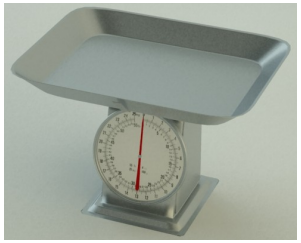
Scale di misurazione

La tecnica adottata per individuare e quantificare i legami tra variabili dipende dalla loro scala di misurazione.

Noi ipotizziamo che per ogni unità (riga della matrice dei dati) possano essere presenti variabili del tipo

- ***Binarie simmetriche e asimmetriche***
- ***Politome sconnesse***
- ***Politome ordinate***
- ***Gradautorie***
- ***Scale a rapporti o intervallari***



In inglese: scale=bilancia

Scale di misurazione

Molte analisi multidimensionali si collegano all'idea di dissomiglianza o distanza tra una coppia di entità.

Pertanto la scelta della scala di misurazione è essenziale e condizionerà pesantemente i risultati.

La scala è una utile caratterizzazione della variabile ed è legata al contesto dell'indagine.

Gli stessi dati possono essere interpretati su scale diverse secondo che interessi solo distinguere le unità oppure graduarne gli scarti.

- *Scala di misura*
  - *Quantitativa*
    - *Rapporti*
    - *Intervalli*
    - *Ranghi*
    - *Graduatorie*
  - *Qualitativa*
    - *Nominali*
    - *Sconnesse*
  - *Variabili binarie*

Scale proporzionali o a rapporti

Ad un incremento relativo nella misura, corrisponde un incremento relativo in ciò che si misura

Ad esempio, per la lunghezza di un segmento, la misura di due centimetri è -senza incertezze- il doppio di uno con lunghezza di un centimetro.

Per queste scale esiste un elemento minimo che si può far coincidere univocamente con l'assenza completa di ciò che si misura.

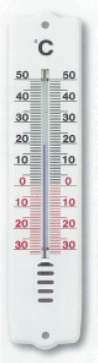
Nelle scale proporzionali rientrano volume, altezza, area, inclinazione di un piano, resistenza alla tensione dei materiali, prezzo di un bene, durata delle componenti di un sistema di controllo.

La caratteristica distintive di queste scale è l'invarianza dei rapporti se le misure sono moltiplicate per una costante:

$$Y_1 = bX_1; \quad Y_2 = bX_2 \Rightarrow \frac{Y_1}{Y_2} = \frac{bX_1}{bX_2} = \frac{X_1}{X_2} \quad se \quad b \neq 0$$

Scale ad intervalli

Nelle scale intervallari si valuta ciò che succede al fenomeno ponendolo, in relazione con un movimento a scansione prefissata lungo un'asta graduata.



Le tacche sono separate -al livello minimo- da una unità convenzionale che può essere dilatata o contratta senza interferire con ciò che si misura o sulla sua interpretazione.

L'origine agisce solo come riferimento e può essere cambiata a piacere. In questo tipo di scala un incremento assoluto tra due misurazioni ha lo stesso significato qualunque sia il livello da cui si calcola l'incremento.

$$30\,C \rightarrow \frac{9*30}{5}+32 = 86F; \quad 40\,C \rightarrow \frac{9*40}{5}+32 = 104\,F; \quad 20\,C \rightarrow \frac{9*20}{5}+32 = 68\,F;$$

Le differenze tra le temperature non sono cambiate: 40-30=30-20 come 104-86=86-68 e non sono cambiati I rapporti tra gli scarti delle due diverse scale: (40-30)/(30-20)=(104-86)/(86-68).

Ciò che si modifica è lo scarto tra le misure della stesa scala che passa da 10 a 18 perché è stato moltiplicato per il fattore 9/5.

Operazioni ammesse

Level	Examples	Numerical Operations	Descriptive Statistics
Interval	100-Point Job Performance Ratings Assigned by Supervisors: 0% = Worst Performers 100% = Best Performers Temperature Type Attitude Scales: Low Temperature = Bad Attitude High Temperature = Good Attitude	Common Arithmetic Operations	• Mean • Median • Variance • Standard Deviation
Ratio	Amount Purchased Salesperson Sales Volume Likelihood of performing some act: • 0%=No Likelihood to • 100%=Certainty Number of stores visited Time spent viewing a particular web page Number of web pages viewed	All Arithmetic Operations	• Mean • Median • Variance • Standard Deviation

Uso dei numeri

la codifica delle modalità porta ad usare dei numeri. Questo però non significa che siano lecite delle operazioni aritmetiche:

i ruoli di una squadra di calcio sono indicati con dei numeri, ma non si può dire che l'ala sinistra ("11") sia maggiore dello stopper ("5") o che l'unità di misura "1" dei calciatori sia il portiere;

ESEMPI

il numero civico delle abitazioni:

Non ha alcun significato operativo la eventuale progressione delle modalità;

Variabili nominali o politòme

Le modalità di queste variabili esprimono categorie, qualità, status: le $\{X_i\}$ in " hanno la sola funzione di etichettare le unità per formarne un elenco o per raggrupparle in classi omogenee:

ESEMPI:

La variabile "Regione" si manifesta con le usuali 20 modalità S={Calabria, Sicilia, ..., Val d'Aosta, Piemonte}.

Un'impresa può ricadere nel settore {agricoltura, industria, altre attività}.

Le differenze possono essere accertate, ma non ordinate né misurate: si possono scambiar di posto senza che ciò influisca sulla validità dei dati così raccolti

Variabili nominali o polìtome

Il livello di misura della variabile è tale che, date due qualsiasi modalità: x_r, x_s, è possibile affermare solo che:

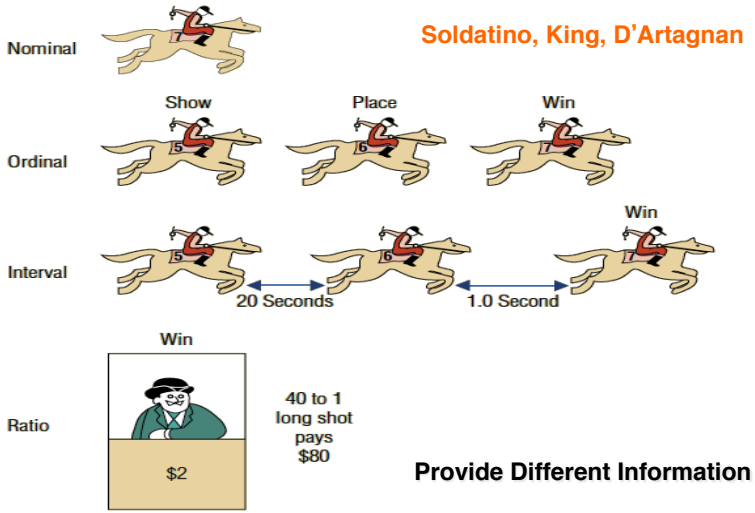
$$x_r = x_s \quad \text{oppure} \quad x_r \neq x_s$$

La scelta di una sola tra le bevande incluse nelle modalità è una politomia

Operazioni ammesse/2

Level	Examples	Numerical Operations	Descriptive Statistics
Nominal	Yes – No Female – Male Buy – Did Not Buy Postal Code: ____	Counting	• Frequencies • Mode
Ordinal	Rankings Choose from the Following: • Dissatisfied • Satisfied • Very Satisfied • Delighted Indicate Your Level of Education: • HS Diploma • Some College • Bachelor's Degree • Graduate Degree	Counting and Ordering	• Frequencies • Mode • Median • Range

Differenza di informazioni



Ordinamenti

Il termine "scala" ha senso se tra le modalità di "S" sono possibili degli ordinamenti.

$$1)\ X_i < X_j \ \ oppure \ \ X_i > X_j \ \ per \ \ ogni \ \ i \neq j$$

$$2)\ X_i < X_j \Rightarrow X_i \neq X_j$$

$$3)\ X_i < X_j \ \ e \ \ X_j < X_k \Rightarrow X_i < X_k \ \ \ per \ \ ogni \ \ i < j < k$$

Maggiore è il contenuto di "fenomeno" maggiore è la modalità che la rappresenta; esiste perciò una disposizione delle modalità che non può essere alterata senza che ne risulti modificata la rilevazione.

Il dominio si esprime con interi consecutivi:

$$S = \{a, a+1,\ a+2,\ \ldots, a+k-1\}$$

Variabili ordinali

i ranghi sono dei voti che esprimono la stima della proprietà posseduta: ogni unità è confrontata con una linea di valutazione che incasella l'unità in una data categoria di valore a prescindere da quello che succede alle altre unità.

Spesso, le modalità di una variabile ordinale esprimono soglie di vicinanza ad un ideale che fungerebbe da "metro" o "campione" di misurazione del concetto.

Invarianza rispetto a trasformazioni monotòne

ESEMPI:

a) Voti di un giudice: S={0, 1, 2, …, 10}; $f(X_i) < f(X_j) \ \ se \ \ X_i < X_j$

b) Ammontare di punti da ripartire: {0 -100};

c) Voti grafici: {++, +0, 0+, 00, -0, 0-, - -};

d) Quantificatore verbale: { pianura, collina, montagna}

Numero di modalità e posizione

Non esiste un numero ottimale di livelli: k=7±2 o k=6 sono considerati uno standard nelle ricerche di mercato (Kinnear eTaylor, 1979, p. 30, Malhotra 1996, p. 298).

3 o 4 gradini comportano risultati confusi per l'accorpamento di giudizi eterogenei; d'altra più di sei è utile solo per acquisire variazioni di quantità molto piccole di cui non sempre si ha bisogno.

Anche la disposizione deve essere equilibrata:

ESEMPIO:

Quale delle tre seguenti moltiplicazioni

$P1.\ 9*7*8*6*5*4*3$

$P2.\ 3*4*5*6*7*8*9$

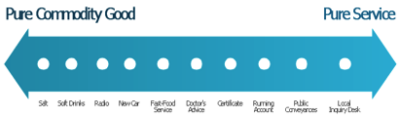
$P3.\ 7*3*8*4*6*5*9$

Effetto posizione

darà il risultato più alto?

Differenziale semantico

Per attenuare le ambiguità delle scale ordinali si possono usare delle scale bipolari in cui siano inserite solo valutazioni opposte dell'aspetto indagato collocando tra di esse, ad opportune interdistanze, una serie di riquadri.



Chi risponde dovrà poi indiviudare il punto più prossimo al suo giudizio ovvero indicare quale descrizione numerica o verbale si adatti al proprio sentire.

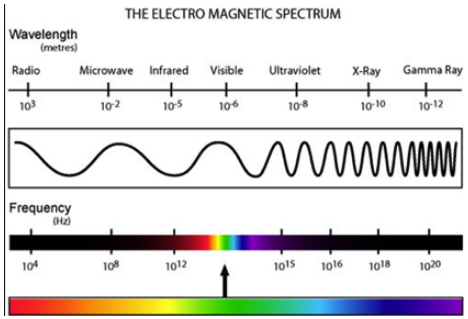
ESEMPIO:
"Come giudicate l'operato dei rappresentanti degli studenti nel Senato accademico?"

Continuo percettivo

L'intensità con cui si avverte una sensazione varia in una sequenza continua di stati: al tessuto uniforme del concetto si sovrappone una griglia più o meno regolare

una unità che sia prima X_i e poi X_j con $X_i<X_j$ sarà passata per tutti gli stati intermedi tra X_i ed X_j.

Le suddivisioni non sono però oggettive come nelle misurazioni fisiche: osservatori diversi scelgono divisioni diverse ovvero lo stesso punto di separazione ha senso diverso.



N.B. Talvota la proprietà studiata ha natura discontinua: si modifica con una scansione non frazionabile per un numero finito di stati che sono i soli a poter essere osservati.

Rilevazione diretta dei ranghi

Un certo insieme di *n* oggetti o situazioni sono ordinate secondo il grado con cui presentano una certa caratteristica X.

Supponiamo …

Che la caratteristica sia un *mix* di immaterialità graduabile, ma non misurabile.

Che le valutazioni siano espresse con i voti {1,2,…,n} così ottenendo la permutazione $\{s_1, s_2,\ldots,s_n\}$

Ripetiamo la rilevazione per una *Y* rilevata allo stesso modo e che produce la permutazione: $\{r_1, r_2,\ldots,r_n\}$

Condizione di ansia e stress
Prima e dopo una separazione

Gli indici di correlazione tra ranghi cercano di quantificare l'intensità del legame tra i due insiemi di giudizi

Esempio: giudizi degli esperti

Ad un esperto è stato chiesto di pronunciarsi sulla posizione che le 20 squadre di un campionato di calcio occuperanno alla fine: $\{s_1, s_2,\ldots,s_{20}\}$.

Alla fine della stagione i giudizi sono comparati con le posizioni reali: $\{r_1, r_2,\ldots,r_{20}\}$.

Per semplificare il calcolo possiamo disporre le due serie di posizioni secondo l'ordine crescente della prima.

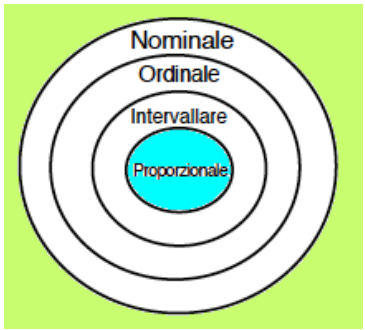
Squadra	A	B	C	D	E	F	G	H	I	L	M	N	O	P	Q	R	S	T	U	V	Totale
Prima	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	210
Dopo	9	2	4	7	5	1	3	8	6	11	13	10	14	18	15	12	16	20	17	19	210

L'esperto ha dato un buon giudizio sebbene sembri più in grado di indovinare le squadre che avranno una cattiva stagione rispetto a quelle che l'avranno buona.

Gerarchia tra scale di misurazione

Se una variabile è su scala proporzionale, con un processo di arrotondamenti è possibile riportarla su scala intervallare; questa a sua volta instaura un ordinamento che è anche utilizzabile per valutare la similarità delle categorie a quella di riferimento (scala nominale).

Fra le scale esiste perciò una gerarchia:.



Strutture di dati

Matrice dei dati: organizzazione righe-colonne (nxm) collocata al centro di molte tecniche di analisi multivariata

n	numero di entità
m	numero di variabili

Matrice rettangolare

$$\mathbf{X} = \begin{array}{c|cccc} & V_1 & V_2 & V_j & V_m \\ \hline U_1 & x_{11} & x_{12} & x_{1j} & x_{1m} \\ U_2 & x_{21} & x_{22} & x_{21} & x_{2m} \\ U_i & x_{i1} & x_{i2} & x_{ij} & x_{im} \\ U_n & x_{n1} & x_{n2} & x_{in} & x_{nm} \end{array}$$

Matrice delle dissimilarità o distanze di ordine (nxn)

Spesso la matrice delle dissimilarità o delle affinità è la prima elaborazione della matrice dei dati per poi servire da base per ulteriori analisi.

