Session
Multiple Correspondence 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*

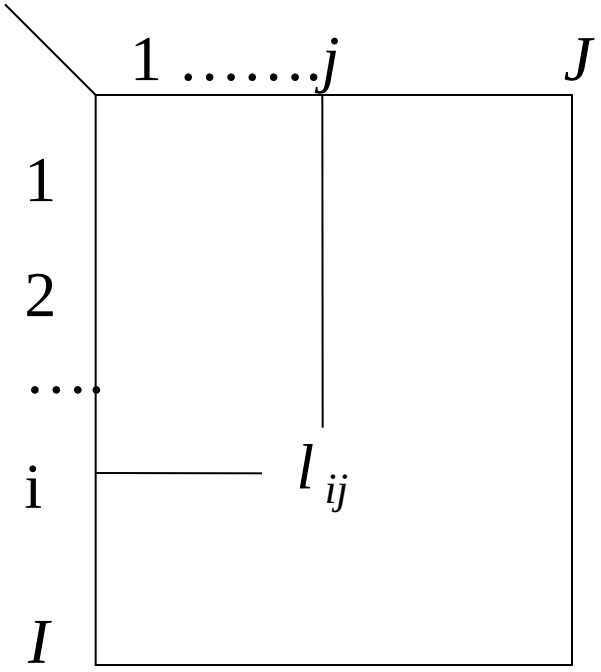
FIB UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Data: base

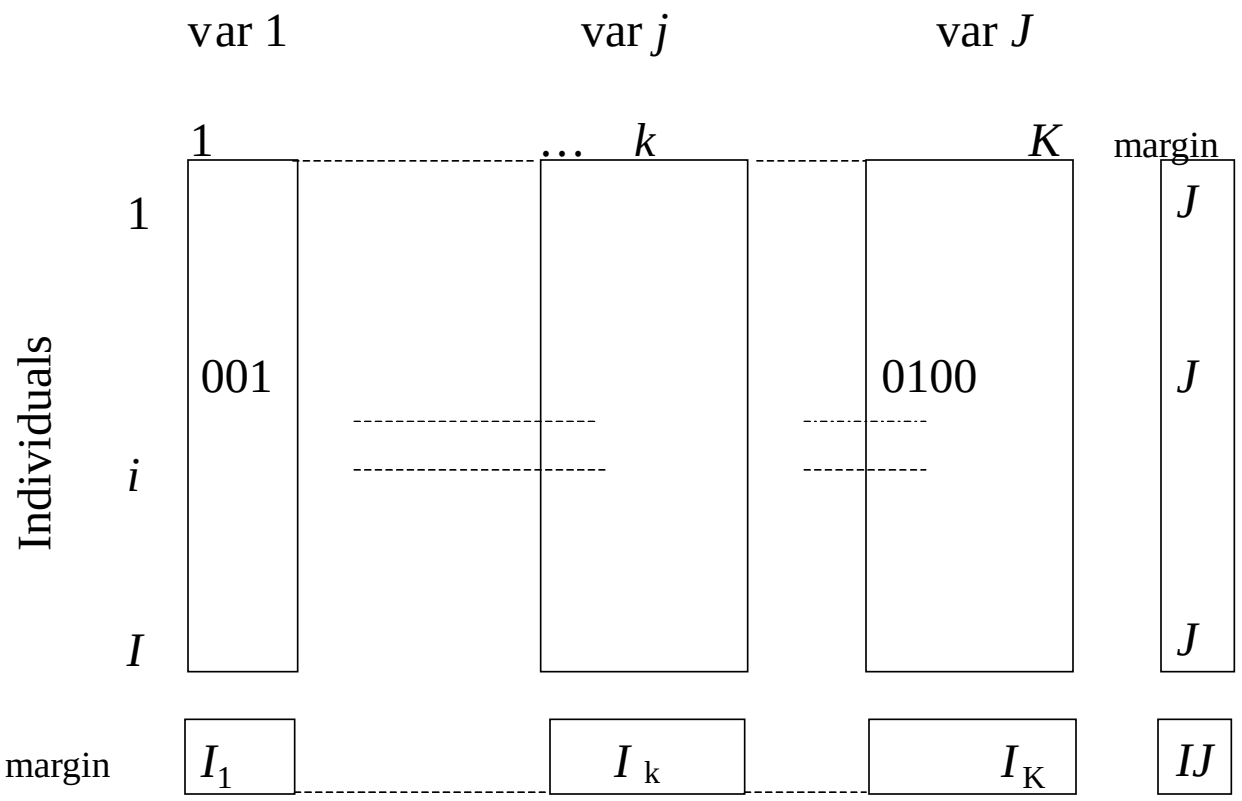
	row.names	tamaÃ±o	peso	velocidad	inteligencia	afectividad	agresividad	funciÃ³n	velo.cont.
1	pachón	Tam-pequeño	Pes-bajo	poco veloz	Inteligencia-baja	no-afectuoso	si-agresivo	caza	10
2	beauceron	Tam-grande	Pes-mediano	muy veloz	Inteligencia-media	si-afectuoso	si-agresivo	vigilancia	80
3	pastor alemán	Tam-grande	Pes-mediano	muy veloz	Inteligencia-alta	si-afectuoso	si-agresivo	vigilancia	90
4	boxer	Tam-mrediano	Pes-mediano	medio veloz	Inteligencia-media	si-afectuoso	si-agresivo	compañía	60
5	bull-dog	Tam-pequeño	Pes-bajo	poco veloz	Inteligencia-media	si-afectuoso	no-agresivo	compañía	20
6	bull-mastif	Tam-grande	pes-alto	poco veloz	Inteligencia-alta	no-afectuoso	si-agresivo	vigilancia	30
7	caniche	Tam-pequeño	Pes-bajo	medio veloz	Inteligencia-alta	si-afectuoso	no-agresivo	compañía	65
8	chihuahua	Tam-pequeño	Pes-bajo	poco veloz	Inteligencia-baja	si-afectuoso	no-agresivo	compañía	10
9	cocker	Tam-mrediano	Pes-bajo	poco veloz	Inteligencia-media	si-afectuoso	si-agresivo	compañía	15
10	colley	Tam-grande	Pes-mediano	muy veloz	Inteligencia-media	si-afectuoso	no-agresivo	compañía	85
11	dálmata	Tam-mrediano	Pes-mediano	medio veloz	Inteligencia-media	si-afectuoso	no-agresivo	compañía	55
12	doberman	Tam-grande	Pes-mediano	muy veloz	Inteligencia-alta	no-afectuoso	si-agresivo	vigilancia	90
13	alano alemán	Tam-grande	pes-alto	muy veloz	Inteligencia-baja	no-afectuoso	si-agresivo	vigilancia	90
14	podenco bretón	Tam-mrediano	Pes-mediano	medio veloz	Inteligencia-alta	si-afectuoso	no-agresivo	caza	55
15	podenco francés	Tam-grande	Pes-mediano	medio veloz	Inteligencia-media	no-afectuoso	no-agresivo	caza	60
16	fox-hound	Tam-grande	Pes-mediano	muy veloz	Inteligencia-baja	no-afectuoso	si-agresivo	caza	85
17	fox-terrier	Tam-pequeño	Pes-bajo	medio veloz	Inteligencia-media	si-afectuoso	si-agresivo	compañía	55
18	grand-bleu de Gascogne	Tam-grande	Pes-mediano	medio veloz	Inteligencia-baja	no-afectuoso	si-agresivo	caza	50
19	labrador	Tam-mrediano	Pes-mediano	medio veloz	Inteligencia-media	si-afectuoso	no-agresivo	caza	55
20	galgo	Tam-grande	Pes-mediano	muy veloz	Inteligencia-baja	no-afectuoso	no-agresivo	caza	100
21	mastiff	Tam-grande	pes-alto	poco veloz	Inteligencia-baja	no-afectuoso	si-agresivo	vigilancia	20
22	pekinés	Tam-pequeño	Pes-bajo	poco veloz	Inteligencia-baja	si-afectuoso	no-agresivo	compañía	15
23	pointer	Tam-grande	Pes-mediano	muy veloz	Inteligencia-alta	no-afectuoso	no-agresivo	caza	95
24	setter	Tam-grande	Pes-mediano	muy veloz	Inteligencia-media	no-afectuoso	no-agresivo	caza	90
25	San-Bernardo	Tam-grande	pes-alto	poco veloz	Inteligencia-media	no-afectuoso	si-agresivo	vigilancia	30
26	teckel	Tam-pequeño	Pes-bajo	poco veloz	Inteligencia-media	si-afectuoso	no-agresivo	compañía	20
27	terranova	Tam-grande	pes-alto	poco veloz	Inteligencia-media	no-afectuoso	no-agresivo	vigilancia	20

