Dalle distanze tra unità alle coordinate (Scaling multidimensionale)

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

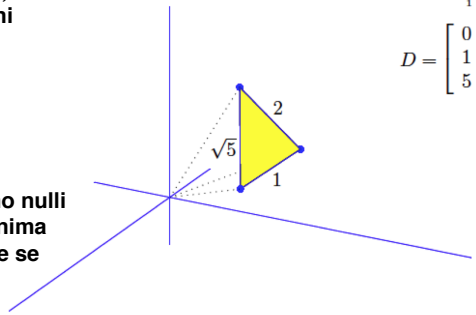
$$D = [d_{ij}] = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{bmatrix} = \begin{bmatrix} 0 & d_{12} & d_{13} \\ d_{12} & 0 & d_{23} \\ d_{13} & d_{23} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 5 \\ 1 & 0 & 4 \\ 5 & 4 & 0 \end{bmatrix}$$

La matrice D ha n^2=9 valori, ma solo n(n-1)/2=3 informazioni distinte dato che

La matrice delle distanze è simmetrica.

I valori sulla diagonale sono nulli per indicare la distanza minima che sussiste tra una unità e se stessa

$$D = \begin{array}{c} \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{bmatrix} 0 & 1 & 5 \\ 1 & 0 & 4 \\ 5 & 4 & 0 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \end{matrix}$$



Matrice di distanza metrica

Consideriamo la matrice delle dissimilarità/distanze D=(d_{ij}) in cui sono state aggregate le (nxn) possibili comparazioni in un set di n unità.

La matrice D è detta metrica se d_{ij} si comporta come una metrica e cioè.

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

Positività, identità, simmetria, disuguaglianza triangolare

La matrice metrica quindi è simmetrica, ha elementi positivi tranne che sulla diagonale che può contenere solo degli zeri.

Inoltre, ogni sua terna di elementi, comunque scelta, verifica la disuguaglianza triangolare.

Matrice dei prodotti incrociati

La seguente matrice ha un ruolo speciale nell'ambito dello scaling metrico.

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

La matrice scarto verifica il vincolo

$$\hat{X}C = 0$$

e quindi il rango della matrice di scarti potrà al massimo essere (n-1)

L'elemento generico della matrice B è

$$b_{i,j} = \hat{x}_i^{\,t}\hat{x}_j = \sum_{r=1}^{m} \hat{x}_{i,r}\hat{x}_{j,r}$$

$$\text{dove } C = \left(I_n - \frac{1}{n}u_n u_n^t\right) \quad u_n^t = \overbrace{[1,1,\ldots,1]}^{\text{n-volte}}$$

Dove m è il numero di variabili nella matrice dei dati.

B 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.

Esempio

$$X = \begin{bmatrix} 2 & -1 \\ -3 & 8 \\ 3 & 4 \\ -4 & 6 \\ 7 & -2 \end{bmatrix}; \quad C = \begin{bmatrix} 4/5 & -1/5 & -1/5 & -1/5 & -1/5 \\ -1/5 & 4/5 & -1/5 & -1/5 & -1/5 \\ -1/5 & -1/5 & 4/5 & -1/5 & -1/5 \\ -1/5 & -1/5 & -1/5 & 4/5 & -1/5 \\ -1/5 & -1/5 & -1/5 & -1/5 & 4/5 \end{bmatrix}$$

La matrice dei prodotti incrociati ha dimensione (n x n)

$$\hat{X} = CX = \begin{bmatrix} 4/5 & -1/5 & -1/5 & -1/5 & -1/5 \\ -1/5 & 4/5 & -1/5 & -1/5 & -1/5 \\ -1/5 & -1/5 & 4/5 & -1/5 & -1/5 \\ -1/5 & -1/5 & -1/5 & 4/5 & -1/5 \\ -1/5 & -1/5 & -1/5 & -1/5 & 4/5 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -3 & 8 \\ 3 & 4 \\ -4 & 6 \\ 7 & -2 \end{bmatrix} = \begin{bmatrix} 1 & -4 \\ -4 & 5 \\ 2 & 1 \\ -5 & 3 \\ 6 & -5 \end{bmatrix}$$

$$B = \hat{X}\hat{X}^t = \begin{bmatrix} 1 & -4 \\ -4 & 5 \\ 2 & 1 \\ -5 & 3 \\ 6 & -5 \end{bmatrix} \begin{bmatrix} 1 & -4 & 2 & -5 & 6 \\ -4 & 5 & 1 & 3 & -5 \end{bmatrix} = \begin{bmatrix} 17 & -24 & -2 & -17 & 26 \\ -24 & 41 & -3 & 35 & -49 \\ -2 & -3 & 5 & -7 & 7 \\ -17 & 35 & -7 & 34 & -45 \\ 26 & -49 & 7 & -45 & 61 \end{bmatrix}$$

Trasformazione di Gower

Esiste una precisa relazione

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

dove b=diag(B) cioè contiene gli elementi posti sulla diagonale della matrice B.

Il significato della relazione è che il complesso di informazioni presente nella matrice dei dati può essere espresso sia come matrice dei prodotti incrociati che come distanze euclidee (al quadrato).

Ricordiamo che D^2 è formata con il quadrato delle distanze euclidee e non con le distanze euclidee tra le unità.

$$\boldsymbol{D}^2 = \boldsymbol{D} \otimes \boldsymbol{D}$$

In questo caso lilprodotto tra matrici è realizzato elemento per elemento (prodotto di Hadamard o prodotto di Kronecker)

Esempio

$$X=\begin{bmatrix}-1&2\\4&-1\\3&0\\6&3\end{bmatrix};\quad u=\begin{bmatrix}1\\1\\1\\1\end{bmatrix}\qquad C=\begin{bmatrix}3/4&-1/4&-1/4&-1/4\\-1/4&3/4&-1/4&-1/4\\-1/4&-1/4&3/4&-1/4\\-1/4&-1/4&-1/4&3/4\end{bmatrix}\quad \hat{X}=CX=\begin{bmatrix}3/4&-1/4&-1/4&-1/4\\-1/4&3/4&-1/4&-1/4\\-1/4&-1/4&3/4&-1/4\\-1/4&-1/4&-1/4&3/4\end{bmatrix}\begin{bmatrix}-1&2\\4&-1\\3&0\\6&3\end{bmatrix}=\begin{bmatrix}-4&1\\1&-2\\0&-1\\3&2\end{bmatrix}$$

$$B=\hat{X}\hat{X}'=\begin{bmatrix}17&-6&-1&-10\\-6&5&2&-1\\-1&2&1&-2\\-10&-1&-2&13\end{bmatrix};b=\begin{bmatrix}17\\5\\1\\13\end{bmatrix}\Rightarrow\begin{bmatrix}17\\5\\1\\13\end{bmatrix}\begin{bmatrix}1&1&1&1\end{bmatrix}+\begin{bmatrix}1\\1\\1\\1\end{bmatrix}\begin{bmatrix}17&5&1&13\end{bmatrix}-2\begin{bmatrix}17&-6&-1&-10\\-6&5&2&-1\\-1&2&1&-2\\-10&-1&-2&13\end{bmatrix}$$

$$=\begin{bmatrix}17&17&17&17\\5&5&5&5\\1&1&1&1\\13&13&13&13\end{bmatrix}+\begin{bmatrix}17&5&1&13\\17&5&1&13\\17&5&1&13\\17&5&1&13\end{bmatrix}-\begin{bmatrix}34&-12&-2&-20\\-12&10&4&-2\\-2&4&2&-4\\-20&-2&-4&26\end{bmatrix}=\begin{bmatrix}0&34&20&50\\34&0&2&20\\20&2&0&18\\50&20&18&0\end{bmatrix}$$

$$d_{12}=(-1-4)^2+(2-(-4))^2=25+9=34;\ d_{34}=(3-6)^2+(0-3)^2=9+9=18;\ d_{41}=(6-(-1))^2+(3-2)^2=49+1=50$$

$$D^2=\begin{bmatrix}0&34&20&50\\34&0&2&20\\20&2&0&18\\50&20&18&0\end{bmatrix}$$

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

La matrice delle distanze euclidee al quadrato può essere ottenuta dalla matrice dei prodotti incrociati.