Altre volte è il punto di partenza delle analisi

Matrice quadrata

$$D = \begin{array}{c|cccc} & U_1 & U_2 & U_i & U_n \\ \hline U_1 & 0 & d_{12} & d_{1i} & d_{1n} \\ U_2 & d_{12} & 0 & d_{2i} & d_{2n} \\ U_i & d_{1i} & d_{2i} & 0 & d_{in} \\ U_n & d_{1n} & d_{2n} & d_{in} & 0 \end{array}$$

Confronto delle unità

Dalla matrice dei dati si passa alla matrice delle affinità e da questa alla matrice delle dissimilarità o distanze (è possibile pure il passaggio diretto)

$$\mathbf{X} = \begin{array}{c|cccc} & V_1 & V_2 & V_j & V_m \\ \hline U_1 & x_{11} & x_{12} & x_{1j} & x_{1m} \\ U_2 & x_{21} & x_{22} & x_{21} & x_{2m} \\ U_i & x_{i1} & x_{i2} & x_{ij} & x_{im} \\ U_n & x_{n1} & x_{n2} & x_{in} & x_{nm} \end{array}$$

$$D = \begin{array}{c|cccc} & U_1 & U_2 & U_i & U_n \\ \hline U_1 & 0 & d_{12} & d_{1i} & d_{1n} \\ U_2 & d_{12} & 0 & d_{2i} & d_{2n} \\ U_i & d_{1i} & d_{2i} & 0 & d_{in} \\ U_n & d_{1n} & d_{2n} & d_{in} & 0 \end{array}$$

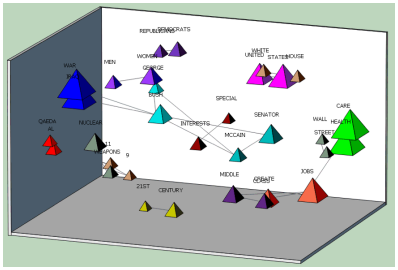
La prima struttura riporta analiticamente tutti i dati, la seconda ne condensa alcune caratteristiche.

La matrice delle dissimilarità può sia essere costruita in base alle variabii di una matrice di dati, ma anche essere ottenuta dal confronto diretto delle unità rispetto a caratteristiche immateriali o indirette.

Matrice delle distanze/dissimilarità

Una matrice quadrata (n x n) in cui ogni elemento rappresenta una misura di come differiscono tra di loro le due unità a confronto.

$$D = \begin{array}{c|cccc} & U_1 & U_2 & U_i & U_n \\ \hline U_1 & 0 & d_{12} & d_{1i} & d_{1n} \\ U_2 & d_{12} & 0 & d_{2i} & d_{2n} \\ U_i & d_{1i} & d_{2i} & 0 & d_{in} \\ U_n & d_{1n} & d_{2n} & d_{in} & 0 \end{array}$$



Lo scopo di molte tecniche di analisi multivariata è di ottenere una configurazione geometrica di punti.

Questa dovrebbe essere di aiuto a comprendere come i soggetti valutano gli aspetti sottoposti rispetto a poche dimensioni latenti

Esempio: *heart disease*

N=303, m=14

1. **Eta'**
2. **Genere**
3. **Tipo di dolore toracico *(1:angina tipica; 2:angina atipica; 3:non anginoso; 4:Asintomatico)***
4. **Pressione del sangue a riposo *(in mm Hg alla accettazione in ospedale)***
5. **Livello di colesterolo (mg/dl)**
6. **Glicemia >120 mg (dicotomica)**
7. **Elettrocardiogramma a riposo *(0:normale, 1:con anomalie dell'onda ST-T*, 2:ipertrofia probabile o definita ventricolare sinistra secondo i criteri di Estes)***
8. **Numero massimo di battiti**
9. **Comparsa di dolore anginoso sotto stress (categorica)**
10. **Diminuzione della distanza ST sotto stress rispetto a quella a riposo**
11. **Pendenza del segmento ST sotto stress *(1: ascendente; 2: piatto; 3: Discendente)***
12. **Numero di vasi principali evidenziati dalla fluoroscopia (da zero a tre)**
13. **Thal *(3: normale; 6: difetto irreversibile; 7: difetto reversibile)***
14. **Diagnosi di cardiopatia ischemica (rispetto all'angiografia): *(0: volume > 50%,1: volume < 50%); i valori da 1 a 4 si riferiscono al numero di vasi principali risultati occlusi all'analisi angiografica.***

Esempio: *heart disease/2*

Varianza e norme delle variabili di Cleveland con indicazione della scala di misurazione

Variabili	σ_i^2	$\|\cdot\|_\infty$	$\|\cdot\|_1$	$\|\cdot\|_2$
1 - numerale	81,8977	77	16199	952,77
2 - dicotomica	0,2195	1	201	14,177
3 – a intervalli	0,9310	4	938	56,903
4 - numerale	31,.5173	200	39113	2290
5 - numerale	$2737 * 10^3$	564	73463	4355,6
6 - dicotomica	0,1242	1	43	6,5574
7 – categorica	0,9899	2	296	24,249
8 - numerale	526,3153	202	44431	2608,2
9 - categorica	0,2207	1	97	9,8489
10 – numerale	1,3597	6,2	313,5	27,082
11 – categorica	0,3822	3	476	29,597
12 – numerale?	0,8817	3	201	19,925
13 - categorica	3,7583	7	1405	88,085
14 – numerica	1,5241	4	281	26,777

Ogni tipologia di variabile necessita di una specifica misura di affinità o di dissimilarità.

Alle variabili deve poi essere assegnato un peso: in blocco o singolarmente

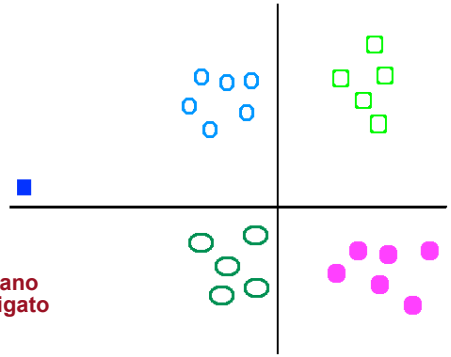
Il tema dell'affinità

La percezione della similarità rappresenta forse l'asse portante del nostro pensiero.

La nozione di similarità ha un ruolo fondamentale in contesti quali diagnosi delle malattie, economia, psicologia, marketing, teoria delle decisioni, sociologia e scienza della politica,

Il tema si affronta valutando su di un piano il modo in cui dei soggetti organizzano degli item al fine di collocare vicini quelli somiglianti e lontani quelli diversi.

Il grafico aiuta a visualizzare le percezioni



La rappresentazione geometrica su piano cartesiano è un passaggio quasi obbligato per lo studio dell'affinità, prossimità, similarità tra item distinti.

La misura dell'affinità

Il problema del confronto nasce quando si considerano almeno due unità rispetto ad una caratteristica suscettibile di almeno due valori.

Il caso minimale è appunto n=2 unità, m=1 variabile con 2 modalità

$$S(x) = \{x_1, x_2\}$$

Il confronto dei due soggetti i e j rispetto alla variabile X potrà configurare le seguenti situazioni

unità i \ unità j	x_1	x_2
x_1	=	≠
x_2	≠	=

N.B. il confronto a due su di una variabile alla volta è una astrazione. In realtà i confronti coinvolgono molte più unità e più aspetti

La situazione di indagine è tale che per ogni rilevazione potrà verificarsi una ed una sola delle celle previste nella suddetta tabella tetracorica

La misura dell'affinità/2

L'affinità (o prossimità o contiguità oppure somiglianza) tra due unità è la percezione di qualche loro tratto palese o latente che porta a collocarle in un'unica categoria piuttosto che in categorie diverse.

Se i due stati possibili della X sono due località o due periodi allora le entità potrebbero essere giudicate affini se sono coeve oppure occupano lo stesso sito (contiguità temporale o spaziale).

Se le due modalità indicano due aspetti che siano parti, componenti, organi, etc. allora l'affinità potrebbe derivare da qualcosa che le unità hanno in comune o su cui hanno un comune effetto

Se invece X_1 è la presenza di una caratteristica X_2 la sua assenza, l'affinità sarà legata alla condivisione di quella presenza e/o dell'assenza.

All'aumentare del numero di modalità ed all'aumentare delle relazioni che si possono instaurare tra le modalità stesse (tenuto conto della loro scala di misurazione), l'idea di affinità diventa più articolata e la sua misura sempre più ricca di possibilità.

Rilevazione dell'affinità

Ecco tre diversi modi per rilevare l'affinità

- **RATING. Un modo immediato di acquisire un valore per l'affinità e di richiedere ai soggetti di assegnare un voto: (1-10) (18-30) (0-100) in cui l'estremo inferiore implica zero somiglianza e l'estremo superiore la massima somiglianza.**

- **PROBABILITA' DI CONFUSIONE. L'affinità si può rilevare ipotizzando la probabilità di confusione tra gli item. Se la probabilità di confusione è nulla allora gli item sono estremamente dissomiglianti. Se è certa la confusione allora gli item sono identici.**

- **SORTING. Gli item a confronto sono disposi in gruppo da un certo numero di soggetti. L'affinità tra due item è data dalla proporzione in cui gli item sono stati collocati nello stesso gruppo.**

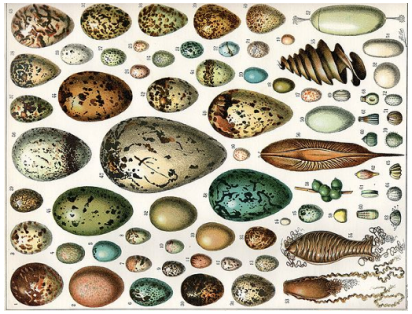
"anything which, by an act of faith, can be considered a similarity" (Shepard)

Rilevazione dell'affinità/2

Occorre dedicare molta cura alla definizione della metrica: non è facile e ci vuole inventiva oltre che tecnica.

I benefici di una metrica ben fatta sono però altamente remunerativi

Il grado di affinità tra due uova è dato dal confronto tra i volumi d'acqua spostati.



Se il numero di item da comparare è elevato diventa difficile ottenere giudizi attendibili.

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

Spesso, ai "giudici" sono presentati set diversi di coppie.

Se possibile occorre basarsi su elementi neutri.

Nel caso delle figure, ogni affinità riferita alla forma dell'oggetto può dare buoni risultati.

Esempio: codice Morse

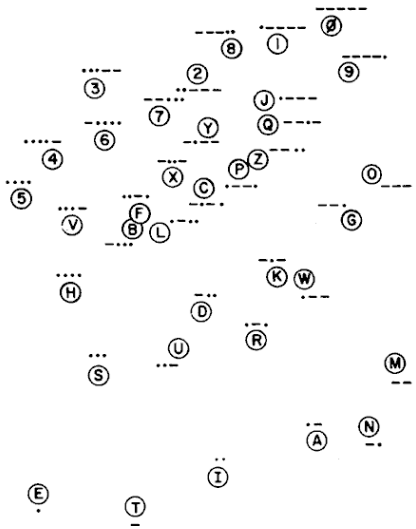
Il codice Morse consiste brevi segnali composti da punti e linee per rappresentare 10 cifre e 26 lettere.

Per valutare la loro confondibilità i segnali sono stati tradotti in suoni (brevi e lunghi) e sottoposti a circa 600 soggetti che non conoscevano il Morse.

I segnali sono stati presentati a coppie 2 volte (prima uno e poi un altro e viceversa)

L'affinità è rilevata con il numero di mancate corrispondenze.

Nel grafico c'è il risultato di una delle procedure di scaling



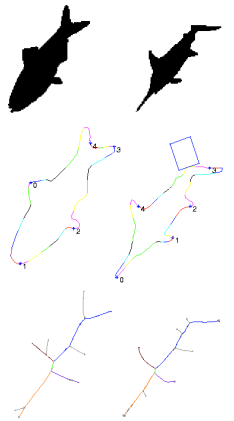
Rilevazione dell'affinità/3

La definizione della metrica è la fase cruciale in una vastissima gamma di ricerche.

In molte occasioni è necessario definire una metrica ad hoc

Che tipo di metrica adottare per valutare la somiglianza-dissomiglianza ne confronto di due pesci?

a) La conformazione del bordo
b) La lisca
c) La combinazione di genere, peso, colore
d) Lo sforzo comunque misurato di trasformarne uno in un altro.



Esempio di Gordon (1990)

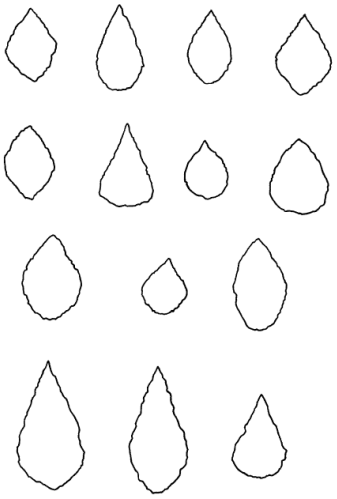
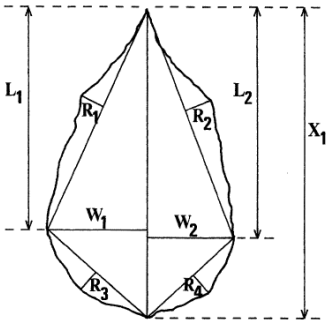


Figure 1. Outlines of 14 leaf-shaped flint arrowheads recovered during archaeological investigations in northeast Scotland.



Definition of variables used to describe the flint arrowheads; see also Figure 2

X_1 = maximum length of arrowhead.

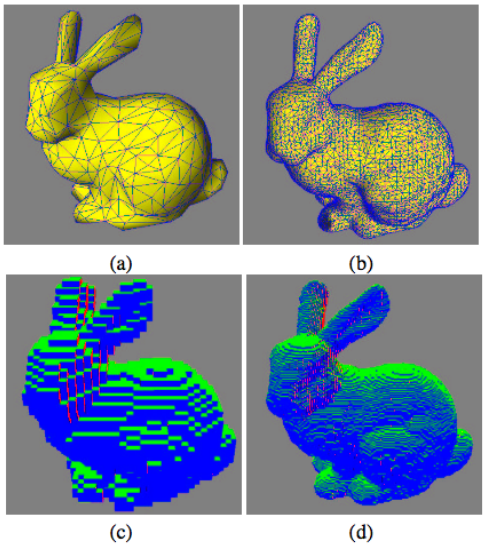
$X_2 = (W_1 + W_2)/X_1$, ratio of maximum width to maximum length.

$X_3 = (L_1 + L_2)/X_1$, location of maximum width.

$X_4 = (R_1 + R_2)/X_1$, convexity of upper part of arrowhead.

$X_5 = (R_3 + R_4)/X_1$, convexity of lower part of arrowhead.

Affinità e dettagli



il riconoscimento delle forme e la loro modellazione sono riferite al numero di dettagli che si riescono a considerare.

Maggiori sono i dettagli maggiore è la accuratezza della ricostruzione.

Alcuni dettagli possono essere tralasciati come l'artista che nel dipingere un quadro ignora molte cose che sono invece presenti in una foto.

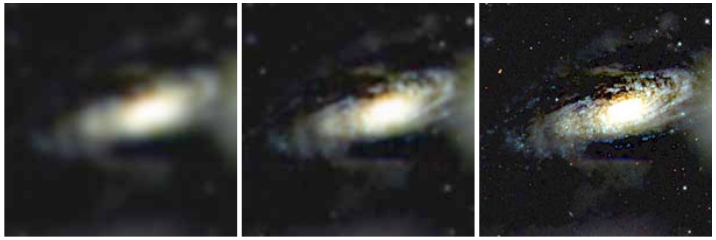
Analogia del telescopio

bassa sensitività

alta sensitività

La sensitività è il livello minimo di affinità che l'indice riesce a distinguere rispetto al rumore di fondo

Maggiore è la risoluzione cioè più numerose sono le variabili, tante più solo differenze che la misura ègrado di percepire tra le entità a confronto.



Un esempio

Fitch and Margoliash (Science, 1967) hanno determinato la dissimilarità tra specie in base al numero di posizioni nella molecola della proteina cytocrome-c in cui le due specie avevano aminoacidi diversi.

