Capitolo IV

DISTANZE ED INDICI DI SIMILARITÀ

> *« Localiser un objet, cela veut dire simplement se représenter les mouvements qu'il faudrait faire pour l'atteindre. Mais, étant donné un objet, on peut concevoir plusieurs séries différentes de mouvements qui permettent également de l'atteindre »*
>
> (Jules-Henry Poincaré, *La valeur de la science*).

1. *Introduzione*

Nei capitoli precedenti abbiamo esaminato le relazioni tra variabili (che corrispondono ai vettori colonna nella matrice dei dati) ed i metodi che, partendo dalla matrice di covarianza o dalla matrice di correlazione, si proponevano di conseguire una riduzione delle dimensioni dello spazio R^p.

Passiamo ora ad illustrare la « prossimità » tra unità statistiche, alle quali corrispondono vettori riga nella matrice dei dati (1). Con il termine generico di « prossimità » intendiamo riferirci sia al concetto di « rassomiglianza » tra le unità, sia a quello antitetico di « diversità », essendo chiaro che è equivalente affermare che due unità sono molto simili, ovvero poco diverse.

Un indice di prossimità tra due generiche unità statistiche u_i e u_j è definito come una funzione dei rispettivi vettori riga nella matrice dei dati:

---

(1) In psicometria l'analisi delle relazioni tra le variabili viene talvolta indicata col termine di « tecnica R », mentre lo studio della prossimità tra gli elementi viene denominato « tecnica Q ».

$$IP_{ij} = f(\mathbf{x}'_i, \mathbf{x}'_j) \qquad i,j = 1,2,...,n \tag{4.1}$$

Si supponga, ad esempio, di aver rilevato per n tipi di autovetture p variabili (prezzo, cilindrata, velocità massima, ecc.). La conoscenza degli indici di prossimità per ciascuna delle possibili coppie di autovetture consente di individuare quelle tra loro più simili (meno diverse). Un secondo esempio è fornito dai comuni d'una regione, per i quali si possono costruire molti indicatori demografici, economici e sociali, in base ai dati del censimento della popolazione e di quello dell'industria e dei servizi: gli indici di prossimità tra coppie di comuni permettono di « misurare » la diversità tra gli stessi, sotto gli aspetti di volta in volta considerati.

In una matrice dei dati, com'è noto, possono considerarsi anche caratteri qualitativi, le cui modalità sono tradotte da opportuni codici numerici. Un esempio è fornito dalla matrice dei dati ottenuta dallo spoglio di n questionari d'una inchiesta che contempli p domande con risposte chiuse di tipo qualitativo (ad esempio, la domanda sulla professione dell'individuo, con categorie prefissate di risposta) ed eventualmente anche risposte di tipo quantitativo (ad esempio, l'ammontare della spesa mensile). Per due generici rispondenti all'inchiesta può essere interessante valutare in che misura le relative « batterie » di risposte si rassomigliano, ovvero differiscono.

Le informazioni fornite dagli indici di prossimità tra coppie d'elementi costituiscono una premessa per l'individuazione di gruppi di unità omogenee (in senso relativo). Nel primo esempio sopra citato si identificano gruppi di autovetture tra loro simili, che possono rappresentare i « segmenti » del mercato automobilistico; nel secondo caso si riconoscono aree omogenee, formate da insiemi di comuni, che presentano nel loro interno valori simili per le variabili utilizzate nella classificazione; nel terzo caso si costruiscono tipologie di rispondenti all'inchiesta, con riferimento ai sottoinsiemi d'individui che hanno fornito risposte analoghe.

La formazione di gruppi omogenei di unità — che sarà oggetto del capitolo successivo — può interpretarsi come una sorta di « riduzione delle dimensioni » dello spazio R^n, poiché si riuniscono le n unità in g sottoinsiemi (e solitamente $g \ll n$), che dovrebbero corrispondere a categorie, o classi, delle stesse.

