UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

DEGREE IN ARTIFICIAL INTELLIGENCE

Correspondence Analysis (CA)

Prof. Dante Conti

Credits:Prof. Tomàs Aluja

Count tables

Tables formed by crossing two categorical variables.
Every cell express the number of times modalities *i* and *j* cooccur.

Examples:

Ecological: Sites by species.
Lexical: Individuals by words
Marketing: product by adjective
Urbanism: Neighborhood by Occupation
Politics: crossing census units by political parties (cells expressing the votes)
. . .

Goal: *Automatic revealing of the information contained in the table :*

1. *which rows are similar,*
2. *which columns are similar,*
3. *which relations exist between rows and columns*

Miguel Hernández poems - Example

```
> poems_MH <- read.table("poemas_Miguel_Hernandez.txt",header=T,sep="\t")
```

Miguel Hernandez, 1910-1942

```
> poems_MH
                              AMOR CORAZON HUERTO MUERTE SANGRE LUZ HOMBRE
La Morada                       41       3     32     21      8  52      5
Perito en Lunas                  4       1      3      3      1  12      0
Oda a la Higuera                37       6     11     27     14  35      6
Rayo que no cesa                17      26      0      8     12   1      1
Mi sangre es un camino           7      16      0      9     26   1      2
Vientos del pueblo               3      23      2     61     35   3     22
Romancero de ausencias          44      20      2     38     25  19     19
Hijo de la luz y de la sombra   14      11      2     15     13  25      8
```

rows and columns are exchangeable

Checking some concepts with Contingency Tables

Tabla de contingencia (frecuencias absolutas)

Modalidades en las columnas

Modalidades en las filas

$X \setminus Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_J	
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet J}$	

Checking some concepts with Contingency Tables

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Perfiles fila

	y_1	y_2	$\cdots$	y_j	$\cdots$	y_J	
$Y / X = x_1$	$f_{1/1}$	$f_{2/1}$	$\cdots$	$f_{j/1}$	$\cdots$	$f_{J/i}$	1
$Y / X = x_2$	$f_{1/2}$	$f_{2/2}$	$\cdots$	$f_{j/2}$	$\cdots$	$f_{J/2}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Y / X = x_i$	$f_{1/i}$	$f_{2/i}$	$\cdots$	$f_{j/i}$	$\cdots$	$f_{J/i}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Y / X = x_I$	$f_{1/I}$	$f_{2/I}$	$\cdots$	$f_{j/I}$	$\cdots$	$f_{J/I}$	1

$$f_{j/i} = \frac{n_{ij}}{n_{i\cdot}} = \frac{f_{ij}}{f_{i\cdot}}$$

A cada perfil fila i le corresponde una *masa* igual a su frecuencia relativa, $f_{i\cdot}$, que representa el peso de dicho perfil en el análisis.

Perfiles columna

	$X/Y=y_1$	$X/Y=y_2$	$\dots$	$X/Y=y_j$	$\dots$	$X/Y=y_J$
x_1	$f_{1/1}$	$f_{1/2}$	$\dots$	$f_{1/j}$	$\dots$	$f_{1/J}$
x_2	$f_{2/1}$	$f_{2/2}$	$\dots$	$f_{2/j}$	$\dots$	$f_{2/J}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	$f_{i/1}$	$f_{i/2}$	$\dots$	$f_{i/j}$	$\dots$	$f_{i/J}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_I	$f_{I/1}$	$f_{I/2}$	$\dots$	$f_{I/j}$	$\dots$	$f_{I/J}$
	1	1	$\dots$	1	$\dots$	1

$$f_{i/j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{f_{ij}}{f_{\cdot j}}$$

A cada perfil columna j le corresponde una *masa* igual a su frecuencia relativa, $f_{\cdot j}$, que representa el peso de dicho perfil en el análisis.

Centroides

Los elementos del centroide o vector de medias de los perfiles fila coinciden con las frecuencias relativas de las modalidades en las columnas,

$$C_f = (f_{\cdot 1}, \ldots, f_{\cdot j}, \ldots, f_{\cdot J})$$

De forma similar, el centroide (vector de medias) de los perfiles columna está formado por las frecuencias relativas de las modalidades en las filas,

$$C_c = (f_{1 \cdot}, \ldots, f_{i \cdot}, \ldots, f_{I \cdot})$$

A partir de una tabla de contingencia en frecuencias relativas (f_{ij}), sumando por filas se obtiene el centroide de las columnas y sumando por columnas se obtiene el centroide de las filas.

Tabla de contingencia (frecuencias relativas)

Modalidades en las columnas

Modalidades en las filas

$X \setminus Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_J	Centroide de los perfiles columna
x_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1J}	$f_{1\bullet}$
x_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2J}	$f_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{iJ}	$f_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_I	f_{I1}	f_{I2}	$\cdots$	f_{Ij}	$\cdots$	f_{IJ}	$f_{I\bullet}$
Centroide de los perfiles fila	$f_{\bullet 1}$	$f_{\bullet 2}$	$\cdots$	$f_{\bullet j}$	$\cdots$	$f_{\bullet J}$	1

Distancia chi-cuadrado

La distancia chi-cuadrado es la distancia natural cuando se trabaja con datos cualitativos, entre otros motivos porque mantiene las distancias entre las filas cuando se agrupan dos columnas con el mismo perfil (y las distancias entre las columnas cuando se agrupan dos filas con el mismo perfil). Así, la distancia chí-cuadrado entre el perfil fila i y el perfil fila i' es

$$d_{\chi^2}(i, i') = \sqrt{\sum_{j=1}^{J} \left(f_{j/i} - f_{j/i'}\right)^2 \frac{1}{f_{\cdot j}}}$$

Inercia

La inercia total es la cantidad total de información de la tabla de contingencia, esto es, su variabilidad o dispersión. Puede obtenerse como la suma de las inercias de las filas o como la suma de las inercias de las columnas.

La inercia de la fila i es la masa de esa fila ($f_{i\cdot}$) multiplicada por el cuadrado de la distancia chi-cuadrado entre el perfil de fila i y su centroide, $d^2_{\chi^2}(i, C_f)$

$$Inercia\ fila\, i = f_{i\cdot}\, d^2_{\chi^2}(i, C_f)$$

donde,

$$d^2_{\chi^2}(i, C_f) = \sum_{j=1}^{J} \left(f_{j/i} - f_{\cdot j}\right)^2 \frac{1}{f_{\cdot j}}$$

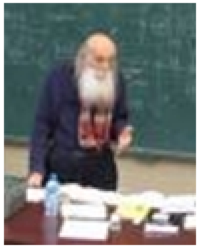
La inercia total coincide con el valor del estadístico chi-cuadrado del contraste de independencia, χ^2, dividido por el número total de observaciones, n, donde

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n}\right)^2}{\frac{n_{i.} n_{.j}}{n}}$$

Reducción de la dimensión

El objetivo del ACS es determinar un espacio de baja dimensión (un plano a ser posible) que recoja la mayor parte de la inercia (variabilidad) de los datos. Es decir, que reproduzca lo mejor posible las distancias entre los perfiles fila y su centroide (y, por tanto, las distancias entre perfiles fila), y análogamente para los perfiles columna.

Realmente, el ACS permite obtener p coordenadas numéricas (o *puntuaciones*) para cada fila y cada columna, de modo que la distancia euclidea entre las coordenadas de dos filas (o de dos columnas) coincide con la distancia chi-cuadrado entre los respectivos perfiles fila (o perfiles columna). Pero como los ejes del nuevo espacio, que reciben el nombre de *dimensiones*, se determinan de modo que estén ordenados de mayor a menor inercia, en general se consigue una buena representación de los perfiles en el subespacio formado por las dos o tres primeras dimensiones, lo que permite la visualización de los mismos.



Jean Paul Benzecri,
Analyse des Données father

as a particular case of PCA

Let K be a count table

$$n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$$

$$n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$$

$$K= \begin{array}{c|ccc|c} & 1 & j & J & \\ \hline 1 & & & & \\ i & & n_{ij} & & n_{i.} \\ I & & & & \\ \hline & & n_{.j} & & n \end{array}$$

$$n = \sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} = \sum_{j=1}^{J} n_{\cdot j} = \sum_{i=1}^{I} n_{i\cdot}$$

Let F the matrix of relative frequencies

$$F = \frac{1}{n} K$$

$$F= \begin{array}{c|ccc|c} & 1 & & J & \\ \hline 1 & & & & \\ i & & f_{ij} & & f_{i.} \\ I & & & & \\ \hline & & f_{.j} & & 1 \end{array}$$

Table of frequencies

```
> K <- poems_MH
> F <- K/sum(K)
> F
                               AMOR  CORAZON HUERTO MUERTE SANGRE    LUZ HOMBRE
La Morada                    0.0481  0.0035 0.0376 0.0246 0.0094 0.0610 0.0059
Perito en Lunas              0.0047  0.0012 0.0035 0.0035 0.0012 0.0141 0.0000
Oda a la Higuera             0.0434  0.0070 0.0129 0.0317 0.0164 0.0411 0.0070
Rayo que no cesa             0.0200  0.0305 0.0000 0.0094 0.0141 0.0012 0.0012
Mi sangre es un camino       0.0082  0.0188 0.0000 0.0106 0.0305 0.0012 0.0023
Vientos del pueblo           0.0035  0.0270 0.0023 0.0716 0.0411 0.0035 0.0258
Romancero de ausencias       0.0516  0.0235 0.0023 0.0446 0.0293 0.0223 0.0223
Hijo de la luz y de la sombra 0.0164  0.0129 0.0023 0.0176 0.0153 0.0293 0.0094

> fi <- rowSums(F)
                    La Morada                         0.190
              Perito en Lunas                         0.028
             Oda a la Higuera                         0.160
             Rayo que no cesa                         0.076
       Mi sangre es un camino                         0.072
           Vientos del pueblo                         0.175
       Romancero de ausencias                         0.196
Hijo de la luz y de la sombra                         0.103

> fj <- colSums(F)
     AMOR     CORAZON      HUERTO      MUERTE      SANGRE         LUZ      HOMBRE
    0.196       0.124       0.061       0.214       0.157       0.174       0.074
```

The F and the diagonal matrices of weights

$$
\begin{array}{rcl}
 & J & \\
F = & \begin{bmatrix} & \vdots & \\ \cdots & f_{ij} & \cdots \\ & \vdots & \end{bmatrix} & f_{i\cdot} \\
I & f_{\cdot j} &
\end{array}
$$

Diagonal matrix of weights of rows

$$
D_I = \begin{bmatrix} \ddots & & 0 \\ & f_{i\cdot} & \\ 0 & & \ddots \end{bmatrix}
$$

$$
D_J = \begin{bmatrix} \ddots & & 0 \\ & f_{\cdot j} & \\ 0 & & \ddots \end{bmatrix}
$$

Diagonal matrix of weights of columns

Comparison of rows ?

$$row\ i\ of\ F\ :\ \left(f_{ij}\right)\ j = 1, \ldots, J$$

	AMOR	CORAZON	HUERTO	MUERTE	SANGRE	LUZ	HOMBRE	fi
La Morada	0.0481	0.0035	0.0376	0.0246	0.0094	0.0610	0.0059	0.190
Perito en lunas	0.0047	0.0012	0.0035	0.0035	0.0012	0.0141	0.0000	0.028

or

$$row\ i\ :\ \left(\frac{f_{ij}}{f_{i\cdot}}\right)\ j = 1, \ldots, J$$

	AMOR	CORAZON	HUERTO	MUERTE	SANGRE	LUZ	HOMBRE
La Morada	0.2531	0.0185	0.1975	0.130	0.0494	0.3210	0.0309
Perito en lunas	0.1667	0.0417	0.1250	0.125	0.0417	0.5000	0.0000