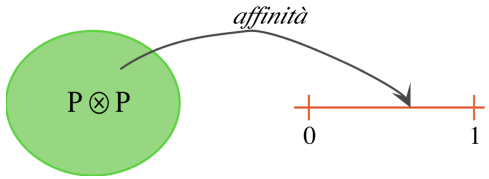
La matrice qui riportata ne include una parte

	Man	Monkey	Horse	Pig	Pigeon	Tuna	Mould	Fungus
Man	0	1	17	13	16	31	63	66
Monkey	1	0	16	12	15	32	62	65
Horse	17	16	0	5	16	27	64	68
Pig	13	12	5	0	13	25	64	67
Pigeon	16	15	16	13	0	27	59	66
Tuna	31	32	27	25	27	0	72	69
Mould	63	62	64	64	59	72	0	61
Fungus	66	65	68	67	66	69	61	0

Da notare la non negatività dei valori, la diagonale nulla e la simmetria

Confronti e affinità

Per ragioni di semplicità preferiamo pensare all'affinità come ad una misura normalizzata compresa nell'intervallo unitario.



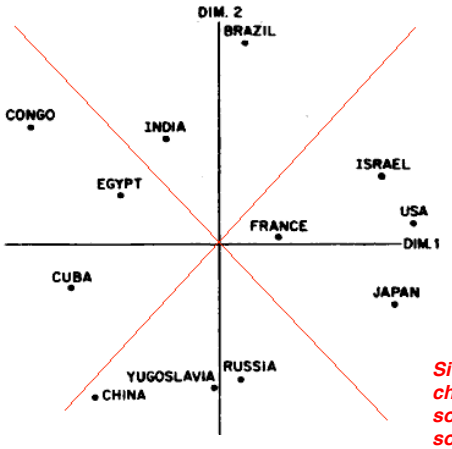
Dove $P = \{u_1, u_2, \ldots, u_n\}$ **è l'insieme delle "n" entità considerate nell'indagine.**

Ad ogni elemento del prodotto cartesiano di insiemi $P \otimes P$ **è associato un numero reale $a_{i,j}$ che esprime il grado di affinità, di prossimità, di contiguità comunque percepito tra le entità a confronto.**

L'affinità o similarità è un numero reale non negativo a_{ij}, variamente ottenuto, in grado di quantificare in modo univoco le differenze che separa due entità, se differenti.

Altro esempio

Un gruppo di soggetti (18) ha classificato n=12 Paesi in base ad un differenziale di 9 livelli dove 1= estremamente diversi e 9=Identici



STIMULI	DIM. 1	DIM. 2
BRAZIL	0.15	1.22
CONGO	-1.15	0.71
CUBA	-0.90	-0.29
EGYPT	-0.60	0.29
FRANCE	0.36	0.02
INDIA	-0.33	0.64
ISRAEL	0.96	0.40
JAPAN	1.04	-0.39
CHINA (MAINLAND)	-0.76	-0.96
RUSSIA	0.12	-0.85
U.S.A.	1.14	0.12
YUGOSLAVIA	-0.03	-0.90

Si possono notare alcuni cluster di Paesi che sono percepiti simili. Gli assi non sono solo un oggetto geometrico, ma sono variabili latenti dietro la percezione dei soggetti che li hanno classificati

Requisiti per le misure di affinità

Indistinguibilità degli identici: $\text{Se } u_i = u_j \Rightarrow a_{ij} = 1$

Questo implica che $a_{ii} = a_{jj} = 1$ ***cioè se si confrontano una entità con se stessa questo deve risultare in un valore unitario (o massimo) dell'affinità.***

Distinguibilità dei diversi: $\text{Se } u_i \neq u_j \Rightarrow a_{ij} < 1$

Il valore uno deve essere riservato all'identità tra i due soggetti a confronto. Ogni altra comparazione deve dar luogo ad un grado o misura dell'affinità diverso da quello massimo. Non sempre questo requisito può essere rispettato

Monotonicità: se $a_{i,j} < a_{r,s}$ allora le entità i e j sono meno affini di quanto non lo siano le unità r ed s

Questa proprietà garantisce la possibilità di ordinare in sequenza diversi casi di affinità riscontrati

$$0 \le a_{i_1 i_2} \le a_{i_2 i_3} \le \ldots \le a_{i_{n-1} i_n} \le 1$$

Requisiti per le misure di affinità/2

Simmetria: $a_{ij} = a_{ji} \;\; \forall i,j$

Talvolta questa condizione deve essere abbandonata dato che è incongrua con certi aspetti intuitivi dell'affinità:
il viaggio da una località A ad un'altra località B non è necessariamente lo stesso che da B ad A dato che i due lati della stessa strada potrebbero avere condizioni di usura diverse.

In alcune soluzioni chimiche l'ordine di miscelazione degli ingredienti implica la formazione di composti differenti

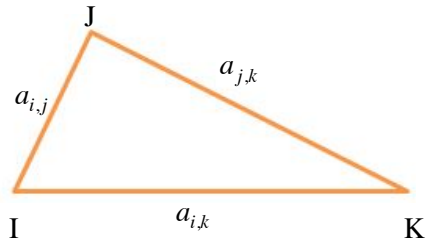
Requisiti per le misure di affinità/3

L'affinità che si può rilevare dal confronto di due unità non può superare quella riscontrabile in una triade di entità comparate due a due:

$$a_{ij} \le a_{ik} + a_{jk} \;\; \forall i,j,k$$

Se i valori delle variabili fossero univocamente rappresentabili come proiezioni su degli assi ortogonali allora ogni unità sarebbe un punto.

La disuguaglianza triangolare assicura che tali punti, in triade, formino dei triangoli: scaleni, isoscele, equilateri.

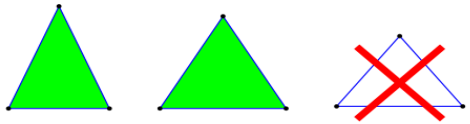


Sebbene la disuguaglianza triangolare sia considerata fondamental, altri autori la considerano una condizione valida in generale, ma che per un numero ristretto di comparazioni può essere violata.

Requisiti per le misure di affinità/4

Ultrametricità: $a_{ij} \le \max\left\{a_{ik}, a_{jk}\right\} \;\; \forall i,j,k$

In questo caso le triadi di unità, se viste come punti dello spazio, possono formare solo un triangolo isoscele.



Se la la condizione di ultrametricità è verificata, allora anche la disuguaglianza triangolare è verificata in quanto entrambe rientrano nello schema:

$$a_{ij} \le \left[a_{ik}^{\alpha} + a_{jk}^{\alpha}\right]^{\frac{1}{\alpha}} \;\; \forall \; i,j,k$$

Se α=1 si ottiene la disuguaglianza triangolare, ma se α va all'infinito si ottiene la disuguaglianza ultrametrica.

Inoltre, è possibile dimostrare che se la ultra metricità è vera per α allora è anche vera $\beta<\alpha$; ne consegue che la disuguaglianza ultrametrica è una condizione più stringente della triangolare, in quanto la implica e non ne è implicata.

Variabili binarie o dicotome

La variabile binaria esprime la dicotomia tra due possibili stati in cui può trovarsi l'unità.

Di solito si rileva la presenza o l'assenza di una proprietà e talvolta si fa riferimento alla condizione ON/OFF di un circuito logico

Si possono verificare quattro eventi

u_i \ u_j	1	x_1	x_2
x_1	1	0	
x_2	0	0	

u_i \ u_j	2	x_1	x_2
x_1	0	0	
x_2	0	1	

u_i \ u_j		x_1	x_2
x_1	0	0	
x_2	1	0	

u_i \ u_j	m	x_1	x_2
x_1	0	1	
x_2	0	0	

Ogni cella può contribuire in modo diverso alla misura dell'affinità:

a) contemporanea presenza: le due unità sono più simili o vicine tra di loro perché hanno in comune un aspetto.
b) contemporanea assenza: le due unità sono più simili o più vicine tra di loro perché sono entrambe privi di una caratteristica.

c) e d) le due unità sono meno simili o vicine perché hanno un comportamento diverso rispetto alla presenza/assenza, di una proprietà.

Esempio

Comune	X1	X2	X3
Roccasecca	0	1	0
Cajaniello	0	1	1
Capri	1	1	1
Sorrento	1	0	1

Il computo dell'affinità passa per la valutazione della compresenza e della coassenza ovvero della non presenza in una o entrambe le unità delle caratteristiche misurate dalle tre variabili

I comuni più affini sono Capri e Sorrento dato che la somma delle affinità binarie raggiunge il massimo.
I meno affini sono Roccasecca e Sorrento che non hanno alcuna affinità rispetto agli aspetti considerati nell'esempio.

X1	RS		Caj		Cap		Sor	
RS	1	0	0	0	0	1	0	1
	0	0	0	1	0	0	0	0
Caj	0	0	1	0	0	1	0	1
	0	1	0	0	0	0	0	0
Cap	0	1	0	1	1	0	1	0
	0	0	0	0	0	0	0	0
Sor	0	1	0	1	1	0	1	0
	0	0	0	0	0	0	0	0

X2	RS		Caj		Cap		Sor	
RS	1	0	1	0	1	0	0	0
	0	0	0	0	0	0	1	0
Caj	1	0	1	0	1	0	1	0
	0	0	0	0	0	0	0	0
Cap	1	0	1	0	1	0	0	0
	0	0	0	0	0	0	0	0
Sor	0	0	1	0	0	0	1	0
	1	0	0	0	1	0	0	0

X3	RS		Caj		Cap		Sor	
RS	1	0	0	1	0	1	0	1
	0	0	0	0	0	0	0	0
Caj	0	1	1	0	1	0	1	0
	0	0	0	0	0	0	0	0
Cap	0	1	1	0	1	0	1	0
	0	0	0	0	0	0	0	0
Sor	0	1	1	0	1	0	1	0
	0	0	0	0	0	0	0	0

Tot	RS		Caj		Cap		Sor	
RS	3	0	1	1	1	2	0	2
	0	0	0	1	0	0	1	0
Caj	1	1	3	0	2	1	2	1
	0	1	0	0	0	0	0	0
Cap	1	2	2	1	3	0	2	0
	0	0	0	0	0	0	0	0
Sor	0	2	2	1	2	0	3	0
	1	0	0	0	1	0	0	0

Variabili binarie o dicotome/2

La valutazione del legame tra unità passa di solito per molte variabili binarie perché sarebbe troppo riduttivo affrontare una comparazione solo in base allo stato attivo/passivo di un'unica caratteristica.

Per ogni variabile si può produrre una delle combinazioni: 00, 11, 01, 10 e l'esito del confronto risulterà dalla aggregazione dei singoli confronti parziali.

Sulla singola variabile binaria

k	x_1	x_2
x_1	a_k (11)	b_k (10)
x_2	c_k (01)	d_k (00)

Riassuntiva

u_i \ u_j	a	b
c	d	

Aggregazione su m variabili binarie

$$\begin{matrix} \sum_{k=1}^{m} a_k & \sum_{k=1}^{m} b_k \\ \sum_{k=1}^{m} c_k & \sum_{k=1}^{m} d_k \end{matrix}$$

$$a = \sum_{k=1}^{m} a_k; \quad b = \sum_{k=1}^{m} b_k; \quad c = \sum_{k=1}^{m} c_k; \quad d = \sum_{k=1}^{m} d_k$$

$$T_k = a_k + b_k + c_k + d_k$$

Coefficienti di affinità/binarie

In questo contesto assume grande rilevanza il problema di come considerare la assenza congiunta.

Se il fatto di non possedere un attributo è irrilevante ai fini della somiglianza allora la cella "d" non deve entrare nella misura.

Se invece la dicotomia è fra due stati complementari aventi uguale rilevanza allora "a" e "d" entrano nello stesso modo nell'indice. Nel primo caso si parla di varabili binarie asimmetriche e nel secondo le variabili sono dette binarie simmetriche.

Il leone ed il coniglio sono privi di ali, ma questo non li rende più affini.

Esempi di coefficienti per binarie

Coefficiente	*Tipo*	*Formula*
Jaccard	*T*	$\frac{a}{a+b+c}$
Andenberg	*T*	$\frac{a}{a+2(b+c)}$
Czekanowsi	*T*	$\frac{2a}{2a+b+c}$
Ochiai	*T*	$\frac{a}{\sqrt{(a+b)(a+c)}}$
Sokal – Sneath	*S*	$\frac{a+d}{a+0.5b+0.5c+d}$
Hamann	*S*	$\frac{(a+d)-(b+c)}{a+b+c+d}$
Rogers – Tanimoto	*S*	$\frac{a+d}{a+2b+2c+d}$
Simple Matching	*S*	$\frac{a+d}{a+b+c+d}$
Russell – Rao	*S*	$\frac{a}{a+b+c+d}$

Se la compresenza o la coassenza nell'attributo sono tanto peculiari nel generare molta somiglianza tra le due unità allora "a" e/o "d" devono entrare nel coefficiente con peso doppio o comunque maggiore degli altri.

Nella tabella sono distinti gli indici che non includono il conteggio della coassenza (T) con quelli che la includono (S).

In caso di valori nella forma 0/0 si può porre, per convenzione, il coefficiente pari a zero.

I coefficienti qui inseriti variano tra zero ed uno e generano una matrice delle distanze euclidea.

Esempio

Entità	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
A	1	1	0	1	0	1	0	0	1	1
B	0	0	0	0	0	1	1	1	1	1
C	1	0	1	0	1	0	1	0	1	0
D	1	0	1	1	0	0	1	0	1	0
E	0	1	0	1	0	1	0	1	0	1
F	1	1	1	1	1	0	0	0	0	0
G	1	0	0	1	0	0	0	0	1	0
H	0	0	1	1	1	0	0	1	0	0

La matrice di dissimilarità risulta diversa. Speso è molto diversa.

Jaccard

	A	B	C	D	E	F	G	H
A	0.0000	0.7906	0.8819	0.7906	0.6547	0.7906	0.7071	0.9428
B	0.7906	0.0000	0.8660	0.8660	0.7559	1.0000	0.9258	0.9354
C	0.8819	0.8660	0.0000	0.5774	1.0000	0.7559	0.8165	0.8452
D	0.7906	0.8660	0.5774	0.0000	0.9428	0.7559	0.6325	0.8452
E	0.6547	0.7559	1.0000	0.9428	0.0000	0.8660	0.9258	0.8452
F	0.7906	1.0000	0.7559	0.7559	0.8660	0.0000	0.8165	0.7071
G	0.7071	0.9258	0.8165	0.6325	0.9258	0.8165	0.0000	0.9129
H	0.9428	0.9354	0.8452	0.8452	0.8452	0.7071	0.9129	0.0000

Simple matching

	A	B	C	D	E	F	G	H
A	0.0000	0.7071	0.8367	0.7071	0.5477	0.7071	0.5477	0.8944
B	0.7071	0.0000	0.7746	0.7746	0.6325	1.0000	0.7746	0.8367
C	0.8367	0.7746	0.0000	0.4472	1.0000	0.6325	0.6325	0.7071
D	0.7071	0.7746	0.4472	0.0000	0.8944	0.6325	0.4472	0.7071
E	0.5477	0.6325	1.0000	0.8944	0.0000	0.7746	0.7746	0.7071
F	0.7071	1.0000	0.6325	0.6325	0.7746	0.0000	0.6325	0.5477
G	0.5477	0.7746	0.6325	0.4472	0.7746	0.6325	0.0000	0.7071
H	0.8944	0.8367	0.7071	0.7071	0.7071	0.5477	0.7071	0.0000

Dalla affinità alla dissimilarità

Ad ogni coefficiente di affinità/prossimità è associato un indice di dissimilarità o dissomiglianza o distanza.

Se a_{ij} è simmetrico e non negativo allora la dissomiglianza d_{ij} dovrà avere la stessa proprietà; inoltre, deve diminuire quando la prima aumenta e viceversa.

