

# Principle Component Analysis

Sébastien Lê

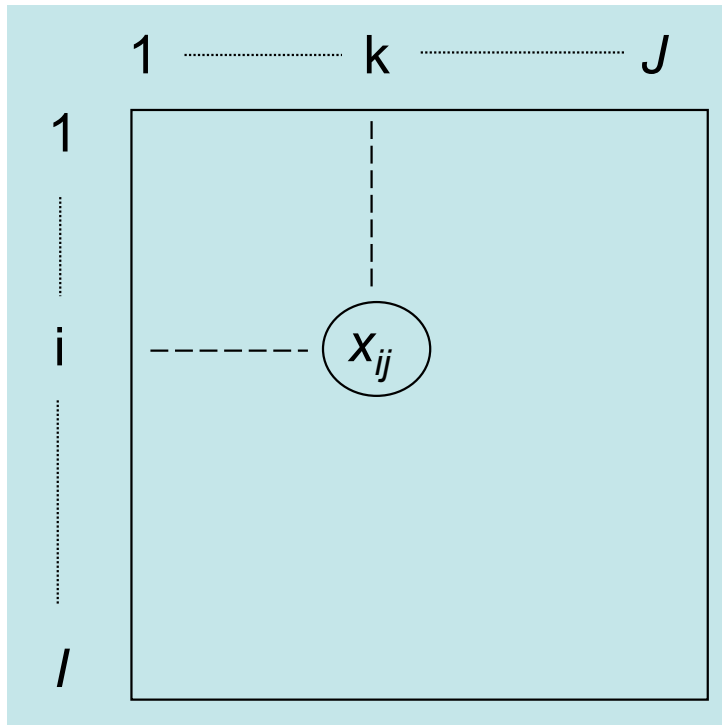
# Data, notations, examples

- Individuals – objects described by the set of data (**rows** in the dataset or data matrix)
- Variables – any characteristics of an individual (**columns** in the dataset or data matrix)

# Data, notations, examples

$I$  individuals

$J$  quantitative variables



Average:

$$\bar{x}_j = \frac{1}{I} \sum_{i=1}^I x_{ij}$$

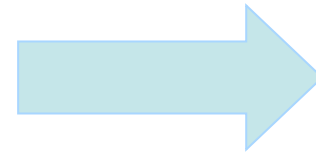
Standard deviation:

$$s_j = \sqrt{\frac{1}{I} \sum_{i=1}^I (x_{ij} - \bar{x}_j)^2}$$

# Problematics: individuals

Resemblance between 2 individuals?

Summary of the resemblances



Typology

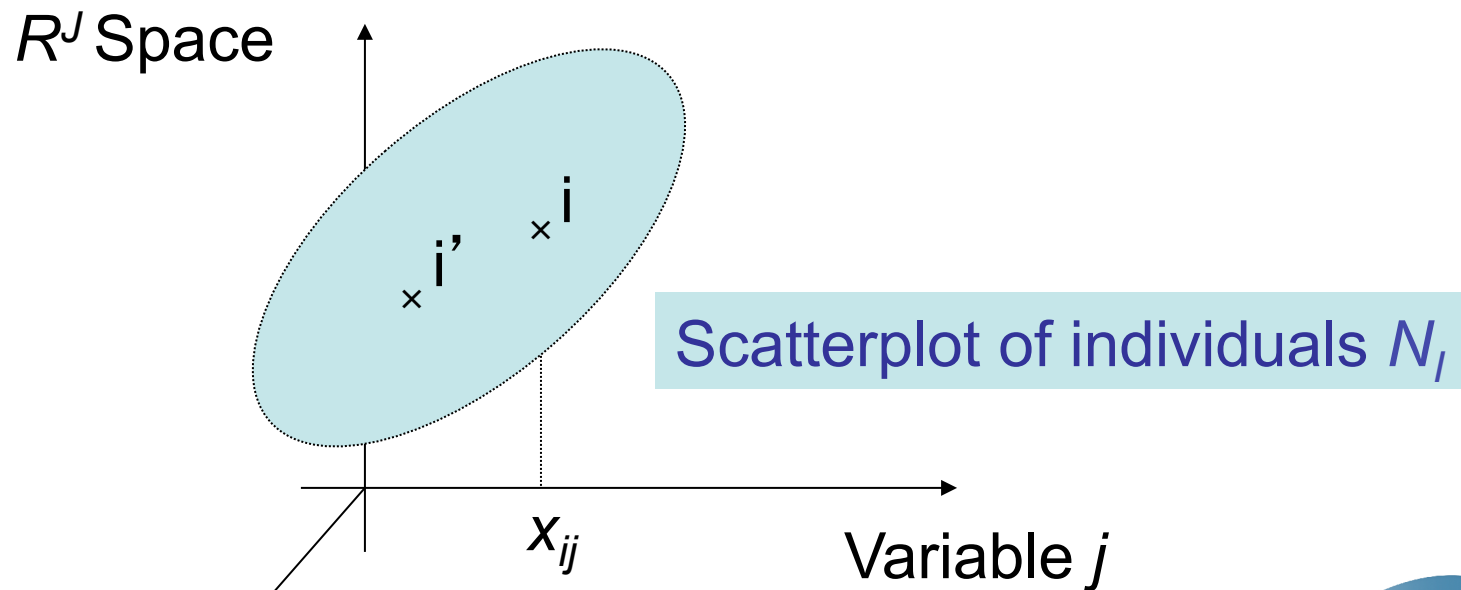
# Problematics: variables

Resemblance between 2 variables?

