Clustering

Anàlisi de Dades i Explotació de la Informació

Grau d'Enginyeria Informatica.

Information System track

Prof. Mónica Bécue Bertaut & Lidia Montero

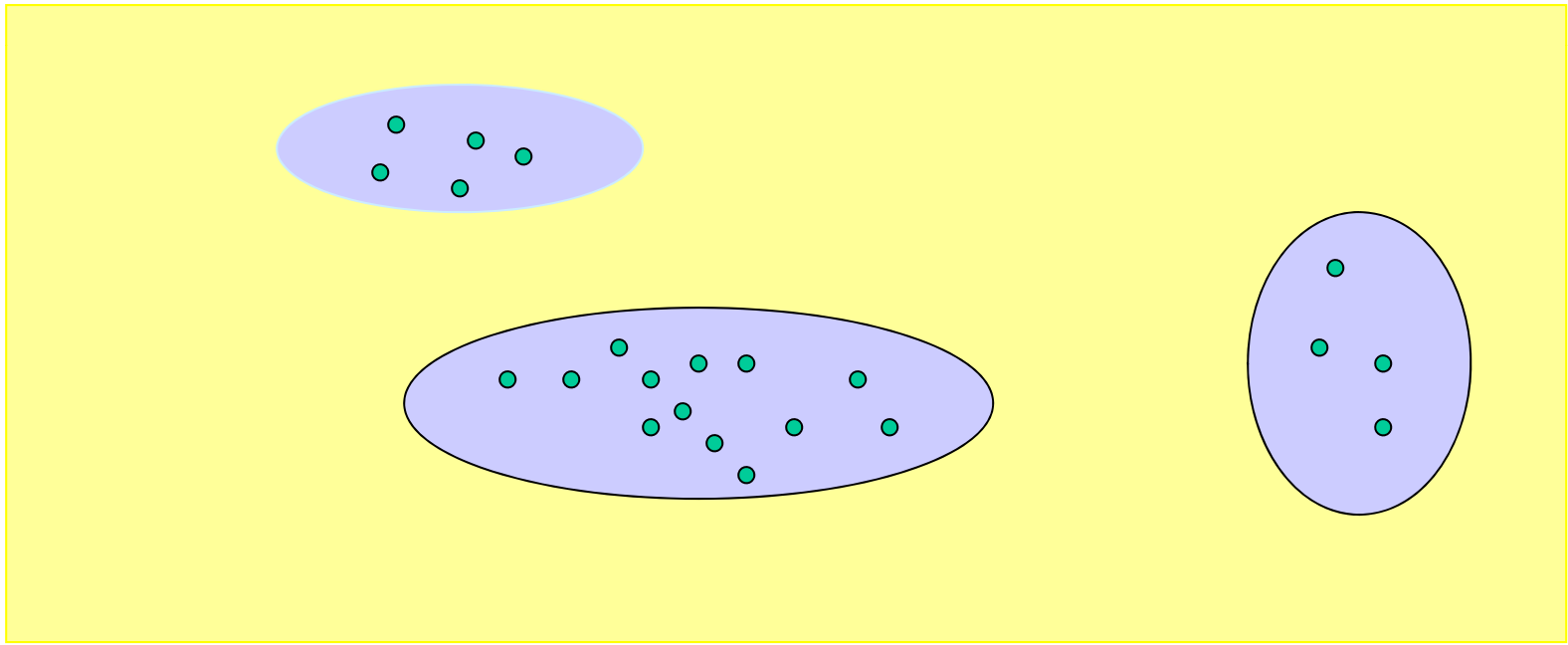
Monica.becue@upc.edu *lidia.montero@upc.edu*

Outline

1. Principles
2. Direct partitioning
3. Hierarchical clustering
4. Clustering and principal axes methods
5. Clusters description

1. Principles

Objective: to group the elements into “homogeneous and well separated clusters”



Two main families of clustering methods:

- Direct partitioning methods
- Hierarchical methods

2. Direct partitioning

A distance is supposed to be defined

on the set of elements to cluster

- Manhattan

$$d(i,l) = \sum_{k=1}^{K} |x_{ik} - x_{lk}|$$

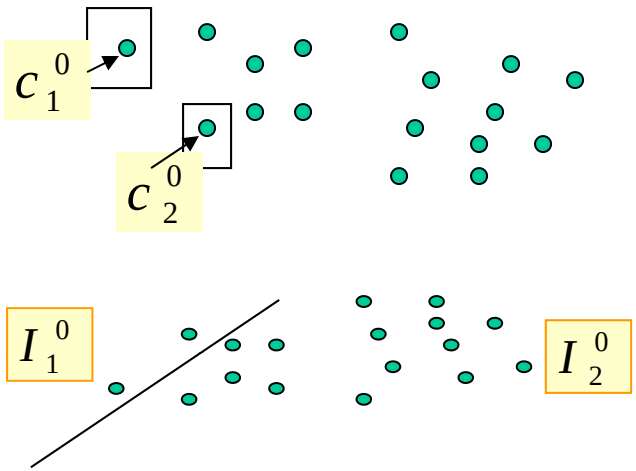
- Euclidean

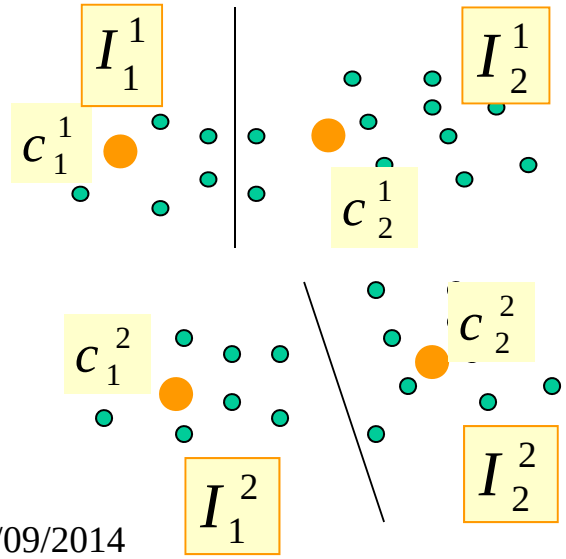
$$d^2(i,l) = \sum_{k=1}^{K} (x_{ik} - x_{lk})^2$$

used with factorial analysis (orthogonal projection)