Data: Individuals × Categorical Variables

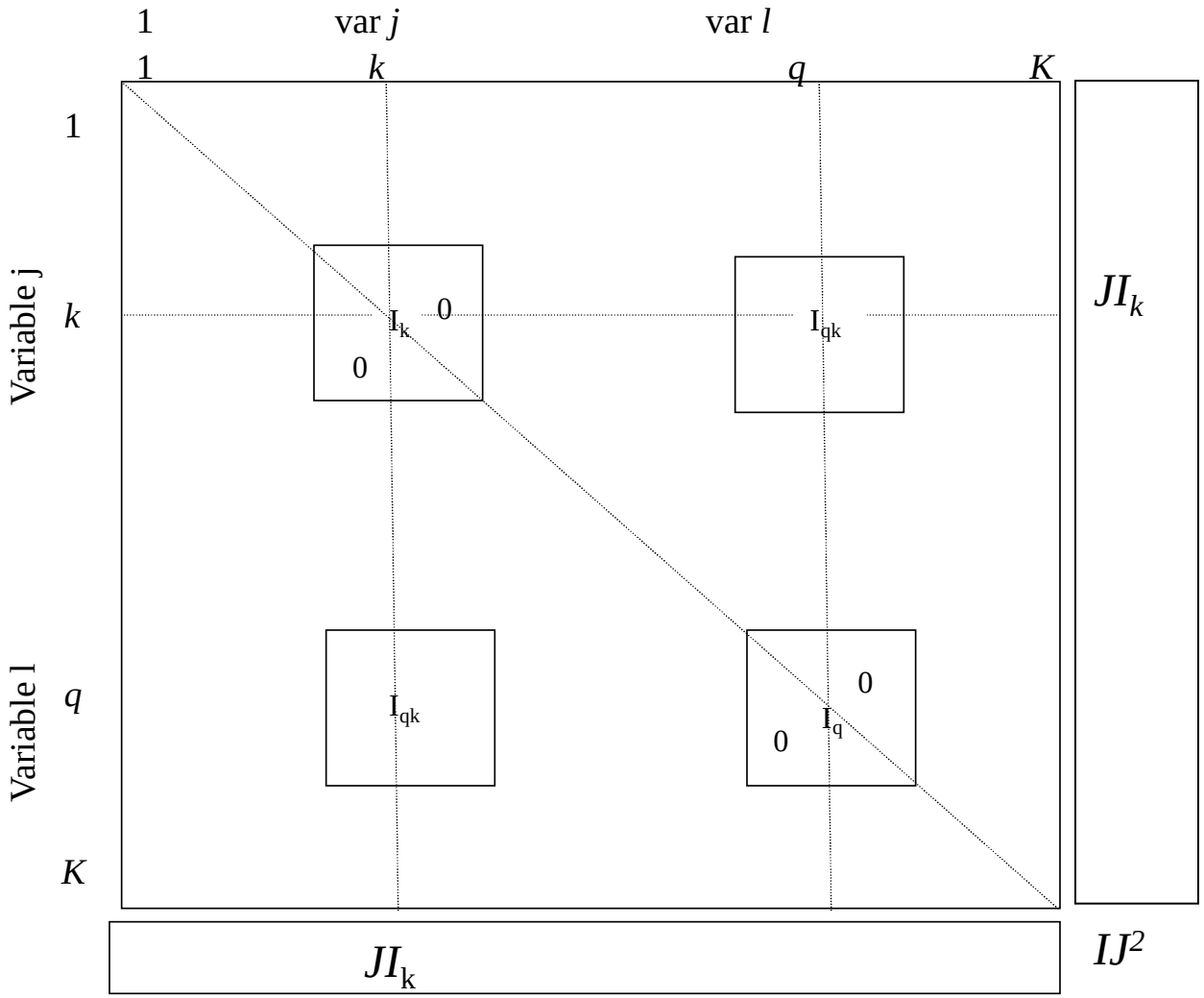
a. Condensed encoding



b. Complete disjunctive encoding: complete disjunctive table CDT



c. Burt table


1
var *j*
var *l*
1
k
q
K
1
Variable j
k
I_k
0
0
I_{qk}
JI_k
I_{qk}
Variable l
q
0
I_q
0
K
JI_k
IJ^2

Objectives

- Close to PCA objectives : similarity among individuals/ similarity among variables
- Generalization of CA objectives: relationships among the categories

3 types of objects:

- Individuals
- Variables
- Categories

Analysis of the individuals

- Typology of the individuals

  Close individuals: those who share many categories

Analysis of the variables

- Relationship among the variables through the relationships among the categories

- Summarizing the categorical variables through quantitative variables.

Analysis of the categories

Two points of view

Indicador variable= CDT column

Two categories are close if they are simultaneously present (absent) for large number of individuals

Category of individuals= row of Burt table

Two categories are close if they are associated to the same other categories

Synthesis of the objectives

3 types of objects. The classical problematic (typology of the rows, typology of the columns, relationships between both) is more complex

However, uniqueness of the table…

The analysis is mainly driven by the typology of the categories which gives account of:

- the relationships among the variables through the relationships among the categories :
- the average behavior of the individuals through the study of the categories that are “average individuals”

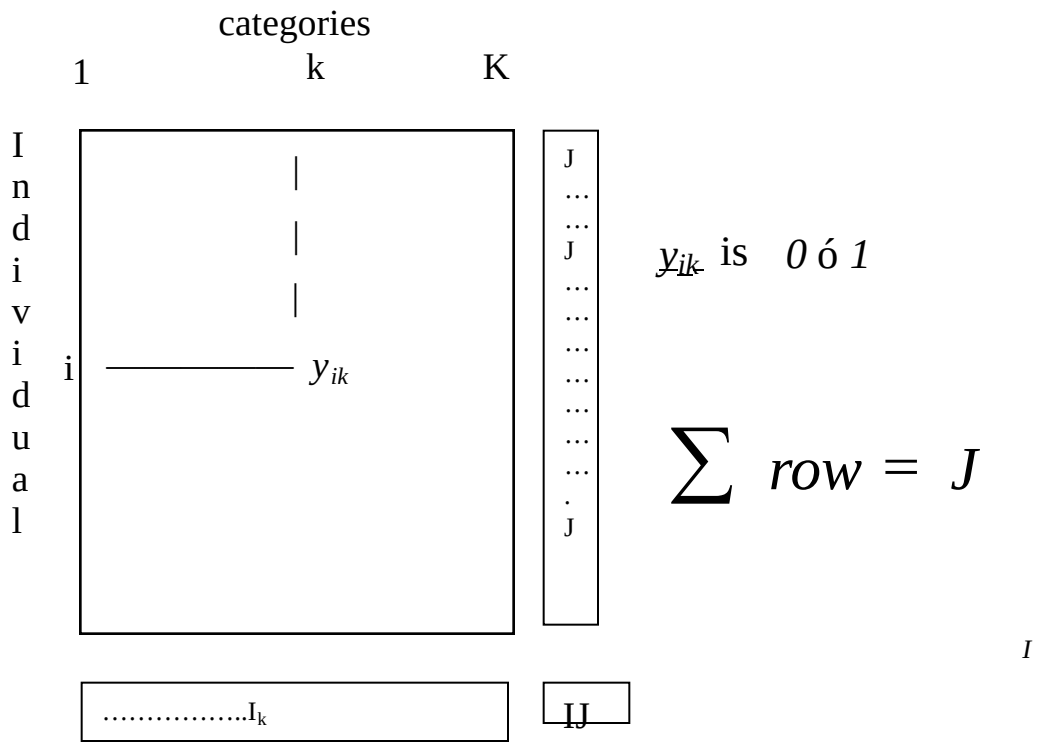
MCA= CA of the disjunctive table

CA, suitable method for analyzing a contingency or frequency table, cannot be applied as a method to analyze the CDT

MCA computing is totally equivalent to CA applied to CDT

however

Own interpretation rules have to be developed, which leads to an original method


categories
1
k
K
Individual
i
y_{ik}
J
…
…
J
…
…
…
…
…
…
…
…
.
J
y_{ik} is 0 ó 1
$\sum$ row = J
I
……..I_k
IJ

Cloud of individuals

$$d^2(i,l) = \sum_k \frac{IJ}{I_k}\left(\frac{y_{ik}}{J} - \frac{y_{lk}}{J}\right)^2 = \frac{1}{J}\sum_k \frac{I}{I_k}(y_{ik} - y_{lk})^2$$

0 or 1. 1 if only one individual presents category *k*

<u>Distance</u>: the distance increases depending on the numer of categories whch differ from an individual to another. Category *k* intervenes with weight n/I_k, inverse of its frequency. The presence of a rare category moves individuals that present it away of the other individuals.

For these properties, the distance induced y CA applied to the CDT is satisfactory.

