let $X$ be the matrix with the initial data.

let $X_c$ be the matrix with the centered data

let $X_n$ be the matrix with the normalized data

let

$$T = \begin{cases} \frac{1}{n} X_c^T X_c & : \text{non normalized PCA} \\ \frac{1}{n} X_n^T X_n & : \text{normalized PCA} \end{cases}$$

let denote $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ : the eigenvalues of $T$ in decreasing order

let denote $a_k$ a unitary eigenvector of $T$ associated to $\lambda_k$

Rk: the fact that $a_k$ is a unitary vector is just $^t a_k a_k = 1$

Let denote $V_1, \dots V_p$ the p initial variables (the p columns of $X_c$ or $X_n$)

$Z_k = X_c a_k$ ← non normalized PCA

$\underline{X_n a_k}$ ← normalized PCA

the new variable number k

$$Var(z_k) = \frac{1}{n} {}^t z_k z_k$$

If we consider the non normalized PCA

$$Var(z_k) = \frac{1}{n} {}^t a_k {}^t X_c X_c a_k$$

$$= {}^t a_k \left( \frac{1}{n} {}^t X_c X_c \right) a_k$$

$$= {}^t a_k V a_k \quad \text{with } V \text{ the covariance matrix}$$

but $Va_k = \lambda_k \cdot a_k$

Thus

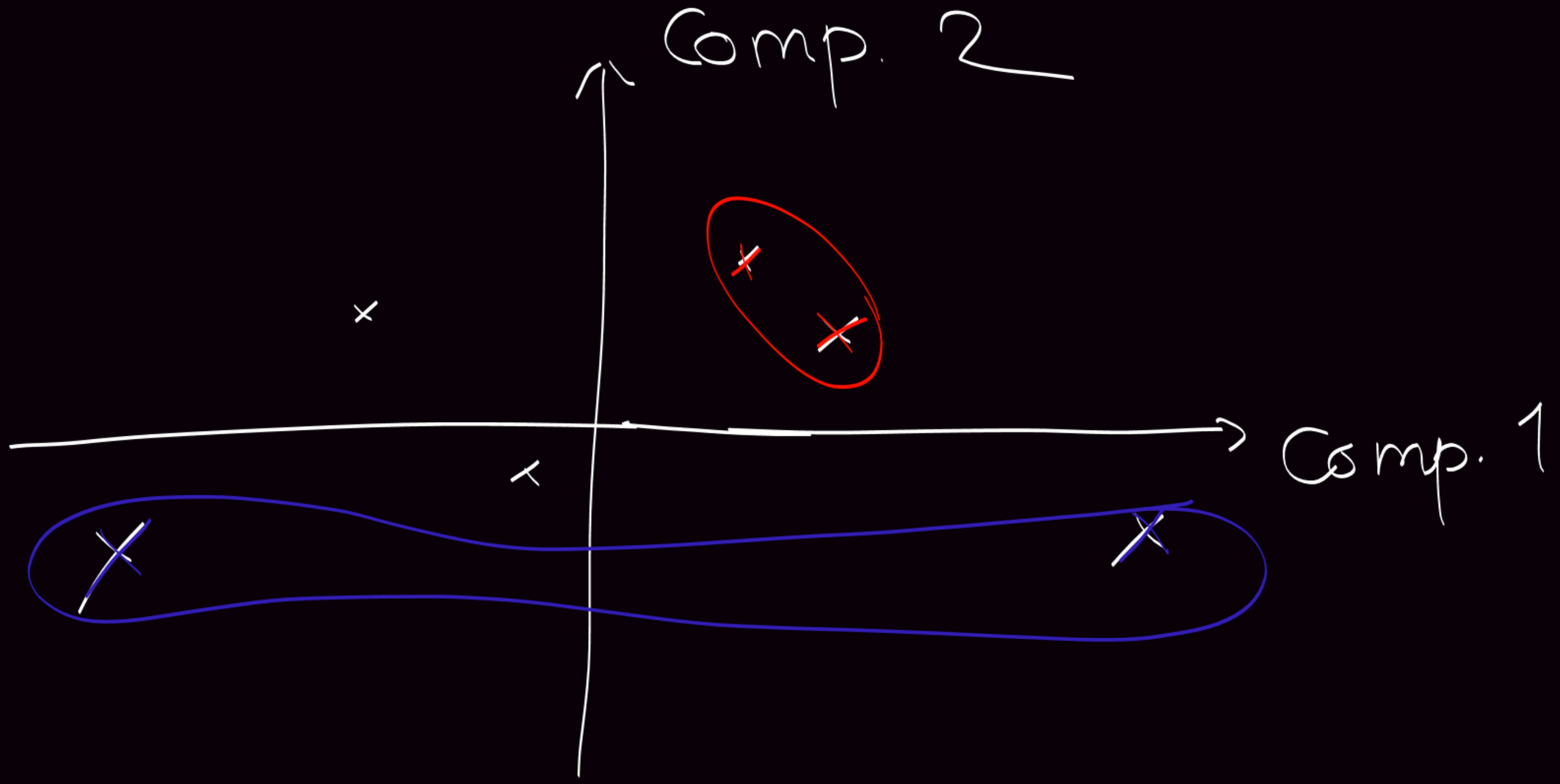
$$Var(z_k) = {}^t a_k \lambda_k a_k$$

$$= \lambda_k {}^t a_k a_k$$

$$= \lambda_k$$

$\sqrt{\lambda_k}$ = standard deviation of the new variable number $k$

This explains why in P$ sdev (where P = princomp(X)) we see the term standard deviation.


Comp. 2
Comp. 1

be careful:

let $v_i$ and $v_j$ be two individuals

let $h_i$ and $h_j$ be their projections

$$d^2(h_i, h_j) \leq d^2(v_i, v_j)$$

If on the first factorial plan, two projected points are far then, the two individuals are far in the global space

But if to projected points are closed in the first factorial plan, the associated individuals may be far in the global space.

⇒ we need to quantify the quality of the representation of an individual on an axis or

on a plan.

the idea is to consider the angle formed by the vectors $\overrightarrow{Ou_i}$ and $\overrightarrow{Oh_i}$

If this angle is small $\Rightarrow$ the projection is a good representation of the individual

For simplicity, we consider in fact the cosinus of the angle, and better the cosinus square!

