Scaling Metrico

Le variabili possono essere riassunte con un numero ridotto di componenti o variabili latenti e per classificare, conoscere, verificare le relazioni tra le unità e determinarne i tratti più salienti è sufficiente (almeno entro un definito gradi di approssimazione) studiare le componenti trascurando le variabili originali.

Le componenti stanno alle variabili come le variabili stanno alle unità: sono elementi caratterizzanti che si dispongono trasversalmente rispetto alle variabili (sebbene non su tutte allo stesso modo) e laddove il gruppo o classe è il comune oggetto delle unità, la componente è il comune oggetto delle variabili. C'è però una differenza importante: esiste un solo set di unità, ma possono esistere più di una componente e lo scaling multidimensionale ha come scopo la scoperta e la determinazione del numero minimo di componenti (che hanno un massimo nel numero delle variabili originali) nonché il contributo di ciascuna variabile alla formazione della componente considerata come proiezione della variabile sull'asse-dimensione. Naturalmente, se le unità subiscono modifiche in funzione delle variabili e queste, a loro volta, subiscono modifiche in funzione delle componenti, ne consegue che queste hanno incidenza anche sulle unità ed è per questo che lo scalino (o riduzione) può essere riferito sia alle unità che alle variabili.

Una volta scelto tra lo scaling delle unità e lo scaling delle variabili (e per questo che si usa la dizione generica "entità") le entità sono rappresentate come punti nello spazio e la loro distanza è funzione diretta della loro dissimilarità nel senso che entità simili sono vicine ed entità diverse sono distanti. Ne consegue che la procedura di scaling multidimensionale ha due fasi autonome, studiabili separatamente:

1. Lo studio delle somiglianze tra entità e la possibilità di convertirle in distanze in modo che entità simili occupino lo stesso punto nello spazio ed entità molto dissimili siano più distanti di entità molto simili.
2. Lo studio della dimensionalità dello spazio che tenta di convertire le dissimilarità/distanze in coordinate di rappresentazione per le entità.

Dalla prima fase si esce con una matrice che condensa tutte le informazioni sulle entità in termini di relazioni di dissomiglianza o affinità tra ogni coppia.

$$\mathbf{D} = \left(d_{ij} \right)$$

Tale matrice è l'input della fase due che provvede al calcolo multivariato delle coordinate.

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 r=2 modalità possibili della variabile, cioè detta X la variabile, il dominio di quest'ultima è almeno l'insieme

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

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

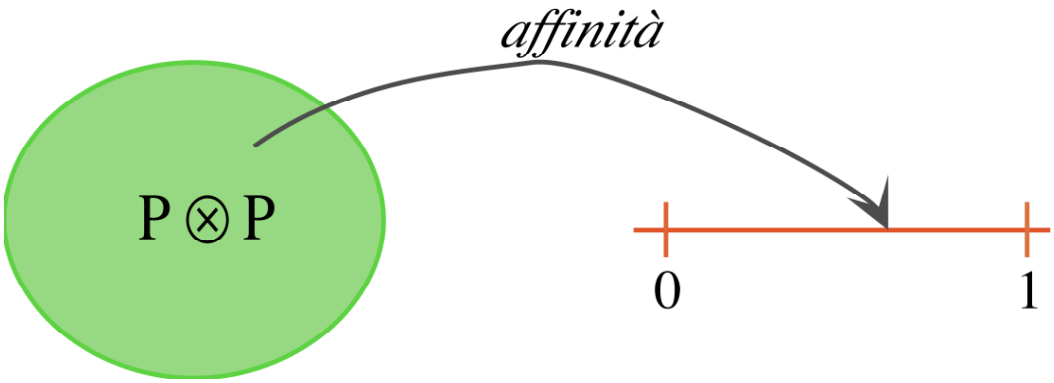
		unità j	
		x_1	x_2
unità i	x_1	=	≠
	x_2	≠	=

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

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

In generale, considerato un data set di n entità, sulle quali si acquisiscono m informazioni, l'affinità è una misura reale definita sul prodotto cartesiano: $P \otimes P$ dove $P = \{u_1, u_2, \ldots, u_n\}$ è l'insieme delle "n" entità considerate nell'indagine. Possono trattarsi di persone, enti, paesi, specie, gruppi ma anche variabili, stimoli, situazioni. Ad ogni elemento di $P \otimes P$ indicato genericamente come (i,j) è associato un numero non negativo a(i,j) o a_{ij} che esprime il grado di affinità, di prossimità, di contiguità comunque percepito tra le entità a confronto.

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



La letteratura statistica è ricca di misure che si propongono di esprimere e quantificare l'affinità. I requisiti richiesti a tali misure sono:

1. Indistinguibilità degli identici: Se $u_i = u_j \Rightarrow a_{ij} = 1$. Questo implica, ad esempio, che $a_{ii} = a_{jj} = 1$ cioè se si confronta una entità con se stessa questo deve risultare in un valore unitario (cioè il valore massimo) dell'affinità.

2.Distinguibilità dei diversi: 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 tale requisito è rispettato e può succedere che

$$a_{ij} \neq a_{ik} \quad \text{anche se } k = j.$$

3.Simmetria: $a_{ij} = a_{ji} \quad \forall i, j$ e cioè l'affinità è la stessa sia che si ragguagli l'entità i-esima alla j-esima che viceversa. 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.

4. Univocità. Se $a_{ij} < a_{rs}$ Allora le entità i e j sono meno affini di quanto non lo siano le unità r ed s

5. Transitività Se $a_{ij} \le a_{rs}$ e $a_{rs} \le a_{pq} \Rightarrow a_{ij} \le a_{pq}$ Questa proprietà garantisce la possibilità di ordinare in sequenza i diversi casi di affinità riscontrati in $P \otimes P$

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

e l'insieme delle affinità sarà dotato di un minimo ed un massimo oltre che degli estremi 0 ed 1.

6. Disuguaglianza triangolare
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:

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

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. Se vale la 6, la 2 è superflua. Sebbene la disuguaglianza triangolare sia considerata indispensabile, altri autori la considerano una condizione valida in generale, ma che per un numero ristretto di comparazioni può essere violata.

7. Ultrametricità

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

In questo caso le triadi di unità, se viste come punti dello spazio, possono solo formare un triangolo isoscele. Se la la condizione di ultrametricità è verificata, allora anche la disuguaglianza triangolare è verificata in quanto entrambe rientrano nello schema

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

Se α=1 si ottiene la disuguaglianza triangolare, ma se $h \to \infty$ 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.

La funzione adottata per misurare l'affinità e le proprietà che la funzione stessa possiede dipende dalla scala di misurazione delle variabili. Noi ipotizziamo che per ogni unità siano presenti variabili del tipo

1. Binarie
2. Politome sconnesse
3. Politome ordinate
4. Scale a rapporti o intervallari

Affinità per variabili binarie

La variabile binaria è l'espressione di una 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 bianco/nero, oppure ON/OFF di un circuito logico ed il confronto di due soggetti si traduce nella predisposizione di una tabella tetracorica. Nel comparare le due unità rispetto alla variabile binaria X l'esito ricade certamente in una delle quattro configurazioni seguenti

		u_j	
	1	x_1	x_2
u_i	x_1	1	0
	x_2	0	0

		u_j	
	2	x_1	x_2
u_i	x_1	0	0
	x_2	0	1

		u_j	
		x_1	x_2
u_i	x_1	0	0
	x_2	1	0

		u_j	
	m	x_1	x_2
u_i	x_1	0	1
	x_2	0	0

Ognuna di queste situazioni contribuisce all' affinità tra i due soggetti in modo autonomo e separato ed è subito evidente una sorta di gerarchia di importanza:
a) contemporanea presenza: le due unità sono più simili tra di loro perché hanno in comune un aspetto.
b) contemporanea assenza: le due unità sono più simili tra di loro perché sono entrambe privi di una caratteristica .
c) e d) le due unità sono meno simili perché hanno un comportamento diverso rispetto alla presenza/assenza, di una proprietà.

La valutazione dell'affinità tra soggetti passa di solito per più variabili binarie perché sarebbe troppo riduttivo affrontare una comparazione solo in base allo stato attivo/passivo di un unico tratto che riguardi le unità. Nella maggior parte delle applicazioni sono presenti diverse variabili binarie, diciamo m, e per ciascuna si deve predisporre una tabella tetracorica di confronto e questo per ogni coppia di soggetti rientranti nel data set. 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 (cioè relativa ad una sola variabile). Al fine di ottenere la massima flessibilità di giudizio sul grado di affinità che scaturisce dai confronti di variabili binarie, conviene associare ad

ogni tabella di confronto, una tabella di valutazione che specifica il contributo che il confronto stesso dà all'affinità complessiva rispetto alle variabili binarie. Ad esempio, per la variabile X_k si realizza la valutazione:

Tabella di confronto

k	x_1	x_2
x_1	=	≠
x_2	≠	=

Tabella di giudizio

k	x_1	x_2	
x_1	a_k	b_k	$a_k + b_k$
x_2	c_k	d_k	$c_k + d_k$
	$a_k + c_k$	$b_k + d_k$	T_k

Dove a_k, b_k, c_k, d_k sono dei numeri non negativi ed eventualmente frazionari che quantificano il peso da dare al singolo confronto e cioè: se la variabile X_k è presente in entrambe le unità, il contributo che questa duplice presenza dà all'affinità complessiva è a_k; se è per entrambe assente sarà d_k e sarà invece b_k ovvero c_k se in una e presente e nell'altra è assente. È evidente che molto spesso, ma non sempre, sarà $b_k=c_k=0$ e che $d_k=0$ se la contemporanea assenza di una proprietà, nulla può aggiungere alla somiglianza delle unità sotto esame. Ad esempio, nel paragonare un leone ad un criceto l'assenza di ali in entrambi gli animali non aiuta a comprenderne la somiglianza. Se invece si fa una comparazione sugli effetti collaterali di un farmaco sia la compresenza che la coassenza di un sintomo sono importanti.

Il totale $T_k = a_k + b_k + c_k + d_k$ costituisce il peso complessivo che la variabile k-esima ottiene nella comparazione di due soggetti ed il suo contributo al formarsi della valutazione della loro affinità. Il confronto complessivo produce la tabella:

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

Per semplificare la simbologia scriviamo:

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

La tabella riassuntiva dei confronti sarà:

u_i \ u_j		
	a	b
	c	d

le cui entrate possono anche essere dei valori frazionari.

Esempio
Si intende valutare l'affinità di un data set di comuni rispetto ad alcune variabili binarie dando ad ognuna un peso diverso.

Variabile	Peso Variabile	Peso delle binarie	Peso specifico binarie
X1=Litoraneità	45	30	13.5
X2=Ufficio turistico	15	30	4.5
X3=Alberghi 5 stelle	40	30	12.0
	100		

La tabella si riferisce alla presenza/assenza della caratteristica indicata nella variabile e tiene conto che nella analisi di riferimento, l'affinità derivata dalle variabili binarie ha un peso pari al 30% dell'affinità totale riscontrabile tra le unità. I pesi delle variabili sono così ripartiti in base a considerazioni a priori.

X_1

	x_1	x_2	
x_1	45	0	45
x_2	0	0	0
	45	0	45

X_2

	x_1	x_2	
x_1	7.5	0	7.5
x_2	0	7.5	7.5
	7.5	7.5	15

X_3

	x_1	x_2	
x_1	25	0	25
x_2	0	10	10
	25	5	35

Per la X_2 e la X_3 si è ritenuto che l'assenza della caratteristica generasse un grado di affinità positivo, ma inferiore a quello della comune presenza per la X_3 e di pari livello per la X_2. Invece, l'assenza di litoraneità nella X_1 non è stata considerata elemento per una affinità positiva. Ecco i dati per la rilevazione su quattro comuni:

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 assenza congiunta ovvero della non presenza in una o entrambe le unità delle caratteristiche misurate dalle tre variabili.

X1	Roccasecca	Cajaniello	Capri	Sorrento
Roccasecca		d1=1	b1=1	b1=1
Cajaniello	d1=1		b1=1	b1=1
Capri	b1=1	b1=1		a1=1
Sorrento	b1=1	b1=1	a1=1	

X2	Roccasecca	Cajaniello	Capri	Sorrento
Roccasecca		a2=1	a2=1	c2=1
Cajaniello	a2=1		a2=1	a2=1
Capri	a2=1	a2=1		c2=1
Sorrento	c2=1	a2=1	c2=1	

X3	Roccasecca	Cajaniello	Capri	Sorrento
Roccasecca		b3=1	b3=1	b3=1
Cajaniello	b3=1		a3=1	a3=1
Capri	b3=1	a3=1		a3=1
Sorrento	b3=1	a3=1	a3=1	

X1,X2,X3	Roccasecca	Cajaniello	Capri	Sorrento
Roccasecca		45+7.5+0	0+7.5+0	0+0+0
Cajaniello	45+7.5+0		0+7.5+25	0+0+25
Capri	0+7.5+0	0+7.5+25		45+0+25
Sorrento	0+0+0	0+0+25	45+0+25	

X1,X2,X3	Roccasecca	Cajaniello	Capri	Sorrento
Roccasecca		52.5	7.5	0.0
Cajaniello	52.5		32.5	25.0
Capri	7.5	32.5		70.0
Sorrento	0.0	25.0	70.0	

Dalla tabella emerge che i comuni più affini sono Capri e Sorrento dato che la somma delle affinità binarie raggiunge il massimo 70 ed i comuni meno affini sono Roccasecca e Sorrento che non hanno alcuna affinità rispetto agli aspetti considerati nell'esempio.

Coefficienti di affinità per variabili binarie

Gli elementi della tabella tetracorica riassuntiva possono essere utilizzati per definire un indice di affinità tra due soggetti che combini in vario modo i conteggi delle celle {a,b,c,d}. A questo fine esistono numerosi coefficienti: il comand sim del pachetto R *simba* ne include 56 (uno studio di Choi et al. ne elenca 76) ed a noi non resta che l'imbarazzo della scelta.

