FIB

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Artificial Intelligence Degree

PMAAD

# Multiple Correspondence Analysis (MCA)

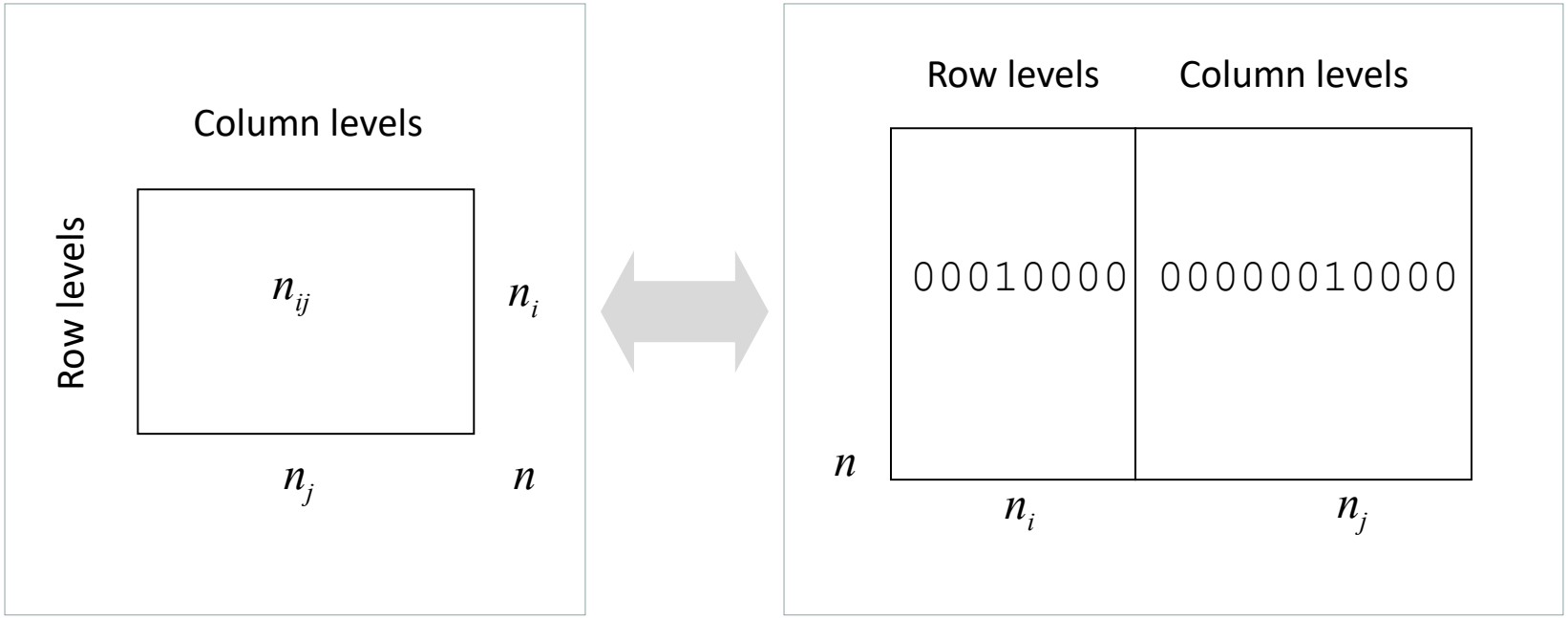
**Prof. Dante Conti / Sergi Ramírez**

**Credits – Prof. Tomàs Aluja**

# CA generalization: Multiple Correspondence Analysis

Ludovic Lebart
French statistician, promoter of MCA

Count table: = Individuals × (categorical) variables table:



# CA generalization: Multiple Correspondence Analysis

DISJUNCTIVE TABLE & BURT TABLE

| Individuos | Género | Años | Ingreso |
|---|---|---|---|
| 1 | Mujer | 5 | Medio |
| 2 | Mujer | 3 | Alto |
| 3 | Hombre | 4 | Bajo |
| 4 | Mujer | 1 | Bajo |
| 5 | Mujer | 2 | Medio |
| 6 | Hombre | 5 | Alto |
| 7 | Mujer | 2 | Medio |
| 8 | Hombre | 3 | Bajo |
| 9 | Hombre | 1 | Alto |
| 10 | Mujer | 4 | Medio |