Per indici di affinità normalizzati (0,1) la trasformazione in una dissimilarità è semplice

1. $\delta_{ij} = 1 - a_{ij}^{\alpha}$
2. $\delta_{ij} = \left(1 - a_{ij}\right)^{\alpha} \quad \alpha > 0$
3. $\delta_{ij} = \frac{\ln\left(1 + a_{ij}\right) - \ln(2)}{-\ln(2)}$
4. $\delta_{ij} = \frac{\alpha - a_{ij}}{\alpha + a_{ij}}$
5. $\delta_{ij} = \frac{e^{\left(1-a_{ij}\right)} - 1}{e - 1}$

Ogni funzione f definita nell'intervallo (0,1), tale che f(0)=1 e f(1)=0 e che in tale intervallo abbia derivata negativa è idonea a trasformare un indice di affinità in un indice di dissomiglianza.

Affinità/dissimilarità in ambiente R

stats, vegan, ade4, cluster, FD, arules.

Although the literature provides similarity as well as distance measures, in R all similarity measures are converted to distances to compute a square matrix of class "dist" in which the diagonal (distance between each object and itself) is 0 and can be ignored.

The conversion formula varies with the package used, and this may not be without consequences:

In stats, FD and vegan, the conversion from similarities S to dissimilarities D is D = 1 − S.

In ade4, it is computed as $D = (1- S)^{0.5}$. This allows some indices to become Euclidean

Distance matrices computed by other packages that are not Euclidean can often be made Euclidean by computing D <- sqrt(D).

In cluster, all available measures are distances, so no conversion has to be made.

Esempio: SPECTF heart data

The dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images.

Each of the patients is classified into two categories: normal and abnormal.

The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images.

Package ade4: simple matching

	1	2	3	4	5
1	0.0000	0.2182	0.5774	0.6172	0.4880
2	0.2182	0.0000	0.5345	0.6547	0.5345
3	0.5774	0.5345	0.0000	0.6547	0.5345
4	0.6172	0.6547	0.6547	0.0000	0.5774
5	0.4880	0.5345	0.5345	0.5774	0.0000

Package stats: Russell Rao

	1	2	3	4	5
1	0.0000	0.1667	0.7778	1	0.7143
2	0.1667	0.0000	0.6667	1	0.7500
3	0.7778	0.6667	0.0000	1	0.7500
4	1.0000	1.0000	1.0000	0	1.0000
5	0.7143	0.7500	0.7500	1	0.0000

Package vegan, jaccard

	1	2	3	4	5
1	0.0000	0.4366	0.4513	0.2975	0.4951
2	0.4366	0.0000	0.2997	0.5395	0.3659
3	0.4513	0.2997	0.0000	0.3633	0.4436
4	0.2975	0.5395	0.3633	0.0000	0.6109
5	0.4951	0.3659	0.4436	0.6109	0.0000

Politome sconnesse o multistato

Il dominio è formato da modalità che distinguono gli aspetti, le categorie, gli attributi che le unità possiedono in vario modo senza che tra le modalità possa essere stabilito un ordinamento univoco in termini quantitativi.

Il livello di misura della variabile è tale che, date due qualsiasi modalità: x_r, x_s, è possibile solo affermare che:

$$x_r = x_s \quad \text{oppure} \quad x_r \neq x_s$$

Le modalità hanno qui la sola funzione di etichettare le unità per formarne una lista o per raggrupparle in categorie omogenee.

Le differenze tra le unità possono essere accertate, ma non ordinate né misurate.

L'ordinamento alfabetico con cui sono spesso presentate le modalità semplifica l'esposizione, ma non stabilisce una gerarchia.

Politome sconnesse o multistato/2

Supponiamo che la variabile X sia rilevata con il dominio.

$$S(X) = \{x_1, x_2, \ldots, x_m\}$$

Per ciascun confronto tra due unità può verificarsi una qualsiasi delle combinazioni di modalità

$$\{(x_r, x_s),\ r, s = 1, 2, \ldots, n\}$$

Per ciascuna combinazione di modalità occorre esprimere un giudizio di somiglianza: a_{ij}

	x_1	x_2	…	x_j	…	x_m
x_1	1	a_{12}	…	a_{1j}	…	a_{1m}
x_2	a_{21}	1	…	a_{2j}	…	a_{2m}
…	…	…	…	…	…	…
x_i	a_{i1}	a_{i2}	…	1	…	a_{im}
…	…	…	…	…	…	…
x_m	a_{m1}	a_{m2}	…	a_{mj}	…	1

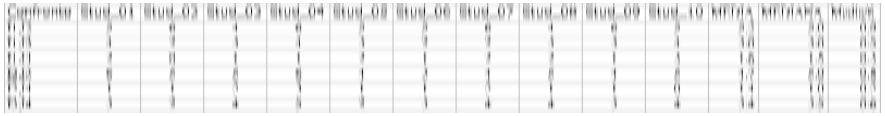
Le entrate a_{ij} esprimono la valutazione della affinità tra due unità in cui si verificano le due modalità della coppia considerata.

Non è necessario che verifichino la disuguaglianza triangolare, ma è preferibile

Esempio

Ad un gruppo di 10 studenti scelti accuratamente per posizione curricolare, estrazione sociale, formazione secondaria, etc. è stato chiesto di esprimere un giudizio (voto da 1 a 6) su di un gruppo di quattro insegnanti:

0= diametralmente opposti e 6= del tutto equivalenti



La sintesi dei giudizi può avvenire con la media aritmetica, con la mediana, il voto minimo, il voto massimo, etc.

Qui si è scelta la media aritmetica, rapportata a 6 per avere numeri tra zero ed uno. A questo punto la matrice dei giudizi è pronta per comparare due qualsiasi soggetti

	$I1$	$I2$	$I3$	$I4$
$I1$	1	0.87	0.50	0.27
$I2$		1	0.78	0.23
$I3$			1	0.57
$I4$				1

Coefficiente di Beijnen

Un coefficiente che risponde al requisito di disuguaglianza triangolare è stato proposto da Beijnen (1973).

$$a_{ij} = \sqrt{\frac{\sum_{r=1}^{m} \alpha\left(x_{ri}, x_{rj}\right)}{m}} \qquad con \quad \alpha\left(x_{ri}, x_{rj}\right) = \begin{cases} \frac{L_r}{p} & se \quad x_{ri} = x_{rj} \\ 0 & altrimenti \end{cases}$$

dove L_r è il numero di modalità previsto per la variabile multistato r-esima con totale:

$$p = \sum_{r=1}^{m} L_r$$

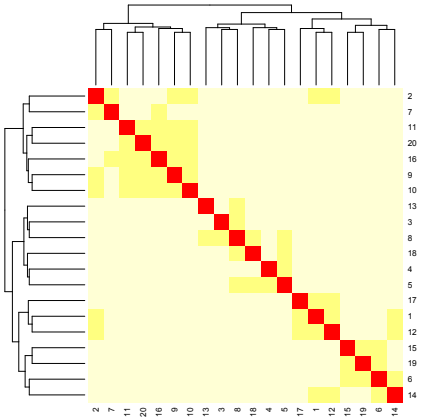
Il coefficiente proposto somma una frazione pari al numero di modalità della variabile politoma rispetto a tutte le modalità delle diverse politomie presenti nel data set.

L'affinità è tanto maggiore quanto maggiore è il numero di stati nella politomia.

Esempio

Consideriamo dei dati simulati su n=15 unità e su k=3 politomie con 3, 4, 5 diversi stati.

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15]
[1,] 0.00 0.91 1.00 0.96 0.96 0.96 0.96 1.00 0.96 0.96 1.00 0.87 1.00 0.91 0.96
[2,] 0.91 0.00 1.00 1.00 1.00 0.96 0.91 1.00 0.91 0.91 0.96 0.91 1.00 0.96 0.96
[3,] 1.00 1.00 0.00 1.00 0.96 0.96 1.00 0.91 1.00 1.00 1.00 1.00 0.96 1.00 0.96
[4,] 0.96 1.00 1.00 0.00 0.91 1.00 1.00 0.96 1.00 1.00 1.00 0.96 0.96 0.96 0.96
[5,] 0.96 1.00 0.96 0.91 0.00 1.00 1.00 0.91 1.00 1.00 1.00 0.96 0.96 0.96 1.00
[6,] 0.96 0.96 0.96 1.00 1.00 0.00 0.96 0.96 1.00 1.00 0.96 0.96 0.96 0.91 0.91
[7,] 0.96 0.91 1.00 1.00 1.00 0.96 0.00 1.00 0.96 0.96 0.96 0.96 0.96 0.96 0.96
[8,] 1.00 1.00 0.91 0.96 0.91 0.96 1.00 0.00 1.00 1.00 1.00 1.00 0.91 1.00 0.96
[9,] 0.96 0.91 1.00 1.00 1.00 1.00 0.96 1.00 0.00 0.87 0.91 0.96 1.00 1.00 1.00
[10,] 0.96 0.91 1.00 1.00 1.00 1.00 0.96 1.00 0.87 0.00 0.91 0.96 1.00 1.00 1.00
[11,] 1.00 0.96 1.00 1.00 1.00 0.96 0.96 1.00 0.91 0.91 0.00 1.00 1.00 0.96 1.00
[12,] 0.87 0.91 1.00 0.96 0.96 0.96 0.96 1.00 0.96 0.96 1.00 0.00 1.00 0.91 0.96
[13,] 1.00 1.00 0.96 0.96 0.96 0.96 0.96 0.91 1.00 1.00 1.00 1.00 0.00 1.00 0.96
[14,] 0.91 0.96 1.00 0.96 0.96 0.91 0.96 1.00 1.00 1.00 0.96 0.91 1.00 0.00 0.96
[15,] 0.96 0.96 0.96 0.96 1.00 0.91 0.96 0.96 1.00 1.00 1.00 0.96 0.96 0.96 0.00
[16,] 1.00 0.96 1.00 1.00 1.00 1.00 0.91 1.00 0.91 0.91 0.91 1.00 0.96 1.00 1.00
[17,] 0.91 0.96 0.96 0.96 0.96 1.00 1.00 1.00 0.96 0.96 1.00 0.91 1.00 0.96 1.00
[18,] 1.00 0.96 0.96 0.96 0.91 1.00 0.96 0.91 0.96 0.96 0.96 1.00 0.96 1.00 1.00
[19,] 0.96 0.96 0.96 0.96 1.00 0.91 0.96 0.96 1.00 1.00 1.00 0.96 0.96 0.96 0.87
[20,] 1.00 0.96 1.00 1.00 1.00 0.96 0.96 1.00 0.91 0.91 0.87 1.00 1.00 0.96 1.00
```



La tecnica heatmap raggruppa le unità con maggiore prossimità

Scomposizione della politomia in variabili binarie

La rilevazione della variabile politoma è suddivisa in vari confronti binari (tante quanto sono le modalità del suo dominio.

$$S(X) = \{x_1,\ x_2,\ \ldots,\ x_m\}$$

l'accertamento della modalità presentata da una qualsiasi unità si realizza valutando una sequenza di valori binari (ad esempio zero ed uno

	X_1	X_2	...	X_i	...	X_m
Presente	1	1	...	1	...	1
Assente	0	0	...	0	...	0

Il confronto di due unità generiche: i e j passa per m variabili binarie di tipo asimmetrico (perché manca la complementarità)

Politomia in variabili binarie

In sintesi, per ogni indicatore di stato del dominio S si presenta una delle quattro configurazioni: 00, 01, 10, 11 che poi vanno adeguatamente sommate come si è fatto per il confronto su variabili binarie autonome.

	u_j	
u_i	X_1	1	0
	1	a	b
	0	c	d

	u_j	
u_i	X_2	1	0
	1	a	b
	0	c	d

$\bullet\bullet\bullet$

	u_j	
u_i	X_k	1	0
	1	a	b
	0	c	d

Dove $X_1, X_2, \ldots, X_k$ non sono variabili vere e proprie, ma pseudo-variabili derivate dalle modalità della politomia.

A queste tabelle, si può applicare uno degli indici di somiglianza tra variabili binarie.

Esempio

Consideriamo dei dati simulati su n=15 unità e su k=3 politomie a diversi stati

unità	X1	X2	X3	Range	Xa1	Xa2	Xa3	Xb1	Xb2	Xb3	Xb4	Xc1	Xc2	Xc3	Xc4	Xc5
u_1	2	3	1	X1=1:3	0	1	0	0	0	1	0	1	0	0	0	0
u_2	3	3	1	X2=1:4	0	0	1	0	0	1	0	1	0	0	0	0
u_3	1	2	2	X3=1:5	1	0	0	0	1	0	0	0	0	0	0	0
u_4	2	1	4		0	1	0	1	0	0	0	0	0	0	1	0
u_5	2	1	2		0	1	0	1	0	0	0	0	0	0	0	0
u_6	1	3	5		1	0	0	0	0	1	0	0	0	0	0	1
u_7	3	3	3		0	0	1	0	0	1	0	0	0	1	0	0
u_8	1	1	2		1	0	0	1	0	0	0	0	0	0	0	0
u_9	3	4	1		0	0	1	0	0	0	1	1	0	0	0	0
u_10	3	4	1		0	0	1	0	0	0	1	1	0	0	0	0
u_11	3	4	5		0	0	1	0	0	0	1	0	0	0	0	1
u_12	2	3	1		0	1	0	0	0	1	0	1	0	0	0	0
u_13	1	1	3		1	0	0	1	0	0	0	0	0	1	0	0
u_14	2	3	5		0	1	0	0	0	1	0	0	0	0	0	1
u_15	1	3	4		1	0	0	0	0	1	0	0	0	0	1	0

Le tre politomie sono diventate 15 dicotomie e la similarità tra i soggetti pu essere calcolata con uno dei coefficienti di tipo T ad esempio lo Jaccard o lo Ochiai.

Esempio/continua

Con il frazionamento in binarie e coefficiente di Jaccard

```
     [,1]   [,2]   [,3]   [,4]   [,5]
[1,] 0.0000 0.7071 1.0000 0.8944 0.8944
[2,] 0.7071 0.0000 1.0000 1.0000 1.0000
[3,] 1.0000 1.0000 0.0000 1.0000 0.8944
[4,] 0.8944 1.0000 1.0000 0.0000 0.7071
[5,] 0.8944 1.0000 0.8944 0.7071 0.0000
```

Coefficiente di Beijnen