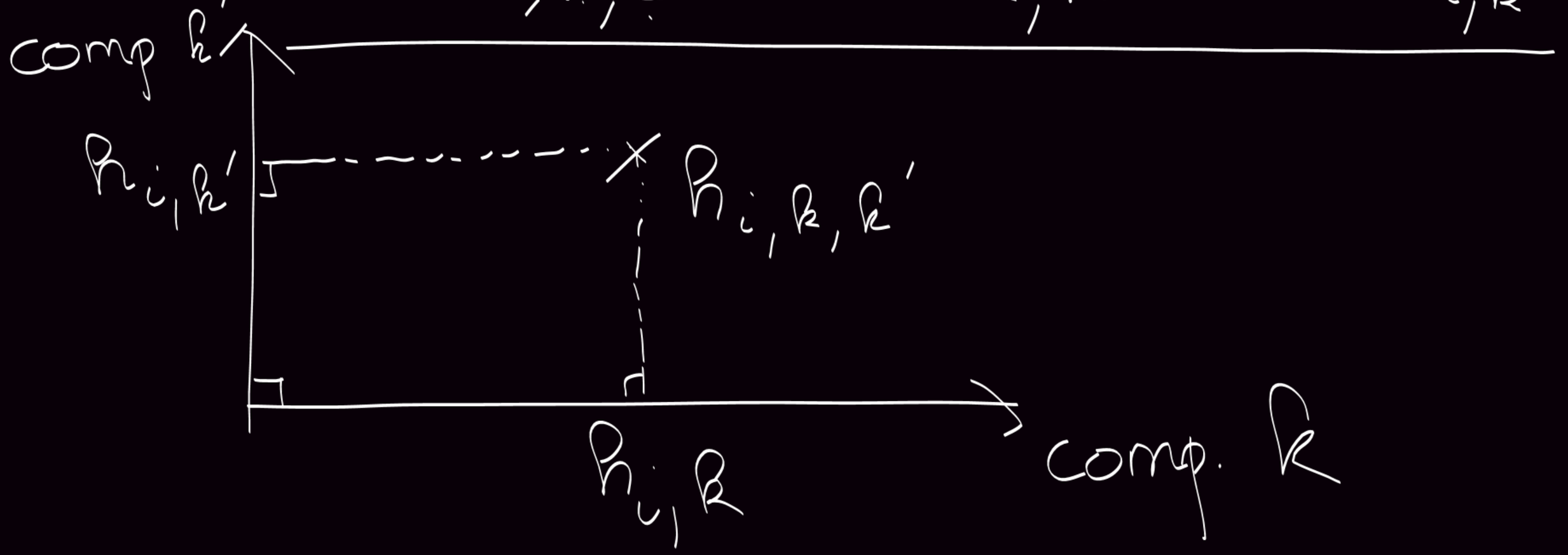
$\Rightarrow$ if $\cos^2$ near 1 $\Rightarrow$ good representation!

Rk :

because the factorial plan are generated by 2 orthogonal vectors The $\cos^2$ of the angle with the plan is equal to sum $\cos^2$ of the angle with each generated vector

consider the factorial plan $(k, k')$
let denote $\alpha_{i,k,k'}$ the angle between
the individual $v_i$ and its projection
on the plan $(k, k')$
the denote $\alpha_{i,k}$ the angle between $v_i$
and its projection on the component $k$

$$\cos^2 \alpha_{i,k,k'} = \cos^2 \alpha_{i,k} + \cos^2 \alpha_{i,k'}$$



Rk: how to compute $\cos^2 \alpha_{i,k}$ ?

$$\cos^2 \alpha_{i,k} = \frac{\langle \overrightarrow{O u_i}, \overrightarrow{O a_k} \rangle^2}{\|\overrightarrow{O u_i}\|^2 \|\overrightarrow{O a_k}\|^2}$$

where $U_c$ : is the vector associated to the individual i in the matrix $X_c$

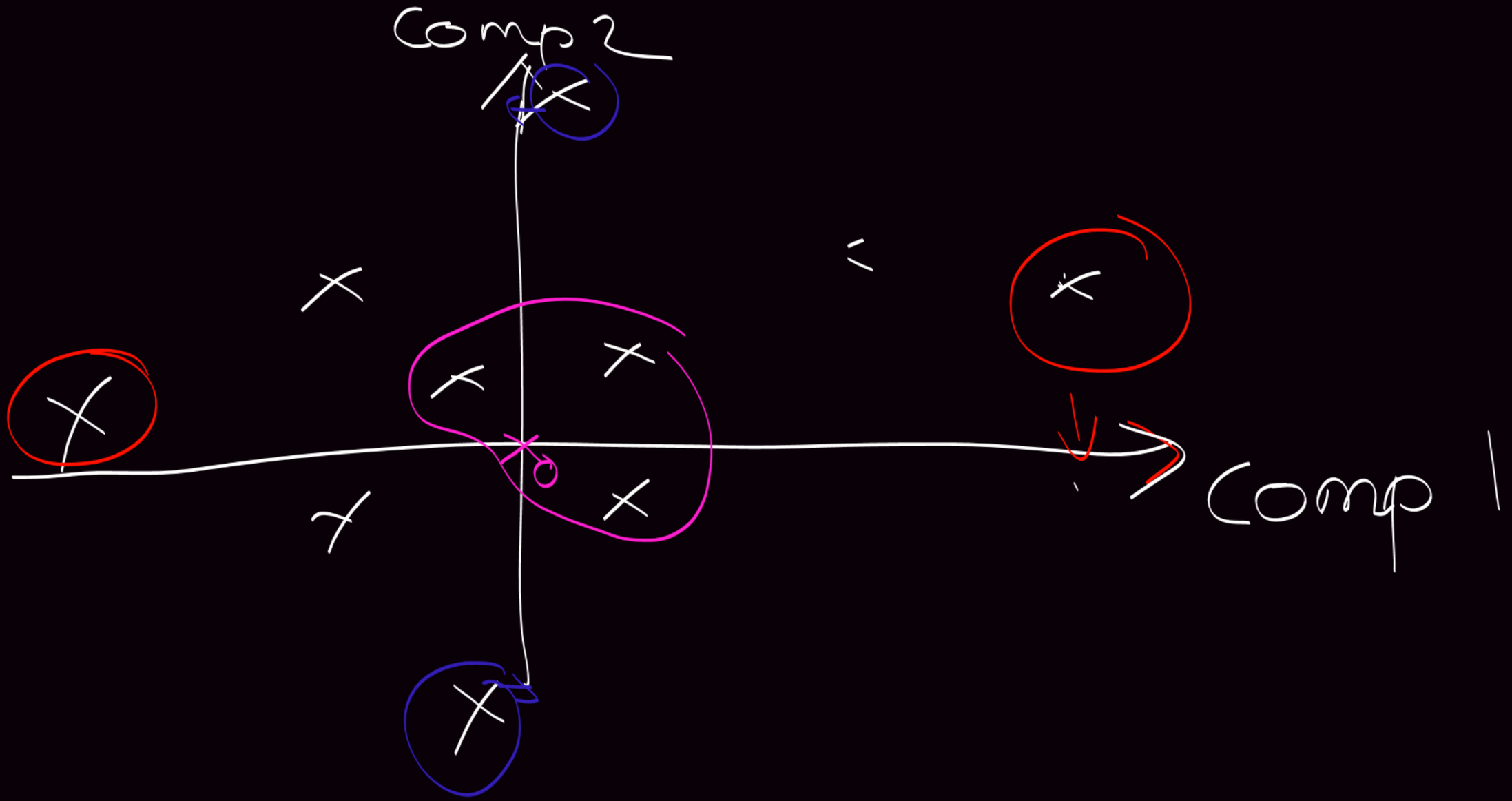
$$= \frac{(t_{a_k} U_{c_i}) \times (t U_c, a_k)}{t U_{c,i} \times U_{c_i}}$$