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

Tenuto conto della finalità dello scaling metrico, tra i tanti indici menzionati in letteratura, ci si può limitare ai coefficienti che generano una matrice delle distanze euclidea e che variano tra zero ed uno. Nella tabella seguente sono stati distinti gli indici che non includono il conteggio della coassenza (T) con quelli che la includono (S)

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*	$\frac{a}{a+b+c+d}$

Esempio

Attributi binari

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

Coefficiente che ignora le coassenze (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

Coefficiente che include le coassenze (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

Affinità per variabili politome (nominali o multistato)

Il dominio è formato da modalità che distinguono gli aspetti, le categorie, gli attributi, stati 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 affermare che:

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

Le modalità devono derivare da un elenco di scelte mutualmente esclusive e potrebbero anche essere dei numeri, ma senza che ad essi sia possibile assegnare l'usuale significato di conteggio o misurazione. Le modalità hanno qui la sola funzione di etichettare le unità per formarne una lista o per raggrupparle in categorie omogenee: ad esempio il comune di residenza è una nominazione che raggruppa le unità in cui è denunciata la nascita; la data di pubblicazione sulla Gazzetta Ufficiale e il numero cronologico di una legge identificano le norme emanante dal Parlamento. Le differenze tra le unità possono essere accertate, ma non ordinate né misurate. Quello che si deve garantire è che se le unità presentano la stessa caratteristica allora debbono essere rappresentate dalla stessa categoria e, categorie diverse debbono essere attribuite ad unità con modalità differenti. Tuttavia, all'interno di S le modalità potrebbero essere scambiate di posto senza che ciò influisca sulla validità della classificazione o possa avere effetto sui risultati delle elaborazioni.

Per misurare la somiglianza delle unità secondo questo tipo di scala ci sono due possibilità: matrice dei giudizi predefinita, frammentazione in variabili binarie.

Definizione della matrice dei giudizi

Supponiamo che la variabile X sia rilevata con il dominio

$$S(X) = [x_1, x_2, \ldots, x_m].$$

dove m è il numero di modalità esaustive ed esclusive previsto nel dominio. L'ordinamento alfabetico con cui sono spesso presentate le nominazioni e le variabili nominali semplifica l'esposizione, ma non stabilisce una gerarchia.

Consideriamo due generiche unità: i e j. Per ogni confronto tra due unità può verificarsi una qualsiasi delle combinazioni di modalità

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

dove n rappresenta la dimensione del data set rispetto alle unità. Per ciascuna combinazione di modalità occorre esprimere un giudizio di somiglianza:

$$\left[\begin{array}{c|cccccc} & x_1 & x_2 & \dots & x_j & \dots & x_m \\ \hline x_1 & 1 & a_{12} & \dots & a_{1j} & \dots & a_{1m} \\ x_2 & a_{21} & 1 & \dots & a_{2j} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_i & a_{i1} & a_{i2} & \dots & 1 & \dots & a_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_m & a_{m1} & a_{m2} & \dots & a_{mj} & \dots & 1 \end{array}\right]$$

Le entrate a_{ij} esprimono la valutazione della prossimità tra due unità in cui si verificano le due modalità della coppia considerata. Potenzialmente potremmo far variare a piacimento i numeri da dare alle a_{ij}, ma è meglio semplificare seguendo alcune regole minimali:

1. Intervallo unitario:

$$0 \le a_{ij} \le 1$$

Questo è apprezzato da tutti gli utenti delle tecniche delle tecniche di scaling, ma è meno gradito dai teorici che vi avvertono una forzatura. Gli estremi sono ovvi:

$$a_{ij} = \begin{cases} 0 \text{ significa che, nel metro di giudizio adottato, le due unità sono all'opposto (se questo ha senso)} \\ 1 \text{ significa che le due unità sono identiche (rispetto alla variabile)} \end{cases}$$

ogni valore intermedio esprimerà una valutazione di somiglianza crescente all'aumentare della affinità a_{ij}.

2. Simmetria Questa è una seconda regola su cui c'è poca controversia:

$$a_{ij} = a_{ji}$$

e cioè il giudizio di affinità, prossimità, similarità non cambia se si compara l'unità i con la j oppure si procede al contrario; quindi, solo una parte, quella superiore o quella inferiore della matrice dei giudizi, dovrà essere riempita formando m*(m-1)/2 valutazioni.

3. Monotonicità. L'affinità deve aumentare quando le due unità a confronto presentano una combinazione di modalità che rivela una maggiore vicinanza rispetto alle altre. I numeri da inserire sono soggettivi e possono anche essere dei valori frazionari. Quello che si richiede è che, ad esempio, nel confronto della posizione politica le codifiche "liberale" e "moderato" siano più affini di quanto non lo siano "liberale" e "conservatore". In generale, alle misure di affinità non è richiesto di verificare la disuguaglianza triangolare, dato che si preferisce privilegiare le proprietà ordinali dei coefficienti piuttosto che la loro metricità. Tuttavia, ai fini dello scaling metrico la condizione di generare una matrice euclidea delle dissimilarità, la scelta dei coefficienti che quantificano l'affinità tra variabili politome è cruciale come lo era per le binarie.

Esempio_1:
Giudizi personali sul confronto delle cinque province calabresi rispetto alla attività prevalente:
S(X)={Commercio, Turismo, Amministrazione pubblica, Industria manufatturiera, Agricoltura}

Matrice o tabella dei giudizi

	Commercio	Turismo	Amm. Pubblica	Industria Mani	Agricoltura
Commercio	1	0.7	0.8	0.4	0.1
Turismo		1	0.6	0.3	0.5
Amm. Pubblica			1	0.4	0.2
Industria Manif.				1	0.1
Agricoltura					1

L'idea è di sottoporre poi a delle unità (un campione oppure un censimento di soggetti in posizioni strategiche) una domanda del tipo "Quale considera sia l'attività prevalente della sua provincia?"

L'incrocio dei dati produrrà poi valori in base alla tabella delle compresenze. Naturalmente, la formazione della matrice dei giudizi potrebbe aver richiesto, a sua volta, interviste e/o valutazioni di esperti oppure i giudizi possono derivare da acquisizioni empiriche consolidate.

Esempio_2
Ad un gruppo di 10 studenti scelti accuratamente per posizione curricolare, estrazione sociale, formazione secondaria, etc. è stato chiesto di esprimere con un voto da 1 a 6 un gruppo di quattro insegnanti: 0= diametralmente opposti e 6= del tutto equivalenti. Ecco i risultati

Confronto	Stud_01	Stud_02	Stud_03	Stud_04	Stud_05	Stud_06	Stud_07	Stud_08	Stud_09	Stud_10	MEDIA	MEDIANA	Media/6
I1,I2	5	6	4	6	5	5	6	4	6	5	5.2	5.0	0.9
I1,I3	4	2	3	4	3	2	2	3	3	4	3.0	3.0	0.5
I1,I4	2	0	1	1	2	1	2	2	1	3	1.6	1.5	0.3
I2,I3	6	5	5	6	4	3	4	5	4	5	4.7	5.0	0.8
I2,I4	1	1	2	3	1	1	2	1	1	0	1.4	1.0	0.2
I3,I4	3	2	4	5	2	3	4	4	3	4	3.4	3.5	0.6

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 studenti (e non solo i dieci giudici serviti per formare la matrice dei giudizi) possono essere confrontati rispetto alla scelta fatta.

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

Un coefficiente di dissimilarità (complementare alla affinità) che risponde al requisito di disuguaglianza triangolare ed anche a quella di euclideità della corrispondente matrice è sato proposto da Beijnen (1973).

$$d_{ij} = \sqrt{\frac{\sum_{\substack{h_{sij}=1 \\ s\in[1,..,m]}} \delta\left(x_{si},x_{sj}\right)}{m}} \quad con \quad \delta\left(x_{si},x_{sj}\right) = \begin{cases} \frac{L_s}{p} & se \quad x_{si} \neq x_{sj} \\ 0 & altrimenti \end{cases}$$

dove L_s è il numero di modalità previsto per la variabile s-esima e

$$p = \sum_{s=1}^{m} L_s$$

è il numero totale di modalità attinenti alle variabili politome. Inoltre, l'indicatore h_{sij} è uno se il confronto tra le due unità: i e j è ammissibile rispetto alla variabile s-esima (ad esempio nessuno dei due campi è vuoto). In caso contrrio si ha h_{sij} =0. 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.

La matrice delle distanze formata con il suddetto coefficiente verifica la condizione di euclideità.

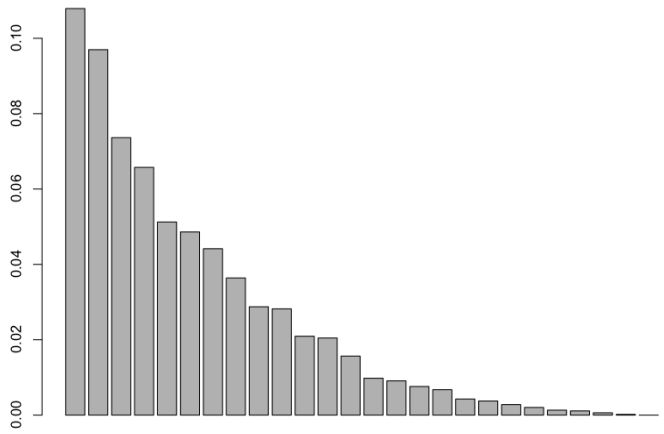
Esempio

Consideriamo il data set artificiale "nominal_data" inserito nel pacchetto cluster Sim di R

```
library(clusterSim);library(ade4)
##########################
DistPoly<-function(ir,jc){
```

```
 Iam<-which(is.na(Redata[ir,Ipo]));Ibm<-which(is.na(Redata[jc,Ipo]))
 if (length(Iam)>0) Ime1<-Ipo[-Iam] else Ime1<-Ipo
 if (length(Ibm)>0) Ime2<-Ipo[-Ibm] else Ime2<-Ipo
 eqv<- Ime1 %in% Ime2;eqp<-which(eqv)
 if (length(eqp)==0) {dista<-NA;return(dista)}
 co<-which(Redata[ir,Ipo[eqp]]!=Redata[jc,Ipo[eqp]])
 if (length(co)>0) Dista<-sum(Lep[co])/sum(Lep) else Dista<-0
return(Dista)}
#########################
data(data_nominal)
Redata<<-data_nominal
m<-ncol(Redata);n<-nrow(Redata)
Ipo<<-1:12;nm1<-n-1;p<-length(Ipo)
Lep<<-c(4,4,5,4,4,4,3,4,2,3,3,3)
Dp<-matrix(0,n,n)
for (i in 1:nm1){k<-i+1
        for (j in k:n){Dp[i,j]<-sqrt(DistPoly(i,j)/p);Dp[j,i]<-Dp[i,j]}}
Test<-is.euclid(as.dist(Dp), plot = FALSE, print = TRUE, tol = 1e-
07);print(Test)
```

Si può constatare che gli autovalori della matrice di distanze ottenute dal nostro coefficiente sono tutti non negativi e solo l'ultimo è zero.



Scomposizione della politomia in variabile binarie

In questo caso la rilevazione della variabile politoma è suddivisa in vari confronti (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)

	X1	X2	Xi	Xm
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. 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. Poiché le due modalità presentate dalla unità i e dalla j possono essere uguali o diverse le tabelle riassuntive possono solo essere due:

u_i \ u_j	X_1	1	0
	1	a	b
	0	c	d

u_i \ u_j	X_2	1	0
	1	a	b
	0	c	d

u_i \ u_j	X_k	1	0
	1	a	b
	0	c	d

A queste tabelle, una per ogni politomia, si può applicare uno degli indici di somiglianza tra variabili binarie, adeguatamente ponderato.

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. Ecco quella ottenuta con il secondo coefficiente:

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_10
u_1	0.0000	0.5774	1.0000	0.8165	0.7693	0.8165	0.8165	1.0000	0.8165	0.8165
u_2	0.5774	0.0000	1.0000	1.0000	1.0000	0.8165	0.5774	1.0000	0.5774	0.5774
u_3	1.0000	1.0000	0.0000	1.0000	1.0000	0.7693	1.0000	0.7071	1.0000	1.0000
u_4	0.8165	1.0000	1.0000	0.0000	0.4284	1.0000	1.0000	0.7693	1.0000	1.0000
u_5	0.7693	1.0000	1.0000	0.4284	0.0000	1.0000	1.0000	0.7071	1.0000	1.0000
u_6	0.8165	0.8165	0.7693	1.0000	1.0000	0.0000	0.8165	0.7693	1.0000	1.0000
u_7	0.8165	0.5774	1.0000	1.0000	1.0000	0.8165	0.0000	1.0000	0.8165	0.8165
u_8	1.0000	1.0000	0.7071	0.7693	0.7071	0.7693	1.0000	0.0000	1.0000	1.0000
u_9	0.8165	0.5774	1.0000	1.0000	1.0000	1.0000	0.8165	1.0000	0.0000	0.0000
u_10	0.8165	0.5774	1.0000	1.0000	1.0000	1.0000	0.8165	1.0000	0.0000	0.0000
u_11	1.0000	0.8165	1.0000	1.0000	1.0000	0.8165	0.8165	1.0000	0.5774	0.5774
u_12	0.0000	0.5774	1.0000	0.8165	0.7693	0.8165	0.8165	1.0000	0.8165	0.8165
u_13	1.0000	1.0000	0.7693	0.8165	0.7693	0.8165	0.8165	0.4284	1.0000	1.0000
u_14	0.5774	0.8165	1.0000	0.8165	0.7693	0.5774	0.8165	1.0000	1.0000	1.0000
u_15	0.8165	0.8165	0.7693	0.8165	1.0000	0.5774	0.8165	0.7693	1.0000	1.0000

	u_11	u_12	u_13	u_14	u_15
u_1	1.0000	0.0000	1.0000	0.5774	0.8165
u_2	0.8165	0.5774	1.0000	0.8165	0.8165
u_3	1.0000	1.0000	0.7693	1.0000	0.7693
u_4	1.0000	0.8165	0.8165	0.8165	0.8165
u_5	1.0000	0.7693	0.7693	0.7693	1.0000
u_6	0.8165	0.8165	0.8165	0.5774	0.5774
u_7	0.8165	0.8165	0.8165	0.8165	0.8165
u_8	1.0000	1.0000	0.4284	1.0000	0.7693
u_9	0.5774	0.8165	1.0000	1.0000	1.0000
u_10	0.5774	0.8165	1.0000	1.0000	1.0000
u_11	0.0000	1.0000	1.0000	0.8165	1.0000
u_12	1.0000	0.0000	1.0000	0.5774	0.8165
u_13	1.0000	1.0000	0.0000	1.0000	0.8165
u_14	0.8165	0.5774	1.0000	0.0000	0.8165
u_15	1.0000	0.8165	0.8165	0.8165	0.0000

La dicotomizzazione delle politomie genera un gran numero di variabili binarie e potrebbe ingenerare somiglianze anche nei confronti di unità in cui queste sono assenti. Inoltre, nel confronto di unità rispetto a variabili miste (che studieremo più avanti) potrebbe esagerare oltre misura il ruolo delle variabili binarie.