Conditional distribution of words *(row)* within the poems *(columns)* $\equiv$ Row-profiles

Defining the cloud of rows:

1. Row profile: $$f_{j/i} = \frac{f_{ij}}{f_{i\cdot}} = \frac{n_{ij}}{n_{i\cdot}},\ j = 1,\ldots,J$$

Conditional frequencies of rows

2. Row's weights: $f_{i\cdot}$

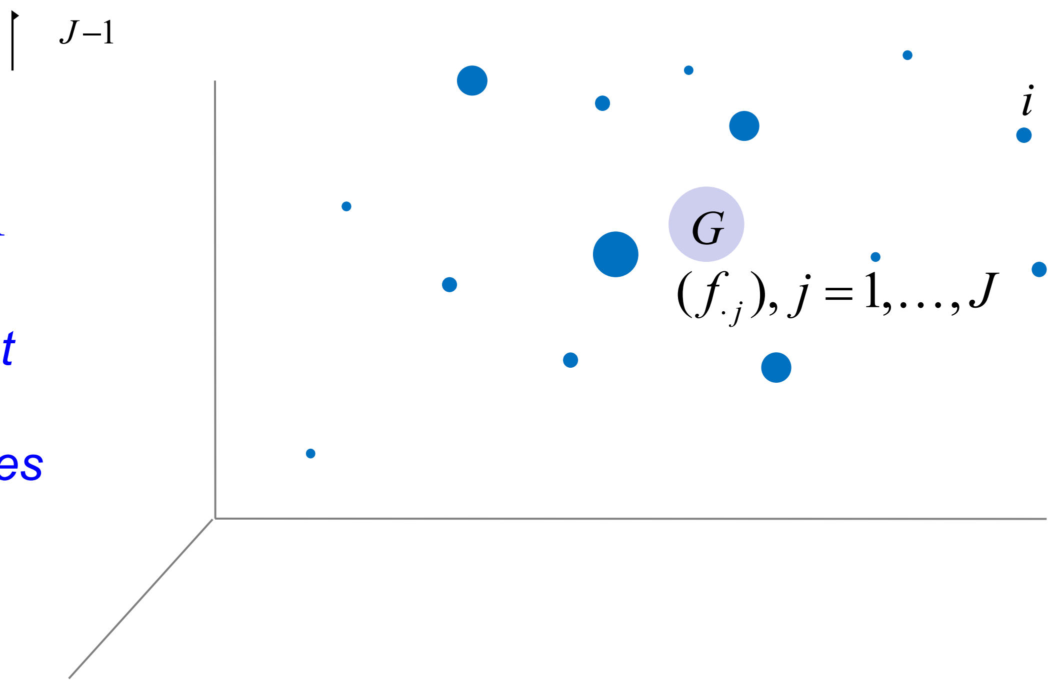
$$F_{J/I} = D_I^{-1}F = \begin{array}{|ccc|l}\hline & & & \\ \cdots & \dfrac{f_{ij}}{f_{i\cdot}} & \cdots & 1 \\ & & & \\ \hline \end{array}$$

Matrix of *I* row-profiles *(I,J)*

$$\begin{array}{|ccc|}\hline \cdots & f_{\cdot j} & \cdots \\ \hline \end{array}$$

cdg of row profiles: $$\sum_{i=1}^{I} \frac{f_{ij}}{f_{i\cdot}} f_{i\cdot} = f_{\cdot j}$$

Cloud of row profiles



Coordinates of points

$$i \quad \left(\frac{f_{ij}}{f_{i\cdot}}\right) j = 1, \ldots, J$$

$$\sum_{j=1}^{J} \frac{f_{ij}}{f_{i\cdot}} = 1$$

Constraint upon the coordinates

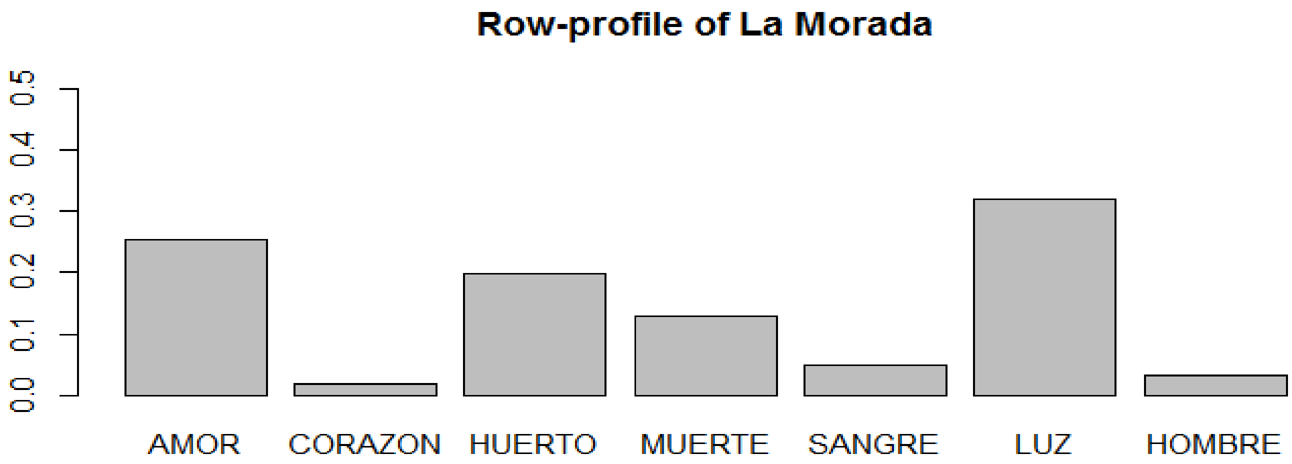
Computing the row-profiles

```
> fi <- rowSums(F)
> Di <- diag(fi)
> Fi <- solve(Di) %*% as.matrix(F)
> print(Fi,digits=3)
                             AMOR CORAZON HUERTO MUERTE SANGRE    LUZ HOMBRE
La Morada                  0.2531  0.0185 0.1975  0.130 0.0494 0.3210 0.0309
Perito en Lunas            0.1667  0.0417 0.1250  0.125 0.0417 0.5000 0.0000
Oda a la Higuera           0.2721  0.0441 0.0809  0.199 0.1029 0.2574 0.0441
Rayo que no cesa           0.2615  0.4000 0.0000  0.123 0.1846 0.0154 0.0154
Mi sangre es un camino     0.1148  0.2623 0.0000  0.148 0.4262 0.0164 0.0328
Vientos del pueblo         0.0201  0.1544 0.0134  0.409 0.2349 0.0201 0.1477
Romancero de ausencias     0.2635  0.1198 0.0120  0.228 0.1497 0.1138 0.1138
Hijo de la luz y de la sombra 0.1591  0.1250 0.0227  0.170 0.1477 0.2841 0.0909
```

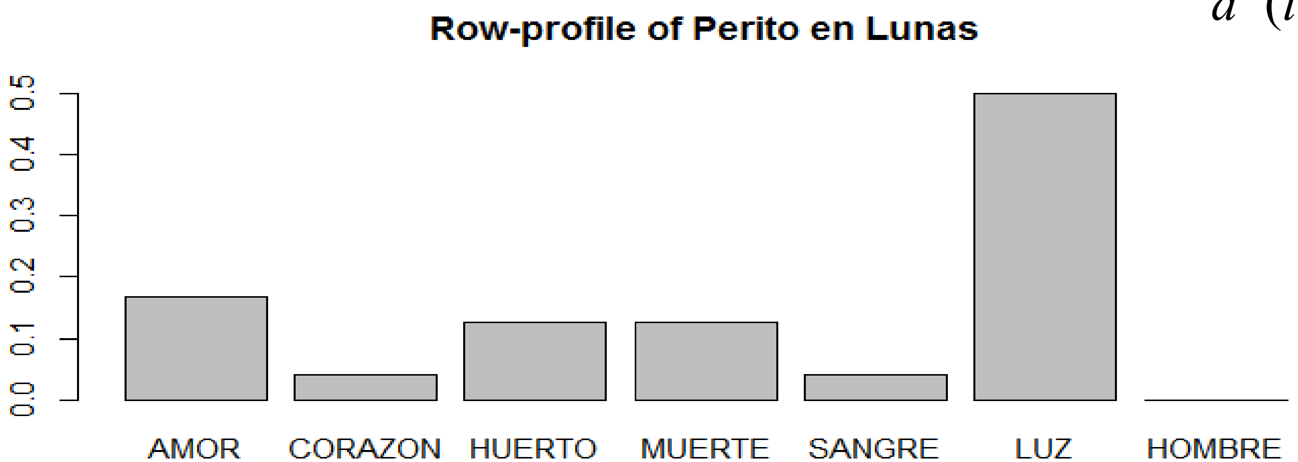
```
> apply(Fi,2,weighted.mean, w=fi)
      AMOR    CORAZON     HUERTO     MUERTE     SANGRE        LUZ     HOMBRE
0.19600939 0.12441315 0.06103286 0.21361502 0.15727700 0.17370892 0.07394366
```

```
> fj
      AMOR    CORAZON     HUERTO     MUERTE     SANGRE        LUZ     HOMBRE
0.19600939 0.12441315 0.06103286 0.21361502 0.15727700 0.17370892 0.07394366
```

How to measure the distance between two row-profiles?



$$d^2(i,i') = \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$$



classical euclidean ?

The Chi-square distance (L_2)

$$d^2(i,i') = \sum_{j=1}^{J} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$$

Chi-square distance

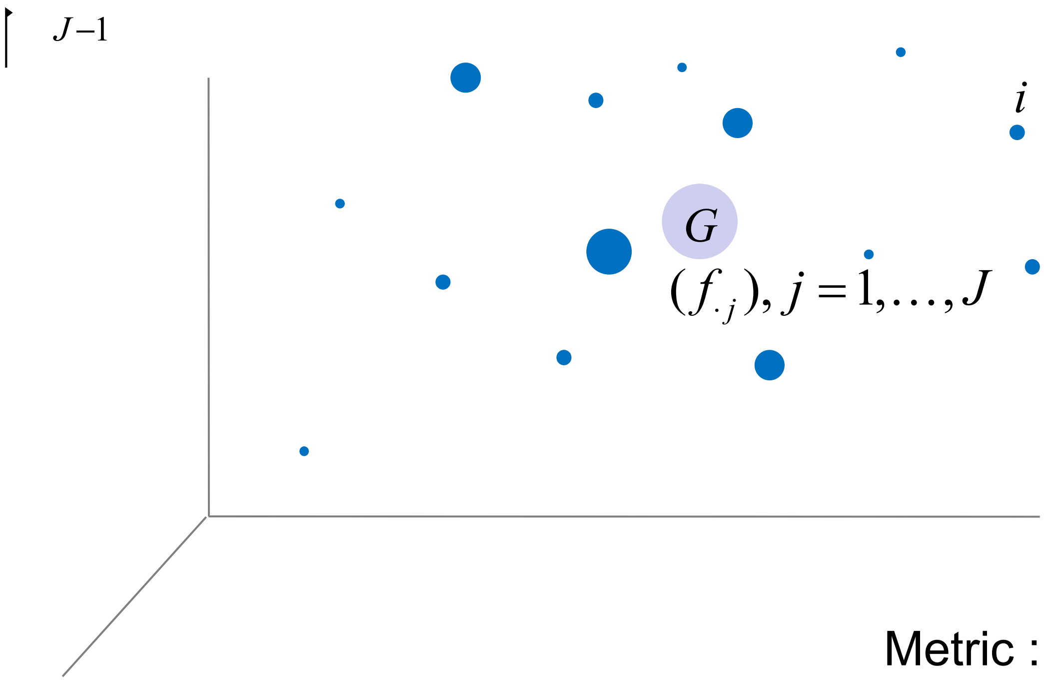
It overweights the rare events

$$M = \begin{pmatrix} \ddots & & \\ & \frac{1}{f_{\cdot j}} & \\ & & \ddots \end{pmatrix}$$

```
> fj <- colSums(F)
     AMOR    CORAZON     HUERTO     MUERTE     SANGRE        LUZ     HOMBRE
    0.196      0.124      0.061      0.214      0.157      0.174      0.074

> 1/fj
     AMOR   CORAZON    HUERTO    MUERTE    SANGRE       LUZ    HOMBRE
 5.101796  8.037736 16.384615  4.681319  6.358209  5.756757 13.523810
```