Inversione

Il senso dello scaling metrico è di invertire tale relazione e cioè partire dalla matrice delle distanze ed arrivare alla matrice dei prodotti incrociati.

$$\underset{nxn}{B} = \frac{1}{2}\left[\underset{nx1}{b}\,\underset{1xn}{u^t} + \underset{1xn}{u}\,\underset{nx1}{b^t} - \underset{nxn}{D^2}\right]$$

Notate che B compare sia a sinistra che a destra della relazione. Dobbiamo superare questa contraddizione.

Qui può essere d'aiuto lo sviluppo del quadrato della distanza euclidea

$$d_{ij}^2 = \left(\tilde{x}_i - \tilde{x}_j\right)^t\left(\tilde{x}_i - \tilde{x}_j\right) = \tilde{x}_i^t\tilde{x}_i + \tilde{x}_j^t\tilde{x}_j - 2\tilde{x}_i^t\tilde{x}_j$$
$$= b_{ii} + b_{jj} - 2b_{ij}$$

Trattandosi di prodotti interni (cioè che danno uno scalare come risultato) si ha

$$\hat{\boldsymbol{x}}_{\boldsymbol{i}}{}^t\hat{\boldsymbol{x}}_{\boldsymbol{j}} = \hat{\boldsymbol{x}}_{\boldsymbol{j}}{}^t\hat{\boldsymbol{x}}_{\boldsymbol{i}}$$

Inversione/2

Per comodità consideriamo la distanza tra le due unità "I" e "j" facendo intervenire una unità di comodo: la "k"

$$b_{ij} = x_i^t x_j = \left(x_i - x_k\right)^t\left(x_j - x_k\right) = \left\|x_i - x_k\right\|\left\|x_j - x_k\right\|\cos\left(\theta_{ij}\right)$$

Deriva dalla legge dei coseni (interpretazione del prodotto scalare

Ovvero $\boldsymbol{b_{ij} = d_{ki}d_{kj}\cos\left(\theta_{ij}\right)}$

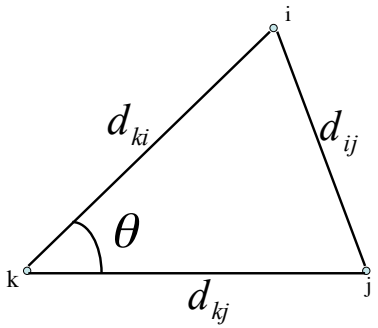
Ripetiamo la stessa operazione sviluppando il quadrato

$$d_{ij}^2 = \left\|x_i - x_j\right\|^2 = \left\|\left(x_i - x_k\right) - \left(x_j - x_k\right)\right\|^2 = \left\|\left(x_i - x_k\right)\right\|^2 + \left\|\left(x_j - x_k\right)\right\|^2 - 2\left(x_i - x_k\right)^t\left(x_j - x_k\right)$$
$$= \left\|\left(x_i - x_k\right)\right\|^2 + \left\|\left(x_j - x_k\right)\right\|^2 - \left\|\left(x_i - x_k\right)\right\|\left\|\left(x_i - x_k\right)\right\|\cos\left(\theta_{ij}\right)$$
$$= d_{ki}^2 + d_{kj}^2 - 2d_{ki}d_{kj}\cos\left(\theta_{ij}\right)$$

Inversione/3

Sostituendo l'espressione di b_{ij} si ottiene $d_{ij}^2 = d_{ki}^2 + d_{kj}^2 - 2b_{ij}$

Portando a sinistra il prodotto incrociato abbiamo



$$b_{ij} = \frac{1}{2}\left(d_{ki}^2 + d_{kj}^2 - d_{ij}^2\right)$$

Quindi, per descrivere i prodotti incrociati è necessario solo conoscere le distanze euclidee

La matrice H

Definiamo la matrice H di ordine (nxn) con elementi $h_{ij} = -0.5d_{ij}^2$

$$\boldsymbol{H} = -\frac{1}{2}\boldsymbol{D}^2$$

La matrice H si ottiene agevolmente dalla matrice dei quadrati delle distanze dividendole per due e cambiando il segno.

$$\boldsymbol{D} = \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{bmatrix} \rightarrow \boldsymbol{H} = \begin{bmatrix} 0 & & & & \\ -2 & 0 & & & \\ -18 & -12.5 & 0 & & \\ -50 & -40.5 & -8 & 0 & \\ -40.5 & -32 & -12.5 & -4.5 & 0 \end{bmatrix}$$

Bicentramento delle distanze

Si può agevolmente verificare che

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

Questi sono nulli perché è nulla la somma di riga (o di colonna) della matrice di centramento: $u^tC=0$, $Cu=0$.

$$\begin{aligned} CHC &= C\left[-\frac{1}{2}D^2\right]C = C\left[-\frac{1}{2}\left(bu^t + ub^t - 2B\right)\right]C \\ &= -\frac{1}{2}C\left(bu^t\right)C - \frac{1}{2}C\left(ub^t\right)C + CBC \\ &= -0 - 0 + C\left(\hat{X}\hat{X}^t\right)C = C\left(CXX^tC\right)C \\ &= CXX^tC = \hat{X}\hat{X}^t = B \end{aligned}$$

CC=C

$CX = \hat{X}$

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

Esempio

$$D^2 = \begin{bmatrix} 0 & 34 & 20 & 50 \\ 34 & 0 & 2 & 20 \\ 20 & 2 & 0 & 18 \\ 50 & 20 & 18 & 0 \end{bmatrix} \Rightarrow H = -\frac{1}{2}\begin{bmatrix} 0 & 34 & 20 & 50 \\ 34 & 0 & 2 & 20 \\ 20 & 2 & 0 & 18 \\ 50 & 20 & 18 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -17 & -10 & -25 \\ -17 & 0 & -1 & -10 \\ -10 & -1 & 0 & -9 \\ -25 & -10 & -9 & 0 \end{bmatrix}$$

$$C = \begin{bmatrix} 3/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \end{bmatrix}$$

$$\tilde{B} = CHC$$

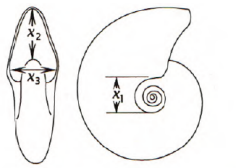
Si usa la tilde perché è una matrice dei prodotti incrociati ricostruita

$$CX = \begin{bmatrix} 3/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \end{bmatrix}\begin{bmatrix} 0 & -17 & -10 & -25 \\ -17 & 0 & -1 & -10 \\ -10 & -1 & 0 & -9 \\ -25 & -10 & -9 & 0 \end{bmatrix} = \begin{bmatrix} 13 & -10 & -5 & -14 \\ -4 & 7 & 4 & 1 \\ 3 & 6 & 5 & 2 \\ -12 & -3 & -4 & 11 \end{bmatrix};$$

$$CXC = \begin{bmatrix} 13 & -10 & -5 & -14 \\ -4 & 7 & 4 & 1 \\ 3 & 6 & 5 & 2 \\ -12 & -3 & -4 & 11 \end{bmatrix}\begin{bmatrix} 3/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \end{bmatrix} = \begin{bmatrix} 17 & -6 & -1 & -10 \\ -6 & 5 & 2 & -1 \\ -1 & 2 & 1 & -2 \\ -10 & -1 & -2 & 13 \end{bmatrix} = \tilde{B}$$

