Principal Component Analysis (PCA)

Aim: to create new variables to summarize the data. We would like to have less variables than previously.

→ be careful: no clustering with PCA!

Remember:

the data X:

$$\underset{n \text{ rows}}{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ & & & \\ x_{n1} & x_{n2} & & x_{np} \end{pmatrix}}$$

p columns

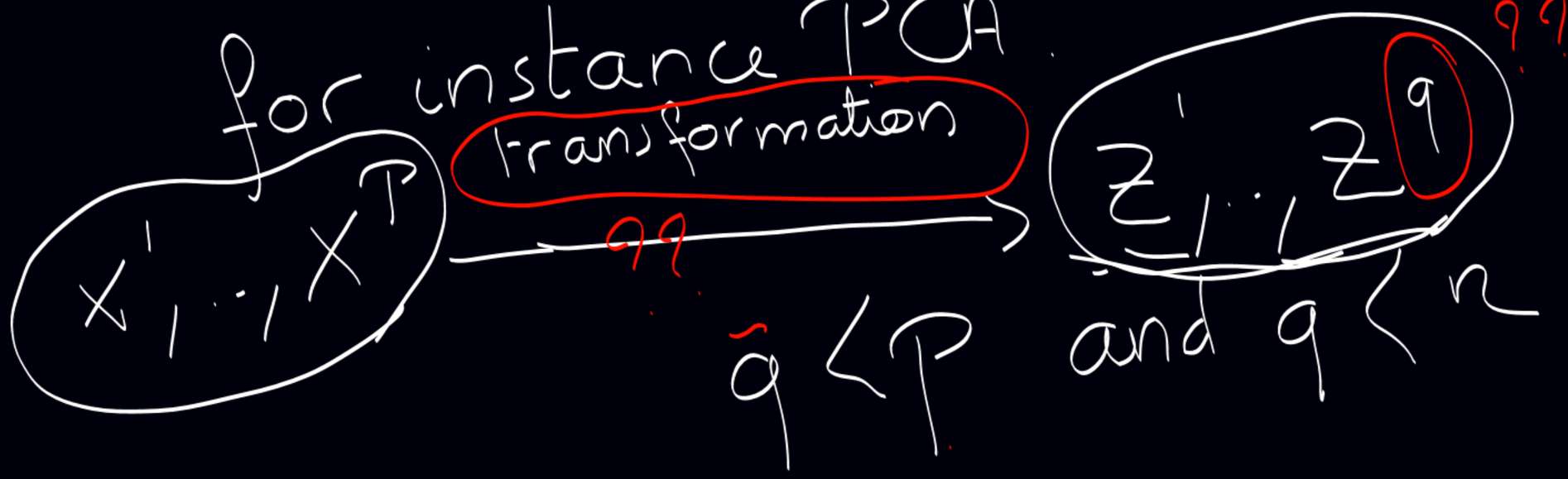
Now, we are often in the case where $p > n$.

If we want to perform linear regression

$\Rightarrow p^b$ because $p > n \Rightarrow$ unicity

A solution:

we are going to reduce the number of variables by using for instance PCA.



