Laboratori de Modelització i Anàlisi de la Informació

Session
Principal component analysis

Anàlisi de Dades i Explotació de la Informació

Grau d'Enginyeria Informatica.

Information System tracking

Prof. Mónica Bécue Bertaut & Lidia Montero

Monica.becue@upc.edu *lidia.montero@upc.edu*

Universitat Politècnica de Catalunya BarcelonaTech

Principal axes methods

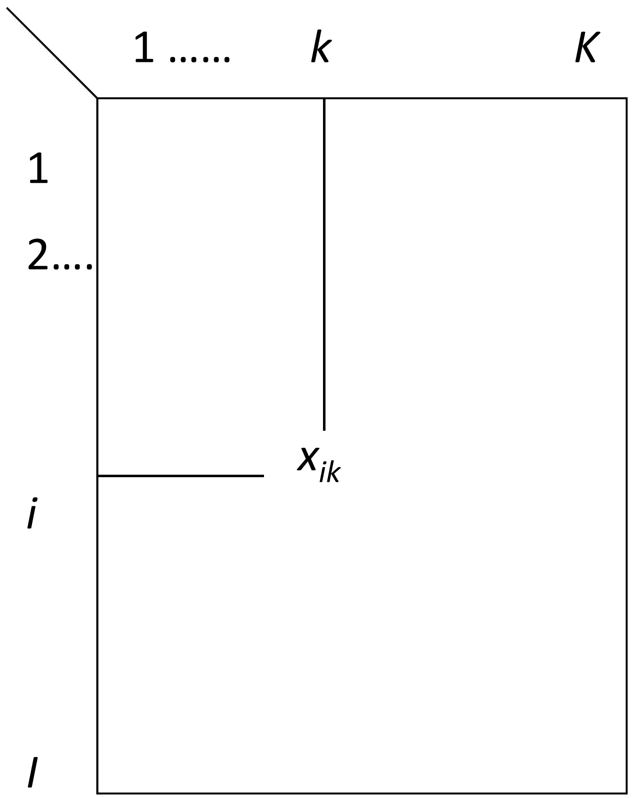
- *Principal components analysis*
- *Correspondence analysis*
- *Multiple Correspondence Analysis*

Principal component analysis

Dimensionality reduction: PCA allows us to describe a dataset **with a smaller number of variables**

Used for data compression, data reconstruction, preprocessing, describing a dataset, reduction of the dimension, etc.

Principal component analysis



Data

Individuals×Quantitative variables

Principal component analysis

Objetives

We want to put to the fore

• the structure of the row-individuals through an Euclidean representation to **detect the individuals that are similar from the point of view of the active variables**

• the structure of the column-variables a representation that **evidences the variables highly correlationated**

This method aims at discovering the data structure, the underlying system, the patterns, the general rules but also the clues towards hidden information

And explain

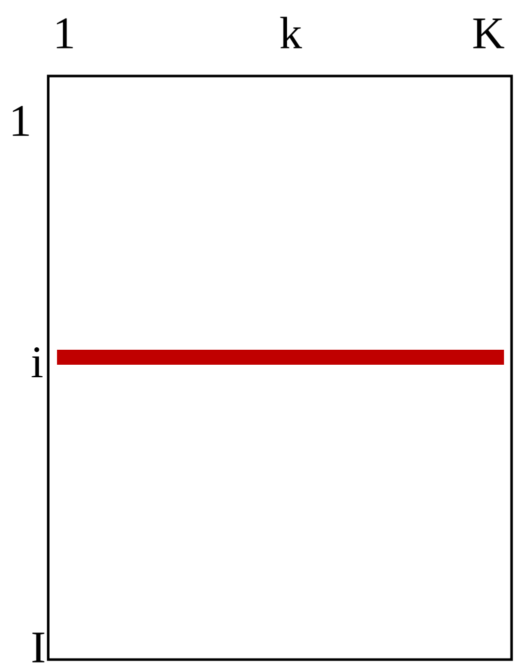
• the variability of the individuals from the variables point of view

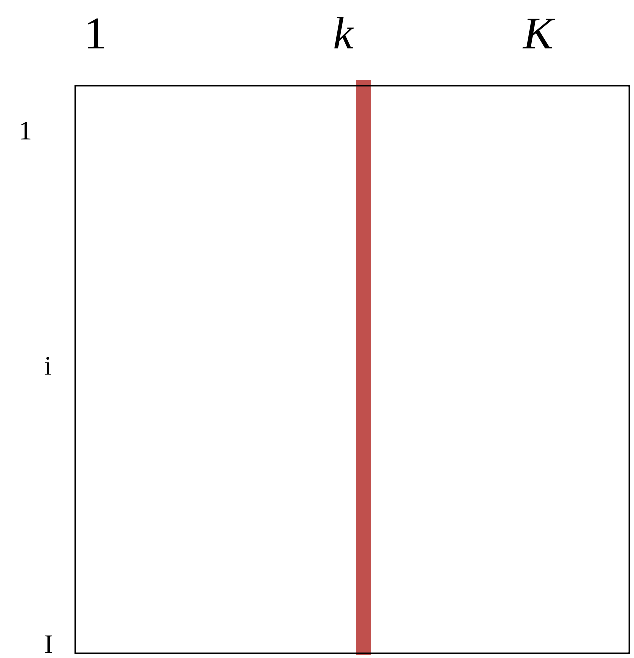
• the dispersion of the variables from the individuals point of view

Principal component analysis

Variables

Duality of the table





Two points of view: individuals or variables

Principal component analysis

Weights of the individuals

The individuals can be endowed with weights that intervenen in computing mean, standard deviation, correlation, etc. In the following, $p_i=1/I$

Weights of the variables

not very used, but...if variable *k* is provided with weight *k*, then

$$d^2(i,l)=\sum_{k\in K} m_k\left(x_{ik}-x_{lk}\right)^2$$

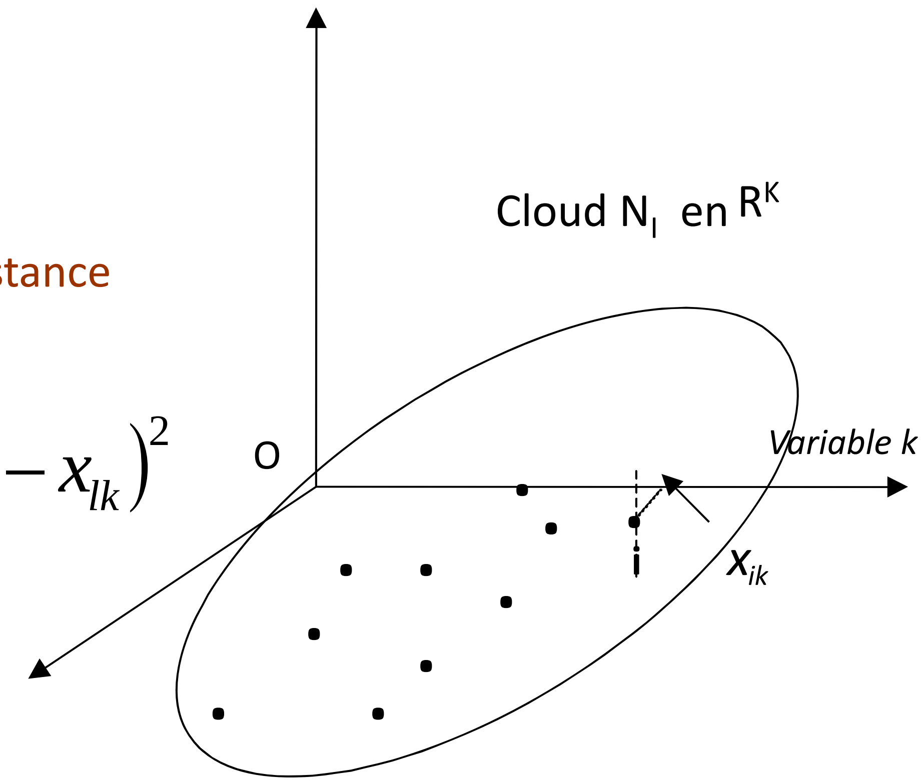
We will see only the case $m_k=1$

Principal component analysis

Cloud of individuals

Classical Euclidean distance

$$d^2(i,l) = \sum_{k \in K} \left(x_{ik} - x_{lk} \right)^2$$

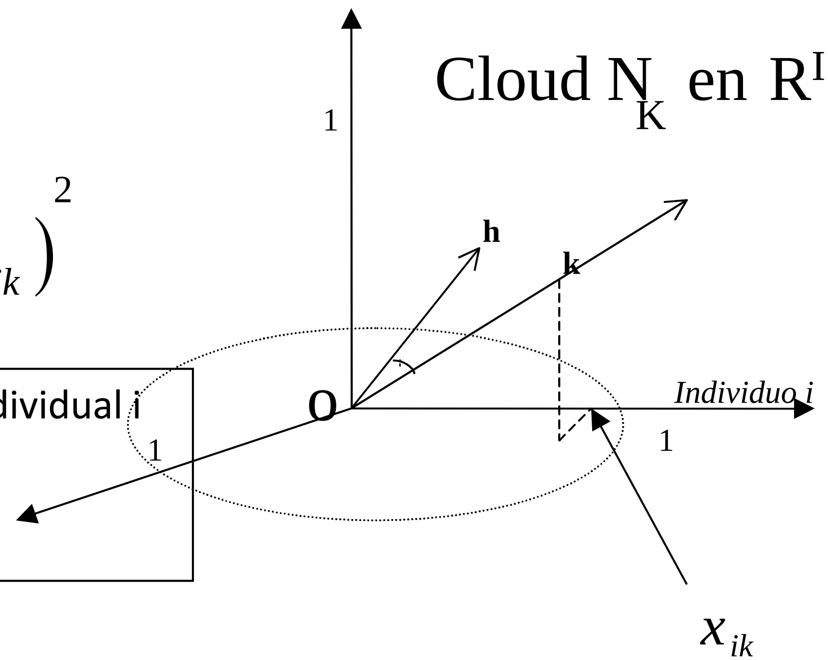


Principal component analysis

Cloud of variables

$$d^2(h,k) = \sum_{i \in I} \frac{1}{I} \left(x_{ih} - x_{ik} \right)^2$$

Weight of individual i



However, it is usual to analyze the relationships between variables through the linear correlation coefficient

