Multiple Correspondence Analysis

Karina Gibert[1,2] & Artur Palomino[1,2,3]

[1]Department of Statistics and Operation Research

[2] Knowledge Engineering and Machine Learning group-at-Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya, Barcelona

[3]Kantar

Universitat Politècnica de Catalunya

Factorial Methods

Data	Factorial Method
Continuous variables	PCA - Principal Component Analysis
Count variables	CA - (Simple) Correspondence Analysis
Categorical variables	MCA - Multiple Correspondence Analysis

MCA

- Objective

1. Close to PCA objectives: similarity among indiv/ similarity among variables
2. Generalization of CA objectives: relationships among the categories

- Types of objects

1. Individuals
2. Variables
3. Categories

MCA

- Multiple Correspondence Analysis
  - Only qualitative variables
  - Find the most informative projection planes
    (factorial planes, maximize projected inertia)

- Given <X,M,D>
  - A data matrix X (nxp) the logic table
  - A matrix of individuals weights D (nxn)
  - Assume chi2 metrics to compare individuals (M)

MCA

- Given <X,M,D>

Same algorithm as PCA

  - Matrix $M^{1/2} X'DXM^{1/2}$
  - – Catches relationships and oppositions of data

Diagonalize matrix $M^{1/2} X'DXM^{1/2}$

  - Get r eigen values λ_α and sort decreasingly

$$\{\lambda_1\}_{\alpha=1:r} \qquad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r$$

  - – Corresponding eigenvectors $u_\alpha = (u_{\alpha 1} \ldots u_{\alpha p})$

$|u_\alpha|=1$

$u_\alpha u_{\alpha'} = 0$

$\{u_\alpha\}_{\alpha=1:r}$ orthonormal base for individuals

MCA

- Given <X,M,D>

  Diagonalize matrix $M^{1/2} X'DXM^{½}$

  - The subspace generated by $\{u_{\alpha}\}_{\alpha=1:r}$ is the same as the subspace generated by the rows of X

  $u^*_{\alpha} = M^{-½} u_{\alpha}$ are the principal factors of X :

  - good rotation directions

  $U^*=([u^*_1]\ [u^*_2] \ldots\ [u^*_r])$ is the basis for the projection space

MCA

- Two possibilities:
  - Analysis of Logic Table (X logic table)
  - Analysis of Burt Table (X Burt table)

MCA

Analysis of Logic Table

- Given <X,M,D>
- *X =Z*

  Z: logic matrix (complete disjunctive form)

- $M^{1/2} X'DXM^{1/2}$ *is sparse (high useless dimensions)*
- *Centroids of modalities ARE NOT CDG of their individuals (systematic bias of 1/sqrt(λ))*
- *Traditional dual analysis*

Systematic bias

```
Mca0<- MCA(X, method="Indicator")    ;  FactoMineR
```

MCA

- Original data set (only qualitative variables)

χ_{nxQ} =

n	χ_1 factor A	χ_2 factor B	χ_3 factor C	χ_4 factor D	χ_5 factor E
1	A2	B4	C3	D3	E4
2	A2	B4	C3	D2	E4
3	A3	B2	C2	D3	E4
4	A3	B9	C2	D2	E2
5	A5	B1	C2	D2	E3
6	A2	B4	C3	D4	E2
7	A1	B3	C2	D3	E2
8	A2	B5	C3	D4	E1
9	A2	B4	C2	D3	E4
1	A2	B4	C1	D2	E4
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
2000	A5	B2	C2	D2	E3

n=2000
Q=5

MCA

Logic Table - Complete Disjunctive Form CDF

$Z = [Z_1, Z_2, \ldots, Z_Q]$, $Z_q, \forall q \in [1:Q]$, qualitative var

J_q = number of categories for Z_q

Each χ_q coded into the binary variables Z_1, Z_q

	χ_1
	...tor A
	A2
	A2
3	A3
4	A3
5	A5
6	A2
7	A1
8	A2
9	A2
1	A2
…	…
…	…
…	…
…	…
2000	A5

$$Z_{n \mathrm{x} \sum_{q=1}^{Q} J_q} =$$

Z1				
A1	A2	A3	A4	A5
0	1	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	1	0	0
0	0	0	0	1
0	1	0	0	0
1	0	0	0	0
0	1	0	0	0
0	1	0	0	0
0	1	0	0	0
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

…

Z_1

n=2000

J_1 *=5*

10

n=2000

Q=8

MCA

Logic Table - Complete Disjunctive Form CDF

$Z = [Z_1 , Z_2, . . . , Z_Q]$, $Z_q, \forall q \in [1:Q]$, qualitative var

Each χ_q coded into the binary variables Z_1, Z_q

J_q = number of categories for Z_q

$$Z_{n\mathrm{x}\sum_{q=1}^{Q} J_q} =$$

	Z1				Z2						Z3		
	A2	A3	A4	A5	B1	B2	B3	B4	B5	B9	C1	C2	C3
0	1	0	0	0	0	0	0	1	0	0	0	0	1
0	1	0	0	0	0	0	0	1	0	0	0	0	1
0	0	1	0	0	0	1	0	0	0	0	0	1	0
0	0	1	0	0	0	0	0	0	0	1	0	1	0
0	0	0	0	1	1	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	1	0	0	0	0	1
1	0	0	0	0	0	0	1	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	1	0	0	0	1
0	1	0	0	0	0	0	0	1	0	0	0	1	0
0	1	0	0	0	0	0	0	1	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...

...

Z_1
n=2000
J_1 *=5*

Z_2
n=2000
J_2 *=6*

Z_3
n=2000
J_3 *=3*

n=2000
Q=8

MCA

- Methodology

1. Project centroids of all qualitative variables
2. Use one color for each variable
3. Add legend to the map
4. Link centroids of same variable with lines for ordinal variables

5. Evenctually project individuals

MCA

- Analysis of the categories

1. Two categories close if
   - Tend to be simultaneously present (absent) for large number of individuals
   - We say that categories are associated

MCA

- Analysis relationships on individuals

1. Typology of the individuals
2. Close individuals: those who share many categories

- Analysis of interactions among variables

1. Relationship among the variables through the relationships among their categories

MCA

- Process to interpret a factorial map
- Forget about variables bad represented in the factorial plan (X≈0, Y≈0)
- Which are the variables with relevant direct contribution to Factor in Axis X ? •
- Which are the variables with relevant inverse contribution to Factor in Axis X
- Supplementary variables: later introduce info on qualitative variables as well
- Analyze profiles opposed in two extremes of Axis X
- Induce a label for the Factor that represents the concept
  • Repeat with Factor in Axis Y