Dalla matrice delle distanze euclidee al quadrato si può passare alla matrice dei prodotti incrociati e viceversa. Qui si ha $\tilde{B} = B$

Esempio



	Umbilical diameter (mm)	Height of whorl (mm)	Width of whorl (mm)
	X_1	X_2	X_3
Species A	4	27	18
Species B	12	25	12
Species C	10	23	16
Species D	14	21	14
Means	10	24	15

	scarti			XX^t			
	X1	X2	X3	X1	X2	X3	
A	4	27	18	-6	3	3	54 -18 0 -36
B	12	25	12	2	1	-3	-18 14 -4 8
C	10	23	16	0	-1	1	0 -4 2 2
D	14	21	14	4	-3	-1	-36 8 2 26
Media	10	24	15				

D

	A	B	C	D
A	0.0000	10.1980	7.4833	12.3288
B	10.1980	0.0000	4.8990	4.8990
C	7.4833	4.8990	0.0000	4.8990
D	12.3288	4.8990	4.8990	0.0000

H

	A	B	C	D
A	0	-52	-28	-76
B	-52	0	-12	-12
C	-28	-12	0	-12
D	-76	-12	-12	0

B

	[,1]	[,2]	[,3]	[,4]
[1,]	54	-18	0	-36
[2,]	-18	14	-4	8
[3,]	0	-4	2	2
[4,]	-36	8	2	26

Dalle coordinate arriviamo alle distanze e da queste ai prodotti incrociati per poi tornare alle distanze ma su un numero minore di coordinate.

Buscar el levante por el poniente

Applicazione: beer data

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.

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

Ogni giudice effettua 11*10/2=55 confronti a coppia.

La sintesi dei valori per i 55 confronti e per tutti i giudici assaggiatori è riportata nella matrice

Beer data/2

La matrice necessaria per arrivare allo scaling è la seguente

$$\tilde{B} = CHC$$

Tale matrice è SIMILE al risultato del prodotto di una matrice di dati per la sua trasposta

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

anche se il calcolo è avvenuto con una procedura differente.

	[,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

Matrice dei quadrati delle distanze

Il risultato teorico che non si deve perdere di vista è che la matrice dei prodotti interni può essere ottenuta in due modi diversi

- **come prodotto della matrice dei dati (scarti) per la sua trasposta:** $B = \hat{X}\hat{X}^t$
- **Come doppio centramento della matrice H cioè** $\tilde{B} = CHC$

Se fossero già 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 fornisce una rappresentazione nello spazio euclideo a due o tre dimensioni anche per distanze che non nascono da delle coordinate.

Le coordinate principali

La matrice dei prodotti incrociati può essere espressa nella sua scomposizione in valori singolari

$$\tilde{B} = \sum_{r=1}^{p} \lambda_r v_r v_r^{\,t} = VLV,\ con\ \ V^tV = I_n,\ \ L = diag\left(\lambda_1, \lambda_2, \cdots, \lambda p\right)$$

Dove $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ sono gli autovalori non nulli di $\tilde{B}$ disposti in ordine decrescente di grandezza e v_r è l'autovettore ortogonale associato.

Possiamo accostare le due matrici di prodotti incrociati

$$\tilde{B} = VLV = VL^{0.5}L^{0.5}V = \left(VL^{0.5}\right)\left(VL^{0.5}\right)^t \quad con\ \ L^{0.5} = diag\left(\sqrt{\lambda}_1, \sqrt{\lambda_2}, \cdots, \lambda p\right)$$

dalla quale ricaviamo la definizione $VL^{0.5} = \tilde{X}$

Gli autovettori per la radice degli autovalori danno delle pesudo-coordinate ovvero le coordinate principali

Caratteristica delle pseudo-variabili

Se la matrice dei dati è sconosciuta non rimane che agire sulla matrice $\tilde{B}$ ottenuta dalle matrice delle distanze

$$\tilde{X} = VL^{0.5}$$

Le varianze-covarianze delle pseudo-variabili coincide con la matrice diagonale degli autovalori

$$\tilde{X}^t\tilde{X} = \left(VL^{0.5}\right)^t\left(VL^{0.5}\right) = L^{0.5}V^tVL^{0.5} = L^{0.5}IL^{0.5} = L^{0.5}L^{0.5} = L$$

Ogni autovalore rappresenta una varianza. Maggiore è l'autovalore più grande è la variabilità della pseudo-variabile.

Inoltre, le pseudo-variabili sono tra di loro incorrelate.

Infine a causa del centramento, le pseudo-variabili hanno media zero.

Esempio

Consideriamo la seguente matrice delle distanze euclidee

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

$$\lambda_1 = 16;\ \lambda_2 = 4;\ \lambda_3 = \lambda_4 = \lambda_5 = 0;$$

$$\sqrt{\lambda_1} = 4;\ \sqrt{\lambda_2} = 2$$

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

Queste sono le pseudo-coordinate ovvero le coordinate principali

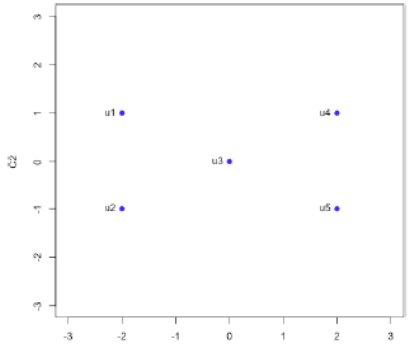
Gli autovettori della scomposizione in valori singolari sono

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

Esempio (continua)

In questo caso sono bastate solo due dimensioni per rappresentare tutti i punti.

Inoltre la ricostruzione è esatta dato che sono solo due gli autovalori diversi da zero.

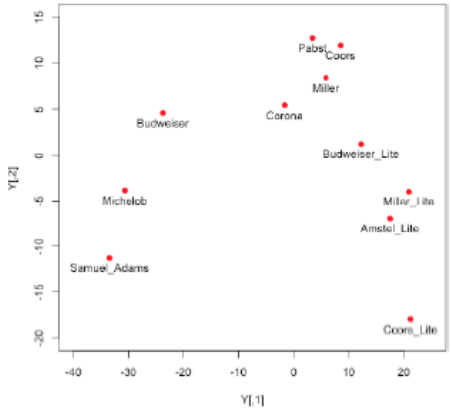


$$\tilde{X}\tilde{X}^t = \begin{bmatrix} -2 & 1 \\ -2 & -1 \\ 0 & 0 \\ 2 & 1 \\ 2 & -1 \end{bmatrix}\begin{bmatrix} -2 & -2 & 0 & 2 & 2 \\ 1 & -1 & 0 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 4 & 5 & 16 & 20 \\ 4 & 0 & 5 & 20 & 16 \\ 5 & 5 & 0 & 5 & 5 \\ 16 & 20 & 5 & 0 & 4 \\ 20 & 16 & 5 & 4 & 0 \end{bmatrix}$$

Esempio sulle marche di birra

Coors Lite e Samuel Adams sono percepite come le meno prossime rispetto alle altre marche. Pabst e Coors sembrano invece quelle più simili.