Summary of the resemblances  Typology

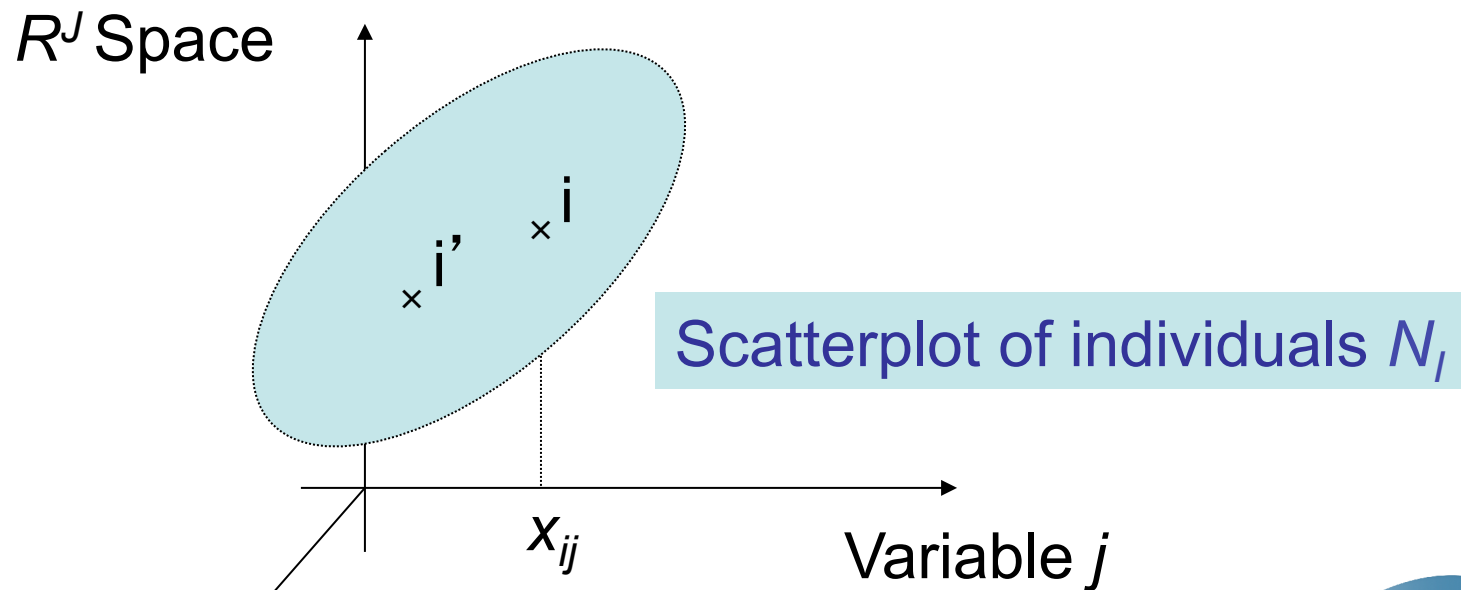
# Scatterplot of individuals $N_i$

Scatterplot of individuals: graphical and geometrical representation of the  $I$  individuals described by the  $J$  variables

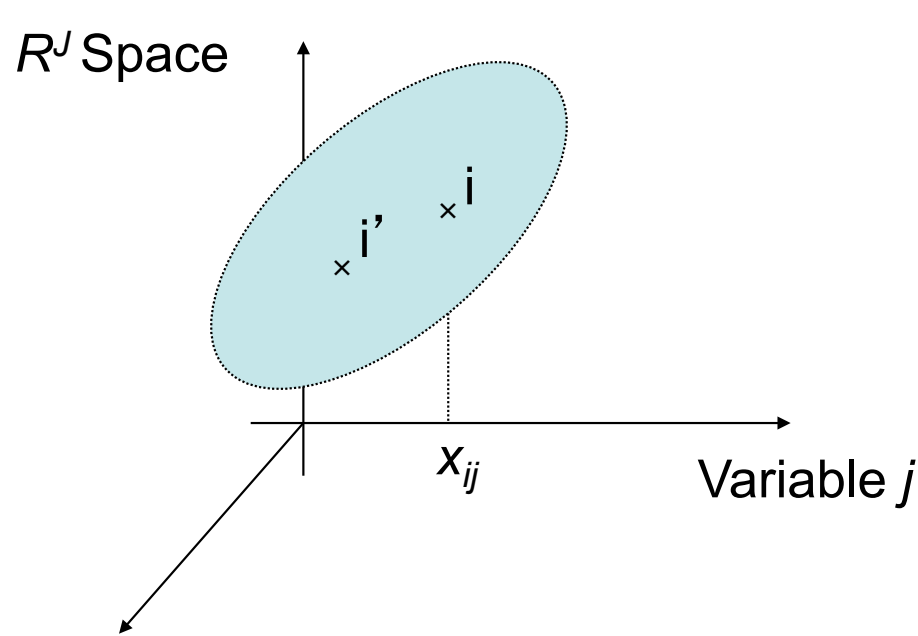


# Scatterplot of individuals $N_i$

Studying individuals  $\longleftrightarrow$  Studying shape of scatterplot



# Scatterplot of individuals $N_i$



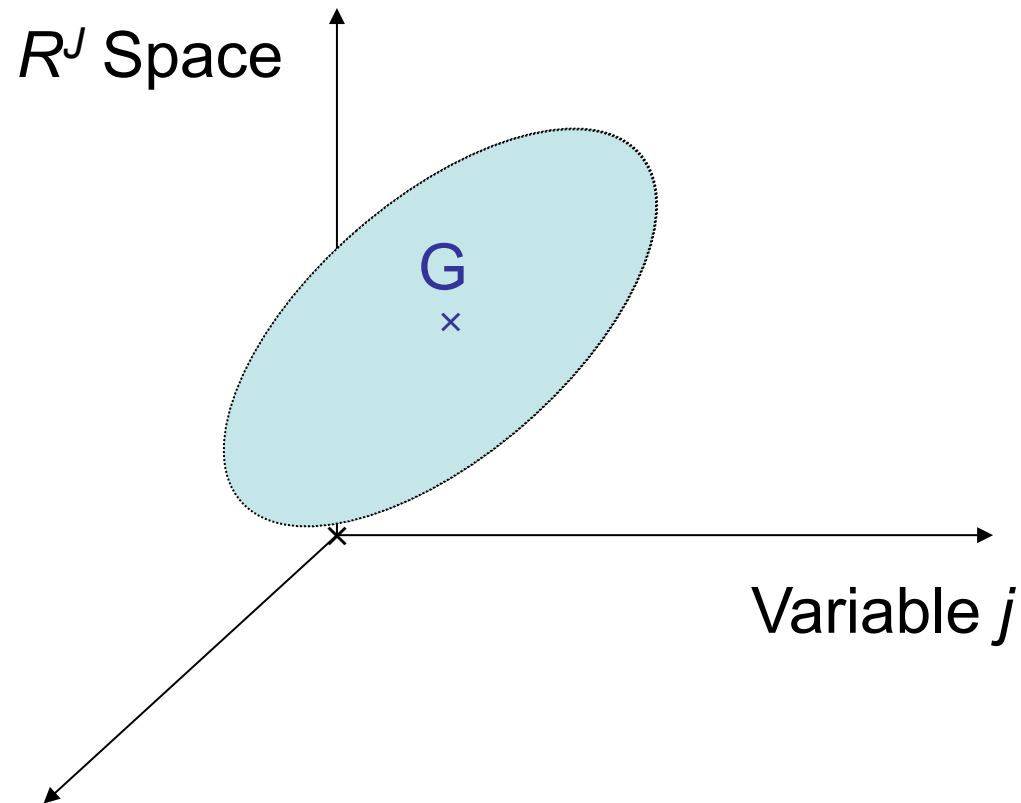
$$G = \text{Barycentre} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_j \\ \bar{x}_J \end{pmatrix}$$

$$\text{Inertia} = \sum_{i=1}^I d^2(G - i)^2$$

Resemblance between 2 individuals:

$$d^2(i, i') = \sum_{j=1}^J (x_{ij} - x_{i'j})^2$$

# Scatterplot of individuals $N_i$

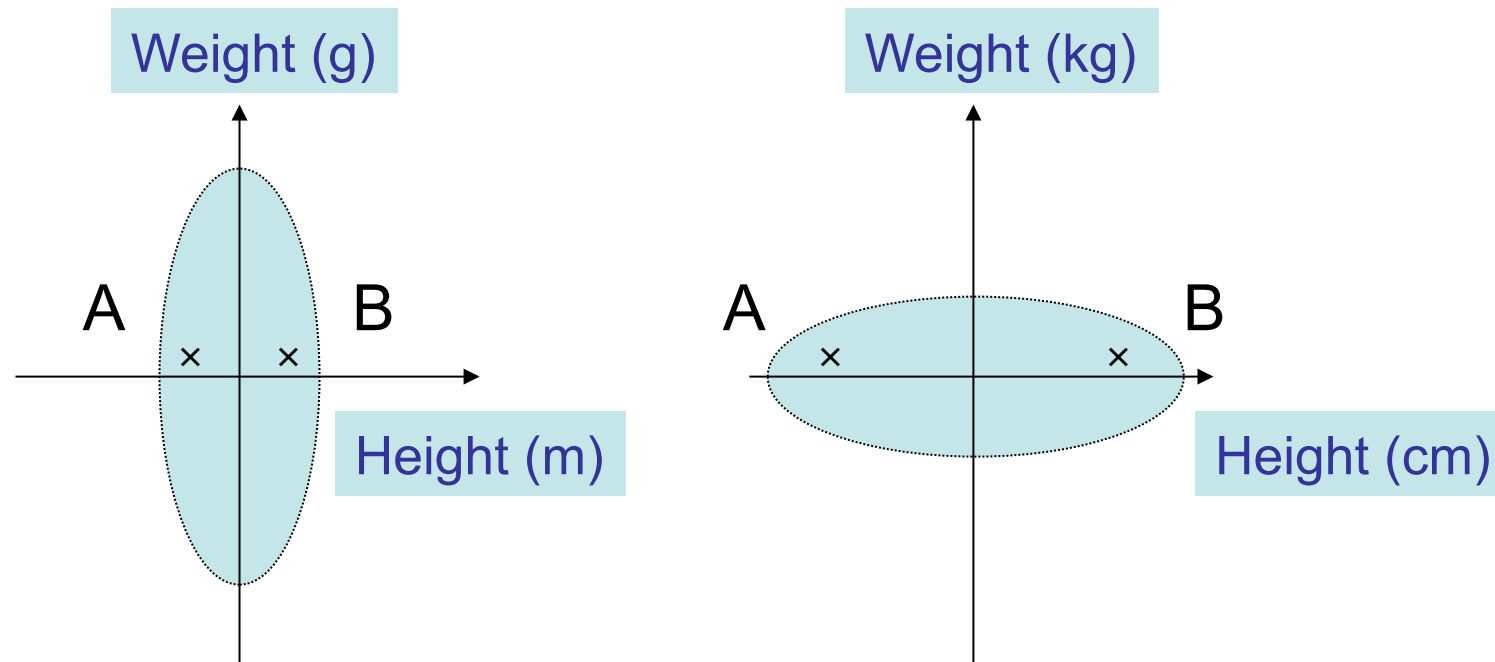


$$G = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Mean-centering the data:  $x_{ij} \Rightarrow x_{ij} - \bar{x}_j$

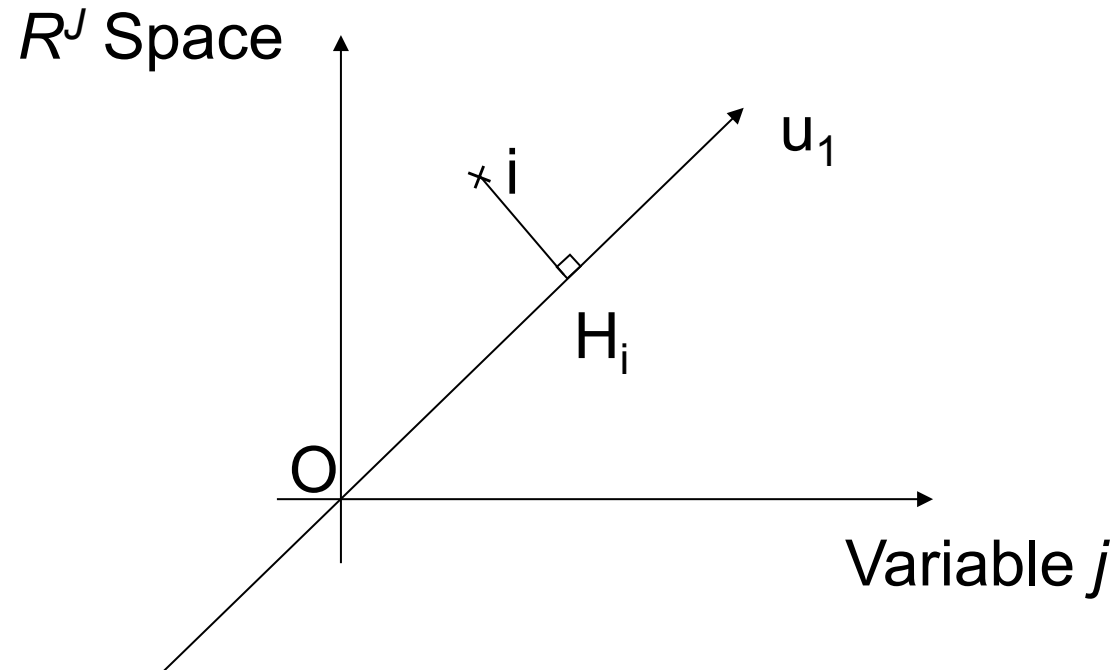
# Scatterplot of individuals $N_i$

Problems related with unit of measurement



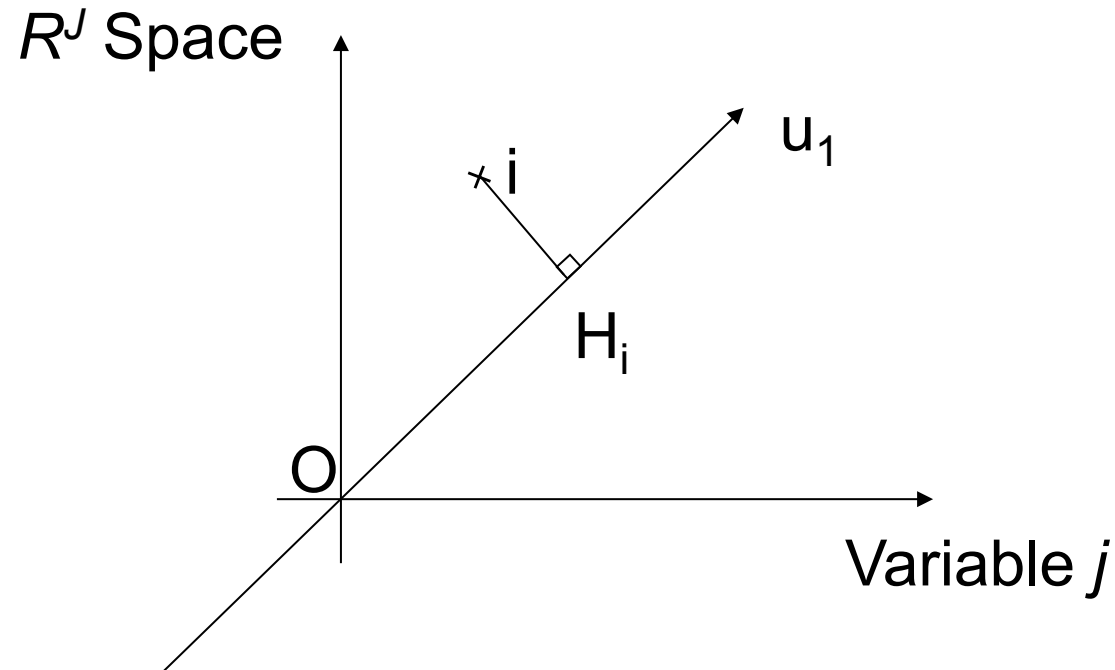
Standardizing the data:  $x_{ij} \Rightarrow \frac{x_{ij} - \bar{x}_j}{s_j}$