29/09/2014 Clustering 6

Mobile centres





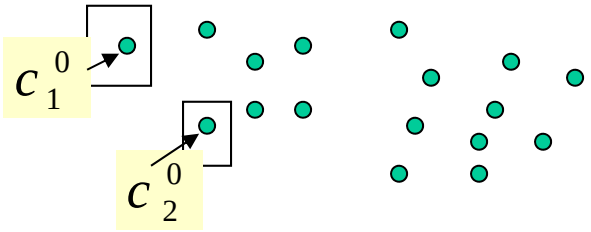
1. Random selection of q centres at stp 0 (here, q=2)

2. These centres determine 2 zones, which are considered as the two clusters at this step 0

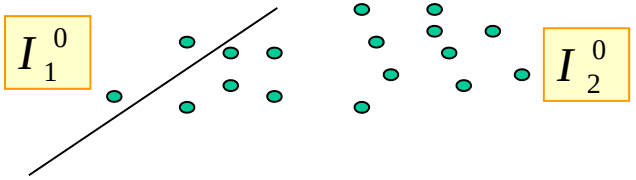
3. The "true" centroids of the clusters computed at step 0 are computed. Thus a new partition in two zones (clusters) is determined (=clusters at step 1)

4. The "true" centroids of the clusters computed at step 1 are computed. Thus a new partition in two zones (clusters) is determined (=clusters at step 2).

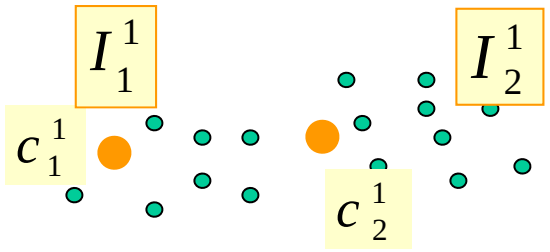
… This process stabilizes



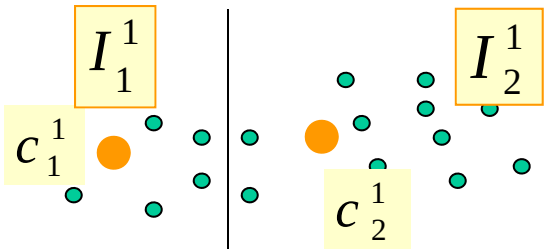
Step 0: Random selection of 2 centers



Step 0: formation of partition P^0



Step 1: Centroids of P^0



Step 1: Formation ofpartition P^1

. . .

Close techniques

K-means (MacQueen, 1967)

q centroids are randomly selected

but the centroids are newly computed at every re-assignation of an individual to a center

In only one iteration, a partition of good enough quality is obained although the final results (=clusters) depend on the order of the individuals in the file.

Close techinques

Strong forms and stable groups(Diday, 1972)

To try to have a better solution (closer to the global optimum), several partitions are computed in the following way:

Strong forms and stable groups(Diday, 1972)

Partition 1

Partition 2

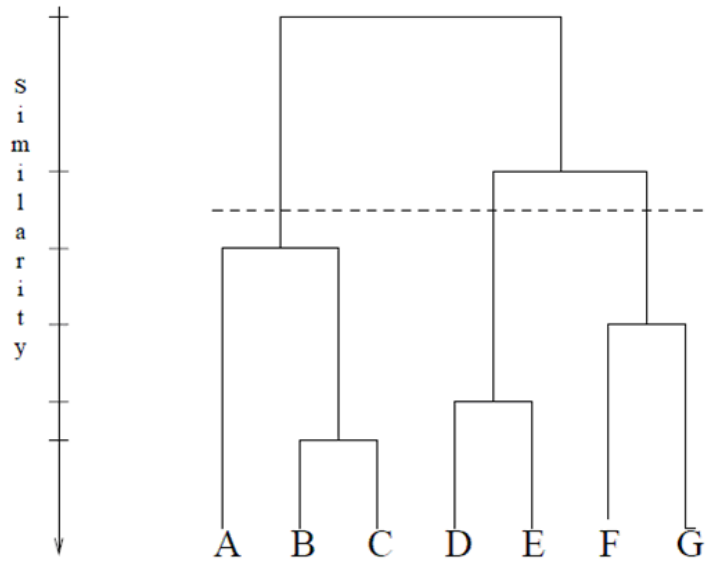
113	**38**	**35**	**40**
30	**5**	**25**	**0**
43	**30**	**8**	**5**
40	**3**	**2**	**35**

Strong forms

Product partition

3. Hierarchical clustering

Hierarchical clustering



- **A distance is supposed to be defined**
  - on the set of elements to cluster
  - on “groups” of elements (aggregation criterium/ method)

