

## Lezione : Distanze

*Docente: Aldo Solari*

### 1 Distanze

L'analisi di raggruppamento (*cluster analysis*) ha per scopo far emergere dall'insieme dei dati a disposizione gruppi di unità statistiche “simili” tra loro e “dissimili” da quelle degli altri gruppi. Che cosa si intende per unità statistiche “simili”, o equivalentemente, “dissimili”? Dobbiamo quantificare con un numero la “diversità” tra due unità statistiche

- Variabili Quantitative: Diversità = Distanza (Metrica e Indice di Distanza)
- Variabili Qualitative: Diversità = Indice di Dissimilarità

#### 1.1 La funzione distanza

Consideriamo misurazioni su  $p$  variabili tutte quantitative. Quanto sono “distanti” due unità statistiche  $u_i$  e  $u_l$ ? Dipende da come definiamo la “distanza”.

In generale, una distanza è una funzione

$$d : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$$

che associa ad una coppia di unità statistiche  $u_i$  e  $u_l$  un numero reale

Proprietà di una distanza:

(D1) Non negatività

$$d(u_i, u_l) \geq 0$$

(D2) Identità

$$d(u_i, u_l) = 0 \Leftrightarrow u_i = u_l$$

(D3) Simmetria

$$d(u_i, u_l) = d(u_l, u_i)$$

(D4) Disuguaglianza triangolare

$$d(u_i, u_l) \leq d(u_i, u_t) + d(u_t, u_l)$$

- METRICA: valgono (D1), (D2), (D3) e (D4)
- INDICE DI DISTANZA: valgono (D1), (D2) e (D3)

- Distanza Euclidea  $d_2$  tra due unità statistiche  $\begin{matrix} u'_i & u'_l \\ 1 \times p & 1 \times p \end{matrix}$

$$d_2(u_i, u_l) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{lj})^2}$$

$d_2$  soddisfa le proprietà (D1), (D2), (D3) e (D4), quindi è una metrica

- Distanza di Manhattan (o della città a blocchi)  $d_1$  tra due unità statistiche  $\begin{matrix} u'_i & u'_l \\ 1 \times p & 1 \times p \end{matrix}$

$$d_1(u_i, u_l) = \sum_{j=1}^p |x_{ij} - x_{lj}|$$

$d_1$  soddisfa le proprietà (D1), (D2), (D3) e (D4), quindi è una metrica

- Distanza di Lagrange  $d_\infty$  tra due unità statistiche  $\begin{matrix} u'_i & u'_l \\ 1 \times p & 1 \times p \end{matrix}$

$$d_\infty(u_i, u_l) = \max_{j \in \{1, \dots, p\}} |x_{ij} - x_{lj}|$$

$d_\infty$  soddisfa le proprietà (D1), (D2), (D3) e (D4), quindi è una metrica

**Example 1.1.**  $\begin{matrix} u'_1 & u'_2 \\ 1 \times 2 & 1 \times 2 \end{matrix} = \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix}$

$$d_2(u_1, u_2) = \sqrt{(1-2)^2 + (1-3)^2} = \sqrt{5}$$

$$d_1(u_1, u_2) = |1-2| + |1-3| = 3$$

$$d_\infty(u_1, u_2) = \max\{|1-2|, |1-3|\} = 2$$

## 1.2 Distanza di Minkowski

Distanza di Minkowski  $d_m$  (di ordine  $m \geq 1$ ) tra due unità statistiche  $\begin{matrix} u'_i & u'_l \\ 1 \times p & 1 \times p \end{matrix}$

$$d_m(u_i, u_l) = \left[ \sum_{j=1}^p |x_{ij} - x_{lj}|^m \right]^{1/m}$$

- Per  $m \geq 1$ ,  $d_m$  soddisfa le proprietà (D1), (D2), (D3) e (D4), quindi è una metrica
- Casi particolari: distanza Euclidea ( $m = 2$ ), di Manhattan ( $m = 1$ ) e di Lagrange ( $m = \infty$ )
- E' funzione non crescente dell'indice  $m$