```
     [,1] [,2] [,3] [,4] [,5]
[1,] 0.00  0.5 1.00 0.75 0.75
[2,] 0.50  0.0 1.00 1.00 1.00
[3,] 1.00  1.0 0.00 1.00 0.75
[4,] 0.75  1.0 1.00 0.00 0.50
[5,] 0.75  1.0 0.75 0.50 0.00
```

La dicotomizzazione delle politomie genera molte variabili binarie e potrebbe ingenerare somiglianze anche nei confronti di unità in cui queste siano assenti.

Inoltre, nel confronto di unità rispetto distanze multiscala (che studieremo più avanti) potrebbe esagerare oltre misura il ruolo delle variabili binarie.

Affinità per i ranghi

Riguarda le variabili riportate in scala quantitativa ordinale.

- **Perché non esiste una vera misura, ma solo un punteggio o valutazione**
- **Perché le misurazioni su sono imprecise o viziate da errore**
- **Perché sono presenti dei valori remoti**

Le modalità sono poste in corrispondenza con dei numeri naturali (ranghi)

Per ogni unità si osserva una coppia di modalità che si trasforma poi in una coppia di ranghi

$$(X_i, Y_i) \longrightarrow (\pi_i, \sigma_i)$$

E' possibile che i ranghi siano rilevati direttamente come risposte ad un quantificatore verbale

Due tipologie

Le situazioni che coinvolgono i ranghi sono di due tipi

Politome ordinate.
La o le variabili hanno come dominio un insieme di numeri naturali ovvero quantificatori verbali descritti con dei numeri.

Su ogni unità si rileva una singola modalità

Graduatorie.
Ciò che si rileva sulle unità è una graduatoria che considera un gruppo prefissato di oggetti.

La graduatoria è una singola variabile a se, anche se si presenta con diverse modalità in forma di ranghi

Applicazione politome ordinate

Response Pattern	No of Cases
2 2 2 2 2 2	97
3 3 2 3 3 3	70
3 2 2 2 2 2	49
3 3 3 3 3 3	45
3 3 2 2 2 2	45
3 2 2 3 3 3	40
3 3 2 3 2 2	32
3 3 2 3 2 3	31
2 2 1 2 2 2	25
1 1 1 1 1 1	23
3 2 2 3 2 2	20
2 2 1 1 1 1	18
3 3 2 2 2 3	18
2 3 2 2 2 2	17
3 2 2 3 2 3	16
3 3 3 3 2 3	16
3 3 2 3 3 2	15
3 2 3 3 3 3	15
2 1 2 2 2 2	14
2 2 2 3 2 2	13

agree strongly (1), agree (2), disagree (3), disagree strongly (4) per 6 domande sulla partecipazione alla vita politica sottoposte a 1554 soggetti.

NOSAY	People like me have no say in what the government does
VOTING	Voting is the only way that people like me can have any say about how the government runs things
COMPLEX	Sometimes politics and government seem so complicated that a person like me cannot really understand what is going on
NOCARE	I don't think that public officials care much about what people like me think
TOUCH	Generally speaking, those we elect to Congress in Washington lose touch with the people pretty quickly
INTEREST	Parties are only interested in people's votes but not in their opinions description

Le combinazioni di risposte possibili sono 4^6=4096, ma molte sono ripetute ed altre non si sono realizzate.

La matrice delle affinità potrà avere dimensioni 1554x1554 se si misura l'affinità tra i soggetti, oppure 6 x 6 se si misura l'affinità tra le domande.

Appliczione politome ordinate/2

Per ogni unità si osserva un valore ordinale (o rango) di una variabile con scala di misurazione che consente di stabilire delle priorità univoche tra le modalità osservate.

I valori riportati sono dei numeri interi che variano in un intervallo limitato e valori diversi indicano stati diversi in cui uno precede l'altro.

Detersivi. Graduazione a somma costante

Attributo	Giudice1	Giudice2	Giudice3
Fragranza	19	23	27
Confezione	14	16	18
Prezzo	31	24	28
Detergenza	26	17	19
Schiumosita	9	11	6
Consistenza	1	9	2
	100	100	100

Gli attributi sono delle unità e i giudici sono delle variabili

Affinità/dissimilarità per politome ordinate

Gli indici che si possono utilizzar si basano sugli scarti tra modalità

Arcavata distance

$$d_{i,j} = \left(\frac{1}{m}\right)\left[\sum_{i=1}^{m}\left(\frac{|\pi_i - \sigma_i|}{d_i}\right)^{\alpha}\right]^{\frac{1}{\alpha}} \quad dove \begin{cases} \pi_i \ \ rango \ \ i-esimo \ \ di \ \ A \\ \sigma_i \ \ rango \ \ i-esimo \ \ di \ \ B \\ d_i \ \ scarto \ \ massimo \ \ possibile \ \ in \ \ i-esima \ \ posizione \end{cases}$$

Il coefficiente varia tra zero (le due entità coincidono esattamente su tutte le variabili di questa tipologia) ed uno (per ogni variabile si riscontra lo scarto massimo ammesso per quella variabile.

L'equivalente misura della affinità si ottiene considerando il complemento ad uno del coefficiente di dissimilarità

$$a_{i,j} = 1 - \left(\frac{1}{m}\right)\left[\sum_{h=1}^{m}\left(\frac{|\pi_h - \sigma_h|}{d_h}\right)^{\alpha}\right]^{\frac{1}{\alpha}}$$

Esempio

Consideriamo due soggetti che abbiano dato le seguenti risposte su m=6 domande politome ordinare con 4 modalità potenziali

3	2	2	3	2	1	A
1	2	3	3	3	3	B

Come si può misurare l'affinità/vicinanza?

Qui non sembra logico riscalare le risposte in una graduatoria unica del tipo (1,2,…,n) dato che si tratta di domande diverse.

Si può utilizzare il valore numerico del rango e valutare la quantità:

$$d_{i,j} = \left(\frac{1}{6}\right)\sum_{i=1}^{6}\left(\frac{|\pi_i - \sigma_i|}{d_i}\right) \quad dove \quad \alpha = 1, \ m = 6$$

$$\frac{|3-1|}{4-1} + \frac{|2-2|}{4-1} + \frac{|2-3|}{4-1} + \frac{|3-3|}{4-1} + \frac{|2-3|}{4-1} + \frac{|1-3|}{4-1} =$$

$$\frac{2}{3} + \frac{0}{3} + \frac{1}{3} + \frac{0}{3} + \frac{1}{3} + \frac{2}{3} = \frac{6}{3} = 2 \Rightarrow \frac{2}{6} = \frac{1}{3}; \ 1 - \frac{1}{3} = \frac{2}{3}$$

Distanza

Situazione di studio

Un insieme fisso di n oggetti è graduato da un giudice rispetto a due specifici attributi ovvero due giudici graduano gli n oggetti rispetto ad un solo attributo.

Lo stesso scenario vale per lo studio di variabili metriche poi trasformate in ranghi

Ci troviamo di fronte due permutazioni dei primi n numeri naturali

$$\begin{array}{cccccccc} 1 & 2 & \dots & i & \dots & n-1 & n & \\ \sigma_1 & \sigma_2 & \dots & \sigma_i & \dots & \sigma_{n-1} & \sigma_n & X_1 \\ \pi_1 & \pi_2 & \dots & \pi_i & \dots & \pi_{n-1} & \pi_n & X_2 \end{array}$$

Sono possibili le parità (ex aequo)

Convenience Store	Distance from CAM (m)	Rank distance	Price of 50cl bottle (€)	Rank price
1	50	10	1.80	2
2	175	9	1.20	3.5
3	270	8	2.00	1
4	375	7	1.00	6
5	425	6	1.00	6
6	580	5	1.20	3.5
7	710	4	0.80	9
8	790	3	0.60	10
9	890	2	1.00	6
10	980	1	0.85	8

Un indice di correlazione tra ranghi riassume l'intensità e la direzione del legame tra le due graduatorie.

Esempio

Un gruppo di clienti di una banca classificato per reddito e per importo del prestito. Convertiamo i valori osservati in ranghi.

Cliente	Reddito (x)	Rango (x)	Prestito (y)	Rango (y)
A	25 600	6	8 600	5
B	17 800	9	8 800	4
C	167 200	1	500	8
D	44 200	3	6 600	6
E	36 400	4	10 500	3
F	27 400	5	74 400	1
G	83 600	2	0	9
H	18 600	8	6 300	7
I	24 500	7	12 100	2

E' evidente la perdita di informazione. Lo scarto tra i ranghi in X per i clienti H ed I è 9-7=3 e sarebbe questo per qualunque coppia di valori compresi tra 18'600 e 24'500.

In breve, conoscere i ranghi poco ci dice sui valori originari

Altro esempio

Ad un campione di soggetti è sottoposto un elenco di situazioni o item in cui debbono indicare il grado di problematicità ovvero assegnare un livello di priorità:

K.	**Aumentare l'occupazione**	**R.**	**Assistenza agli anziani**
A.	**Dare speranze ai giovani**	**C.**	**Tutelare l'infanzia**
J.	**Ridurre la criminalità**	**L.**	**Costruire case a basso costo**
L.	**Ridurre l'orario di lavoro**	**W.**	**Dare un salario ai disoccupati**
B.	**Migliorare la sanità**	**F.**	**Investire in ricerca scientifica**
Z.	**Incrementare la solidarietà**	**H.**	**Dare dignità alla condizione umana**
D.	**Eliminare la povertà**	**T.**	**Migliorare il sistema carcerario**

Ogni soggetto esprime una gerarchia (per comodità, crescente) delle posizioni che gli item occupano sulla propria scala di priorità.

Sono possibili le parità di posizione.

C'è un potenziale di 14!=87'178'291'200 possibili configurazioni

U1) 10 8 1 2 14 6 12 7 4 9 3 5 11 13
U2) 5 4 14 11 7 2 13 10 9 3 8 6 12 1
U3) 4 12 9 8 3 7 10 5 13 6 1 2 11 14

Requisiti degli indici

1. *Normalizzazione*: $-1 \le r(\sigma,\pi) \le 1$

   $con\ \ r(\sigma,\sigma) = r(\pi,\pi) = 1,\ r(\sigma,\sigma^*) = r(\pi,\pi^*) = -1$

   $dove\ \ \pi^* = n+1-\pi,\ \sigma^* = n+1-\sigma$
2. *Simmetria*: $r(\sigma,\pi) = r(\pi,\sigma)$
3. *Antisimmetria dopo l'antiteticità*: $r(\sigma,\pi^*) = r(\sigma^*,\pi) = -r(\sigma,\pi)$
4. *Invarianza a destra*: $r(\sigma,\pi) = r[\sigma(\theta),\pi(\theta)]\ \ \forall\ \sigma,\pi,\theta$
5. *Valore atteso nullo in caso di indipendenza*

   $E_{\sigma,\pi\in S_n}[r(\sigma,\pi)] = 0,\ dove\ \ S_n\ \ è\ \ l'insieme\ \ delle\ \ n!\ \ permutazioni$

I desiderata in questo elenco non sono tutti ritenuti cogenti. Ad esempio gli indici top-rank sono asimmetrici in quanto pesano di più i confronti per i ranghi iniziali della graduatoria (web-search)

Importanza dei requisiti

La presenza dei due limiti consente di valuare l'indice rispetto alle situazioni estreme a -1 e 1 assegnate a situationi estreme.

E' poco utile un indice che cambia valore secondo l'ordine di considerazione delle permutazioni

$$r(\sigma,\pi) = r(\pi,\sigma)$$

E' poco utile un indice che cambia valore oltre che segno se il confronto avviene con la permutatione antitetica

$$r(\sigma,\pi^*) = r(\sigma^*,\pi) = -r(\sigma,\pi)$$

Il valore dell'indice non deve cambiare se cambia l'ordine di considerazione delle varie coppie di ranghi

$$(\sigma_i,\pi_i),\ i = 1,2,\ldots,n$$

Se le n! permutazioni sono equiprobabili l'indice deve avere aspettativa zero altrimenti i test risultano viziati

Indici ammissibili

Spearman (r_1)	$\left(\frac{3}{n^3-n}\right)\left\{\sum_{i=1}^{n}\lvert\pi_i^*-\sigma_i\rvert^2-\sum_{i=1}^{n}\lvert\pi_i-\sigma_i\rvert^2\right\}$	1906
Gini (r_2)	$\left(\frac{1}{\lfloor n^2/2\rfloor}\right)\left\{\sum_{i=1}^{n}\lvert\pi_i^*-\sigma_i\rvert-\sum_{i=1}^{n}\lvert\pi_i-\sigma_i\rvert\right\}$	1914
Kendall (r_3)	$\left(\frac{1}{n^2-n}\right)\sum_{i=1}^{n} sign\left\{(\sigma_i-\sigma_j)sign(\pi_i-\pi_j)\right\}$	1938
Blomqvist	$\left(\frac{1}{n}\right)\sum_{i=1}^{n} sign\left\{[\sigma_i-median(\sigma)][\pi_i-median(\pi)]\right\}$	1950
Gideon – Hollister	$\left(\frac{1}{\lfloor n/2\rfloor}\right)\left\{\max_{1\le i\le n}\sum_{j=1}^{n}\left[I(\sigma_j\le n-\pi_i)\right]-\max_{1\le i\le n}\sum_{j=1}^{n}\left[I(\sigma_j>\pi_i)\right]\right\}$	1987

Robustezza: l'indice rimane stabile se i dati da cui derivano i ranghi cambiano poco.

Sensitività: l'indice è in grado di differenziare permutazioni diverse sia pure somiglianti. r_1 r_2 r_3 hanno un range di valori ridotto risultando poco informativi. Gli ultimi due indici sono troppo "robusti" per essere utili in molti contesti.

ρ (rho) di Spearman

La misura forse più popolare della dipendenza tra i ranghi è la seguente

$$r_1 = \frac{\sum_{i=1}^{n}\left(\pi_i-\frac{n+1}{2}\right)\left(\sigma_i-\frac{n+1}{2}\right)}{\sqrt{\sum_{i=1}^{n}\left(\pi_i-\frac{n+1}{2}\right)^2\sum_{i=1}^{n}\left(\sigma_i-\frac{n+1}{2}\right)^2}}$$

Caso delle n coppie di valori senza posizioni di parità.

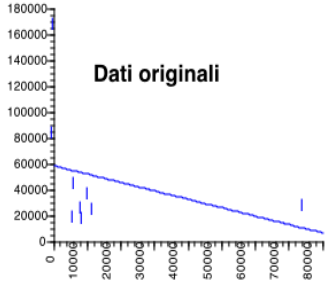
La definizione di r_1 è la stessa del coefficiente di correlazione. Comunque il particolare tipo di dati coinvolti consente delle semplificazioni. Ad esempio:

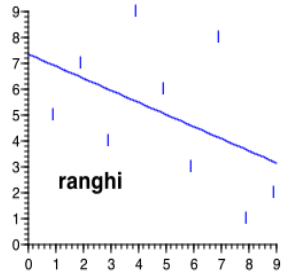
$$r_S = 1-\frac{6\sum_{i=1}^{n}(r_i-s_i)^2}{n(n^2-1)}$$

Esempio

Cliente	Reddito	r_i	Prestito	s_i	$(r_i - s_i)^2$
A	25600	6	8600	5	1
B	17800	9	8800	4	25
C	167200	1	500	8	49
D	44200	3	6600	6	9
E	36400	4	10500	3	1
F	27400	5	74400	1	16
G	83600	2	0	9	49
H	18600	8	6300	7	1
I	24500	7	12100	2	25
					176

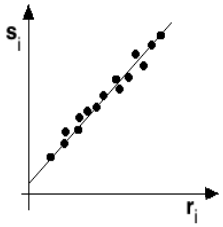
$$r_S = 1 - \frac{6*176}{9*(81-1)} = -0.4667$$

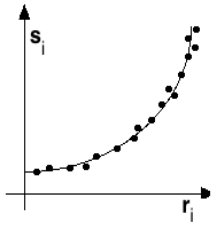




Considerazione sul rho di Spearman

- Per costruzione l'indice rho varia tra -1 ed 1
- Misura la dipendenza monotonica tra le due variabili. Assume il valore massimo (minimo) quando gli ordinamenti sono perfettamente concordi (discordi)





$r_s = \pm 1$ se tra la Y e X esiste una relazione monotona, crescente o decrescente, senza vincolo di linearità

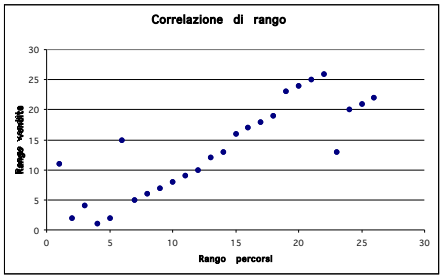
- Rho mette tutte le osservazioni sullo stesso piano (considera solo l'ordine) e perciò annulla l'effetto dei valori remoti
- L'indice rimane invariato se una o entrambe le variabili subiscono una trasformazione monotona, ad esempio Y'=Log(Y)e/o X'=exp(X)

Esempio

Unità	X Percorsi	Y Vendite	Rank(X)	Rank(Y)
A	121.5	373	21	25
B	151.5	314	25	21
C	146.2	301	24	20
D	106.7	263	16	17
E	98.9	204	11	9
F	95.1	176	9	7
G	90.1	138	4	1
H	115.5	329	19	23
I	71.7	225	1	11
J	111.7	300	18	19
K	93.6	164	7	5
L	109.6	284	17	18
M	105.3	252	15	16
N	125.0	400	22	26
O	91.7	239	6	15
P	88.7	161	3	4
Q	101.9	226	13	12
R	162.3	322	26	22
S	96.4	185	10	8
T	90.7	143	5	2
U	100.0	212	12	10
V	102.6	232	14	13
X	94.5	171	8	6
Y	88.6	143	2	2
W	119.4	358	20	24
Z	142.9	232	23	13

rho=	0.850084703
gdl=	25
tc=	7.907679188
p-Value	2.90161E-08

Venditori porta-a-porta per vendite e km percorsi



La correlazione è elevata sebbene si notino diversi disturbi

tau di Kendall

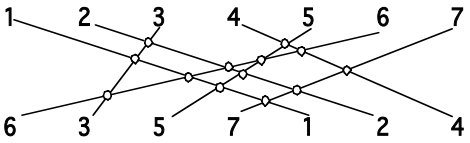
E' una misura alternativa di dipendenza tra ranghi

$$\tau = 1 - \frac{4C}{n(n-1)} \quad con \quad -1 \le \tau \le 1$$

"C" è il numero minimo di scambi necessari per trasformare una graduatoria nell'altra. Gli estremi sono interpretabili come nel rho di Spearman

ESEMPIO

Calcolo con il netodo di Holmes (1920)



Le linee che congiungono i ranghi nelle due graduatorie si incrociano C volte

$$\tau = 1 - \frac{4(13)}{7(6)} = 1 - 1.2381 = -0.2381$$

Presenza di parità

Se i ranghi graduano misure soggette a errori è utile dare rango eguale a valori molto prossimi.

Anche in caso di giudizi succede di non riuscire a stabilire una preferenza tra due aspetti egualmente percepiti.

Le parità o ex aequo richiedono l'uso di pseudo-ranghi nelle posizioni coinvolte e modifiche nelle formule di calcolo

e	σ	s/r.m.	σ/max
1	1	1	1
2	2	3	4
3	2	3	3
4	2	3	2
5	3	5	5
6	4	6	6
7	5	7	7
8	6	8.5	9
9	6	8.5	8
10	7	11.5	13
11	7	11.5	12
12	7	11.5	11
13	7	11.5	10
14	8	14	14
15	9	15	15

Esempi:

1) Rango medio

2) Sub-graduatoria che rende massima o minima la correlazione di rango. Media tra le due

3) Boostrap di sub-permutazioni

Il problema è nella determinazione una tantum del massimo possibile in caso di ex aequo

Formula di rho in caso di parità

$$r_1 = \frac{\left(n^3-3\right) - 6\sum_{i=1}^{n} d_i^2 - \frac{1}{2}\left\{\sum_{j=1}^{n_x}\left[\left(t_j^x\right)^3 - \left(t_j^x\right)\right] + \sum_{j=1}^{n_y}\left[\left(t_j^y\right)^3 - \left(t_j^y\right)\right]\right\}}{\sqrt{\left[\left(n^3-3\right) - \sum_{j=1}^{n_x}\left[\left(t_j^x\right)^3 - \left(t_j^x\right)\right]\right]\left[\left(n^3-3\right) - \sum_{j=1}^{n_y}\left[\left(t_j^y\right)^3 - \left(t_j^y\right)\right]\right]}}$$

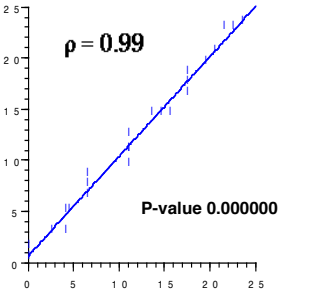
$$dove \begin{cases} n_x = numero\ \ di\ \ gruppi\ \ di\ \ X\ \ con\ \ parità \\ t_j^x = numero\ \ di\ \ valori\ \ uguali\ \ per\ \ la\ \ j-esima\ \ parità\ \ in\ \ X \\ n_y = numero\ \ di\ \ gruppi\ \ di\ \ Y\ \ con\ \ parità \\ t_j^y = numero\ \ di\ \ valori\ \ uguali\ \ per\ \ la\ \ j-esima\ \ parità\ \ in\ \ Y \end{cases}$$

Esempio

E	E-or	r_E	B	B-or	r_B	$(r_E)^2$	$(r_B)^2$	$r_E * r_B$
10.3	9.2	0.5	-1.3	-1.8	1	0.25	1	0.5
9.9	9.2	0.5	-1.1	-1.6	2	0.25	4	1
10.3	9.3	3.0	1.3	-1.5	3.5	9	12.25	10.5
9.7	9.4	4.5	0.5	-1.5	3.5	20.25	12.25	15.75
9.6	9.4	4.5	1.8	-1.3	5.5	20.25	30.25	24.75
9.5	9.5	5.0	-0.4	-1.3	5.5	25	30.25	27.5
10.8	9.6	7.0	-1.6	-1.1	7	49	49	49
9.7	9.6	7.0	-1.8	-0.8	8	49	64	56
9.4	9.6	7.0	1.0	-0.7	9	49	81	63
10.1	9.7	11.5	-1.5	-0.5	10	132.25	100	115
9.4	9.7	11.5	-0.8	-0.4	11.5	132.25	132.25	132.25
10.4	9.7	11.5	-1.3	-0.4	11.5	132.25	132.25	132.25
9.6	9.7	11.5	-0.7	-0.2	13	132.25	169	149.5
9.7	9.8	14.0	-0.4	0.5	15	196	225	210
10.6	9.9	15.0	0.5	0.5	15	225	225	225
9.3	10.1	16.0	1.3	0.5	15	256	225	240
10.5	10.3	18.0	1.1	0.7	17	324	289	306
10.7	10.3	18.0	0.9	0.8	18	324	324	324
9.2	10.3	18.0	0.5	0.9	19	324	361	342
9.7	10.4	20.0	0.8	1.0	20	400	400	400
9.2	10.5	21.0	-1.5	1.1	21	441	441	441
9.6	10.6	22.0	-0.2	1.3	23.5	484	552.25	517
9.8	10.7	23.0	0.7	1.3	23.5	529	552.25	540.5
10.3	10.8	24.0	-0.5	1.8	24	576	576	576
						4830	4988	4898.5

Accertamento di una relazione d'ordine tra il tasso di interesse effettivo "E" dei BOT trimestrali e l'indice di borsa "B"

$$r_S = \frac{\frac{4898.5}{24} - \frac{(25)^2}{4}}{\sqrt{\left[\frac{4830}{24} - \frac{(25)^2}{4}\right] * \left[\frac{4988}{24} - \frac{(25)^2}{4}\right]}}$$



Esempio

Distanza da un punto inquinante e concentrazione dell'agente nell'aria

Distanza (X)	Concen. (Y)	ranghi(X)	ranghi (Y)	d(x,y)
0	510	1	12	121
50	380	2	9	49
300	450	3.5	10	42.25
300	480	3.5	11	56.25
800	300	5	7.5	6.25
900	300	6	7.5	2.25
1000	170	7	6	1
1500	94	9	3.5	30.25
1500	94	9	3.5	30.25
1500	108	9	5	16
2000	45	11	1	100
5000	89	12	2	100
				554.5

Il rango medio sottostima la variabilità nei ranghi.

Induce a ritenere che sia presente più correlazione di quanto non ve ne sia in realtà

$$\rho_S = \frac{1725 - 3327 - 0.5\left\{\left[(8-2)+(27-3)\right]+\left[(8-2)+(8-2)\right]\right\}}{\sqrt{[1725-30][1725-12]}} = -0.95$$

Formula di tau in caso di parità

$$\tau_b = \frac{S}{\sqrt{\left[\binom{n}{2} - \sum_{j=1}^{n_x}\binom{t_j^x}{2}\right]\left[\binom{n}{2} - \sum_{j=1}^{n_y}\binom{t_j^y}{2}\right]}}$$

N.B. $\binom{n}{2} = \frac{n*(n-1)}{2}$

$$dove \begin{cases} n_x = numero\ \ di\ \ gruppi\ \ di\ \ X\ \ con\ \ parità \\ t_j^x = numero\ \ di\ \ valori\ \ uguali\ \ per\ \ la\ \ j-esima\ \ parità\ \ in\ \ X \\ n_y = numero\ \ di\ \ gruppi\ \ di\ \ Y\ \ con\ \ parità \\ t_j^y = numero\ \ di\ \ valori\ \ uguali\ \ per\ \ la\ \ j-esima\ \ parità\ \ in\ \ Y \\ s = numero\ \ minimo\ \ di\ \ interscambi\ \ che\ \ trasforma\ \ X\ \ in\ \ Y \end{cases}$$

$$S = \sum_{i=1}^{n-1}\sum_{j=1}^{i} \mathrm{sgn}\left(r_i - r_j\right)\mathrm{sgn}\left(s_i - s_j\right) \quad dove \quad \mathrm{sgn}(x) = \begin{cases} 1 & se\ \ x > 0 \\ 0 & se\ \ x = 0 \\ -1 & se\ \ x < 0 \end{cases}$$

Applicazione

Una selezione di n=50 giudici ha disposto secondo l'ordine di preferenza 15 versioni di uno stesso prodotto

Giudice	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
J_01	10	6	9	7	8	4	5	2	1	3	15	11	14	13	12
J_02	10	7	9	8	6	15	13	11	12	14	1	3	4	5	2
J_03	1	5	4	3	2	6	7	9	8	10	15	13	12	11	14
J_04	10	8	9	6	7	4	5	1	2	3	12	14	13	11	15
J_05	7	8	10	9	6	15	12	13	11	14	4	1	2	5	3
J_06	9	6	10	7	8	1	3	4	2	5	15	11	12	13	14
J_07	8	7	9	6	10	1	2	4	5	3	13	11	14	12	15
J_08	14	15	13	12	11	7	6	9	8	10	5	2	1	3	4
J_09	8	6	7	10	9	5	4	3	1	2	15	14	11	12	13
J_10	9	8	6	7	10	13	11	15	12	14	5	4	3	1	2
J_11	7	6	8	10	9	2	5	3	1	4	11	13	12	15	14
J_12	9	7	10	8	6	1	2	5	4	3	13	15	11	12	14
J_13	4	3	2	5	1	6	8	9	10	7	14	15	12	11	13
J_14	10	9	8	6	7	4	2	5	1	3	15	12	13	14	11
J_15	3	1	2	4	5	9	6	8	7	10	11	12	14	13	15
J_16	8	10	6	9	7	2	5	1	4	3	15	13	11	14	12
J_17	10	6	7	9	8	1	4	2	3	5	11	15	13	12	14
J_18	10	7	6	9	8	15	12	11	14	13	3	4	1	2	5
J_19	9	8	6	7	10	12	11	13	14	15	5	1	4	3	2
J_20	2	1	5	3	4	7	6	8	9	10	15	11	12	14	13

Non sono presenti parità, ma è raro con tanti gradi di giudizio

In questi casi non si considerano 15 distinte variabili, ma una sola variabile: la graduatoria fissata dal giudice (che si articola in quindici valori non scindibili)

Applicazione/2

Riportiamo una parte della matrice delle dissimilarità