<u>Weights</u>: uniform weights for all individuals

Cloud of categories

$$d^2(k,h) = \sum_i I\left(\frac{y_{ik}}{I_k} - \frac{y_{ih}}{I_h}\right)^2$$

$$= \frac{I}{I_k \cdot I_h} \cdot \left[I - (I_{(1,kh} + I_{(0,kh)}\right]$$

Distance: the distance increases depending on the number of individuals presenting only one of both categories. The distance decreases depending on the size of each category. Category *k* intervenes with weight I/I_k, inverse of its frequency. Two categories of a same variable tend to be far; Two categories chosen by the same individuals lie at the same position; **the rare categories lie far away of the others**.

Weights: weight of category *k*: $\frac{I_k}{IJ}$

Distance from a category to the centroid

Inertia de *k* relatively to G_K

$$d^2(k, G_K) = \frac{I}{I_k} - 1$$

$$\frac{1}{J} \cdot \left(1 - \frac{I_k}{I}\right)$$

For a rare category, the weak weight is not enough to compensate the distance from the centroid

Total inertia of the cloud

$$\left(\frac{K}{J}\right) - 1$$

Inertia of the K_j categories of variable *j*

$$\frac{K_j - 1}{J}$$

Maximum number if eigenvalues

Min (*I*-1, *K* – *J)*

Transition relationships

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k \frac{y_{ik}}{J} G_s(k)$$

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{y_{ik}}{I_k} \cdot F_s(i)$$

The variables through their categories

- Centroid of the categories of one variable

lies on the global centroid)

- The proportion of inertia associated to each factor is weak if the variables have many categories
- Even if a variable is highly linked to a factor, all its categories cannot be well represented on this factor
- Although there are many individuals, it is not useful to divide a variable in too many categories

Synthesis of the categorical variables

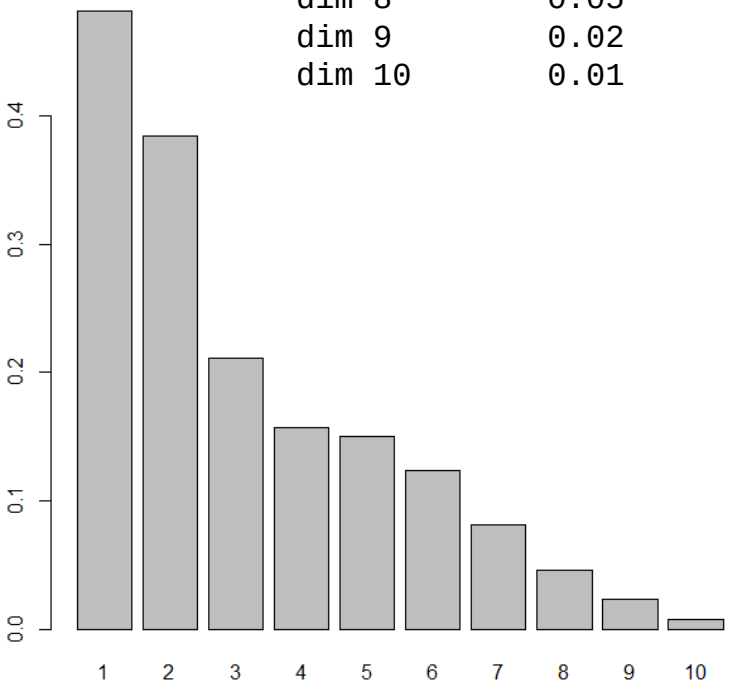
The “factors on the individuals” are a synthesis of the categorical variables that are quantitative variables.

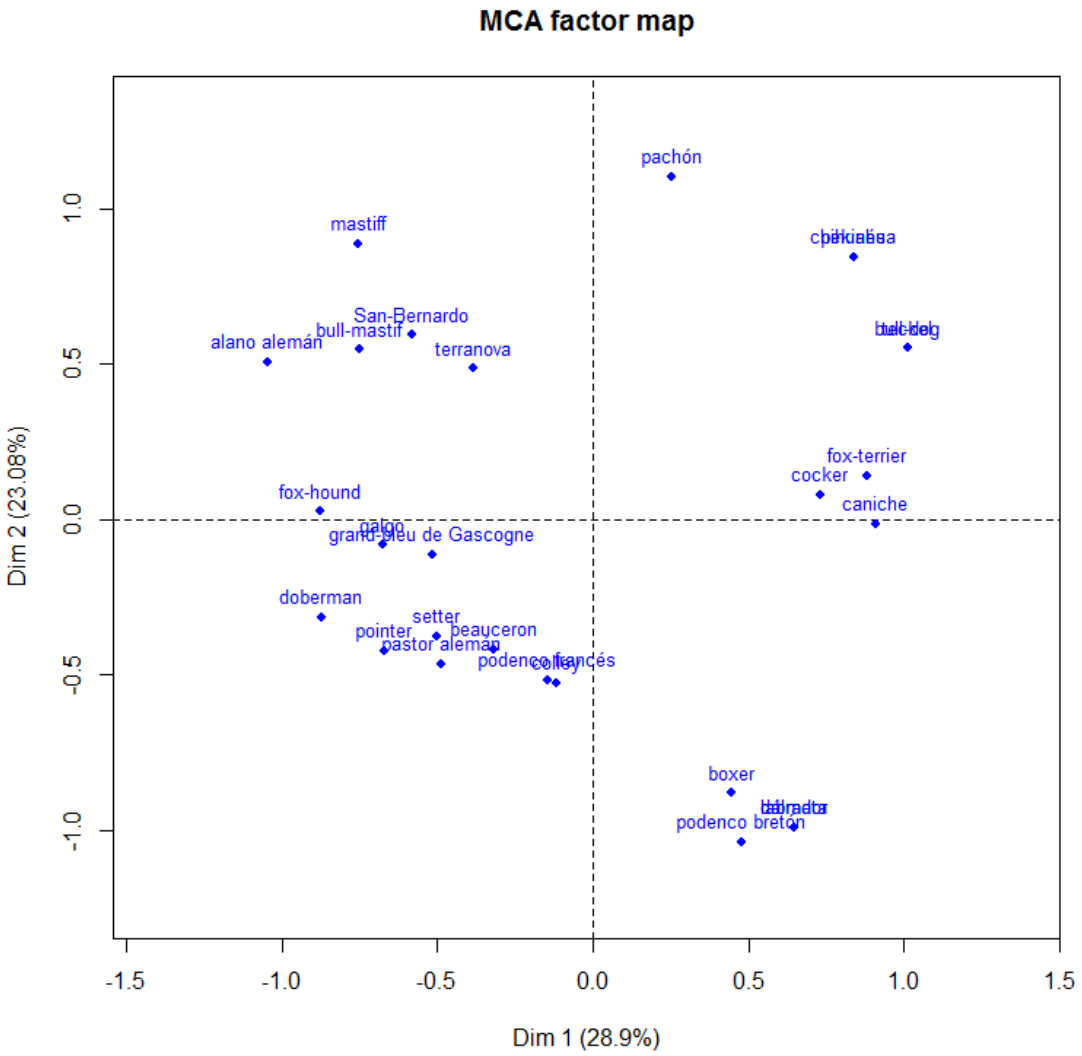
To measure the association between the qualitative variables and the factors, the correlation ratio can be used:

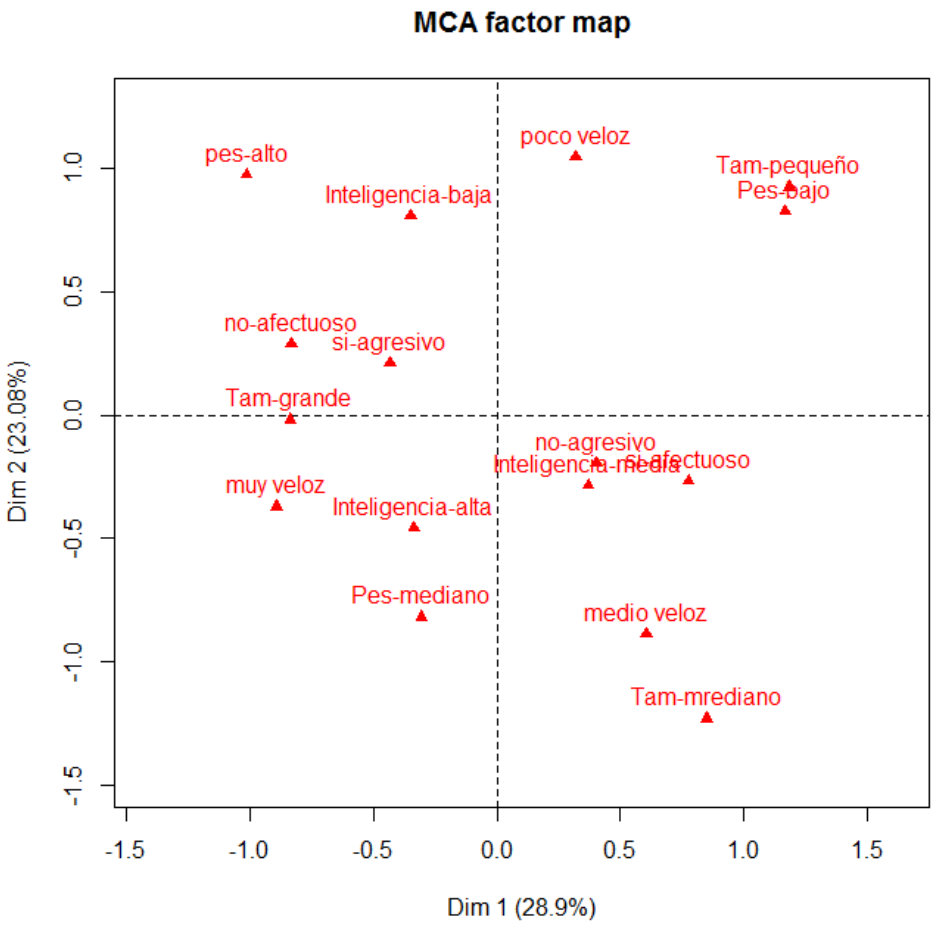
$$\eta^2(F_s, j) = \frac{\sum_{k \in K_j} \left(I_k / I \right) \left(F_s(centroid_k) \right)^2}{\lambda_s}$$