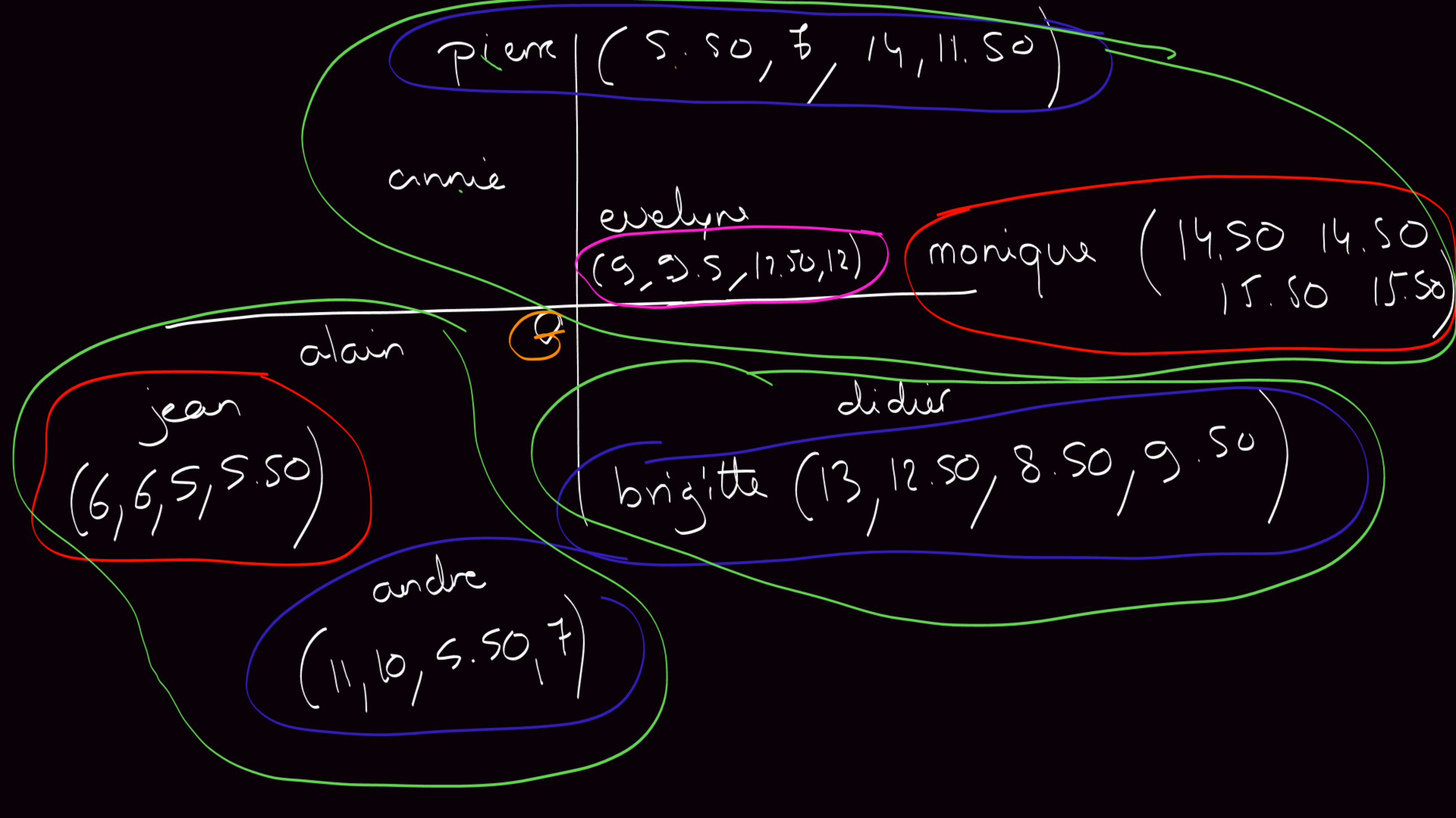
# Contribution of an individual to an axis $\Delta$

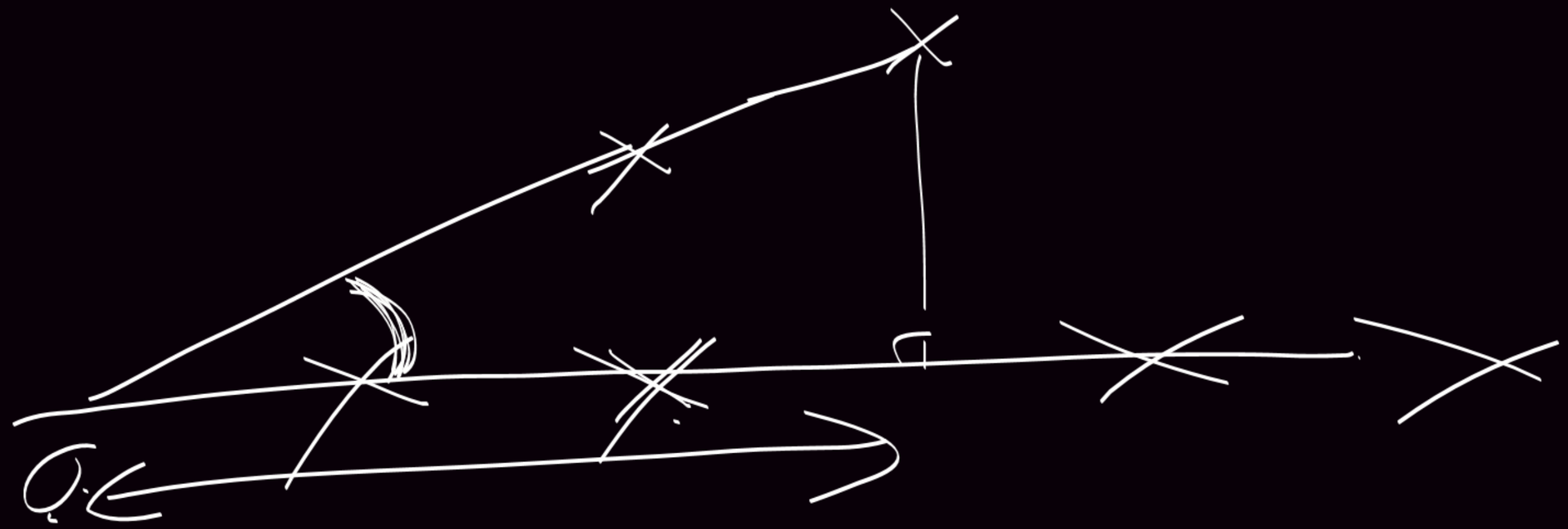
The contribution of the individual $i$ on $\Delta$ is

$$\frac{1}{n} d^2(O, P_{\Delta,i})$$

$P_{\Delta,i}$: the projection of individual $i$ and $\Delta$


Comp 2
=
Comp 1


pierre (5.50, 7, 14, 11.50)
annie
evelyne
(9, 9.5, 12.50, 12)
monique (14.50 14.50
15.50 15.50
alain
jean
(6, 6, 5, 5.50)
didier
brigitte (13, 12.50, 8.50, 9.50)
andre
(11, 10, 5.50, 7)


Q

# What happens for the variable?

$V_1, \dots, V_p \longrightarrow Z_1, \dots, Z_p$ $\qquad Z_k = X_c a_k$

(centered)