# Fitting the scatterplot $N_j$



Minimizing  $(iH_i)^2 \iff$  Maximizing  $(OH_i)^2$

# Fitting the scatterplot $N_j$



Minimizing  $(iH_i)^2 \iff$  Maximizing  $(OH_i)^2$

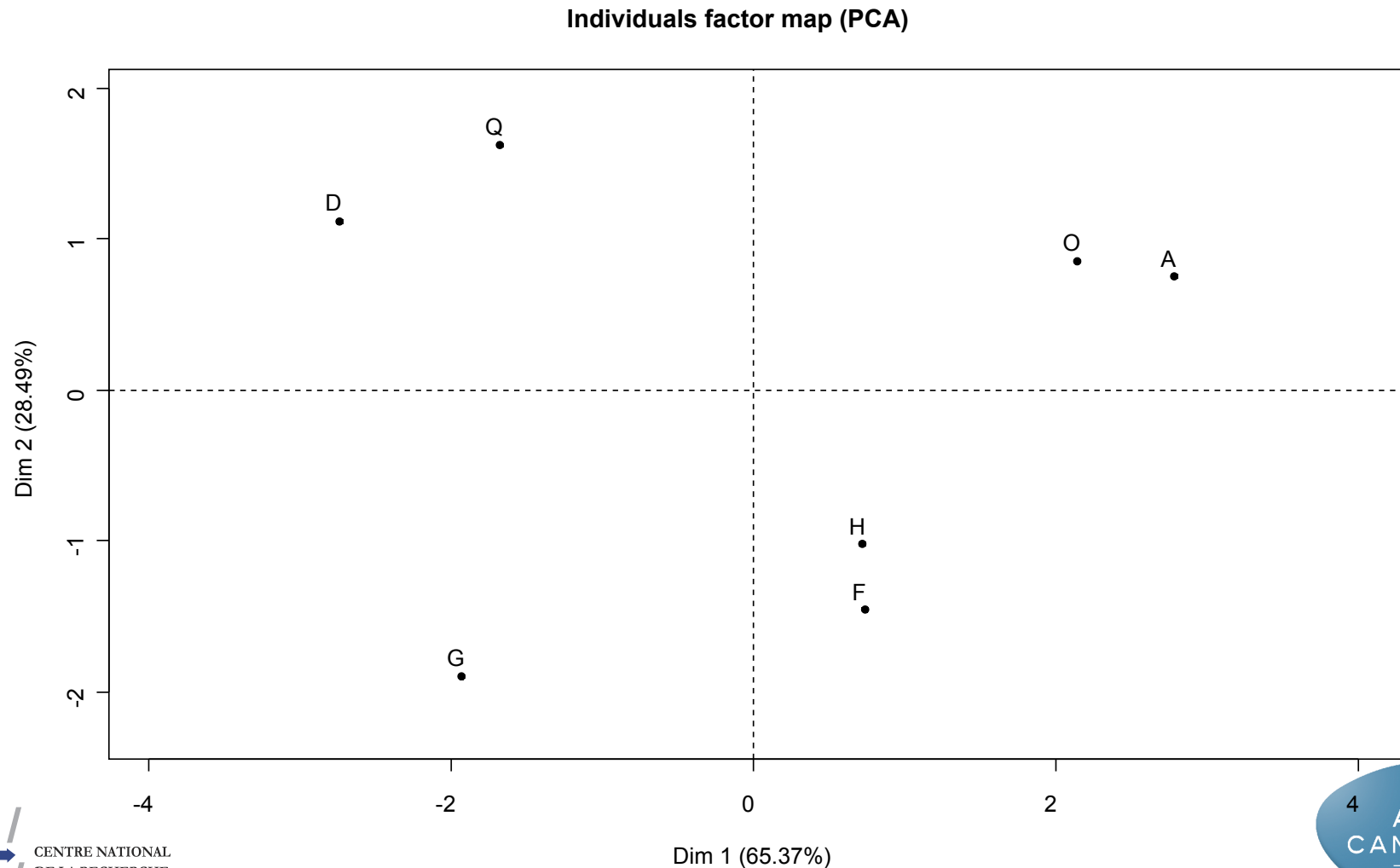
Maximizing projected inertia:  $\sum_i (OH_i)^2$