Gli indici di prossimità vengono abitualmente distinti a seconda che essi si applichino a fenomeni quantitativi oppure qualitativi. Con riferimento al primo caso considereremo le distanze, gli indici di di-

stanza e gli indici di dissimilarità, che corrispondono a famiglie via via più ampie di indici, in cui ciascuna classe comprende le precedenti; per i caratteri qualitativi illustreremo gli indici di similarità, ponendo in luce anche le relazioni tra questi e le distanze. Esamineremo poi il caso misto, in cui compaiono sia fenomeni quantitativi sia caratteri qualitativi.

2. *Definizione di distanza*

In statistica il concetto di distanza è mutuato dalla geometria, ove si fa riferimento alla distanza tra due punti in R^p.

Definizione. Si dice distanza (o metrica) tra due punti corrispondenti ai vettori $\mathbf{x}, \mathbf{y} \in \mathrm{R}^p$ una funzione $d(\mathbf{x}, \mathbf{y})$ che gode delle seguenti proprietà:

1) *non negatività*:

$$d(\mathbf{x}, \mathbf{y}) \geq 0 \qquad \forall \quad \mathbf{x}, \mathbf{y} \in R^p$$

2) *identità*:

$$d(\mathbf{x}, \mathbf{y}) = 0 \qquad \Leftrightarrow \quad \mathbf{x} = \mathbf{y}$$

3) *simmetria*:

$$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}) \qquad \forall \quad \mathbf{x}, \mathbf{y} \in R^p$$

4) *disuguaglianza triangolare*:

$$d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z}) \qquad \forall \quad \mathbf{x}, \mathbf{y}, \mathbf{z} \in R^p$$

Uno spazio con riferimento al quale si sia definita una distanza è detto *spazio metrico.*

Nel caso d'una matrice dei dati con variabili tutte quantitative, la distanza tra due generiche unità statistiche corrisponde ad una classe particolare, ma molto importante, degli indici di prossimità introdotti nel paragrafo precedente; essa è calcolata sui vettori riga $\mathbf{x}'_i$ e $\mathbf{x}'_j$ e sarà indicata nella maniera seguente:

$$d(\mathbf{x}'_i, \mathbf{x}'_j) = d_{ij} \tag{4.2}$$

Osservazione. Una funzione per poter essere qualificata in generale come distanza deve godere delle quattro proprietà di non negatività, identità, simmetria e disuguaglianza triangolare per tutti i vettori in R^p. Non sarebbe quindi corretto introdurre la definizione di distanza in termini degli indici d_{ij} calcolabili in una matrice dei dati, riguardante n unità statistiche e p variabili. Può accadere, infatti, che la funzione prescelta soddisfi le quattro proprietà con riferimento ad una certa matrice dei dati, ma non in generale.

Le quattro proprietà che definiscono la distanza non sono tra loro indipendenti: si può dimostrare che se valgono le proprietà di identità e di disuguaglianza triangolare sussistono anche le proprietà di non negatività e di simmetria (Leti, 1979a, pp. 11-12; Landenna, 1994, p. 288).

3. *Alcuni tipi di distanza*

Il tipo più noto di distanza tra due punti è la distanza euclidea, su cui si fonda la geometria consueta (2). Con riferimento ad una matrice dei dati, con n unità statistiche e p variabili, essa risponde alla seguente:

Definizione. Si dice distanza euclidea tra due unità statistiche i e j la norma (euclidea) della differenza tra i rispettivi vettori:

$$_2d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\| = \left[\sum_{s=1}^{p}(x_{is} - x_{js})^2\right]^{1/2} \tag{4.3}$$