$\tilde{q} < p$ and $q < n$

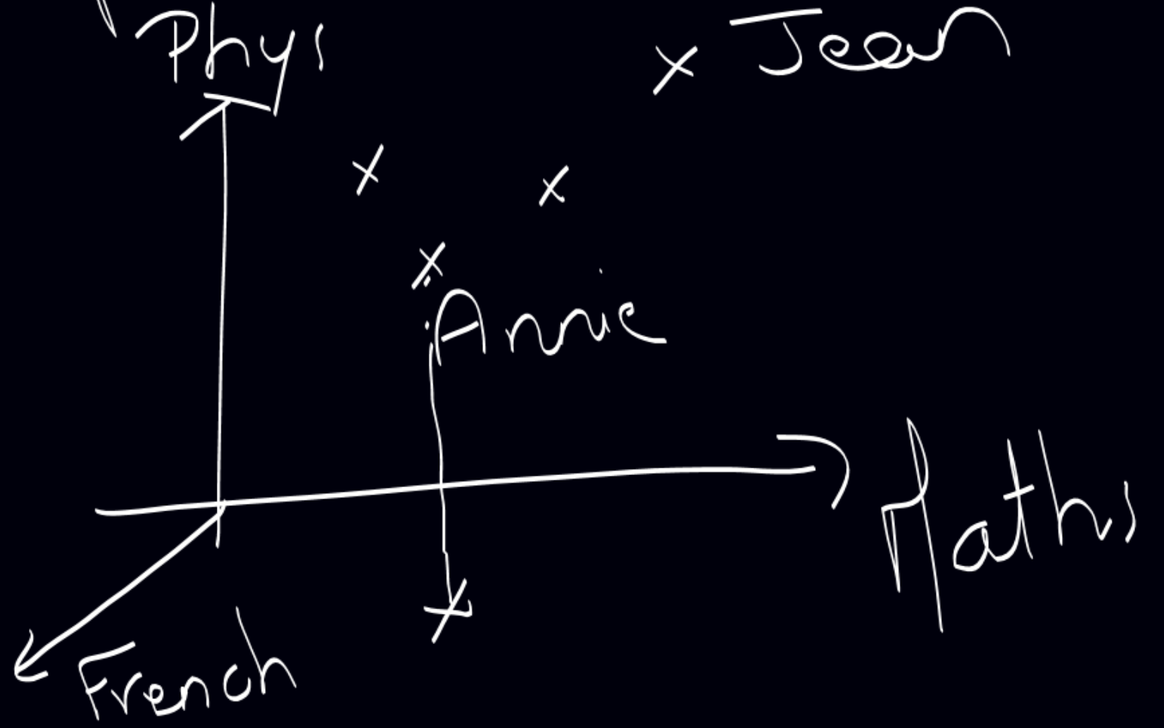
Example:

variables

students	Maths	Phys.	French	English
Jean	6.00	6.00	5.00	5.50
Alan	8.00	8.00	8.00	8.00
Annie	6.00	7.00	11.00	9.50
Monique	14.50	14.50	15.50	15.00
Didier	14.00	14.00	12.00	12.50
André	11.00	10.00	5.50	7.00
Pierre	5.50	7.00	14.00	11.50
Brigitte	13.00	12.50	8.50	9.50
Evelyne	9.00	9.50	12.50	12.00

An individual is a point of $\mathbb{R}^4$.

$\Rightarrow$ I am not able to make a drawing to represent the individuals



1°) Computation of usual statistics

	Mean	Standard deviation	Min.	Max.
Maths	9.67	3.37	5.50	14.50
Phys.	9.83	2.99	6.00	14.50
French	10.22	3.47	5.00	15.50
English	10.06	2.81	5.50	15.00

Is homogeneity among the variable!

Coefficients of correlation

$$Var(X^i) = c \times \sum_{k=1}^{n} \left(X_k^i - \bar{X}^i\right)^2$$

$$cov(X^i, X^j) = c \times \sum_{k=1}^{n} \left(X_k^i - \bar{X}^i\right)\left(X_k^j - \bar{X}^j\right)$$

$$cor(X^i, X^j) = \frac{c \times \sum_{k=1}^{n} \left(X_k^i - \bar{X}^i\right)\left(X_k^j - \bar{X}^j\right)}{c\sqrt{\left(\sum \left(X_k^i - \bar{X}^i\right)^2\right)\left(\sum \left(X_k^j - \bar{X}^j\right)^2\right)}}$$

Let X be the table with the data:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$\leftarrow$ observations for the individual $\downarrow$

$\uparrow$ observations for the variable $\downarrow$

two preprocessings.

- to center the data: always done!

$$X_c = \begin{pmatrix} x_{11} - x_{\cdot 1} & x_{12} - x_{\cdot 2} & \cdots & x_{1p} - x_{\cdot p} \\ \vdots & \vdots & & \vdots \\ x_{n1} - x_{\cdot 1} & x_{n2} - x_{\cdot 2} & & x_{np} - x_{\cdot p} \end{pmatrix}$$

$$\begin{pmatrix} a & b & c & d \\ a & b & c & d \\ \vdots & & & \\ a & b & c & d \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (a \; b \; c \; d)$$

where $x_{\cdot j} = \underbrace{\frac{1}{n}\sum_{k=1}^{n} x_{kj}}$

empirical mean of the variable number j

RB: The empirical mean of each variable of X_c is equal to 0.