# Fitting the scatterplot $N_i$

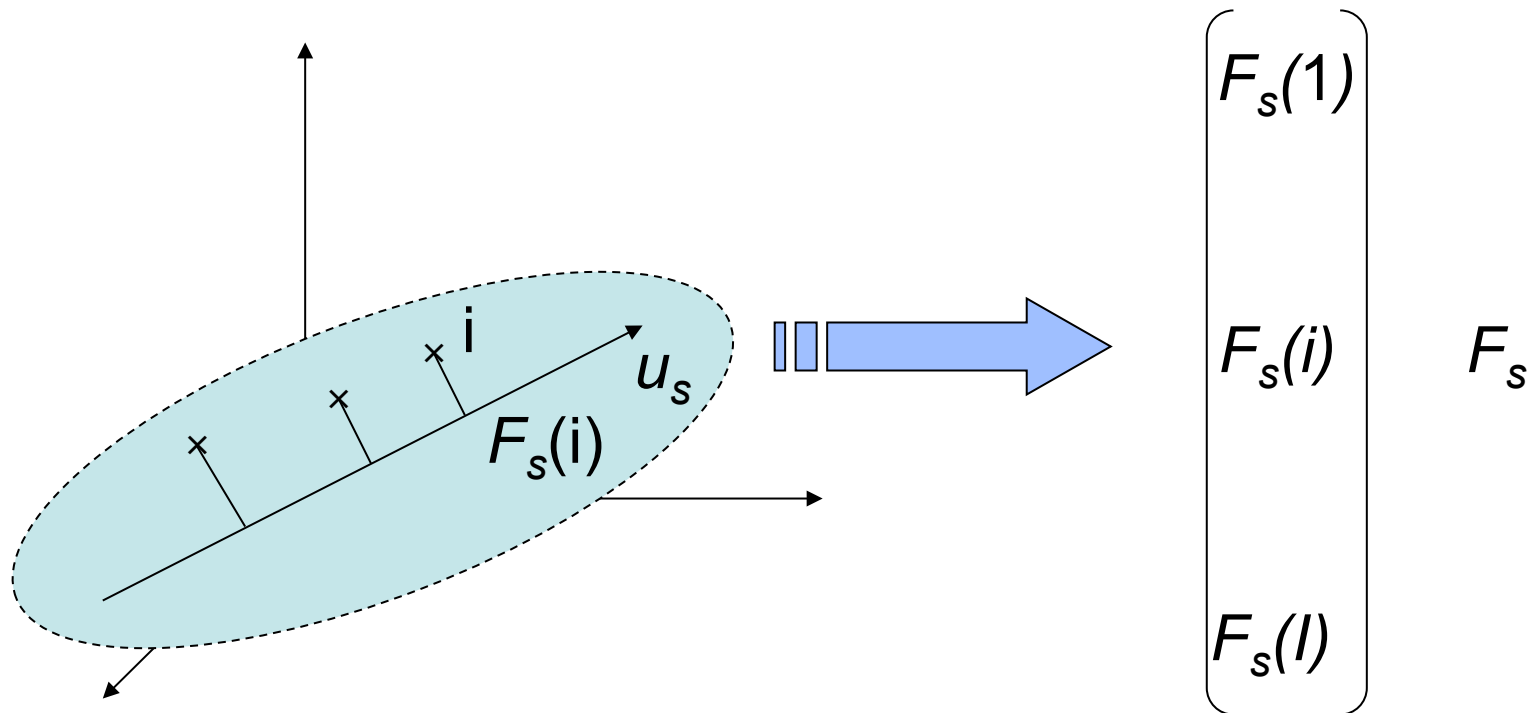
$N_i$  is projected on a sequence  $(u_s)$   
of orthogonal axes of maximum inertia

Axes: main factors of variability

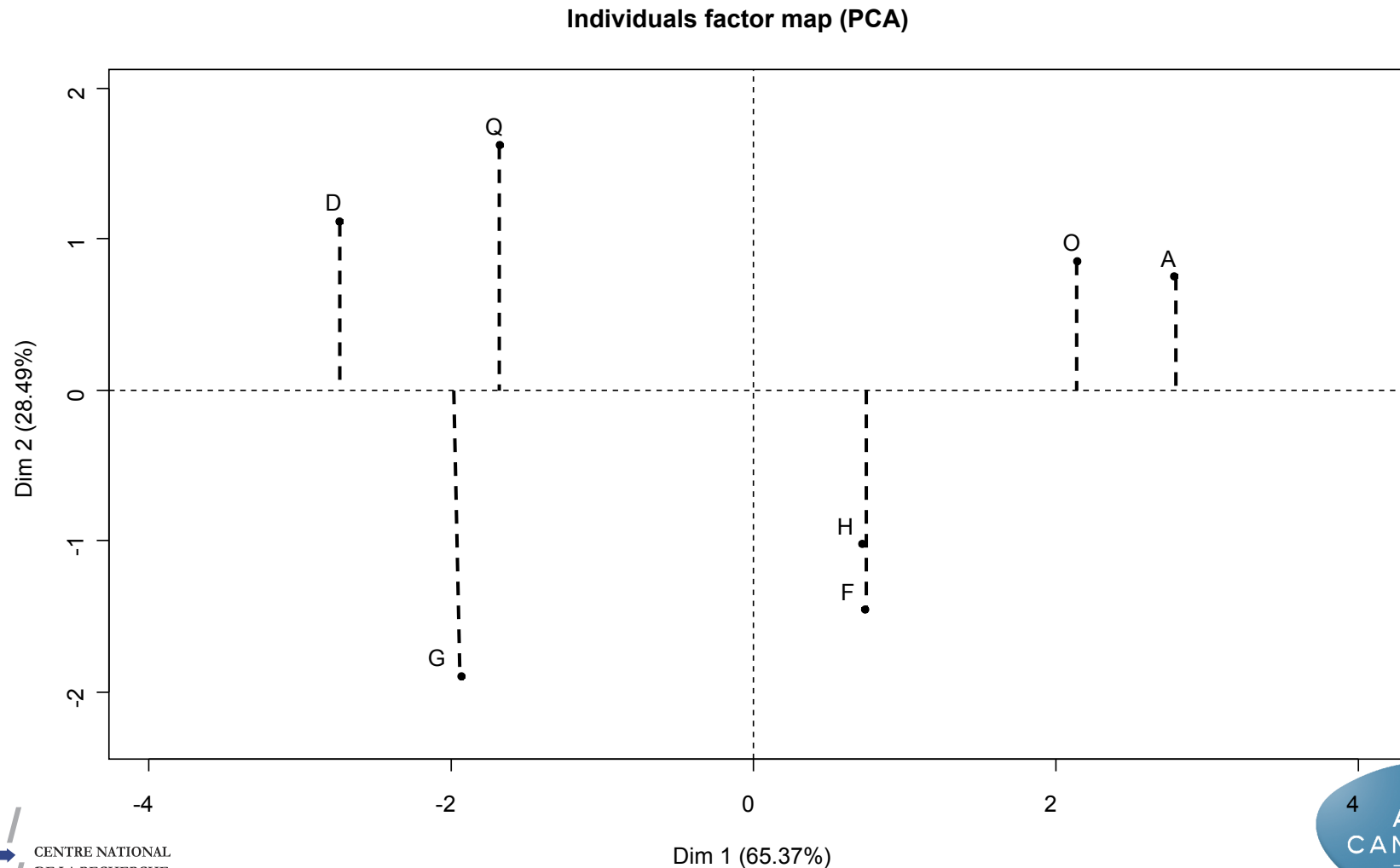
# Fitting the scatterplot $N_i$



# Representation of the variables



# Fitting the scatterplot $N_i$



# Fitting the scatterplot $N_1$

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor	Dim,1	Dim,2
A	1,31	0,28	-0,30	-1,11	1,66	1,60	2,78	0,76
D	-1,11	-1,47	1,32	1,81	-0,45	-0,94	-2,74	1,12
F	0,22	1,13	-1,07	-0,56	-0,34	-0,14	0,73	-1,45
G	-1,46	0,81	-0,64	1,15	-1,22	-1,19	-1,94	-1,90
H	0,29	0,77	-0,87	-0,14	-0,29	0,40	0,71	-1,02
O	1,24	-0,03	-0,12	-0,86	1,35	1,11	2,14	0,86
Q	-0,49	-1,49	1,68	-0,30	-0,71	-0,85	-1,68	1,63