$$d_{m'}(u_i, u_l) \geq d_{m''}(u_i, u_l) \quad 1 \leq m' < m''$$

### 1.3 Distanza di Canberra

Distanza di Canberra  $d_C$  tra due unità statistiche  $u_i$  e  $u_l$   
 $1 \times p \quad 1 \times p$

$$d_C(u_i, u_l) = \sum_{j=1}^p \frac{|x_{ij} - x_{lj}|}{|x_{ij} + x_{lj}|}$$

- Sono esclusi dalla somma i termini in cui il denominatore si annulla
- E' utilizzata per variabili quantitative non negative
- E' una versione pesata della distanza di Manhattan

$$d_C(u_i, u_l) = \sum_{j=1}^p w_j |x_{ij} - x_{lj}|$$

con pesi  $w_j = \frac{1}{|x_{ij} + x_{lj}|}$ ,  $j = 1, \dots, p$

### 1.4 Distanza Euclidea al quadrato

- Distanza Euclidea al quadrato

$$d_2^2(u_i, u_l) = \sum_{j=1}^p (x_{ij} - x_{lj})^2$$

- La distanza Euclidea al quadrato  $d_2^2$  soddisfa (D1), (D2) e (D3) ma non soddisfa (D4), quindi è un indice di distanza
- Tuttavia,  $d_2^2$  gode della proprietà di addittività (mentre la distanza Euclidea  $d_2$  no):

per due insiemi  $K$  e  $Q$  tali che  $K \cup Q = \{1, \dots, p\}$  e  $K \cap Q = \emptyset$

$$d_2^2(u_i, u_l) = \sum_{j=1}^p (x_{ij} - x_{lj})^2 = \sum_{k \in K} (x_{ik} - x_{lk})^2 + \sum_{q \in Q} (x_{iq} - x_{lq})^2$$

**Example 1.2.**  $u_1' = [10 \ 5]$   
 $1 \times 2$

$$u_2' = [13 \ 9]$$

$$u_3' = [11 \ 7]$$

$$d_2^2(u_1, u_2) = (10 - 13)^2 + (5 - 9)^2 = 25$$

$$d_2^2(u_1, u_3) = (10 - 11)^2 + (5 - 7)^2 = 5$$

$$d_2^2(u_3, u_2) = (11 - 13)^2 + (7 - 9)^2 = 8$$

$$25 = d_2^2(u_1, u_2) > d_2^2(u_1, u_3) + d_2^2(u_3, u_2) = 5 + 8 = 13$$

## 2 Distanza di Mahalanobis

Tra due unità statistiche  $\underset{1 \times p}{u'_i}$  e  $\underset{1 \times p}{u'_l}$

$$d_M(u_i, u_l) = \sqrt{\underset{1 \times p}{(u_i - u_l)'} \underset{p \times p}{S^{-1}} \underset{p \times 1}{(u_i - u_l)}}$$

Tra l' $i$ -sima unità statistica  $\underset{1 \times p}{u'_i}$  e il baricentro  $\underset{1 \times p}{\bar{x}'}$ :

$$d_M(u_i, \bar{x}) = \sqrt{\underset{1 \times p}{(u_i - \bar{x})'} \underset{p \times p}{S^{-1}} \underset{p \times 1}{(u_i - \bar{x})}}$$