Distance on the set of elements

- Manhattan

$$d(i,l) = \sum_{k=1}^{K} |x_{ik} - x_{lk}|$$

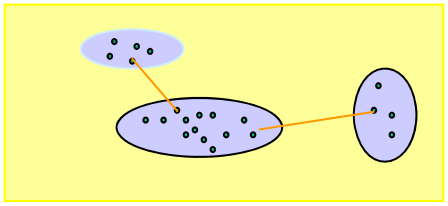
- Euclidean

$$d^2(i,l) = \sum_{k=1}^{K} (x_{ik} - x_{lk})^2$$

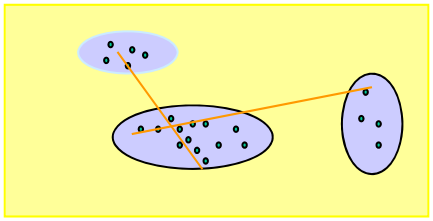
used with factorial analysis (orthogonal projection)

Usual aggregation criteria

- **Minimum linkage:** Chain effects are possible



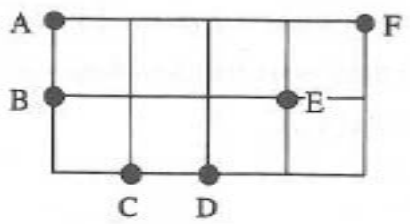
- **Complete linkage:** Compact clusters



- **Average linkage:** Compromise between the formers. Individual weights are taken into account

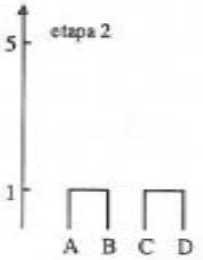
Example of clustering

- Distance=manhattan and criteria=complete linkage

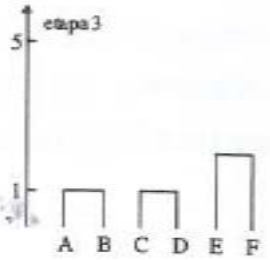


	A	B	C	D	E	F
A	0					
B	1	0				
C	3	2	0			
D	4	3	1	0		
E	4	3	3	2	0	
F	4	5	5	4	2	0

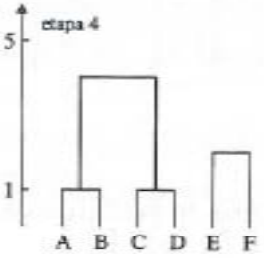
	AB	C	D	E	F
AB	0				
C	3	0			
D	4	1	0		
E	4	3	2	0	
F	5	5	4	2	0



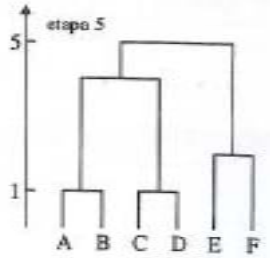
	AB	CD	E	F
AB	0			
CD	4	0		
E	4	3	0	
F	5	5	2	0



	AB	CD	EF
AB	0		
CD	4	0	
EF	5	5	0



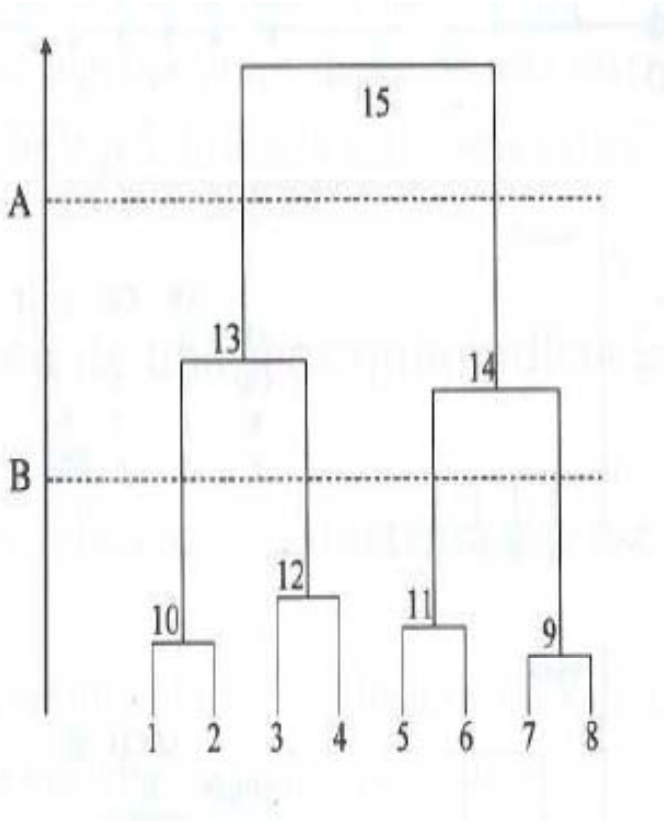
	ABCD	EF
ABCD	0	
EF	5	0



Final: **hierarchy** that can be cutted to provide a partition

Hierarchy and partition

- *n* elements to be classified are the terminal nodes (from 1 to *I*)
- The following nodes, group several elements, are numbered as *I*+1, *I*+2,….
- Cutting the tree provides a partition
- A hierarchy of partitions is obtained
- Remember that clustering aims to group the elements into "homogeneous and well separated clusters"



Ward criterion (I)

- Euclidean distance
- **Aim:** To find a good partition maximizing quality of the partition

  - A good partition means: inside clusters elements are homogenous and between clusters elements are diferent