3 CATEGORICAL VARIABLES WITH A TOTAL NUMBER OF MODALITIES = 10 (SEX(2); YEARS(5) & INCOME(3))

# CA generalization: Multiple Correspondence Analysis

DISJUNCTIVE TABLE (Z)

| Género | | Años | | | | | Ingresos | | |
|---|---|---|---|---|---|---|---|---|---|
| Mujer | Hombre | 1 | 2 | 3 | 4 | 5 | Bajo | Medio | Alto |
| 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |

# CA generalization: Multiple Correspondence Analysis

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

**BURT TABLE**

$Z =$

| | | Género | | Años | | | | | Ingresos | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | H | 1 | 2 | 3 | 4 | 5 | B | M | A |
| → Género → | M | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| | H | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| Años ↦ | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| | 3 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| | 5 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| | B | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| Ingresos ↦ | M | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | A | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |

$Z' =$

| | | Género | | Años | | | | | Ingresos | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | H | 1 | 2 | 3 | 4 | 5 | B | M | A |
| → Género → | M | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |
| | H | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| Años ↦ | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| | 4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| | B | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| Ingresos ↦ | M | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| | A | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |

$$B = Z'.Z$$

# CA generalization: Multiple Correspondence Analysis

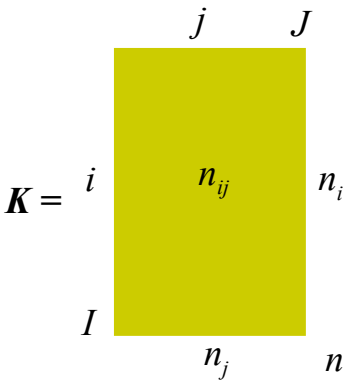
BURT TABLE

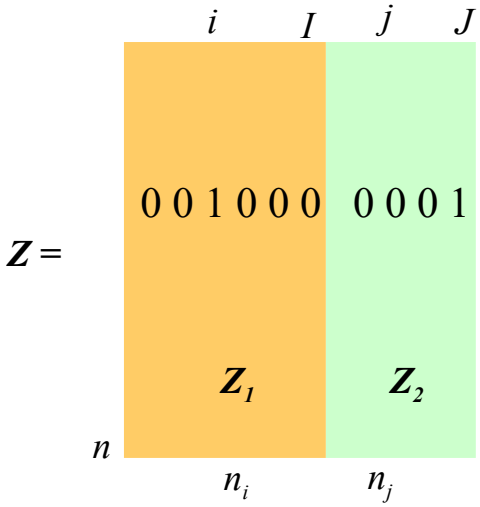
MATRIZ DE BURT

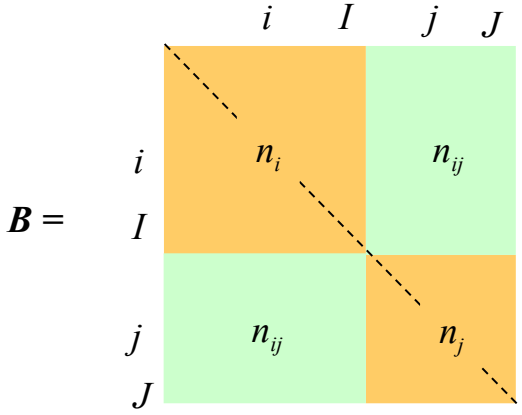
$Z'.Z$ =

| | | Género | | Años | | | | | Ingresos | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | H | 1 | 2 | 3 | 4 | 5 | B | M | A |
| Género | M | 6 | 0 | 1 | 2 | 1 | 1 | 1 | 1 | 4 | 1 |
| | H | 0 | 4 | 1 | 0 | 1 | 1 | 1 | 2 | 0 | 2 |
| Años | 1 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| | 2 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 0 |
| | 3 | 1 | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 1 |
| | 4 | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 1 | 0 |
| | 5 | 1 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 1 |
| Ingresos | B | 1 | 2 | 1 | 0 | 1 | 1 | 0 | 3 | 0 | 0 |
| | M | 4 | 0 | 0 | 2 | 0 | 1 | 1 | 0 | 4 | 0 |
| | A | 1 | 2 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 3 |