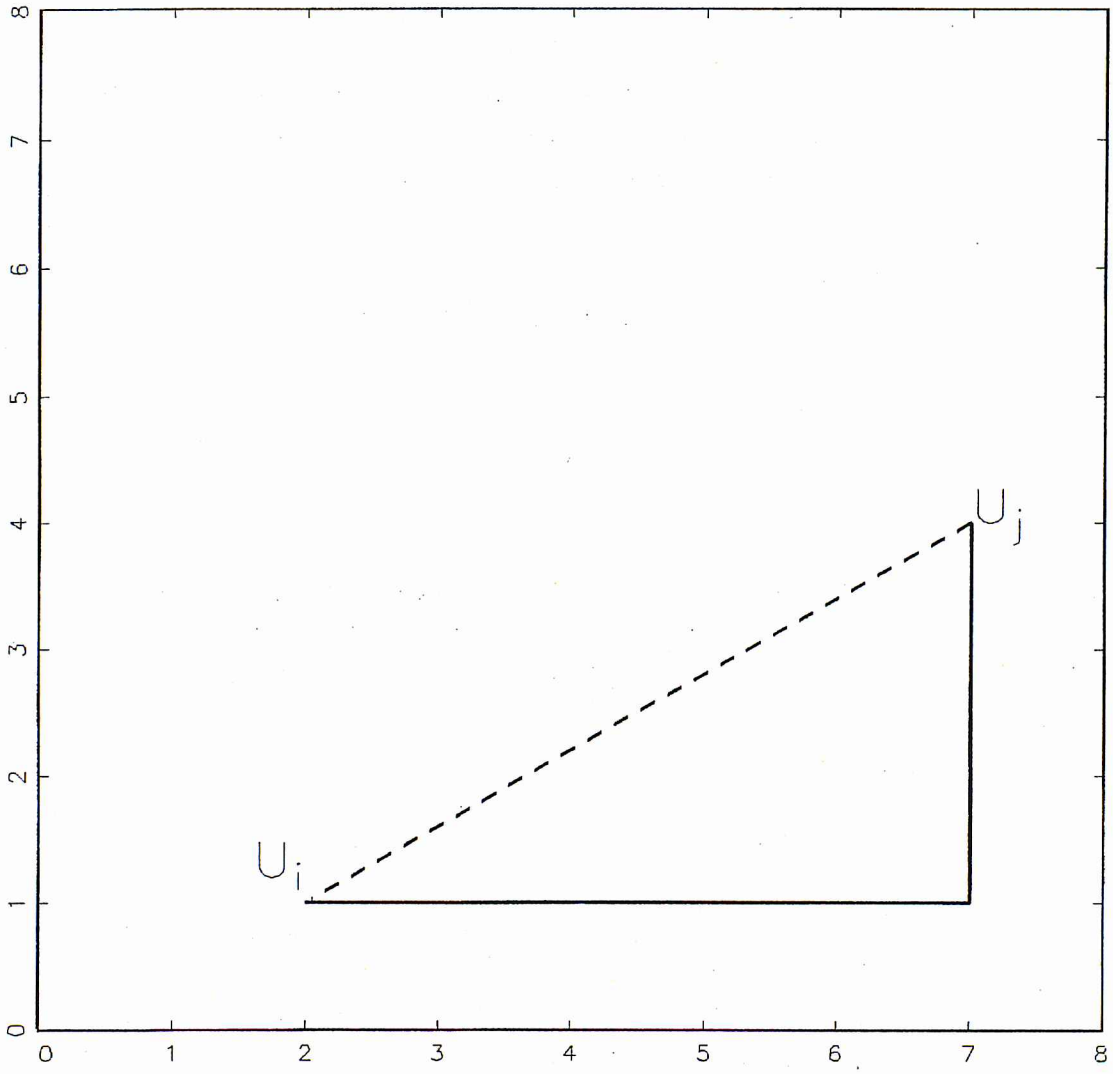
Nel caso particolare di due sole variabili, X_1 e X_2, è possibile rappresentare nel piano cartesiano i punti corrispondenti alle unità statistiche. La distanza euclidea tra due punti è allora uguale alla lunghezza del segmento che li unisce, com'è posto in evidenza nella fig. 4.1, in cui sono rappresentati i punti di coordinate $(2, 1)$ e $(7, 4)$.

Siano $\mathbf{x}_i = [x_{i1}, x_{i2}]'$ e $\mathbf{x}_j = [x_{j1}, x_{j2}]'$ i vettori corrispondenti a due generiche unità statistiche. La distanza euclidea tra essi è:

$$_2d_{ij} = \left[(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2\right]^{1/2} \tag{4.4}$$

(2) Trattazioni molto approfondite dei vari tipi di distanza sono state proposte da Leti (1979b), Arabie (1991), Critchley *et al.* (1992).