# Fitting the scatterplot $N_1$

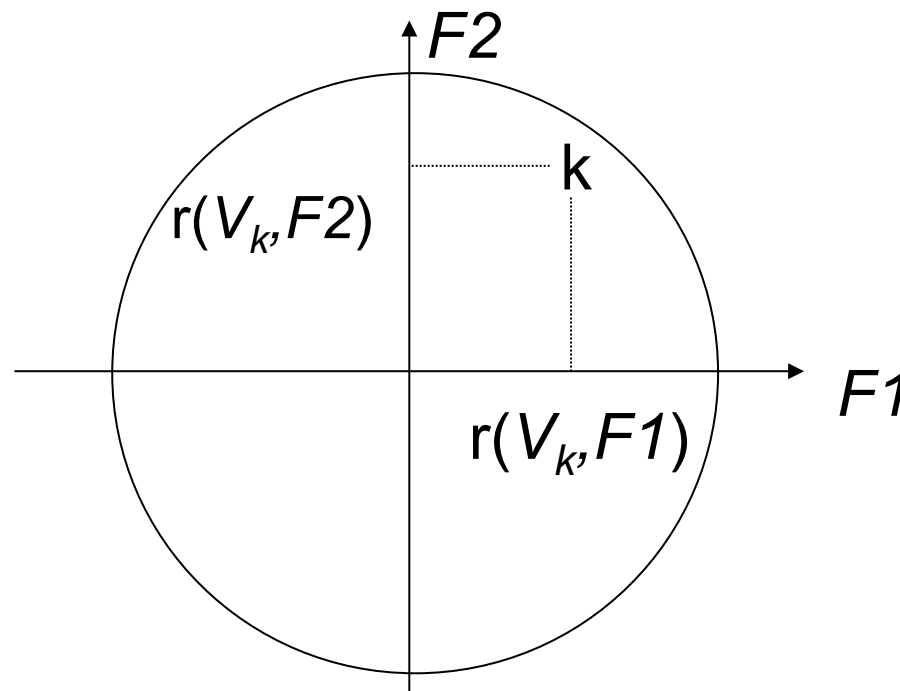
Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor	Dim,1	Dim,2
A	1,31	0,28	-0,30	-1,11	1,66	1,60	2,78	0,76
D	-1,11	-1,47	1,32	1,81	-0,45	-0,94	-2,74	1,12
F	0,22	1,13	-1,07	-0,56	-0,34	-0,14	0,73	-1,45
G	-1,46	0,81	-0,64	1,15	-1,22	-1,19	-1,94	-1,90
H	0,29	0,77	-0,87	-0,14	-0,29	0,40	0,71	-1,02
O	1,24	-0,03	-0,12	-0,86	1,35	1,11	2,14	0,86
Q	-0,49	-1,49	1,68	-0,30	-0,71	-0,85	-1,68	1,63

	Dim,1	Dim,2
Ext_Color	0,96	0,24
Firm	0,51	-0,86
Melty	-0,56	0,82
Mealy	-0,88	-0,14
Sweet	0,85	0,42
Tomato_Flavc	0,96	0,19

# Representation of the variables

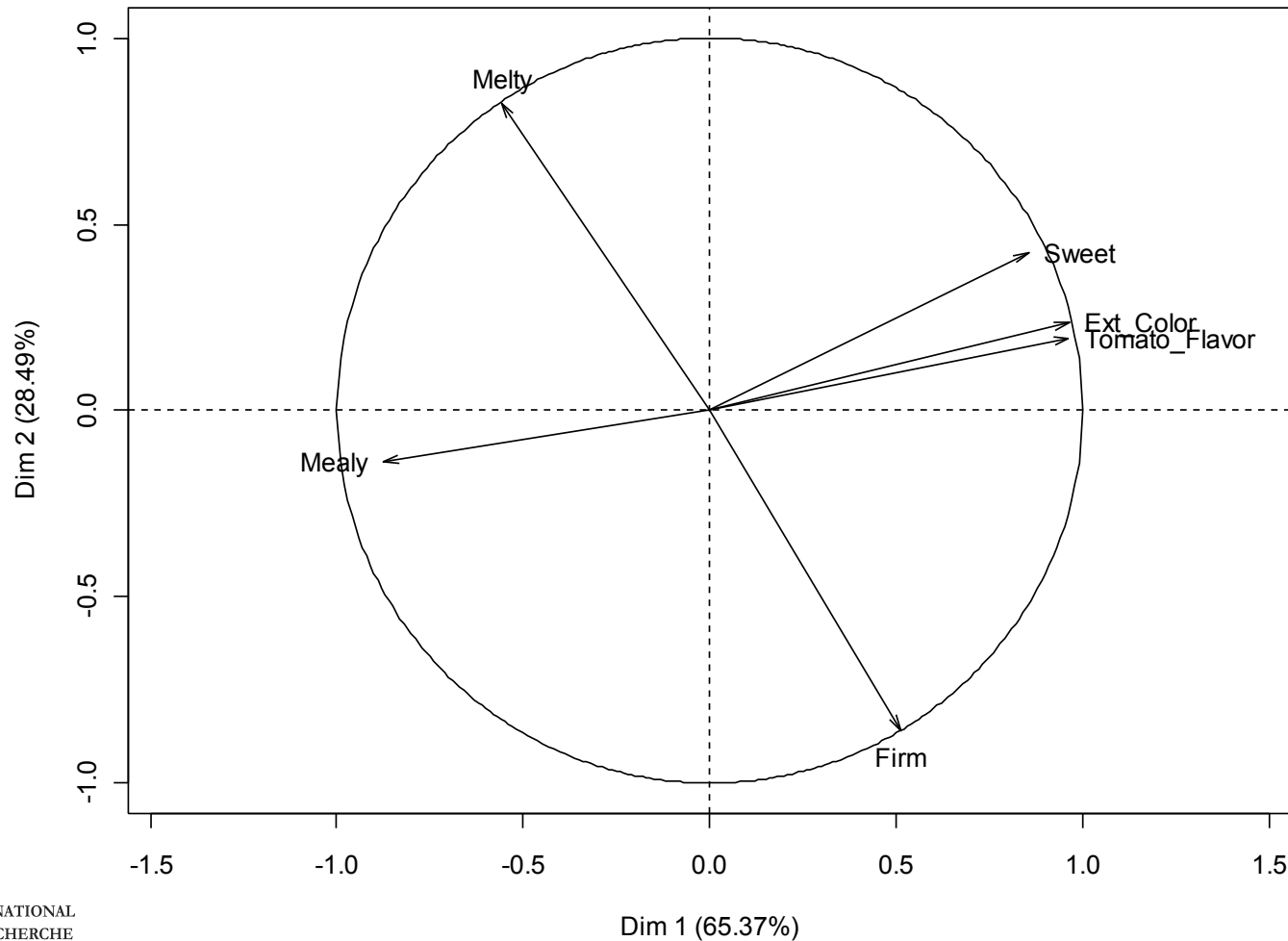
**F1**: vector of individuals coordinates on the first axis