with $centroid_k$= centroid of the individuals presenting category *k*

```
> round(res.mca$eig[1:10,],2)
       eigenvalue percentage of variance cumulative percentage of variance
dim 1        0.48                 28.90                               28.90
dim 2        0.38                 23.08                               51.98
dim 3        0.21                 12.66                               64.64
dim 4        0.16                  9.45                               74.09
dim 5        0.15                  9.01                               83.10
dim 6        0.12                  7.40                               90.50
dim 7        0.08                  4.89                               95.38
dim 8        0.05                  2.74                               98.12
dim 9        0.02                  1.41                               99.54
dim 10       0.01                  0.46                              100.00
```




MCA factor map
pachón
mastiff
San-Bernardo
bull-mastif
alano alemán
terranova
fox-terrier
cocker
caniche
fox-hound
grand bleu de Gascogne
doberman
setter
pointer
beauceron
pastor aleman
boxer
podenco breton
Dim 2 (23.08%)
Dim 1 (28.9%)
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

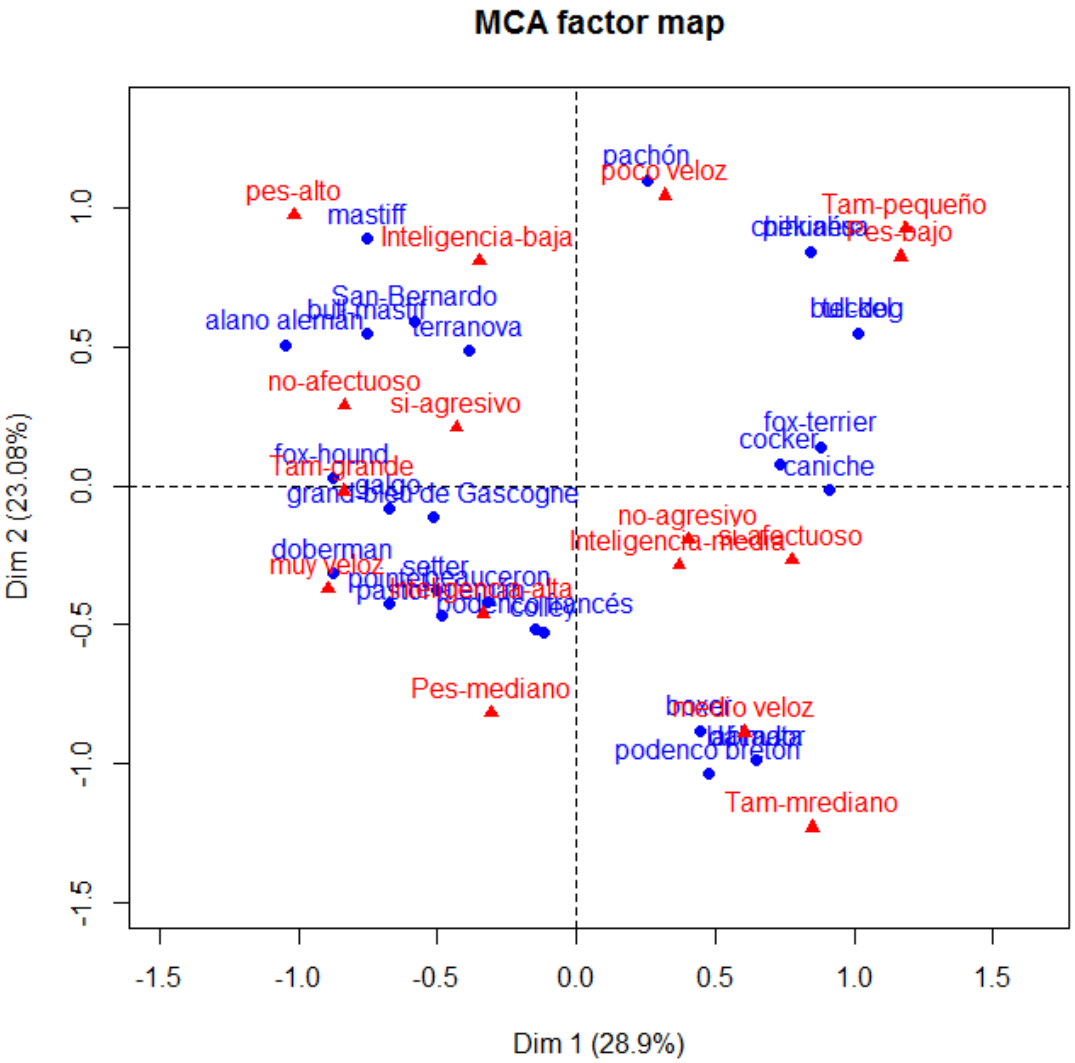

MCA factor map
pes-alto
Inteligencia-baja
poco veloz
Tam-pequeño
Pes-bajo
no-afectuoso
si-agresivo
Tam-grande
no-agresivo
Inteligencia-media
muy veloz
Inteligencia-alta
Pes-mediano
medio veloz
Tam-mrediano
Dim 2 (23.08%)
Dim 1 (28.9%)
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