- Inertia decomposition: $I_{total}=I_{inter}+I_{intra}$

$$\sum_{q=1}^{Q}\sum_{i=1}^{I_q}\sum_{k=1}^{K}\left(y_{iqk}-\bar{y}_k\right)^2=\sum_{q=1}^{Q}\sum_{k=1}^{K}I_q\left(\bar{y}_{qk}-\bar{y}_k\right)^2+\sum_{q=1}^{Q}\sum_{i=1}^{I_q}\sum_{k=1}^{K}\left(y_{iqk}-\bar{y}_{qk}\right)^2$$

  - Q: clusters, I_q: Elements in cluster q, K: variables
  - y_{iqk}: Value of element i from cluster q for variable k

Ward criterion (II)

- At every step, intra-cluster inertia increases
- Ward criterion means to aggregate, at every step, the two elements or the two nodes that leads to minimum increase in total intra-cluster inertia after merging
- For clusters p and q, Ward criterion has the following value:

$$\Delta(p,q) = \frac{I_p I_q}{I_p + I_q} d^2(g_p, g_q)$$

I_p: Elements in cluster p, I_q: Elements in cluster q

g_p: Gravity center of cluster p, g_q: Gravity center of cluster q

- Minimize $\Delta(p,q)$ means to choose:
  - Clusters with gravity centers closest ($d^2(g_p, g_q)$ low)
  - Clusters with few elements ($\frac{I_p I_q}{}$ low)

Choosing a partition

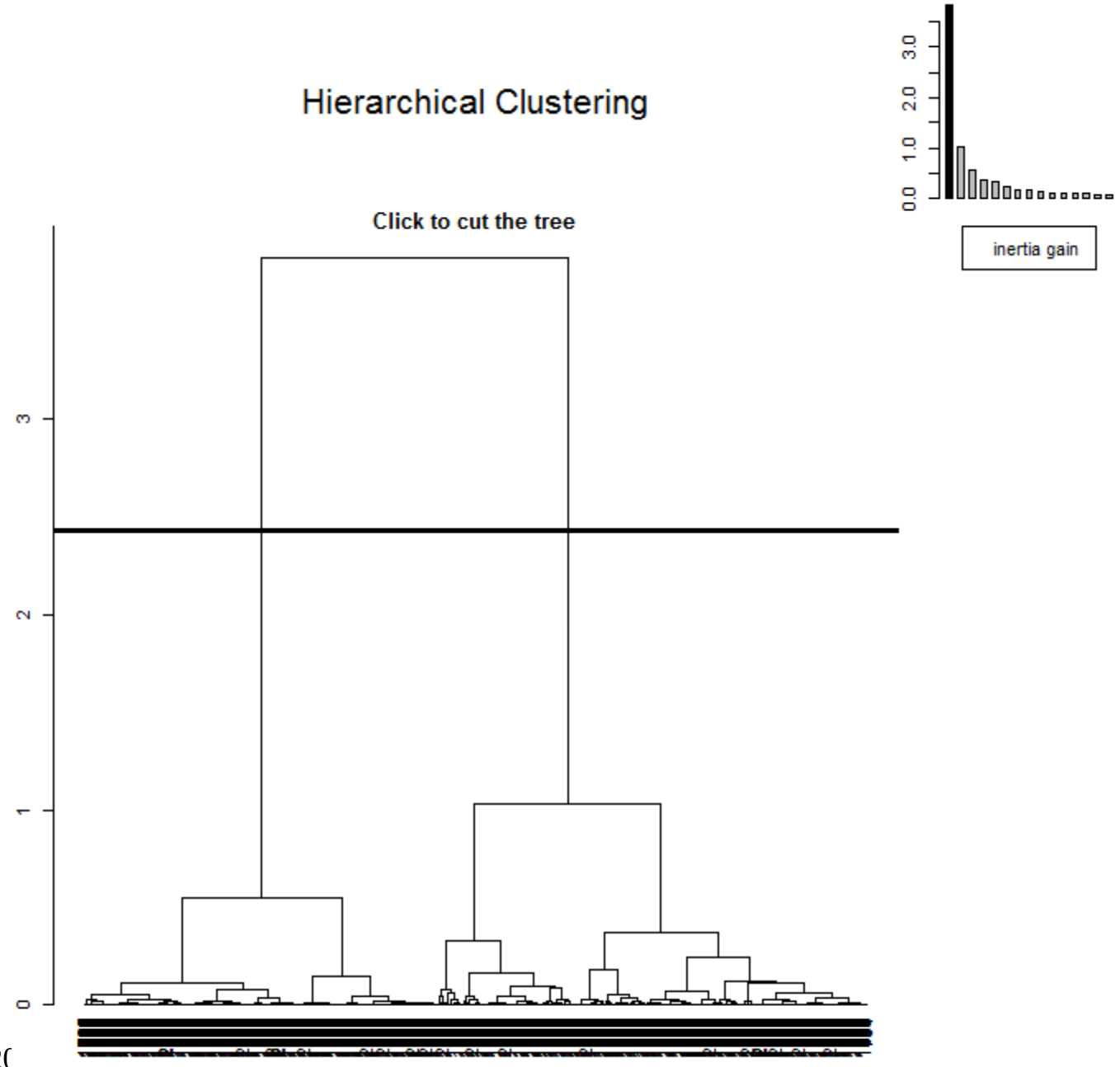
- Not too much clusters
- Take a partition that you can interpret
- Criterion to choose an optimum partition