$$r(k,h) = \frac{1}{I} \sum_{i \in I} \left(\frac{x_{ik} - \bar{x}_k}{s_k} \right) \cdot \left(\frac{x_{ih} - \bar{x}_h}{s_h} \right) = \frac{s_{k,h}}{s_k s_h}$$

Principal component analysis

Objective: to obtain the “best representations” in a low dimension space of

- the cloud of individuals
- the cloud of variables
- in such a way that both representations are linked and jointly interpreted

In PCA, either matrix *Y* or matriz Z are used with general terms:

$$y_{ik} = (x_{ik} - \bar{x}_k) \qquad z_{ik} = \frac{(x_{ik} - \bar{x}_k)}{s_k}$$

The best representation

in a two-dimensions space

Find the subspace which better sums up the data

Why?

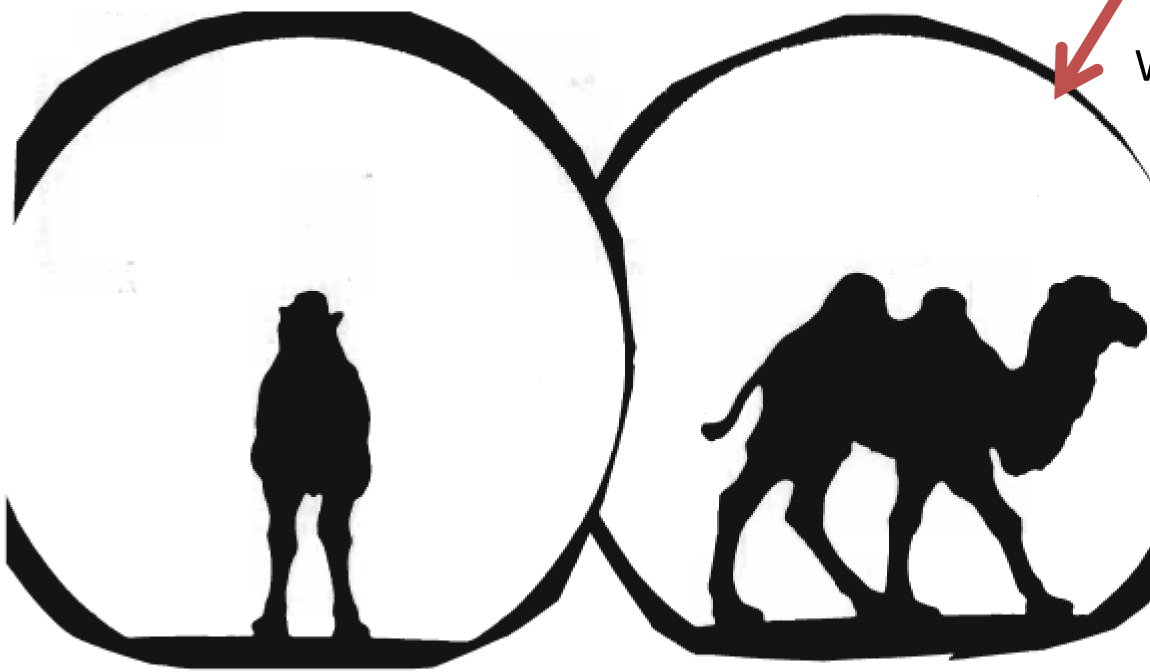


Figure: Camel vs dromedary?

11

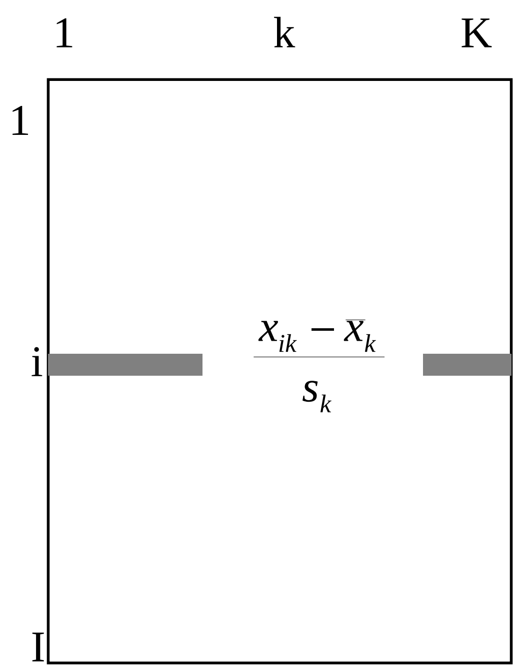
Principal component analysis

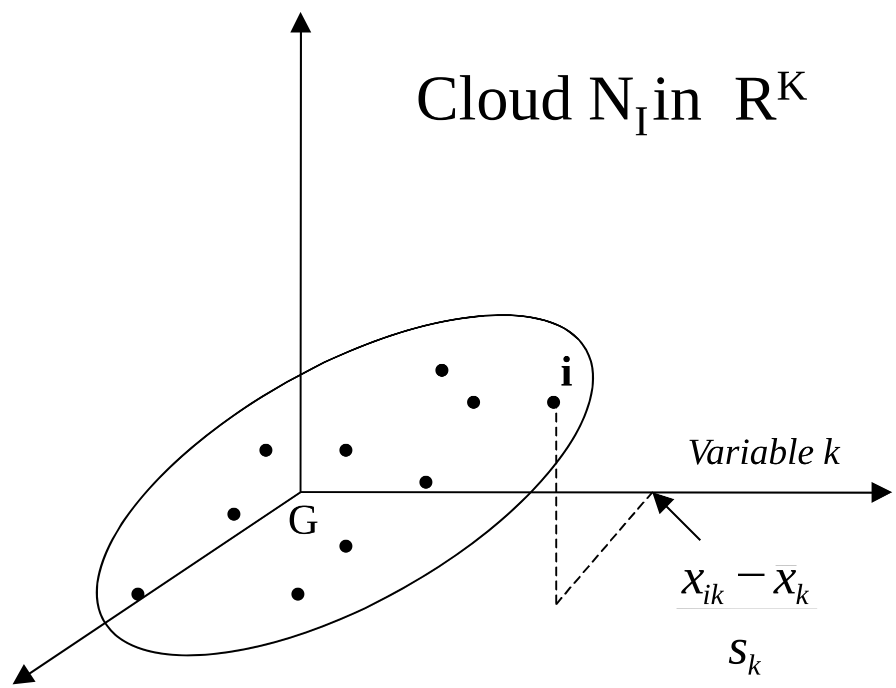
In the individuals space

Principal component analysis

In the individual space

$$d^2(i,l)=\sum_k\left(\frac{x_{ik}-x_{lk}}{s_k}\right)^2=\sum_k\left(z_{ik}-z_{lk}\right)^2$$





