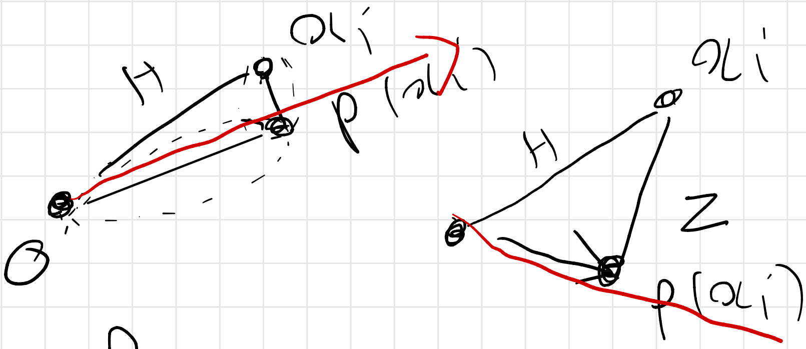
$$= \sum_i \|x_i - p(x_i)\|^2 + \|p(x_i)\|^2 + 2 \underbrace{\langle x_i - p(x_i), p(x_i) \rangle}_{0}$$

$$\boxed{I_T = I_{F^\perp} + I_E}$$

NOTRE PROBLÈME :

MINIMISER $I_E^{\perp}$ OU

MAXIMISE I_E (DE MAN



$$I_E = \sum_{i=1}^{n} \| p(x_i) \|^2$$

$$= \sum_{i=1} \| U c_i \|^2$$

$$= \sum_{i=} \| U U^T x_i \|^2$$

$$= \sum \langle U U^T x_i , U U^T x_i \rangle$$

$$= \sum_{i=1}^{n} x_i^T U \underbrace{U^T U}_{I} U^T x_i$$

$$= \sum_{i=1}^{n} x_i^T U U^T x_i$$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix} n \xrightarrow{ACP} C = \begin{bmatrix} c_1^T \\ \vdots \\ c_n^T \end{bmatrix} n$$

p (columns of X), q (columns of C)

$$q < p$$