MCA
EXAMPLE

- Dog breeds

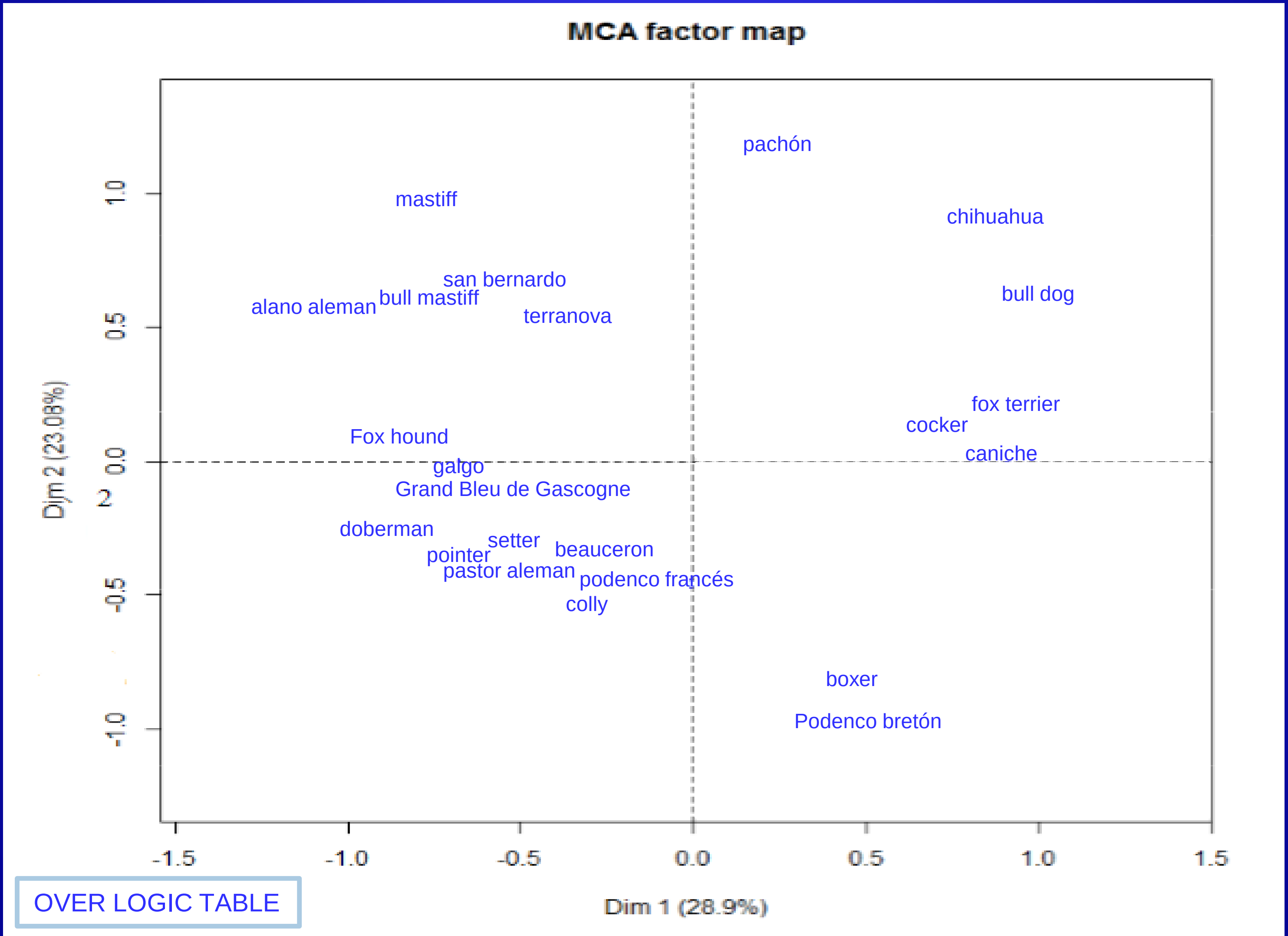
  >library(FactoClass); data("DogBreeds")

1. Data represents different dog breeds which contains 6 different qualitative variables
    1. Tamaño(SIZE): Pequeño(sma), Mediano(med), largo(lar)
    2. Peso(WEIG): ligero(lig), Medio(med), Pesado(hea)
    3. Velocidad(SPEE): Bajo(low), Media(med), Alto(hig)
    4. Inteligencia(INTE): Bajo(low), Media(med), Alto(hig)
    5. Afectividad(AFFE): Bajo(low), Alto(hig)
    6. Agresividad(AGGR): Bajo(low), Alto(hig)

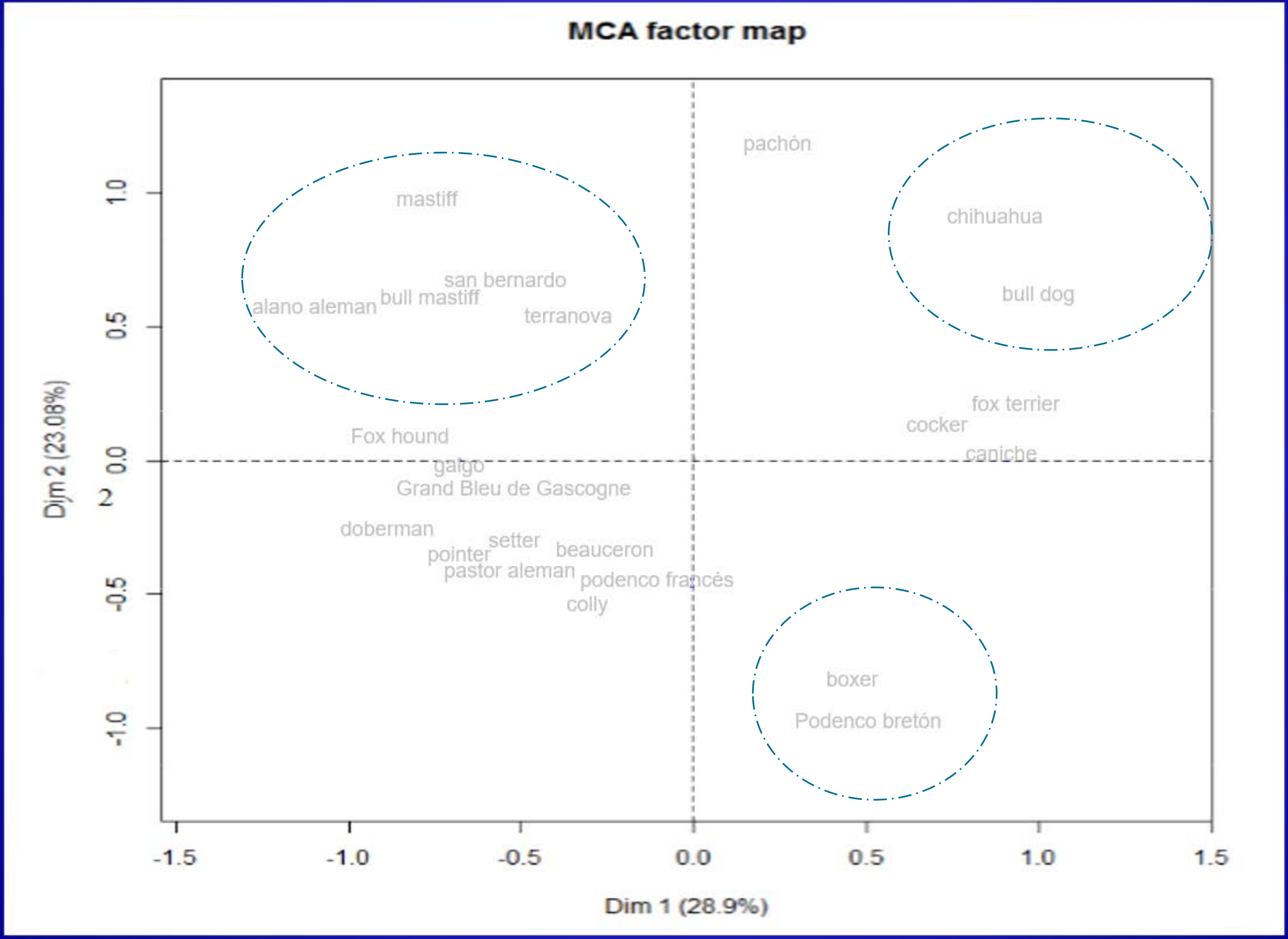
https://upc0-my.sharepoint.com/:x:/g/personal/karina_gibert_upc_edu/EStnc3D4xmVGrnlVIzbwz18Bf_AlK1vIaFKtRXrllDuKeA?e=nFVwh4