FIB UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH



Coordinates of points

$$i \quad \left(\frac{f_{ij}}{f_{i\cdot}}\right) \; j = 1, \ldots, J$$

$$i \text{ weight} : f_{i\cdot} \qquad \left(N = D_I\right)$$

Metric : $$d^2(i,i') = \sum_{j=1}^{J} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$$

$$\left(M = D_J^{-1}\right)$$

(Chi-square metric)

The row-profiles coordinates including the metric

Transforming the chi-square metric to the canonical euclidean

$$d^2(i,i') = \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \frac{f_{i'j}}{f_{i'\cdot}\sqrt{f_{\cdot j}}} \right)^2$$

$$F_I = D_I^{-1} F D_J^{-1/2} = \boxed{\cdots \quad \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} \quad \cdots}$$

$$\sqrt{f_{\cdot j}}$$

Coordinates of rows with embedded metric

cdg of rows $\quad \sum_{i=1}^{I} \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} f_{i\cdot} = \sqrt{f_{\cdot j}}$

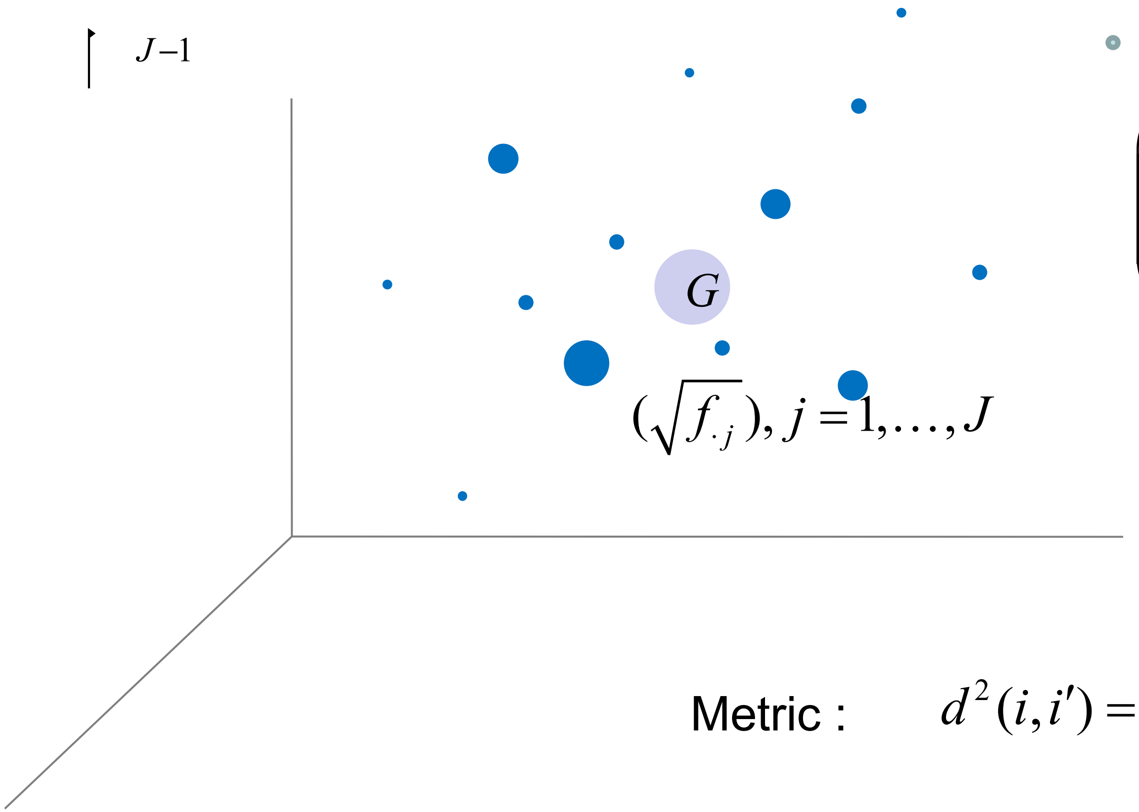
The row-profiles cloud with the metric effect

```
> Fi_m <- Fi %*% diag(1/sqrt(fj))
> Fi_m
 AMOR CORAZON HUERTO MUERTE SANGRE   LUZ  HOMBRE
La Morada                     0.572   0.053  0.800   0.28   0.12 0.770   0.114
Perito en Lunas               0.376   0.118  0.506   0.27   0.11 1.200   0.000
Oda a la Higuera              0.615   0.125  0.327   0.43   0.26 0.617   0.162
Rayo que no cesa              0.591   1.134  0.000   0.27   0.47 0.037   0.057
Mi sangre es un camino        0.259   0.744  0.000   0.32   1.07 0.039   0.121
Vientos del pueblo            0.045   0.438  0.054   0.89   0.59 0.048   0.543
Romancero de ausencias        0.595   0.340  0.048   0.49   0.38 0.273   0.418
Hijo de la luz y de la sombra 0.359   0.354  0.092   0.37   0.37 0.682   0.334


> apply(Fi_m, 2, weighted.mean, w=fi)
     AMOR   CORAZON    HUERTO    MUERTE    SANGRE       LUZ    HOMBRE
0.4427295 0.3527225 0.2470483 0.4621851 0.3965816 0.4167840 0.2719258


> sqrt(fj)
     AMOR   CORAZON    HUERTO    MUERTE    SANGRE       LUZ    HOMBRE
0.4427295 0.3527225 0.2470483 0.4621851 0.3965816 0.4167840 0.2719258
```

The row-profiles with M=I



Coordinates of points

$$\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}\right) \quad j = 1, \ldots, J$$

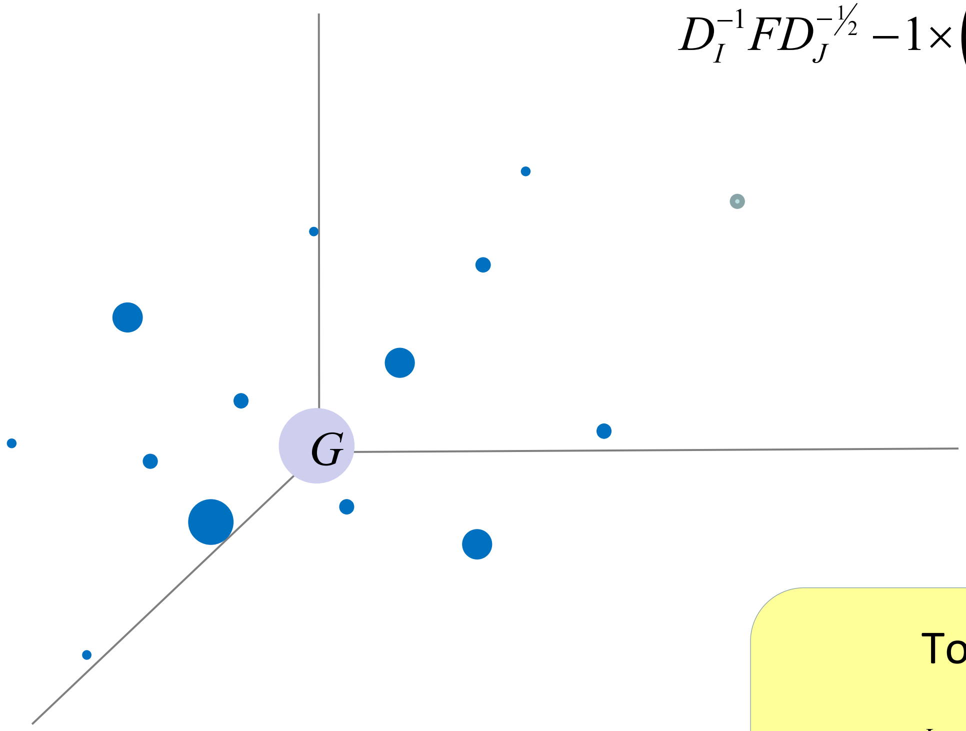
Weight: $f_{i\cdot}$

Metric : $$d^2(i,i') = \sum_{j=1}^{J}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \frac{f_{i'j}}{f_{i'\cdot}\sqrt{f_{\cdot j}}}\right)^2$$

The PCA solution: First we center the data

$$D_I^{-1} F D_J^{-1/2} - 1 \times \left(\sqrt{f_{.j}}\right) = \boxed{\cdots \quad \frac{f_{ij} - f_{i.} f_{.j}}{f_{i.} \sqrt{f_{.j}}} \quad \cdots} = X$$

0



Total inertia of the cloud of row points

$$I_I = \sum_{i=1}^{I} f_{i.} \sum_{j=1}^{J} \left(\frac{f_{ij} - f_{i.} f_{.j}}{f_{i.} \sqrt{f_{.j}}} \right)^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{f_{ij} - f_{i.} f_{.j}}{\sqrt{f_{i.} f_{.j}}} \right)^2$$

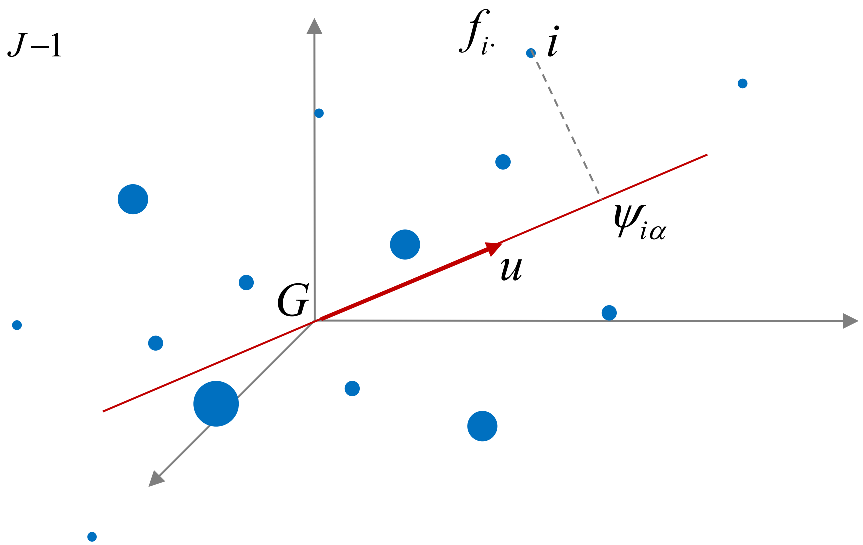
Then we find the directions of maximal inertia