Coefficienti di affinità per variabili ordinali

Da definire

Dalla affinità alla dissimilarità

Ad ogni indice 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 cioè con valori in un intervallo limitato (di solito quello unitario) la trasformazione in una dissimilarità è abbastanza semplice

$$1.\quad \delta_{ij} = 1 - a_{ij}^{\alpha} \qquad 2.\quad \delta_{ij} = \left(1 - a_{ij}\right)^{\alpha} \qquad \alpha > 0$$

$$3.\quad \delta_{ij} = \frac{\ln\left(1 + a_{ij}\right) - \ln(2)}{-\ln(2)} \qquad 4.\quad \delta_{ij} = \frac{\alpha - a_{ij}}{\alpha + a_{ij}}$$

$$5.\quad \delta_{ij} = \frac{e^{\left(1-a_{ij}\right)} - 1}{e - 1}$$

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 indice di dissomiglianza.

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 sarebbe:

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

che porta i valori da (-1,1) in (0,1). La 6. ha però 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, tuttavia 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.

Per misure di affinità illimitate superiormente si possono comunque adottare trasformazioni del tipo:

$$7.\quad \delta_{ij} = \bar{e}^{a_{ij}} \qquad 8.\quad \delta_{ij} = \frac{\alpha_1}{\alpha_2 + a_{ij}}\ ,\ \alpha_1, \alpha_2 > 0$$

$$9.\quad \delta_{ij} = \frac{\bar{e}^{a_{ij}}}{1 + \bar{e}^{a_{ij}}} \qquad 10.\quad \delta_{ij} = \left[1 - \frac{a_{ij}}{\max\left\{a_{ij}\right\}}\right]^{0.5}$$

che non solo convertono l'affinità in dissomiglianza, ma operano in un intervallo limitato. Consideriamo ad esempio la conversione numero uno con α=1 e cioè

$$d_{ij} = 1 - a_{ij} \quad con \quad 0 \le a_{ij} \le 1$$

La corrispondente matrice di distanze sarà

$$\mathbf{D} = \mathbf{u}\mathbf{u}^{\mathrm{t}} - \mathbf{A} \quad \Rightarrow \quad -0.5d_{ij}^2 = -0.5\left(1 - a_{ij}^2\right)^2$$

Se a_{ij} è compreso nell'intervallo unitario, lo sarà anche d_{ij} e quando a_{ij}=1 allora d_{ij}=0. Se la A è (semi)positiva definita, la fattorizzazione della forma $\mathbf{A} = \mathbf{X}\mathbf{X}^t$ produrrà una matrice X di coordinate reali tali che il quadrato della distanza euclidea tra l'unità i-esima e quella j.esima è pari a $2\left(1 - a_{ij}\right)$ che non può essere negativo se loè anche d_{ij}. Ne consegue che, la matrice delle dissimilarità con elementi ottenuti come $d_{ij} = \sqrt{\left(1 - a_{ij}\right)}$ è un matrice euclidea purchè a_{ij} sia compreso tra zero ed uno, estremi inclusi.

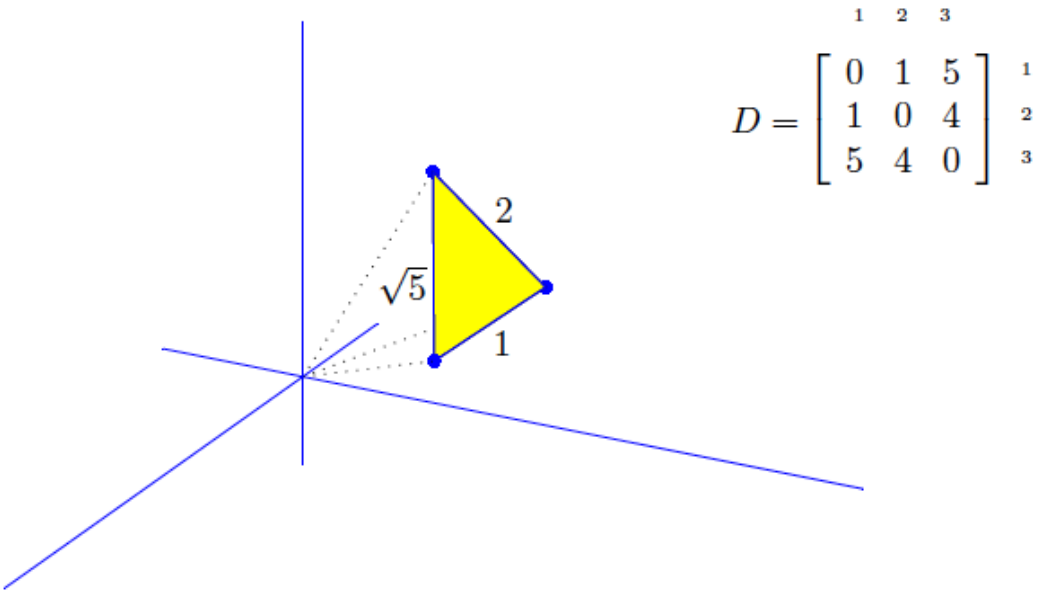
E' stato giustamente osservato che questa condizione è sufficiente per assicurare la verifica della condizione di euclideità per cui potrebbe rivelarsi troppo stringente in alcune applicazioni.

Dalle distanze tra unità alle coordinate nello spazio

Consideriamo una semplice matrice delle distanze tra un gruppo di n=3 unità

$$D = [d_{ij}] = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{bmatrix} = \begin{bmatrix} 0 & d_{12} & d_{13} \\ d_{12} & 0 & d_{23} \\ d_{13} & d_{23} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 5 \\ 1 & 0 & 4 \\ 5 & 4 & 0 \end{bmatrix}$$

La rappresentazione grafica è la seguente



La matrice **D** ha n^2=9 valori, ma solo n(n-1)/2=3 informazioni distinte dato che la matrice delle distanze è simmetrica ed i valori sulla diagonale sono nulli per costruzione. Disponiamo della matrice delle possibili distanze tra tutte le coppie di unità facenti parte di un data set di *n* unità. È possibile risalire alle coordinate (cioè i valori delle variabili) che l'hanno determinata? La risposta a questo problema costituisce il nucleo intorno al quale si è sviluppata la tecnica di rappresentazione nota come "scaling metrico" nata nel 1938 per merito di Young e Householder (alcuni risultati erano però già noto da molto prima). Notiamo subito che, come si è visto nello studio delle rotazioni, la distanza tra i punti rimane invariata se le coordinate originali sono moltiplicate per una matrice di rotazione ortogonale per cui se da una **X** si arriva ad una ed una sola **D**, dalla **D** si possono ottenere infinite matrici di coordinate **X**.

Scarti e prodotti interni

Sia $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})^t$ il vettore delle osservazioni sulle m variabili rilevate su scala metrica associato all'unità i-esima. Una funzione che misuri la distanza tra

le unità $\mathbf{x}_i$ deve verificare le condizioni che consentono la loro rappresentazione in uno spazio metrico e cioè

$$d_{ij} \geq 0 \ \ \forall \ \ i \neq j \qquad \textit{non negatività}$$

$$d_{ij} \geq 0 \ \ \textit{se e solo se} \ \ \mathbf{x}_i = \mathbf{x}_j \qquad \textit{esclusività dello zero}$$

$$d_{ij} = d_{ji} \qquad \textit{simmetria}$$

$$d_{ij} \leq d_{ik} + d_{kj} \ \ \forall \ \ i,j,k \qquad \textit{Disuguaglianza triangolare}$$

Una matrice delle distanze è detta metrica se i valori che contiene derivano dal calcolo di una metrica cioè un meccanismo che rispetti i requisiti precedenti. Inoltre tre qualsiasi delle sue entrate devono rispettare la disuguaglianza triangolare: se due certe unità hanno distanza vicina allo zero allora ogni altra unità differirà poco da entrambe. La matrice metrica quindi è simmetrica ed ha elementi non negativi.

E' da osservare che se anche gli elementi della matrice **D** verificassero la diseguaglianza triangolare non necessariamente **D** sarebbe semidefinita positiva. Consideriamo, in prima istanza, il quadrato della distanza euclidea tra due unità qualsiasi i e j:

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

Come è noto, gli scarti rimangono gli stessi anche se sono riferiti alle medie aritmetiche globali nel data set cioè

$$\hat{\mu} = \frac{1}{n}\mathbf{X}^t\mathbf{u} \quad \text{dove} \ \ \mathbf{u} = (1,1,\ldots,1)^t \ \ \textit{formato da n uno}$$

In questo senso conviene eliminare le differenze di livello tra le variabili e ragionare con scarti dalle medie aritmetiche. Infatti, come si è già visto a proposito della distanza euclidea, si ha

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

Posto perciò: $\hat{\mathbf{x}}_i = \mathbf{x}_i - \hat{\mu} \qquad i = 1,2,\ldots,n$ il quadrato delle distanze euclidee sarà:

$$d_{ij}^2 = \left(\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\right)^t \left(\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\right) \qquad \forall i,j$$

e la matrice dei dati **X** risulta trasformata nella matrice degli scarti in base alla relazione:

$$\hat{\mathbf{X}} = \mathbf{X}\left(\mathbf{I} - \frac{1}{\mathrm{n}}\mathbf{u}\mathbf{u}^{\mathbf{t}}\right) = \mathbf{X}\mathbf{C}$$

che verifica il vincolo: $\hat{\mathbf{X}}^t\mathbf{u} = \mathbf{0}$ e quindi il rango della $\hat{\mathbf{X}}$ potrà al massimo essere pari a n-1 dato che il sistema: $\hat{\mathbf{X}}^{\mathrm{t}}\mathbf{u} = \lambda\mathbf{u}$ ha una soluzione con almeno un $\lambda = 0$. Su questo ritorneremo più avanti. Ricordiamo che **C** è la matrice di centramento.

La matrice $\hat{\mathbf{X}}\hat{\mathbf{X}}^t$ di ordine (n x n) è anche detta "matrice dei prodotti incrociati" in quanto aggrega tutti i prodotti scalari tra gli scarti relativi alla unità i-esima e quelli della unità j-esima per ogni i ed ogni j. La matrice $\mathbf{B} = \hat{\mathbf{X}}\hat{\mathbf{X}}^t$ può essere espressa in termini della matrice delle distanze euclidee **B**

$$\underset{nxn}{\mathbf{D}} = \underset{nx1}{\mathbf{b}}\,\underset{1xn}{\mathbf{u}^t} + \underset{1xn}{\mathbf{u}}\,\underset{nx1}{\mathbf{b}^t} - 2\,\underset{nxn}{\mathbf{B}}$$

dove **b**=diag(**B**). Ricordiamo che **D** contiene il quadrato delle distanze euclidee e non le distanze euclidee tra le unità. Il senso dello scaling metrico è di invertire tale relazione e cioè partire dalla matrice delle distanze ed arrivare alla matrice dei prodotti incrociati. Questo sarà semplice e immediato se si dispone delle coordinate delle variabili, ma sarà più complesso se si dispone solo della matrice delle distanze.

Sviluppiamo il quadrato della distanza euclidea tra due vettori di valori osservati su due particolari unità: i e j.

$$d_{ij}^2 = \left(\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\right)^t\left(\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\right) = \hat{\mathbf{x}}_i{}^t\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_i{}^t\hat{\mathbf{x}}_j - \hat{\mathbf{x}}_j{}^t\hat{\mathbf{x}}_i + \hat{\mathbf{x}}_j{}^t\hat{\mathbf{x}}_j$$

Trattandosi di prodotti interni (cioè che danno uno scalare come risultato) succede che $\hat{\mathbf{x}}_i{}^t\hat{\mathbf{x}}_j = \hat{\mathbf{x}}_j{}^t\hat{\mathbf{x}}_i$ e la precedente relazione si può scrivere

$$d_{ij}^2 = \hat{\mathbf{x}}_i{}^t\hat{\mathbf{x}}_i + \hat{\mathbf{x}}_j{}^t\hat{\mathbf{x}}_j - 2\hat{\mathbf{x}}_i{}^t\hat{\mathbf{x}}_j$$

La somma delle quantità sulla sinistra d_{ij}^2 rispetto ad uno o entrambi gli indici è esprimibile come una cumulata dei prodotti scalari sulla destra. Ad esempio, la somma rispetto all'indice di riga comporta:

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n}\hat{\mathbf{x}}_i{}^t\hat{\mathbf{x}}_i + \sum_{i=1}^{n}\hat{\mathbf{x}}_j{}^t\hat{\mathbf{x}}_j - 2\sum_{i=1}^{n}\hat{\mathbf{x}}_j{}^t\hat{\mathbf{x}}_j$$

$$= \sum_{i=1}^{n}\hat{\mathbf{x}}_i{}^t\hat{\mathbf{x}}_j + n\,\hat{\mathbf{x}}_j{}^t\hat{\mathbf{x}}_j - 2\left[\sum_{i=1}^{n}\hat{\mathbf{x}}_i{}^t\right]\hat{\mathbf{x}}_j = \sum_{i=1}^{n}\hat{\mathbf{x}}_i{}^t\hat{\mathbf{x}}_i + n\,\hat{\mathbf{x}}_j{}^t\hat{\mathbf{x}}_j$$

In particolare, dalla relazione:

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} \left(\hat{\mathbf{x}}_i{}^t \hat{\mathbf{x}}_i + n\, \hat{\mathbf{x}}_j{}^t \hat{\mathbf{x}}_j \right) \Rightarrow \frac{\sum_{i=1}^{n} d_{ij}^2}{n} = \frac{\sum_{i=1}^{n} \hat{\mathbf{x}}_i{}^t \hat{\mathbf{x}}_i}{n} + \hat{\mathbf{x}}_j{}^t \hat{\mathbf{x}}_j$$

Si ottiene:

$$\hat{\mathbf{x}}_j{}^t \hat{\mathbf{x}}_j = \frac{1}{n} \sum_{i=1}^{n} d_{ij}^2 - \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{x}}_i{}^t \hat{\mathbf{x}}_i$$

Ad un analogo risultato si arriva sommando rispetto all'indice j che permette poi di definire:

$$\hat{\mathbf{x}}_i{}^t \hat{\mathbf{x}}_i = \frac{1}{n} \sum_{j=1}^{n} d_{ij}^2 - \frac{1}{n} \sum_{j=1}^{n} \hat{\mathbf{x}}_j{}^t \hat{\mathbf{x}}_j$$

Cosa si ottiene con la sommatoria doppia?