FIG. 4.1. *Tipi di distanza tra due punti u_i e u_j di coordinate (2, 1) e (7, 4).*



- - - distanza euclidea; — distanza della città a blocchi.

L'espressione precedente è la radice quadrata della somma dei quadrati costruiti sui cateti del triangolo rettangolo che compare nella fig. 4.1, che, per il teorema di Pitagora, equivale all'ipotenusa del triangolo stesso.

La distanza euclidea è indubbiamente quella più conosciuta ed anche in statistica essa risulta di impiego molto frequente.

Un altro tipo di distanza di notevole interesse applicativo è fornito dalla seguente:

Definizione. Si dice *distanza della città a blocchi* tra due unità i e j l'espressione:

$$_1d_{ij} = \sum_{s=1}^{p} |x_{is} - x_{js}| \qquad (4.5)$$

Nella fig. 4.1 questa distanza corrisponde alla somma dei due cateti ed il nome le deriva proprio dal fatto che essa è la lunghezza che si deve percorrere per spostarsi da $\mathbf{x}_i$ a $\mathbf{x}_j$ qualora sia consentito muoversi solo nelle direzioni parallele agli assi, come avviene in una città con una griglia regolare di strade che s'intersecano ad angolo retto. Per tale motivo essa viene anche chiamata *distanza di Manhattan* o *metrica del taxi* (3).

I due tipi di distanza precedenti possono ottenersi entrambi da una formula più generale.

Definizione. Si dice distanza di Minkowski di ordine k tra le unità i e j l'espressione seguente:

$$ {}_kd_{ij} = \left[\sum_{s=1}^{p} |x_{is} - x_{js}|^k \right]^{1/k} \qquad k \geq 1 \tag{4.6} $$

Si ricava facilmente che la distanza euclidea può interpretarsi come la metrica di Minkowski per $k = 2$ e la distanza della città a blocchi per $k = 1$ e questo giustifica la simbologia ${}_1d_{ij}$ e ${}_2d_{ij}$ adottata in precedenza. Inoltre:

$$ \lim_{k \to \infty} {}_kd_{ij} = max_s|x_{is} - x_{js}| = {}_\infty d_{ij} \tag{4.7} $$

definisce un altro tipo di distanza, chiamata *distanza lagrangiana* o anche *distanza di Chebychev.*

Si noti che nel caso d'una sola variabile X la metrica di Minkowski per qualunque valore di k risulta:

$$ d_{ij} = |x_i - x_j| \tag{4.8} $$

cioè è uguale alla distanza tra due punti in R^1.

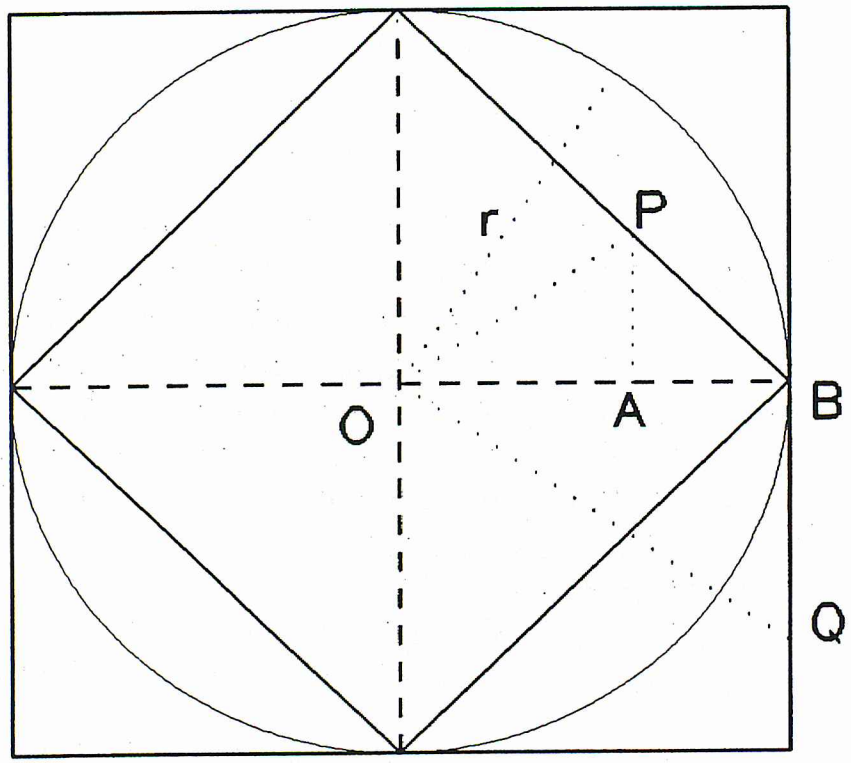
È interessante far notare che ciascuno dei tipi precedenti di distanza definisce una corrispondente geometria. In particolare, nel caso di due sole variabili X_1 e X_2, consideriamo il luogo dei punti equidistanti da un centro O (fig. 4.2). Adottando la distanza euclidea tale