Then we project the cloud of points upon the direction u maximising the inertia

$$u \in \mathbb{R}^J \qquad \mathbb{R}^{J-1}$$

$$u'u = 1$$

$$X = \begin{bmatrix} \cdots & \dfrac{f_{ij} - f_{i\cdot} f_{\cdot j}}{f_{i\cdot}\sqrt{f_{\cdot j}}} & \cdots \end{bmatrix}$$



$$\psi = Xu$$

$$Max \sum_{i=1}^{I} f_{i\cdot} \psi_{i\alpha}^2 \qquad Max\ u'X'D_I Xu$$

Diagonalizing $X'D_IX$

Diagonalize: $$\underset{(I,I)}{X'D_IX = \left(X'D_I^{1/2}\right)\left(D_I^{1/2}X\right) = Z'Z}$$

$$Z = D_I^{1/2}\left(D_I^{-1}FD_J^{-1/2} - 1\times\left(\sqrt{f_{j}}\right)\right) = D_I^{-1/2}FD_J^{-1/2} - \left(\sqrt{f_{i\cdot}}\right)\times\left(\sqrt{f_{j}}\right) =$$

$$\boxed{\cdots \quad \sqrt{f_{i\cdot}}\,\frac{f_{ij}-f_{i\cdot}f_{\cdot j}}{f_{i\cdot}\sqrt{f_{\cdot j}}} \quad \cdots} = \boxed{\cdots \quad \frac{f_{ij}-f_{i\cdot}f_{\cdot j}}{\sqrt{f_{i\cdot}f_{\cdot j}}} \quad \cdots} = Z$$

$$tr(Z'Z) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{f_{ij}-f_{i\cdot}f_{\cdot j}}{\sqrt{f_{i\cdot}f_{\cdot j}}}\right)^2 = I_I$$

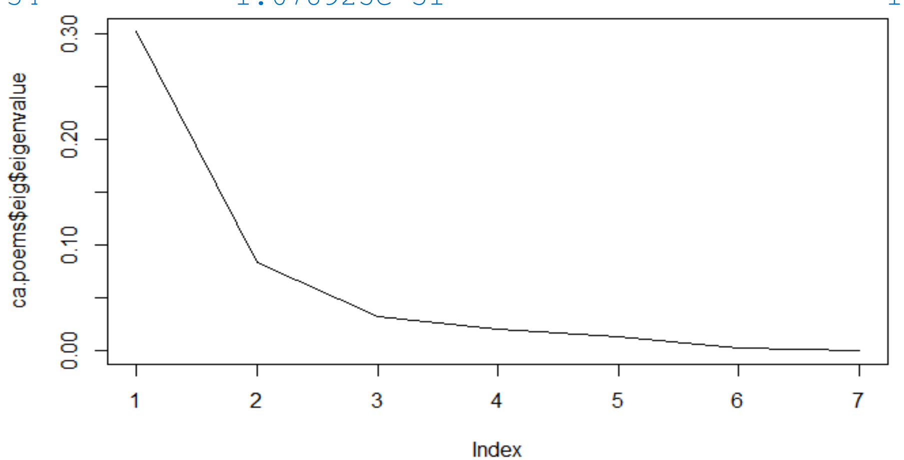
The eigenvalues

```
> pca.poems <- PCA(Fi_centered, scale.unit=FALSE, row.w=fi)
> pca.poems$eig
       eigenvalue percentage of variance cumulative percentage of variance
comp 1 3.021171e-01           6.619164e+01                          66.19164
comp 2 8.392238e-02           1.838678e+01                          84.57842
comp 3 3.241546e-02           7.101990e+00                          91.68040
comp 4 2.105919e-02           4.613914e+00                          96.29432
comp 5 1.392728e-02           3.051365e+00                          99.34568
comp 6 2.986480e-03           6.543160e-01                         100.00000
comp 7 2.769580e-33           6.067948e-31                         100.00000
```

CA in as a PCA of the row-profiles

```
> ca.poems <- CA(K)
> ca.poems$eig
      eigenvalue percentage of variance cumulative percentage of variance
dim 1 3.021171e-01           6.619164e+01                          66.19164
dim 2 8.392238e-02           1.838678e+01                          84.57842
dim 3 3.241546e-02           7.101990e+00                          91.68040
dim 4 2.105919e-02           4.613914e+00                          96.29432
dim 5 1.392728e-02           3.051365e+00                          99.34568
dim 6 2.986480e-03           6.543160e-01                         100.00000
dim 7 7.663080e-34           1.678925e-31                         100.00000
```

CA in R



Projecting the row-profiles

$$\psi_{i\alpha} = \sum_{j=1}^{J} \frac{f_{ij} - f_{i\cdot} f_{\cdot j}}{f_{i\cdot}\sqrt{f_{\cdot j}}} u_{j\alpha} = \sum_{j=1}^{J} \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} u_{j\alpha} = x_i' u_\alpha$$

ψ is called a **factor**

$$\sum_{i=1}^{I} f_{i\cdot} \psi_{i\alpha} = 0$$

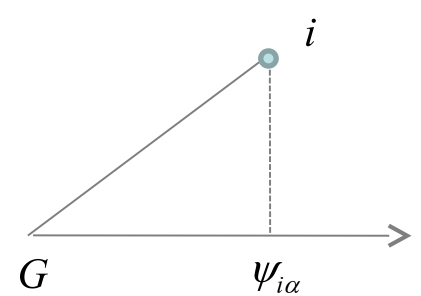
ψ is centered

$$\sum_{i=1}^{I} f_{i\cdot} \psi_{i\alpha}^2 = \lambda_\alpha$$

ψ inertia is equal to its eigenvalue

Contribution of a row i to a factor: $\dfrac{f_{i\cdot}\psi_{i\alpha}^2}{\lambda_\alpha}$

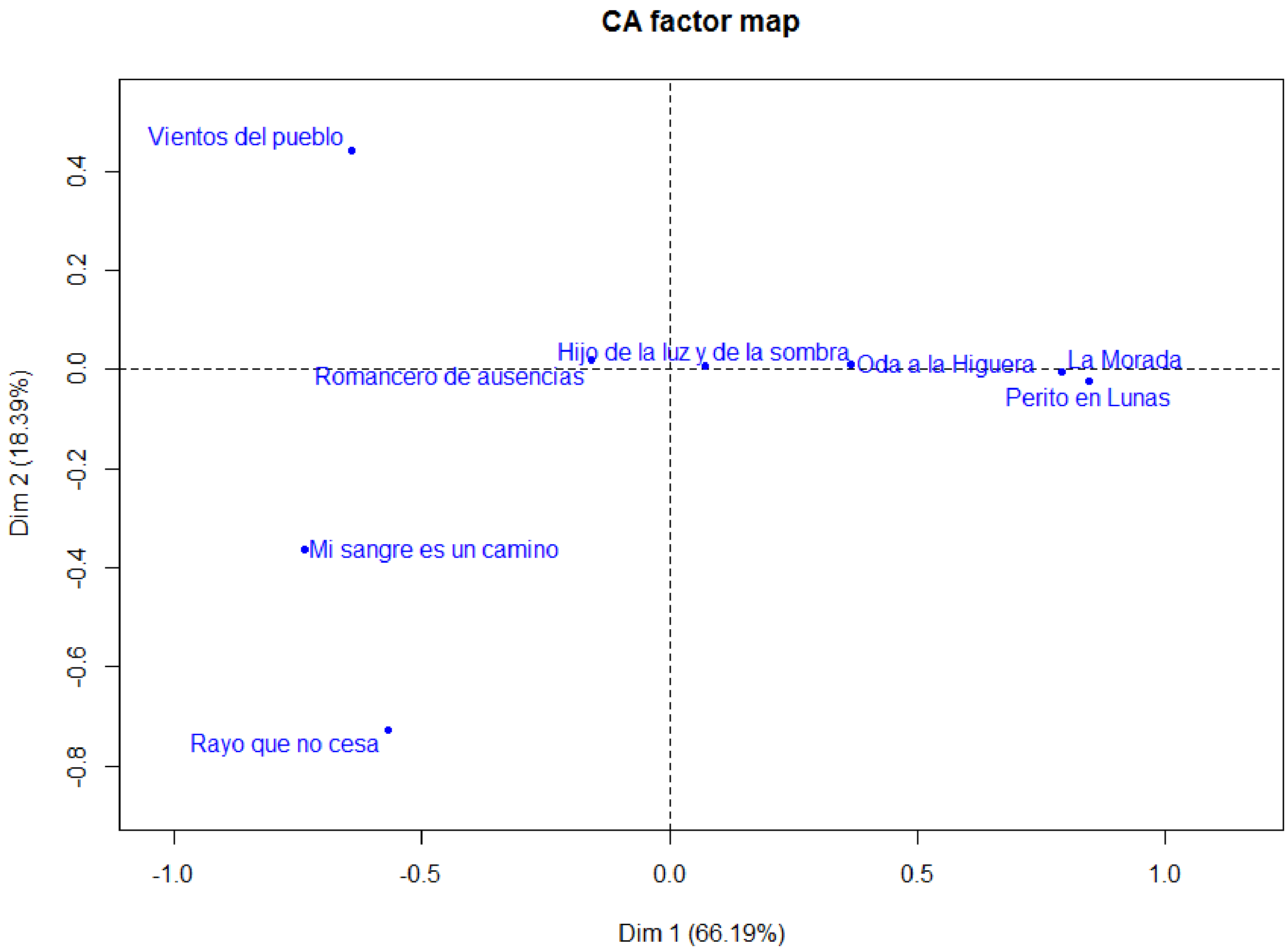
Interpreting the factors: Select the rows with **contribution** > 100/*I*



Quality of representation: $\dfrac{\psi_{i\alpha}^2}{d^2(i,G)}$

The row-profile plot

```
> plot(ca.poems, invisible="col", label="row")
```



Coordinates, contributions and cos2

```
> ca.poems$row$coord
                                Dim 1        Dim 2       Dim 3       Dim 4        Dim 5
La Morada                  0.79095206 -0.004357592  0.14808103  0.14062857 -0.002906936
Perito en Lunas            0.84594367 -0.023857687  0.16315014 -0.35677760  0.238702899
Oda a la Higuera           0.36583833  0.009240982 -0.09822331 -0.03210386 -0.073750687
Rayo que no cesa          -0.56987300 -0.727076551 -0.05749139  0.15920775  0.211226036
Mi sangre es un camino    -0.73836595 -0.363497483  0.36939296 -0.13290371 -0.248593298
Vientos del pueblo        -0.64311037  0.441087724  0.10127366  0.07031227  0.074585269
Romancero de ausencias    -0.15926192  0.018729150 -0.26666464  0.02361306 -0.086118783
Hijo de la luz y de la sombra  0.07171999  0.006877548 -0.04430875 -0.30129916  0.107673614

> ca.poems$row$contrib
                               Dim 1        Dim 2      Dim 3      Dim 4       Dim 5
La Morada                 39.3731794  0.004302201 12.8624028 17.8558655  0.01153666
Perito en Lunas            6.6723571  0.019105154  2.3130999 17.0264918 11.52446345
Oda a la Higuera           7.0713522  0.016242682  4.7509039  0.7812179  6.23397620
Rayo que no cesa           8.2007707 48.056954621  0.7779062  9.1824918 24.44006414
Mi sangre es un camino    12.9198747 11.272370728 30.1380592  6.0051357 31.76894216
Vientos del pueblo        23.9410089 40.543274606  5.5333419  4.1055111  6.98532268
Romancero de ausencias     1.6456046  0.081928540 42.9987247  0.5189668 10.43773629
Hijo de la luz y de la sombra  0.1758524  0.005821469  0.6255614 44.5243194  8.59795841

> ca.poems$row$cos2
                               Dim 1        Dim 2       Dim 3       Dim 4        Dim 5
La Morada                 0.93509099 2.838224e-05 0.032775733 0.029559745 1.263062e-05
Perito en Lunas           0.76532557 6.087234e-04 0.028466774 0.136131461 6.093667e-02
Oda a la Higuera          0.83686143 5.339632e-04 0.060325961 0.006444507 3.401005e-02
Rayo que no cesa          0.35025588 5.701506e-01 0.003564798 0.027337431 4.811984e-02
Mi sangre es un camino    0.61032806 1.479186e-01 0.152755635 0.019773987 6.918291e-02
Vientos del pueblo        0.65738121 3.092404e-01 0.016301939 0.007857935 8.842038e-03
Romancero de ausencias    0.23876418 3.302033e-03 0.669385286 0.005248675 6.981376e-02
Hijo de la luz y de la sombra 0.04498964 4.137131e-04 0.017171613 0.794013459 1.014030e-01
```