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = \sum_{j=1}^{n} \left[\sum_{i=1}^{n} \hat{\mathbf{x}}_i{}^t \hat{\mathbf{x}}_i + n \hat{\mathbf{x}}_j{}^t \hat{\mathbf{x}}_j \right] = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} \hat{\mathbf{x}}_i{}^t \hat{\mathbf{x}}_i \right) + n \sum_{j=1}^{n} \hat{\mathbf{x}}_j{}^t \hat{\mathbf{x}}_j$$

$$= n \sum_{i=1}^{n} \hat{\mathbf{x}}_i{}^t \hat{\mathbf{x}}_i + n \sum_{j=1}^{n} \hat{\mathbf{x}}_j{}^t \hat{\mathbf{x}}_j = 2n \sum_{i=1}^{n} \hat{\mathbf{x}}_i{}^t \hat{\mathbf{x}}_i$$

Quest'ultima relazione dimostra che il prodotto scalare dei vettori relativi alle unità i e j ha una importante relazione con la loro distanza euclidea al quadrato:

$$\hat{\mathbf{x}}_i^t \hat{\mathbf{x}}_j = \frac{1}{2} \left[\hat{\mathbf{x}}_i^t \hat{\mathbf{x}}_i + \hat{\mathbf{x}}_j^t \hat{\mathbf{x}}_j - d_{ij}^2 \right]$$

La somma di queste ultime quantità è la stessa indipendentemente dall'indice secondo il quale si somma:

$$\sum_{i=1}^{n} \hat{\mathbf{x}}_i{}^t \hat{\mathbf{x}}_i = \sum_{j=1}^{n} \hat{\mathbf{x}}_j{}^t \hat{\mathbf{x}}_j = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2$$

Il significato di questa relazione è che il complesso di informazioni presente nella matrice dei dati, esprimibile come matrice di varianze-covarianze, può anche essere ricondotto alla somma del quadrato delle distanze euclidee.

A questo punto dobbiamo evidenziare il ruolo del prodotto scalare tra le due generiche unità i e j. Tenuto conto del suo legame con il quadrato della distanza euclidea, possiamo scrivere:

$$\begin{aligned}
\hat{\mathbf{x}}_i^t \hat{\mathbf{x}}_j &= \frac{1}{2}\left[\frac{1}{n}\left(\sum_{j=1}^{n} d_{ij}^2 - \sum_{j=1}^{n} \hat{\mathbf{x}}_j^t \hat{\mathbf{x}}_j\right) + \frac{1}{n}\left(\sum_{i=1}^{n} d_{ij}^2 - \sum_{i=1}^{n} \hat{\mathbf{x}}_i^t \hat{\mathbf{x}}_i\right) - d_{ij}^2\right] \\
&= \frac{1}{2}\left[\frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{1}{n}\left(\sum_{j=1}^{n} \hat{\mathbf{x}}_j^t \hat{\mathbf{x}}_j\right) + \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{1}{n}\left(\sum_{j=1}^{n} \hat{\mathbf{x}}_i^t \hat{\mathbf{x}}_i\right) - d_{ij}^2\right] \\
&= \frac{1}{2}\left[\frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{1}{n}\left(\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2\right) + \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{1}{n}\left(\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2\right) - d_{ij}^2\right] \\
&= \frac{1}{2}\left[\frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 + \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{2}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2\right) - d_{ij}^2\right] \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(\frac{d_{ij}^2}{2}\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{d_{ij}^2}{2}\right) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{d_{ij}^2}{2}\right) - \frac{d_{ij}^2}{2}
\end{aligned}$$

Adottiamo ora una simbologia semplificativa

$$h_{ij} = -\frac{1}{2} d_{ij}^2;\; h_{i0} = \frac{1}{n}\sum_{j=1}^{n}\left(-\frac{1}{2} d_{ij}^2\right);\; h_{0j} = \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{1}{2} d_{ij}^2\right);\; h_{00} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(-\frac{1}{2} d_{ij}^2\right)$$

Il prodotto scalare dei valori associati a due unità qualsiasi è sintetizzato dalla espressione:

$$\hat{\mathbf{x}}_i^t \hat{\mathbf{x}}_j = h_{ij} - h_{i0} - h_{0j} + h_{00}$$

Il significato geometrico della operazione diventa più chiaro se si riprende il ben noto legame tra prodotto scalare di vettori e le norme dei vettori:

$$\hat{\mathbf{x}}_i^t \hat{\mathbf{x}}_j = \left\|\hat{\mathbf{x}}_i\right\| \cdot \left\|\hat{\mathbf{x}}_j\right\| \cdot \cos\left(\theta_{ij}\right)$$

ovvero, il prodotto scalare è pari al coseno dell'angolo da essi formato per un fattore proporzionale costituito dalle lunghezze dei due vettori. Inoltre, tenuto conto che le coordinate sono espresse come scarti dalla media, le norme dei vettori corrispondono a delle misure di variabilità: se si usa la norma euclidea il coseno è proporzionato al prodotto degli scarti quadratici medi. Se la distanza tra due punti non cambia se si sposta l'origine degli assi (si sottraggono le coordinate del nuovo punto-origine) il prodotto scalare invece cambia perché le

lunghezze dei vettori sono appunto distanze dal punto (0,0) al punto vettore $\mathbf{x}_i$. Tale indeterminatezza è stata risolta centrando le coordinate nel punto medio: $(\bar{\mathbf{x}}_1,\cdots,\bar{\mathbf{x}}_m)$ le cui coordinate sono costituite dalle medie aritmetiche totali delle variabili. In verità, nella formula:

$$\hat{\mathbf{x}}_i^t\hat{\mathbf{x}}_j = h_{ij} - h_{i0} - h_{0j} + h_{00} = h_{ij} - \left(h_{i0} + h_{0j} - h_{00}\right)$$

non si realizza altro che lo scorporo dello scarto differenziale dell'effetto di riga e di colonna dalle distanze (cioè l'effetto complessivo congiunto riga/colonna).

Se le h_{ij} sono raccolte in una matrice **H** di ordine (nxn) ed i prodotti scalari nella matrice **B,** pure di ordine (nxn), allora: **B=CHC** dove

$$\mathbf{C} = \left[\mathbf{I} - \frac{1}{\mathrm{n}}\mathbf{u}\mathbf{u}^{\mathrm{t}}\right] \text{ e } \mathbf{H} = -\frac{1}{2}\mathbf{D}$$

Da notare che la matrice **H** si ottiene agevolmente dalla matrice dei quadrati delle distanze euclidee dividendole per due e cambiando il segno da positivo in negativo. Poi, per esprimere la matrice dei prodotti interni **B**, basterà applicare due volte la matrice di centramento. Il risultato teorico che non si deve perdere di vista è che la matrice dei prodotti interni è stata ottenuta in due modi diversi: come prodotto della matrice dei dati per la sua trasposta: $\mathbf{B} = \hat{\mathbf{X}}\hat{\mathbf{X}}^t$ ed anche come doppio centramento della matrice **H** cioè **B=CHC.** Se fossero note le coordinate e si calcolassero le distanze euclidee in **D**, allora ci si troverebbe di fronte alla semplice scelta di esprimere in due modi alternativi le stesse informazioni. Il fascino dello scaling metrico è che la **B** ottenuta come doppio centramento di **H** consente una rappresentazione nello spazio euclideo a due o tre dimensioni anche di distanze basate su dati che non sono delle coordinate.

La riuscita della conversione tra distanze e coordinate è però legata alla condizione di euclideità della matrice **D** delle distanze. Tale vincolo si pone meglio sulla versione **H**. Se succede che la matrice

$$\mathbf{B} = \left[\mathbf{I} - \frac{1}{\mathrm{n}}\mathbf{u}\mathbf{u}^{\mathrm{t}}\right]\begin{bmatrix} 0 & -0.5d_{ij}^2 & -0.5d_{ik}^2 \\ -0.5d_{ij}^2 & 0 & -0.5d_{jk}^2 \\ -0.5d_{ik}^2 & -0.5d_{jk}^2 & 0 \end{bmatrix}\left[\mathbf{I} - \frac{1}{\mathrm{n}}\mathbf{u}\mathbf{u}^{\mathrm{t}}\right]$$

è (semi)positiva definita per ogni terna i, j, k allora la matrice **D** è considerata euclidea. In altre parole la condizone di euclideità è verificata se

$$2d_{ij}^2d_{ik}^2 + 2d_{ij}^2d_{jk}^2 + 2d_{ik}^2d_{jk}^2 - d_{jk}^4 - d_{ik}^4 - d_{ij}^4 \geq 0$$

che risulta più facile da controllare dato che non richiede il calcolo degli autovalori come nella condizione precedente.

Esempio

Ad un gruppo di giudici assaggiatori maschi è stato chiesto di comparare e valutare n=11 marche di birra diffuse sul mercato statunitense e di giudicarne soprattutto la somiglianza. Ogni giudice effettua ben 11*10/2=55 confronti a coppia. Si presume che il confronto di apprezzamento tra la birra A e la birra B dia sistematicamente lo stesso risultato, per tutti i giudici, del confronto della birra A con la birra B. La sintesi dei valori per i 55 confronti e per tutti i giudici assaggiatori è riportata nella seguente matrice

```
Samuel_Adams      0
Michelob          4     0
Budweiser        17    12   0
Corona           38    31  25   0
Coors            47    41  32  13   0
Budweiser_Lite   48    42  35  16   1   0
Miller_Lite      53    51  46  26   7   6  0
Amstel_Lite      52    49  43  23  11   8  9   0
Coors_Lite       55    54  50  36  37  21  5  18  0
Miller           44    39  30  15   3  10 19  28 27  0
Pabst            45    40  29  22  14  20 24  33 34  2  0
```

La matrice $\mathbf{B} = \mathbf{CHC}$ necessaria per arrivare allo scaling metrico è

```
        [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]
 [1,] 1234.64 1083.18  760.68  -37.18 -412.27 -470.55 -598.55 -518.05
 [2,] 1083.18  947.73  689.73   60.86 -291.73 -344.00 -638.00 -510.00
 [3,]  760.68  689.73  575.73   42.86 -149.23 -260.50 -581.50 -420.00
 [4,]  -37.18   60.86   42.86  135.00   57.91    3.64  -81.86   19.64
 [5,] -412.27 -291.73 -149.23   57.91  149.82  138.55  239.05  231.05
 [6,] -470.55 -344.00 -260.50    3.64  138.55  128.27  234.77  248.77
 [7,] -598.55 -638.00 -581.50  -81.86  239.05  234.77  377.27  364.77
 [8,] -518.05 -510.00 -420.00   19.64  231.05  248.77  364.77  433.27
 [9,] -504.23 -593.18 -571.18 -189.55 -218.64  234.59  567.09  445.59
[10,] -287.55 -223.50  -99.00   18.14  133.55   77.27   71.27 -112.23
[11,] -250.14 -181.09   12.41  -29.45  121.95    9.18   45.68 -182.82
        [,9]   [,10]   [,11]
 [1,] -504.23 -287.55 -250.14
 [2,] -593.18 -223.50 -181.09
 [3,] -571.18  -99.00   12.41
 [4,] -189.55   18.14  -29.45
 [5,] -218.64  133.55  121.95
 [6,]  234.59   77.27    9.18
 [7,]  567.09   71.27   45.68
 [8,]  445.59 -112.23 -182.82
 [9,]  781.91   89.59  -42.00
[10,]   89.59  126.27  206.18
[11,]  -42.00  206.18  290.09
```

Tale matrice è simile al risultato del prodotto di una matrice di dati per la sua trasposta $\hat{\mathbf{X}}\hat{\mathbf{X}}^t$ e noi faremo finta che sia stata ottenuta proprio in questo modo anche se la determinazione di **D** è avvenuta con una procedura interamente differente.