Principal component analysis

Find the subspace which better sums up the data

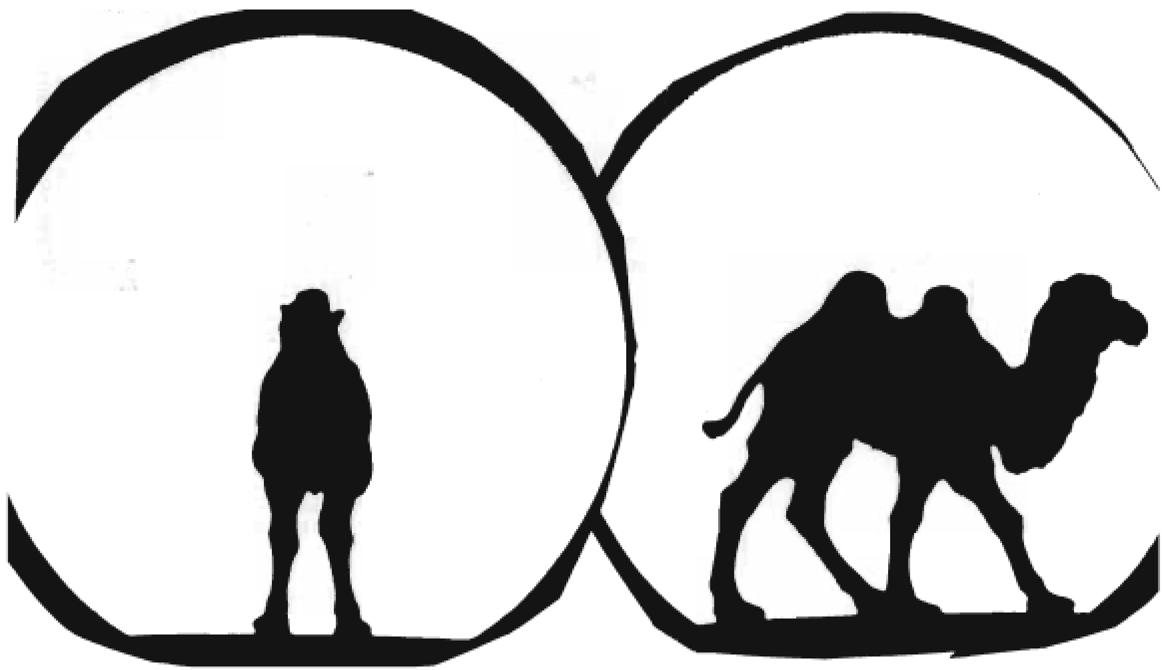
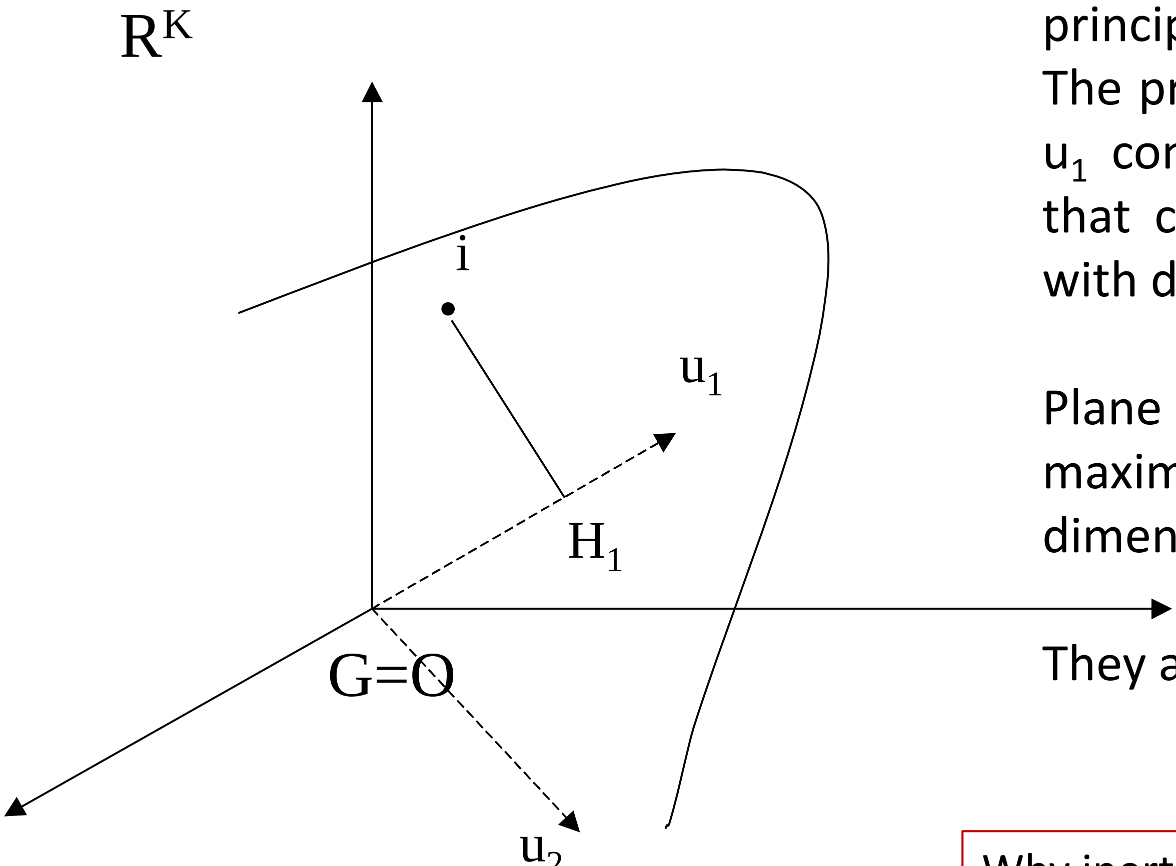


Figure: Camel vs dromedary?

Principal component analysis



$$\boxed{Max \sum_i OH_i^2}$$

We look for the maximum dispersion axes in R^K, called principal axes $\{u_s;\ s=1,\ldots,S\}$.
The projection of the individuals on u_1 conserves the maximum inertia that can be conserved on a space with dimesion 1;

Plane (u_1,u_2) conserves the maximum inertia on a space with dimension 2 , etc.

They are orthogonal directions.

Why inertia and not the sum of all the squared interdistances?
The total inertia of the cloud is….?

Principal component analysis

Computing u_1, u_2, etc. is performed through diagonalizing the matrix with general term:

$$c_{kk'} = p_i \sum_{i=1}^{I} \frac{(x_{ik} - \overline{x}_k)(x_{ik'} - \overline{x}_{k'})}{s_k s_{k'}} = cor(k, k')$$

diagonalizing **Z'DZ** **Z**=standardized data matrix;

D: diagonal matrix with the individual weights (=1/I)

$\rightarrow$ $\lambda_1 > \ldots > \lambda_s > \ldots > \lambda_S$ eigenvalues with $S \leq Min(I,K)$

u_1 u_s u_S standardized eigen vectors

Orthogonality of the vectors u_s

Principal component analysis

Coordinates of the individuals on u_s: $$\mathbf{F_s} = \mathbf{Z u_s}$$

$\boldsymbol{F_s}$ is:

•a lineal combination of the original variables (coefficients = elements of u_s)

• centred with variance $\lambda_s i$

F_s is called the s_{th} principal component

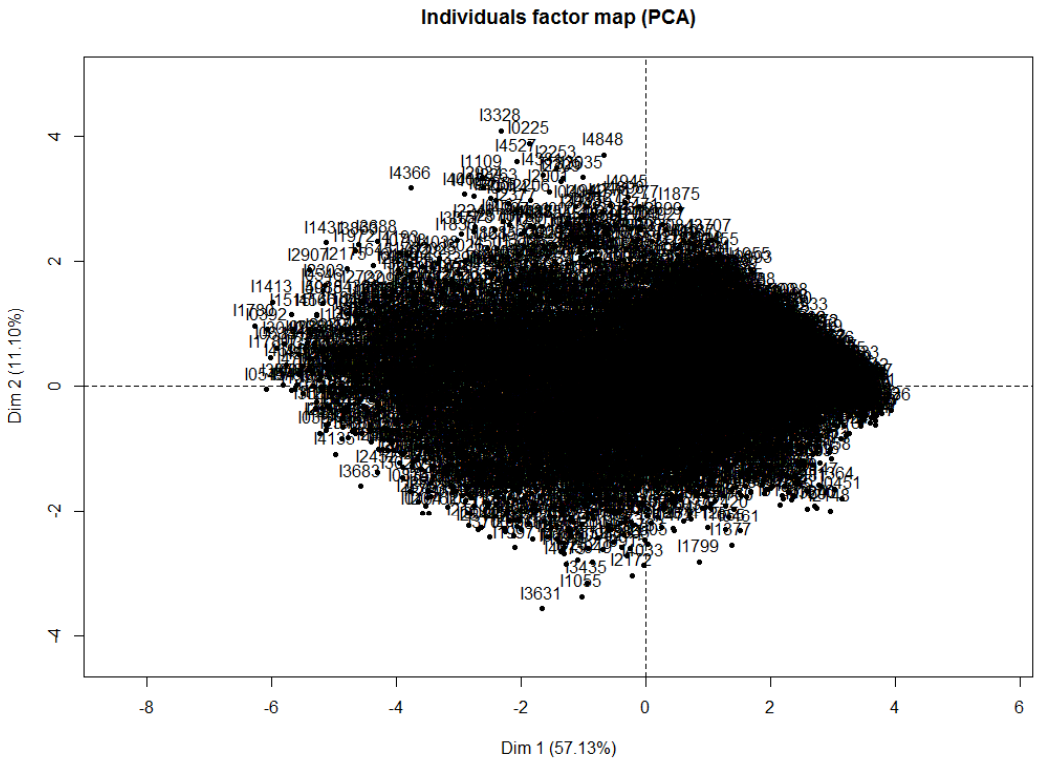
The principal axes $\boldsymbol{u_s}$ are orthogonal.
Thus a series of synthetic variables is defined, called principal components.
They are uncorrelated; they constitute the best summary of the initial variables.