$\mathrm{Var}(Z_k) = \lambda_k$

$$\mathrm{cov}(V_j, Z_k) = \frac{1}{n} {}^t a_k {}^t X_c V_j$$

$$= \frac{1}{n} {}^t a_k {}^t X_c X_c \times \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow j$$

$$\mathrm{cov}(Z_k, V_j) = {}^{t}a_k V \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

matrice de covariance de $X$ or matrice de covariance de $X_c$

$$= {}^{t}\left({}^{t}V a_k\right) \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$= L\left(V a_k\right)\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}$$

$$= L\left(\lambda_k a_k\right)\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}$$

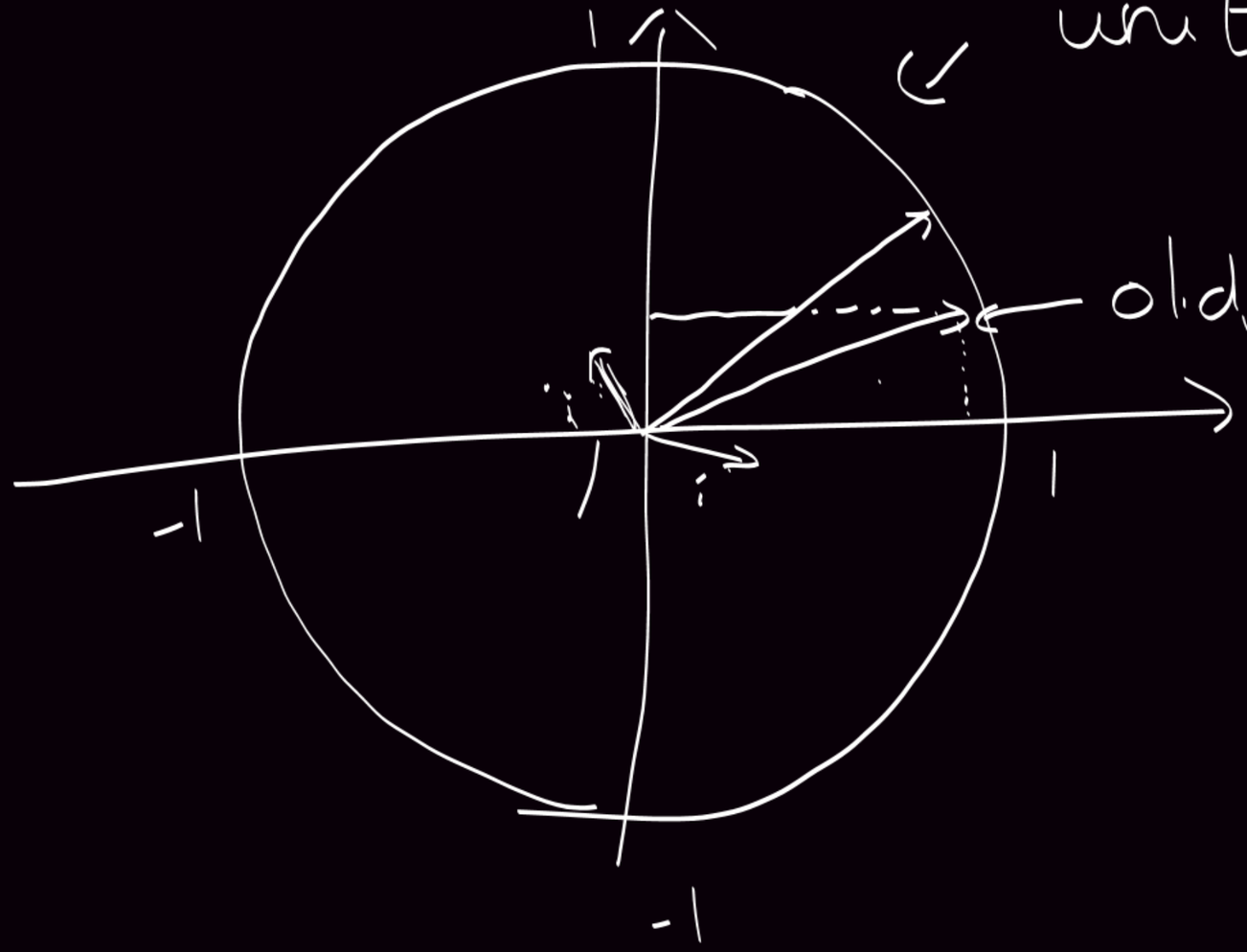
$$= \lambda_k L a_k \begin{pmatrix}0\\0\\\vdots\\1\\0\end{pmatrix} = \lambda_k \underbrace{a_{kj}}_{\text{element } j \text{ of } a_k}$$

Thus

$$Cor(Z_k, V_j) = \frac{\lambda_k a_{kj}}{\sqrt{\lambda_k}\sqrt{Var\, V_j}} = \sqrt{\lambda_k}\frac{a_{kj}}{\sqrt{Var\, V_j}}$$

$$Var(V_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{\cdot j})^2 \Rightarrow \text{square of a norm}$$

correlation : scalar product


1
unitary circle
old variables
-1
1
-1

$$\underline{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & & x_{np} \end{pmatrix}$$

$p >> n$

regression on PCA

PCA

$q < n$

$$\begin{pmatrix} z_{11} & \cdots & z_{1q} \\ \vdots & & \vdots \\ z_{n1} & & z_{nq} \end{pmatrix}$$

# Tradeoff : Partial least Square

# Clustering

aim: We have n individuals
can we determine several
groups among the n individuals?

Several methods

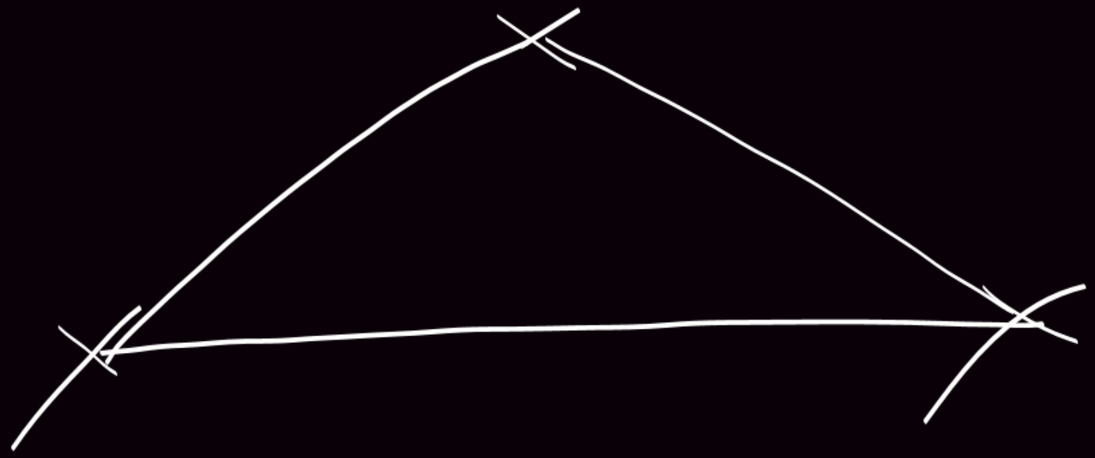
↳ all are based on the notion of distance

$T^b$: several distances!