# From CA to Multiple Correspondence Analysis



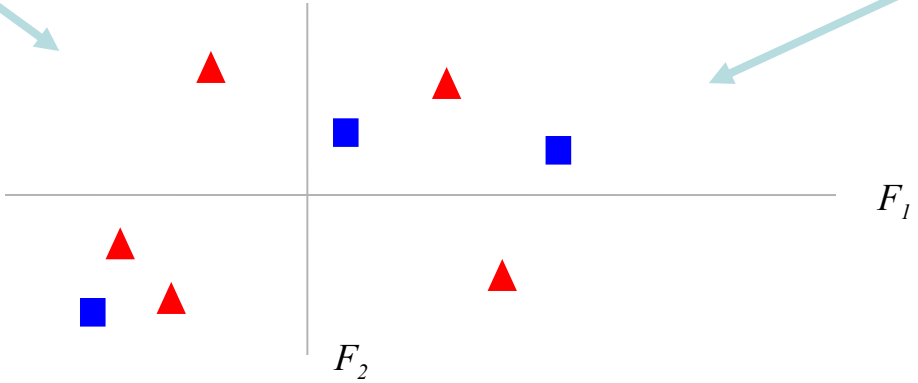

Burt table

$$K = Z_1'Z_2 \longrightarrow \lambda_K$$

$$Z \longrightarrow \lambda_Z = \frac{1+\sqrt{\lambda_K}}{2}$$

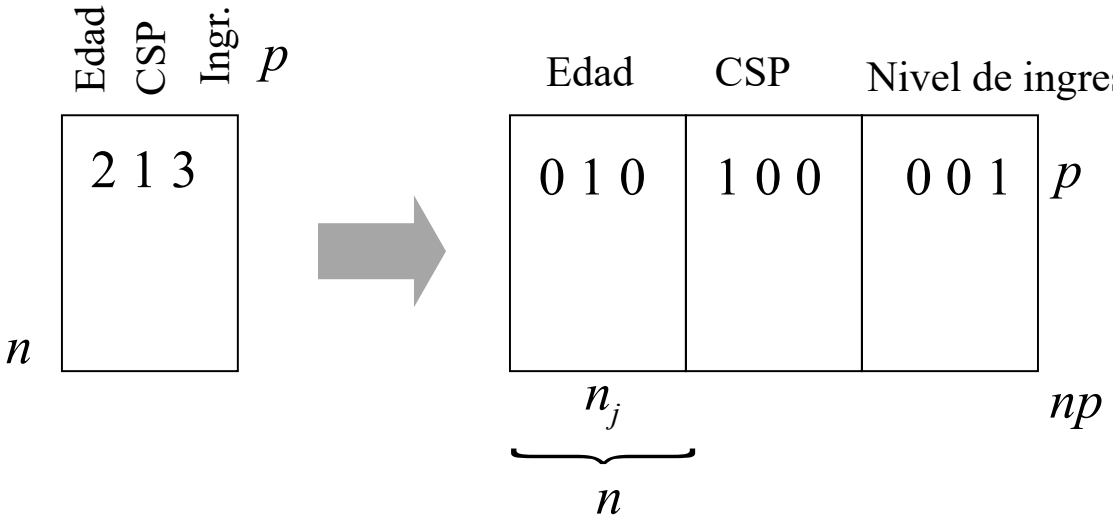
$$B = Z'Z \longrightarrow \lambda_B = \lambda_Z{}^2$$