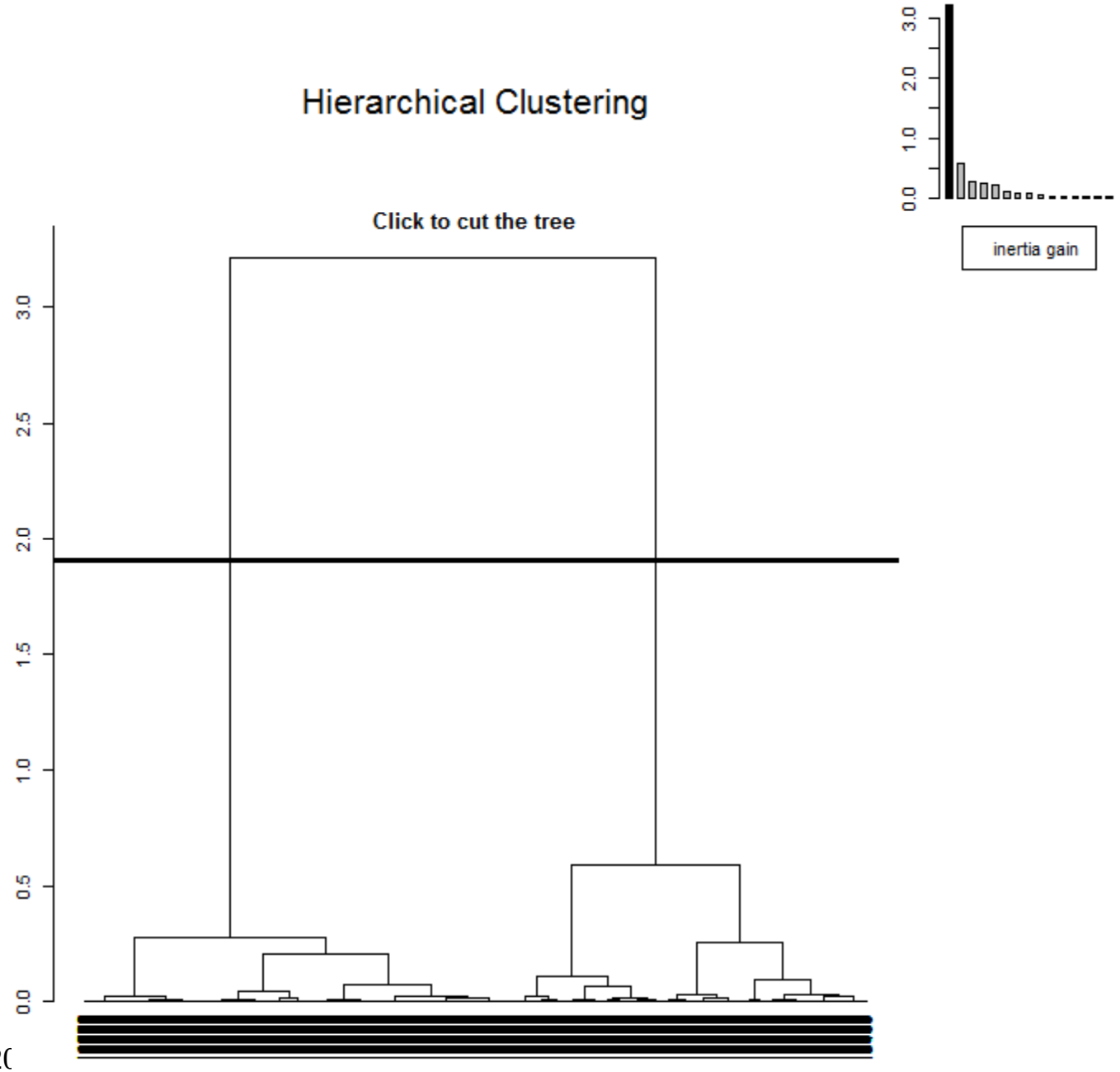
$$\min_{q_{min} \le q \le q_{max}} \frac{\Delta(q)}{\Delta(q+1)}$$