**F2**: vector of individuals coordinates on the 2<sup>nd</sup> axis



# Representation of the variables

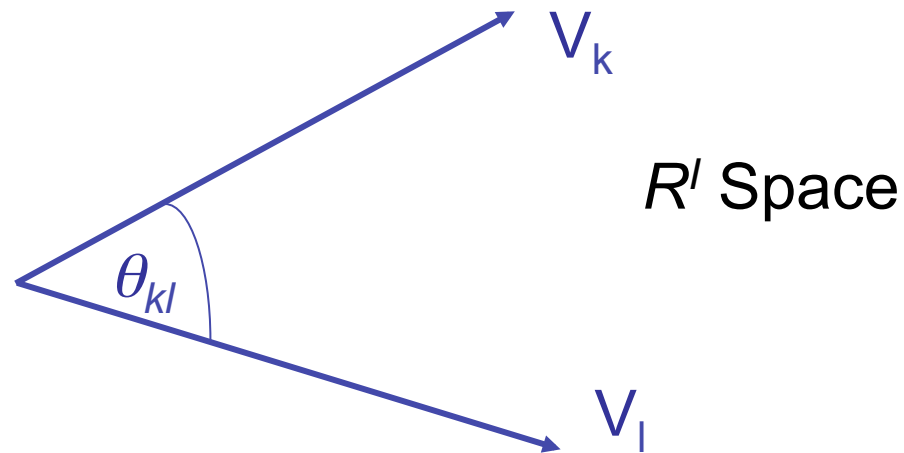
Variables factor map (PCA)



# Scatterplot of variables $N_j$

1 variable = 1 column

1 variable = 1 vector in the space  $R^l$



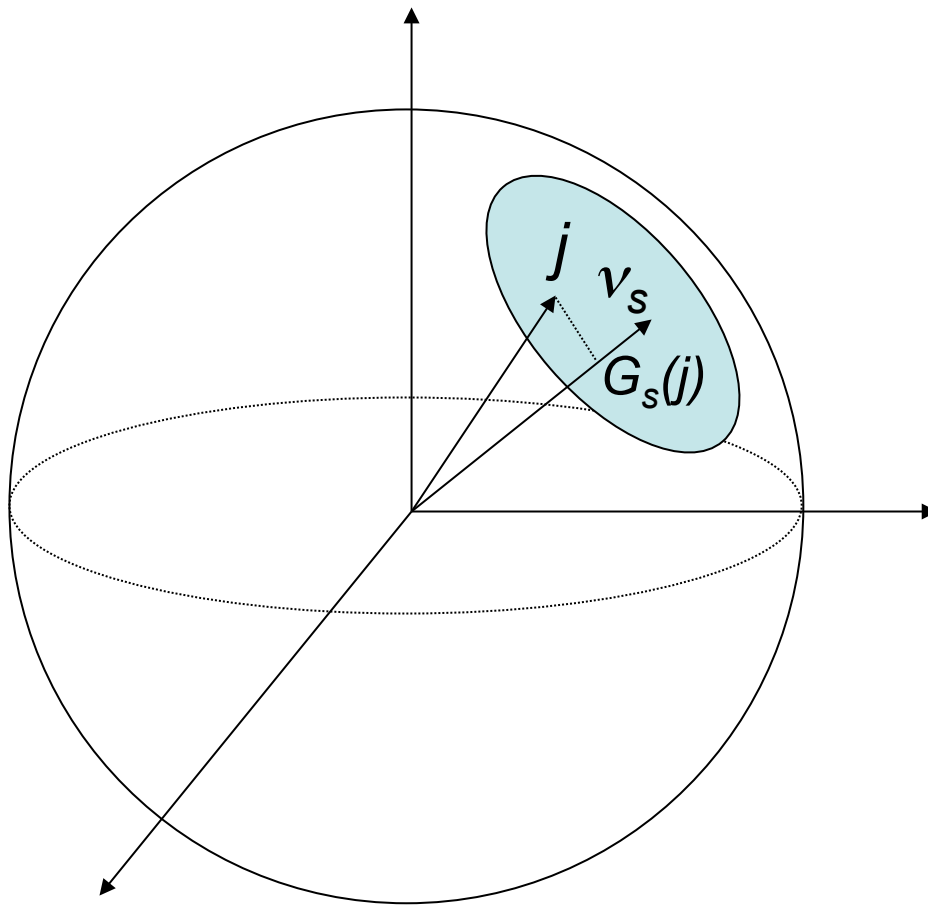
$$\text{Property : } \cos (\theta_{kl}) = r (V_k, V_l)$$

# Fitting the scatterplot $N_J$

$N_J$  is projected on a sequence  $(v_s)$   
of orthogonal axes of maximum inertia