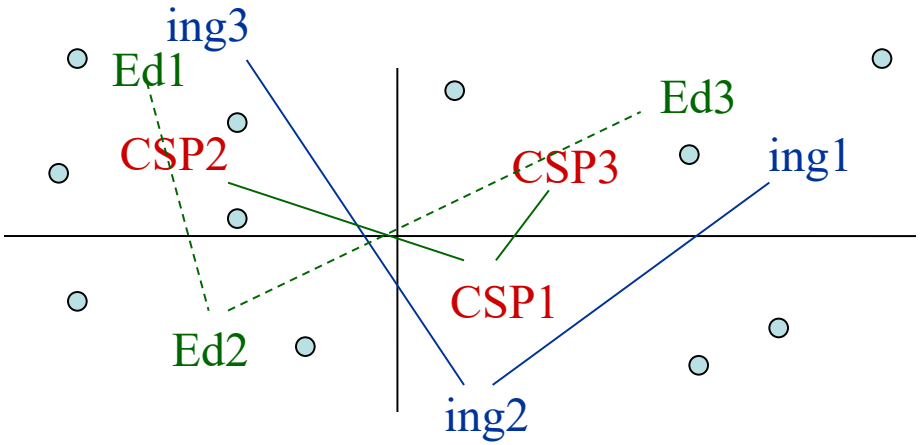
same display



# *Multiple Correspondence Analysis (p>2)*

Tables of the form: individuals × *categorical* variables



# *Multiple Correspondence Analysis (p>2)*

Tables of the form: individuals × *categorical* variables

Información de partida
Los datos observados, de los que parte el ACM, deben disponerse en una tabla en la que las filas representen los N individuos analizados, y las columnas, las p variables de interés. Cada celda de la tabla, producto del cruce de una fila y una columna, contiene la categoría que presenta el individuo correspondiente a esa fila en la variable recogida en esa columna. Las categorías de cada variable deben ser exhaustivas y mutuamente excluyentes.

A partir de esa tabla inicial, el ACM comienza elaborando otra, conocida como la tabla disyuntiva completa, que contiene la misma información que la anterior. Sin embargo, ahora las columnas corresponden a las categorías (y no a las variables) y el contenido de la tabla está codificado mediante unos y ceros, dependiendo de si el individuo de la fila presenta o no, respectivamente, la categoría de la columna. Posteriormente, se aplica a esta nueva tabla un análisis de correspondencias simple (ACS). Este análisis de correspondencias simple constituye la esencia del ACM.

## GETTING INFORMATION FROM MCA (VARIABLES)

(1) La distancia de una categoría al centro de la nube de puntos de las categorías será mayor cuanto menor sea el número de individuos que presenten esa categoría.

(2) La distancia entre dos categorías de dos variables distintas es pequeña cuando la mayoría de los individuos que presentan una de ellas presenta también la otra.

(3) La distancia entre dos categorías de la misma variable es pequeña cuando los individuos que presentan una y los que presentan la otra –individuos que, necesariamente, han de ser diferentes entre sí– tienden a presentar categorías idénticas en el resto de las variables.

(4) Si las posiciones respecto el origen de dos categorías de variables distintas son perpendiculares entre sí, ambas categorías serán independientes.

## GETTING INFORMATION FROM MCA (INDIVIDUALS)

(1) Los individuos con categorías poco frecuentes tenderán a estar lejos del centro de la nube de puntos de los individuos.

(2) Dos individuos con bastantes categorías comunes se encontrarán a poca distancia entre sí.

(3) Un individuo con una categoría poco frecuente se encontrará lejos de los individuos que no la posean.

(4) Dos individuos que compartan una categoría poco frecuente estarán a poca distancia.

(5) En general, estarán más cerca entre sí dos individuos que compartan una categoría poco frecuente que dos que compartan una categoría bastante frecuente.

PSEUDO-ALGORITHM

Los ejes en cada nube se seleccionan uno a uno, de manera secuencial.

