LiAM Laboratori de Modelització i Anàlisi de la Informació

Session
Multiple correspondence analysis

Anàlisi de Dades i Explotació de la Informació

Grau d’Enginyeria Informatica.

Information System tracking

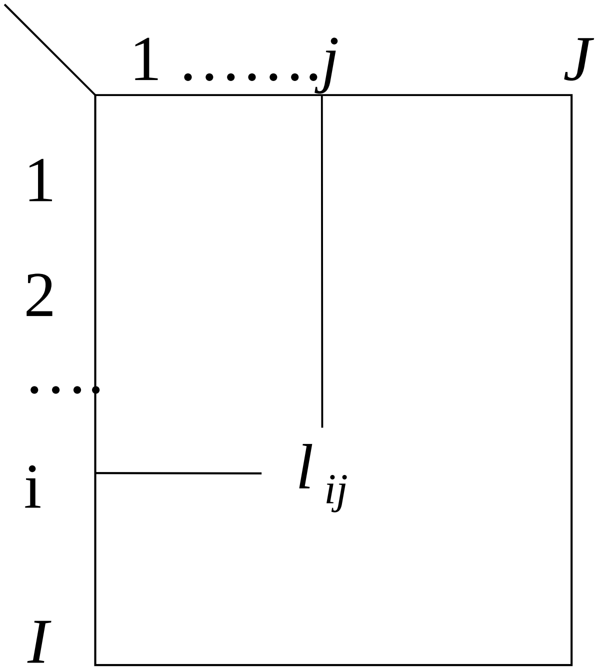
Prof. Mónica Bécue Bertaut & Lidia Montero

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

FIB UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Data: Individuals × Categorical Variables