$$\frac{1}{n}\sum_{k=1}^{n}\left(x_{kj} - x_{\cdot j}\right) = \frac{1}{n}\sum_{k=1}^{n} x_{kj} - \frac{1}{n}\sum_{k=1}^{n} x_{\cdot j}$$

$$= x_{\cdot j} - \frac{1}{n} \times n\, x_{\cdot j} = 0$$

2nd preprocessing : normalization

(not always done)

$$X_n = \begin{pmatrix} \frac{x_{11} - x_{.1}}{s_1} & \frac{x_{12} - x_{.2}}{s_2} & \cdots & \frac{x_{1p} - x_{.p}}{s_p} \\ \frac{x_{21} - x_{.1}}{s_1} & \frac{x_{22} - x_{.2}}{s_2} & & \frac{x_{2p} - x_{.p}}{s_p} \\ \vdots & & & \vdots \\ \frac{x_{n1} - x_{.1}}{s_1} & \frac{x_{n2} - x_{.2}}{s_2} & & \frac{x_{np} - x_{.p}}{s_p} \end{pmatrix}$$

where $s_j^2 = \frac{1}{n} \sum_{k=1}^{n} (x_{kj} - x_{\cdot j})^2$

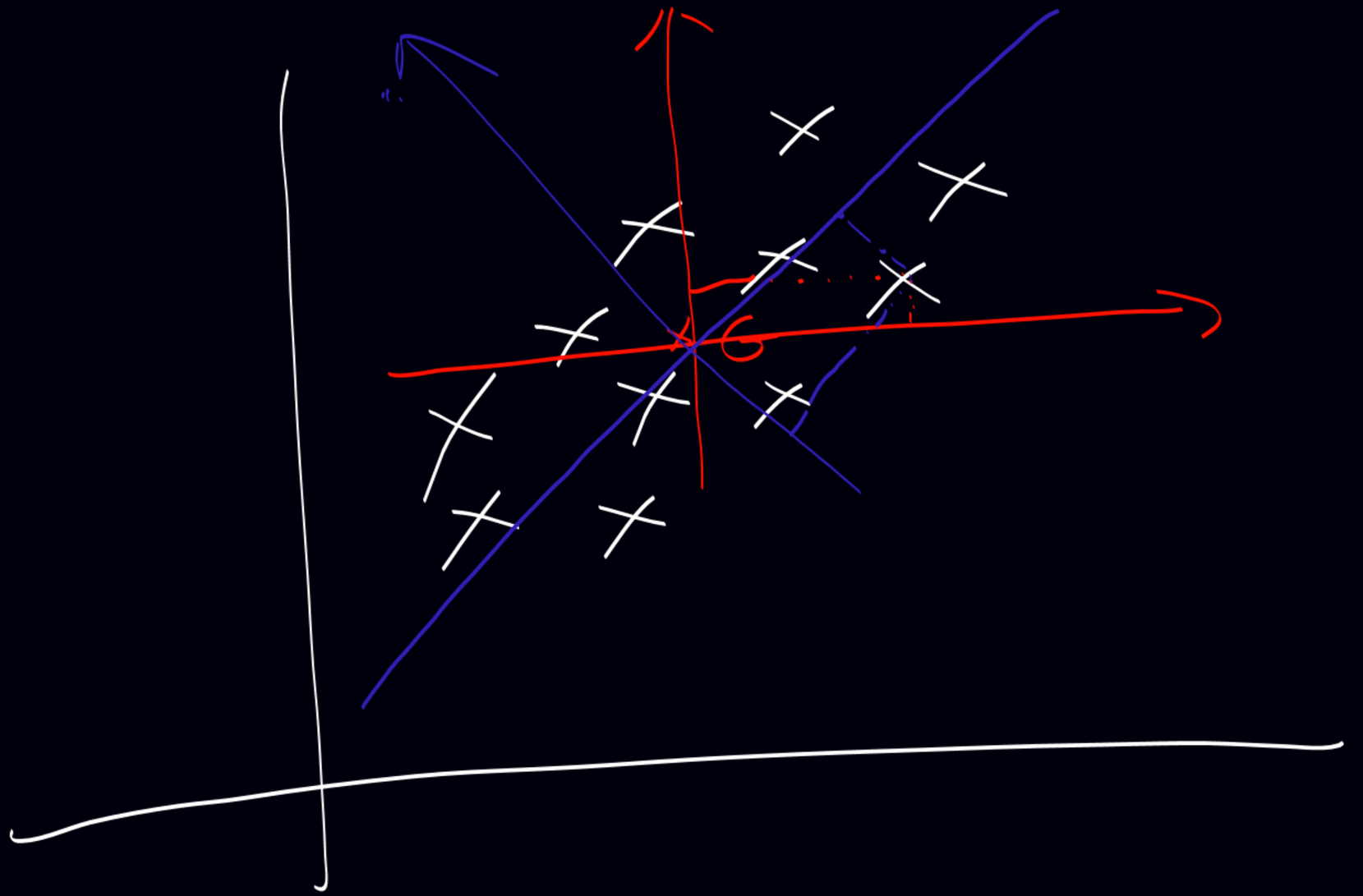
Matrix of data X

X_c : the centered version

$V = \frac{1}{n} {}^t X_c X_c \rightarrow$ the covariance matrix

X_n : the normalized version

$R = \frac{1}{n} {}^t X_n X_n \rightarrow$ correlation matrix



Compute the eigenvalues and the eigenvectors of

V and R

$V = \frac{1}{n} {}^{t}X_c X_c$ and $R = \frac{1}{n} {}^{t}X_n X_n$

↳ eigen

$$cov(X^i, X^j) = \frac{1}{n} \sum_{k=1}^{n} (x_{ki} - x_{\cdot i})(x_{kj} - x_{\cdot j})$$

$$\left(\frac{1}{n} X_c^t X_c\right)_{ij} = \frac{1}{n} \times \text{product between the the column } i \text{ and the column } j \text{ of } X_c$$

$\times \quad \mathbb{R}^3$

$$A:(x,y,z) = x \times (1,0,0) + y \times (0,1,0) + z \times (0,0,1)$$