```
      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] 0.0000 0.9097 0.9634 0.9825 0.9520 0.9479 0.9572
[2,] 0.9097 0.0000 0.9769 0.9640 0.9654 0.9214 0.9267
[3,] 0.9634 0.9769 0.0000 0.9681 0.9996 0.9807 0.9835
[4,] 0.9825 0.9640 0.9681 0.0000 0.9934 0.9566 0.9582
[5,] 0.9520 0.9654 0.9996 0.9934 0.0000 0.9190 0.9316
[6,] 0.9479 0.9214 0.9807 0.9566 0.9190 0.0000 0.9309
[7,] 0.9572 0.9267 0.9835 0.9582 0.9316 0.9309 0.0000
```

Gli autovalori sono negativi tranne il primo (quello massimo in valore assoluto) che è positivo.

La traccia della matrice di dissimilarità ovviamente è nulla

```
 [1] 47.363 -0.706 -0.742 -0.759 -0.782 -0.785 -0.813 -0.818 -0.829 -0.843 -0.858
[12] -0.860 -0.863 -0.868 -0.884 -0.890 -0.896 -0.905 -0.911 -0.912 -0.917 -0.922
[23] -0.926 -0.931 -0.935 -0.938 -0.942 -0.951 -0.956 -0.959 -0.968 -0.970 -0.978
[34] -0.988 -0.994 -1.000 -1.008 -1.021 -1.024 -1.056 -1.070 -1.084 -1.089 -1.105
[45] -1.147 -1.204 -1.256 -1.348 -1.366 -1.389
```

Conversione per le correlazioni

Se la misura di affinità è ottenuta come coefficiente di associazione quale ad esempio il coefficiente di correlazione di rango, la conversione richiede una riflessione in più in quanto la presenza del segno specifica la direzione in cui si muove una entità al variare dell'altra.

La trasformazione più ovvia per portare i valori in zero/uno sarebbe:

$$6.\quad \delta_{ij} = \left[\frac{1-a_{ij}}{2}\right]^{\alpha} \qquad \alpha > 0$$

Ha il difetto logico di far corrispondere la dissomiglianza massima alla dissociazione (cioè unità con modalità opposte) e non alla mancanza di affinità: a_{ij}=0.

Se pensiamo alla associazione negativa come ad una forma più dettagliata di affinità che oltre a dare la misura dell'intensità del legame tra le due unità è in grado di specificare il grado di opposizione, la (6.) torna ad essere intellegibile.

Conversione Affinità/Dissimilarità

Per trasformare le correlazioni in dissimilarità o distanze esistono varie formule

1) $$d_{ij} = \sqrt{1 - \frac{r_{ij}+1}{2}},\ \forall i,j$$

Le distanze hanno valore zero se la r_{ij} è+1.

Hanno valore massimo 1 se r_{ij} è -1.

2) $$d_{ij} = \frac{1-r_{ij}}{2}\ \forall i,j$$

L'incertezza è sul significato da dare a r_{ij}=0.5

L'uso del coefficiente in valore assoluto o Del quadrato elimina le incertezze

3) $$d_{ij} = \sqrt{1-r_{ij}}\ \forall i,j$$

$$d_{i,j} = \sqrt{1-\left|r_{i,j}\right|}$$

$$d_{i,j} = \sqrt{1-\left(r_{i,j}\right)^2}$$

Variabili metriche

I valori sono veri e propri numeri adoperati per registrare l'esito di conteggi (variabili discrete) o di rapporti di misurazione (variabili continue).

In entrambi i casi le modalità del dominio si presentano come una successione più o meno fitta di valori

$$S(x) = \{x_1, x_2, \ldots, x_p, \ldots\}$$

I cui valori delle continue sono stati arrotondati ed eventualmente moltiplicati per una potenza del dieci. Il dominio può essere finito o infinito, anche se, per ogni data applicazione è possibile proporre dei ragionevoli limiti estremi.

Vettore delle medie (centroide) — **Matrice di varianze-covarianze e di correlazione**

$$X_{4\times 2} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 8 \\ 58 & 3 \end{bmatrix} \quad \bar{x} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}, \quad S_n = \begin{bmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & -0.36 \\ -0.36 & 1 \end{bmatrix}$$

Una distanza elementare

Verifichiamo le potenzialità come metrica della seguente funzione

$$d(x_i, x_j) = \begin{cases} 1 & se \quad i \neq j \\ 0 & se \quad i = j \end{cases}$$

Un punto non può essere più vicino ad un altro di quanto non lo sia a se stesso.

E' immediato accertare che possiede le caratteristiche della identità (dato che è zero solo se i=j), della positività (dato che è maggiore e uguale a zero) e della simmetria (dato che è 1 se i≠j e 1 se j≠i).

Per la disuguaglianza triangolare dobbiamo considerare tre punti: i, j, k.

Se sono tutti diversi, allora $d(x_i, x_k) + d(x_k, x_j) \geq d(x_i, x_j) \Leftrightarrow 1 + 1 > 1$

Se due o tutti sono uguali, allora

$$x_i = x_j \neq x_k \Rightarrow d(x_i, x_k) + d(x_k, x_j) \geq d(x_i, x_j) \Leftrightarrow 1 + 1 \geq 0$$
$$x_i = x_k \neq x_j \Rightarrow d(x_i, x_k) + d(x_k, x_j) \geq d(x_i, x_j) \Leftrightarrow 0 + 1 \geq 1$$
$$x_i \neq x_j = x_k \Rightarrow d(x_i, x_k) + d(x_k, x_j) \geq d(x_i, x_j) \Leftrightarrow 1 + 0 \geq 1$$
$$x_i = x_j = x_k \Rightarrow d(x_i, x_k) + d(x_k, x_j) \geq d(x_i, x_j) \Leftrightarrow 0 + 0 \geq 0$$

Distanze tra entità

Ipotizziamo, per semplicità, che le unità siano descritte da un vettore di valori X

$d(X_i, X_j) = 0$ se e solo se $X_i = X_j$;	Identità
$d(X_i, X_j) > 0$ se $X_i \neq X_j$;	Positività
$d(X_i, X_j) = d(X_j, X_i)$;	Simmetria
$d(X_i, X_k) + d(X_k, X_j) \geq d(X_i, X_j)$;	Disuguaglianza triangolare

La quarta proprietà porta alla distinzione tra dissimilarità e distanza nel senso che alle prime non è richiesta la verifica della disuguaglianza triangolare.

L'insieme delle distanze forma una matrice simmetrica e nonnegativa avente degli zeri sulla diagonale.

Se per ognuna delle permutazioni degli n indici presi a tre a tre si verifica la quarta proprietà, la matrice si dice euclidea e gode di particolari proprietà

Differenze di livello

Consideriamo il quadrato della distanza euclidea tra due unità qualsiasi i e j:

$$d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^t (\mathbf{x}_i - \mathbf{x}_j) = \sum_{r=1}^{m} (x_{ir} - x_{jr})^2 \qquad \forall i, j$$

Se si pone $\hat{x}_i = (x_i - \mu)$ **si ottiene**

$$(\mathbf{x}_i - \mathbf{x}_j) = [(\mathbf{x}_i - \bar{\mu}) - (\mathbf{x}_j - \bar{\mu})] = (\mathbf{x}_i - \bar{\mu} - \mathbf{x}_j + \bar{\mu}) \qquad \forall i, j$$

Gli scarti rimangono gli stessi anche se sono riferiti alla media globale (centroide) del data set

La matrice dei dati X si trasforma nella matrice degli scarti in base alla relazione:

$$\hat{X} = CX = \left(I - \frac{1}{n} u u^t\right) X \qquad \overbrace{\mathbf{u} = (1, 1, \cdots, 1)}^{n \ elementi}$$

Dove C è la matrice di centramento è u un vettore di soli uno.

Esempio

$$\begin{bmatrix} 1 & 1 & -3 \\ 4 & 0 & 10 \\ 2 & 2 & 5 \\ 1 & 1 & 0 \end{bmatrix}; \quad \hat{\mu}^t = (2, \ 1, \ 3) \Rightarrow \hat{x} = \begin{bmatrix} -1 & 0 & -6 \\ 2 & -1 & 7 \\ 0 & 1 & 2 \\ -1 & 0 & -3 \end{bmatrix}$$

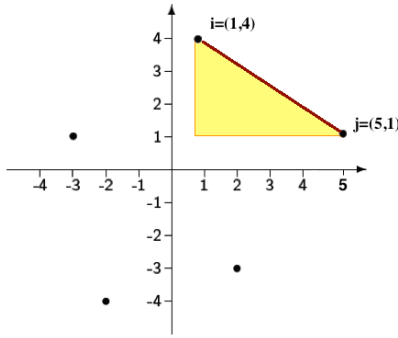
$$\left\{ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \right\} \begin{bmatrix} 1 & 1 & -3 \\ 4 & 0 & 10 \\ 2 & 2 & 5 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -3 \\ 4 & 0 & 10 \\ 2 & 2 & 5 \\ 1 & 1 & 0 \end{bmatrix} - \frac{1}{4} \begin{bmatrix} 8 & 4 & 12 \\ 8 & 4 & 12 \\ 8 & 4 & 12 \\ 8 & 4 & 12 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & -3 \\ 4 & 0 & 10 \\ 2 & 2 & 5 \\ 1 & 1 & 0 \end{bmatrix} - \begin{bmatrix} 2 & 1 & 3 \\ 2 & 1 & 3 \\ 2 & 1 & 3 \\ 2 & 1 & 3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & -6 \\ 2 & -1 & 7 \\ 0 & 1 & 2 \\ -1 & 0 & -3 \end{bmatrix}$$

Notare che la somma di colonna è nulla in ogni caso

Significato

Due punti che sono entrambi vicini ad un altro debbono pure essere in qualche modo vicini

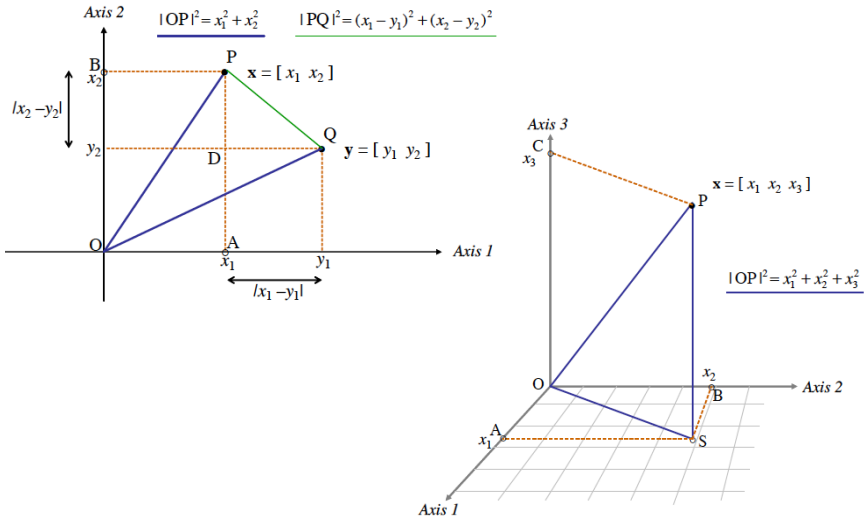


$$d(x_i, x_j) = \sqrt{(1-5)^2 + (4-1)^2}$$
$$= \sqrt{16+9} = \sqrt{25} = 5$$

Il percorso diretto è sempre più breve o al massimo uguale che con una svolta intermedia.

Questa nozione di distanza ingloba l'idea di invarianza rispetto alle traslazioni ed alle rotazioni degli assi

Teorema di Pitagora e distanza 2D e 3D



Esercizio

Step 1: Sum of pairwise squared distances

First, show the sum of squared interpoint distances is proportional to the sum of dot-products:

$$\frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = \sum_{k=1}^{n} x_k^\top x_k$$

Assume the mean of X is at the origin:

$$\sum_{j=1}^{n} x_j = 0$$

Then:

$$\begin{aligned} \frac{1}{2n} \sum_{i}^{n} \sum_{j}^{n} d_{ij}^2 &= \frac{1}{2n} \sum_{i}^{n} \sum_{j}^{n} \|x_i - x_j\|_2^2 \\ &= \frac{1}{2n} \sum_{i,j} (x_i - x_j)^\top (x_i - x_j) \\ &= \frac{1}{2n} \sum_{i,j} (x_i^\top x_i + x_j^\top x_j - 2 x_i^\top x_j) \\ &= \frac{1}{2n} \sum_{i,j} 2 x_i^\top x_i - 2 x_i^\top x_j \\ &= \frac{1}{n} \sum_{i} n x_i^\top x_i - \frac{1}{n} \sum_{i} x_i^\top \sum_{j} x_j \\ &= \sum_{i} x_i^\top x_i - \frac{1}{n} \sum_{i} x_i^\top 0 = \boxed{\sum_{i} x_i^\top x_i} \end{aligned}$$

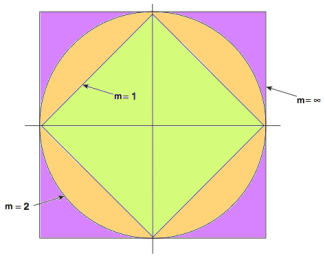
This step also shows that we can compute the sum of squared interpoint distances in a Euclidean space in time linear in the number of points.

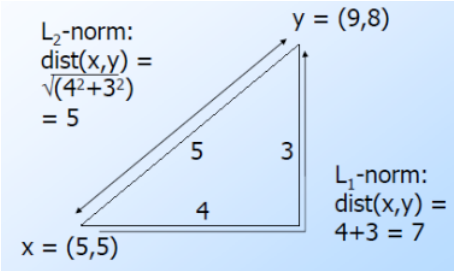
Metriche di Minkowski

$$d_{ij} = \left[\sum_{r=1}^{m} |x_{ri} - x_{rj}|^p \right]^{1/p}$$

In generale, non si può dire che le metriche di Minkowski siano monotone rispetto al loro indice

p=1 (city block o Manhattan)
p=2 (Euclidea)
$p \to \infty$ **(Tchebycheff o Max)**





Il valore numerico delle metriche di Minkowski aumenta con il numero di variabili.

Sono molto sensibili ai valori remoti. Se questi fossero presenti sarebbe opportuno adoperare metodi robusti per il calcolo di queste distanze.

Esempio

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

La scelta tra metrica euclidea e city-block implica la decisione se le variabili sono da considerarsi separate oppure interagiscono

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

La city-block porta ad un giudizio sulla distanza in cui le variabili agiscono in modo indipendente per sommarsi nel giudizio di prossimità. Ad esempio valutando una persona per peso ed altezza la somiglianza tra persone potrebbe avvenire giudicandole separatamente per i due aspetti.

Con la euclidea il giudizio deriva da un mix tra due aspetti collegati.

Metriche di Minkowski/2

Le metriche di Minkowski verificano la disuguaglianza triangolare e ciò deriva proprio dalla disuguaglianza di Minkowski

$$\left[\sum_{r=1}^{m} |x_{ri} + x_{rj}|^p \right]^{1/p} \le \left[\sum_{r=1}^{m} |x_{ri}|^p \right]^{1/p} + \left[\sum_{r=1}^{m} |x_{rj}|^p \right]^{1/p}$$

Poniamo $y_{ri} = x_{ri} - x_{rk},\ k \neq i,\ k \neq j$ **Ne consegue:**

$$\left[\sum_{r=1}^{m} |y_{ri} - y_{rj}|^p \right]^{1/p} \le \left[\sum_{r=1}^{m} |y_{ri} + y_{rj}|^p \right]^{1/p} \le \left[\sum_{r=1}^{m} |y_{ri}|^p \right]^{1/p} + \left[\sum_{r=1}^{m} |y_{rj}|^p \right]^{1/p}$$

$$\left[\sum_{r=1}^{m} |y_{ri} - y_{rj}|^p \right]^{1/p} \le \left[\sum_{r=1}^{m} |y_{ri} + y_{rj}|^p \right]^{1/p} \le \left[\sum_{r=1}^{m} |x_{ri} - x_{rk}|^p \right]^{1/p} + \left[\sum_{r=1}^{m} |x_{rj} - x_{rk}|^p \right]^{1/p}$$

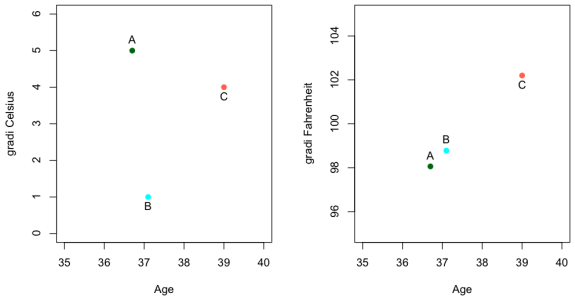
$$\left[\sum_{r=1}^{m} |x_{ri} - x_{rk} - (x_{rj} - x_{rk})|^p \right]^{1/p} \le \left[\sum_{r=1}^{m} |y_{ri} + y_{rj}|^p \right]^{1/p} \le d_{ik}^p + d_{jk}^p$$