Let $\mathcal{X}$ be the space of the observations.

A distance $d$ on $\mathcal{X}$ is a application from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}^+$ such that:

- $\forall x, y \in \mathcal{X} \quad d(x,y) = d(y,x)$
- $\forall x, y \in \mathcal{X}, \; d(x,y) = 0 \Leftrightarrow x = y$
- $\forall x, y, z \in \mathcal{X}, \; d(x,y) \leq d(x,z) + d(z,y)$



$$X = \mathbb{R}^p$$

**euclidian distance:**

$$\forall x, y \in \mathbb{R}^p, \quad d_2^2(x, y) = \sum_{i=1}^{p} (x_i - y_i)^2$$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}$$

## $l_1$ - distance

$$d_1(x,y) = \sum_{i=1}^{p} |x_i - y_i|$$

## $l_\infty$ - distance

$$d_\infty(x,y) = \max_{i \in \{1,\dots,p\}} \{|x_i - y_i|\}$$

# Exercise:

We take $p = 2$

Define $B(0,1) = \{x = (x_1, x_2) \in \mathbb{R}^2 / d(0,x) \leq 1\}$

for $d = \begin{vmatrix} d_1 \\ d_2 \\ d_\infty \end{vmatrix}$

$d_2$ :

$B(0;1)$

center, radius

$d_2(0,x) \leq 1 \iff d_2^2(0,x) \leq 1$

$\iff (x_1-0)^2 + (x_2-0)^2 \leq 1$

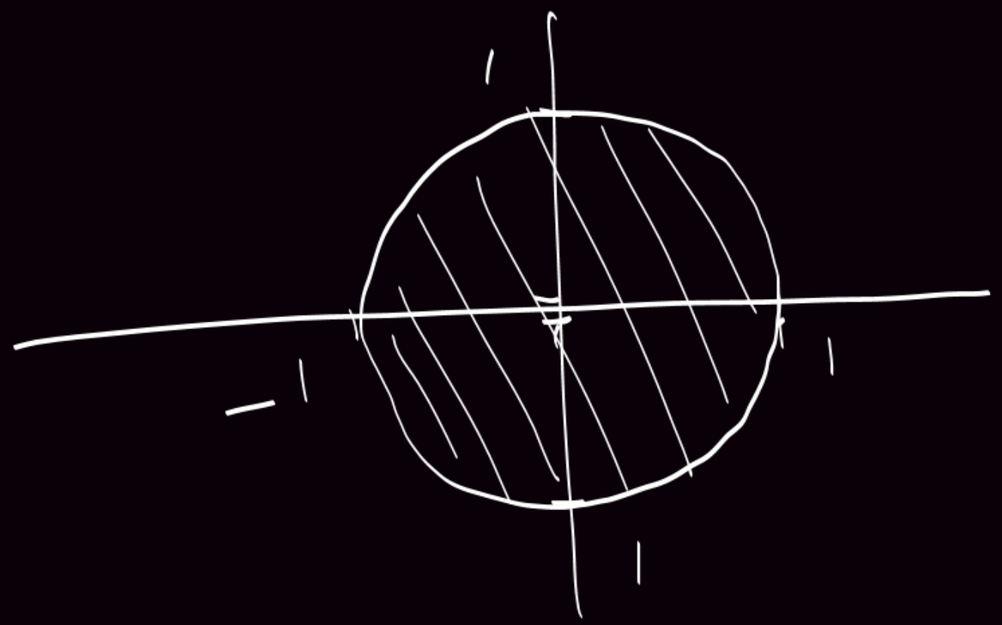
$\iff x_1^2 + x_2^2 \leq 1$

$$x_1^2 + x_2^2 = 1$$

↳ equation of the circle whose center is 0 and radius 1

Rk: $(x_1 - 2)^2 + (x_2 - 4)^2 = 2$

⇒ circle whose center is $\begin{pmatrix} 2 \\ 4 \end{pmatrix}$ and radius $\sqrt{2}$




euclidian distance

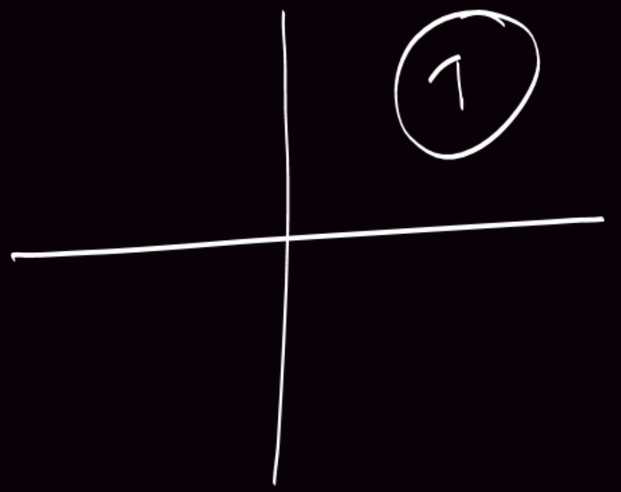
$d_1$ :

$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ such that $|x_1 - 0| + |x_2 - 0| \leq 1$