$\{(1,0,0),(0,1,0),(0,0,1)\}$: all the points of $\mathbb{R}^3$ can be written as a linear combination of $\{e_1, e_2, e_3\}$

$e_1 = (1,0,0)$, $e_2 = (0,1,0)$, $e_3 = (0,0,1)$

the family $\{e_1, e_2, e_3\}$ generates $\mathbb{R}^3$.

To be able to say that $\{e_1, e_2, e_3\}$ is a basis of $\mathbb{R}^3$, the linear combination should be unique.

A way to show this is to prove that

if $(0,0,0) = \lambda_1 \times e_1 + \lambda_2 e_2 + \lambda_3 e_3 \Rightarrow \lambda_1 = \lambda_2 = \lambda_3 = 0$

$A \rightarrow$ | eigenvalues $\lambda_1, \ldots, \lambda_p$

| eigenvectors $u_1, \ldots, u_p$

If $\lambda_i \neq \lambda_j$ $u_i \perp u_j$

$$\sum_{k=1}^{n} u_{i,k} \times u_{j,k} = 0$$

If $\lambda_i = \lambda_j$

v_i and v_j is not necessary orthogonal

There exists a tecnic named Gram-Schmidt that allow to create v_i and v_j eigenvectors of A associated to the eigenvalue λ_i such that $v_i \perp v_j$

$$u_1 = (1,0)$$
$$u_2 = (1,1)$$

$$\Rightarrow \quad v_1 = (1,0)$$
$$v_2 = (0,1)$$

$$V = PDP^{-1}$$

with P composed of the eigenvectors

$$P^{-1} = {}^t(P)$$

Now what is PC scores?

$$X_c = \begin{pmatrix} \text{ind1} \\ \text{ind2} \\ \vdots \\ \text{indn} \end{pmatrix}$$

Eigenvectors

$$\begin{pmatrix} v1 & v2 & v3 & v4 \\ | & | & | & | \\ | & | & | & | \end{pmatrix}$$

X_c % * % Eigenvectors

Rk:

A is a matrix

λ is an eigenvalue for A if there exists a vector X which is not null if $AX = \lambda X$