Dual CA. Projection of columns?

```
> pca.poems$var$coord
              Dim.1        Dim.2       Dim.3         Dim.4         Dim.5
AMOR     0.12810222 -0.115245396 -0.10825126  0.0357832243 -0.0471857898
CORAZON -0.23249984 -0.162030892  0.01741327  0.0151471730  0.0699090257
HUERTO   0.26286009  0.013513857  0.09837850  0.0871461144  0.0001654327
MUERTE  -0.12190152  0.164097197 -0.01140329  0.0308068294  0.0217025576
SANGRE  -0.21160156 -0.021550954  0.08529634 -0.0426614334 -0.0724381871
LUZ      0.30543200  0.005550967  0.01034790 -0.0958420727  0.0327864271
HOMBRE  -0.09813954  0.129541875 -0.05659498 -0.0003258489  0.0044989609
```

PCA of the row-profiles ≠ CA

```
> ca.poems$col$coord
             Dim 1       Dim 2       Dim 3        Dim 4         Dim 5
AMOR     0.2893465 -0.26030658 -0.24450882  0.080824129 -0.1065792821
CORAZON -0.6591580 -0.45937218  0.04936818  0.042943600  0.1981983863
HUERTO   1.0640028  0.05470127  0.39821565  0.352749294  0.0006696372
MUERTE  -0.2637505  0.35504653 -0.02467257  0.066654751  0.0469564245
SANGRE  -0.5335637 -0.05434179  0.21507890 -0.107572891 -0.1826564325
LUZ      0.7328304  0.01331857  0.02482798 -0.229956207  0.0786652689
HOMBRE  -0.3609055  0.47638678 -0.20812654 -0.001198301  0.0165448084
```

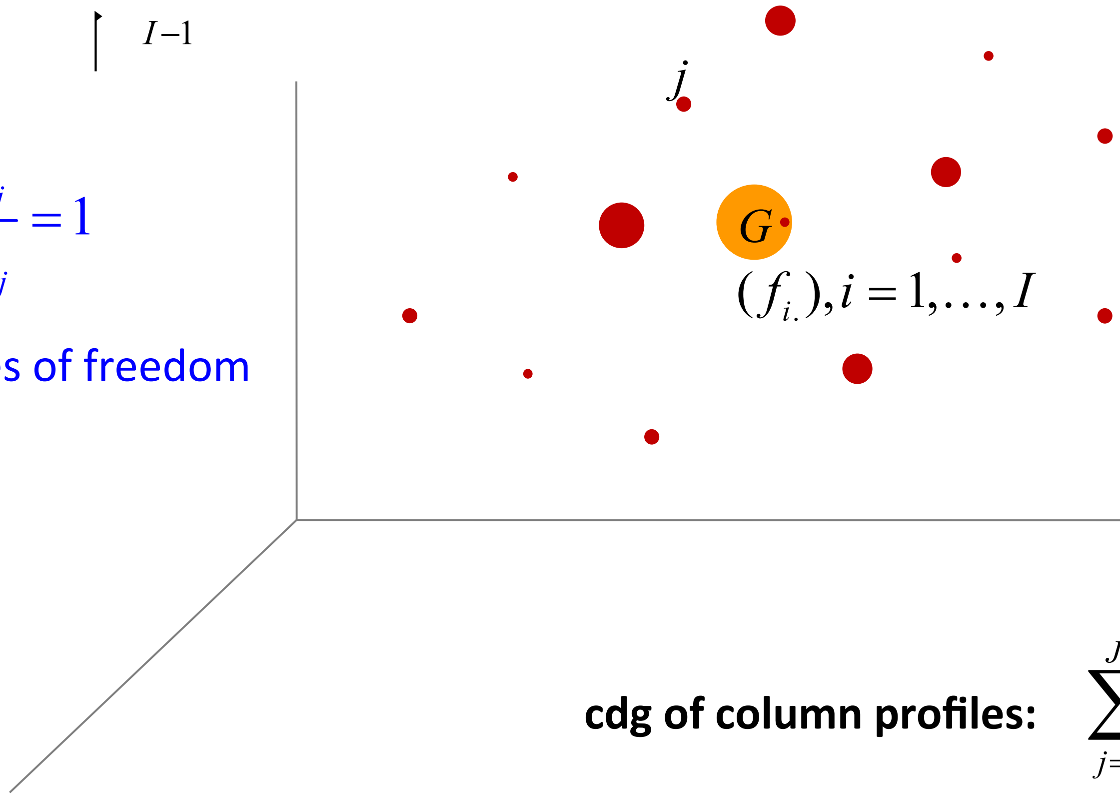
CA in R

CA as a double PCA,
one PCA of the the row-profiles
and another PCA for the colum-profiles

Cloud of column profiles

$$F = \begin{array}{c|c|c} & \begin{matrix} & \vdots & \\ \cdots & f_{ij} & \cdots \\ & \vdots & \end{matrix} & f_{i\cdot} \\ & f_{\cdot j} & \end{array}$$

J column profiles



$$\sum_{j=1}^{J} \frac{f_{ij}}{f_{\cdot j}} = 1$$

Degrees of freedom = *I*-1

Coordinates

$$\left(\frac{f_{ij}}{f_{\cdot j}}\right) i = 1, \ldots, I$$

Weight: $f_{\cdot j}$

$$N = D_J$$

cdg of column profiles: $\sum_{j=1}^{J} \frac{f_{ij}}{f_{\cdot j}} f_{\cdot j} = f_{i\cdot}$

Analysis of column-profiles

Matrix of column-profiles

$$F_{I/J} = FD_J^{-1} = \begin{array}{c} \boxed{\begin{matrix} \vdots \\ \dfrac{f_{ij}}{f_{\cdot j}} \\ \vdots \end{matrix}} \; f_{i\cdot} \\ 1 \end{array}$$

cdg of column profiles

$$d^2(j, j') = \sum_{i=1}^{I} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2$$

Chi-square metric

$$M = \begin{pmatrix} \ddots & & \\ & \dfrac{1}{f_{i\cdot}} & \\ & & \ddots \end{pmatrix}$$

Column-profile formulas

PCA of column-profiles:

Coordinates of row-profiles including the metric: $\left\{ \dfrac{f_{ij}}{\sqrt{f_{i\cdot} f_{\cdot j}}} \right\}, i = 1, \ldots, I$

Centered coordinates: $\left\{ \dfrac{f_{ij} - f_{i\cdot} f_{\cdot j}}{\sqrt{f_{i\cdot} f_{\cdot j}}} \right\}, i = 1, \ldots, I$

Looking for directions v_α of maximal inertia: $Max \sum_{j=1}^{J} f_{\cdot j} \varphi_{j\alpha}^2 \qquad \varphi_{j\alpha} = \sum_{i=1}^{I} \dfrac{f_{ij} - f_{i\cdot} f_{\cdot j}}{\sqrt{f_{i\cdot} f_{\cdot j}}} v_{i\alpha}$

$\rightarrow$ Diagonalize: $\underset{(J,J)}{ZZ'}$

$$Z = \boxed{\begin{matrix} \vdots \\ \sqrt{f_{\cdot j}} \dfrac{f_{ij} - f_{i\cdot} f_{\cdot j}}{\sqrt{f_{i\cdot} f_{\cdot j}}} \\ \vdots \end{matrix}} = \boxed{\begin{matrix} \vdots \\ \dfrac{f_{ij} - f_{i\cdot} f_{\cdot j}}{\sqrt{f_{i\cdot} f_{\cdot j}}} \\ \vdots \end{matrix}}$$

Projecting the column-profiles

$$\varphi_{j\alpha} = \sum_{i=1}^{I} \frac{f_{ij} - f_{i\cdot} f_{\cdot j}}{\sqrt{f_{i\cdot} f_{\cdot j}}} v_{i\alpha} = \sum_{i=1}^{I} \frac{f_{ij}}{\sqrt{f_{i\cdot} f_{\cdot j}}} v_{i\alpha}$$

φ is called a **factor**

$$\sum_{j=1}^{J} f_{\cdot j} \varphi_{j\alpha} = 0$$

φ is centered

$$\sum_{j=1}^{J} f_{\cdot j} \varphi_{j\alpha}^2 = \lambda_\alpha$$

φ inertia is equal to its eigenvalue

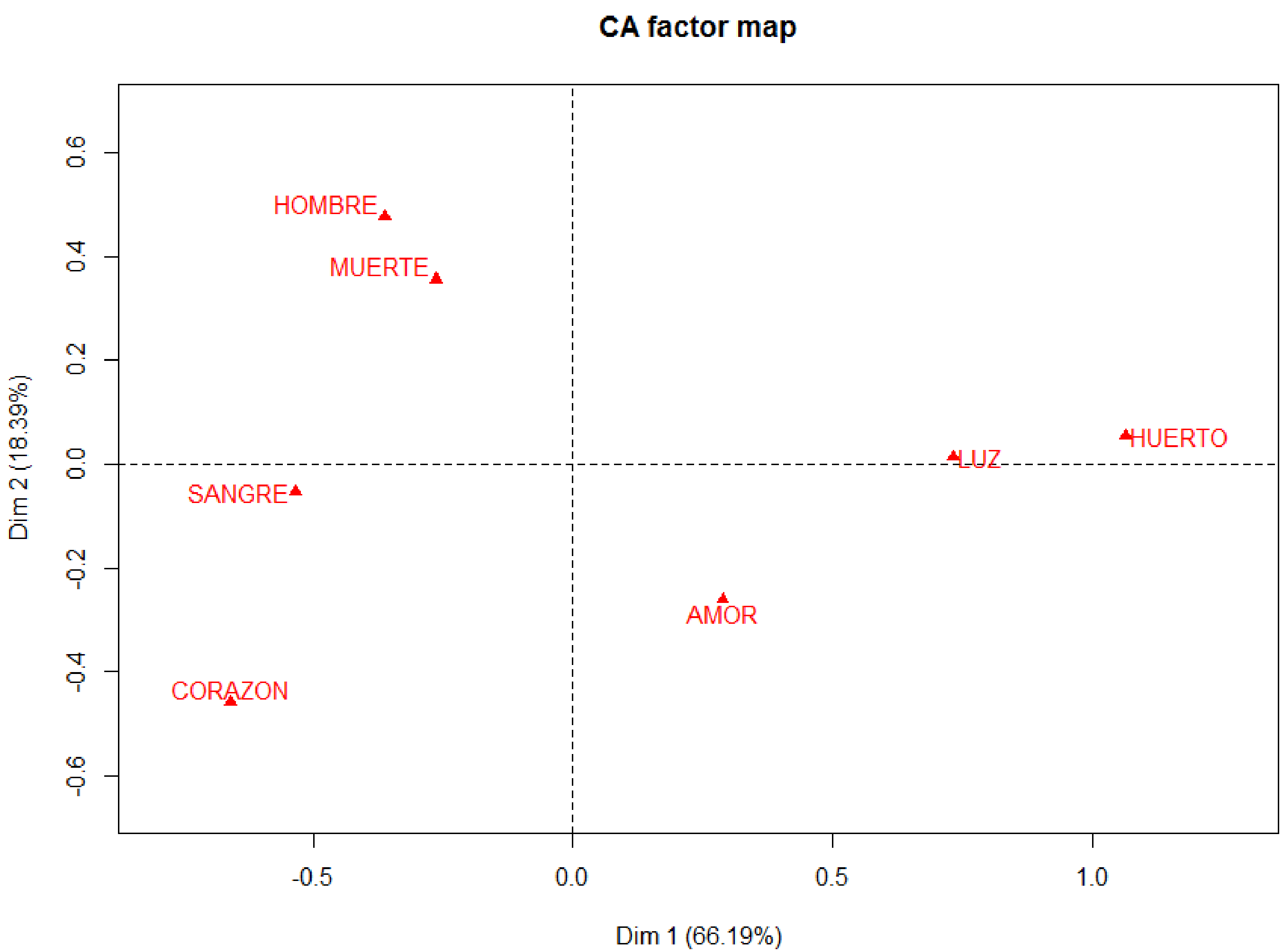
Contribution of a column j to a factor: $\dfrac{f_{\cdot j} \varphi_{j\alpha}^2}{\lambda_\alpha}$