L'insieme dei punti  $p$ -dimensionali  $\underset{1 \times p}{u'}$  con distanza di Mahalanobis costante  $c > 0$  dal baricentro  $\underset{1 \times p}{\bar{x}'}$  soddisfano l'equazione

$$\underset{1 \times p}{(u - \bar{x})'} \underset{p \times p}{S^{-1}} \underset{p \times 1}{(u - \bar{x})} = c^2$$

che definisce un iper-ellissoide

Si noti che l'insieme di punti  $p$ -dimensionali  $\underset{1 \times p}{u'}$  con distanza Euclidea costante  $c > 0$  dal baricentro  $\underset{1 \times p}{\bar{x}'}$  soddisfano l'equazione

$$\underset{1 \times p}{(u - \bar{x})'} \underset{p \times 1}{(u - \bar{x})} = c^2$$

che definisce una ipersfera di raggio  $c$  dal baricentro

### 2.1 Distanza di Mahalanobis e valori anomali

Se si può assumere che le righe della matrice  $X$  sono realizzazioni indipendenti generate dalla medesima distribuzione Normale  $p$ -variata, possiamo definire l' $i$ -sima unità statistica  $\underset{1 \times p}{u'_i}$  un outlier se

$$d_M^2(u_i, \bar{x}) > q_{0.95}$$

dove  $q_{0.95}$  è il 0.95 quantile di una distribuzione  $\chi^2$  con  $p$  gradi di libertà

$p$	$q_{0.95}$
1	3.8415
2	5.9915
3	7.8147
5	11.0705
10	18.3070
100	124.3421

Se le  $n$  unità statistiche sono realizzazioni indipendenti generate dalla medesima distribuzione Normale  $p$ -variata

$$\text{valore atteso di outliers} = n \times 0.05$$

Qualora osserviamo un numero sostanzialmente più elevato di quello atteso, abbiamo un eccesso di *outliers*

**Example 2.1. Dati Animals**

$i$	$d_M^2(u_i, \bar{x})$	$i$	$d_M^2(u_i, \bar{x})$
1 Mountain beaver	3.339	15 African elephant	5.445
2 Cow	1.529	16 Triceratops	22.392
3 Grey wolf	0.134	17 Rhesus monkey	4.12
4 Goat	0.328	18 Kangaroo	0.014
5 Guinea pig	3.533	19 Golden hamster	8.51
6 Dipliodocus	26.092	20 Mouse	14.908
7 Asian elephant	2.963	21 Rabbit	2.243
8 Donkey	0.343	22 Sheep	0.082
9 Horse	1.348	23 Jaguar	0.169
10 Potar monkey	2.087	24 Chimpanzee	0.795
11 Cat	2.573	25 Rat	6.249
12 Giraffe	1.346	26 Brachiosaurus	38.629
13 Gorilla	0.423	27 Mole	11.303
14 Human	2.084	28 Pig	0.782

### 3 Distanze e trasformazioni lineari

#### 3.1 Invarianza di $d_M$ rispetto alle trasformazioni lineari

La trasformazione lineare dell'  $i$ -sima unità statistica  $u_i' = x_i'$

$$y_i' = x_i' A' + b'$$

è definita da

- la matrice  $A$   
 $p \times p$
- il vettore  $b$   
 $p \times 1$

La matrice dei dati linearmente trasformati risulta

$$Y = X A' + 1 b'$$

con con vettore delle medie e matrice di varianze/covarianze

$$\bar{y}_{p \times 1} = A_{p \times p} \bar{x}_{p \times 1} + b_{p \times 1}, \quad S^Y_{p \times p} = A_{p \times p} S^X_{p \times p} A'_{p \times p}$$

Si noti che ci stiamo limitando a considerare trasformazioni lineari delle unità statistiche da  $\mathbb{R}^p$  a  $\mathbb{R}^p$ , ovvero trasformazioni lineari che non riducono la dimensionalità originale  $p$ .