a. Condensed encoding



$1 \ldots\ldots j$ $\quad J$

1

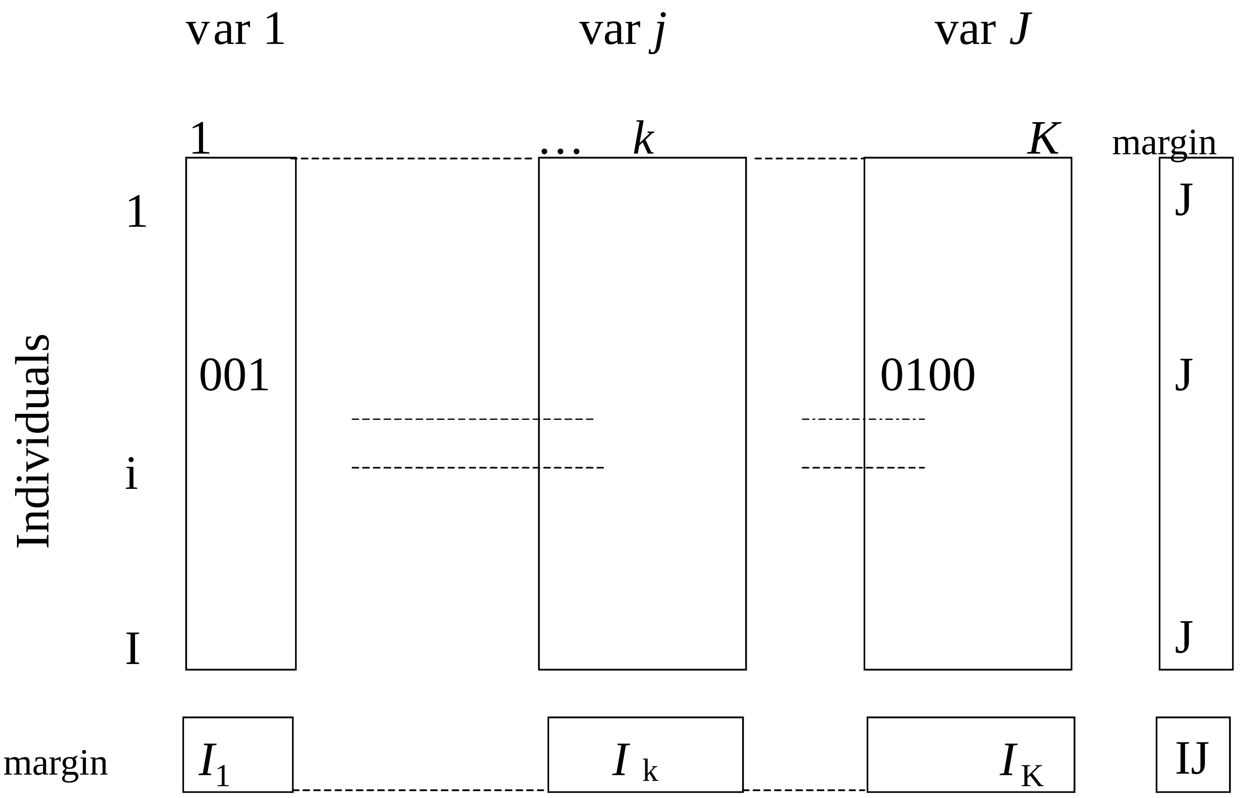
2

$\ldots.$

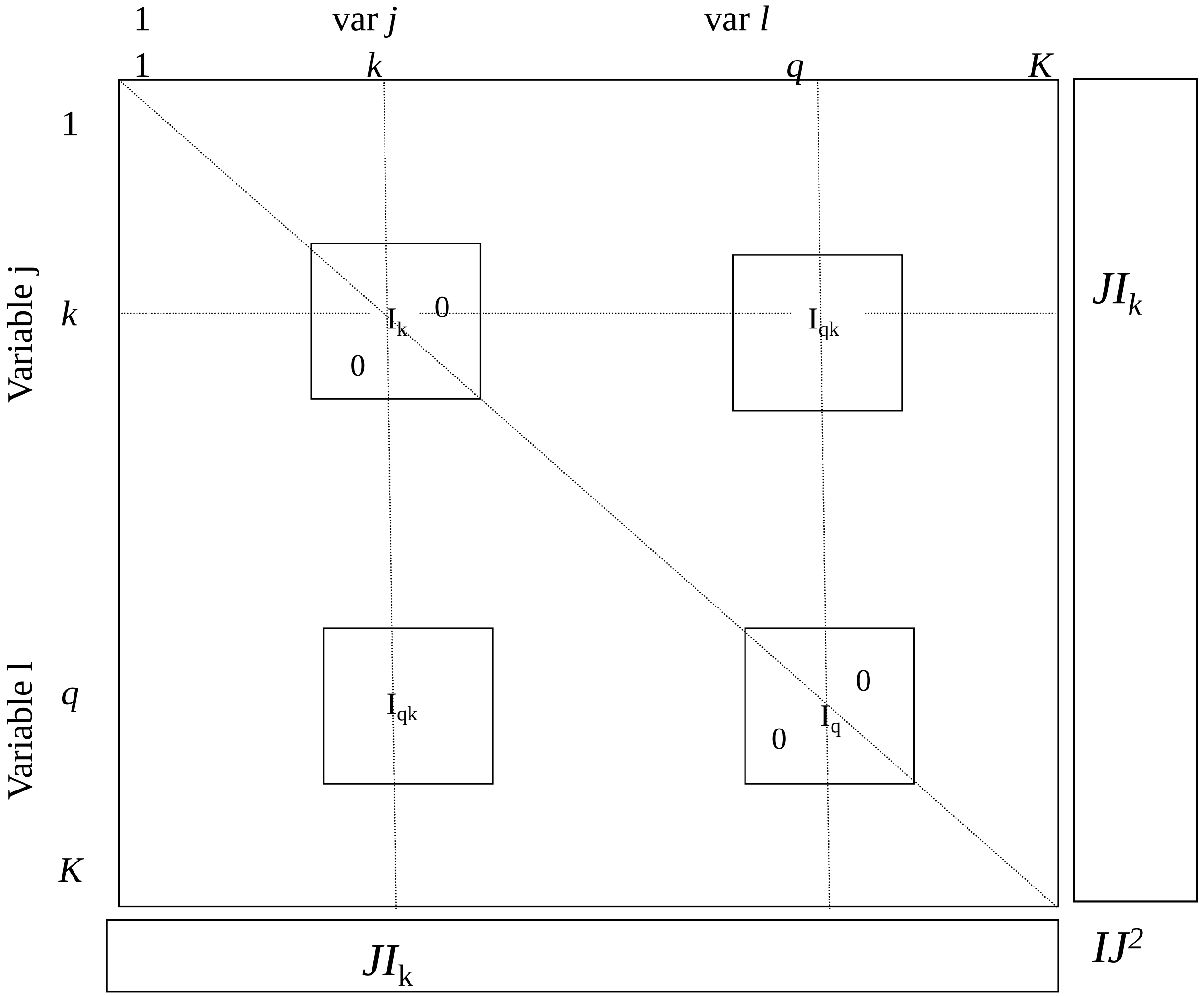
i $\quad l_{ij}$

I

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}
Variable l
q
I_{qk}
0
I_q
0
K
JI_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 categories and variables

- Typology of the categories: close if they are shared by many individuals but also if they are associated to the same other categories

- 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”

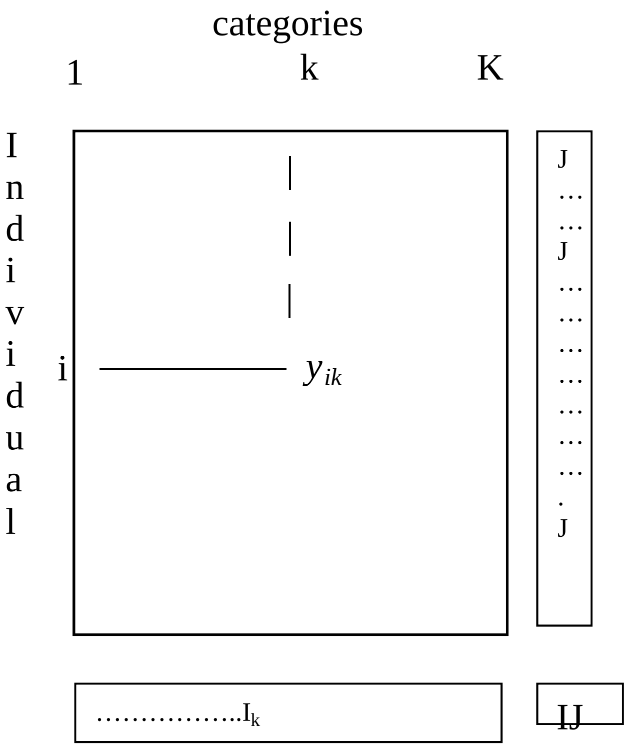
MCA= CA of the disjunctive table

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

BUT MCA computing is totally equivalent to CA applied to CDT

although

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



y_{ik} is *0 ó 1*

$$\sum row = J$$

I

Cloud of individuals

$$d^2(i,l) = \sum_k \frac{IJ}{I_k}\left(\frac{y_{ik}}{J} - \frac{y_{lk}}{J}\right)^2 = \frac{1}{J}\sum_k \frac{I}{I_k}(\underbrace{y_{ik} - y_{il}}_{\text{0 or 1. 1 if only one individual presents category } k})^2$$

Distance: the distance increases depending on the number of categories which differs from an individual to another. Category *k* intervenes with weight I/I_k, inverse of its frequency. The presence of a rare category moves individuals who present it far fromthe other individuals.

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

Weights: uniform weights for all individuals (all the rows have equal sum *J*)

Cloud of categories

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

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

<u>Distance</u>: 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.

<u>Weights</u>: weight of category *k*: $\frac{I_k}{IJ}$

Distance from a category to the centroid $$d^2(k, G_K) = \frac{I}{I_k} - 1$$

Inertia de *k* relatively to G_K $$\frac{1}{J} \cdot \left(1 - \frac{I_k}{I} \right)$$

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

Total inertia of the cloud $$\left(\frac{K}{J} \right) - 1$$

Inertia of the K_j categories of variable *j* $$\frac{K_j - 1}{J}$$

Transition relationships

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

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

The variables through their categories

- Centroid of the categories of one variable

$$\sum_{k \in K_j} \frac{I_k}{I} \frac{y_{ik}}{I_k} = \frac{1}{I}$$

(The centroid of the categories of one variable lies on the global centroid)

- *Subspace generated by the categories of one variable has a dimension equal to r-1:*
- 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
- Inertia of a variable with K_j categories:$(K_j-1)/J$

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:

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

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