$\Rightarrow |x_1| + |x_2| \leq 1$

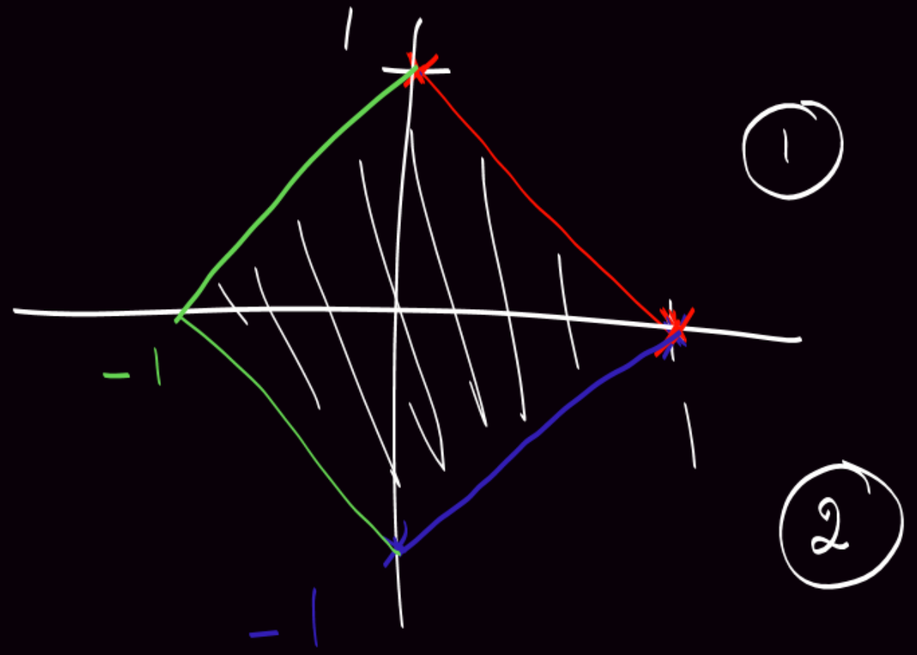
1st case: $x_1 \geq 0, x_2 \geq 0$

$\Rightarrow x_1 + x_2 \leq 1$



We solve $x_1 + x_2 = 1$

$\Leftrightarrow \quad x_2 = 1 - x_1$



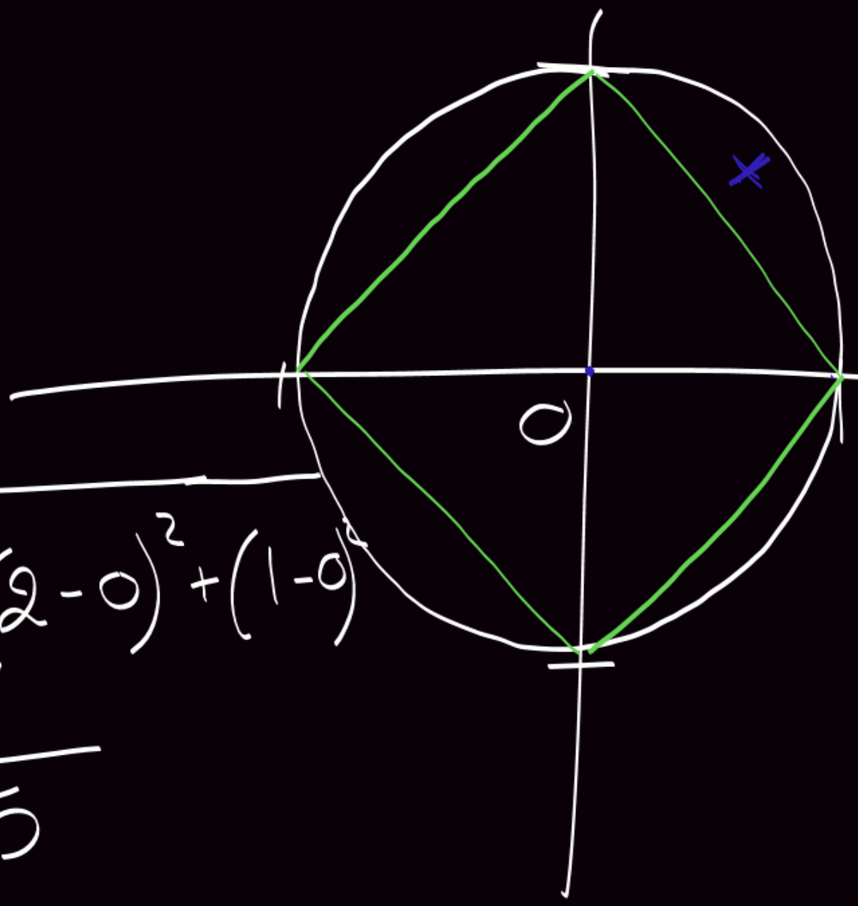

<u>2 nd case</u>: $x_1 \geqslant 0$ and $x_2 \leqslant 0$

we solve: $|x_1| + |x_2| = 1$

$x_1 + (-x_2) = 1$

# R2:

$$x = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$




$$d_2(0,x) = \sqrt{(2-0)^2 + (1-0)^2}$$

$$= \sqrt{5}$$

$$d_1(0,x) = |2-0| + |1-0|$$

$$= 3$$

$\underline{d_\infty}$

$$d_\infty(0, x) = \max\{|x_1 - 0|, |x_2 - 0|\}$$

$$= \max\{|x_1|, |x_2|\}$$

$$\leq 1$$

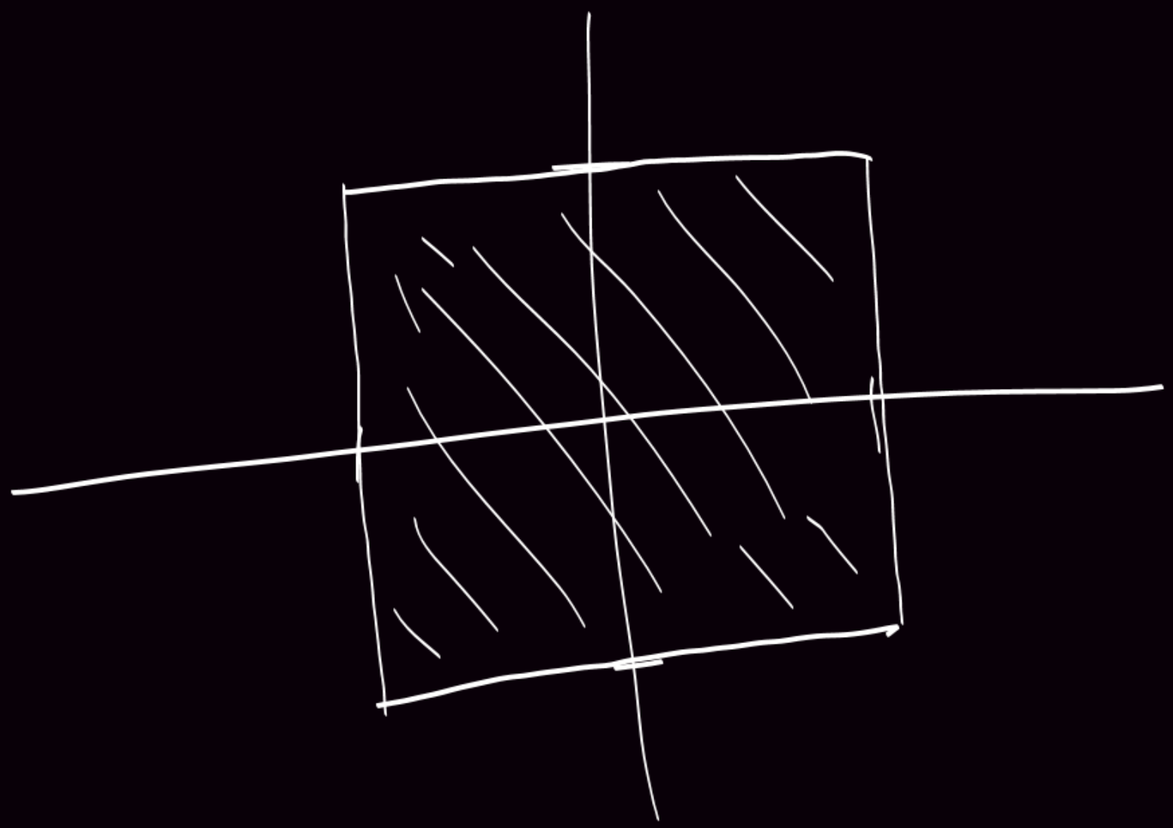
$$\max\{|x_1|, |x_2|\} = 1$$

$$\Leftrightarrow \quad |x_1| = 1 \text{ and } |x_2| \leq 1$$

$$\text{or}$$

$$|x_1| \leq 1 \text{ and } |x_2| = 1$$

$$\Leftrightarrow \Big( \big( (x_1 = 1 \text{ or } x_1 = -1) \text{ and } (-1 < x_2 < 1) \big) \text{ or } \big( (-1 < x_1 < 1) \text{ and } (x_2 = -1 \text{ or } x_2 = 1) \big) \Big)$$



- Ascending Hierarchical Classification
- k-means

# Ascending Classification

This is a reccursive method that depends on to distances:

- individual distance
- group distance

- individual distance is used to compute distance between elementary individuals.