Factorial Methods

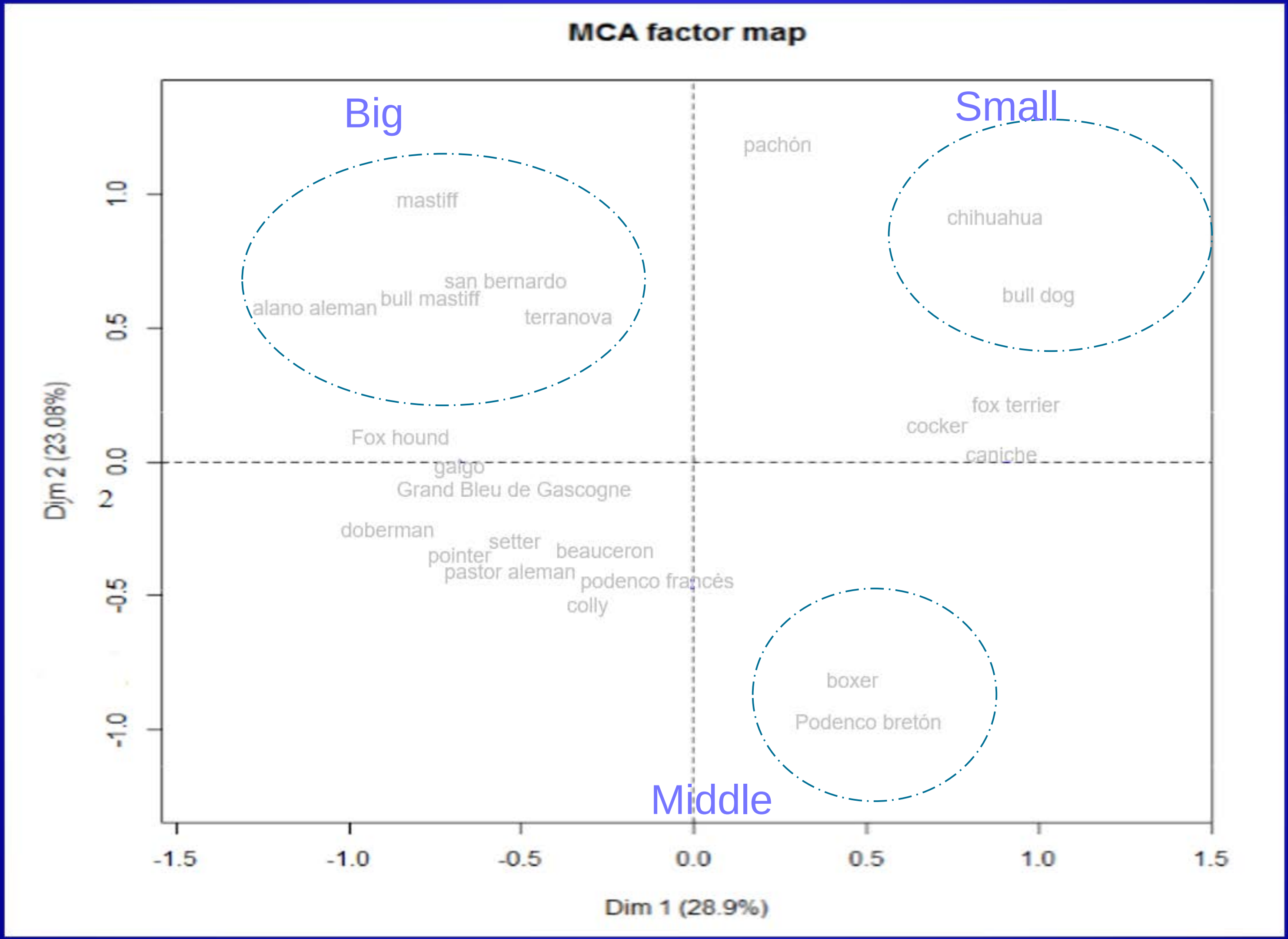
Projection of individuals



Factorial Methods
Projection of individuals

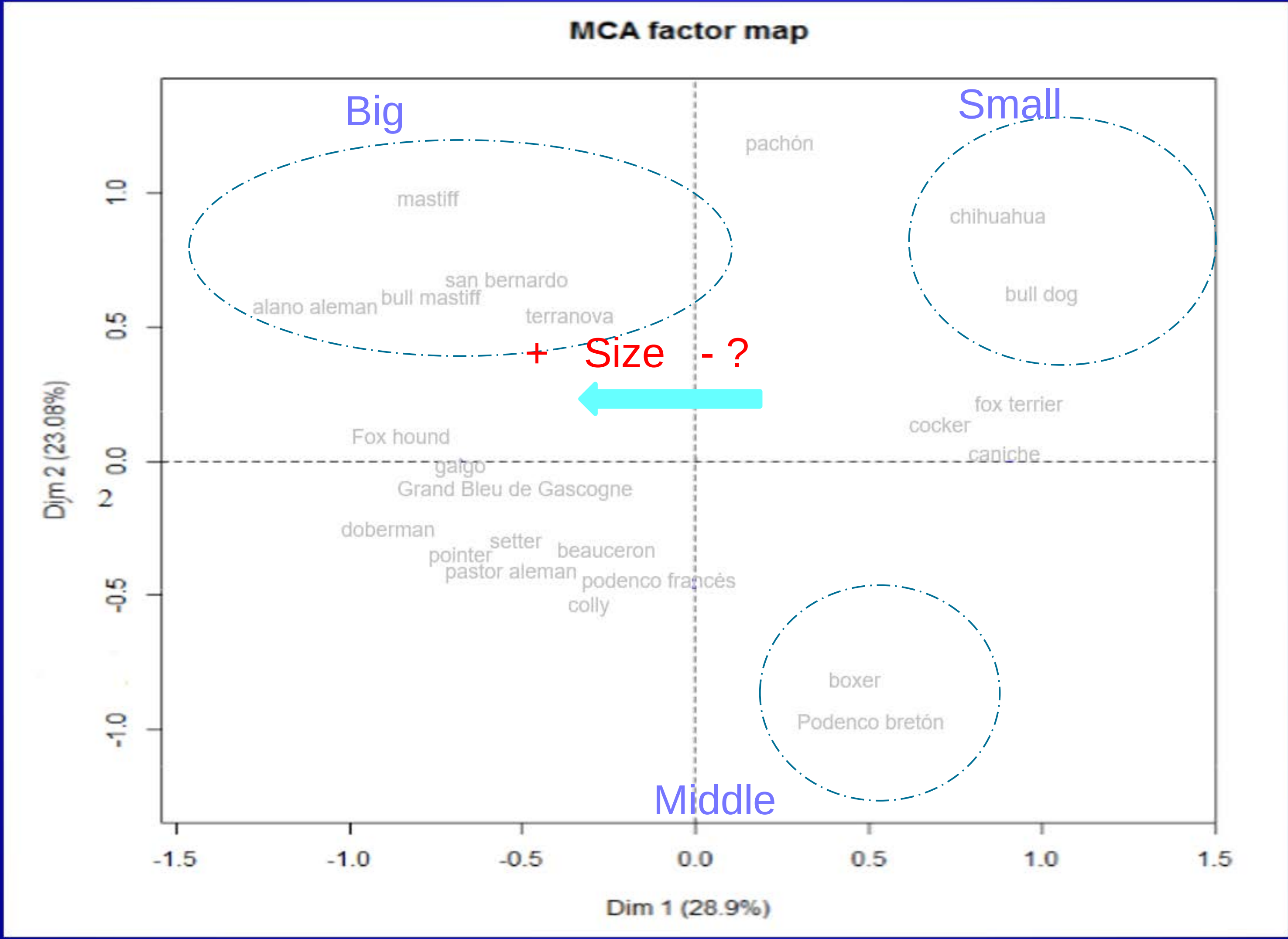


Factorial Methods
Projection of individuals



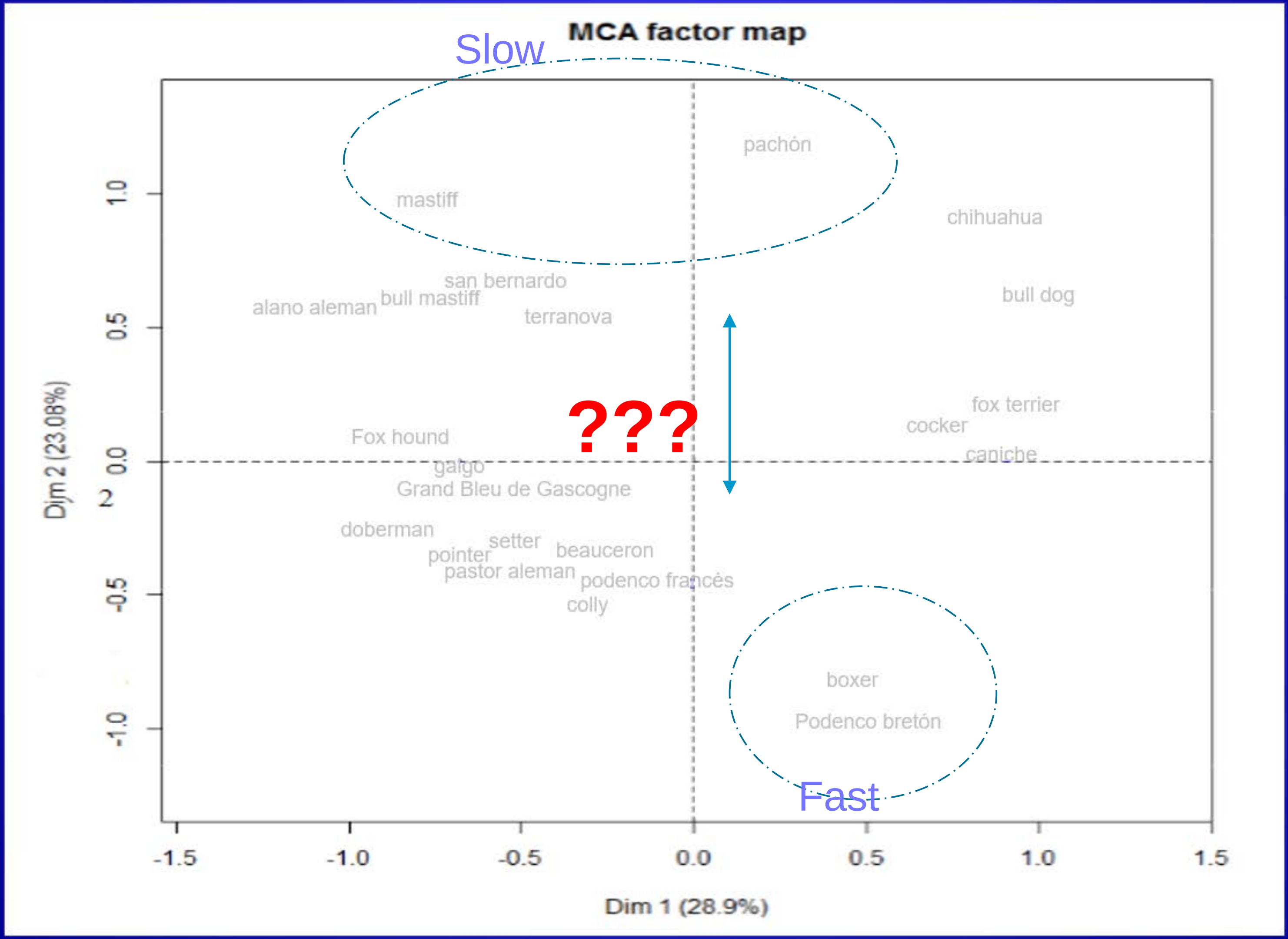
Factorial Methods

Projection of individuals



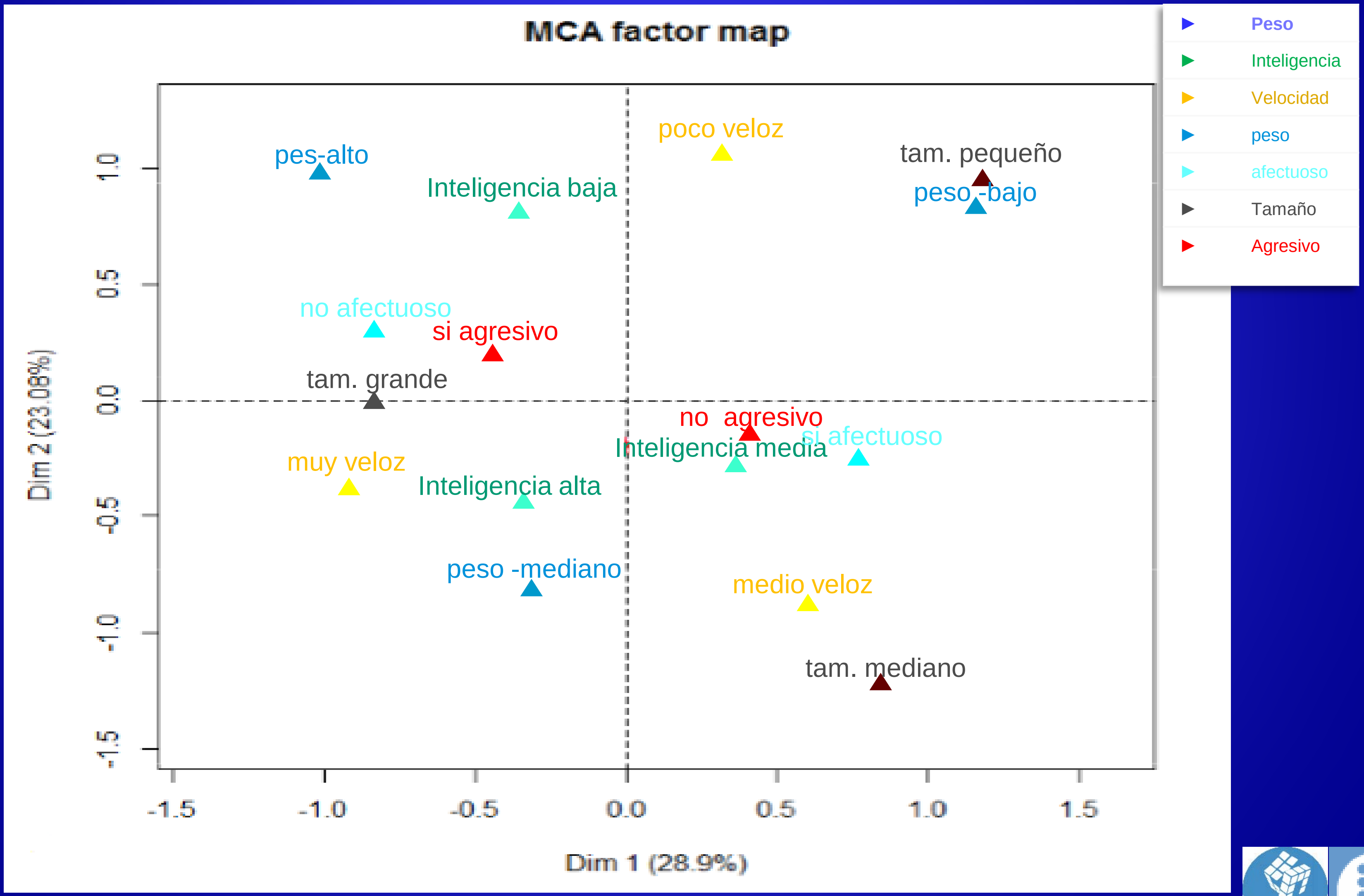
Factorial Methods

Projection of individuals



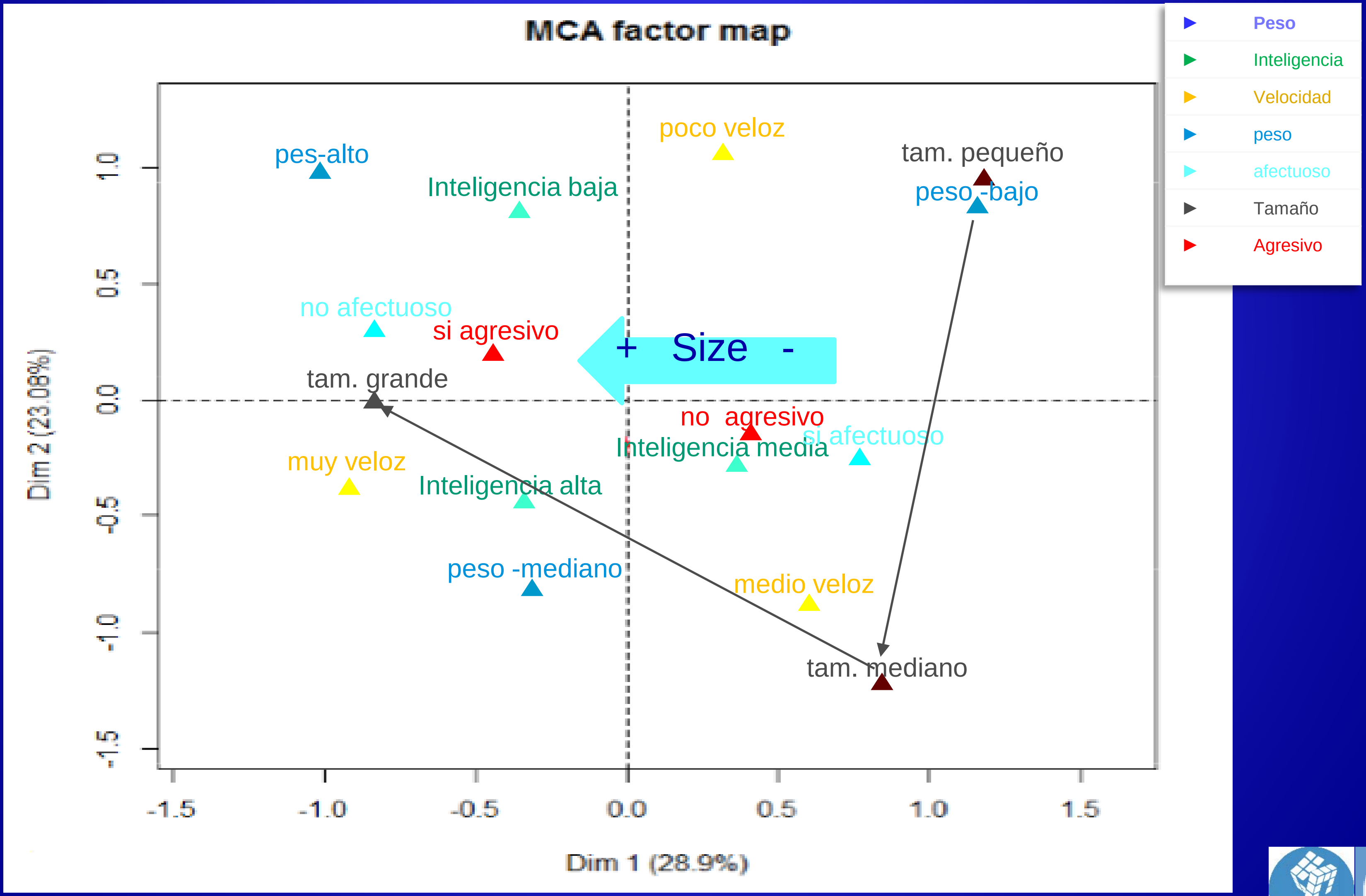
Factorial Methods

Projection of variables



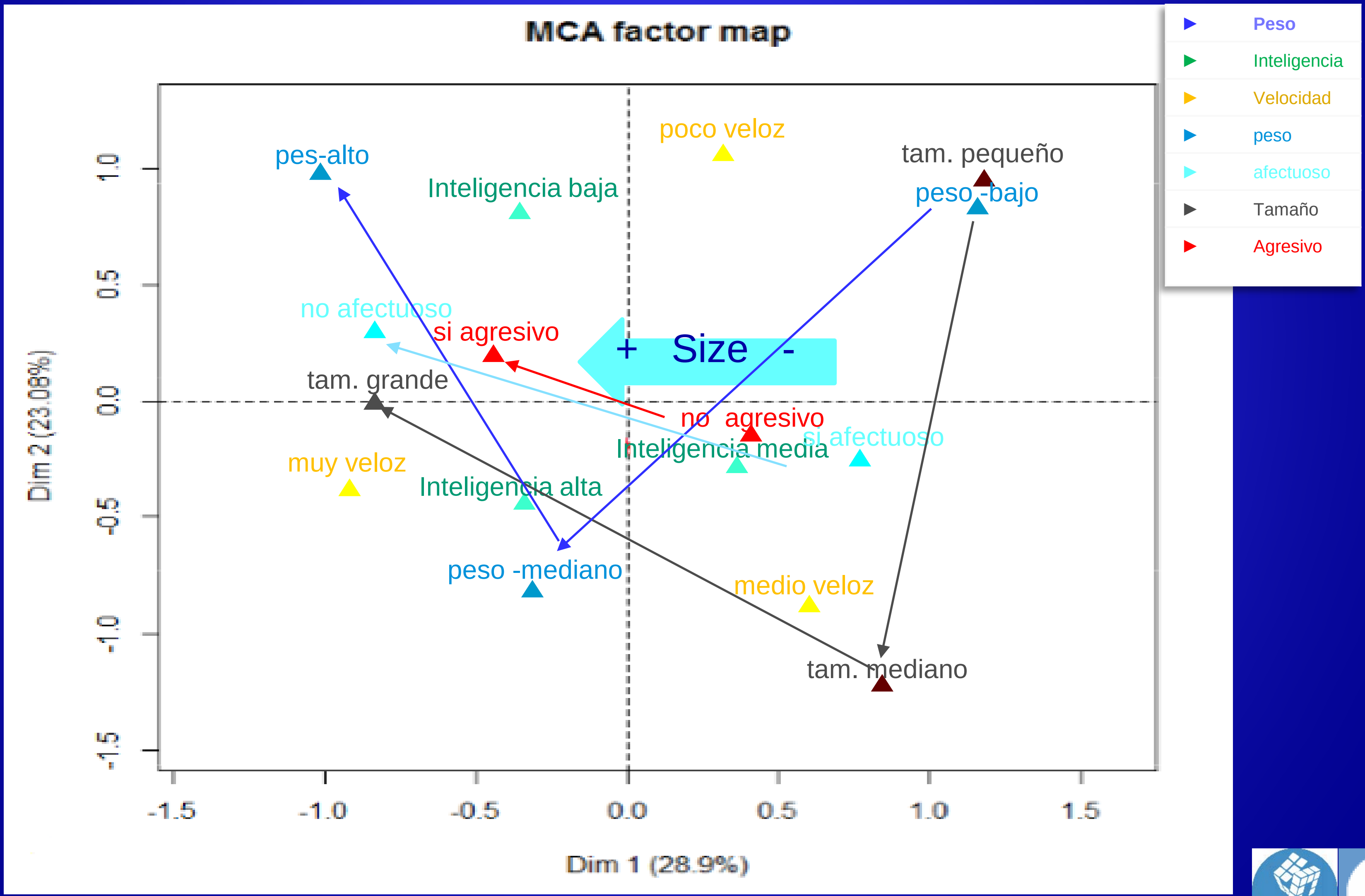
Factorial Methods

Projection of variables



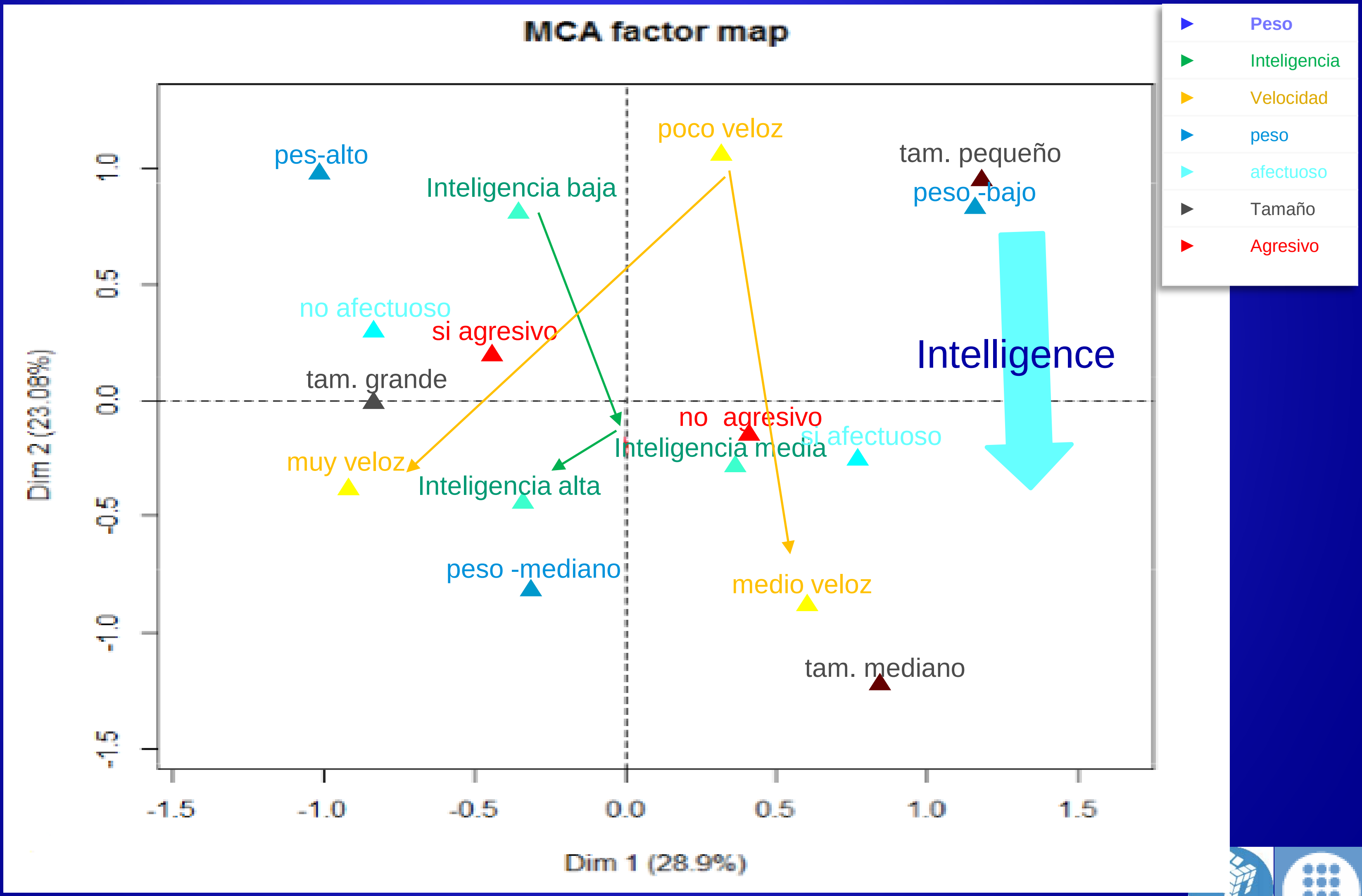
Factorial Methods