**Proposition 3.1.** Siano  $y'_i = x'_i A' + b'$  e  $y'_l = x'_l A' + b'$  con  $A$  non singolare. La distanza di Mahalanobis  $d_m$  è invariata rispetto alle trasformazioni lineari (non singolari):

$$\begin{aligned} d_M(y_i, y_l) &= \sqrt{(y_i - y_l)' [S^Y]^{-1} (y_i - y_l)} \\ &= \sqrt{(Ax_i + b - Ax_l - b)' [AS^X A']^{-1} (Ax_i + b - Ax_l - b)} \\ &= \sqrt{[A(x_i - x_l)]' [AS^X A']^{-1} [A(x_i - x_l)]} \\ &= \sqrt{(x_i - x_l)' A' [A']^{-1} [S^X]^{-1} [A]^{-1} A (x_i - x_l)} \\ &= d_M(x_i, x_l) \end{aligned}$$

ricordando che  $(AB)^{-1} = B^{-1}A^{-1}$

### 3.2 Invarianza di $d_m$ rispetto alle traslazioni

Una traslazione è un caso particolare di trasformazione lineare con

- $A_{p \times p} = I_{p \times p}$
- $b'_{1 \times p}$  arbitraria

Traslazione della matrice dei dati  $X$

$$Y_{n \times p} = X_{n \times p} + \mathbf{1}_{n \times 1} b'_{1 \times p}$$

con vettore delle medie e matrice di varianze/covarianze

$$\bar{y}_{p \times 1} = \bar{x}_{p \times 1} + b_{p \times 1}, \quad S^Y_{p \times p} = S^X_{p \times p}$$

**Proposition 3.2.** Siano  $\underset{1 \times p}{y'_i} = \underset{1 \times p}{x'_i} + \underset{1 \times p}{b'}$  e  $\underset{1 \times p}{y'_l} = \underset{1 \times p}{x'_l} + \underset{1 \times p}{b'}$ . La distanza di Minkowski  $d_m$  è invariata rispetto alle traslazioni:

$$\begin{aligned} d_m(y_i, y_l) &= \left[ \sum_{j=1}^p |y_{ij} - y_{lj}|^m \right]^{1/m} \\ &= \left[ \sum_{j=1}^p |(x_{ij} + b_j) - (x_{lj} + b_j)|^m \right]^{1/m} \\ &= \left[ \sum_{j=1}^p |x_{ij} - x_{lj}|^m \right]^{1/m} \\ &= d_m(x_i, x_l) \end{aligned}$$

### 3.3 Invarianza di $d_2$ rispetto alle trasformazioni ortogonali

- $A$  matrice ortogonale:  $\underset{p \times p}{A'} = \underset{p \times p}{A}^{-1}$  e  $\underset{p \times p}{A'} \underset{p \times p}{A} = \underset{p \times p}{A} \underset{p \times p}{A'} = \underset{p \times p}{I}$
- $\underset{1 \times p}{b'} = \underset{1 \times p}{0}$

Trasformazione ortogonale della matrice dei dati  $X$

$$\underset{n \times p}{Y} = \underset{n \times p}{X} \underset{p \times p}{A'}$$

con vettore delle medie e matrice di varianze/covarianze

$$\underset{p \times 1}{\bar{y}} = \underset{p \times p}{A} \underset{p \times 1}{\bar{x}}, \quad \underset{p \times p}{S^Y} = \underset{p \times p}{A} \underset{p \times p}{S^X} \underset{p \times p}{A'}$$

Esempi di trasformazioni ortogonali sono:

- Trasformazione identità:  $\underset{p \times p}{A} = \underset{p \times p}{I}$
- Permutazione:  $\underset{p \times p}{A}$  è una matrice di permutazione che si ottiene permutando le righe (o le colonne) della matrice identità  $\underset{p \times p}{I}$ . In due dimensioni, la seguente matrice di permutazione comporta scambiare l'ordine delle colonne di  $X$ :

$$\underset{2 \times 2}{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

- Rotazione:  $\underset{p \times p}{A}$  è una matrice di rotazione, ovvero  $\underset{p \times p}{A}$  ortogonale con  $\det(A) = 1$  o  $-1$ . In due dimensioni, la seguente matrice di rotazione comporta una rotazione antioraria di angolo  $\theta$  radianti intorno all'origine:

$$\underset{2 \times 2}{A} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

**Proposition 3.3.** Siano  $y_i' = x_i' A'$  e  $y_l' = x_l' A'$  con  $A$  matrice ortogonale. La distanza Euclidea  $d_2$  è invariata rispetto alle trasformazioni ortogonali:

$$\begin{aligned}
 d_2(y_i, y_l) &= \sqrt{(y_i - y_l)'(y_i - y_l)} \\
 &= \sqrt{(Ax_i - Ax_l)'(Ax_i - Ax_l)} \\
 &= \sqrt{[A(x_i - x_l)]'[A(x_i - x_l)]} \\
 &= \sqrt{(x_i - x_l)'A'A(x_i - x_l)} \\
 &= \sqrt{(x_i - x_l)'A^{-1}A(x_i - x_l)} \\
 &= \sqrt{(x_i - x_l)'(x_i - x_l)} \\
 &= d_2(x_i, x_l)
 \end{aligned}$$

### 3.4 Distanza Euclidea calcolata su $\tilde{X}$ , $Z$ e $\tilde{Z}$

- $\tilde{x}_i' = (u_i - \bar{x})'$  è l' $i$ -sima riga di  $\tilde{X} = H X$   
 $1 \times p \quad 1 \times p \quad n \times p \quad n \times n \times p$

$$d_2(\tilde{x}_i, \tilde{x}_l) = \sqrt{(u_i - u_l)'(u_i - u_l)} = d_2(u_i, u_l)$$

$$1 \times p \quad p \times 1$$

- $z_i' = (u_i - \bar{x})' D^{-\frac{1}{2}}$  è l' $i$ -sima riga di  $Z = H X D^{-\frac{1}{2}}$   
 $1 \times p \quad 1 \times p \quad p \times p \quad n \times p \quad n \times n \times p \quad p \times p$

$$d_2(z_i, z_l) = \sqrt{(u_i - u_l)' D^{-1} (u_i - u_l)} = \sqrt{\sum_{j=1}^p \frac{1}{s_{jj}} (x_{ij} - x_{lj})^2}$$

$$1 \times p \quad p \times p \quad p \times 1$$

- $\tilde{z}_i' = (u_i - \bar{x})' S^{-\frac{1}{2}}$  è l' $i$ -sima riga di  $\tilde{Z} = H X S^{-\frac{1}{2}}$   
 $1 \times p \quad 1 \times p \quad p \times p \quad n \times p \quad n \times n \times p \quad p \times p$

$$d_2(\tilde{z}_i, \tilde{z}_l) = \sqrt{(u_i - u_l)' S^{-1} (u_i - u_l)} = d_M(u_i, u_l)$$

$$1 \times p \quad p \times p \quad p \times 1$$

## 4 Indici di similarità

Consideriamo misurazioni su  $p$  variabili, qualitative e/o quantitative. Ciascuna unità statistica presenta misurazioni appartenenti allo spazio campionario  $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_p$ . Ad esempio, se abbiamo  $p = 2$  variabili, Sesso e Posizione geografica, lo spazio campionario è:

$$\mathcal{X} = \mathcal{X}_{\text{Sesso}} \times \mathcal{X}_{\text{Pos.Geog.}} = \{M, F\} \times \{\text{Nord, Centro, Sud}\} = \{(M, \text{Nord}), (F, \text{Nord}), (M, \text{Centro}), (F, \text{Centro}), (M, \text{Sud}), (F, \text{Sud})\}$$



## 4.1 Indice di similarità

In generale, un indice di similarità è una funzione

$$s : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

che associa ad una coppia di unità statistiche  $u'_i$  e  $u'_l$  un numero reale.

Un indice di similarità soddisfa le seguenti proprietà:

$$\begin{array}{ll} \text{(S1) Non negatività} & s(u_i, u_l) \geq 0 \\ \text{(S2) Normalizzazione} & u_i = u_l \Rightarrow s(u_i, u_l) = 1 \\ \text{(S3) Simmetria} & s(u_i, u_l) = s(u_l, u_i) \end{array}$$

dove 1 è il massimo valore assumibile dall'indice di similarità.