Principal component analysis

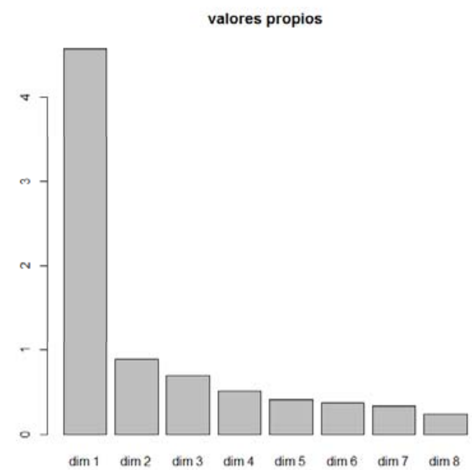
Possibly, supplementary individuals are considered

- They are not used to compute the axes
- Their position is computed after

Principal component analysis



	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.57	57.13	57.13
comp 2	0.89	11.10	68.23
comp 3	0.69	8.61	76.84
comp 4	0.51	6.35	83.19
comp 5	0.40	5.05	88.24
comp 6	0.37	4.63	92.87
comp 7	0.33	4.16	97.03
comp 8	0.24	2.97	100.00



```
> base[which(row.names(base)=="I0536"),45:52]
      PF_Phisica RP_Role.li RE_Role.li SF_Social MH_Mental EV_Energy P_Pain
I0536        100        100        100     88.89       100       100    100
      HP_General
I0536        100

> base[which(row.names(base)=="I3328"),45:52]
      PF_Phisica RP_Role.li RE_Role.li SF_Social MH_Mental EV_Energy P_Pain
I3328        100        100          0     22.22         4         0    100
      HP_General
I3328         45


> base[which(row.names(base)=="I3631"),45:52]
      PF_Phisica RP_Role.li RE_Role.li SF_Social MH_Mental EV_Energy P_Pain
I3631          5          0        100     77.78        92        45      0
      HP_General
I3631         20


> base[which(row.names(base)=="I1780"),45:52]
      PF_Phisica RP_Role.li RE_Role.li SF_Social MH_Mental EV_Energy P_Pain
I1780         25          0          0         0         0         0      0
      HP_General
I1780          0
```

Principal component analysis

In the variables space

Principal component analysis

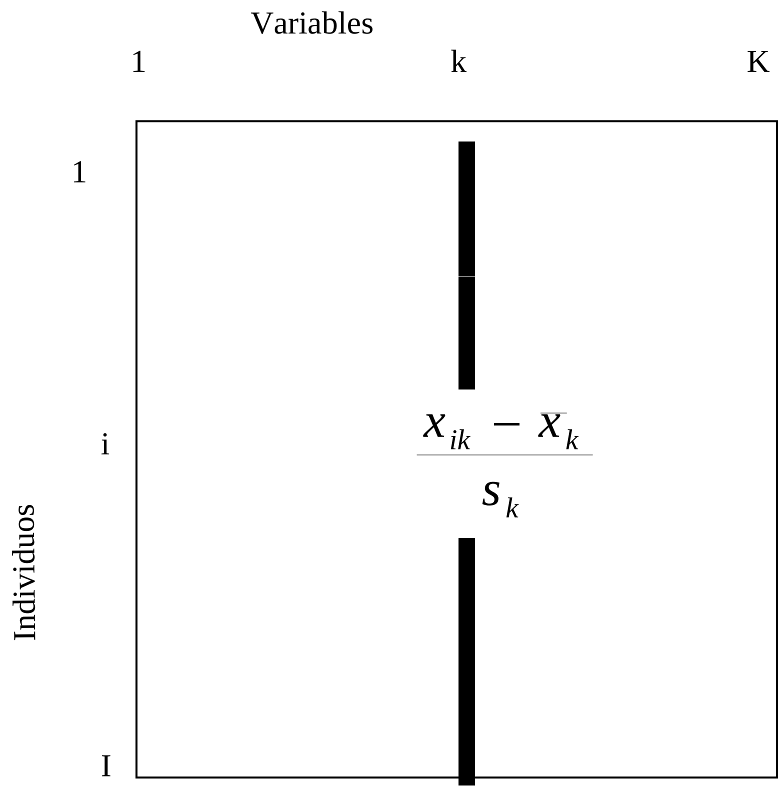
En R^I, because of the variables centring, all the variables lie on the hypersphere with radius 1

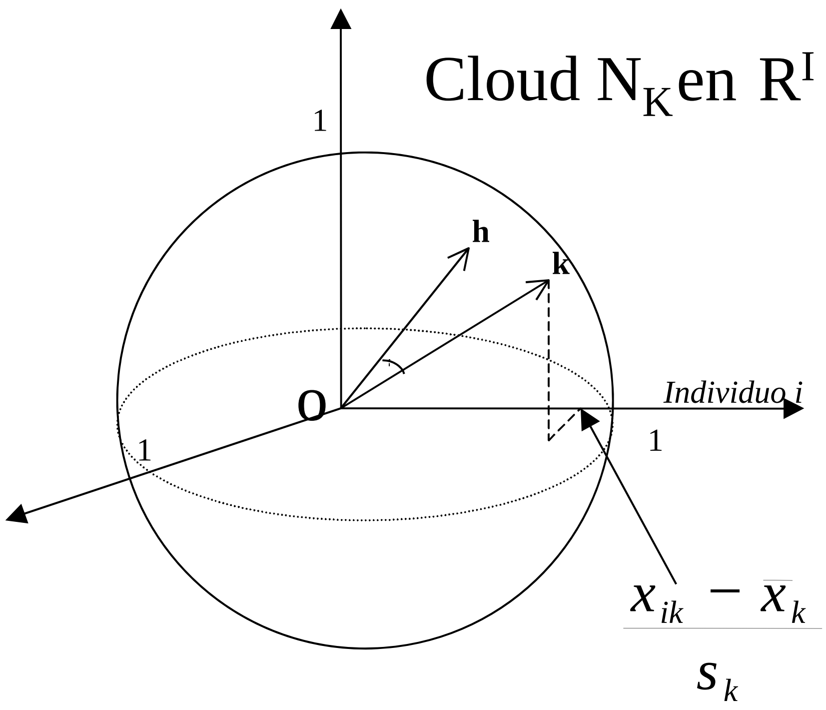
To center the variables means:

- in R^K, moving the centroid
- in R^I, projecting in parallel to the first bisectriz on the hyperplane orthogonal to it