Projection of variables



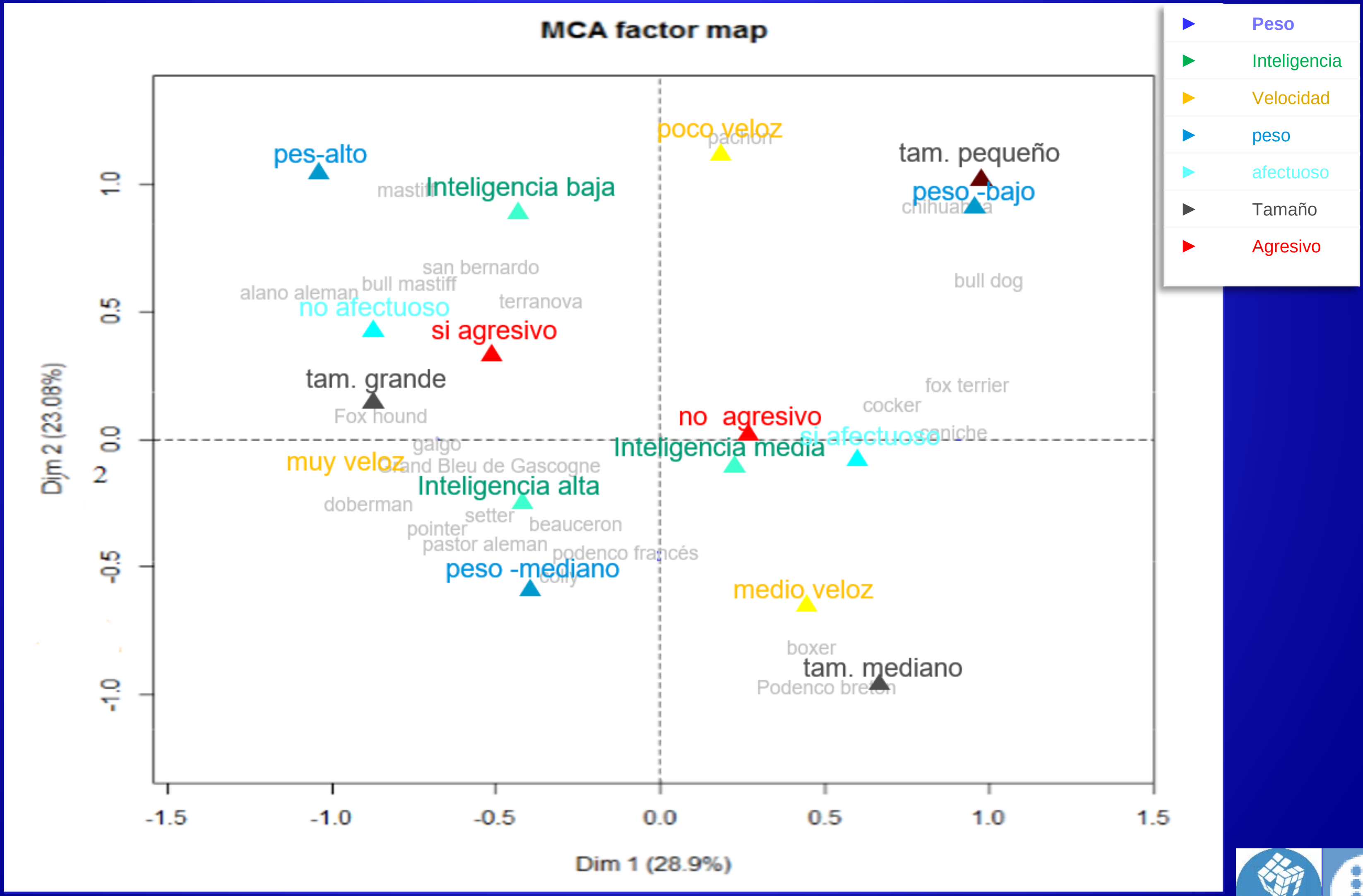
Factorial Methods

Projection of variables



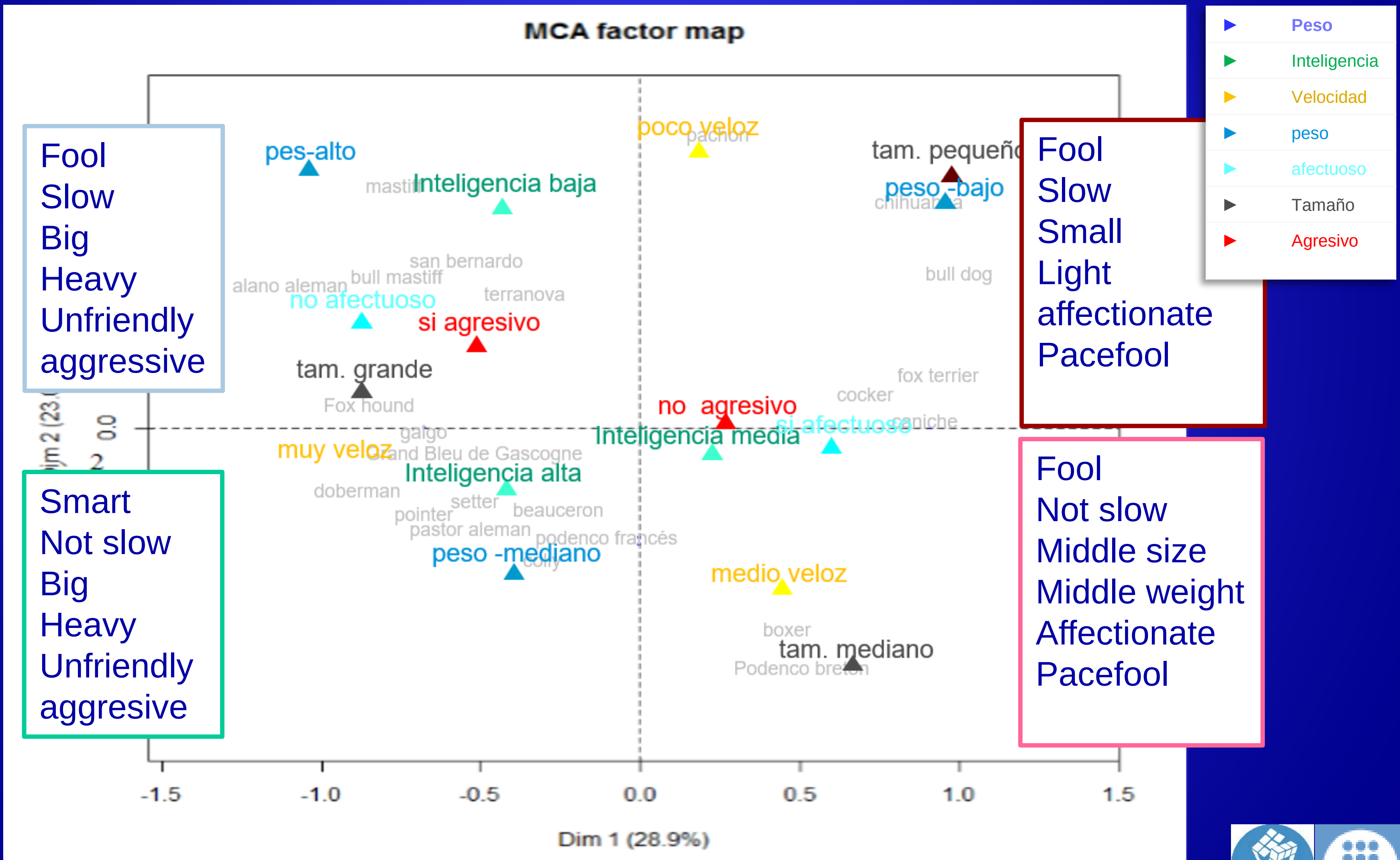
Factorial Methods

Projection of individuals



Factorial Methods

Projection of individuals



Fool
Slow
Big
Heavy
Unfriendly
aggressive

Smart
Not slow
Big
Heavy
Unfriendly
aggresive

Fool
Slow
Small
Light
affectionate
Pacefool

Fool
Not slow
Middle size
Middle weight
Affectionate
Pacefool

MCA

Analysis of Burt Table

- Given <X,M,D>
- *X =B* *B: Burt Table,* $M^{½} X'DXM^{½}$