Un indice di dissimilarità è definito come

$$d(u_i, u_j) = 1 - s(u_i, u_j)$$

e soddisfa (D1) e (D3)

## 4.2 Variabili binarie

Supponiamo che il profilo dell' $i$ -esima unità statistica  $u'_i$  sia composto di sole variabili binarie (o dicotomiche), codificate per comodità come 0 e 1

$$X_{n \times p} = \begin{bmatrix} 0 & 0 & \dots & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & \dots & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & \dots & 0 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 & \dots & 1 \end{bmatrix} = \begin{bmatrix} u'_1 \\ \dots \\ u'_i \\ \dots \\ u'_l \\ \dots \\ u'_n \end{bmatrix}$$

Possiamo costruire, per ciascuna coppia  $u'_i$  e  $u'_l$ , la seguente tabella di contingenza

unit $i$	unit $l$		
	1	0	
1	$a$	$b$	$a + b$
0	$c$	$d$	$c + d$
	$a + c$	$b + d$	$p = a + b + c + d$

dove

- $a$  è la frequenza di variabili binarie con valore 1 per l'unità  $i$  e valore 1 per l'unità  $l$
- $b$  è la frequenza di variabili binarie con valore 1 per l'unità  $i$  e valore 0 per l'unità  $l$

- etc.

**Example 4.1.**

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} u'_1 \\ u'_2 \end{bmatrix}$$

	$u_2$	
$u_1$	1 0	
1	2 1	3
0	1 1	2
	3 2	$p = 5$

**4.2.1 Indice di corrispondenza e di Jaccard**

Consideriamo 1 come ‘presenza’ e 0 come ‘assenza’. Non è ovvio se la contemporanea presenza 1-1 o la contemporanea assenza 0-0 siano egualmente indicativi di somiglianza. Ad esempio, se le unità sono individui e la variabile binaria è “capelli castani (1)/capelli non castani (0)” la contemporanea presenza 1-1 è indubbiamente indicativa di somiglianza, non così la contemporanea assenza 0-0. Si parla in questo caso di variabile binaria *asimmetrica*. Per contro se la variabile binaria è “maschio (1)/femmina (0)” la contemporanea assenza 0-0 ha lo stesso valore della contemporanea presenza 1-1. Si parla in questo caso di variabile binaria *simmetrica*.

Indice di corrispondenza semplice

$$s_c(u_i, u_l) = \frac{a + d}{p}$$

considera allo stesso modo co-presenze 1-1 e co-assenze 0-0, quindi è opportuno per variabili binarie simmetriche

Indice di Jaccard

$$s_J(u_i, u_j) = \frac{a}{a + b + c}$$

ignora le coassenze 0-0 (ed è indeterminato se  $d = p$ ), quindi è opportuno per variabili binarie asimmetriche

Per l'esempio precedente abbiamo

$$s_c(u_1, u_2) = \frac{3}{5} = 0.6, \quad s_J(u_1, u_2) = \frac{2}{4} = 0.5$$

**Example 4.2.**

$$\begin{bmatrix} 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} u'_1 \\ u'_2 \\ u'_3 \end{bmatrix}$$

Per ciascuna coppia di osservazioni calcoliamo la tabella di contingenza, ottenendo le tre tabelle

$u_1 \setminus u_2$	$I$	$O$	$u_1 \setminus u_3$	$I$	$O$	$u_2 \setminus u_3$	$I$	$O$
$I$	2	2	$I$	1	3	$I$	1	2
$O$	1	0	$O$	0	1	$O$	0	2

$$s_c(u_1, u_2) = 2/5, s_c(u_1, u_3) = 2/5, s_c(u_2, u_3) = 3/5$$

$$s_J(u_1, u_2) = 2/5, s_J(u_1, u_3) = 1/4, s_J(u_2, u_3) = 1/3$$

*Si noti che  $u_1$  è equi-somigliante a  $u_2$  e  $u_3$  secondo  $s_c$ , mentre è più somigliante a  $u_2$  che a  $u_3$  secondo  $s_J$ , questo poichè la co-assenza che lo accomuna a  $u_3$  non ha peso nell'indice di Jaccard.*