Principal component analysis

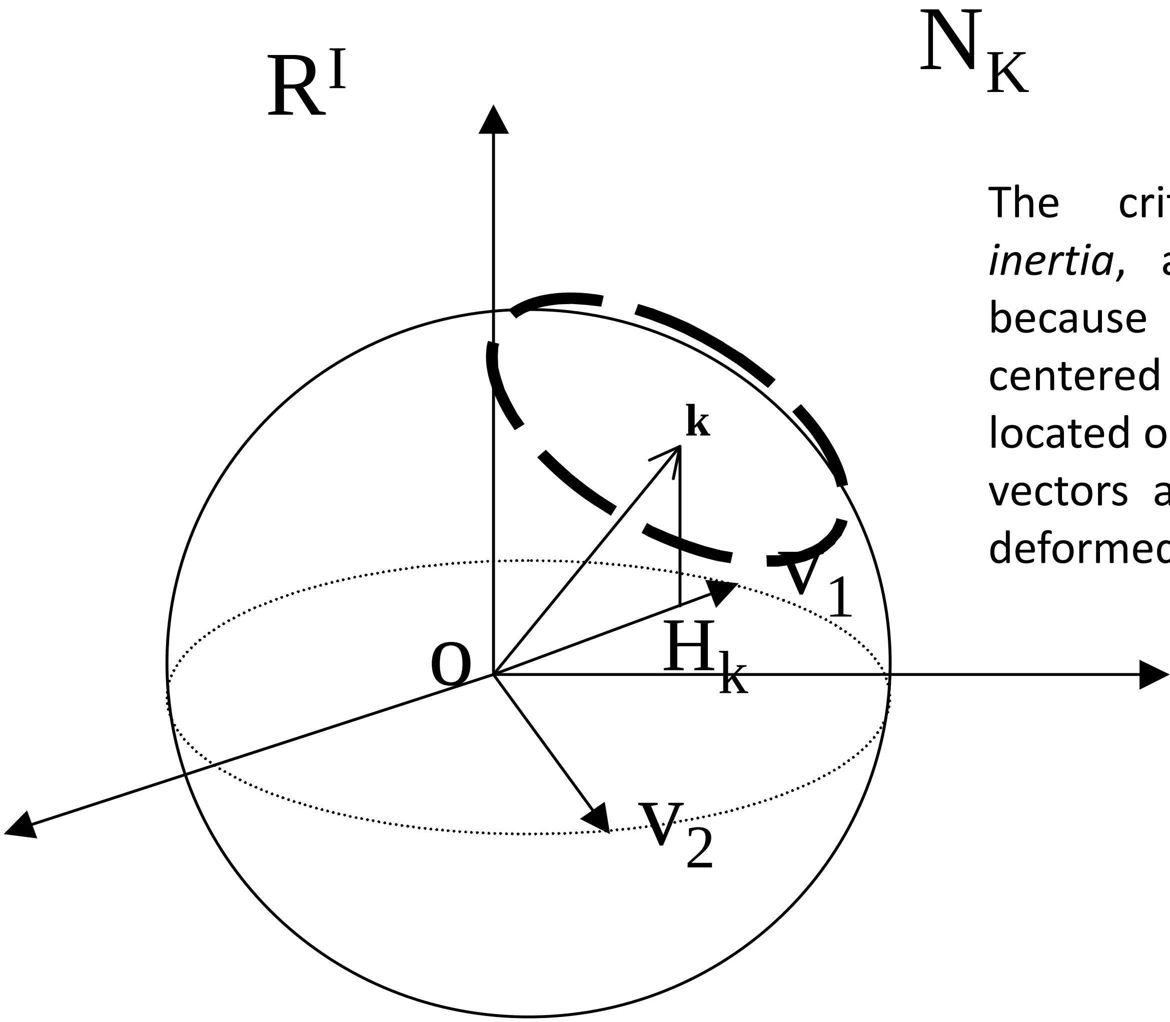
Cloud of variables





Principal component analysis

Cloud of variables



The criterion, *maximum projected inertia*, acquires a different meaning because the variable cloud is not centered and because all the points are located on the unit sphere: the between vectors angles are as little as possible deformed by the projections.

Principal component analysis

Distances between variables

We suppose all the individuals weights igual to 1

$$d^2(k,k') = 2(1 - c_{kk'})$$

$$0 \leq d^2(k,k') \leq 4$$

En R^I, the cosine of the angle between two vectors is equal to the correlation between the two variables.

Maismum inertia axes in the variable space

diagonalizing **ZZ'D** **Z**=standardized data matrix;
D: diagonal matrix with the individual weights (=1/I)

$\lambda_1 > \ldots > \lambda_s > \ldots > \lambda_S$ eigenvalues with $S \leq \mathrm{Min}(I,K)$

v_1 v_s v_S standardized eigen vectors

Orthogonality of the vectors u_s

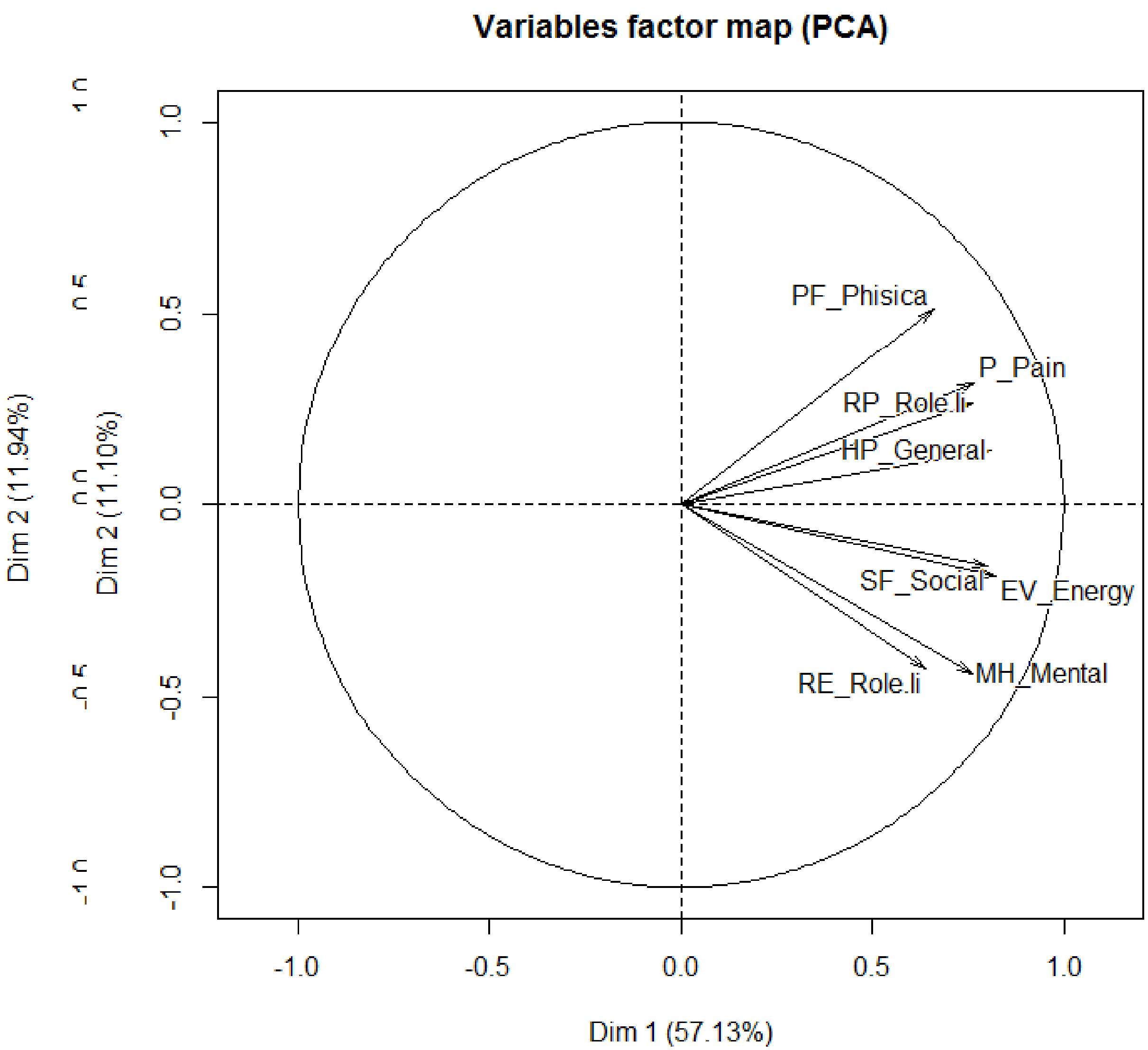
Principal component analysis

The vector v_1 characterizes the maximum inertia direction

As the variables are centered and standardized, **the projection on $\mathbf{v_1}$ of any variable is equal to the coefficient of correlation with this variable**.

The coordinates of the *K* variables on axis s are computed as:

$$G_s = Z' v_s$$


Variables factor map (PCA)
PF_Phisica
P_Pain
RP_Role.li
HP_General
SF_Social
EV_Energy
MH_Mental
RE_Role.li
Dim 1 (57.13%)
Dim 2 (11.94%)
Dim 2 (11.10%)
-1.0
-0.5
0.0
0.5
1.0

Principal component analysis

Transition relationships between both spaces

Principal component analysis

Duality and transition formulas in PCA

The clouds of individuals and variables are two representations of the same table.

There are strong relations between the two representations, called duality relationships.

Principal component analysis

The total inertia of both clouds is the same

$$Inertia = \frac{1}{I}\sum_{k}\sum_{i}\left(\frac{x_{ik}-\bar{x}}{s_{k}}\right)^{2} = \text{sum of variances}$$

Inertia total= nr of variables when the variables are standardized

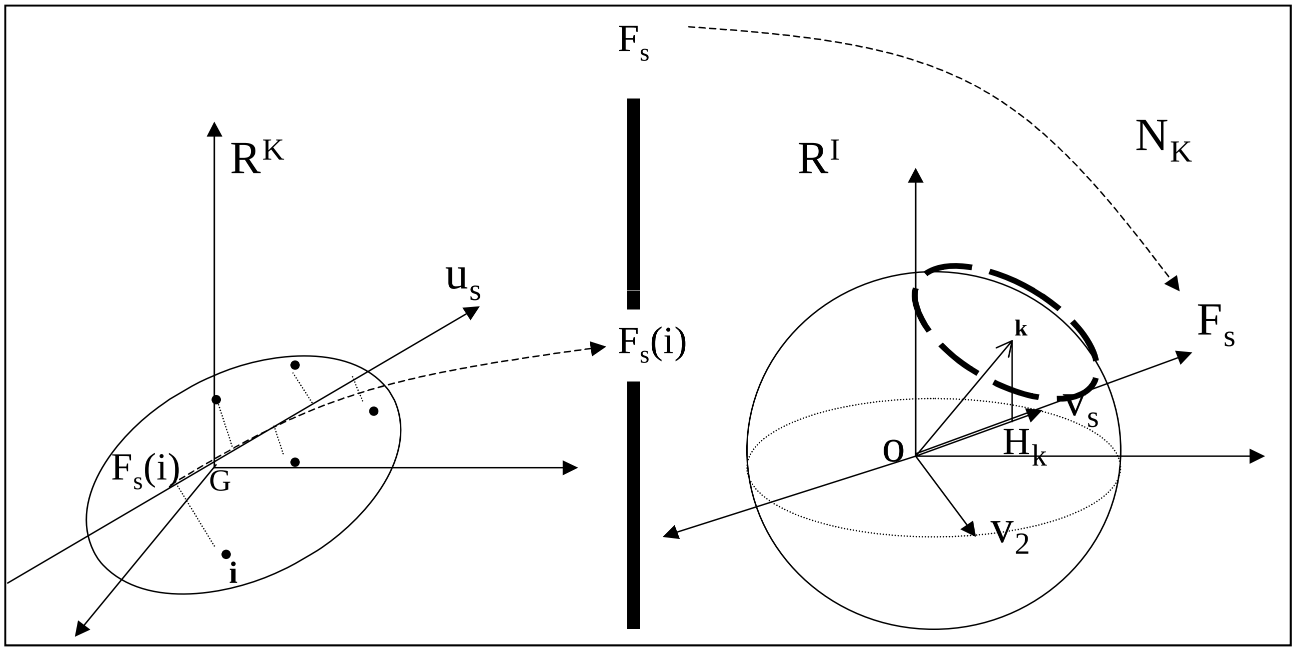
In general, the inertia is equal to
- the sum of variances of the variables
- the sum of the trace of the variance-covariance matrix

Thus, this analysis performs a decomposition of the total inertia equivalent in both spaces. The inertia projected onto the same rank axis are equal.