Il grafico riporta una struttura tipica di questa metodologia: il ferro di cavallo (detto anche "effetto Guttman")-



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 fatto interessante è che per arrivare alla matrice B non è stato necessario conoscere le variabili X.
Ono bastati iquadrati delle distanze.

Applicazione: Distanze città britanniche

L'avvio delle elaborazioni è la costruzione di una matrice di distanza: chilometri, tempi di percorrenza, movimenti dei pendolari, etc.

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

Non ci sono delle vere e proprie variabili.

Il senso dell'esempio è che se fossero disponibili le osservazioni sulle variabili, la matrice dei dati potrebbe essere sintetizzata dalle sue prime due pseudo-coordinate.

In realtà conosciamo solo le distanze.

Applicazione: bicentramento

	Aberystwyth	Brighton	Edinburg	Exetburg	Glasgow	Inverness	Liverpool
Aberystwyth	0.0	-45000.0	-65522.0	-23544.5	-56448.0	-132612.5	-9522.0
Brighton	-45000.0	0.0	-108578.0	-28322.0	-101700.5	-203522.0	-36720.5
Edinburg	-65522.0	-108578.0	0.0	-92880.5	-1104.5	-16200.0	-26220.5
Exetburg	-23544.5	-28322.0	-92880.5	0.0	-86112.5	-177012.5	-27848.0
Glasgow	-56448.0	-101700.5	-1104.5	-86112.5	0.0	-18050.0	-22898.0
Inverness	-132612.5	-203522.0	-16200.0	-177012.5	-18050.0	0.0	-77224.5
Liverpool	-9522.0	-36720.5	-26220.5	-27848.0	-22898.0	-77224.5	0.0

$$B = -\frac{1}{2}\left(I - \frac{1}{n}U\right)D^2\left(I - \frac{1}{n}U\right)$$

Il bicentrameto della matrice delle distanze ne riduce il rango di una untà dato che uno degli autovalori è ora forzatamente zero

Si ritiene che la componente così eliminata sia un fattore ignorabile legato al livello generale della distanza o dissimilarità

Applicazione: coordinate principali

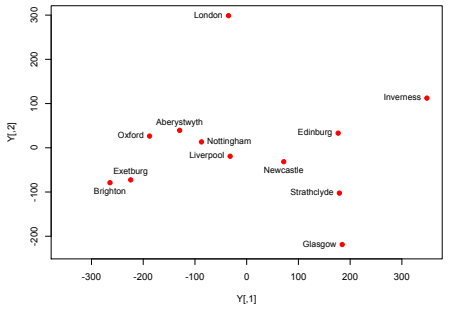
$$\tilde{B} = VLV^t = \left(VL^{0.5}\right)\left(VL^{0.5}\right)^t = \tilde{X}\tilde{X}^t$$

$$\tilde{X} = VL^{0.5}$$

L=(405976.84 176539.17 122412.80 55433.67
38713.33 31805.95 27532.10 14364.26
6035.29 5668.08 101.35 0.00)

Compito semplice: da una mappa si risale alle distanze tra i punti.

Compito dello scaling: dalle distanze tra i punti si risale alla mappa



	[,1]	[,2]
[1,]	-129.60	39.40
[2,]	-264.28	-78.94
[3,]	176.82	33.13
[4,]	-224.25	-72.58
[5,]	184.59	-218.74
[6,]	348.30	112.31
[7,]	-32.01	-19.14
[8,]	-35.15	298.67
[9,]	71.59	-31.52
[10,]	-87.39	13.31
[11,]	-187.80	26.51
[12,]	179.18	-102.40

Effetto Guttman

Consideriamo la matrice (n x n) definita come

$$Exponential: D = uu^t - A, \quad dove \; a_{ij} = e^{0.5|i-j|}$$

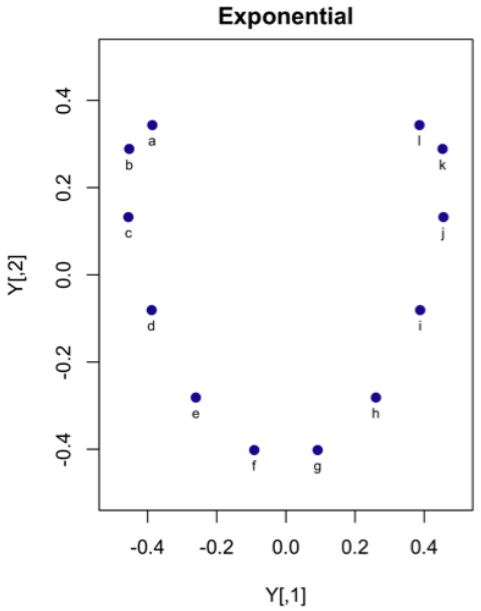
Poiché sia la matrice uu^t che A sono simmetriche lo sarà anche D; inoltre, gli elementi sulla diagonale di D sono pari a zero e gli elementi fuori diagonale sono positivi ed inferiori ad uno.

	a	b	c	d	e
a	0.0000	0.3935	0.6321	0.7769	0.8647
b	0.3935	0.0000	0.3935	0.6321	0.7769
c	0.6321	0.3935	0.0000	0.3935	0.6321
d	0.7769	0.6321	0.3935	0.0000	0.3935
e	0.8647	0.7769	0.6321	0.3935	0.0000

Da notare che gli elementi sulle righe aumentano allontanandosi in termini di posizioni (diriga o di colonn).

Le distanze più grandi confondono e oscurano quelle più piccole.

Effetto Guttman/2



Questa configurazione nota anche come ferro-di-cavallo (horseshoe) è stata rilevata molto spesso e ne sono state date diverse interpretazioni.