But as we are making groups we need to compute distances between groups!

⇒ Group distance

× group distance

× Ward : let A and B be two groups

let $n_A$ the number of elementary individuals in A

and $n_B$ : the same for B

the $G_A$ the mean point of A

$G_B$ ——— B

$$d_W(A,B) = \frac{n_A \times n_B}{n_A + n_B} \times d^2(G_A, G_B)$$

↑ individual distance

- minimal link:

$$d_{mi}(A,B) = \min_{\substack{x \in A \\ y \in B}} d(x,y)$$

- maximal link:

$$d_{ma}(A,B) = \max_{\substack{x \in A \\ y \in B}} d(x,y)$$

- average link

$$d_{av}(A, B) = d(G_A, G_B)$$

# description of the method

Step 1: we have n groups, each group contains 1 individual.

We compute all the possible distance between two individuals

We look at the **two** individuals associated to the smallest computed distance, not located on the diagonal!

→ We put those two individuals in a group

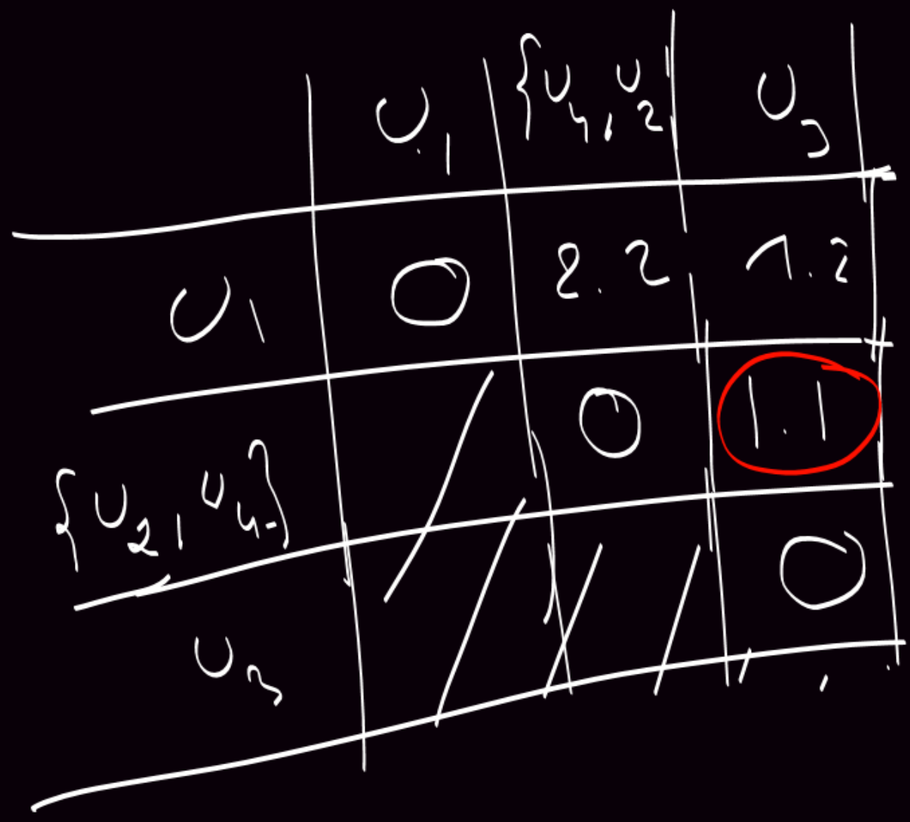
⇒ at the end of step 1, (n-2) groups with 1 individual, and 1 group with 2.

4 individuals.

| | $u_1$ | $u_2$ | $u_3$ | $u_4$ |
|---|---|---|---|---|
| $u_1$ | 0 | 0.5 | 1.2 | 0.42 |
| $u_2$ | | 0 | 1.7 | 0.27 |
| $u_3$ | | | 0 | 1.4 |
| $u_4$ | | | | 0 |

→ we group $u_4$ and $u_2$

4 inclusion chain.




$\Rightarrow$ we group $\{v_2, v_4\}$ with $v_3$

## Step 2:

We compute the distance between the (n-1) groups.

→ smallest distance

→ group

Rk:

Total variance with n points:

$$\frac{1}{n}\sum_{i=1}^{n} d^2(e_i, G)$$

$e_i$ : individual i

G : mean point

$$= \begin{pmatrix} x_{.1} \\ \vdots \\ x_{.p} \end{pmatrix}$$

$$= V_T$$

Suppose that we have K groups

$G_i$ : the mean points of group i.

variance between groups : $V_B$

$$V_B = \frac{1}{n} \sum_{i=1}^{K} d^2(G, G_i)\, n_i$$

$n_i$ : # of individuals inside group i

Variance inside groups: $V_W$

$$V_W = \sum_{i=1}^{K} \frac{n_i}{n} \frac{1}{n_i} \sum_{j=1}^{n_i} d^2(e_{i,j}, g_i)$$