Principal component analysis

Principal components

Factor *s* or principal component *s* on individuals: projection of all points of the cloud of individuals on the axis *s* noted F_s

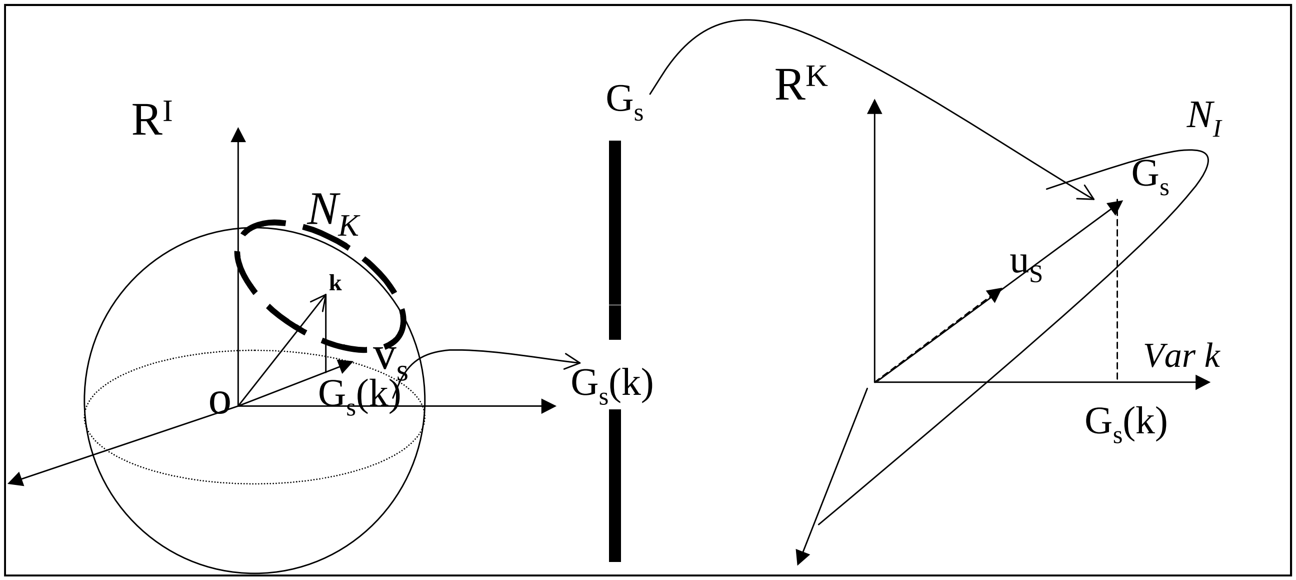


$$v_s = \frac{1}{\sqrt{\lambda_s}} F_s$$

Relationship between the factor F_s and factorial axis v_s

Principal component analysis

Factor s on the variables G_s: projection of the K variables on the factorial axis v_s. The set of values is the factor s on variables noted G_s



Relationship between axes u_s and factor G_s $$u_s = \frac{1}{\sqrt{\lambda_s}} G_s$$

Principal component analysis

Transition relationships

They are deduced from the relationships between axes and factors:

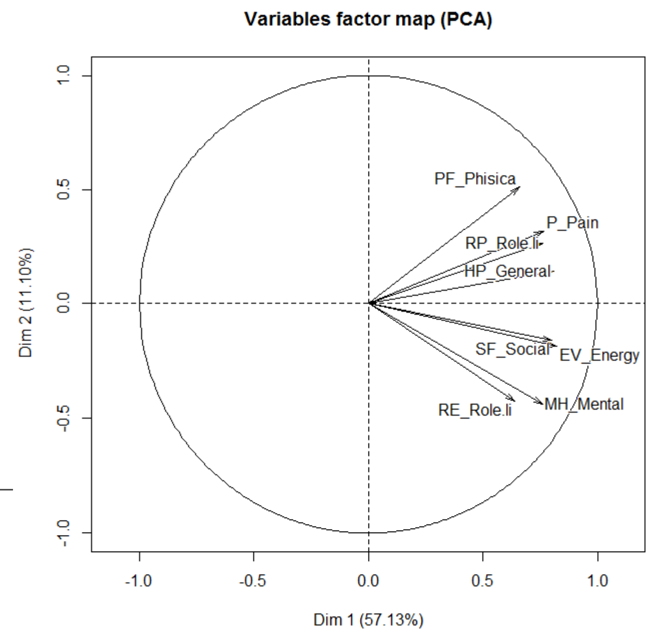
$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k \frac{x_{ik} - \bar{x}_k}{s_k} G_s(k) \qquad G_s(k) = \frac{1}{I} \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{x_{ik} - \bar{x}_k}{s_k} F_s(i)$$

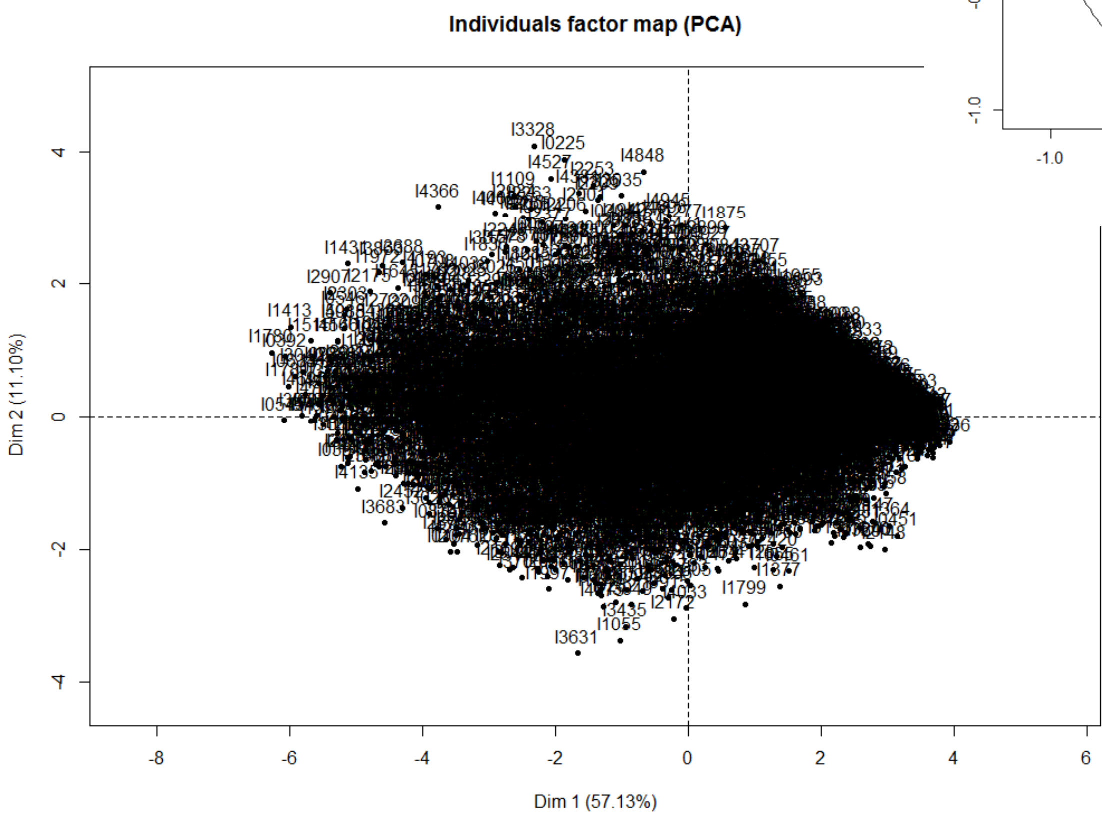
In practice, they are computed :

$$G_s(k) = \sqrt{\lambda_s}\, u_{sk}$$

$$G_s(k) = corr(k, F_s)$$

Simultaneous reading of both graphics





Principal component analysis

Helps to interpretation

Quality representation of the projection of the cloud on an axis, a plane, etc.

$$\frac{\sum_{1}^{q} \lambda_s}{\sum_{1}^{K} \lambda_s}$$

eigenvalue	percentage of variance	cumulative percentage of variance	
comp 1	4.57	57.13	57.13
comp 2	0.89	11.10	68.23