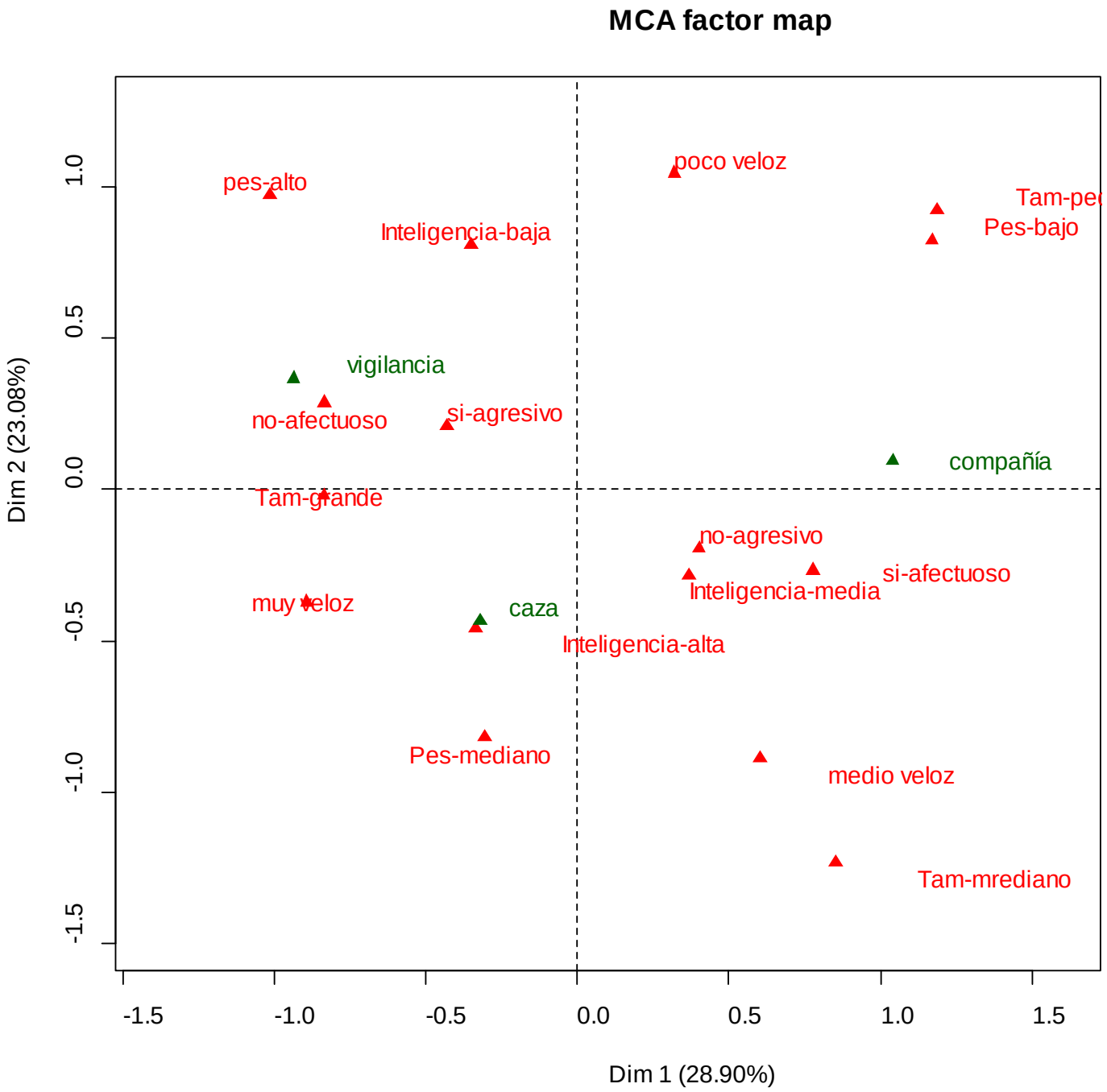
Transition relationships

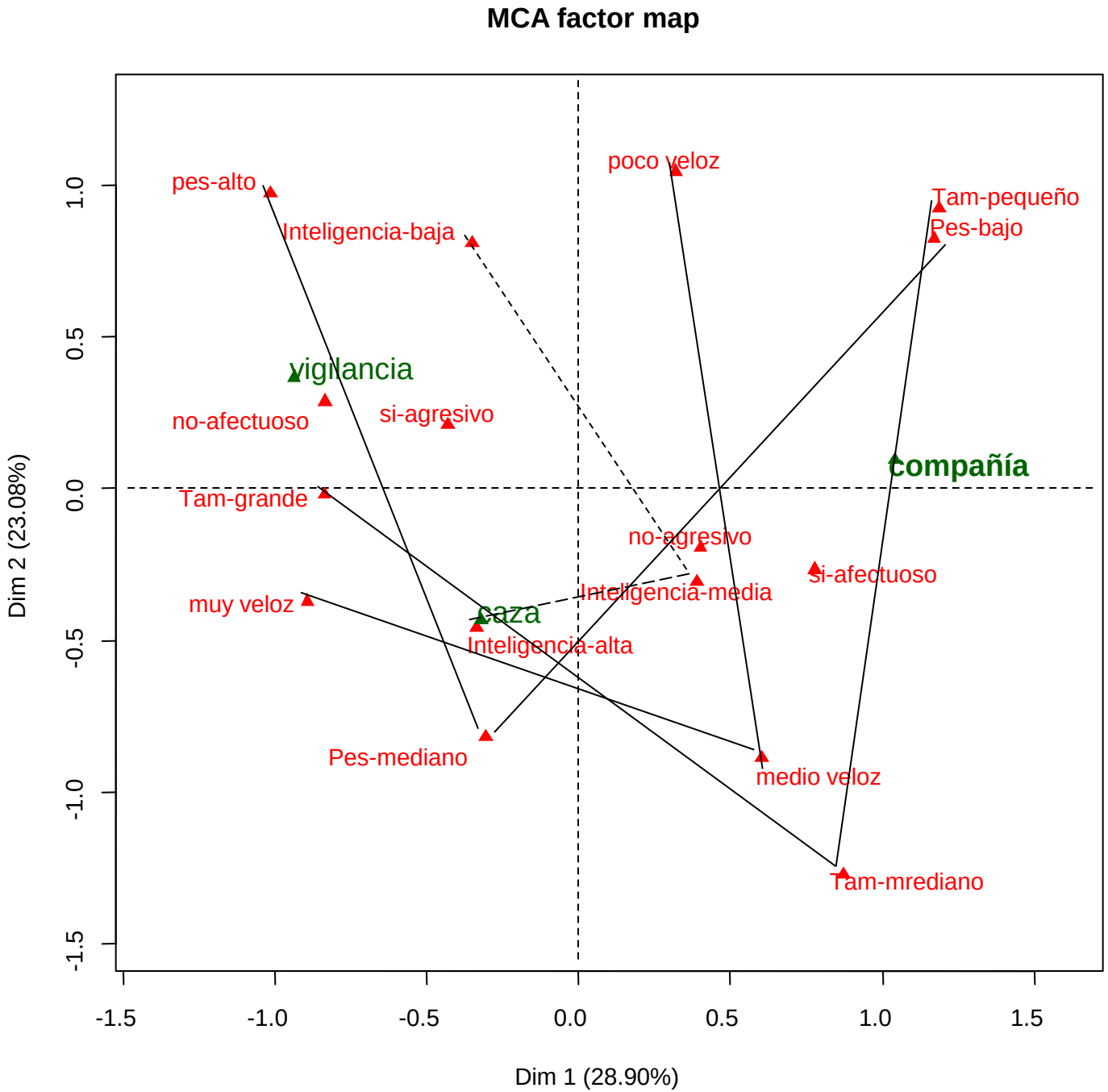
$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k \frac{y_{ik}}{J} G_s(k)$$

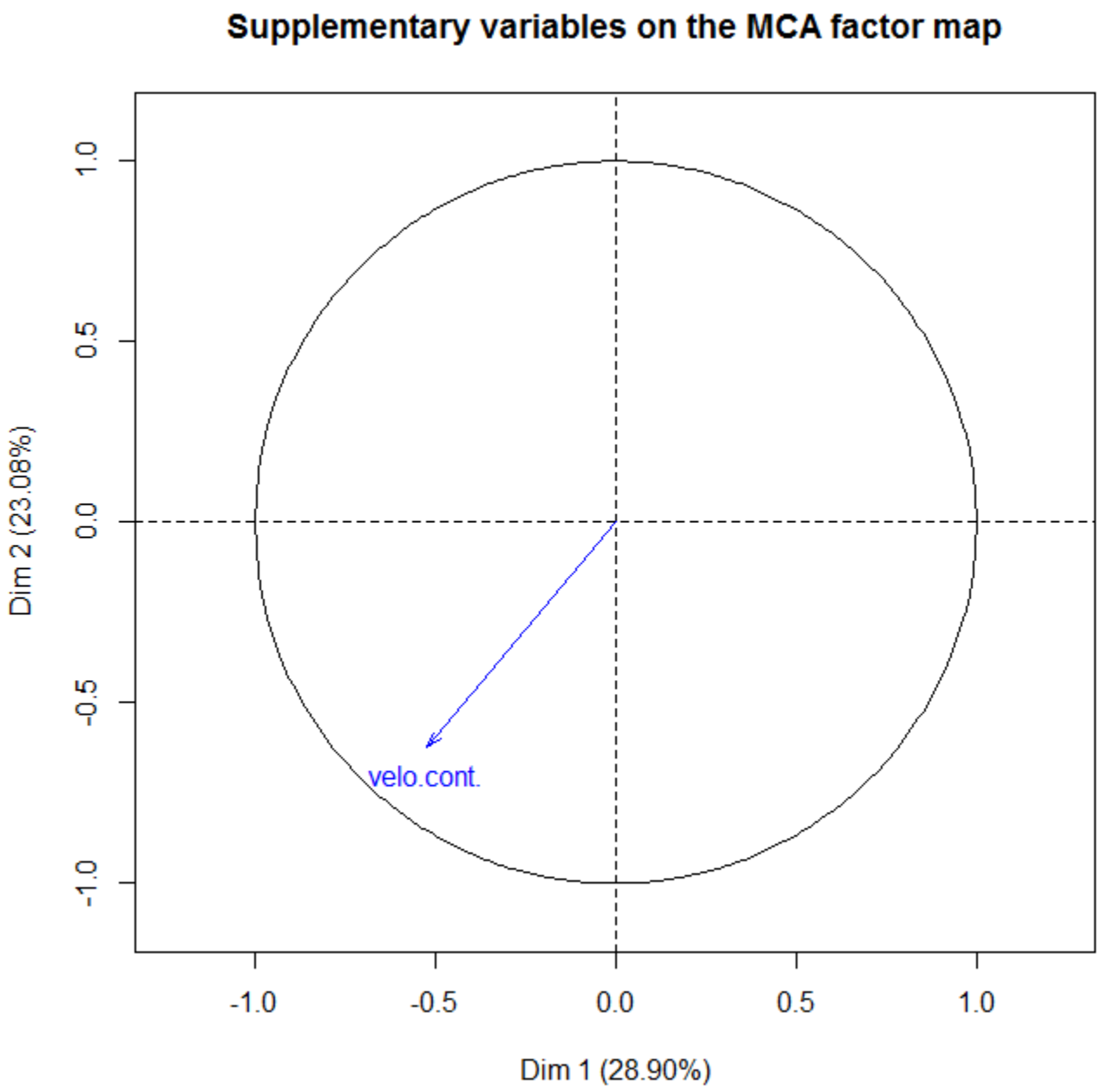
$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{y_{ik}}{I_k} \cdot F_s(i)$$