$$\left[\sum_{r=1}^{m} |x_{ri} - x_{rj}|^p \right]^{1/p} \le d_{ik}^p + d_{jk}^p \Rightarrow d_{ij}^p \le d_{ik}^p + d_{jk}^p$$

Dipendenza dall'unità di misura

Sog.	Temp_C	Temp_F	Age	TC*	TF*	A*
A	36.70	98.06	5	-0.7324	-0.7324	0.8006
B	37.10	98.78	1	-0.4069	-0.4069	-1.1209
C	39.00	102.2	4	1.1393	1.1393	0.3203
	37.600	99.680	3.333	0.000	0.000	0.000
	1.229	2.212	2.082	1.000	1.000	1.000

Le metriche di Minkowski dipendono dall'unità di misura delle variabili



Con i gradi Celsius il soggetto A e C sono più vicini rispetto al soggetto B. Con i gradi Fahrenheit risulta separata la C.

Dietro ogni normalizzazione/standardizzazione c'è sempre una perdita di informazioni

Ponderazione delle variabili

Si intende la trasformazione dei valori di una variabile al fine di contrarre o espandere l'impatto che la variabile stessa ha sulle unità.

La procedura è controversa:

Da un lato c'è chi ritiene di disporre di informazioni sulla importanza relativa delle diverse variabili ed intende farne uso.
Un geologo che prospetta un terreno può decidere che gli strati abbiano rilevanza inversa rispetto alla profondità. Chi analizza un farmaco può disporne in ordine decrescente di valutazione le componenti

Se le variabili sono state inserite per una qualche ragione allora le ragioni sono tutte egualmente valide e non è corretto dare un peso diverso alle variabili.
Attenzione alla ponderazione implicita! Variabili molto correlate in realtà fanno pesare molto di più quello che hanno in comune

Metriche di Minkowski ponderate

$$d_{ij} = \left[\sum_{r=1}^{m} \left|y_{ri} - y_{rj}\right|^{p}\right]^{1/p};\ y_{ri} = w_r^{1/p} x_{ri},\ y_{rj} = w_r^{1/p} x_{rj};\quad w_r \geq 0,\ \sum_{r=1}^{m} w_r = 1$$

Il modo più semplice di determinare i pesi è di scegliere una quantità da collocare al denominatore delle variabili.

$$Standardizzazione: \frac{x_{r,j}}{\sigma\left(x_j\right)}$$

$$Unitarizzazione: \frac{x_{r,j}}{\max\left(x_j\right) - \min\left(x_j\right)}$$

$$Rapporto\ alla\ media: \frac{x_{r,j}}{\left|\mu\left(x_j\right)\right|}$$

$$Scarto\ medio\ assoluto: \frac{n x_{r,j}}{\sum_{r=1}^{n}\left|x_{r,j} - Me\left(x_j\right)\right|}$$

La standardizzazione costringe le nuove variabili ad avere varianza uno.

La unitarizzazione costringe le nuove variabili a variare nell'intervallo [0,1]

La divisione per la media porta al coefficiente di variazione.

Metriche di Minkowski ponderate/2

Se $d_{i,j}$ è una distanza allora lo sono anche:

$$d_{ij}^{+} = \alpha d_{ij}\quad con\quad \alpha > 0;\quad d_{ij}^{+} = \log\left(1 + d_{ij}\right);\quad d_{ij}^{+} = \left(d_{ij}\right)^{\beta}\quad con\quad 0 < \beta \leq 1$$

La scelta della trasformazione (e quindi della ponderazione può favorire alcune variabili a danno di altre.

ESEMPIO: data set di n entità e variabile binaria X. Vi sono "a" entità che hanno valore 1. Misuriamo la distanza con la metrica city-block.

Prima della standardizzazione, lo scarto tra due entità era $0 \leq \left|x_i - x_j\right| \leq 1$

Dopo la standardizzazione è $0 \leq \left|x_i^{*} - x_j^{*}\right| \leq \frac{n}{\sqrt{a\left(n-a\right)}}$

Se a=1 oppure a=(n-1) il valore assunto sarà √(n-1) e sarà pari a 2 se a=1/(2n) cosicché una variabile con pochi valori 1 può ottenere una rilevanza sproporzionata rispetto al ruolo atteso.

Metriche relative

Rendono invariante la distanza rispetto a trasformazioni moltiplicative o di scala, ma non necessariamente rispetto a trasformazioni additive

Equalizzano l'impatto delle diverse variabili che ora contribuscono alla distanza tra entità senza far valere la loro specifica unità di misura

$$Bray - Curtis: \frac{\sum_{r=1}^{m}\left|x_{ri} - x_{rj}\right|}{\sum_{r=1}^{m}\left|x_{ri}\right| + \sum_{r=1}^{m}\left|x_{rj}\right|};\qquad Canberra: \sum_{r=1}^{m}\left(\frac{\left|x_{ri} - x_{rj}\right|}{\left|x_{ri}\right| + \left|x_{rj}\right|}\right)$$

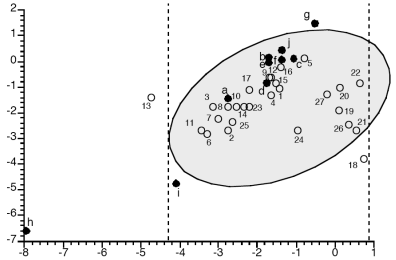
$$Soergel: \frac{\sum_{r=1}^{m}\left|x_{ri} - x_{rj}\right|}{\sum_{r=1}^{m}\left|\max\left(x_{ri}, x_{rj}\right)\right|};\quad Ware - Hedges\ \frac{1}{m}\sum_{k=1}^{m}\left[1 - \frac{\min\left(\left|x_{ki}\right|, \left|x_{kj}\right|\right)}{\max\left(\left|x_{ki}\right|, \left|x_{kj}\right|\right)}\right]$$

Distanza di Mahalanobis

Le metriche di Minkowski ignorano i legami di correlazione tra le variabili.

Un modo per includere le relazioni lineari nella misura della distanza quantitative è la metrica:

$$d_{i,j} = \sqrt{\left(x_i - \mu\right)^t \Omega^{-1} \left(x_j - \mu\right)}$$



Per m=2 variabili si ha

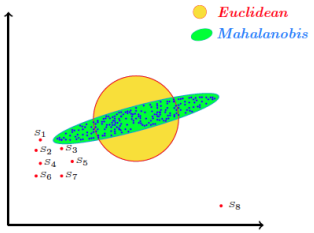
$$d(x,y) = \left[(x_1 - y_1)^2 a_{11}^2 + 2(x_1 - y_1)(x_2 - y_2)a_{12} + (x_2 - y_2)^2 a_{22}\right]^{1/2}$$

La metrica di Mahalanobis è molto utile se si dispone di dati sufficienti per stimare in modo attendibile i suoi parametri.

Se le variabili metriche del data set fossero incorrelate, l'uso della distanza di Mahalanobis equivarrebbe all'impiego delle euclidea per variabili standardizzate.

Distanza di Mahalanobis/2

La distanza di Mahalanobis è la distanza euclidea ponderata ovvero equivale ad una combinazione lineare degli scarti tra le variabili



La distanza di Mahalanobis tiene conto non solo degli scarti tra valori, ma anche della correlazione tra le variabili.

Perchè si possa calcolare le correlazioni non debbono essere perfette.

In questo caso ci sarebbe dipendenza lineare tra le colonne della matrice dei dati ed il determinante della matrice di varianze-covarianze sarebbe zero.

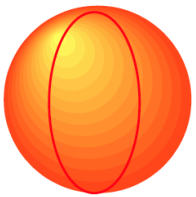
Con variabili correlate la distanza si misura in effetti:

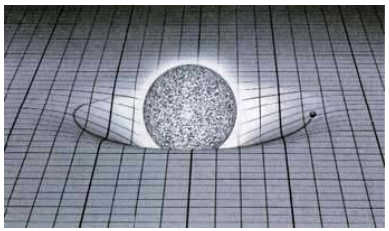
1) *Ruotando gli assi di un angolo legato alla correlazione tra le variabili*
2) *Calcolando al distanza euclidea nel piano degli assi ruotati*

Presupposto euclideo

Le nostre analisi sono sempre riferime allo spazio eulideo multidimensionale privo di curvature.

Non è l'unico possibile. Se ci trovassimo su di una sfera la distanza più breve non sarebbe lungo una linea retta e le rette parallele non potrebbero esistere. Dovremmo ragionare nell'ambito della Geometria ellittica di Riemann.





E' una delle piu' strane, ma piu' vicina delle altre alla geometria del mondo reale

Einstein realized this, and a lot of his relativity work was the development of this different geometry. Much of relativity followed easily once he got the geometry right. N. Vasconcelos, UCSD

Normalizzazione delle distanze

Si possono anche trasformare le distanze in modo da portarle in un intervallo unitario

$$d^*_{i,j} = \frac{d_{i,j} - \min\left\{d_{i,j}\right\}}{\max\left\{d_{i,j}\right\} - \min\left\{d_{i,j}\right\}}; \quad d^*_{i,j} = \sqrt{1 - \frac{d_{i,j}}{\max\left\{d_{i,j}\right\}}};$$

$$d^*_{i,j} = e^{-d_{ij}}; \quad d^*_{i,j} = \frac{2e^{-d_{ij}}}{1 + e^{-d_{ij}}}; \quad d^*_{i,j} = \frac{d_{i,j}}{1 + d_{i,j}}$$

che non solo convertono l'affinità in dissomiglianza, ma sono in grado di trasformare la matrice delle affinità/prossimità in matrici euclidee che hanno un ruolo importante nell'analisi multivariata.

Hanno il difetto di non dipendere dal numero di variabili per cui il raggiungimento degli estremi può essere pilotato scegliendo le variabili opportune.

Distanze multiscala

La matrice dei dati potrebbe contenere variabili misurate su scala diversa. Ad esempio

m_1	*Metriche (rapporti o intervalli)*	$d^1_{i,j}$
m_2	*Ordinali*	$d^2_{i,j}$
m_3	*Politome*	$d^3_{i,j}$
m_4	*Binarie simmetriche*	$d^4_{i,j}$
m_5	*Binarie asimmetriche*	$d^5_{i,j}$

La misura sintetica di distanza può essere definita con una combinazione lineare delle distanze

Poiché l'ordine di grandezza può essere diverso si debbono usare distanze normalizzate (cioè comprese tra zero ed uno).

$$\textit{distanza di Gower}\quad d^g_{i,j} = \sum_{k=1}^{5} \gamma_k d^k_{i,j}; \qquad \gamma_k \geq 0,\ \sum_{k=1}^{5} \gamma_k = 1$$

Scelta dei pesi

Pesi uguali

Accettabile se si ignora se nessuna tipologia è più rilevante delle altre

Pesi come frazione di variabili

Numero di variabili di un tipo sul totale delle variabili. Corretto se ogni tipologia da un proprio contributo che rimane lo stesso all'interno della tipologia.

Pesi per eguagliare la media (o la deviazione standard)

I pesi sono determinati in modo che la media (o la deviazione standard) delle distanze sia la stessa per ciascuna tipologia

Teorema

Appendix

We want to show that if both $d_1(x, y)$ and $d_2(x, y)$ satisfy the triangle inequality, then the same is true for $d(x, y) = \sqrt{d_1^2(x, y) + d_2^2(x, y)}$. Let $a_i = d_i(x, z)$, $b_i = d_i(x, y)$, and $c_i = d_i(y, z)$ so that we have $a_i \leqslant b_i + c_i$, for $i = 1, 2$. Also, let $a = d(x, z) = \sqrt{a_1^2 + a_2^2}$, $b = d(x, y) = \sqrt{b_1^2 + b_2^2}$, and $c = d(y, z) = \sqrt{c_1^2 + c_2^2}$. In order to show that $a \leqslant b + c$ holds, it suffices to show that $a_1^2 + a_2^2 = a^2 \leqslant b^2 + c^2 + 2bc = b_1^2 + b_2^2 + c_1^2 + c_2^2 + 2bc$. This holds if

$$(b_1^2 + c_1^2 + 2b_1c_1) + (b_2^2 + c_2^2 + 2b_2c_2) \leqslant b_1^2 + b_2^2 + c_1^2 + c_2^2 + 2bc,$$

i.e.,

$$b_1c_1 + b_2c_2 \leqslant bc,$$

i.e.,

$$b_1^2c_1^2 + b_2^2c_2^2 + 2b_1c_1b_2c_2 \leqslant b^2c^2 = (b_1^2 + b_2^2)(c_1^2 + c_2^2) = b_1^2c_1^2 + b_2^2c_2^2 + b_1^2c_2^2 + b_2^2c_1^2,$$

i.e.,

$$0 \leqslant (b_1c_2 - b_2c_1)^2,$$

which is clearly true and this completes the proof that $d(x, y)$ satisfies the triangle inequality.

Applicazione

Species	Landolt's indicators: Moisture	Soil ecology	Nitrogen	Light	Temperature	Continentality	Succession
Androsace obtusifolia All.	2	1	1	1	1	2	3
Arabis alpina L.	2	3	1	1	2	2	1
Campanula barbata L.	2	1	1	1	2	2	3
Carex firma Mygind	1	3	1	2	1	3	2
Carex sempervirens Vill.	1	2	1	1	1	3	3
Cerastium uniflorum Clairv.	2	1	1	2	1	2	1
Cirsium spinosissimum (L.) Scop.	3	2	2	1	2	2	2
Dryas octopetala L.	1	3	1	2	1	3	1
Gentiana nivalis L.	2	2	1	1	1	2	3
Geum montanum L.	2	1	1	1	2	2	3
Geum reptans L.	2	1	1	2	1	1	1
Gypsophila repens L.	2	3	1	2	2	2	1
Hedysarum hedysaroides (L.) Schinz & Thell. s.l.	2	3	2	1	2	2	3
Hornungia alpina (L.) Appel s.l.	3	3	1	2	1	1	1
Hypochaeris uniflora Vill.	2	1	1	1	2	2	3
Juncus trifidus L.	1	1	1	2	1	3	2
Ligusticum mutellinoides (Cr.) Vill.	2	2	1	2	1	3	3
Loiseleuria procumbens (L.) Desv.	1	1	1	2	1	3	2
Luzula alpinopilosa (Chaix) Breistr.	3	1	1	2	1	1	2
Peucedanum ostruthium (L.) W.D. Koch	2	2	2	1	2	2	2
Phyteuma betonicifolium Vill. s.l.	2	1	1	1	2	2	3
Phyteuma hemisphaericum L.	1	1	1	1	1	2	2
Ranunculus alpestris L. s.l.	3	3	1	1	1	1	1
Rhododendron ferrugineum L.	2	1	1	1	2	1	3
Saxifraga stellaris L.	3	2	1	2	2	1	1
Sesleria caerulea (L.) Ard.	1	3	1	1	2	3	3
Trifolium alpinum L.	1	1	1	1	2	2	3

Ecological traits and phylogeographic structure for all 27 alpine plant species

Binarie asimmetriche: 3,4,5

Politome ordinate: 1,2,6,7

Applicazione/2