---

(3) Si dice norma L_1 d'un vettore la somma dei valori assoluti delle sue componenti. La distanza della città a blocchi è dunque interpretabile come norma L_1 della differenza tra i vettori $\mathbf{x}_i$ e $\mathbf{x}_j$. L'impiego in statistica di questa norma — in alternativa a quella classica euclidea — ha formato oggetto di numerosi studi. Per un'ampia panoramica si veda Dodge, ed. (1987).

luogo è la consueta circonferenza, che assumeremo con raggio uguale a r.

FIG. 4.2. *Luogo dei punti distanti r dal centro: secondo la distanza euclidea esso è la circonferenza di raggio r; secondo la distanza della città a blocchi è il quadrato inscritto; secondo la distanza lagrangiana è il quadrato circoscritto.*



Con la distanza della città a blocchi il luogo dei punti equidistanti da 0 è il quadrato inscritto, poiché per ogni punto P del medesimo vale la seguente uguaglianza:

$$\overline{OA} + \overline{AP} = r$$

essendo:

$$\overline{AP} = \overline{AB}$$

Con la distanza lagrangiana il luogo dei punti che distano r da 0 è il quadrato circoscritto, poiché per qualunque punto Q del medesimo vale la seguente uguaglianza:

$$\overline{OQ} = \max(\overline{OB}, \overline{BQ}) = r$$

Una distanza che non rientra nella classe delle metriche di Minkowski è la cosiddetta «distanza di Canberra», introdotta da Lance e Williams (1966):

$$_cd_{ij} = \sum_{s=1}^{p} \frac{|x_{is} - x_{js}|}{(x_{is} + x_{js})} \tag{4.9}$$

L'espressione precedente — che può interpretarsi come una particolare traduzione in termini relativi della metrica di Manhattan — risulta una distanza se le variabili assumono solo valori positivi; in caso contrario può non essere soddisfatta la disuguaglianza triangolare (Gower and Legendre, 1986). Il vantaggio di questa metrica deriva dal fatto che in essa si considerano delle somme di rapporti che sono puri numeri, poiché il numeratore ed il denominatore sono espressi nella stessa unità di misura. Pertanto, essa può venire applicata direttamente anche per variabili differenti, senza ricorrere alle operazioni di standardizzazione che illustreremo in seguito. Inoltre la metrica di Canberra è poco sensibile alla asimmetria della distribuzione delle variabili ed alla presenza di *outliers*.

Con riferimento ad n unità statistiche, calcolando la distanza (d'un certo tipo) tra ciascuna delle possibili coppie d'elementi si ottiene una *matrice delle distanze*, di dimensioni $n \times n$, che indicheremo con $\mathbf{D}$:

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12} & \dots & d_{1n} \\ & 0 & \dots & d_{2n} \\ & & \ddots & \vdots \\ & & & 0 \end{bmatrix} \tag{4.10}$$

Per la definizione di distanza, tale matrice è simmetrica e presenta valori tutti nulli sulla diagonale principale. Inoltre, la matrice $\mathbf{D}$ è semidefinita positiva.