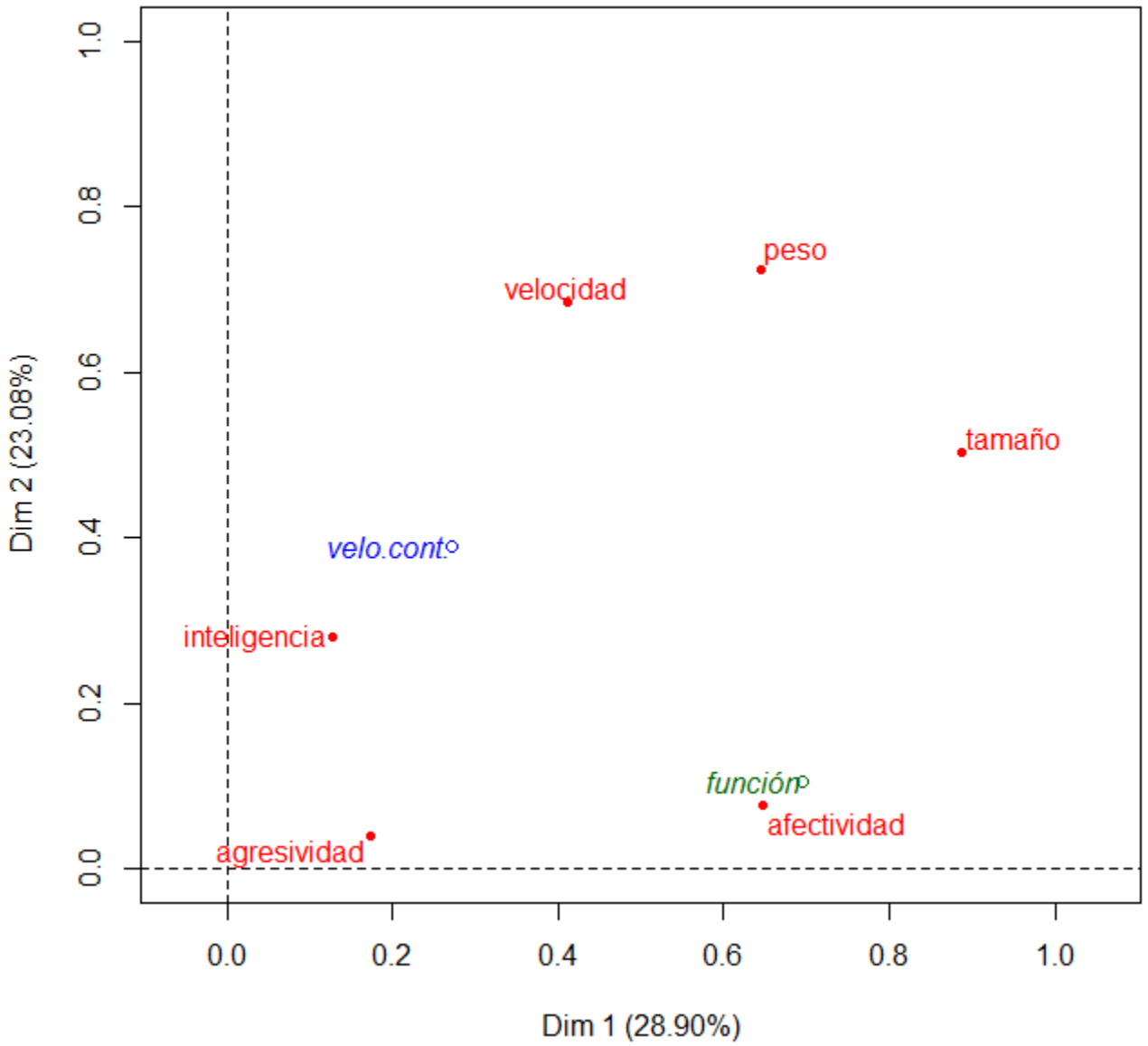
Superimposed representation

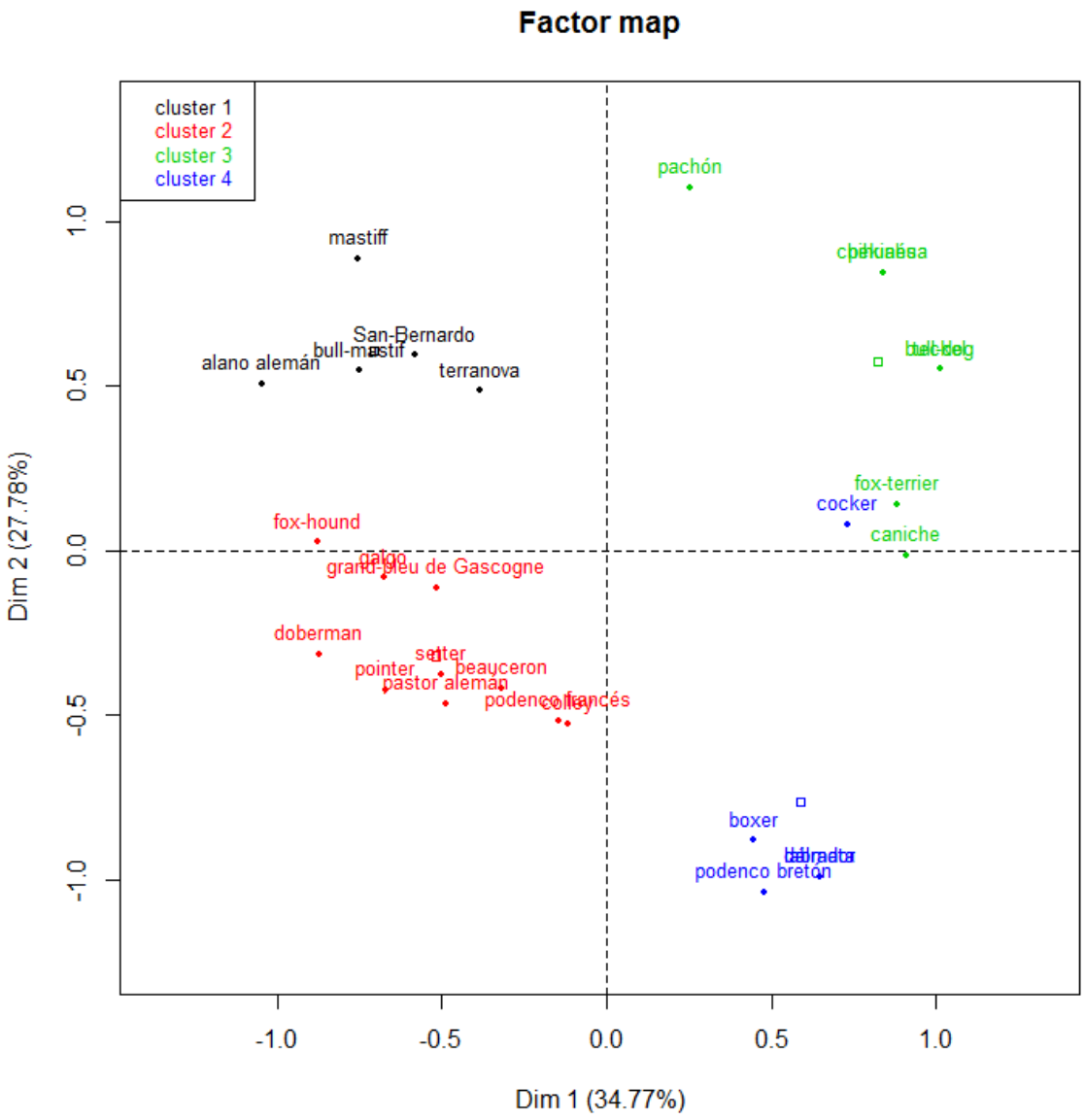



MCA factor map
pes-alto
poco veloz
Tam-pe
Pes-bajo
Inteligencia-baja
vigilancia
no-afectuoso
si-agresivo
compañía
Tam-grande
no-agresivo
si-afectuoso
Inteligencia-media
muy veloz
caza
Inteligencia-alta
Pes-mediano
medio veloz
Tam-mrediano
Dim 2 (23.08%)
Dim 1 (28.90%)
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


MCA factor map
pes-alto
Inteligencia-baja
poco veloz
Tam-pequeño
Pes-bajo
vigilancia
no-afectuoso
si-agresivo
compañía
Tam-grande
no-agresivo
si-afectuoso
Inteligencia-media
muy veloz
caza
Inteligencia-alta
Pes-mediano
medio veloz
Tam-mrediano
Dim 2 (23.08%)
Dim 1 (28.90%)
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


Supplementary variables on the MCA factor map
Dim 2 (23.08%)
1.0
0.5
0.0
-0.5
-1.0
velo.cont.
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (28.90%)


peso
velocidad
tamaño
velo.cont
inteligencia
función
afectividad
agresividad
Dim 2 (23.08%)
Dim 1 (28.90%)
0.0
0.2
0.4
0.6
0.8
1.0