Quality of representation of an element on an axis: ratio between inertia of the projection of the cloud / total inertia

$$QLT_s(i) = \frac{\left(OH_i^s\right)^2}{\left(Oi\right)^2} = \cos^2\theta$$

Principal component analysis

Quality of representation of an element on an axis:

$$QLT_s(i) = \frac{\left(OH_i^s\right)^2}{\left(Oi\right)^2} = \cos^2\theta$$

Contribution of a row-element to the inertia of an axis $\frac{1}{I} \cdot \frac{F_s^2(i)}{\lambda_s}$

Contribution of a column-element to the inertia of an axis $\frac{G_s^2(k)}{\lambda_s}$

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
I0001	3.196	-2.552	0.028	0.637	0.479	0.005	0.022	-1.273	0.047	0.159
I0002	3.381	-1.275	0.007	0.142	2.595	0.151	0.589	-0.299	0.003	0.008
I0003	3.233	-1.676	0.012	0.269	-0.410	0.004	0.016	2.452	0.173	0.575
I0004	2.639	-0.299	0.000	0.013	-1.099	0.027	0.173	0.970	0.027	0.135
I0005	2.507	1.200	0.006	0.229	-0.847	0.016	0.114	-0.721	0.015	0.083
I0006	1.780	-0.624	0.002	0.123	-1.192	0.032	0.449	0.392	0.004	0.048
I0007	4.718	-4.555	0.090	0.932	0.937	0.020	0.039	0.054	0.000	0.000
I0008	4.681	-4.415	0.085	0.890	1.064	0.025	0.052	0.888	0.023	0.036
I0009	2.780	2.649	0.030	0.908	-0.107	0.000	0.001	-0.056	0.000	0.000
I0010	1.562	-0.256	0.000	0.027	-0.171	0.001	0.012	1.150	0.038	0.542

Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
PF_Phisica	0.660	9.536	0.436	0.511	29.386	0.261	-0.047	0.324	0.002
RP_Role.li	0.760	12.638	0.578	0.264	7.859	0.070	0.319	14.730	0.101
RE_Role.li	0.641	8.988	0.411	-0.428	20.647	0.183	0.533	41.285	0.284
SF_Social	0.803	14.104	0.645	-0.161	2.929	0.026	0.137	2.743	0.019
MH_Mental	0.761	12.687	0.580	-0.440	21.778	0.193	-0.308	13.808	0.095
EV_Energy	0.824	14.845	0.678	-0.187	3.958	0.035	-0.349	17.693	0.122
P_Pain	0.766	12.833	0.586	0.316	11.249	0.100	0.074	0.798	0.005
HP_General	0.810	14.369	0.657	0.140	2.195	0.019	-0.244	8.619	0.059