$$B=Z'Z=\begin{bmatrix} Z_1'Z_1 & Z_1'Z_2 & Z_1'Z_3 & \cdots & Z_1'Z_q \\ Z_2'Z_1 & Z_2'Z_2 & Z_2'Z_3 & \cdots & Z_2'Z_q \\ Z_3'Z_1 & Z_3'Z_2 & Z_3'Z_3 & \cdots & Z_3'Z_q \\ & & \cdots & & \\ & & \cdots & & \\ Z_q'Z_1 & Z_q'Z_2 & Z_q'Z_3 & \cdots & Z_q'Z_q \end{bmatrix}$$

IDEAI UPC

MCA BURT TABLE. EXAMPLE

n=7, Q=3, J_1 =5, J_2 =6, J_3 =3

$Z_{(7*14)}$=

Z1					Z2						Z3		
A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	B9	C1	C2	C3
0	1	0	0	0	0	0	0	1	0	0	0	0	1
0	1	0	0	0	0	0	0	1	0	0	0	0	1
0	0	1	0	0	0	1	0	0	0	0	0	1	0
0	0	1	0	0	0	0	0	0	0	1	0	1	0
0	0	0	0	1	1	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	1	0	0	0	0	1
1	0	0	0	0	0	0	1	0	0	0	0	1	0