Determinazione delle coordinate

Nella discussione precedente si è solo stabilita una sorta di corrispondenza tra la matrice delle distanze euclidee e i prodotti scalari tra i vettori delle unità. Per progredire dobbiamo coinvolgere le componenti principali. Conviene, a questo fine esprimere **B** in termini della matrice dei dati centrati

$$\mathbf{B} = \hat{\mathbf{X}}\hat{\mathbf{X}}^t$$

Quindi **B** è simmetrica e non negativa definita. Se il rango di **B** (pari al rango di $\hat{\mathbf{X}}$) è p<m allora **B** può essere proposta nella sua scomposizione spettrale o valori singolari con la diagonalizzazione della matrice (n x n)

$$\hat{\mathbf{X}}\hat{\mathbf{X}}^t = \mathbf{B} = \sum_{r=1}^{p} \lambda_i \mathbf{V}_i \mathbf{V}_i^t = \mathbf{V}\mathbf{L}\mathbf{V}^t \quad \text{con} \quad \mathbf{V}^t\mathbf{V} = \mathbf{I}_n,\ L = diag\left(\lambda_1, \cdots, \lambda_p\right)$$

dove $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ sono gli autovalori non nulli di **B** disposti in ordine decrescente.Possiamo accostare le due espressioni di **B** nel modo che segue:

$$\hat{\mathbf{X}}\hat{\mathbf{X}}^t = \mathbf{V}\mathbf{L}^{0.5} \cdot \mathbf{L}^{0.5}\mathbf{V}^t \quad \text{dove} \qquad \mathbf{L}^{0.5} = \text{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p}\right)$$

dalla quale ricaviamo l'uguaglianza: $\mathbf{V}\mathbf{L}^{0.5} = \hat{\mathbf{X}}$. Se si conoscesse la matrice dei dati centrati $\hat{\mathbf{X}}$ questa relazione non aggiunge nulla, ma se le coordinate sono incognite allora possono essere ricostruite in base agli autovalori ed agli autovettori di **B**. Il fatto interessante è che per avere la **B** non è necessario conoscere le $\hat{\mathbf{X}}$; basta disporre dei quadrati delle distanze euclidee tra le unità. Come poi sia possibile ragionare della matrice delle distanze euclidee tra vettori di cui non si conoscono le coordinate sarà una scoperta interessante.

Ricordiamo che la matrice **B**, come la matrice $\hat{\mathbf{X}}$ da cui deriva, ha almeno un autovalore pari a zero. Infatti

$$\mathbf{B}\mathbf{u} = \hat{\mathbf{X}}\hat{\mathbf{X}}^t\mathbf{u} = \hat{\mathbf{X}} \cdot \mathbf{0} = \mathbf{0} \quad \text{e rango}\left(\mathbf{B}\right) \leq n - 1$$

Ne consegue che è sempre possibile trovare uno spazio R^{n-1} nel quale si possono rappresentare esattamente le distanze osservate in R^n. L'auspicio è, naturalmente, che le dimensioni necessarie siano più molte meno di (n-1), anche a costo di rinunciare alla esattezza della ricostruzione delle distanze in base alle coordinate.

L'insieme delle interdistanze tra punti è ricostruibile da una configurazione p-dimensionale dove p è il numero di autovalori non nulli di $\mathbf{B} = \hat{\mathbf{X}}\hat{\mathbf{X}}^t$ e quindi è il numero di colonne linearmente indipendenti presenti nella matrice dei dati centrati $\hat{\mathbf{X}}$. Tuttavia, **B** ha dimesioni (nxn) ed il calcolo pratico di autovalori ed autovettori non è agevole o numericamente stabile con i valori che assume solitamente n. Possiamo però aggirare questo ostacolo utilizzando la scomposizione in valori singolari della $\hat{\mathbf{X}}$.

Ogni matrice di ordine (nxm) può esssere articolata nel prodotto di tre altre matrici:

$$\hat{\mathbf{X}} = \mathbf{U}\mathbf{L}^{0.5}\mathbf{V}^t \qquad \text{con } \mathbf{L}^{0.5} = \text{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p}\right)$$

dove p è il numero di autovalori non nulli associato alla $\hat{\mathbf{X}}$, **U** è di ordine (nxp), **V** è di ordine (pxm) ed inoltre $\mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbf{I}$. La matrice **B** diventa

$$\mathbf{B} = \hat{\mathbf{X}}\hat{\mathbf{X}}^t = \left(\mathbf{U}\mathbf{L}^{0.5}\mathbf{V}^t\right)\left(\mathbf{V}\mathbf{L}^{0.5}\mathbf{U}^t\right) = \mathbf{U}\mathbf{L}\mathbf{U}^t$$

Per cui il calcolo dei suoi autovalori **L** e dei suoi autovettori, **U** può essere effettuato diagonalizzando **B**. D'altra parte, la matrice:

$$\hat{\mathbf{X}}^t\hat{\mathbf{X}} = \left(\mathbf{V}\mathbf{L}^{0.5}\mathbf{U}^t\right)\left(\mathbf{U}\mathbf{L}^{0.5}\mathbf{V}^t\right) = \mathbf{U}\mathbf{L}\mathbf{V}^t$$

ha gli stessi autovalori di **B** e conviene quindi ottenerli da questa visto che è di dimensioni (mxm). Inoltre, nota la matrice **U** è facile ottenere la **V**:

$$\hat{\mathbf{X}} = \mathbf{U}\mathbf{L}^{0.5}\mathbf{V}^t \Rightarrow \hat{\mathbf{X}}\mathbf{U} = \mathbf{V}\mathbf{L}^{0.5}\mathbf{U}^t\mathbf{U} = \mathbf{V}\mathbf{L}^{0.5} \Rightarrow \mathbf{L}^{0.5}\hat{\mathbf{X}}\mathbf{U} = \mathbf{V}$$

ne consegue che le operazioni su **B** possono essere sostituite con operazioni su $\hat{\mathbf{X}}^t\hat{\mathbf{X}}$ con la tecnica nota come analisi delle componenti principali. La matrice $\hat{\mathbf{X}}^t\hat{\mathbf{X}}$ non è una sconosciuta: essa infatti è la matrice delle devianze-codevianze tra le variabili del data set e qui sono più facilmente riconoscibili le dipendenze o quasi-dipendenze lineari in quanto corrispondono a perfette o forti correlazioni tra variabili.

È chiaro che questo approccio è possibile solo se, oltre alla matrice del interdistanze, è nota la matrice dei dati **X**. Se questa manca sarà necessario analizzare la sola **B** derivata dalla matrice delle distanze con i relativi problemi quando n è grande.

Esempio
Ragioniamo sulla seguente matrice delle distanze euclidee, comunque ottenuta.

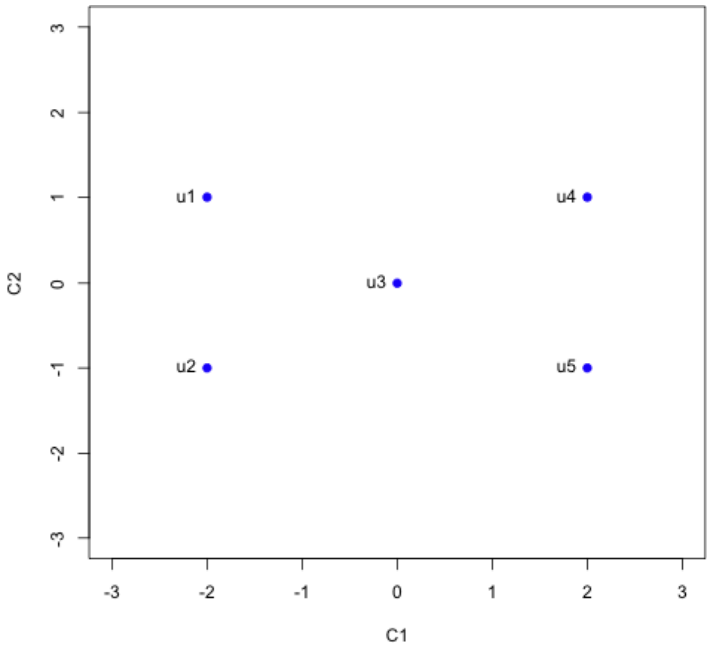
$$\mathbf{D}=\begin{bmatrix}0 & 4 & 5 & 16 & 20\\ 4 & 0 & 5 & 20 & 16\\ 5 & 5 & 0 & 5 & 5\\ 16 & 20 & 5 & 0 & 4\\ 20 & 16 & 5 & 4 & 0\end{bmatrix};\quad \mathbf{H}=-0.5\mathbf{D}=\begin{bmatrix}0 & -2 & -2.5 & -8 & -10\\ -2 & 0 & -2.5 & -10 & -8\\ -2.5 & -2.5 & 0 & -2.5 & -2.5\\ -8 & -10 & -2.5 & 0 & -2\\ -10 & -8 & -2.5 & -2 & 0\end{bmatrix}$$

$$\mathbf{C}=\begin{bmatrix}4/5 & -1/5 & -1/5 & -1/5 & -1/5\\ -1/5 & -4/5 & -1/5 & -1/5 & -1/5\\ -1/5 & -1/5 & -4/5 & -1/5 & -1/5\\ -1/5 & -1/5 & -1/5 & -4/5 & -1/5\\ -1/5 & -1/5 & -1/5 & -1/5 & -4/5\end{bmatrix};\quad \mathbf{CHC}=\begin{bmatrix}5 & 3 & 0 & -3 & -5\\ 3 & 5 & 0 & -5 & -3\\ 0 & 0 & 0 & 0 & 0\\ -3 & -5 & 0 & 5 & 3\\ -5 & -3 & 0 & 3 & 5\end{bmatrix}$$

Ecco gli autovalori: $\lambda_1=16;\ \lambda_2=4;\ \lambda_3=\lambda_4=\lambda_5=0;\ \sqrt{\lambda_1}=4;\ \sqrt{\lambda_2}=2$
Gli autovettori della scomposizione in valori singolari sono

$$\mathbf{V}_1=\begin{bmatrix}-1/2\\ -1/2\\ 0\\ 1/2\\ 1/2\end{bmatrix};\quad \mathbf{V}_2=\begin{bmatrix}1/2\\ -1/2\\ 0\\ 1/2\\ -1/2\end{bmatrix},\quad \mathbf{L}^{0.5}=\begin{bmatrix}4 & 0\\ 0 & 2\end{bmatrix};\quad \mathbf{VL}^{0.5}=\begin{bmatrix}-1/2 & 1/2\\ -1/2 & -1/2\\ 0 & 0\\ 1/2 & 1/2\\ 1/2 & -1/2\end{bmatrix}\begin{bmatrix}4 & 0\\ 0 & 2\end{bmatrix}=\begin{matrix}\hat{\mathbf{x}}_1 & \hat{\mathbf{x}}_2\end{matrix}\begin{bmatrix}-2 & 1\\ -2 & -1\\ 0 & 0\\ 2 & 1\\ 2 & -1\end{bmatrix}=\hat{\mathbf{X}}$$

In questo caso sono bastano solo due dimensioni per rappresentare tutti i punti. Inoltre la rappresentazione è esatta dato che sono solo due gli autovalori diversi da zero.

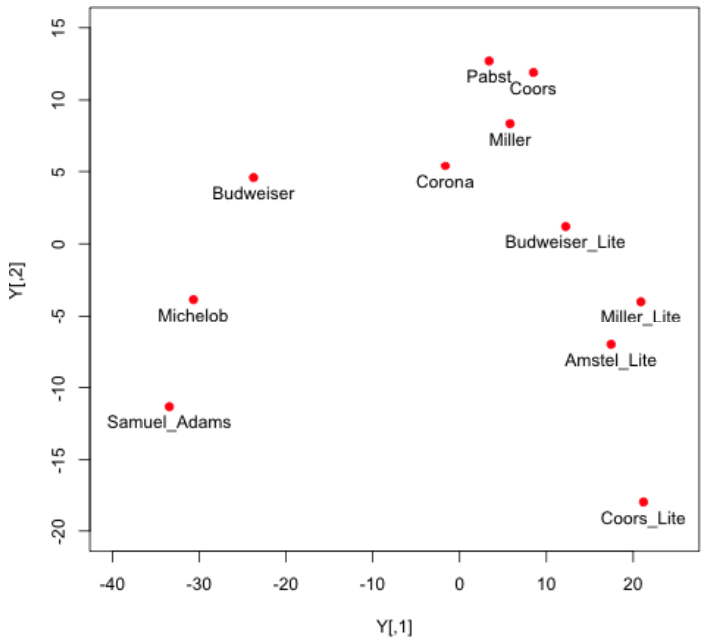


Le distanze tra i punti del grafico corrispondono alle entrate della matrice **D**.
In effetti:

$$
\begin{aligned}
d_{ij}^2 &= -2h_{ij} = h_{ii} + h_{jj} - 2h_{ij} \quad \left(\text{dato che } \; h_{i1} = h_{jj} = 0 \quad \forall i,j\right) \\
&= h_{ii} + h_{jj} - 2\left(h_{i0} + h_{0j} - h_{00} + y_i^t y_j\right) = \left(h_{ii} - 2h_{i0} + h_{00}\right) + \left(h_{jj} - 2h_{0j} + h_{00}\right) - 2\hat{\mathbf{x}}_i^t \hat{\mathbf{x}}_j \\
&= \hat{\mathbf{x}}_i^t \hat{\mathbf{x}}_i + \hat{\mathbf{x}}_j^t \hat{\mathbf{x}}_j - 2\hat{\mathbf{x}}_i^t \hat{\mathbf{x}}_j \quad \left\{\begin{array}{l}\text{in base alle relazioni ottenute con le sommatorie} \\ \text{ed alla definizione di } h_{ij}\end{array}\right\} \\
&= \left(\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\right)^t \left(\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\right)
\end{aligned}
$$

Le posizioni sono ricostruite le une relative alle altre dato che sono possibili rotazioni e traslazioni degli assi senza che si modifichino le distanze. Il senso dell'esempio è che se fossero disponibili le osservazioni sulle variabili come scarto dalla media, la matrice dei dati potrebbe essere sintetizzata senza errori dalle sue prime due componenti principali. In realtà conosciamo solo le distanze euclidee. Il grafico perciò rappresenta la configurazione dei punti che dà luogo alla migliore approssimazione della data matrice delle distanze **D**.

Esempio
Riprendiamo i dati sulla percezione delle marche di birra. Lo scalino metrico produce il seguente grafico



Coors Lite e Samuel Adams sono percepite come le meno prossime rispetto a tutte le altre marche. Pabst e Coors sìsembrano invece quelle più simili. Il grafico riporta una struttura tipica di questa metodologia: il ferro di cavallo (detto anche "effetto Guttman")-

Dissimilarità come distanze (analisi delle coordinate principali)

La discussione precedente, se fosse limitata nei soli termini con cui è stata proposta, sarebbe solo una disquisizione teorica, un divertimento matematico che si compiace delle relazioni tra coordinate e distanze. L'utilità pratica della tecnica di Young-Householder deriva invece dal fatto che la ricostruzione delle coordinate a partire dalle distanze non deve necessariamente basarsi su delle distanze o sul quadrato della distanza euclidea in particolare.