X : eigenvector associated to λ

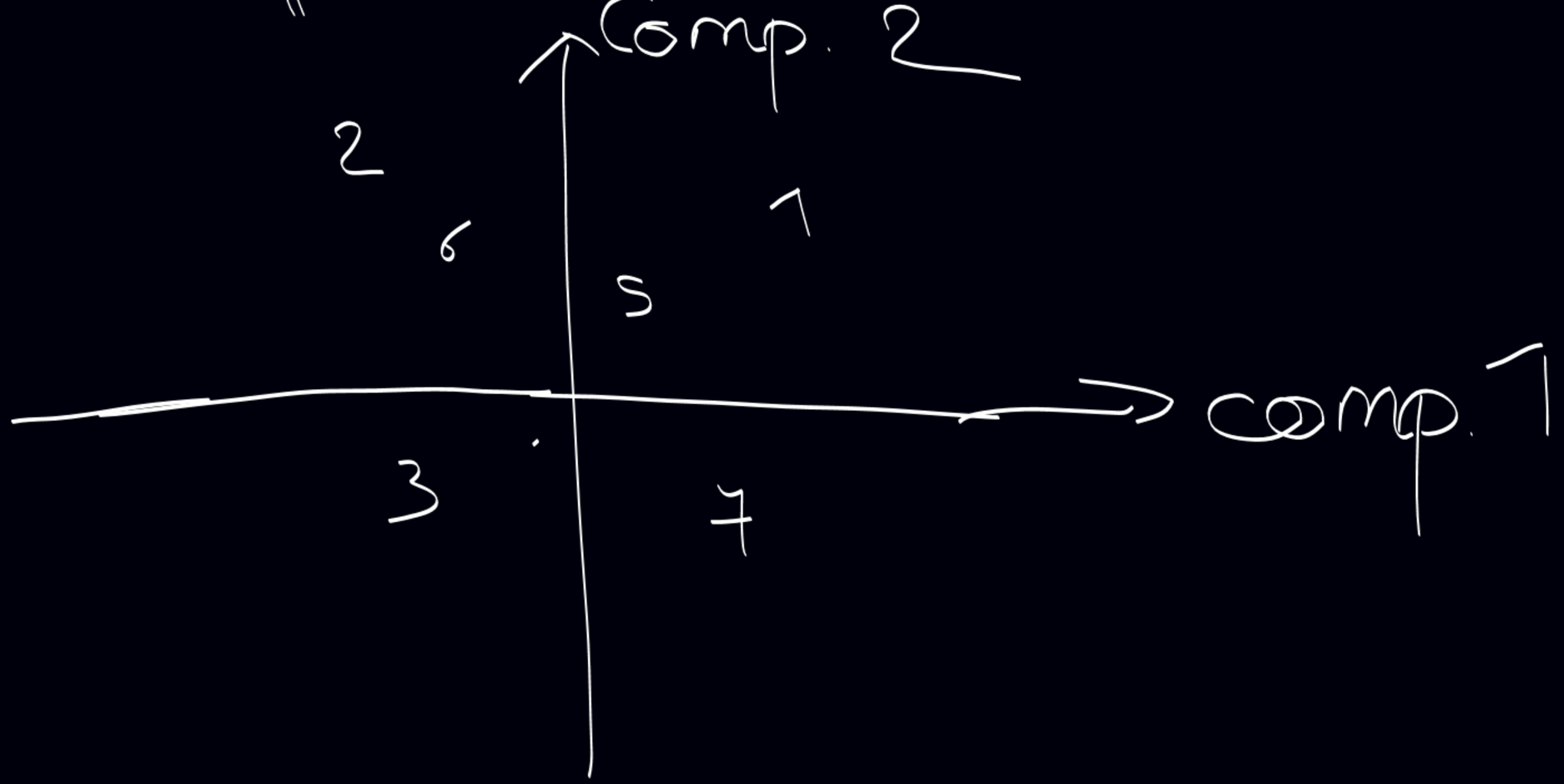
Let μ be a real

If X is an eigenvector for A associated to λ

$\Rightarrow \mu X$ is also an eigenvector for A associated to λ.