### 4.3 Variabili qualitative nominali

Se tutte le variabili sono qualitative nominali (`factor` in R), possiamo considerare come indice di corrispondenza semplice la proporzione di variabili in cui le due unità  $u'_i$  e  $u'_j$  assumono la stessa modalità

$$s_c(u_i, u_j) = \frac{\sum_{j=1}^p I\{x_{ij} = x_{lj}\}}{p}$$

dove  $I\{\cdot\}$  rappresenta la funzione indicatrice

### 4.4 Variabili qualitative ordinali

Variabili qualitative ordinali (`Ord.factor` in R) con modalità ordinate, ad esempio, `mai`  $\prec$  qualche volta  $\prec$  spesso  $\prec$  sempre.

T trattare queste variabili come qualitative non ordinate, sebbene possibile, fa perdere l'informazione relativa all'ordinamento delle modalità (`mai` e qualche volta sono misurate egualmente 'distanti' di `mai` e `sempre`).

Se la  $j$ -sima variabile è qualitativa ordinale, una soluzione alternativa consiste nel trasformare le  $m_j$  modalità ordinate nei corrispondenti numeri interi da 1 a  $m_j$  normalizzando il risultato:

$$y_{ij} = \frac{\text{punteggio}(x_{ij}) - 1}{m_j - 1}$$

e trattare la  $j$ -sima variabile come quantitativa

In questo caso si assume che le 'distanze' tra le categorie ordinate sono le stesse

Ad esempio

Modalità	<code>mai</code>	<code>qualche volta</code>	<code>spesso</code>	<code>sempre</code>
Punteggio	1	2	3	4
$y_{ij}$	0	1/3	2/3	1

### 4.5 Variabili miste: indice di Gower

$$s_G(u_i, u_l) = \frac{\sum_{j=1}^p \delta_{il}(j) s_{il}(j)}{\sum_{j=1}^p \delta_{il}(j)}$$

dove

$$s_{il}(j) = \begin{cases} 1 - \frac{|x_{ij} - x_{lj}|}{\text{range } j\text{-sima variabile}} & \text{se } j\text{-sima variabile quantitativa} \\ I(x_{ij} = x_{lj}) & \text{se } j\text{-sima variabile binaria/nominale} \\ 1 - |y_{ij} - y_{lj}| & \text{se } j\text{-sima variabile ordinale} \end{cases}$$

$$\delta_{il}(j) = \begin{cases} 1 & i, l \text{ confrontabili rispetto } j\text{-sima variabile} \\ 0 & i, l \text{ non confrontabili rispetto } j\text{-sima variabile} \end{cases}$$

dove due unità sono non confrontabili rispetto alla  $j$ -sima variabile se c'è un valore mancante in almeno una delle due o se la  $j$ -sima variabile è binaria asimmetrica e si ha co-assenza 0-0.

## 4.6 Matrice delle distanze/dissimilarità

A  $X_{n \times p}$  si associa una matrice  $D_{n \times n}$  delle distanze/dissimilarità tra le  $n$  unità statistiche

$$D_{n \times n} = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1i} & \cdots & d_{1n} \\ & 0 & \cdots & d_{2i} & \cdots & d_{2n} \\ & & \ddots & \vdots & & \vdots \\ & & & 0 & \cdots & d_{in} \\ & & & & \ddots & \vdots \\ & & & & & 0 \end{bmatrix}$$

dove

- $d_{il} = d(u_i, u_l)$
- $d_{il} = d_{li}$  (la matrice è simmetrica)
- $d_{ii} = 0$