Effetto Guttman/3

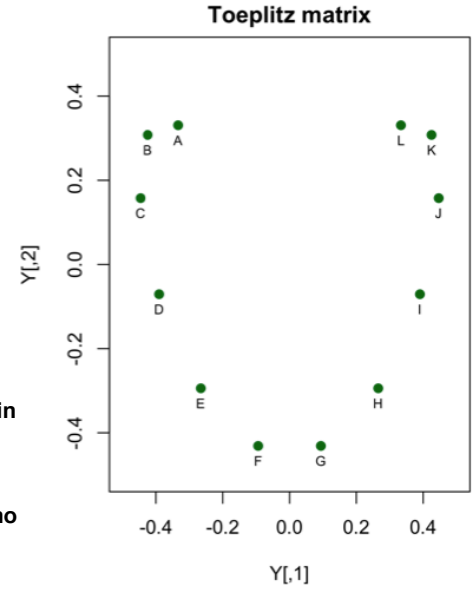
Matrice di Toeplitz

$$D = uu^t - A, \quad a_{ij} = 0.5^{|i-j|}$$

	A	B	C	D	E
A	0.0000	0.500	0.75	0.875	0.9375
B	0.5000	0.000	0.50	0.750	0.8750
C	0.7500	0.500	0.00	0.500	0.7500
D	0.8750	0.750	0.50	0.000	0.5000
E	0.9375	0.875	0.75	0.500	0.0000

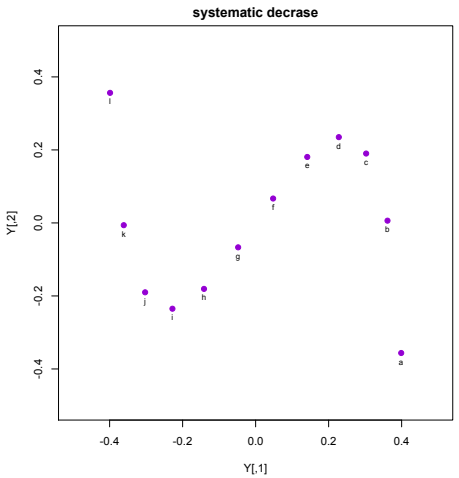
Da notare che gli elementi sulle righe aumentano allontanandosi in termini di posizioni (di riga o di colonna).

Le distanze più grandi confondono e oscurano quelle più piccole.



Effetto Sigma

Un effetto meno noto è quello sigma che si realizza allorché la matrice delle dissimilarità è strutturata in modo che unità prossime nella matrice dei dati sono in realtà distanti nei valori



$$decrease \; D = uu^t - (I - A), \quad a_{ij} = \frac{(i-j)^2}{n^2}$$

	a	b	c	d	e	f
a	0.0000	0.9931	0.9722	0.9375	0.8889	0.8264
b	0.9931	0.0000	0.9931	0.9722	0.9375	0.8889
c	0.9722	0.9931	0.0000	0.9931	0.9722	0.9375
d	0.9375	0.9722	0.9931	0.0000	0.9931	0.9722
e	0.8889	0.9375	0.9722	0.9931	0.0000	0.9931
f	0.8264	0.8889	0.9375	0.9722	0.9931	0.0000

Ricostruzione delle distanze

Il numero di coordinate che riusciamo ad utilizzare per la visualizzazione dei risultati è 2 o al massimo 3.

Se il numero di autovalori non nulli di B è maggiore del numero delle coordinate che si adoperano allora le distanze ricostruite non possono essere esatte, ma solo approssimate.

In generale, la matrice dei prodotti incrociati approssimata è data da

$$\tilde{B} = \tilde{X}\tilde{X}^t \quad dove \quad \tilde{X} = VL^{0.5}$$

La corrispondente matrice delle distanze al quadrato (approssimata) è

$$\tilde{D}^2 = \tilde{b}u^t + u\tilde{b}^t - 2\tilde{B}$$

Dove $\tilde{b}$ è un vettore formato con la diagonale di $\tilde{B}$

Euclideità

La riuscita della conversione tra distanze e coordinate è legata alla euclideità della matrice D delle distanze.

Tale condizione è verificata se per ogni terna si ha

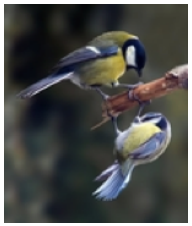
$$\binom{n}{3} = \frac{n(n-1)(n-2)}{6}$$

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

Se la matrice delle distanze D è euclidea allora la corrispondente matrice dei prodotti incrociati è semi-definita positiva e le coordinate principali possono essere ottenute nel modo indicato

Ovvero, se H è semi-definita positiva i punti giacciono in un spazio euclideo e le dissimilarità possono considerarsi in tutto e per tutto delle distanze.

Esempio sulle cinciallegre



La nostra base informativa è costituita solo dalla matrice di distanza (che potrebbe essere una sintesi di altre matrici di distanza)

Percentuale di volte che le cincie sono state viste assieme alla pastura

Ficken et al. *Behav. Ecol. Sociobiol.* 1981

SCAO	1.00						
AOPR	0.18	1.00					
ARPO	0.07	0.27	1.00				
YOSA	0.26	0.12	0.12	1.00			
ROAY	0.21	0.19	0.18	0.31	1.00		
SORA	0.06	0.02	0.03	0.15	0.04	1.00	
BJAO	0.19	0.17	0.09	0.16	0.21	0.28	1.00
	SCAO	AOPR	ARPO	YOSA	ROAY	SORA	BJAO

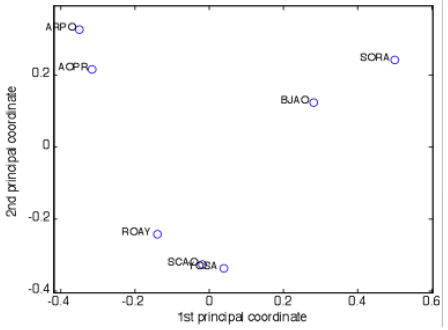
Lo scopo dello scaling multidimensiionale è di sfruttare una analogia tra il concetto di prossimità tra oggetti (sia concreti che astratti) e la distanza tra punti nello spazio.

Esempio (continua)

La percentuale di variabilità delle pseudo-coordinate misura la qualità della rappresentazione

Proporzioni trasformate in distanze $D=\sqrt{(1-X)}$

SCAO	0.00						
AOPR	0.91	0.00					
ARPO	0.96	0.85	0.00				
YOSA	0.86	0.94	0.94	0.00			
ROAY	0.89	0.90	0.91	0.83	0.00		
SORA	0.97	0.99	0.98	0.92	0.98	0.00	
BJAO	0.90	0.91	0.95	0.92	0.89	0.85	0.00
	SCAO	AOPR	ARPO	YOSA	ROAY	SORA	BJAO



Prin Coord	% explained	Cumulative	Eigenvalue
1	22.77	22.77	0.575
2	20.05	42.82	0.507
3	16.63	59.45	0.420
4	15.17	74.62	0.383
5	13.37	87.98	0.338
6	12.02	100.00	0.304

Gli autovalori sono tutti positivi o nulli e quindi B è positiva semi-definita e D è una matrice euclidea.

Il numero di coordinate principali

Lo scaling metrico consente di ottenere una rappresentazione geometrica di punti le cui distanze approssimano le dissomiglianze osservate.

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 corrisponderebbero all'espressione

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

Gli autovalori nulli o quasi-zero sono tali da rendere irrilevante il contributo delle coordinate a cui sono associati. Possiamo senzaltro escluderli.