In effetti, data una qualsiasi matrice di dissimilarità tra unità d_{ij} in cui ogni entrata esprime la dissomiglianza tra le unità i e j. Nulla impedisce di definire

$$h_{ij} = -\frac{1}{2} d_{ij}^2, \quad \forall \ i, \ j$$

per poi chiedersi in quali circostanze sia possibile che la matrice **H** che include tutte queste quantità possa produrre una configurazione di pseudo-coordinate interpretabili come una rappresentazione nello spazio euclideo delle unità analizzate che ne rispetti la distanza originale.

Applichiamo ad **H** la procedura del doppio centramento per ottenere l'equivalente della **B** già vista nel trattamento della matrice dei dati:

$$\mathbf{B} = \mathbf{CHC} \qquad \text{dove} \quad \mathbf{C} = \left(I_n - \frac{1}{n} \mathbf{u}\mathbf{u}^t \right)$$

Siamo quindi pervenuti alla matrice $\mathbf{B} = \widehat{\mathbf{X}}\widehat{\mathbf{X}}^t$ ignorando tutto dei valori delle osservazioni che potrebbero averla prodotta Se **B** fosse almeno non semi definita positiva potremmo realizzare, grazie alla sua simmetria, una comoda approssimazione con la scomposizione in valori singolari

$$\mathbf{B} \cong \sum_{i=1}^{p} \lambda_i \mathbf{U}_i \mathbf{U}_i^t \qquad \lambda_i > 0 \quad i = 1,2,\ldots,p$$

ed ottenere le peusdo-coordinate da $\mathbf{Y} = \mathbf{U}\mathbf{L}^{0.5} \Rightarrow \mathbf{Y}_r = \sqrt{\lambda_r}\mathbf{U}_r \quad r = 1,2,\ldots,p$ dove le $\mathbf{Y}_r$ sono i punti dello spazio p-dimensionale. La distanza tra due punti qualsiasi, espressa con le pseudo-variabili, è:

$$\begin{aligned} \left(\mathbf{Y}_i - \mathbf{Y}_j\right)^t \left(\mathbf{Y}_i - \mathbf{Y}_j\right) &= \mathbf{Y}_i^t \mathbf{Y}_i + \mathbf{Y}_j^t \mathbf{Y}_j - 2\mathbf{Y}_i^t \mathbf{Y}_j \\ &= h_{ii} + h_{jj} - 2h_{ij} \\ &= -2h_{ij} = -2\left(-\frac{1}{2} d_{ij}^2\right) = d_{ij}^2 \end{aligned}$$

In base al teorema di Young e Householder, se **B** è positiva almeno semidefinita, allora i punti giacciono in un spazio euclideo e le dissimilarità da cui siamo partiti possono considerarsi in tutto e per tutto delle distanze. Da notare che le varianza-covarianze delle pseudo-varaibili ricostruite in base alla matrice delle distanze, coincide con la matrice diagonale degli autovalori

$$\mathbf{Y}^t\mathbf{Y} = \mathbf{L}^{0.5}\mathbf{U}^t\mathbf{U}\mathbf{L}^{0.5} = \mathbf{L}^{0.5}\mathbf{I}\mathbf{L}^{0.5} = \mathbf{L}.$$

La configurazione di punti nello spazio euclideo delle pseudo-variabii corrispondente ad una matrice di dissomiglianze può essere ottenuta purché la matrice **B** sia non negativa definita. A questo punto il problema diventa: come scegliere la funzione di distanza in modo che **B** abbia tale caratteristica?

Esempio

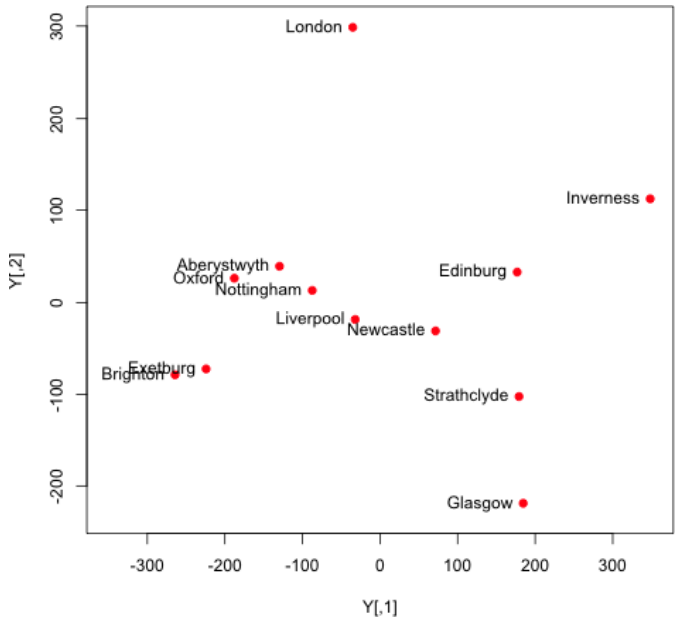
Illustriamo la tecnica dello scaling metrico con un esempio tratto da Cox e Cox (1994). Nella matrice sono riportate i tempi di percorrenza stradali tra n=12 città britanniche. Quindi non ci sono variabili che possono servire per una analisi dei legami tra le unità e la loro collocazione relativa.

Aberystwyth	0											
Brighton	300	0										
Edinburg	362	466										
Exetburg	217	238	431	0								
Glassgow	336	451	47	415	0							
Inverness	515	638	180	595	190	0						
Liverpool	138	271	229	236	214	393	0					
London	88	401	189	386	565	251	206	0				
Newcastle	292	349	139	371	169	316	180	284	0			
Nottingham	206	198	31	211	295	474	130	133	165	0		
Oxford	202	122	378	157	362	542	183	67	268	117	0	
Strathclyde	369	483	155	448	108	299	246	418	202	327	394	0

Gli autovalori della matrice **B** sono i seguenti

[1] 405976.8395 176539.1693 122412.8012 55433.6740 38713.3268 31805.9459
[7] 27532.1006 14364.2649 6035.2906 5668.0771 101.3473 0.000
[1] 45.8947 19.9573 13.8385 6.2666 4.3765 3.5956 3.1124 1.6238 0.6823
[10] 0.6408 0.0115 0.0000
[1] 45.8947 65.8521 79.6905 85.9572 90.3336 93.9292 97.0417 98.6655
[9] 99.3478 99.9885 100.0000 100.0000

La percentuale di variabilità spiegata dalle prime due coordinate è pari al 65.9% che non è entusiasmante, ma accettabile.



I risultati qui ottenuti sono diversi da quanto riportano Cox e Cox i quali trovano una somiglianza tra la posizione nel grafico e la collocazione geografica. Nel nostro caso emerse la posizione isolata di Londra per la sua centralità e di Inverness nel Nord della Scozia.

Il numero di dimensioni

Lo scaling metrico consente di ottenere una rappresentazione geometrica di punti le cui interdistanze coincidono con le dissomiglianze osservate in un collettivo di "n" unità. Quante pseudo-coordinate servono per riprodurre adeguatamente le distanze/dissomiglianze? Se **B** fosse di rango n-1 (il massimo visto che non può avere rango n) allora le distanze corrispondono all'espressione

$$d_{ij}^2 = \sum_{r=1}^{n-1} \lambda_r \left(x_{ir} - x_{jr} \right)^2$$

E dunque gli autovalori nulli o quasi-zero sono tali da rendere irrilevante il contributo delle coordinate a cui sono associati. Dobbiamo perciò sicuramente escludere gli autovalori nulli e considerare solo quelli non nulli.

Poiché per le rappresentazioni grafiche disponiamo di due o al massimo tre dimensioni, conserveremo solo i primi (1, 2, o 3) con l'adeguatezza di rappresentazione misurata da

$$\sum_{i=1}^{p} \lambda_i \Big/ \sum_{i=1}^{n} |\lambda_i| \quad \text{ovvero} \quad \sum_{i=1}^{p} \lambda_i \Big/ \sum_{\lambda_i > 0} \lambda_i$$

Schema dello scaling metrico

1. Calcolare le dissimilarità d_{ij}^2
2. Determinare la matrice **H** con $h_{ij} = -1/2\, d_{ij}^2$
3. Doppio centramento di **H: B=CHC**
4. Scomposizione in valori singolari di **B**. Se ci sono autovalori negativi, si ignorano (ovvero si interviene per attenuarne l'effetto, come vedremo)
5. Si sceglie il numero di dimensioni p=1, 2 o 3
6. Si rappresentano gli pseudo punti di coordinate $\mathbf{Y}_r = \sqrt{\lambda_r}\mathbf{U}_r \quad r = 1,\ldots,p$

Applicazione

Ad un gruppo di n=29 studenti è stato chiesto di esprimere una preferenza nei confronti binari delle seguenti attività

(1) Andare in palestra, (2) stare fuori a discutere con amici e amiche
(3) Andare al cinema, (4) Guardare un programma o un DVD in TV
(5) Parlare al telefono con l'amico/a più caro/a
(6) Papariare in casa

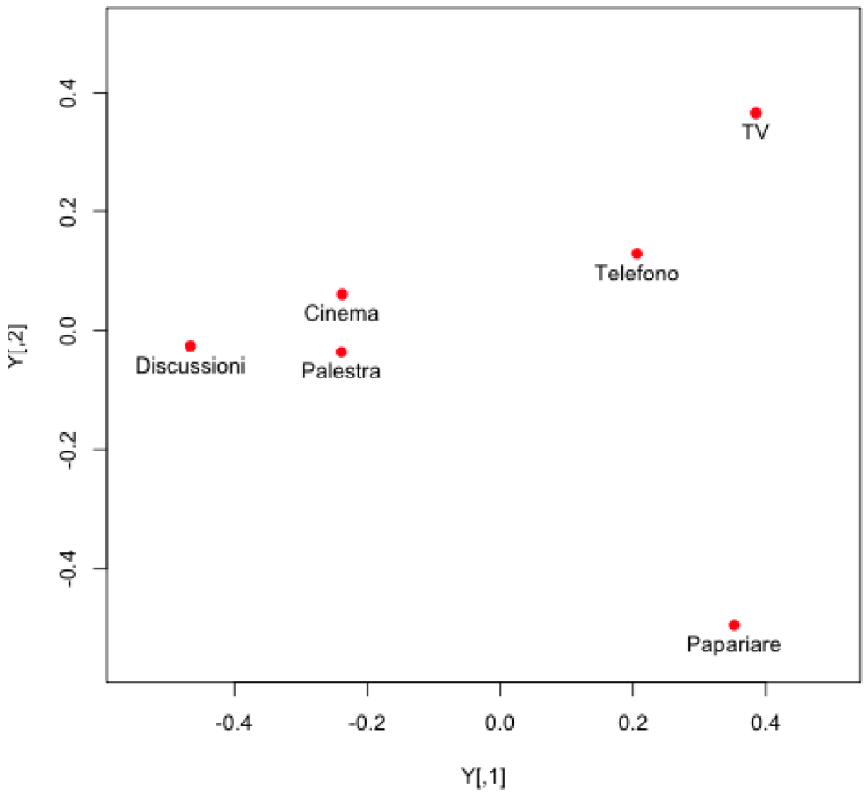
Per ciascuno dei 15 confronti ogni studente ha votato uno o due a seconda di quale delle due proposte preferisse dovendo però esclusivamente scegliere tra le due e senza la scappatoia dell'indifferenza e con l'obbligo di risposta.

Fra	E
(1) Andare in palestra	(2) Andare al Cinema
(1) Stare fuori a discutere in compagnia	(2) Andare al Cinema
(1) Guardare un programma in TV	(2) Andare al Cinema
(1) Parlare al telefono	(2) Andare al Cinema
(1) Papariare in casa	(2) Andare al Cinema
(1) Stare fuori a discutere in compagnia	(2) Andare in palestra
(1) Guardare un programma in TV	(2) Andare in palestra
(1) Parlare al telefono	(2) Andare in palestra
(1) Papariare in casa	(2) Andare in palestra
(1) Guardare un programma in TV	(2) Stare fuori a discutere in compagnia
(1) Parlare al telefono	(2) Stare fuori a discutere in compagnia
(1) Papariare in casa	(2) Stare fuori a discutere in compagnia
(1) Parlare al telefono	(2) Guardare un programma in TV
(1) Papariare in casa	(2) Guardare un programma in TV
(1) Papariare in casa	(2) Parlare al telefono

Nella tabella è riportato il voto medio aritmetico ottenuto da ogni confronto fra i 29 studenti (rapportato a 2 per la normalizzazione)

	Cinema	Palestra	Discussioni	TV	Telefono	Papariare
Cinema	0					
Palestra	0.705	0				
Discussioni	0.550	0.620	0			
TV	0.845	0.880	0.965	0		
Telefono	0.670	0.725	0.860	0.670	0	
Papariare	0.880	0.880	0.965	0.880	0.775	0

Autovalori, Perc. Variabilità spiegata e Variabilità spiegata cumulata
0.6455 0.4007 0.2515 0.2155 0.0903 0.0000
40.2548 24.9894 15.6861 13.4375 5.6323 0.0000
40.2548 65.2442 80.9302 94.3677 100.0000 100.0000



Sulla sinistra dell'asse orizzontale sono disposte le attività svolte all'esterno e sulla destra quelle svolte in casa per cui questo asse può essere inteso come socializzazione. L'asse verticale sembra indicare il gradimento con cui si svolgono le attività. Sembrerebbe che il cinema sia l'attività più gradita tra quelle a maggiore socializzazione e che la TV sia la preferita quando si è da soli. Trattandosi di giovani il semplice e sublime gusto del papariare non è ancora apprezzato.