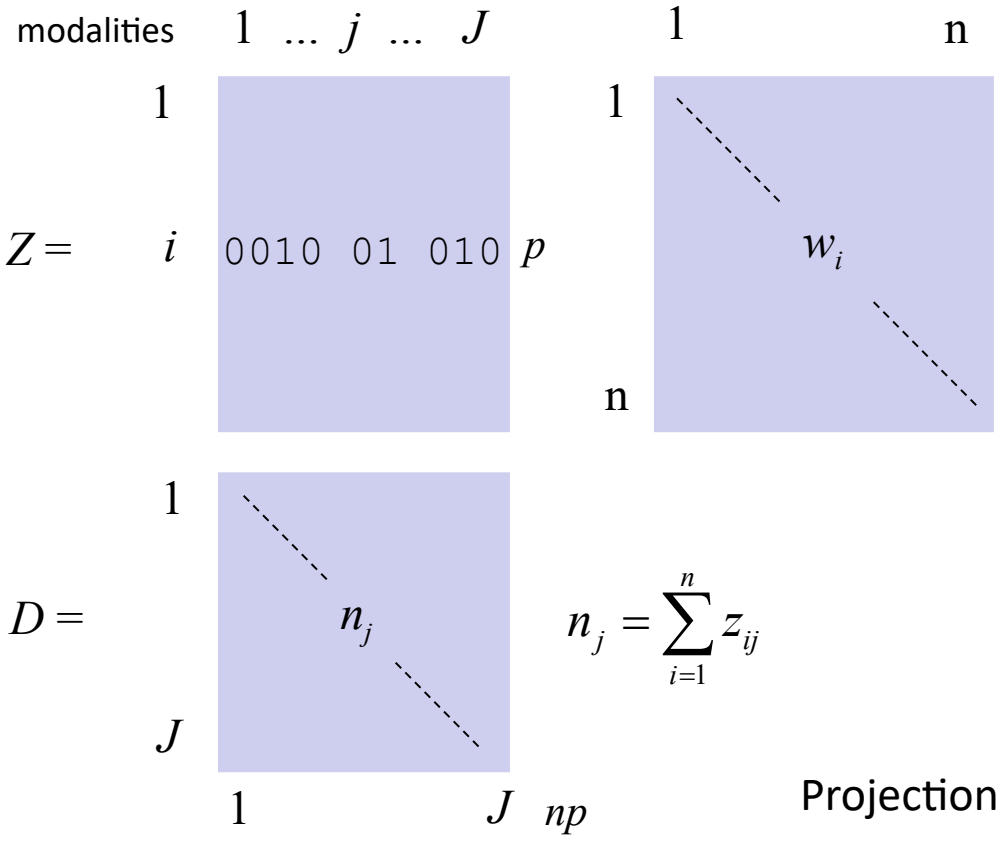
El primer eje que se obtiene en cada nube es tal que la proyección de la nube sobre él es la que más inercia recoge.

Cada uno de los ejes siguientes se escoge de forma que, siendo perpendicular a los anteriores, la proyección de la nube sobre él sea la que recoge más inercia de la nube original.

El número total de ejes seleccionados es K−p en cada nube. Cada eje, solo puede recoger, como máximo, un 100(p/(K−p))% de la inercia total de la nube correspondiente. La inercia recogida por el eje i-ésimo de una nube (inercia del eje) es igual a la recogida por el i-ésimo de la otra nube.

En la práctica, las dimensiones que usualmente se emplean para formar los planos sobre los que proyectar los puntos, son aquellas cuya inercia supera la inercia media por dimensión, ((K−p)/p)/(K−p)=1/p