Poiché per le rappresentazioni grafiche disponiamo di due o al massimo tre dimensioni, conserveremo solo i primi (1, 2, o 3)

P=2 oppure p=3

l'adeguatezza di rappresentazione sarà misurata da $\sum_{i=1}^{p} \lambda_i \Big/ \sum_{i=1}^{n} |\lambda_i|$

Esempio

Differenziazioni migrogeografiche per gli zibellini

Mahalanobis distances among 9 local populations, based on 10 age-adjusted linear measurements of the skulls. Total: 144 individuals

Confirmation of the Euclidean nature of a distance matrix by the Gower's theorem

is.euclid(distmat, plot = FALSE, print = FALSE, tol = 1e-07)

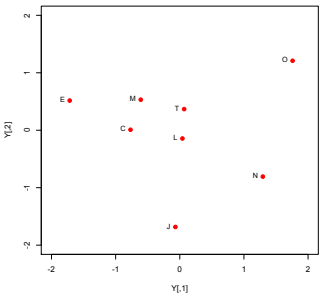
8.69 5.65 2.76 1.86 0.76 0.64 0.04 0.00

%Var= 70.3%

C	0.0000							
E	2.1380	0.0000						
J	2.2713	2.9579	0.0000					
L	1.7135	2.3927	1.7772	0.0000				
M	1.5460	1.9818	2.4575	1.0125	0.0000			
N	2.6979	3.3566	1.9900	1.8520	2.6954	0.0000		
O	2.9985	3.6848	3.4484	2.4272	2.6816	2.3108	0.0000	
T	2.3859	2.3169	2.4666	1.4545	1.7581	2.2105	2.5041	0.0000

	[,1]	[,2]
[1,]	-0.77	0.01
[2,]	-1.72	0.52
[3,]	-0.07	-1.68
[4,]	0.04	-0.14
[5,]	-0.61	0.53
[6,]	1.29	-0.81
[7,]	1.76	1.21
[8,]	0.07	0.37

La sintesi grafica mostra un isolamento della popolazione in "O" ed in "J"

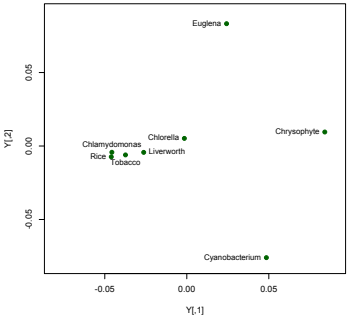


Altro esempio

Evoluzione organismi fotosintetici

LogDet distances among 8 species of photosynthetic organisms, computed from 920 bases from the 16S rRNA of the chloroplasts

1. *Tobacco*	0.0000							
2. *Rice*	0.0258	0.0000						
3. *Liverworth*	0.0248	0.0357	0.0000					
4. *Chlamydomonas*	0.1124	0.1215	0.1014	0.0000				
5. *Chlorella*	0.0713	0.0804	0.0604	0.0920	0.0000			
6. *Euglena*	0.1270	0.1361	0.1161	0.1506	0.1033	0.0000		
7. *Cyanobacterium*	0.1299	0.1390	0.1190	0.1535	0.1128	0.1611	0.0000	
8. *Chrysophyte*	0.1370	0.1461	0.1261	0.1606	0.1133	0.1442	0.1427	0.0000



La matrice delle distanze passa il test di euclidità

0.0163 0.0129 0.0096 0.0087 0.0014
0.0002 0.0001 0.0000

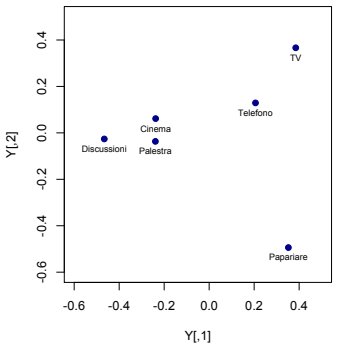
Tre degli organismi sono remoti rispetto agli altri ed ognuo lo è in modo diverso

Applicazione

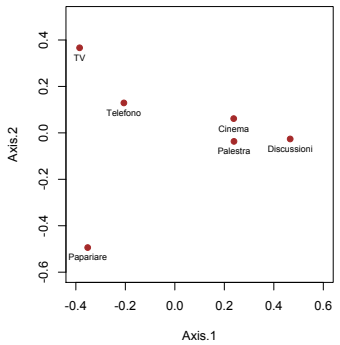
Preferenze tempo libero

	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

Home made



Pcoa in ade4



Trasformazioni delle distanze

Se la matrice D=(d_{ij}) non è euclidea lo scaling metrico deve essere preceduto da una trasformazione delle dissimilarità che la renda euclidea.

AD ESEMPIO

$$\begin{cases} d_{ij}^{*} = \left(\phi_1 + d_{ij}\right)^{\phi_2} \\ d_{ij}^{*} = \left(d_{ij}\right)^{\phi_2} + \phi_1 \end{cases}, \qquad \phi_1, \phi_2 > 0$$

Il problema consiste nella determinazione ottimale dei parametri

Queste trasformazioni non sono necessarie se le dissimilarità provengono da distanze di Minkowski o di Mahalanobis.

Considerazioni

La somma di una costante ϕ_1 preserva l'ordinamento delle distanze: quelle che erano più piccole rimangono più piccole e le più grandi rimangono tali.

Tuttavia si riduce la disuguaglianza tra le distanze dato che

$$\frac{\max\left\{d_{ij}\right\}}{\min\left\{d_{ij}\right\}} > \frac{\max\left\{d_{ij}+\phi_1\right\}}{\min\left\{d_{ij}+\phi_1\right\}} \to 1 \quad se \quad \phi_1 \to \infty$$

Quindi l'attenzione si sposta verso il centro della distribuzione delle distanze più che verso la loro coda.

Un esponente ϕ_2<1 pure lascia inalterato l'ordinamento delle distanze, ma dilata le distanze minori più che le distanze maggiori. Il contrario avviene con ϕ_2>1.

La trasformazione $\left(d_{ij}+\phi_1\right)^{\phi_2}$ è preferibile alla $\left(d_{ij}\right)^{\phi_2}+\phi_1$ perché in presenza di distanze molto piccole, la trasformazione potrebbe attenuare diversità essenziali nella valutazione delle distanze.

Il problema della costante additiva

Se le d_{ij} non formano una matrice euclidea allora è euclidea la matrice

$$d_{ij}^{+} = d_{ij} + \phi, \quad con \quad \phi \geq \max_{i,j,k} \left| d_{ij} - d_{ik} - d_{jk} \right|$$

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.

Attenzione! Aggiungere la stessa costante può avere effetti diversi in base al valore a cui va a sommarsi.

Se D è una matrice di dissimilarità e c una costante additiva allora,

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

All'aumentare di c aumenta l'uguaglianza tra gli autovalori (tendono tutti a zero tranne il primo che tende a c) e questo riduce la capacità descrittiva delle coordinate principali.

Condizione di Lingoes

Dobbiamo ricordare che in alcuni contesti si usano metriche piuttosto diverse e pertanto la condizione di euclideità richiede un supporto.

Se D è una matrice di dissimilarità on euclidea allora esiste una costante c tale che la nuova matrice con elemento generico

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

sia euclidea

Qui λ_n è l'autovalore minore della matrice dei prodotti incrociati basata sulle dissimilarità

$$\tilde{B} = -\frac{1}{2}\left(I - \frac{1}{n}U\right)D^{2}\left(I - \frac{1}{n}U\right)$$

Condizione di Cailliez

Se D è una matrice di dissimilarità allora esiste una costante c tale che la matrice

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

sia euclidea

Qui λ_1 è 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 la correzione è necessario calcolare gli autovalori delle matrici coinvolte nella formula per ottenere c.

Tale calcolo può essere problematico se n è molto grande.

Elegant (De Leeuw, 1975)

Supponiamo che la matrice B ricavata dalle dissomiglianze NON sia semi-definita positiva (ovvero la matrice delle dissoliglianze che la genera NON è euclidea)

Se gli autovalori negativi sono piccoli in valori assoluto si potrebbe ignorarli sommando il valore assoluto dell'autovalore negativo minore (metodo quasi-euclideo), ma esistono altre possibilità quali Lingoes o Caillez.