Effetto Guttman ed effetto sigma

Consideriamo la matrice (n x n) definita come

$$\mathbf{D} = \mathbf{u}\mathbf{u}^t - \mathbf{A} \quad con \quad a_{ij} = \exp\left(-\alpha|i-j|\right), \quad \alpha > 0$$

Poiché sia $\mathbf{u}\mathbf{u}^t$ che **A** sono simmetriche lo sarà anche **D**; inoltre, gli elementi sulla diagonale di **D** sono pari a zero e gli elementi fuori diagonali sono positivi ed inferiori all'unità. Ecco un esempio per a=2 e n=4.

```
       [,1]   [,2]   [,3]   [,4]
[1,] 0.0000 0.8647 0.9817 0.9975
[2,] 0.8647 0.0000 0.8647 0.9817
[3,] 0.9817 0.8647 0.0000 0.8647
[4,] 0.9975 0.9817 0.8647 0.0000
```

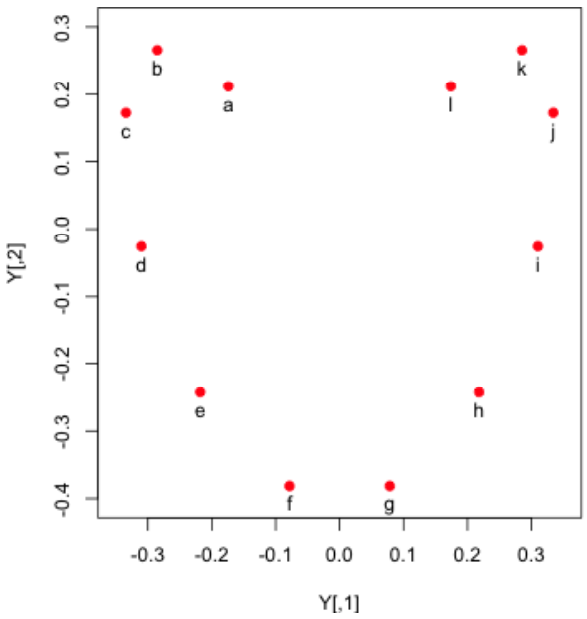
Da notare che gli elementi sulle righe aumentano allontanandosi dalla posizione in diagonale. Le distanze più grandi confondono e oscurano quelle più piccole.

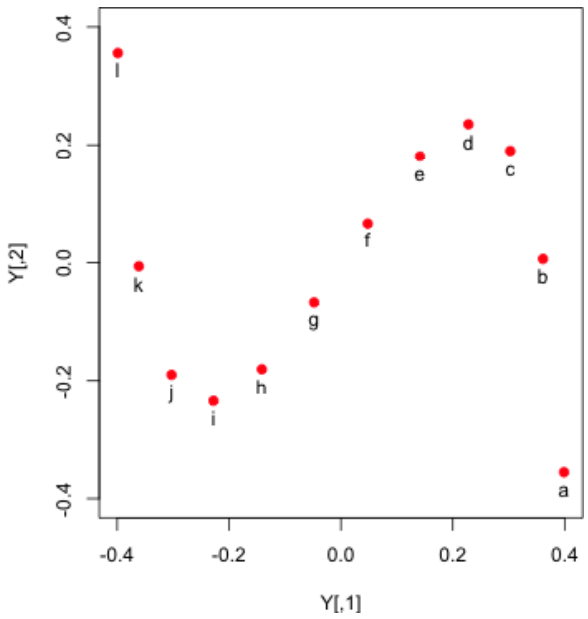
Un effetto meno noto è quello sigma che si realizza allorché la matrice delle affinità **A** è strutturata come

$$\mathbf{D} = \mathbf{u}\mathbf{u}^t - \left(\mathbf{I} - \mathbf{A}\right) \quad con \quad a_{ij} = \frac{|i-j|}{n^2}$$

```
       a      b      c      d
a 0.0000 0.9375 0.7500 0.4375
b 0.9375 0.0000 0.9375 0.7500
c 0.7500 0.9375 0.0000 0.9375
d 0.4375 0.7500 0.9375 0.0000
```

Lo scaling metrico è applicato alla matrice **D** nell'idea che essa sia equipollente ad una matrice euclidea di distanze





Aspetti metrici della dissimilarità

Consideriamo la matrice delle dissimilarità D=(d_{ij}) in cui sono state aggregate le (nxn) possibili dissomiglianze in un data set di n unità. La matrice **D** è detta metrica se la misura d_{ij} si comporta come una metrica e cioè

$$\begin{cases} d_{ij} = 0 \text{ se e solo se } x_i = x_j \\ d_{ij} = d_{ji} \quad \forall i,j \\ d_{ik} + d_{jk} \geq d_{ij} \quad \forall i,j,k \end{cases}$$

La matrice **D** è detta euclidea se gli n punti corrispondenti alle singole unità hanno una rappresentazione geometrica che riproduce delle distanze euclidee coincidenti con le d_{ij}. Quindi, se anche non si parte da coordinate per definire delle distanze euclidee è possibile determinare delle misure di dissimilarità che una volta fatte passare per lo scaling metrico portano comunque ad una configurazione di punti le cui distanze euclidee approssimano al meglio la matrice delle dissomiglianze o dissimilarità.

Risultati teorici importanti

1) Se d_{ij} non è una metrica allora è metrica la distanza:

$$d_{ij}^{+} = d_{ij} + c \quad \text{per } i \neq j$$

con

$$\phi \geq \max_{i,j,k} \left| -d_{ik} - d_{jk} + d_{ij} \right| \qquad (\phi \text{ può anche essere negativo})$$

In altre parole, se d_{ij}non verifica la disuguaglianza triangolare, la d_{ij}^{+} la verifica. Infatti

$$d_{ij}^{+} \leq d_{ik}^{+} + d_{jk}^{+} \Rightarrow d_{ij} + \phi \leq d_{ik} + d_{jk} + 2\phi \Rightarrow d_{ij} \leq d_{ik} + d_{jk} + \phi$$
$$d_{ij} - d_{ik} - d_{jk} \leq \phi$$

che è certamente verificata data la scelta di ϕ.

2) Se la matrice **D** è euclidea allora lo sono anche le matrici con elemento generico dato da

$$d_{ij}^{*} = d_{ij} + \phi^2 \qquad \forall\ \phi \ \text{ reale}$$
$$d_{ij}^{*} = \sqrt[\phi]{d_{ij}} \qquad \phi > 1$$
$$d_{ij}^{*} = d_{ij} \Big/ \left(d_{ij} + \phi^2 \right) \qquad \forall\ \phi \ \text{ real}$$

Queste trasformazioni tuttavia non dovrebbero essere necessarie dato che le dissi-milarità sono di solito normalizzate (variano tra zero ed uno) e le distanze di solito formano matrici euclidee. A questo proposito dobbiamo però ricordare che in alcuni contesti si usano metriche piuttosto diverse ad esempio dalle metriche di Minkowsky e pertanto la condizione di euclideità, particolarmente per quel che concerne la disuguaglianza triangolare, richiede un supporto.

Per rendere più docili dal punto di vista numerico la trattazione dello scaling metrico possono essere di aiuto alcuni risultati teorici.

3) La matrice **D** è euclidea solo se lo è la matrice **CHC** con **C** matrice di centramento e se

$$h_{ij} = -\frac{1}{2} d_{ij}^2$$

cioè **D** è euclidea se lo è la matrice $\mathbf{B} = \widehat{\mathbf{X}}\widehat{\mathbf{X}}^t$

4) Condizione di Lingoes. Se **D** è una matrice di dissimilarità allora esiste una costante c tale che la matrice

$$\mathbf{D}^{*} = \left(d_{ij}^2 \right)$$

è euclidea con

$$d_{ij}^{*} = \left[d_{ij}^2 + 2c \right]^{\frac{1}{2}} \ ; \ c \geq -\lambda_n , \ |c| < \min\left\{ d_{ij}^2 \right\} \quad i \neq j$$

dove λ_n è l'autovalore più piccolo (anche negativo) della matrice $\mathbf{B} = \mathbf{CHC}$. Infatti, la mancanza di euclideità nella matrice delle distanze provoca autovalori negativi.

5) Condizione di Cailliez. Se **D** è una matrice di dissimilarità allora esiste una costante c tale che la matrice

$$\mathbf{D}^{*} = d_{ij}^2 + c \qquad \text{dove} \ \ c \geq \lambda_1 , \quad i \neq j$$

è euclidea con λ_1 autovalore massimo della matrice aggregata

$$\mathbf{G} = \begin{bmatrix} \mathbf{0} & 2\mathbf{H} \\ -\mathbf{I}_n & -4\mathbf{H}_2 \end{bmatrix} \quad \text{dove} \quad \mathbf{H}_2 = -\frac{1}{2} d_{ij}$$

Per applicare le correzioni suddette è necessario calcolare gli autovaloi della matrice da usare come costanti di riferimento; tale calcolo può essere problematico se n è molto grande.

6) Se **D** è una matrice di dissimilarità e c una costante additiva allora, all'aumentare di c, gli autovalori di

$$\mathbf{D}^{+} = \mathbf{D} + c\left(\mathbf{u}\mathbf{u}^{t} - \mathbf{I}\right)$$

Tendono a zero tranne quello massimo che tende a c. L'autovettore associato tende al vettore formato da m uno e non si dà luogo ad alcuna rappresentazione geometrica dato che la matrice delle distanze produrrebbe solo un agglomerato compatto di punti.

Il problema della costante additiva

Come si è visto nell'esempio, non sempre la matrice **B** ricavata dalle dissomiglianze verfica la condizione di essere non negativa definita e qualcuno dei suoi autovalori è negativo. Nell'esempio si è semplicemente scelto di ignorarli, ma esistono altre possibilità.

Il primo approccio è di sommare una costante appropriata a tutte le dissomiglianze (tranne gli zeri)

$$d_{ij} = d_{ij} + c \qquad \forall \; i \neq j$$

tale che $\lambda_r \geq 0 \quad r = 1,2,\ldots,n-1$. Il problema è che non esiste un modo semplice di determinare la costante c. Il vantaggio comunque è chiaro: se **B** fosse non negativa definita allora esisterebbe una rappresentazione geometrica di punti le cui distanze sono le stesse delle dissomiglianze corrette additivamente in uno spazio euclideo. Il problema è ancora aperto e non è sempre determinabile in che modo la scelta di c influenzi la soluzione: all'aumentare di c aumenta l'uguaglianza tra gli autovalori è questo può ridurre la capacità rappresentativa delle coordinate principali. Le soluzioni proposte per il problema della determinazione di c debbono essere approfondite per comprendere quale sia la più efficiente: metricità al minor costo di alterazione delle dissimilarità originali.

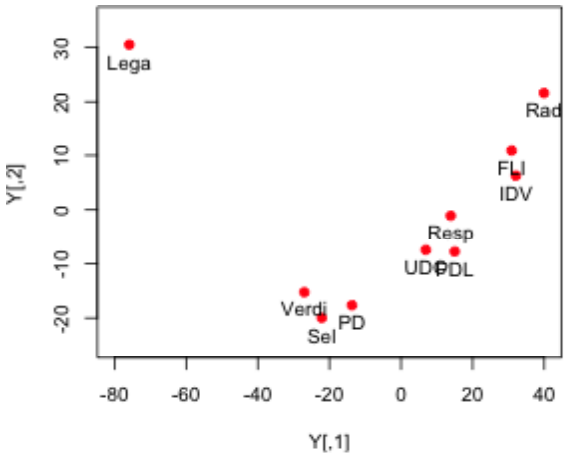
Il secondo approccio parte da una considerazione: se le dissomiglianze verificano la disuguaglianza triangolare allora sono delle distanze e quindi esprimibili in termini di una metrica necessariamente non negativa definita. Questo è però possibile solo per variabili su scala proporzionale; su scala intervallare potrebbe mancare lo zero di riferimento e quindi si può tentare di determinare una trasformazione lineare delle variabili che renda "metrica" la misura di dissomiglianza.

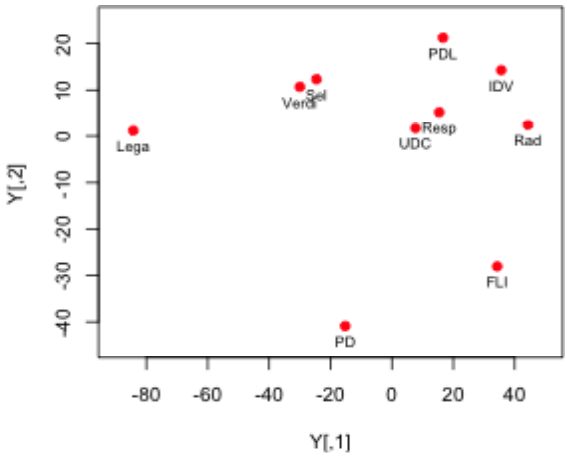
L'unica tecnica immediata che si può suggerire di provare varie scelte di c cercando il valore che avvicina (dal positivo) a zero l'ultimo auto valore. Una procedura del genere è incorporata in alcuni algoritmi che partono dal fatto che se **D** non è una matrice euclidea allora la si può rendere tale in base ai due schemi alternativi già visti di Lingoes e Cailliez. Con il primo metodo di si somma una costante positiva pari al doppio del valore assoluto dell'autovalore negativo più grande (sempre in valore assoluto). Tale costante è sommata agli elementi della matrice delle distanze, tranne che sulla diagonale. Lo scaling metrico si effettua sulla matrice corretta con l'aspettativa che almeno due degli autovalori saranno nulli, ma gli altri positivi. La procedura di Cailliez è simile alla precedente cambiando solo il modo di determinare la costante additiva . Quale preferire è una questione aperta

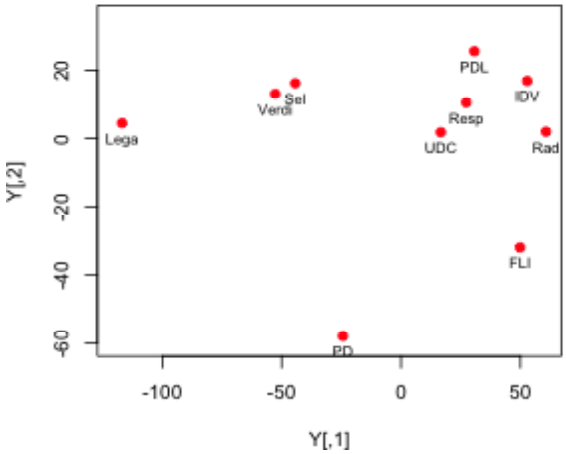
Esempio
Percezione di alcune formazioni politiche da parte di un campione di elettori. La distanza si basa sulla media dei punteggi attribuiti ai confronti a coppia.