# *MCA is equal to the CA of the indicator table of categ. variables Z (Note: J is the same as K)*




$$n_j = \sum_{i=1}^{n} z_{ij}$$

Row profile: $= \frac{1}{p} Z$

Centroid of row profiles: $= \left( \cdots, {n_j}/{np}, \cdots \right)$

$$w_i = \frac{p}{np} = \frac{1}{n}$$

Chi-square Metric: $\left( \frac{D}{np} \right)^{-1}$

Projection of row-profile: $\psi = n\, Z D^{-1} u$

explicit metric

$$\underset{u}{Max} \left( \sum_{i=1}^{n} w_i \psi_i^2 = \frac{1}{n} \psi'\psi = n\, u' D^{-1} Z' Z D^{-1} u \right)$$

$$np\, u' D^{-1} u = 1$$

$$\Rightarrow \quad \text{eig}\, \frac{1}{p} Z' Z D^{-1} u = \lambda u$$

# *Properties of MCA*

**Distance to the centroid of one modality:**

$$d^2_{Chi2}(j,G) = \sum_{i=1}^{n} \frac{1}{1/n}\left(\frac{z_{ij}}{n_j} - \frac{1}{n}\right)^2 = \cdots = \left(\frac{n}{n_j} - 1\right)$$

Distance to the centroide increases in rare modalities

**Inertia of one modality:**

$$I_j = \frac{n_j}{np} d^2_{Chi2}(j,G) = \frac{n_j}{np}\left(\frac{n}{n_j} - 1\right) = \frac{1}{p}\left(1 - \frac{n_j}{n}\right)$$

Inertia of one modality increases as the frequency of the modality decreases

**Inertia of one categorical variable:**

$$I_{j\in \text{var}_q} = \sum_{j=1}^{J_q} \frac{1}{p}\left(1 - \frac{n_j}{n}\right) = \frac{1}{p}\left(J_q - 1\right)$$

Inertia of one categ. variable increases as its modalities increases

**Inertia of all modalities:**

$$I_J = \sum_{j=1}^{J} \frac{1}{p}\left(1 - \frac{n_j}{n}\right) = \frac{J}{p} - 1$$

Average eigenvalue

$$\bar{\lambda}_{MCA} = \frac{1}{p}$$

Total inertia just depends on the ratio of the modalities per categorical variable

⊗ Avoid rare modalities
⊗ Avoid high unbalancing in the number of modalities per variable (if possible)

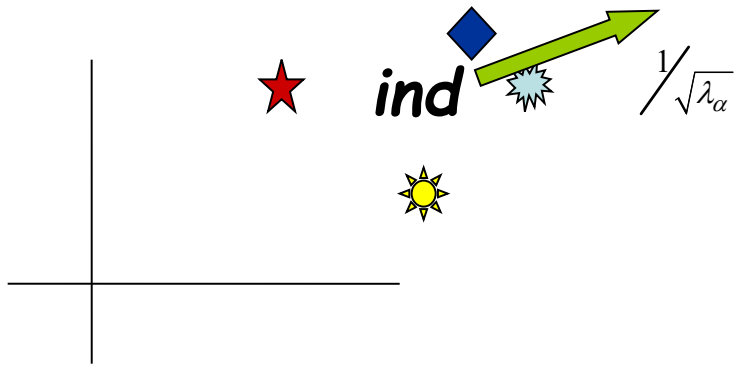
# *Projection of individuals and modalities in MCA*

(from the pseudo-baricentric formulae)

- Every individual is the cdg of their chosen modalities

  (apart from a multiplicative factor)

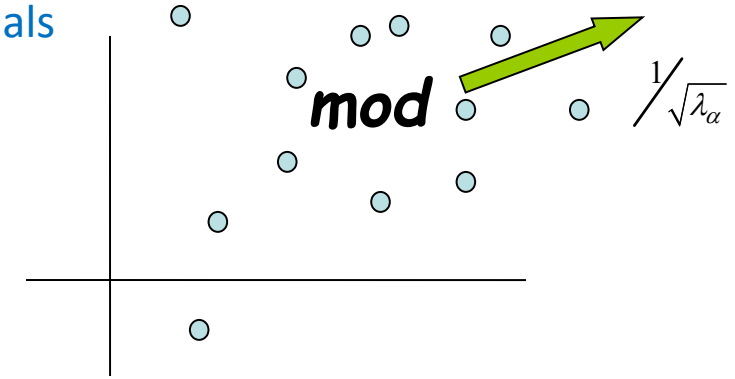
$$\psi_{i\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \frac{\sum_j^J \varphi_{j\alpha} z_{ij}}{p}$$