where $e_{i,1}, \ldots, e_{i,n_i}$ denote individuals inside group $i$

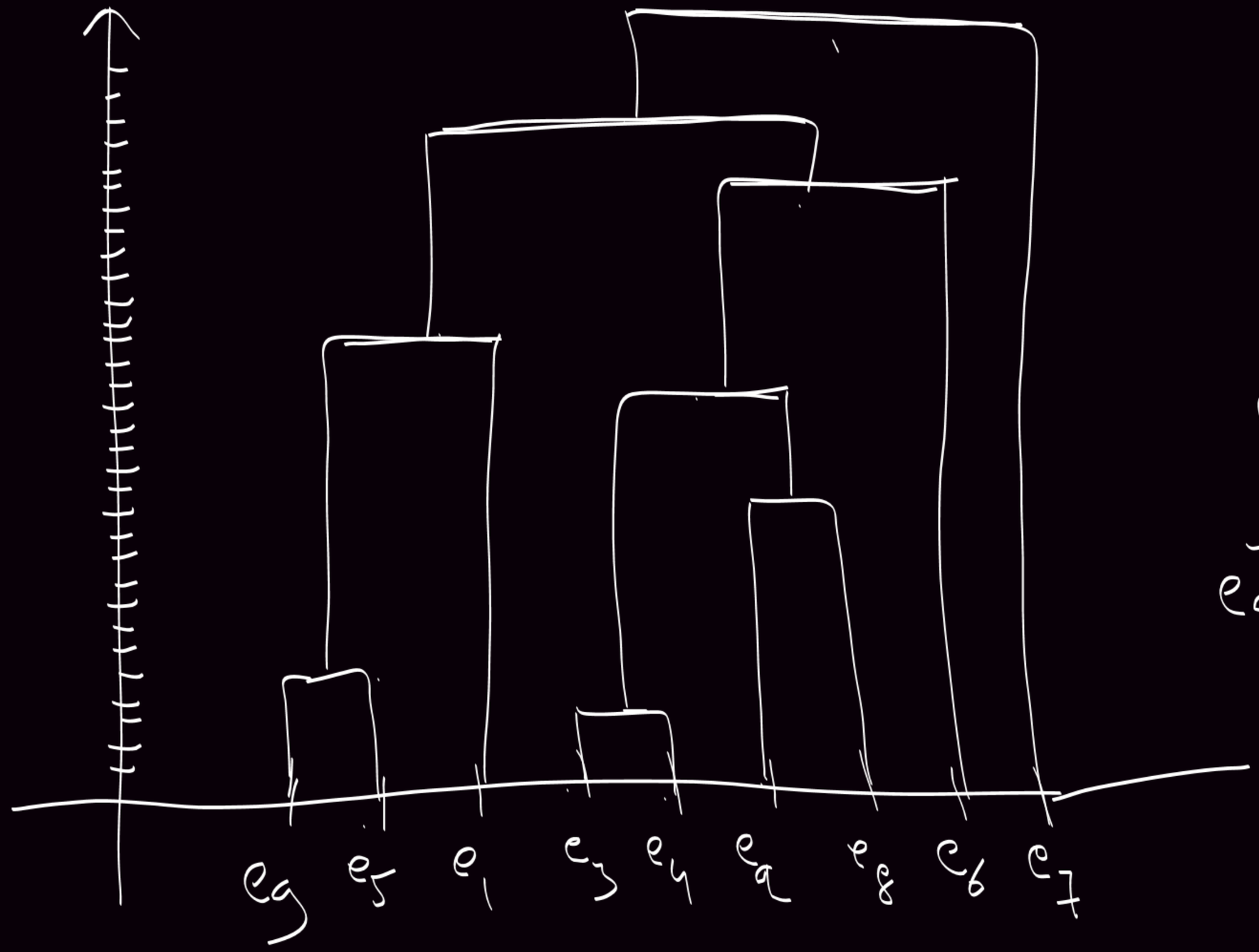
$$V_T = V_B + V_W$$

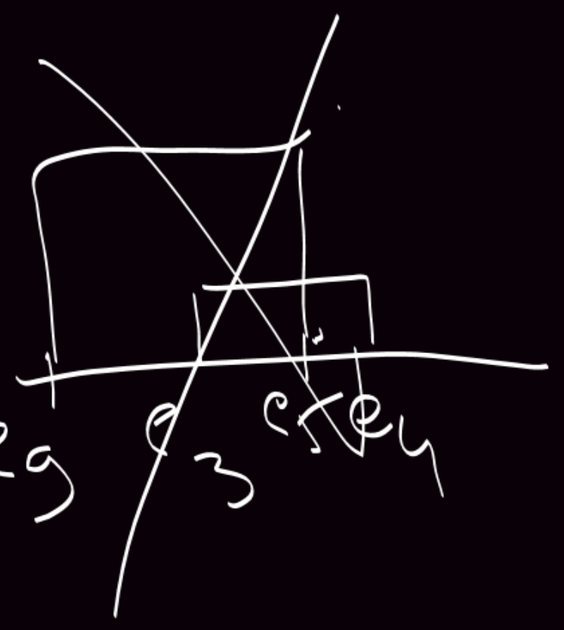
$$V_B \leq V_T$$

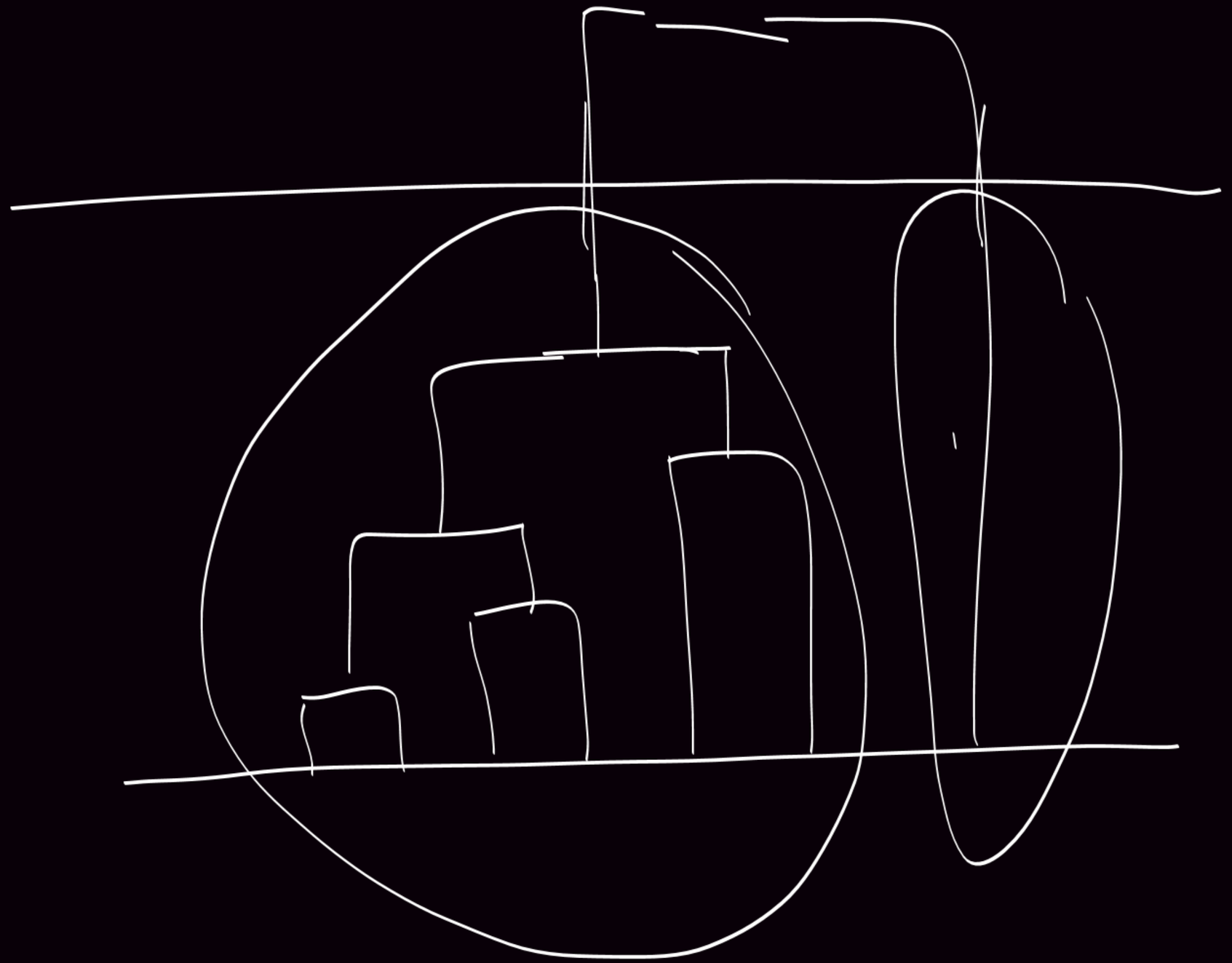
$$V_W \geq 0$$

We stop the ascending method when all the individuals are together!

# drawing of the ascending method: dendrogram


e9 e5 e1 e3 e4 e2 e8 e6 e7


e9 e3 e1 e4



WE ARE BACK!