In particular $-X$ is an eigenvector

To represent the data



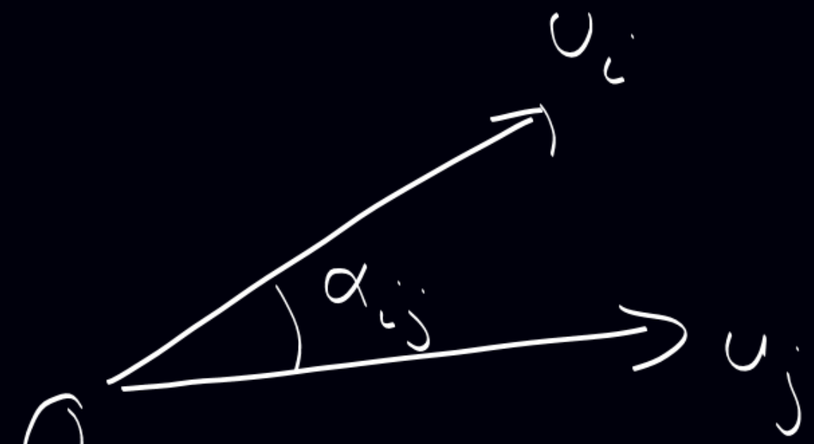
notion of Inertia

↳ associated to the notion of distance

let $u_i = \begin{pmatrix} u_{i1} \\ \vdots \\ u_{ip} \end{pmatrix}$ $u_j = \begin{pmatrix} u_{j1} \\ \vdots \\ u_{jp} \end{pmatrix}$ $d^2(u_i, u_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$

↓ scalar product

$\langle \overrightarrow{Ou_i}, \overrightarrow{Ou_j} \rangle = u_i u_j$



$$\cos(\alpha_{ij}) = \frac{\langle \overrightarrow{Ou_i}, \overrightarrow{Ou_j} \rangle}{\|\overrightarrow{Ou_i}\| \times \|\overrightarrow{Ou_j}\|}$$

with $\|\overrightarrow{Ou_i}\|^2 = \langle \overrightarrow{Ou_i}, \overrightarrow{Ou_i} \rangle$

$$X_c = \begin{pmatrix} {}^tU_1 \leftarrow & x_{11} - x_{\cdot 1} & x_{12} - x_{\cdot 2} & \cdots & x_{1p} - x_{\cdot p} \\ {}^tU_2 \leftarrow & \vdots & \vdots & & \vdots \\ {}^tU_n \leftarrow & x_{n1} - x_{\cdot 1} & x_{n2} - x_{\cdot 2} & & x_{np} - x_{\cdot p} \end{pmatrix}$$

V_1 V_2 V_p

$$U_1 = \begin{pmatrix} x_{11} - x_{\cdot 1} \\ x_{12} - x_{\cdot 2} \\ \vdots \\ x_{1p} - x_{\cdot p} \end{pmatrix}$$

Inertia with respect to the mean point of X_c:

$$\overline{I_O} = \frac{1}{n}\sum_{i=1}^{n} d^2(O, u_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - x_{\cdot j}\right)^2$$

$$= \sum_{j=1}^{p} \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_{\cdot j} \right)^2 \right)}_{\text{Variance of } V_j}$$

$$I_0 = \sum_{j=1}^{p} V[V_j] = tr(\Sigma)$$

where Σ : matrix of variance-covariance

let Δ be an axis such that

$$0 \in \Delta$$

$$I_\Delta = \frac{1}{n} \sum_{i=1}^{n} d^2(h_{\Delta,i}, v_i)$$

with $h_{\Delta,i}$: projection of v_i on Δ

let V denotes a subspace of $\mathbb{R}^p$

and V^* ——— the orthogonal supplementary of V in $\mathbb{R}^p$

Rq: $0 \in V$

$$I_V = \frac{1}{n} \sum_{i=1}^{\hat{}} d(h_{V,i}, v_i)^2$$

with $h_{V,i}$: projection of v_i on V

We can prove that:

$$d^2\left(h_{V,i}, U_i\right) + d^2\left(h_{V^*,i}, U_i\right) = d^2\left(O, U_i\right)$$

$$= d^2\left(O, h_{V_i}\right) + d^2\left(O, h_{V^*_i}\right)$$

Because of this:

$$\overline{I}_O = \overline{I}_V + \overline{I}_{V^*} \overset{\text{generalization}}{=} \overline{I}_{\Delta_1} + \overline{I}_{\Delta_2} + \cdots + \overline{I}_{\Delta_p}$$

all Δ_i are with dimension 1 and orthogonal

1st Step : determination of Δ_1

⇒ idea : Δ_1 should be such that

Δ_1 is as close as possible to

the cloud of points.