Il problema è che c'è un modo semplice di determinare la correzione più appropriata, se mai ne esistesse alcuna.

In genere le matrici di dissimilarità sono positive, simmetriche e con diagonale nulla.

Un approccio alternativo alla costante additiva è la ricerca della matrice euclidea più prossima alla matrice delle dissimilarità.

$D =$

	U_1	U_2	U_i	U_n
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

Elegant/2

In base alla matrice delle dissimilarità e con l'uso del doppio centramento definiamo

$$\tilde{B} = -\left(\frac{1}{2}\right) C D^{2} C$$

La scomposizione in valori singolari produce le variabili approssimate dove p è il numero di autovalori maggiori di zero disposti in ordine decrescente sulla diagonale di L_p e V_p è la matrice (nxp) degi autovettori.

$$\tilde{X} = V_p L_p^{0.5}$$

Lo schema ricorsivo dell'algoritmo elegant è il seguente

Sugg. $\alpha=2/p^2$

$$D^{(k+1)} = \alpha\left[D^{(k)} - S_0 - S^{(k)}\right] + (1-\alpha)\tilde{X}^{(k)}\left[\tilde{X}^{(k)}\right]^{t}, \quad 0 < \alpha < 1$$

$$D^{(0)} = D, \quad S_0 = diag\left(\sum_{j=1}^{n} d_{1j}, \sum_{j=1}^{n} d_{2j}, \cdots, \sum_{j=1}^{n} d_{nj}\right)$$

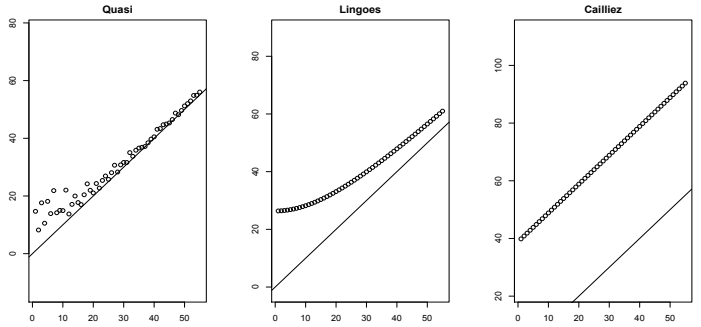
$$\tilde{X}^{(k)} = V_p^{(k)}\left[L_p^{(k)}\right]^{0.5}, \quad S^{(k)} = diag\left(\sum_{j=1}^{n} d_{1j}^{(k)}, \sum_{j=1}^{n} d_{2j}^{(k)}, \cdots, \sum_{j=1}^{n} d_{nj}^{(k)}\right)$$

Lo schema è esigente per i calcoli e quindi molto lento.

Tende a creare outliers

Esempio

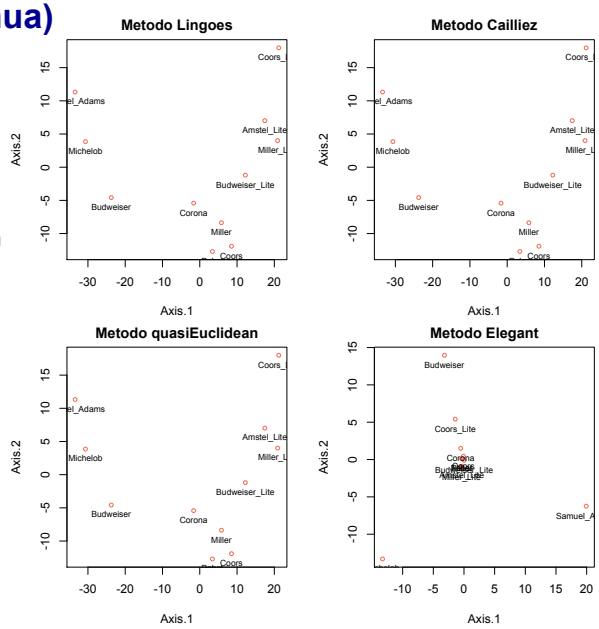
Preferenze per la marca di birre. Il comando kdist di ade4 fornisce tre possibili scelte per la correzione.



In questo caso è suggerito il metodo quasi perché segue meglio la tendenza degli autovalori.

Il metodo quasi è basto sul principio dell'algoritmo Elegant cioè cerca una matrice euclidea prossima a quella originale, ma si realizza ponendo pari a zero gli autovalori negativi.

Esempio (continua)



Le correzioni Lingoes, Caillez e quasiEuclidean sono simili.

La soluzione elegant identifica dei valori remoti ed un centro molto compatto.

Esempio

Prossimità del profilo immunitatio di alcune specie

Il metodo di correzione, come prevedibile, porta a risultati simili

Applicazione: piccolo teatro

Visuale	0								
Acustica	0.133	0.000							
Comodità_poltrone	0.351	0.353	0.000						
Varietà_spettacoli	0.239	0.235	0.267	0.000					
Orario_programmazione	0.304	0.328	0.297	0.352	0.000				
Qualità_spettacoli	0.183	0.155	0.339	0.283	0.287	0.000			
Prezzo	0.256	0.250	0.274	0.284	0.266	0.325	0.000		
Accesibilità/parcheggio	0.371	0.382	0.250	0.412	0.293	0.454	0.350	0.000	
Servizi_Aggiuntivi	0.412	0.427	0.300	0.441	0.344	0.476	0.432	0.267	0.000

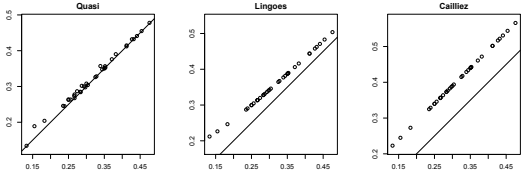
L'indice di affintà è dato dalla percentuale di volte che due item sono stati giudicati della stessa importanza in un differenziale semantico a 4 livelli: per nulla, poco, abbastanza, molto importante.

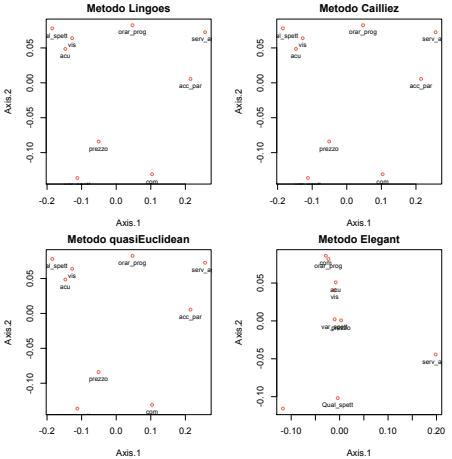
n=306

La dissimilarità è data dal complemento ad uno dell'affinità.