```
     [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 0.000 0.500 0.000 0.625 0.375
[2,] 0.500 0.000 0.500 0.375 0.625
[3,] 0.000 0.500 0.000 0.625 0.375
[4,] 0.625 0.375 0.625 0.000 0.250
[5,] 0.375 0.625 0.375 0.250 0.000
```

Binarie asimmetriche (matrice parziale)

Russell-Rao

```
, , 2

      [,1]   [,2]   [,3]   [,4]   [,5]
[1,] 0.0000 0.6667 0.3333 0.6667 0.6667
[2,] 0.6667 0.0000 0.6667 0.6667 0.6667
[3,] 0.3333 0.6667 0.0000 0.6667 0.6667
[4,] 0.6667 0.6667 0.6667 0.0000 0.3333
[5,] 0.6667 0.6667 0.6667 0.3333 0.0000
```

Politome ordinate (matrice parziale)

Media di scarti relativi

```
      [,1]   [,2]   [,3]   [,4]   [,5]
[1,] 0.0000 0.5952 0.1905 0.6488 0.5417
[2,] 0.5952 0.0000 0.5952 0.5417 0.6488
[3,] 0.1905 0.5952 0.0000 0.6488 0.5417
[4,] 0.6488 0.5417 0.6488 0.0000 0.2976
[5,] 0.5417 0.6488 0.5417 0.2976 0.0000
```

Matrice ponderata con pesi 3/7 e 4/7 in base al numero delle variabili

Applicazione/3

Ipo<-c(1,2,9);Ime<-3:8

Species	Host	Country	Av_Len	Max_Len	Min_Len	Av_Wid	Max_Wid	Min_Wid	Class
I1	Cat	Liberia	64	72	58	41	47	38	uterobilateralis
I2	Cat	Liberia	71	77	65	42	43	40	uterobilateralis
I3	Mongoose	Liberia	70	76	65	42	47	40	uterobilateralis
I4	Mongoose	Gabon	72	78	67	41	44	37	uterobilateralis
I5	Albinos_rat	Liberia	63	70	53	39	43	35	uterobilateralis
I6	Man	Liberia	70	78	62	44	46	39	uterobilateralis
I7	Man	Guinea	72	95	62	44	52	39	uterobilateralis
I8	Man	Congo	66	75	60	44	48	39	uterobilateralis
I9	Cat	Cameroon	93	105	74	48	52	45	africanus
I10	Cat	Cameroon	90	99	85	48	54	46	africanus
I11	Cat	Cameroon	94	103	80	49	55	43	africanus
I12	Civet_cat	Cameroon	89	120	72	50	68	44	africanus
I13	Man	Cameroon	88	112	70	52	62	47	africanus
I14	Cat	Cameroon	98	117	84	57	70	45	westerman
I15	Cat	Cameroon	94	112	71	53	61	47	westerman
I16	Cat	Cameroon	94	110	85	58	75	45	westerman
I17	Man	Gabon	100	112	88	58	67	53	westerman
I18	Man	Cameroon	98	113	84	64	76	51	westerman
I19	Cat	Cameroon	84	93	65	48	55	38	Euparagonimus
I20	Cat	Cameroon	85	95	75	49	55	40	Euparagonimus
I21	Cat_doga	Ivory_Coast	83	102	63	46	54	36	Euparagonimus
I22	Man	Cameroon	85	105	70	49	55	43	Euparagonimus

```
     [,1]  [,2]  [,3]  [,4]
[1,] 0.000 0.143 0.571 1.000
[2,] 0.143 0.000 0.571 1.000
[3,] 0.571 0.571 0.000 0.571
[4,] 1.000 1.000 0.571 0.000
```

```
     [,1]  [,2]  [,3]  [,4]
[1,] 0.000 0.040 0.030 0.041
[2,] 0.040 0.000 0.009 0.013
[3,] 0.030 0.009 0.000 0.019
[4,] 0.041 0.013 0.019 0.000
```