Interpreting the factors: Select the rows with **contribution** > 100/J

Quality of representation: $\dfrac{\varphi_{j\alpha}^2}{d^2(j,G)}$

Plotting the column-profiles

```
plot(ca.poems, invisible="row")
```



Contributions and cos2 of column-profiles

```
> ca.poems$col$coord
             Dim 1       Dim 2       Dim 3        Dim 4         Dim 5
AMOR     0.2893465 -0.26030658 -0.24450882  0.080824129 -0.1065792821
CORAZON -0.6591580 -0.45937218  0.04936818  0.042943600  0.1981983863
HUERTO   1.0640028  0.05470127  0.39821565  0.352749294  0.0006696372
MUERTE  -0.2637505  0.35504653 -0.02467257  0.066654751  0.0469564245
SANGRE  -0.5335637 -0.05434179  0.21507890 -0.107572891 -0.1826564325
LUZ      0.7328304  0.01331857  0.02482798 -0.229956207  0.0786652689
HOMBRE  -0.3609055  0.47638678 -0.20812654 -0.001198301  0.0165448084

> ca.poems$col$contrib
            Dim 1       Dim 2      Dim 3        Dim 4        Dim 5
AMOR     5.431729 15.82593504 36.1504544 6.080192e+00 1.598660e+01
CORAZON 17.892461 31.28368150  0.9354235 1.089486e+00 3.509136e+01
HUERTO  22.870416  0.21761099 29.8571411 3.606238e+01 1.965063e-04
MUERTE   4.918617 32.08666169  0.4011516 4.506635e+00 3.381859e+00
SANGRE  14.820486  0.55342048 22.4444323 8.642298e+00 3.767635e+01
LUZ     30.878332  0.03671635  0.3303335 4.361850e+01 7.718304e+00
HOMBRE   3.187960 19.99597396  9.8810637 5.041863e-04 1.453310e-01

> ca.poems$col$cos2
            Dim 1        Dim 2       Dim 3        Dim 4        Dim 5
AMOR    0.3651205 0.2955085779 0.260728711 2.848931e-02 4.953879e-02
CORAZON 0.6302609 0.3061051898 0.003535375 2.675089e-03 5.698251e-02
HUERTO  0.7969773 0.0021064698 0.111634135 8.759769e-02 3.156745e-07
MUERTE  0.3351842 0.6073900900 0.002933094 2.140718e-02 1.062398e-02
SANGRE  0.7514328 0.0077944495 0.122099316 3.054382e-02 8.806180e-02
LUZ     0.8997002 0.0002971707 0.001032698 8.858928e-02 1.036709e-02
HOMBRE  0.3066376 0.5342661633 0.101974876 3.380415e-06 6.444092e-04
```

Important property: Pseudo-barycentric formulae

Transition relationships

Diagonalization in R^J and R^I

$$R^J \quad Z^{*\prime} Z^* u_\alpha = \lambda_\alpha u_\alpha$$

$$R^I \quad Z^* Z^{*\prime} v_\alpha = \lambda_\alpha v_\alpha$$

$$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} Z^{*\prime} v_\alpha$$

$$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} Z^* u_\alpha$$

$$u_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{I} \frac{f_{ij}}{\sqrt{f_{i\cdot} f_{\cdot j}}} v_{\alpha i}$$

$$v_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{J} \frac{f_{ij}}{\sqrt{f_{i\cdot} f_{\cdot j}}} u_{\alpha j}$$

Projection of row-profiles and column-profiles

$$\psi_\alpha = X^* u_\alpha$$

$$\varphi_\alpha = X^{*\prime} v_\alpha$$

$$\psi_{\alpha i} = \sum_{j=1}^{J} \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} u_{\alpha j}$$

$$\varphi_{\alpha j} = \sum_{i=1}^{I} \frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} v_{\alpha i}$$

Direct formulae

$$\psi_{\alpha i} = \sqrt{\lambda_\alpha} \frac{1}{\sqrt{f_{i\cdot}}} v_{\alpha i}$$

$$\varphi_{\alpha j} = \sqrt{\lambda_\alpha} \frac{1}{\sqrt{f_{\cdot j}}} u_{\alpha j}$$

Indirect formulae

Pseudo-barycentric relantionships:

$$\psi_{i\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{J} \frac{f_{ij}}{f_{i\cdot}} \varphi_{j\alpha}$$

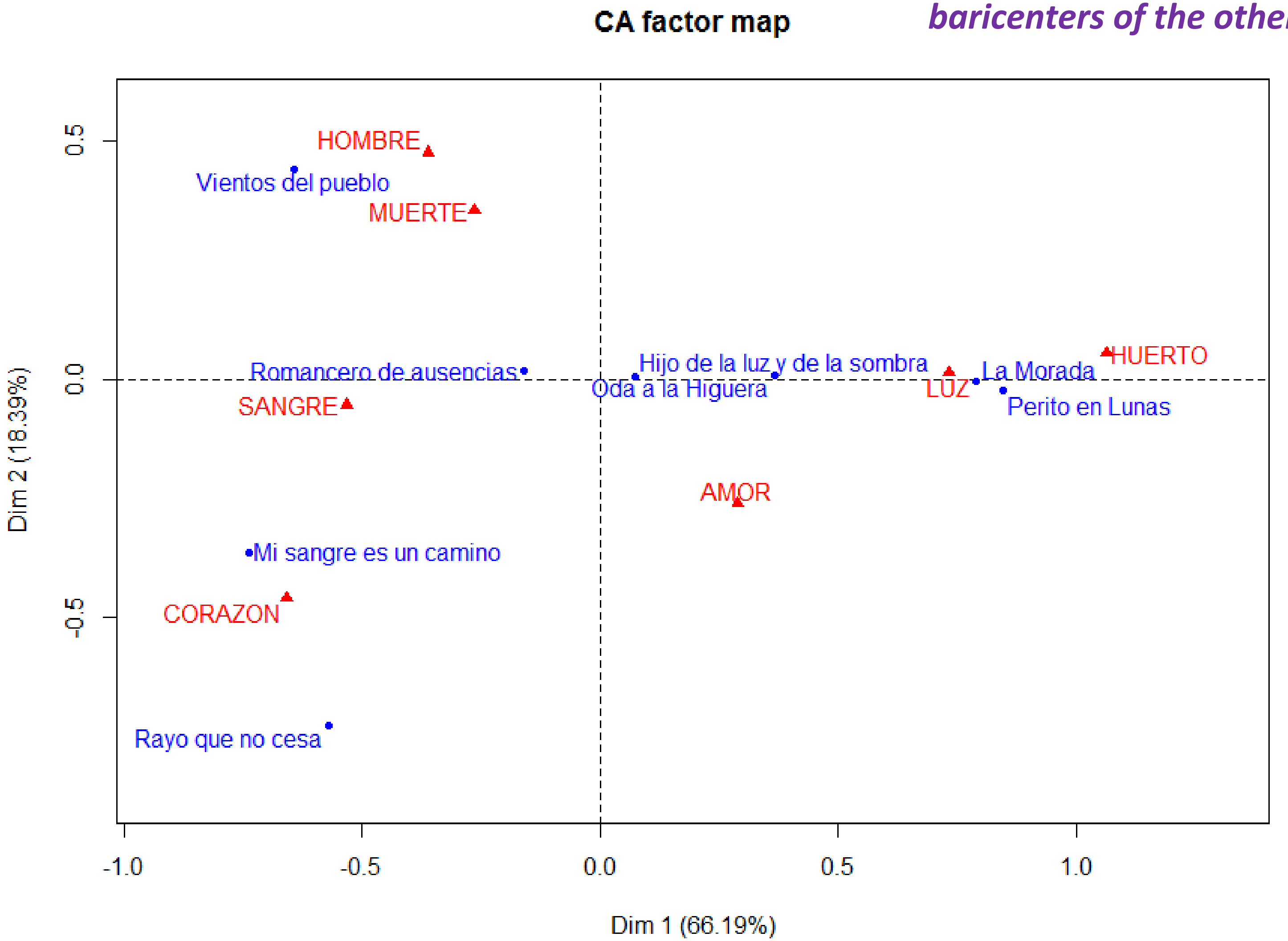
$$\varphi_{j\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{\cdot j}} \psi_{i\alpha}$$

Enlargement coefficient

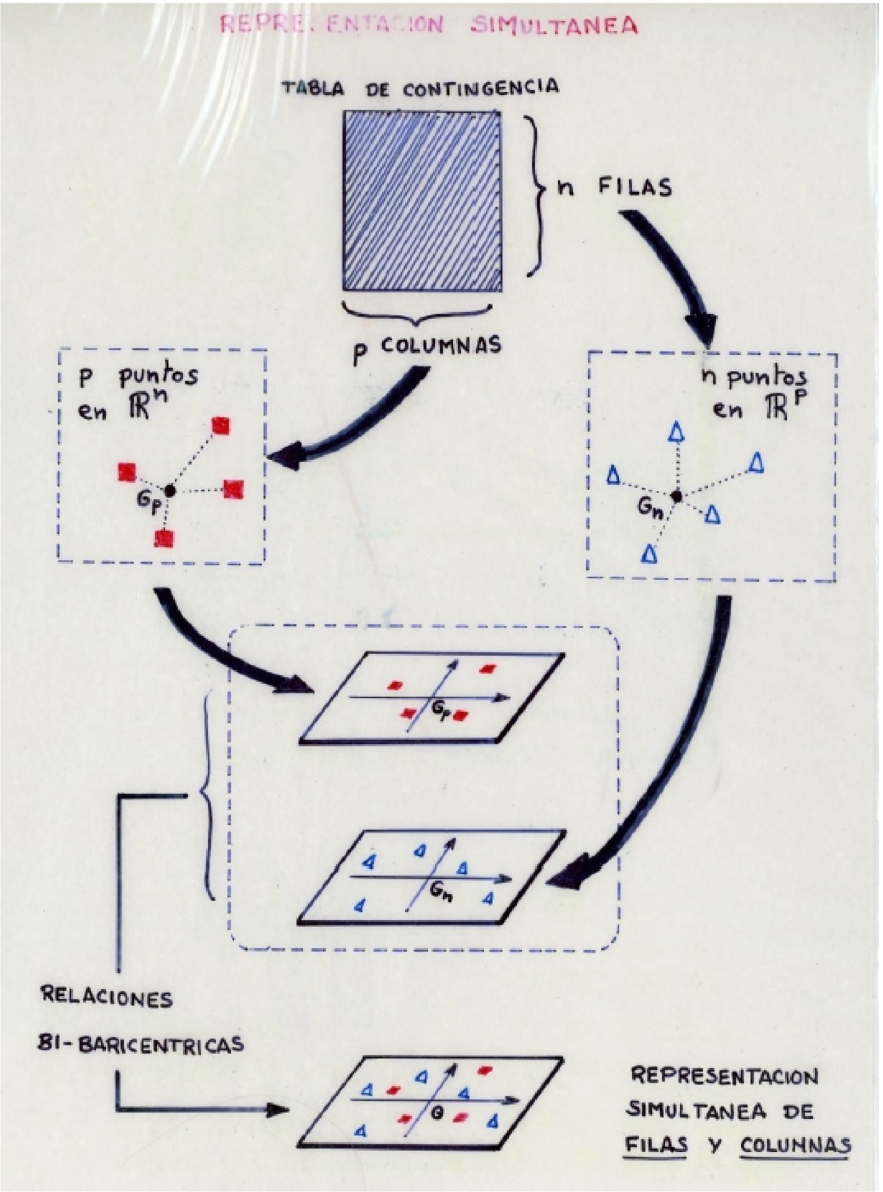
Joint representation in CA

```
> plot(ca.poems)
```

Points interpreted as pseudo-baricenters of the other set



Joint display in CA



Conditions to apply CA

Applying CA means:

- All cells should have positive numbers
- It makes sense adding rows and colums.
- It makes sense to take this sums as weights for rows and columns.
- It makes sense to compare rows from the row-profiles and columns from the column-profiles.
- It makes sense to use the Chi-square metric either for row and column profiles comparison.