Factor map
cluster 1
cluster 2
cluster 3
cluster 4
pachón
mastiff
San-Bernardo
bull-mastif
alano alemán
terranova
fox-terrier
cocker
caniche
fox-hound
galgo
grand-bleu de Gascogne
doberman
setter
pointer
beauceron
pastor alemán
boxer
podenco bretón
Dim 1 (34.77%)
Dim 2 (27.78%)
-1.0
-0.5
0.0
0.5
1.0

with the Croatian survey

First: a short example

with only 4 categorical active variable; these variables are binary

B5_1 HealthPerception because of your physical health you cut down on the amount of time you spent on work

B5_2 HealthPerception because of your physical health accomplished less than you would like

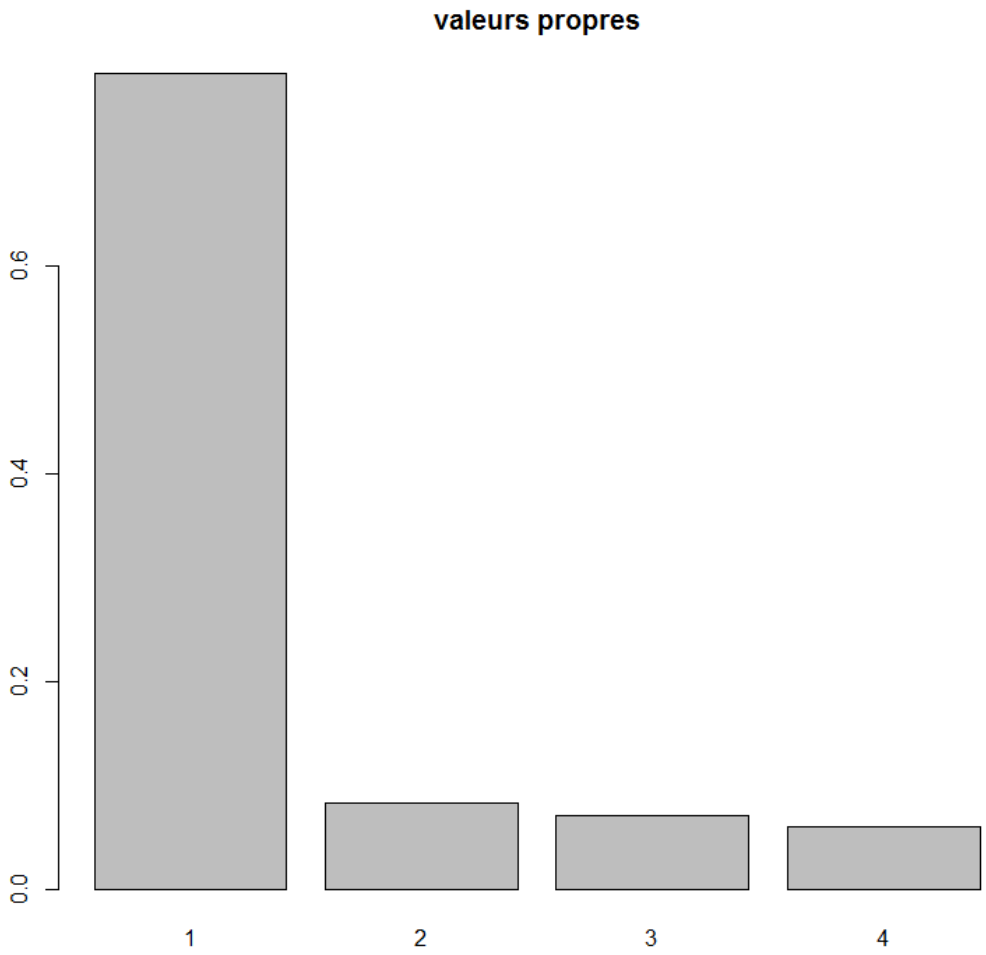
B5_3 HealthPerception because of your physical health were limited in a kind of work

B5_4 HealthPerception because of your physical health had difficulty in performing the work

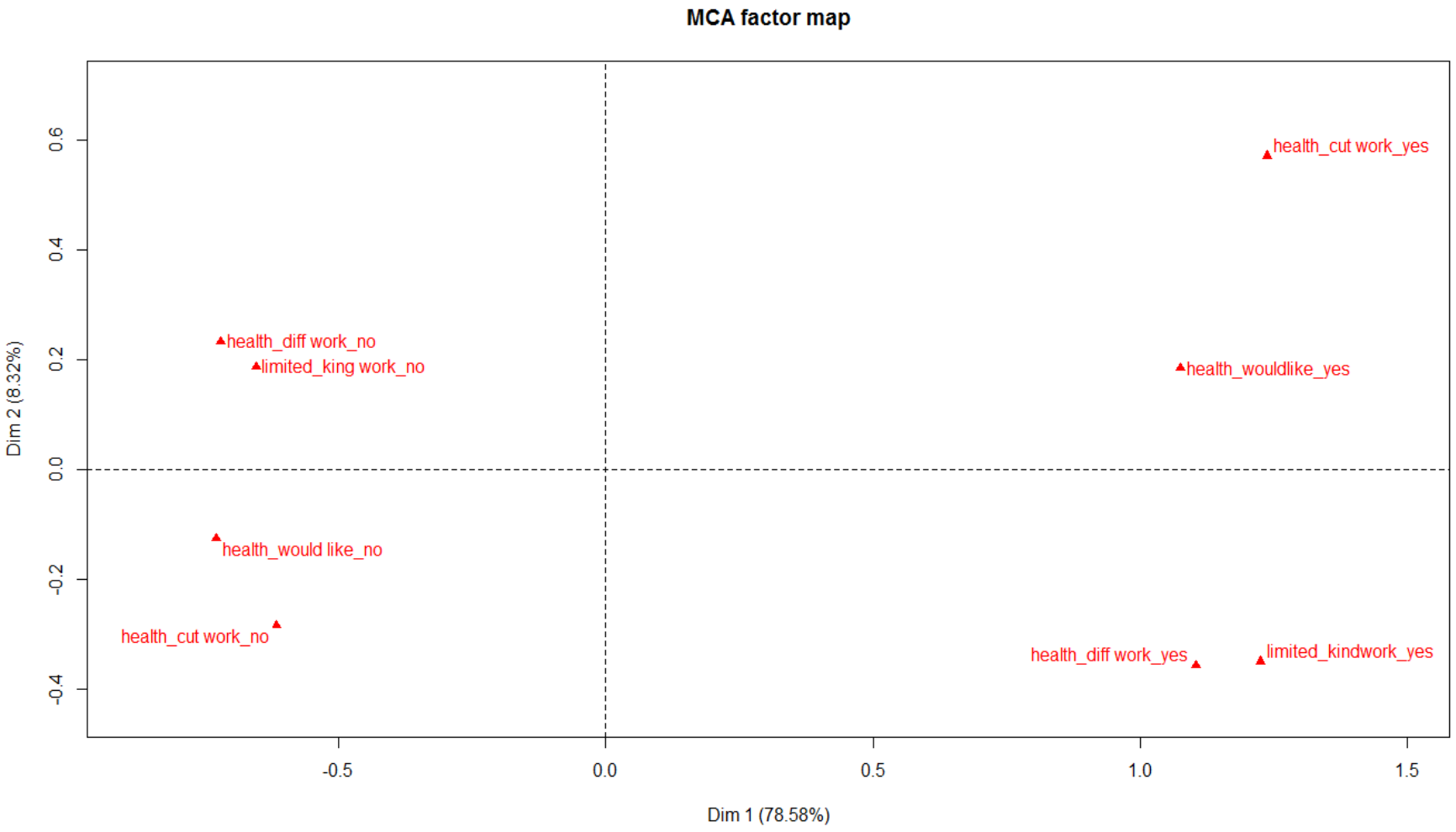
```
> summary(base [,c(18:21)])
                    B51                                B52
 health_cut work_no :3363   health_would like_no:3001
 health_cut work_yes:1674   health_wouldlike_yes:2036
                    B53                                B54
 limited_kindwork_yes:1754   health_diff work_no :3047
 limited_king work_no:3283   health_diff work_yes:1990
```

```
res.mca<-  MCA(base [,c(18:21)],level.ventil = 2)
summary (res.mca)
```