PCI	0									
PSI	18	0								
PSDI	25	2	0							
PRI	25	1	2	0						
DC	43	7	4	6	0					
PLI	47	10	9	5	10	0				
MSI	41	10	6	6	12	6	0			
PR	22	25	31	46	47	46	33	0		
Verdi	23	29	37	56	54	56	42	4	0	
DP	43	73	84	105	116	105	83	18	18	0

Come è ben noto il quadrato della distanza euclidea non è necessariamente euclidea ed infatti quella ottenuta dalla matrice precedente non risulta euclidea. Per rappresentare i Partiti in uno spazio metrico occorre rendere euclidea la matrice delle distanze a mezzo della costante additiva. Abbiamo adoperato entrambe le opzioni di Lingoes e Cailliez







Le configurazioni ottenute con le correzioni additive sono molto diverse da quella calcolata ignorando la condizione di euclideità e riportata nel primo grafico. Questa, tuttavia, non può essere usata perché se la matrice delle distanze non è euclidea, i punti rappresentati nel grafico non hanno le distanze uguali a quelle della matrice originale.

Indeterminatezza della soluzione

Sia **Q** una matrice quadrata e simmetrica di ordine p ortogonale e quindi tale che $\mathbf{QQ}^t = \mathbf{I}$. In altre parole **Q** è una matrice di rotazione che muove gli assi solidalmente in modo che tra di essi si formino sempre degli angoli retti.
Se la soluzione ottenuta con lo scaling metrico

$$\mathbf{Y} = \mathbf{UL}^{0.5}$$

È ruotata a mezzo della matrice **Q**,

$$\mathbf{Y}^* = \mathbf{YQ}^t$$

i valori delle distanze non cambiano

$$\mathbf{Y}^{*}\mathbf{Y}^{*t} = \mathbf{Y}\mathbf{Q}^{t}\mathbf{Q}\mathbf{Y}^{t} = \mathbf{Y}\mathbf{Y}^{t} = \mathbf{B}$$

Per cui non esiste la soluzione dello scaling metrico, ma infinite come infinite sono le possibili rotazioni.

Esempio

Per p=2

$$\mathbf{Q} = \begin{bmatrix} \cos\theta & -sen\theta \\ sen\theta & \cos\theta \end{bmatrix} \Rightarrow \mathbf{Q}^{t}\mathbf{Q} = \begin{bmatrix} \cos\theta & sen\theta \\ -sen\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & -sen\theta \\ sen\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Se θ=30° allora $\cos(\theta) = \sqrt{3}/2,\ sen(\theta) = 1/2$ e quindi si ha

$$\mathbf{Q} = \begin{bmatrix} \sqrt{3}/2 & -\frac{1}{2} \\ \frac{1}{2} & \sqrt{3}/2 \end{bmatrix}$$

Supponiamo che

$$\begin{aligned} P_1 &= \left(\frac{1}{2},\ \frac{1}{2}\right) \\ P_2 &= \left(\frac{1}{4},\ \frac{3}{4}\right) \end{aligned} \Rightarrow \mathbf{Y} = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$$

La rotazione degli assi implica

$$\mathbf{Y}^{*} = \mathbf{Y}\mathbf{Q}^{t} = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} \begin{bmatrix} \sqrt{3}/2 & 1/2 \\ -1/2 & \sqrt{3}/2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{2\sqrt{3}-2}{8} & \frac{2\sqrt{3}+2}{8} \\ \frac{\sqrt{3}-3}{8} & \frac{3\sqrt{3}+1}{8} \end{bmatrix} = \frac{1}{8}\begin{bmatrix} 2\sqrt{3}-2 & 2\sqrt{3}+2 \\ \sqrt{3}-3 & 3\sqrt{3}+1 \end{bmatrix}$$

Le vecchie e nuove distanze sono

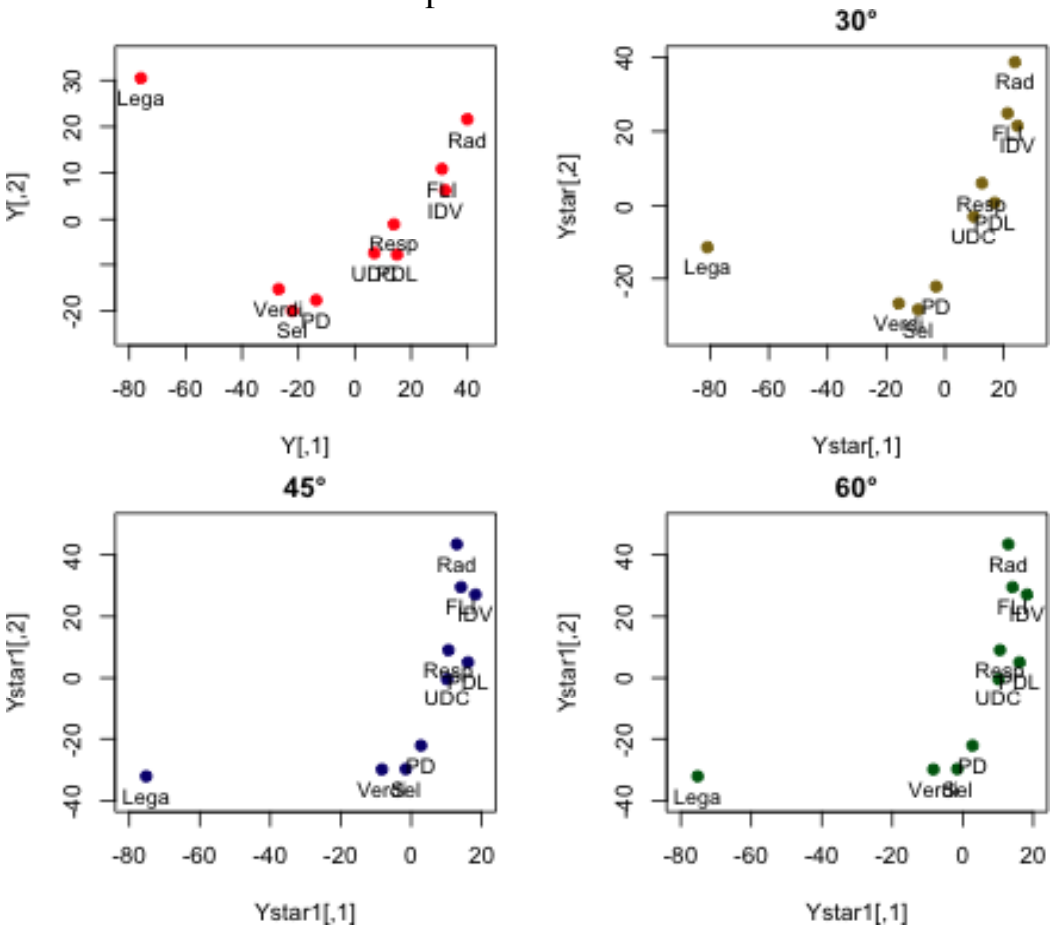
$$\|P_1 - P_2\|^2 = \left[\left(\frac{1}{2}-\frac{1}{4}\right)^2 + \left(\frac{1}{2}-\frac{3}{4}\right)^2\right] = \left[\frac{1}{16}+\frac{1}{16}\right] = \frac{1}{8}$$

$$\left\|P_1^{*} - P_2^{*}\right\|^2 = \left(\frac{1}{64}\right)\left[\left(2\sqrt{3}-2-\sqrt{3}+3\right)^2 + \left(2\sqrt{3}+2-3\sqrt{3}-1\right)^2\right]$$

$$= \left(\frac{1}{64}\right)\left[\left(1+\sqrt{3}\right)^2 + \left(1-\sqrt{3}\right)\right] = \frac{8}{64} = \frac{1}{8}$$

La distanza non cambia, ma la configurazione dei punti potrebbe essere diversa.

Esempio
Rotazione della matrice dei partiti



Nessuna delle rotazioni migliora la leggibilità del risultato: rimane in un'area a parte la lega, ma gli altri gruppi mantengono inalterate le loro posizioni relative.

Esercizio: In Clogg C.C: e Shihadeh F.G. (1994) "Statistical models for ordinal variables", è data la seguente matrice di scarti tra posizioni professionali.

	0.									
Professionisti	1.67	0								
Manager	2.24	1.13	0							
Clero	2.94	1.35	1.09	0						
Artigiani_1	1.96	0.95	0.28	1.24	0					
Artigiani_2	3.57	1.95	1.66	0.63	1.85	0				
Operai non nei trasporti	2.83	1.24	0.99	0.12	1.13	0.75	0			
Operai Agricol.	2.72	1.08	1.10	0.33	1.19	0.87	0.27	0		
Esercito, Polizia, altro	2.67	1.08	0.92	0.27	1.03	0.90	0.16	0.21	0	
Disoccupati	2.56	1.06	1.86	1.32	1.80	1.65	1.28	1.01	1.19	0

Predisporre lo scaling metrico con eventuale rotazione.

Scaling di matrici asimmetriche

La simmetria della matrice delle distanze/dissimilarità è un requisito tecnico, ma non sempre logico. Ad esempio, nelle matrici dei flussi, non è affatto detto che

le importazioni dalla zona A alla zona B sia identico a quello da B ad A: nell'interscambio tra aree partitiche non c'è alcuna ragione di pensare che gli elettori ed elettrici che votavano P1 ed ora votano P2 siano in quantità pari a quelli che prima votavano P2 ed ora votano P1. Lo stesso vale per i costi, le distanze, il traffico tra città, nelle esportazioni ed importazioni, nella mobilità tra professioni, tra posizioni sociali, tra località di emigrazione/immigrazione, etc. In realtà, la simmetria sembra più un requisito logico-matematico che una realtà concreta. Tuttavia, la nostra capacità di trattare matrici asimmetriche è ancora scarsa (ad esempio gli autovalori potrebbero essere complessi) per cui si tenterà di trasformare una matrice asimmetrica in una simmetrica sperando di conservare l'appropriato contenuto informativo.

Sia **G** una matrice asimmetrica (cioè tale che $g_{ij} \neq g_{ji}$.) Allora la matrice

$$G^*_{(\alpha)} = \left[g^*_{ij}(\alpha) \right] \; con \quad g^*_{ij}(\alpha) = \left(\frac{g^{\alpha}_{ij} + g^{\alpha}_{ji}}{2} \right)^{\frac{1}{\alpha}} \qquad \alpha > 0, \; g_{ij}, g_{ji} > 0$$

è simmetrica. In pratica, G* si ottiene realizzando la media potenziata di ordine α tra gli elementi corrispondenti delle matrici **G** e $\mathbf{G}^t$ (la sua trasposta). Tre casi particolari interessanti sono:

$$g^*_{ij} = \max\left\{ g_{ij}, g_{ji} \right\} \; (\alpha \to 0); \qquad g^*_{ij} = \left(g_{ij} + g_{ji} \right) / 2; \qquad g^*_{ij} = \sqrt{g_{ij} * g_{ji}} \; (\alpha \to 0)$$

Per passare da una matrice positiva ad una matrice di dissimilarità, la simmetrizzazione non è però sufficiente perché c'è il problema della diagonale.

Se le misurazioni riguardano delle affinità/prossimità tra unità allora l'elemento sulla diagonale deve essere il maggiore tra tutti gli elementi sulla riga o sulla colonna:

$$g_{ii} = \max_{1 \le j \le n} g_{ij}$$

In questo caso basterà dividere ogni riga e colonna di $\mathbf{G}^*$ per l'elemento sulla diagonale per ottenere una matrice con degli uno sulla diagonale e con elementi inferiori all'unità fuori diagonale. Se definiamo

$$\mathbf{L} = diag\left(1/g_{11}, \; 1/g_{22}, \ldots, 1/g_{nn} \right)$$

la matrice simmetrizzata con diagonale unitaria sarà ottenuta con

$$\mathbf{G}^* = \left[\frac{(\mathbf{GL})^{\alpha} + \left(\mathbf{LG}^t \right)^{\alpha}}{2} \right]^{\frac{1}{\alpha}}$$

La corrispondente matrice di dissomiglianze sarà ottenuta con

$$\mathbf{D} = \left(\mathbf{u}\mathbf{u}^t - \mathbf{G}^*\right)$$

in cui $d_{ij} = 1 - g_{ij}^*$. Se invece la **G*** non ha elemeni sulla diagonale maggiori degli altri possono insorgere problemi sulla sua considerazione come matrice di affinità o prossimità. In questo caso occorrerà approfondire il problema. quando la matrice **G** include misurazioni relative alla dissomiglianza occorre azzerare gli elementi sulla diagonale sottraendo l'elemto sulla diagonale da quelli situati sulla stessa riga e colonna. La positività della matrice sarà preser-vata se

$$g_{ii} = \min_{1 \le j \le n} g_{ij}$$

In pratica si ottiene la matrice di distanza con l'operazione matriciale

$$\mathbf{D} = \mathbf{G}^* - diag\left(1/g_{11}^*,\ 1/g_{22}^*,,,\ldots,1/g_{nn}^*\right)\mathbf{u}\mathbf{u}^t$$

in questo modo **D** ha elementi zero sulla diagonale ed è simmetrica.

Esempio

Flusso di pendolari tra località. Grado di prossimità tra i punti.

	A	B	C	D	E	F
A	100	60	20	30	40	20
B	50	150	25	40	30	10
C	40	70	80	50	20	5
D	30	30	60	70	40	10
E	20	10	20	30	45	10
F	10	5	20	60	10	90

Poiché c'è la diagonale con valori superiori alle altre entrate, si divide ogni colonna/riga per il massimo che si trova sulla loro posizione diagonale. La simmetrizzazione si realizza, ad esempio, con la media geometrica. Si trasformano poi le prossimità in distanze.