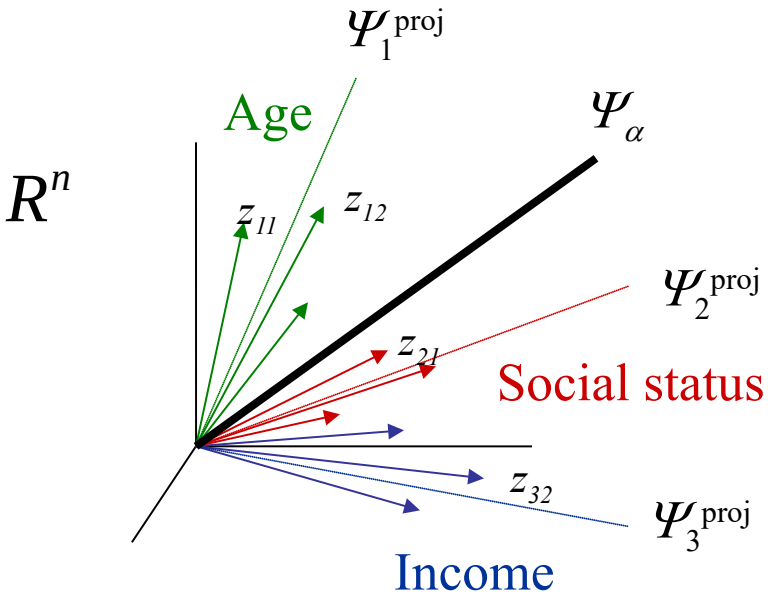

- Une modality (=level) is the cdg of individuals having chosen it

  (apart from a multiplicative factor)

$$\varphi_{j\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \frac{\sum_i^n \psi_{i\alpha} z_{ij}}{n_j}$$



# What are the factors of MCA?




⇒ **Labelling of the dimensions**

# Tips

TIPS

- The inertia for each of the qualitative variables is defined as the sum of the inertias of their categories. The inertia of a variable will be less the fewer categories it presents, so if you want all the variables to have the same weight in the analysis, you will have to consider variables that have the same number of categories.
- The total inertia of the cloud of individuals is the same as that of the cloud of categories, and it only depends on the number of qualitative variables considered and the total number of categories that gather together. Its value is ( K/p)−1 = (K-p)/(p) =sum of eigenvalues (Total Inertia), so Average Inertia is Total Inertia divided K-p = 1/p (Use this to select the "optimal" number of dimensions for Logical/Indicator method)
- The total number of axes selected is K−p in every cloud. Each axis can only collect a maximum of 100(p/(K−p))% of the total inertia of the corresponding cloud.

FIB UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

**Artificial Intelligence Degree**

PMAAD

# FAMD → Factorial Analysis for Mixed Data

# FAMD

Factor analysis of mixed data (FAMD) is a principal component method dedicated to analyze a data set containing both quantitative and qualitative variables (Pagès 2004). It makes it possible to analyze the similarity between individuals by taking into account a mixed types of variables. Additionally, one can explore the association between all variables, both quantitative and qualitative variables.

Roughly speaking, the FAMD algorithm can be seen as a mixed between principal component analysis (PCA) and multiple correspondence analysis (MCA) . In other words, it acts as PCA quantitative variables and as MCA for qualitative variables.

**Pagès, J. 2004. "Analyse Factorielle de Donnees Mixtes." Revue Statistique Appliquee 4: 93–111.**

# *FAMD (Homework and Lab Session)*

VIDEO: https://www.sthda.com/english/articles/21-courses/72-factor-analysis-of-mixed-data-using-factominer/

SCRIPT:

https://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/115-famd-factor-analysis-of-mixed-data-in-r-essentials/