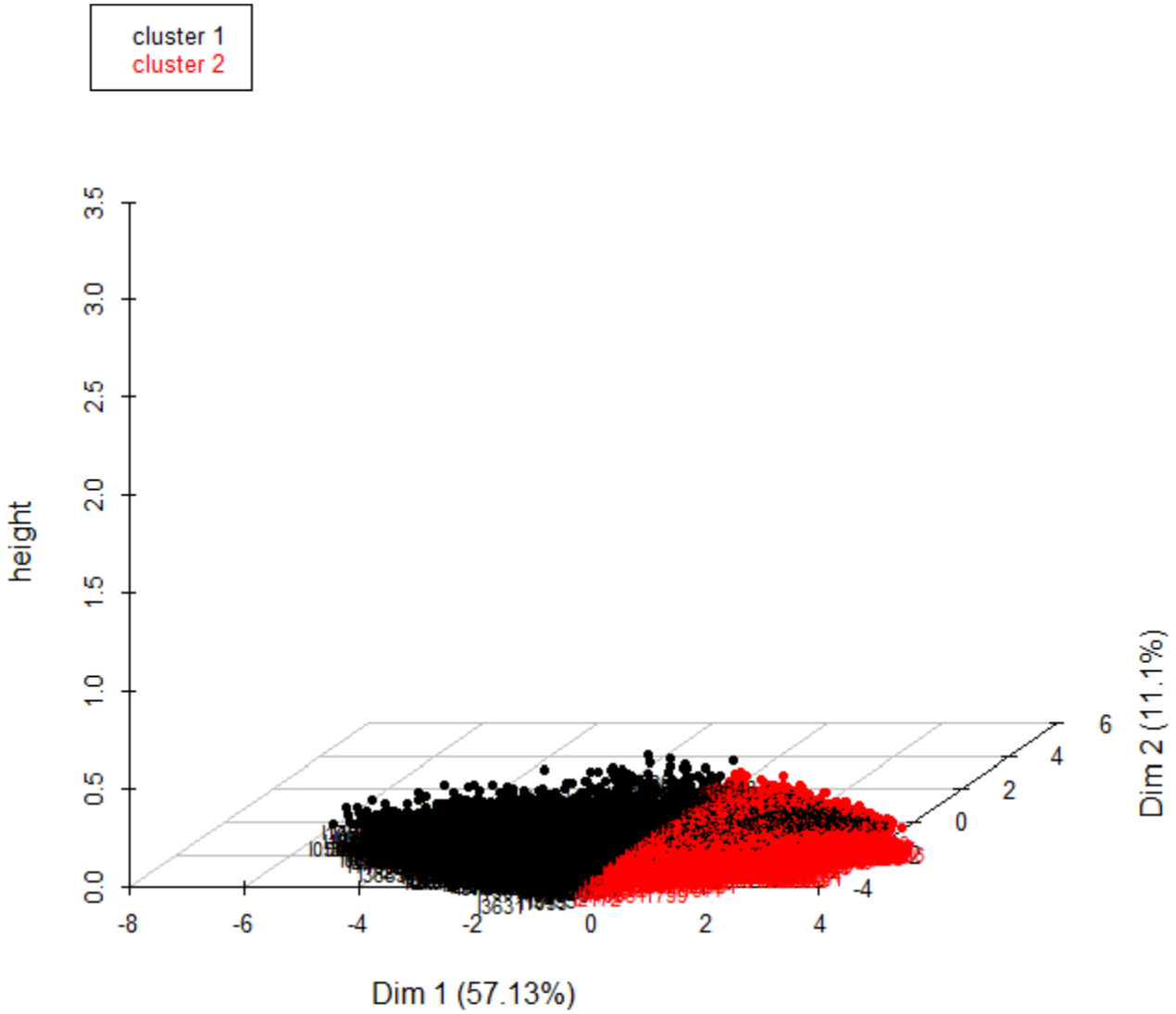
$\Delta(q)$: Gain of inter-cluster inertia going from *q-1* to *q* clusters

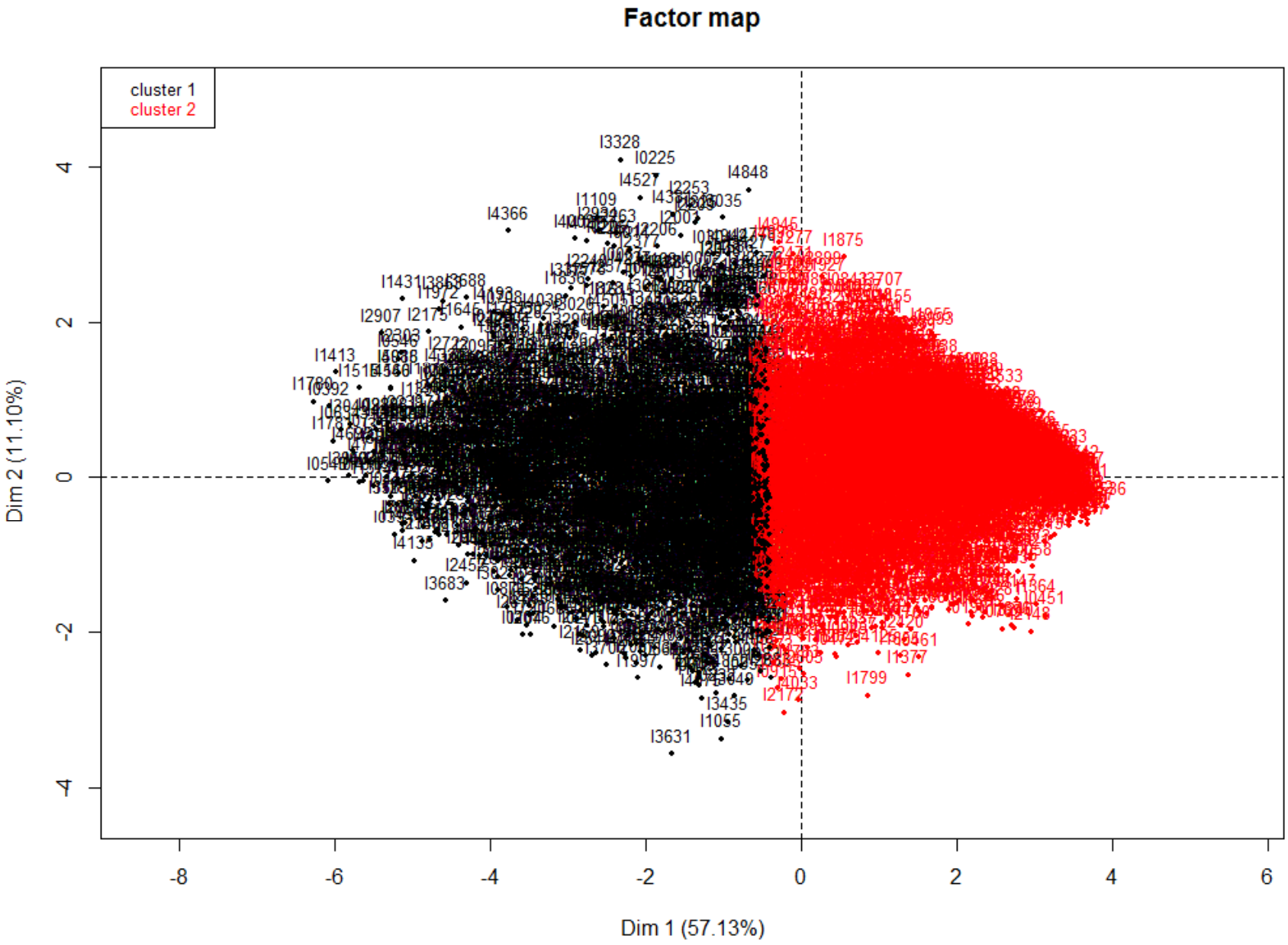
- Improving the final partition with "Consolidation"