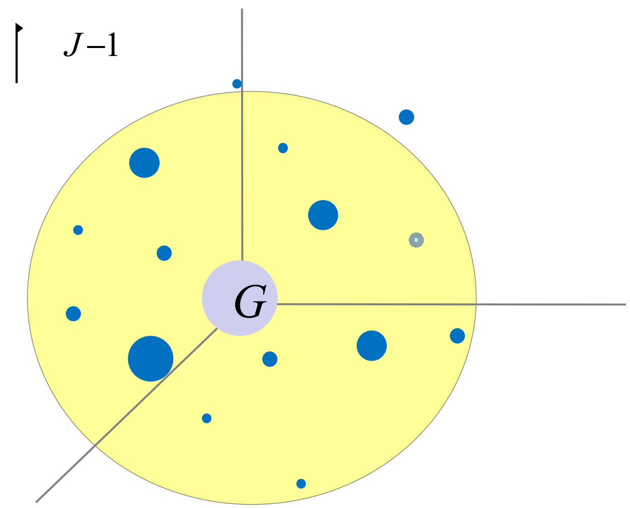
CA and independence of rows and columns

The Chi-square statistic and the Inertia

$$\chi^2_{(I-1)(J-1)} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - {n_{i\cdot}n_{\cdot j}}/{n}\right)^2}{{n_{i\cdot}n_{\cdot j}}/{n}} = n\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{f_{ij} - f_{i\cdot}f_{\cdot j}}{\sqrt{f_{i\cdot}f_{\cdot j}}}\right)^2 = n \cdot I_I$$

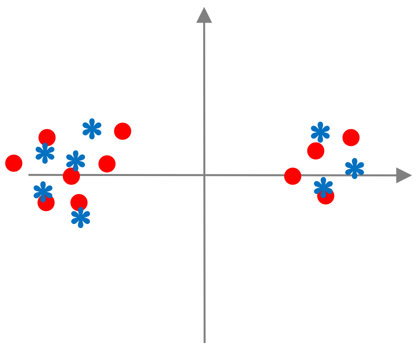
Independence and the cloud of points

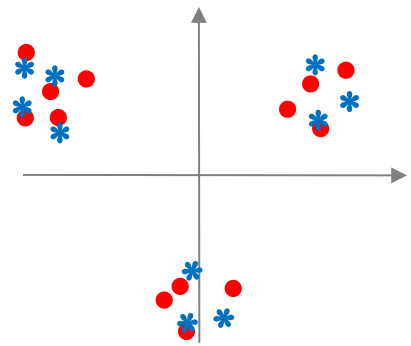
$$f_{ij} = f_{i\cdot}f_{\cdot j} \quad \forall i, j \quad \Rightarrow \quad \frac{f_{ij}}{f_{i\cdot}} = f_{\cdot j}$$

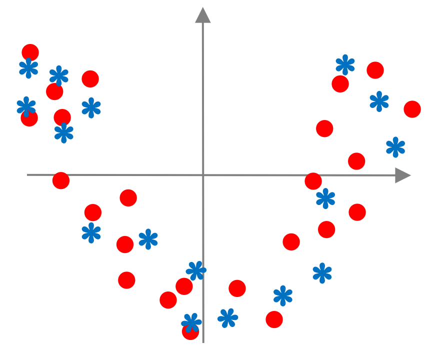


All row-profiles equal to the centroid

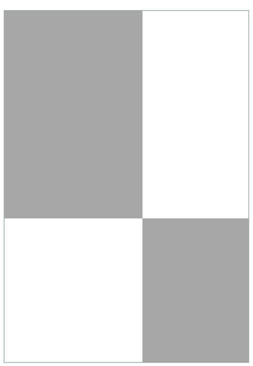
The Guttman effect



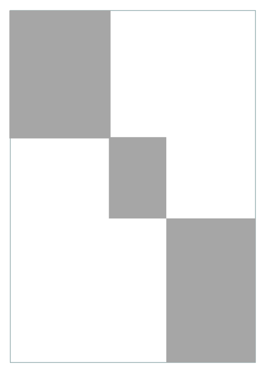




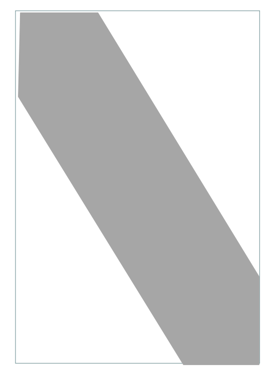
Guttman effect



Two disjoint groups



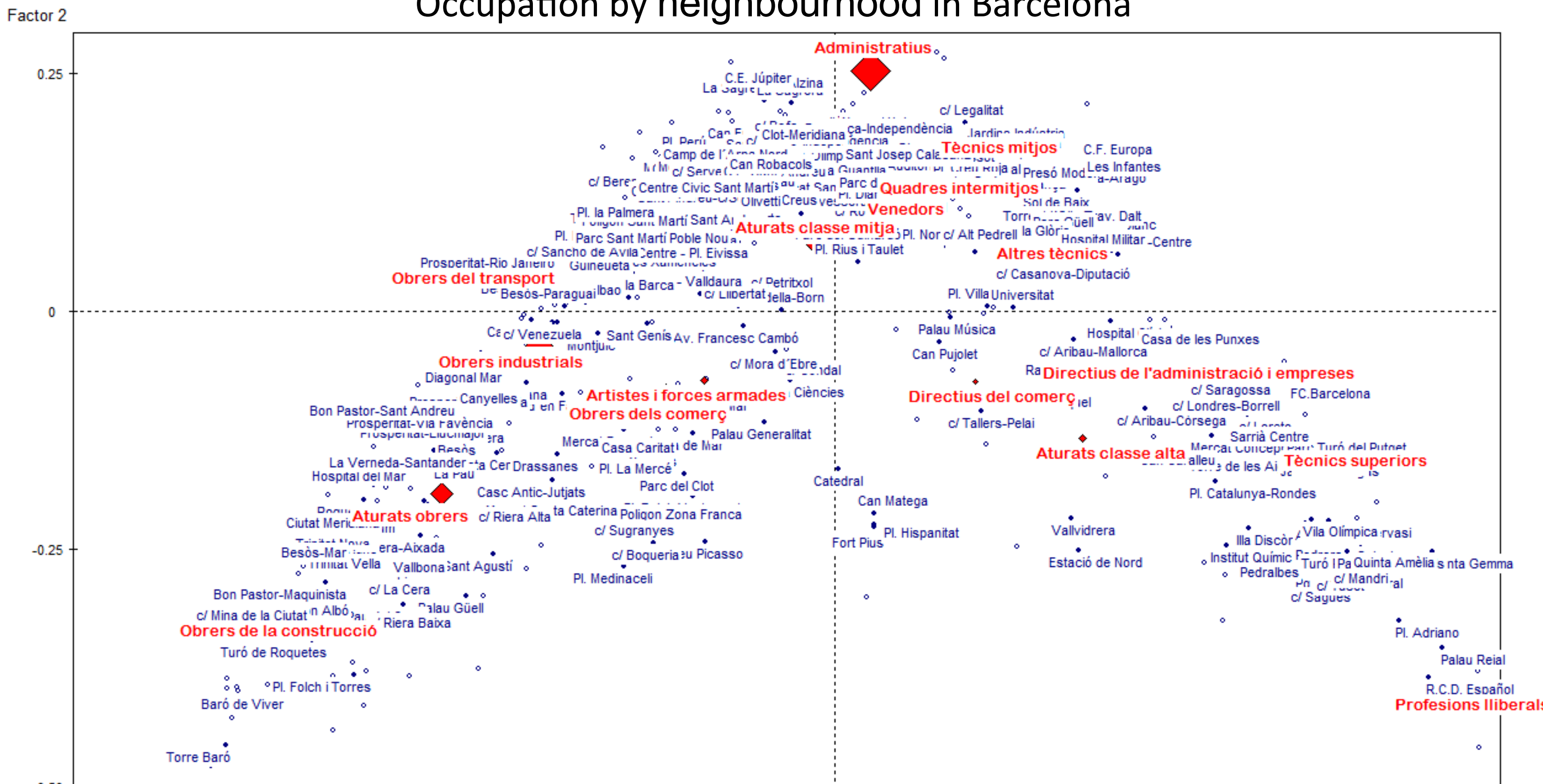
Three disjoint groups



Loaded diagonal

The Guttman effect

Occupation by neighbourhood in Barcelona



Efecto Guttman

El efecto Guttman, también denominado de parábola ocurre por lo general en presencia de variables ordinales, las cuales generalmente se ilustran en tablas de contingencia con la intención de identificar la relación existente entre ellas, sin embargo, este efecto indica la presencia de redundancia entre las categorías de las dos variables analizadas, esto quiere decir, que teniendo conocimiento de la fila en la categoría i es posible predecir o deducir el comportamiento de la columna en la posición j y viceversa.

Projection of illustrative variables

- **Categorical variables**:

  Every modality of the categorical variables is represented as the centroid of the rows having chosen that modality.

- **Continuous vars**:

  We display the correlation of the illustrative variables with the significant factorial axes.

$$\mathrm{cor}(X_{supl}, \psi)$$

Active data

Suppl. vars.

Summary (step by step)

1) Independence Test (Chi-squared Test)

2) Dimensionality Reduction (Max Num of Dimensions = Min(N-1, M-1))

3) Keep two or three dimensions as done in PCA

4) Perform analysis on rows and columns.

5) Use graphs

6) Obtain remarkable conclusions (highlights)

Quick Example

Se desea determinar si la preferencia por el billete cien mil varía dependiendo de estrato socioeconómico del individuo. Para ello se ha tomado una muestra aleatoria de 1000 individuos obteniendo los resultados que se encuentran dispuestos en la siguiente tabla:

```
#Visualización de los datos
datos_billete_cien
```

```
##             Mucho Poco Nada
## Estratos1-2    50   50  300
## Estratos3-4    60  240  100
## Estratos5-6    80   50   70
```

```
#Prueba de independencia.
chisq.test(datos_billete_cien)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  datos_billete_cien
## X-squared = 311.4, df = 4, p-value < 2.2e-16
```

```
#Análisis de correspondencia simple
ACS <- CA(datos_billete_cien, graph = FALSE)
```

Porcentaje de varianza explicado

Con la intención de identificar el número de componentes a utilizar se realiza el estudio del porcentaje de varianza explicado por ejes, es importante mencionar que el número de dimensiones está asociado con la cantidad de columnas, ya que por lo geenral siempre es menor que el número de filas; teniendo en cuenta que dentro del análisis una de las columnas termina siendo combinación lineal de las demás se obtendrán en total p-1 dimensiones y en este caso, son 2 ya que el número total de columnas es 3.