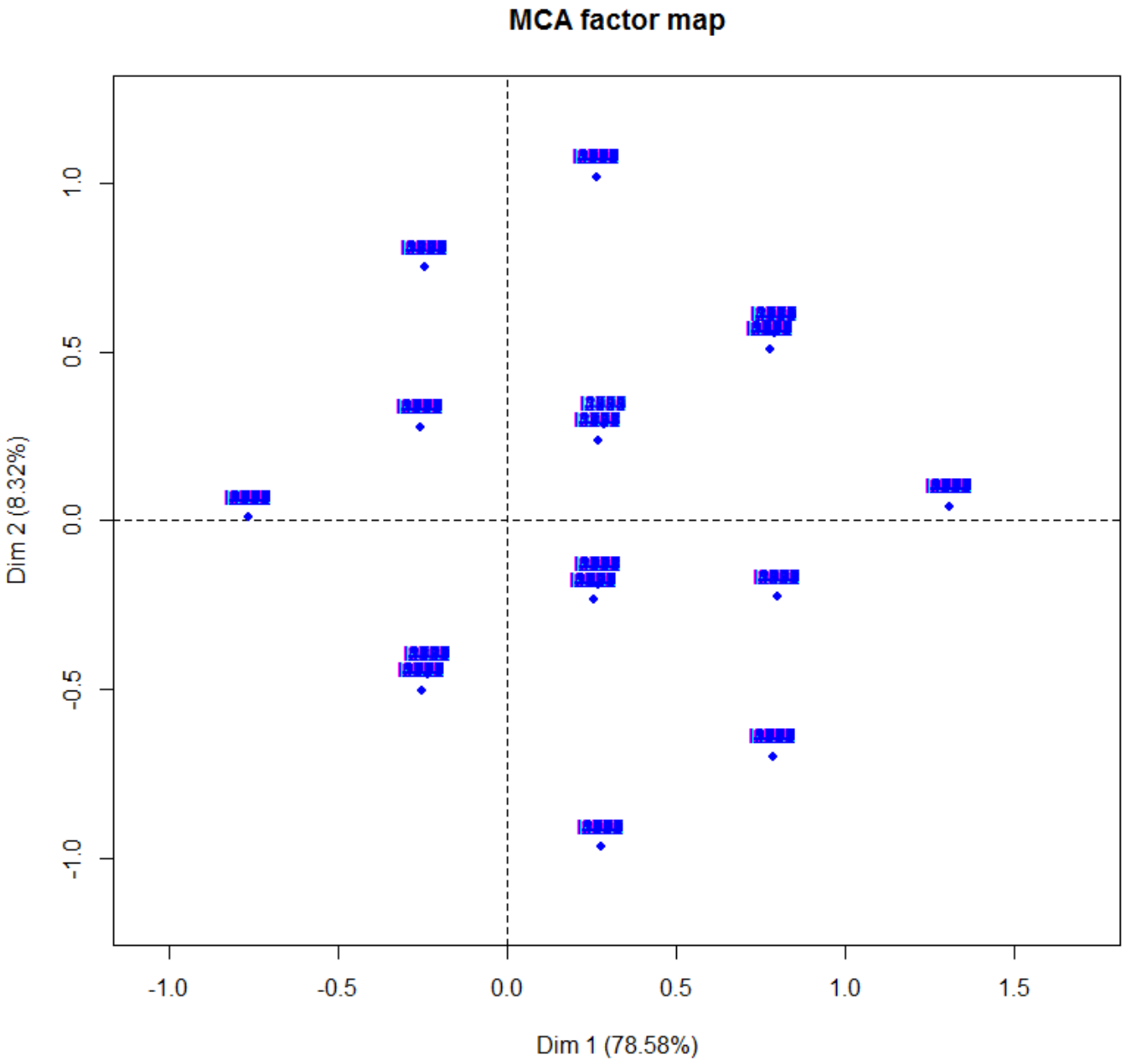
	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.79	78.58	78.58
dim 2	0.08	8.32	86.90
dim 3	0.07	7.07	93.97
dim 4	0.06	6.03	100.00



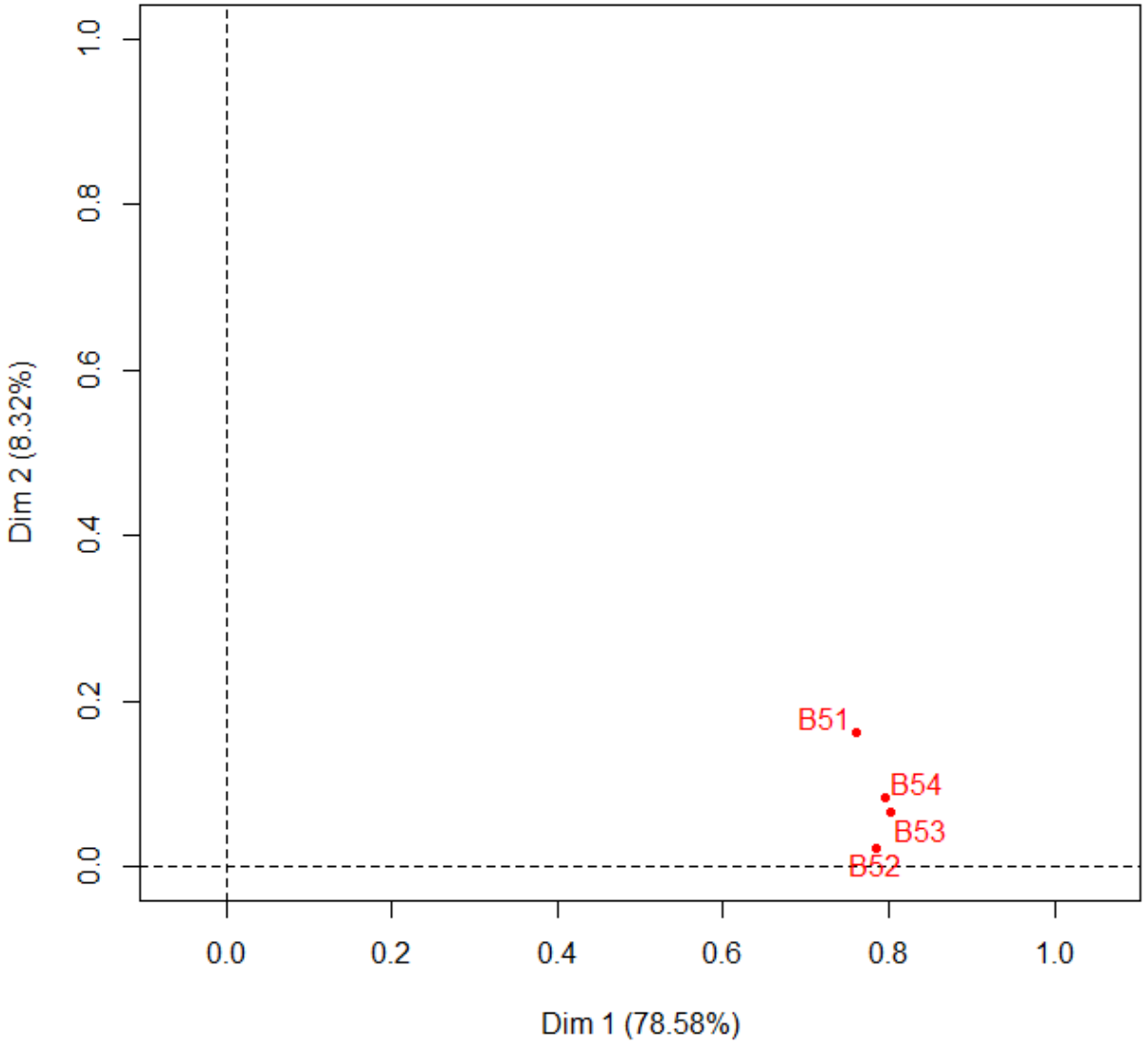
Representación de las categorías



Representación de los individuos



Representación de las variables (cuadrado de correlaciones)



```
> summary(res.mca)
(…)
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4
Variance               0.786   0.083   0.071   0.060
% of var.             78.577   8.321   7.074   6.029
Cumulative % of var.  78.577  86.898  93.971 100.000

Individuals (the 10 first)
                        Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
I0001                |  1.309  0.043  0.995 |  0.042  0.000  0.001 |  0.070  0.001  0.003 |
I0002                | -0.767  0.015  0.996 |  0.009  0.000  0.000 |  0.037  0.000  0.002 |
I0003                | -0.767  0.015  0.996 |  0.009  0.000  0.000 |  0.037  0.000  0.002 |
(…)

Categories
                         Dim.1     ctr    cos2  v.test     Dim.2     ctr    cos2  v.test
health_cut work_no   |  -0.616   8.051   0.761 -61.921 |  -0.284  16.164   0.162 -28.552 |
health_cut work_yes  |   1.237  16.173   0.761  61.921 |   0.570  32.472   0.162  28.552
health_would like_no |  -0.729  10.085   0.784 -62.844 |  -0.125   2.790   0.023 -10.756 |
health_wouldlike_yes |   1.075  14.865   0.784  62.844 |   0.184   4.112   0.023  10.756 |
(…)

Categorical variables (eta2)
                        Dim.1 Dim.2 Dim.3
B51                  | 0.761 0.162 0.077 |
B52                  | 0.784 0.023 0.193 |
B53                  | 0.801 0.065 0.007 |
B54                  | 0.796 0.083 0.007 |
```