4. Principal axes methods and clustering


Hierarchical Clustering
Click to cut the tree
0.0
1.0
2.0
3.0
inertia gain
0
1
2
3


Hierarchical Clustering
Click to cut the tree
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3.0
2.0
1.0
0.0
inertia gain


cluster 1
cluster 2
height
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
-8
-6
-4
-2
0
2
4
Dim 1 (57.13%)
-4
0
2
4
6
Dim 2 (11.1%)


Factor map
cluster 1
cluster 2
Dim 2 (11.10%)
Dim 1 (57.13%)
-8
-6
-4
-2
0
2
4
6
I3328
I0225
I4527
I4848
I2253
I1109
I4366
I1875
I1431
I2907
I2175
I1413
I1760
I0392
I4135
I3683
I1799
I3435
I1055
I3631

6. Cluster description

From the categorical variables

```
> res.hcpc$desc.var$test.chi2
                 p.value df
B1          0.000000e+00  4
Edad_Skol 3.245107e-156  8
Sex         2.257039e-12  1
```

From the categories

```
> res.hcpc$desc.var$category
$`1`
                        Cla/Mod    Mod/Cla    Global       p.value     v.test
B1=health_poor       90.412272 48.0387163 20.706770  0.000000e+00        Inf
Edad_Skol=OldLow     66.666667 37.3917473 21.858249  5.403491e-99  21.118273
B1=health_fair       48.108926 32.3993887 26.245781  3.360483e-15   7.876734
Sex=female           42.852470 66.7345899 60.690887  1.850993e-12   7.045272
Edad_Skol=OldMedium 48.005908 16.5562914 13.440540  2.880527e-07   5.131101
Edad_Skol=MidLow     48.265896  8.5073867  6.869168  2.801679e-04   3.632979
Edad_Skol=OldHigh    46.255507  5.3489557  4.506651  2.254253e-02   2.281100
Edad_Skol=MidMedium 34.720571 14.8751910 16.696446  5.406843e-03  -2.781739
Edad_Skol=MidHigh    27.345845  5.1961284  7.405202  1.065718e-06  -4.879099
Edad_Skol=JovLow     17.801047  1.7320428  3.791940  1.345397e-10  -6.421950
Sex=male             32.979798 33.2654101 39.309113  1.850993e-12  -7.045272
Edad_Skol=JovHigh    11.673152  1.5282731  5.102243  2.449031e-23  -9.952915
B1=health_good       23.189466 16.1487519 27.139170  9.788030e-47 -14.355876
Edad_Skol=JovMedium 16.992188  8.8639837 20.329561  4.079413e-64 -16.905773
B1=health_excellent   1.694915  0.4075395  9.370657  7.632784e-93 -20.438316
B1=health_very good   7.082833  3.0056037 16.537622 1.428371e-114 -22.750203
```

From the quantitative variables

```
> res.hcpc$desc.var$quanti.var
                 Eta2      P-value
PF_Phisica 0.30922824 0.000000e+00
RP_Role.li 0.50533202 0.000000e+00
RE_Role.li 0.29780836 0.000000e+00
SF_Social  0.46725019 0.000000e+00
….etc.
```

```
$`1`
              v.test Mean in category Overall mean sd in category Overall sd      p.value
Edad         8.061905         51.71472     49.08557       18.38630   18.49395 7.511481e-16
RE_Role.li -38.726772         44.86323     72.42412       44.61223   40.35830 0.000000e+00
PF_Phisica -39.462304         48.55833     69.92952       28.22450   30.71124 0.000000e+00
MH_Mental  -43.241226         46.73255     61.69069       17.88921   19.61691 0.000000e+00
P_Pain     -46.759448         39.82454     64.50947       24.40165   29.93736 0.000000e+00
EV_Energy  -47.369413         33.84870     51.83145       16.42762   21.52827 0.000000e+00
SF_Social  -48.508473         44.98724     64.83442       21.02165   23.20236 0.000000e+00
HP_General -48.525317         34.58839     54.26444       16.98673   22.99433 0.000000e+00
RP_Role.li -50.446527         25.01274     63.00377       34.73299   42.70719 0.000000e+00
```