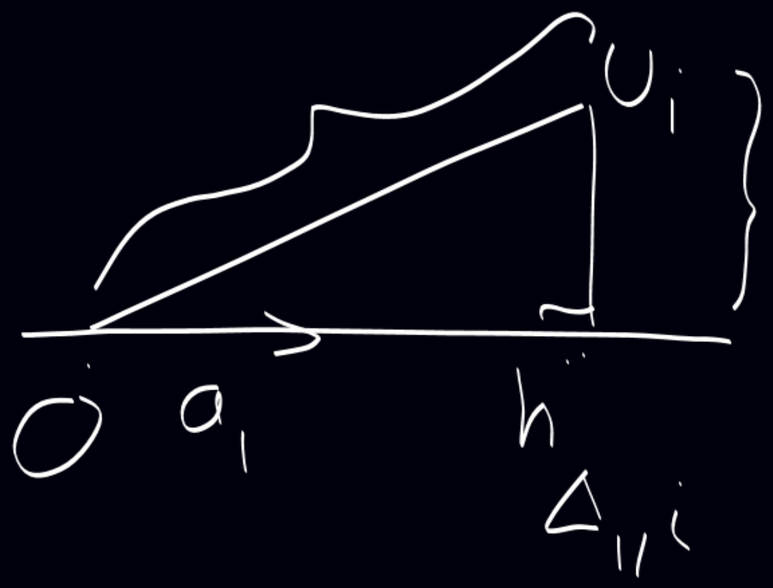
we search Δ_1 such that :

- $O \in \Delta_1$
- I_{Δ_1} minimum ($\Rightarrow I_{\Delta_1^*}$ maximum

Δ_1 is defined by a unitary vector of $\mathbb{R}^p$ denoted $\vec{a}_1$

To determine $\vec{a}_1$ we should have $I_{\Delta_1^*}$ maximum and $\|\vec{Oa}_1\|^2 = 1$

$$\langle \overrightarrow{Oa_1}, \overrightarrow{Oh_{\Delta_{1,i}}} \rangle^2 = d^2\left(O, h_{\Delta_{1,i}}\right)^2$$



$$= \langle \overrightarrow{Oa_1}, \overrightarrow{Ou_i} \rangle$$

$$= {}^t a_1 u_i \; {}^t u_i a_1$$

$$I_{\Delta_i^\varphi} = \frac{1}{n} \sum_{i=1}^{n} \left({}^t a_1 U_i \, {}^t U_i \, a_1 \right)$$

$$= {}^t a_1 \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} U_i \, {}^t U_i \right)}_{\Sigma} a_1$$

Σ : matrice de variance covariance

Optimization p^b:

maximize ${}^t a_1 \Sigma a_1$ with ${}^t a_1 a_1 = 1$

$$g(a_1) = {}^t a_1 \Sigma a_1 - \lambda ({}^t a_1 a_1 - 1)$$

$$\frac{\partial g(a_1)}{\partial a_1} = 2 \Sigma a_1 - 2 \lambda a_1$$

Rk: $a_1 = (a_{11}, \ldots, a_{1p})$

$$\frac{\partial g(a_1)}{\partial a_1} = \begin{pmatrix} \frac{\partial g}{\partial a_{11}} \\ \vdots \\ \frac{\partial g}{\partial a_{1p}} \end{pmatrix}$$

(1) $\Sigma a_1 = \lambda a_1 \iff a_1$ is an eigen-vector of Σ associated to the eigenvalue λ.

$${}^t a_1 a_1 = 1$$

(1) $\Rightarrow \underbrace{{}^t a_1 \Sigma a_1}_{\text{quantity that we want to maximize}} = \lambda \, {}^t a_1 a_1 = \lambda$

$\Rightarrow \lambda$ is the biggest eigenvalue of Σ

In conclusion:

a_1 is a unitary eigenvector associated to the biggest eigenvalue of Σ.

olympic2.txt

TDDSTI.pdf

project1.pdf → groups of 4 persons!