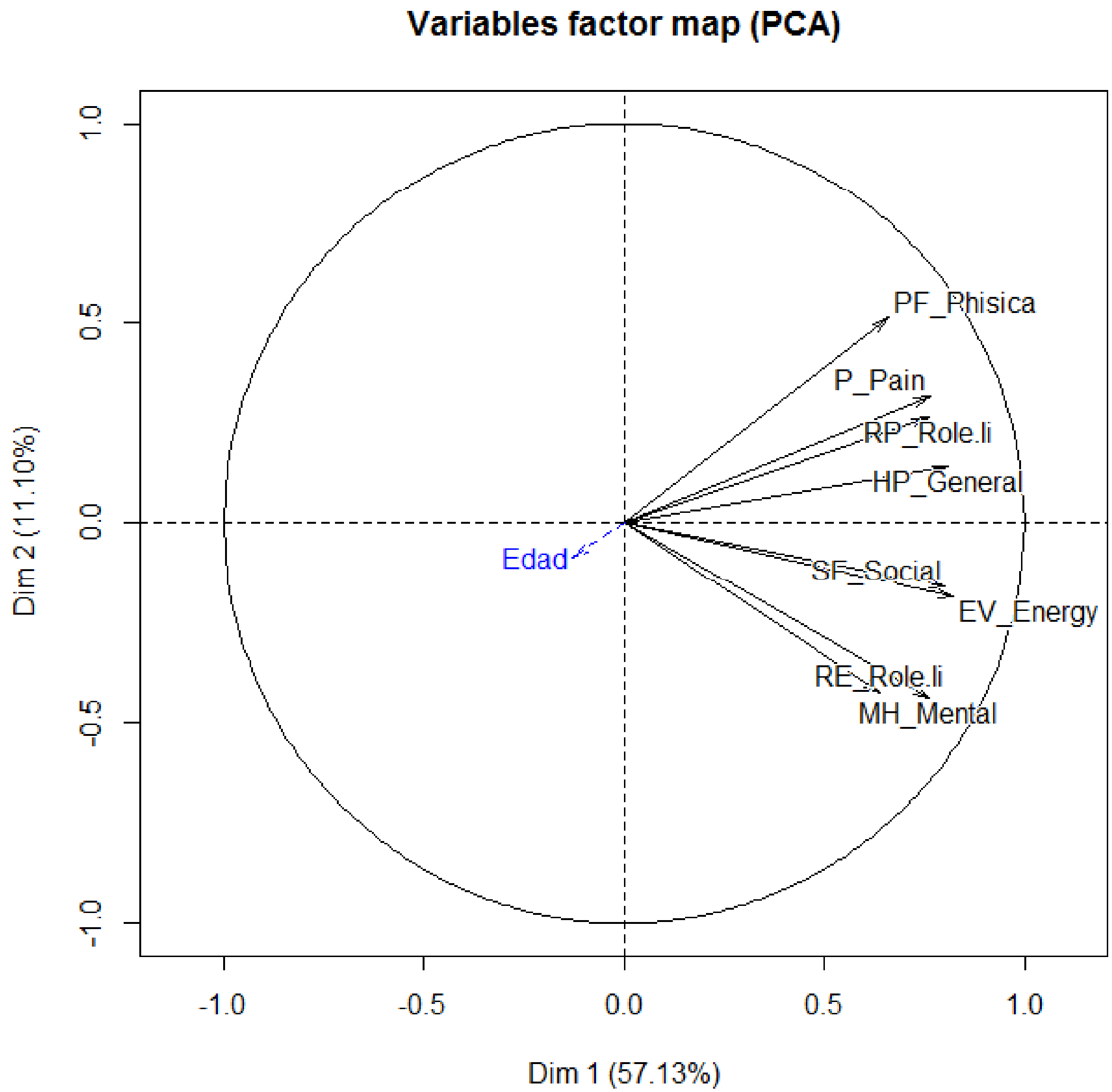
Principal component analysis

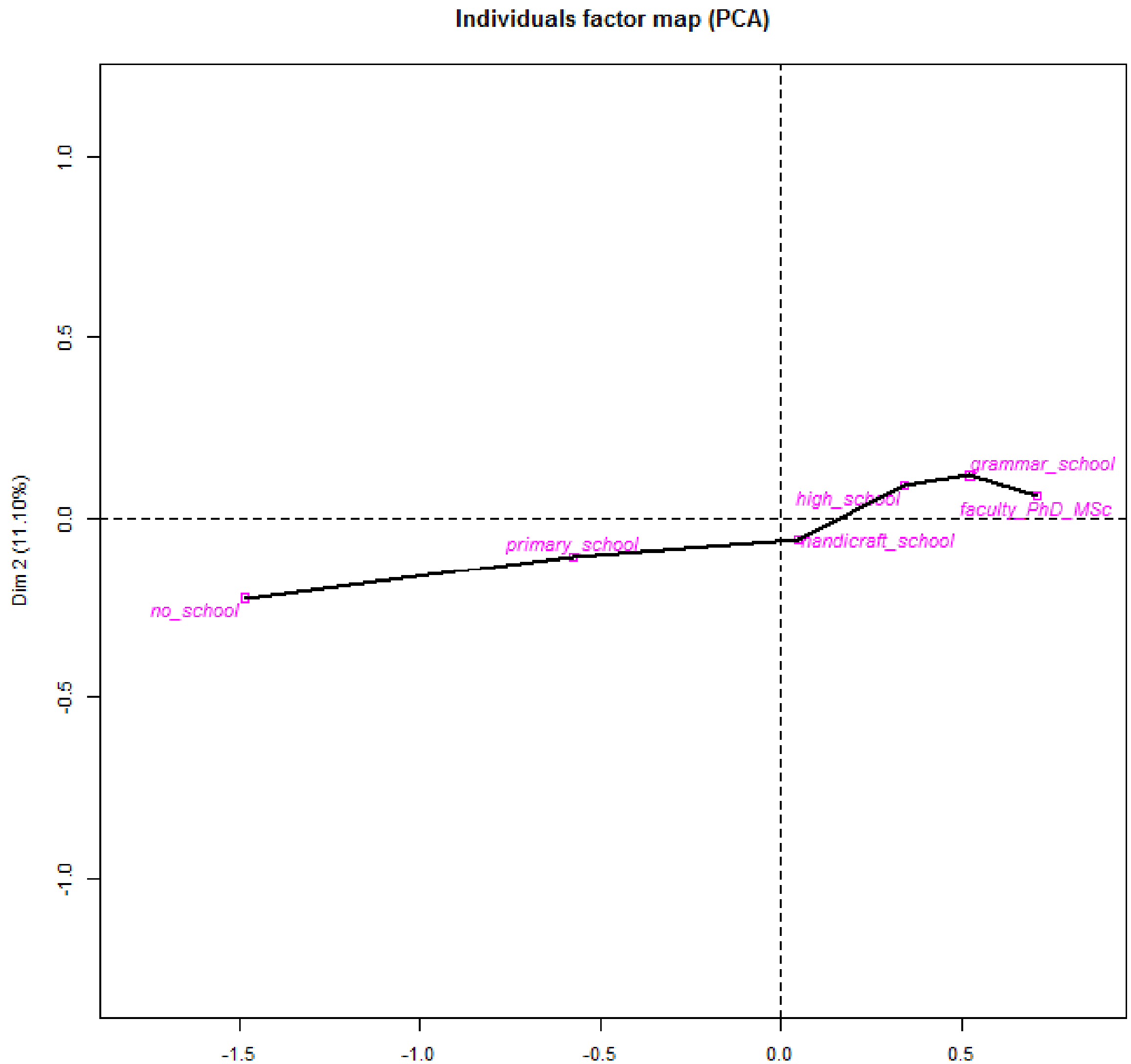
Supplementary individuals and variables

• To project a supplementary or illustrative individual, its values for each of the variables, but centered (centered or standardized) are computed. Then the transition relationships are used to place it on every axis.

• A supplementary quantitative variable is placed on the axes through its correlations with the principal components.

• A category of a supplementary categorical variable (with *m* categories) is placed as centroid of the individuals that belong to this category.


Variables factor map (PCA)
PF_Phisica
P_Pain
RP_Role.li
HP_General
SF_Social
EV_Energy
RE_Role.li
MH_Mental
Edad
Dim 1 (57.13%)
Dim 2 (11.10%)
-1.0
-0.5
0.0
0.5
1.0



The trajectories are very informative

```
> res.pca$quanti.sup
$coord
        Dim.1      Dim.2      Dim.3       Dim.4
Edad -0.1281551 -0.08457451 0.02916865 -0.02786831

$cor
        Dim.1      Dim.2      Dim.3       Dim.4
Edad -0.1281551 -0.08457451 0.02916865 -0.02786831

$cos2
        Dim.1      Dim.2       Dim.3        Dim.4
Edad 0.01642372 0.007152847 0.0008508104 0.0007766429
```

```
Supplementary categories
                   Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test
> res.pca$quali.sup
$coord
                       Dim.1       Dim.2       Dim.3        Dim.4
no_school        -1.48532072 -0.22725895  0.10768854 -0.130207430
primary_school   -0.57712759 -0.10756856  0.06903437 -0.034940085
handicraft_school  0.04573759 -0.06009212  0.01997059  0.023237323
high_school       0.34089128  0.08958263  0.01962310  0.009142059
grammar_school    0.52126938  0.11718080 -0.06698071  0.044732741
faculty_PhD_MSc   0.70842423  0.06198619 -0.05157250  0.026265151

$cos2
                      Dim.1       Dim.2       Dim.3        Dim.4
no_school         0.9628600 0.022540558 0.005061295 0.0073993608
primary_school    0.9224329 0.032045114 0.013198415 0.0033809521
handicraft_school 0.1529803 0.264072968 0.029165593 0.0394876374
high_school       0.8803867 0.060798005 0.002917273 0.0006331836
grammar_school    0.9227651 0.046631526 0.015235835 0.0067954428
faculty_PhD_MSc   0.9416256 0.007209104 0.004990314 0.0012943485

$v.test
                        Dim.1     Dim.2      Dim.3      Dim.4
no_school         -18.6587097 -6.476554  3.4848229 -4.9057692
primary_school     -9.5764156 -4.049283  2.9508446 -1.7388619
handicraft_school   0.5646265 -1.682934  0.6350805  0.8603677
high_school         3.0155781  1.797795  0.4471696  0.2425541
grammar_school     13.6376385  6.954967 -4.5141562  3.5100431
faculty_PhD_MSc     8.0044668  1.588896 -1.5010912  0.8900797

$eta2
          Dim 1     Dim 2       Dim 3       Dim 4
SKOLA 0.1109649 0.0173581 0.006497085 0.006450695
```

Principal component analysis

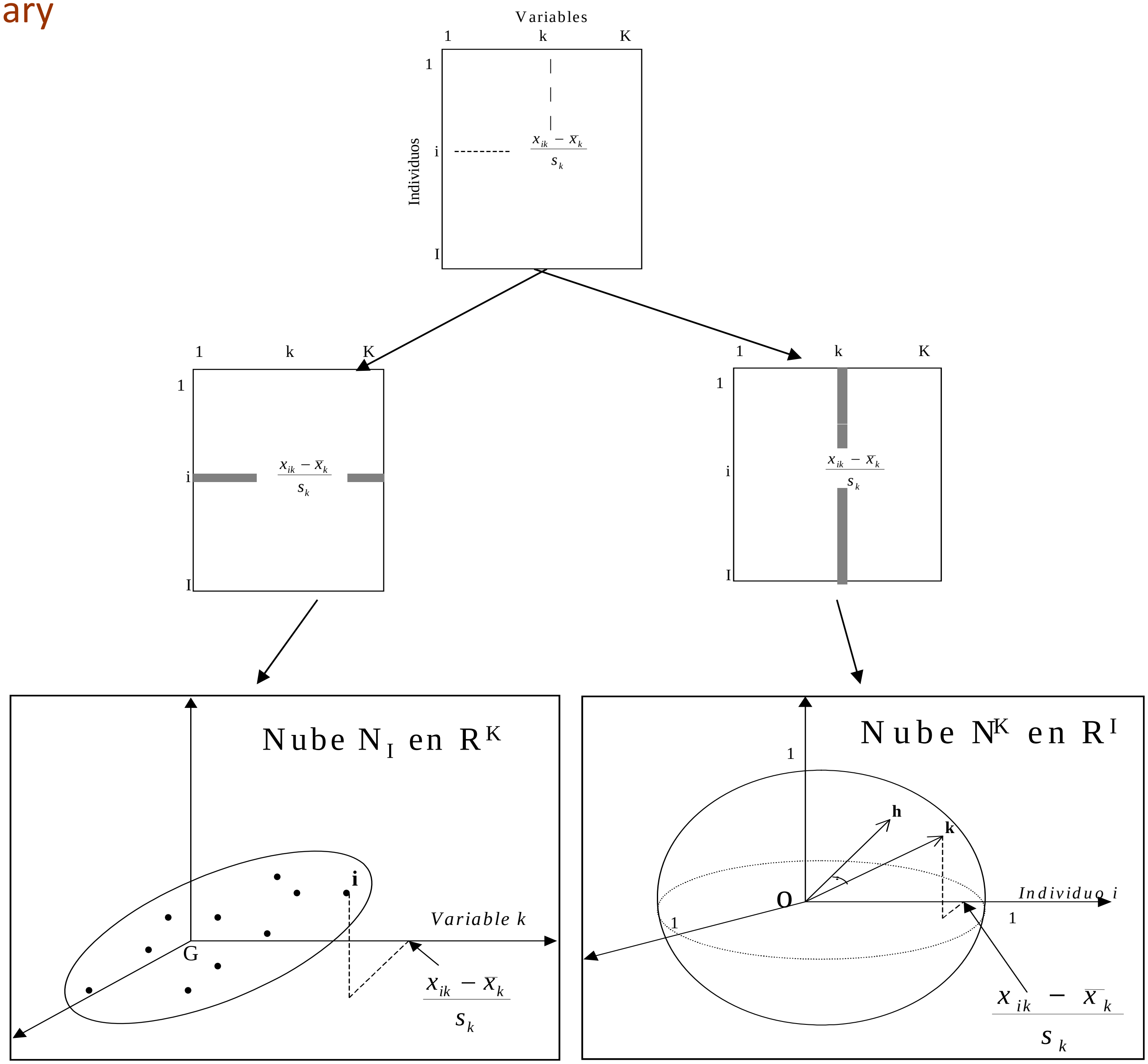
Separate representation of both clouds

The two clouds are not in the same space; They do not have the same referential

The similarities between individuals are interpreted as corresponding to similar behavior as far as the active variables are concerned

The proximity between variables are interpreted as correlations

As a summary



Variables
1
k
K
1
Individuos
i
I
$\frac{x_{ik} - \bar{x}_k}{s_k}$
1
k
K
1
i
I
$\frac{x_{ik} - \bar{x}_k}{s_k}$
1
k
K
1
i
I
$\frac{x_{ik} - \bar{x}_k}{s_k}$
Nube N_I en R^K
i
G
Variable k
$\frac{x_{ik} - \bar{x}_k}{s_k}$
Nube N^K en R^I
1
h
k
O
Individuo i
1
1
$\frac{x_{ik} - \bar{x}_k}{s_k}$