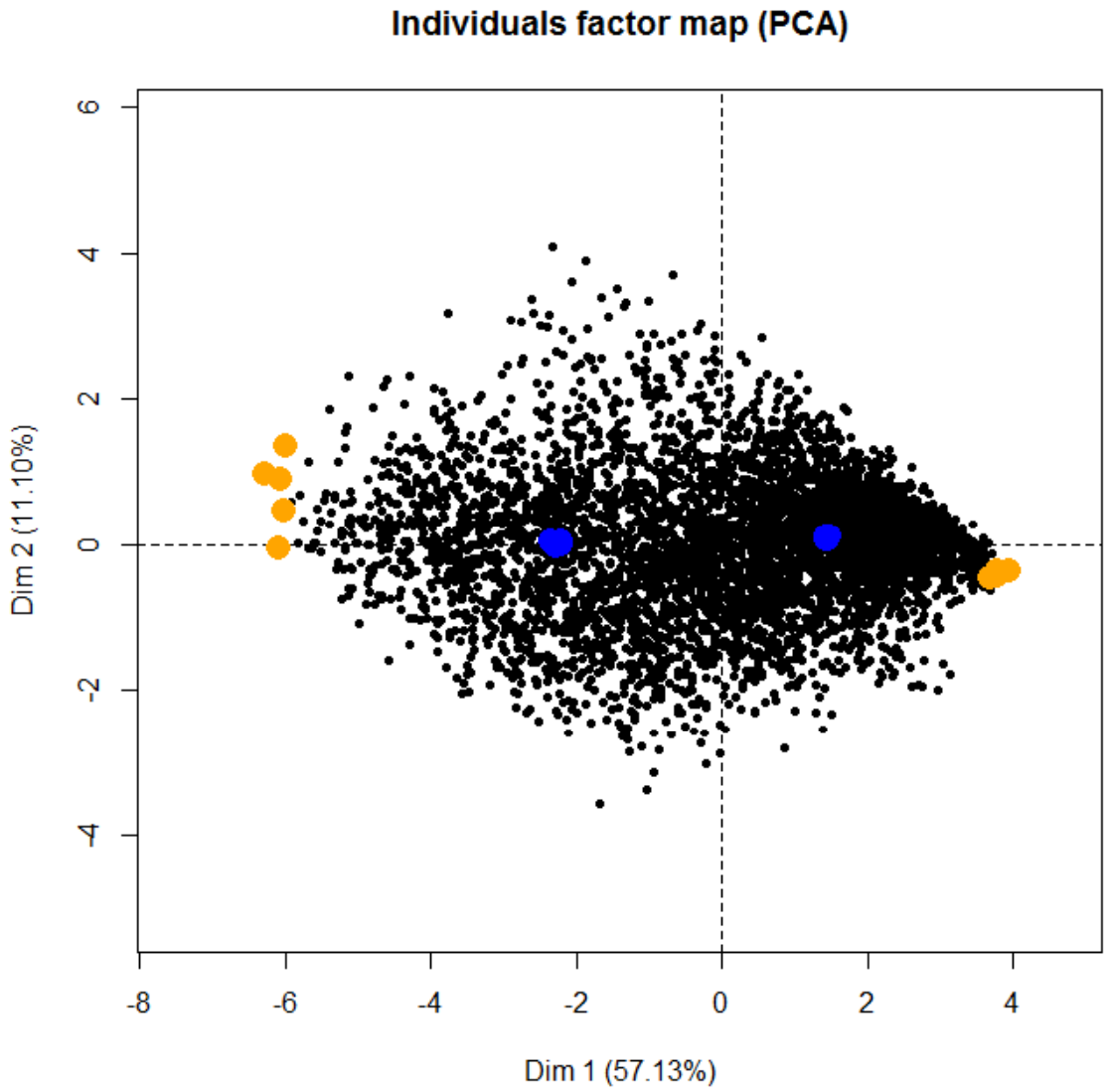
From the individuals

```
> res.hcpc$desc.ind$para
Cluster: 1
    I3045     I2801     I3404     I4561     I2477
0.0140178 0.0454952 0.0748305 0.1016992 0.1131562
--------------------------------------------------------------------------------

Cluster: 2
     I3451      I3063      I3158      I4466      I1166
0.02458791 0.04874171 0.04874171 0.07030856 0.07354237
```

```
> res.hcpc$desc.ind$dist
Cluster: 1
   I1780    I0392    I1413    I0545    I1781
7.783213 7.566054 7.559595 7.543709 7.482944
--------------------------------------------------------------------------------
Cluster: 2
   I0536    I0977    I4932    I3460    I3806
6.219741 6.058947 6.047333 5.978893 5.978893
```

Individuals “para” and “dist” on the principal plane



Variables values of the individuals “para” and “dist” (cluster 1)

```
> res.hcpc$data.clust[which(rownames(res.hcpc$data.clust)%in%names(res.hcpc$desc.ind$para[[1]])),]
         Sex          B1 Edad PF_Phisica RP_Role.li RE_Role.li SF_Social MH_Mental EV_Energy P_Pain HP_General Edad_Skol clust
I2477 female health_poor   33         30         75      66.67     44.44        36         25  44.44          25    OldLow     1
I2801 female health_fair   27         40          0       0.00     44.44        52         50  44.44          45 OldMedium     1
I3045 female health_fair   75         35         25      66.67     44.44        28         45  55.56          30 MidMedium     1
I3404 female health_poor   87         90          0      33.33     44.44        68         25  22.22          30 MidMedium     1
I4561   male health_poor   71         50         25      33.33     44.44        56         30  55.56          25 OldMedium     1


> res.hcpc$data.clust[which(rownames(res.hcpc$data.clust)%in%names(res.hcpc$desc.ind$dist[[1]])),]
         Sex          B1 Edad PF_Phisica RP_Role.li RE_Role.li SF_Social MH_Mental EV_Energy P_Pain HP_General Edad_Skol clust
I0392 female health_poor   49         30          0          0      0.00         4          5   0.00           0    MidLow     1
I0545 female health_poor   67          0          0          0      0.00        24          0   0.00           0    OldLow     1
I1413   male health_poor   61         40          0          0      0.00         0          0  11.11           0    MidLow     1
I1780 female health_poor   26         25          0          0      0.00         0          0   0.00           0    MidLow     1
I1781 female health_poor   52          5          0          0     11.11         8          0  11.11           0 OldMedium     1
```