$$B=\begin{bmatrix} Z_1'Z_1 & Z_1'Z_2 & Z_1'Z_3 & \dots & Z_1'Z_q \\ Z_2'Z_1 & Z_2'Z_2 & Z_2'Z_3 & \dots & Z_2'Z_q \\ Z_3'Z_1 & Z_3'Z_2 & Z_3'Z_3 & \dots & Z_3'Z_q \\ & & \dots & & \\ & & \dots & & \\ Z_q'Z_1 & Z_q'Z_2 & Z_q'Z_3 & \dots & Z_q'Z_q \end{bmatrix}$$

$Z_1'Z_2$ =

	A1	A2	A3	A4	A5	A6
A2	0	0	1	0	0	0
A2	0	0	0	3	0	0
A3	0	1	0	0	0	1
A4	0	0	0	0	0	0
A5	1	0	0	0	0	0

MCA
Analysis of BURT Table
X Burt table (X=B)

Properties of B

- B not sparse
- B smaller than Z

$(\sum_{q=1}^{Q} \mathrm{J_q})$^2 < nxQ

- B Symmetric
- B Unbiased

Properties of the analysis

- Reduced dimensionality
- Direct and dual analysis identical
- Centroids project on CDG of their individuals
- $G_{\propto}^{B(j)} = \sum_{i=1}^{N} \frac{k_{ij}}{k_j F_{\propto}(i)}$

No loss of information

Individuals can project as supplementary

MCA

- Vaps:
  - Of the Burt matrix :

    Square of the vaps obtained by the AC of indicator table

- The matrices on the diagonal of the Burt matrix have maximal inertia, $JQ - 1$ each one

Factorial Methods

- MCA PROPERTIES WITH BURT TABLE

  - Factorial axes build from contingency tables
  - Rows of B are not individuals.
  - Columns of B are not individuals

  BUT

  - Coordinates of projecting individuals as supplementary objects in map obtained with MCA on Burt table are EQUAL to thos projecting individuals on map obtained with MCA on logic table (except for a small corrector factor called systematic bias).
  - The logic table has individuals on rows

MCA

- Projection coordinates:
  - Projection coordinates (if starting from a Burt's table):
    - Over α projection axis of j modality are :
      - Those obtained from the indicator's table multiplied by the square root of the corresponding vap, equivalently.

MCA

- Projection coordinates:
  - Each individual:
  - Can be projected as a supplementary point.
  - How:
    - adding at the bottom of the Burt matrix the complete disjunctive vectors of every individual.

34