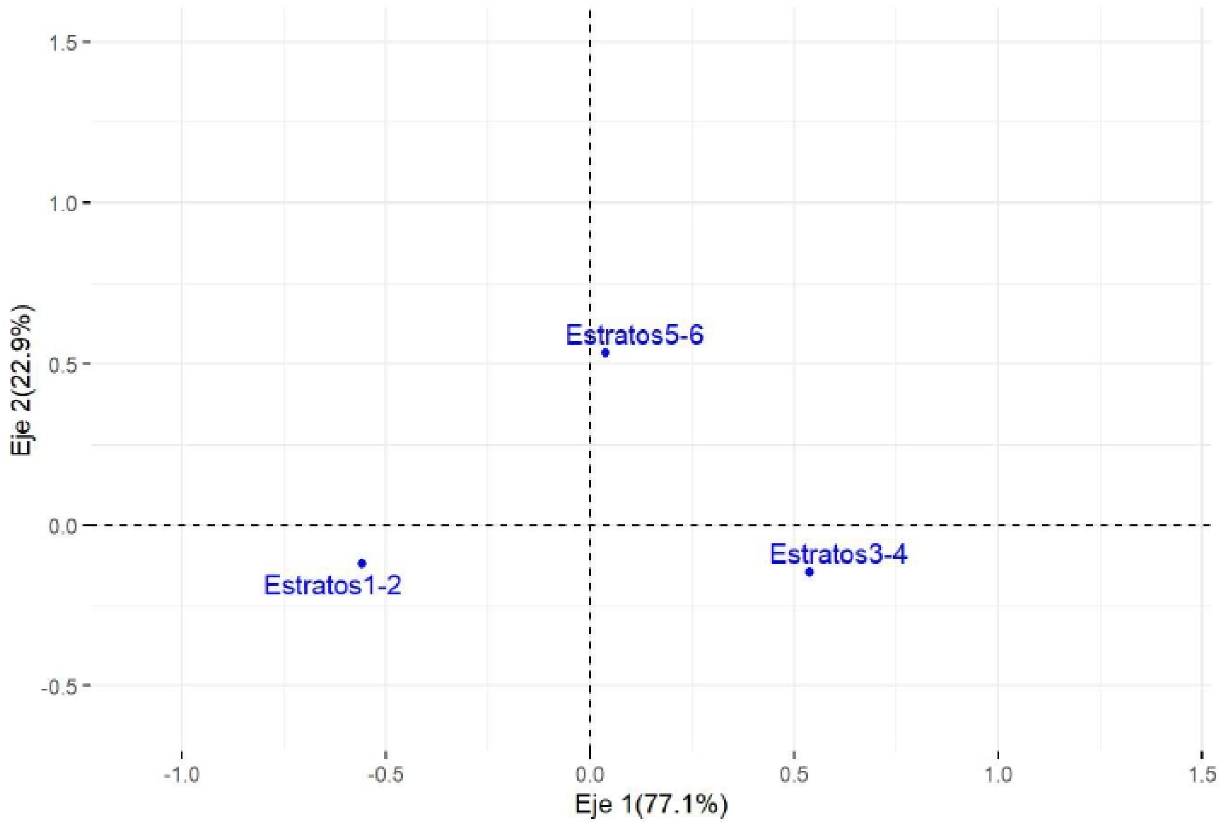
```
#% de varianza explicado
  valores_propios=ACS$eig; valores_propios
```

```
##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1 0.24002086               77.07719                          77.07719
## dim 2 0.07138237               22.92281                         100.00000
```

Análisis puntos fila

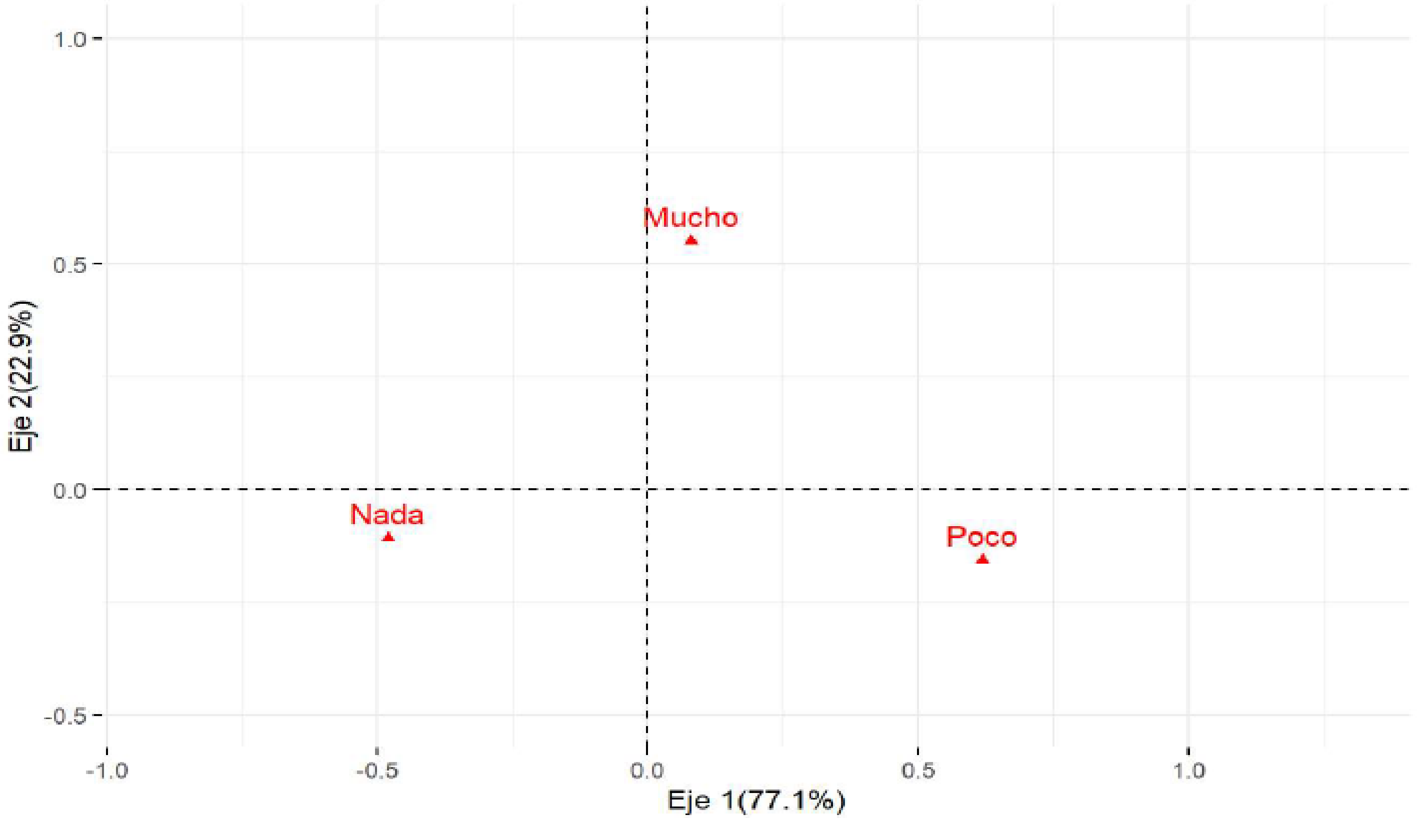
```
#Perfiles fila
variables_fila=get_ca_row(ACS)

#Nube de individuos fila
fviz_ca_row(ACS, repel = TRUE)+ggtitle("") + ylab("Eje 2(22.9%)")+xlab("Eje 1(77.1%)")+ylim(-0.
6,1.5)+xlim(-1.1,1.4)
```



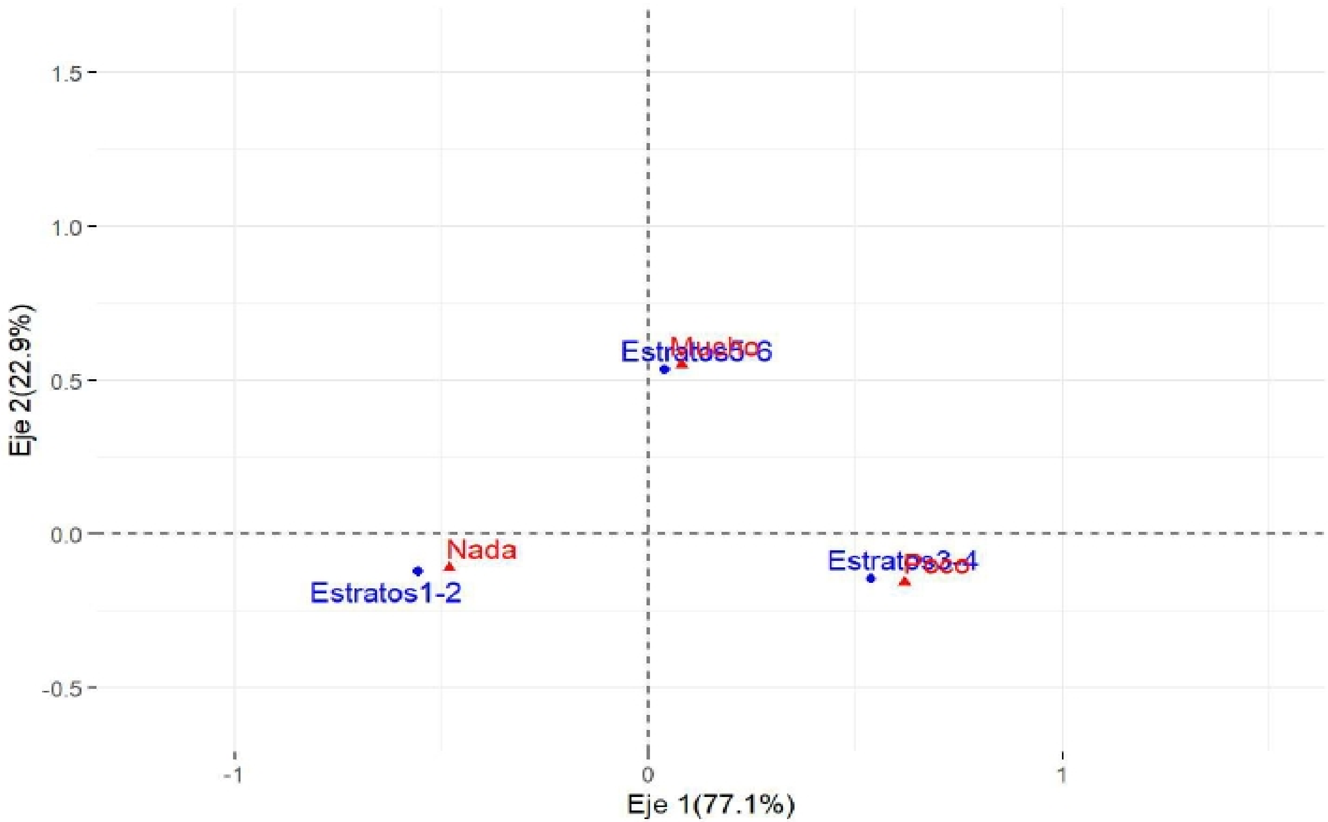
Análisis puntos columna

```
#Perfiles columna
variables_columna=get_ca_col(ACS)
#Nube de individuos columna
fviz_ca_col(ACS)+ggtitle("")+ylab("Eje 2(22.9%)")+xlab("Eje 1(77.1%)")+ylim(-0.5,1)+xlim(-0.9,1.
3)
```



Representación simultánea

```
#Representación simultánea
fviz_ca_biplot(ACS, repel = TRUE)+ggtitle("")+ylab("Eje 2(22.9%)")+xlab("Eje 1(77.1%)")+ylim(-0.
6,1.6)+xlim(-1.2,1.5)
```



Homeworks

Lab Session adapted from:

https://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/120-correspondence-analysis-theory-and-practice/#google_vignette

Lab Session:

https://ramia-lab.github.io/AdvancedModelling/material/04_CorrespondenceAnalysis/laboratorio/ACS.html

Quick Script (Cheat sheet)

https://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/113-ca-correspondence-analysis-in-r-essentials/