Axes: main factors of variability

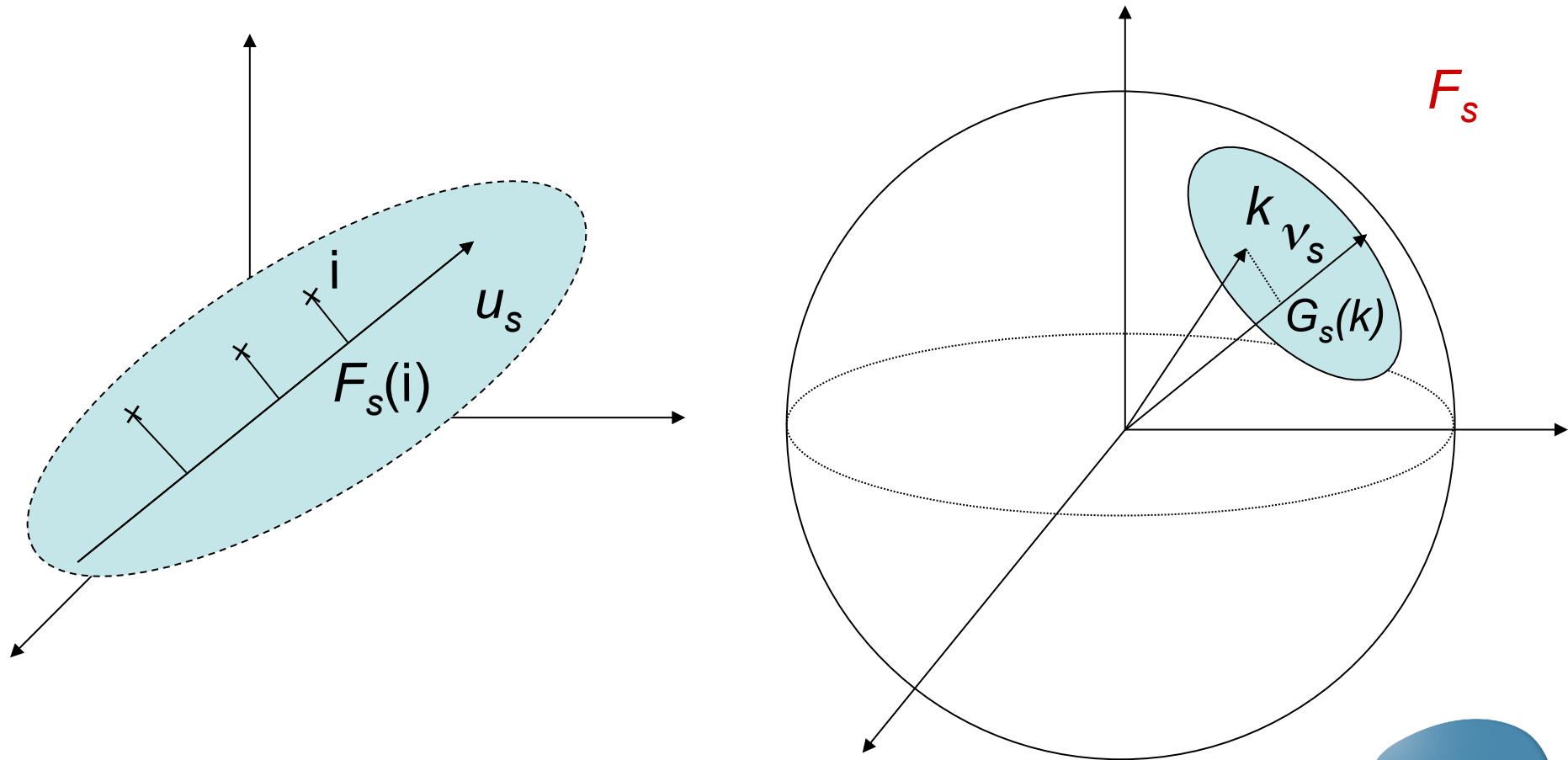
# Fitting the scatterplot $N_j$



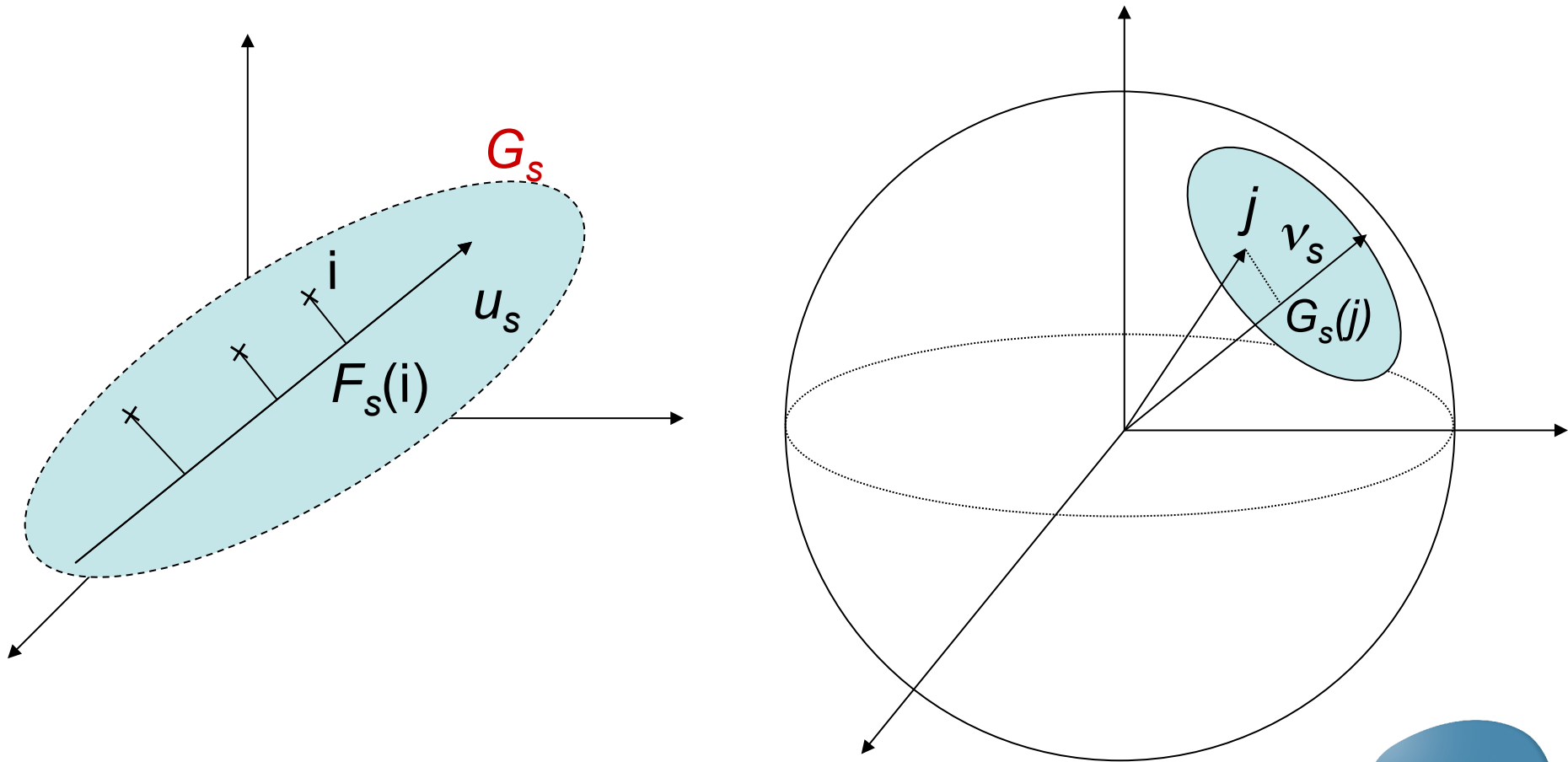
*Sequence of  
latent variables  $v_s$*

$$\begin{pmatrix} G_s(1) \\ \\ G_s(j) \\ \\ G_s(J) \end{pmatrix} G_s$$

# Transition equation



# Transition equation



# Transition equation

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{x_{ij} - \bar{x}_j}{s_j} G_s(j)$$

$$G_s(j) = \frac{1}{I} \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{x_{ij} - \bar{x}_j}{s_j} F_s(i)$$

An individual is on the side of the variables for which it takes high values and on the opposite side of the variables for which it has low values.

# PCA{FactoMineR}

- The main output(s):
  - A representation of the individuals
  - A representation of the variables
  - The “usual” numerical indicators
- The main argument(s):
  - Supplementary categorical variables
  - Supplementary continuous variables
  - Supplementary individuals

# PCA{FactoMineR}

- `library(FactoMineR)`
- `PCA(decathlon, quanti.sup=c(11,12), quali.sup=13)`
- `res <- PCA(decathlon, quanti.sup=c(11,12), quali.sup=13)`
- `names(res)`
- `res$eig`

# plot.PCA{FactoMineR}

- The main output(s):
  - A representation of the individuals
  - A representation of the variables
- The main argument(s):
  - An object issued from PCA{FactoMineR}
  - Supplementary categorical variables
  - Supplementary continuous variables
  - Supplementary individuals

# dimdesc{FactoMineR}

- The main output(s):
  - An automatic description of each axis
- The main argument(s):
  - An object issued from PCA{FactoMineR}

# The End

This presentation is licensed under a  
[Creative Commons Attribution 4.0  
International License.](https://creativecommons.org/licenses/by/4.0/)



By Sébastien Lê