La matrice così ottenuta non è euclidea perché la corrispondente H ha un autovalore negativo λ_g=-0.0136

Piccolo teatro/2

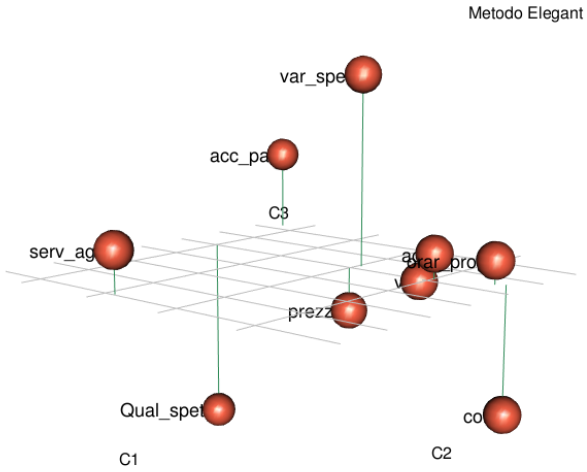




Nella lettura di questi grafici è bene privilegiare le coppie di item vicini in quanto percepiti come sostituti.

N.B. I valori remoti appaiono remoti a due dimensioni, ma potrebbero non esserlo se si usano tre pseudo-variabili.

Piccolo teatro/3



Applicazione

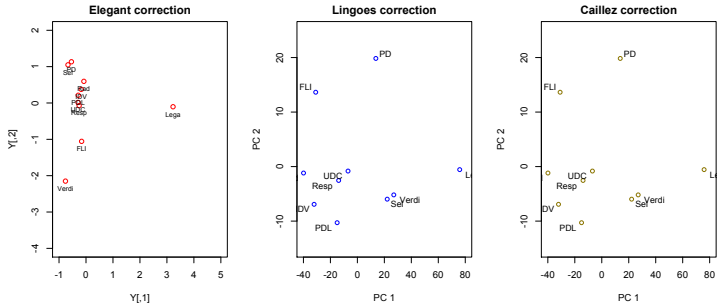
Percezione di alcune formazioni politiche da parte di un campione di elettori ed elettrici. La distanza si basa sulla media dei punteggi nei confronti a coppia

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.

1] 1.120636e+04 8.049433e+02 3.663228e+02 6.460404e+01
[5] 9.952040e+00 -1.078259e-12 -5.344493e+01 -2.440399e+02
[9] -5.895150e+02 -2.614678e+03

Per rappresentare I partiti in uno spazio metrico occorre rendere euclidea la matrice delle distanze.



Indeterminatezza della soluzione

Sia Q una matrice quadrata e simmetrica di ordine p ortogonale ($QQ^t=I$). In altre parole Q è una matrice di rotazione.

Se la soluzione ottenuta con lo scaling metrico $\tilde{X} = VL^{0.5}$

È ruotata a mezzo della matrice Q si ottiene $\tilde{X}^* = \tilde{X}Q^t$

Le distanze non cambiano

$$\tilde{X}^*\left(\tilde{X}^*\right)^t = \left(\tilde{X}Q^t\right)\left(\tilde{X}Q^t\right)^t = \tilde{X}Q^tQ\tilde{X}^t = \tilde{X}\tilde{X}^t = \tilde{B}$$

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

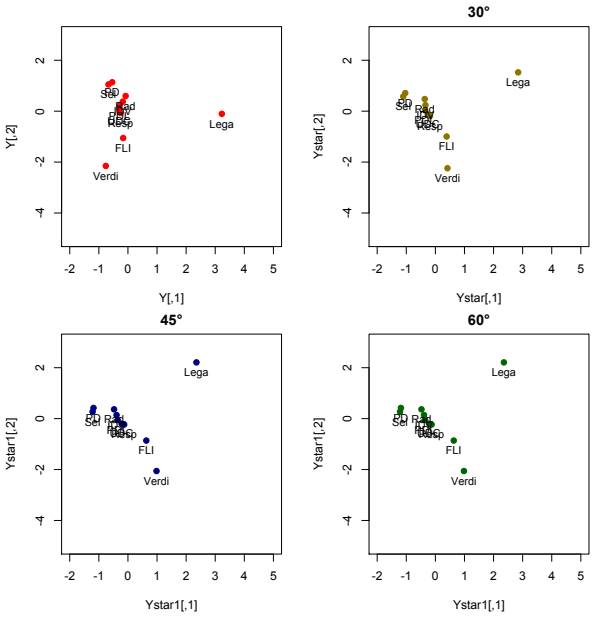
La distanze non cambiano, ma la configurazione dei punti potrebbe essere diversa.

Esempio

Rotazione della soluzione elegant

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.

Esiste una rotazione ottimale? C'è un modo per calcolarla?



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.

Tuttavia, la nostra capacità di trattare matrici asimmetriche è ancora scarsa (ad esempio gli autovalori potrebbero essere complessi)

Quindi tenteremo di trasformare una matrice asimmetrica in una simmetrica sperando di conservarne il contenuto informativo.

Scaling di matrici asimmetriche/2

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

Allora è simmetrica la matrice

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

In pratica, G* si ottiene realizzando la media potenziata di ordine α tra gli elementi corrispondenti delle matrici G e G^t (la sua trasposta).

Tre casi particolari interessanti sono

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

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

Scaling di matrici asimmetriche/3

Se le misurazioni riguardano delle affinità tra unità allora l'elemento sulla diagonale dovrebbe essere il maggiore tra tutti gli elementi sulla riga:

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

In questo caso basterà dividere ogni riga e colonna di G* per l'elemento sulla diagonale per ottenere una matrice con degli uno sulla diagonale e con elementi inferiori all'unità fuori diagonale.

Se invece la G* non ha elementi sulla diagonale maggiori degli altri possono insorgere problemi sulla sua considerazione come matrice di affinità.

la matrice simmetrizzata con diagonale unitaria è $\mathbf{G}^* = \left[\frac{(\mathbf{GL})^{\alpha} + \left(\mathbf{LG}^t\right)^{\alpha}}{2} \right]^{\frac{1}{\alpha}}$

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

La corrispondente matrice di dissomiglianze sarà $\mathbf{D} = \left(\mathbf{u}\mathbf{u}^t - \mathbf{G}^* \right)$

Scaling di matrici asimmetriche/4

Se la matrice G include misurazioni relative alla dissomiglianza occorre azzerare gli elementi sulla diagonale sottraendo l'elemento sulla diagonale da quelli situati sulla stessa riga e colonna (ai fini della simmetria)

La positività della matrice sarà preservata se $g_{ii} = \min_{1 \le j \le n} g_{ij}$

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

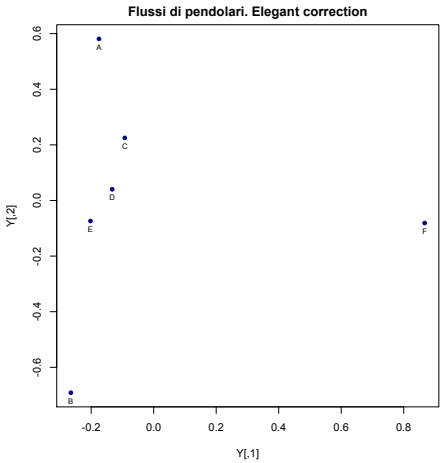
$$\mathbf{D} = \mathbf{G}^* - diag\left(1/g^*_{11}, \;\; 1/g^*_{22}, \ldots, 1/g^*_{nn}\right)\mathbf{u}\mathbf{u}^t$$

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

Esempio

Flusso tra località. Grado di prossimità in termini di numero di pendolari

	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 diagonale.

La simmetrizzazione si realizza, ad esempio, con la media geometrica e posi si trasformano poi le prossimità in distanze.