Esempio. Nella tab. 4.1 sono riportate le ore di trasmissione nel 1997 delle principali reti televisive italiane per tipo di programma (fonte: Istat, *Annuario statistico italiano*, 1998, pp. 213-214). Si vuole

valutare la diversità esistente tra le reti con riferimento alla tipologia di programmazione. A tale scopo calcoliamo la distanza tra le reti con riferimento solo alle prime quattro variabili (film, telefilm, varietà, news), escludendo l'ultima variabile, che, essendo residuale, può avere contenuti molto diversi per le varie reti.

TAB. 4.1. *Ore di trasmissione nel 1997 delle principali reti televisive per tipo di programma.*

	RETE	FILM	TELEFILM	VARIETÀ	NEWS	ALTRO
1	RAIUNO	1158	1280	1577	1703	3042
2	RAIDUE	731	1366	1280	1019	4364
3	RAITRE	1454	675	937	1618	4076
4	RETE4	2053	1289	489	1410	3519
5	CANALE5	582	1193	2166	3372	1447
6	ITALIA1	1167	3119	795	1261	2418

La distanza euclidea tra Rai Uno e Rai Due, ad esempio, si calcola nella maniera seguente:

$$_2d_{12} = \left[(1158 - 731)^2 + \ldots + (1703 - 1019)^2\right]^{1/2} = 863.59$$

La corrispondente distanza della città a blocchi risulta:

$$_1d_{12} = |1158 - 731| + \ldots + |1703 - 1019| = 1494$$

Per calcolare le matrici delle distanze fra tutte le possibili coppie di reti, sempre con riferimento alle sole prime tre variabili, utilizziamo il *package* SPSS, scegliendo la procedura: *correlazione - distanze*. Scegliendo dal menù il tipo di distanza nella casella: *misure*, si può ottenere dapprima la matrice delle distanze euclidee, riportata nella tab. 4.2 e quindi la matrice della distanze della città a blocchi (*city-block*), riportata nella tab. 4.3.

Da tali matrici si evince che le due reti con tipo di programmazione più simile (cioè con minore distanza) risultano, con entrambi i criteri, Rai Uno e Rai Due, mentre quelle più diverse sono Canale 5 e Italia 1.

TAB. 4.2. *Distanze euclidee tra le reti televisive, calcolate sui dati della tab. 4.1.*

	Distanza euclidea					
	1:RAIUNO	2:RAIDUE	3:RAITRE	4:RETE4	5:CANALE5	6:ITALIA1
1:RAIUNO		863.591	932.988	1438.992	1863.284	2046.678
2:RAIDUE	863.591		1215.179	1591.281	2524.626	1885.973
3:RAITRE	932.988	1215.179		989.831	2369.740	2490.606
4:RETE4	1438.992	1591.281	989.831		2972.344	2061.488
5:CANALE5	1863.284	2524.626	2369.740	2972.344		3222.990
6:ITALIA1	2046.678	1885.973	2490.606	2061.488	3222.990	

TAB. 4.3. *Distanze della città a blocchi tra le reti televisive, calcolate sui dati della tab. 4.1.*

	Distanza City Block					
	1:RAIUNO	2:RAIDUE	3:RAITRE	4:RETE4	5:CANALE5	6:ITALIA1
1:RAIUNO		1494.000	1626.000	2285.000	2921.000	3072.000
2:RAIDUE	1494.000		2356.000	2581.000	3561.000	2916.000
3:RAITRE	1626.000	2356.000		1869.000	4373.000	3230.000
4:RETE4	2285.000	2581.000	1869.000		5206.000	3171.000
5:CANALE5	2921.000	3561.000	4373.000	5206.000		5993.000
6:ITALIA1	3072.000	2916.000	3230.000	3171.000	5993.000	

3.1. *Confronto tra distanza euclidea e distanza della città a blocchi*

La distanza euclidea è influenzata più fortemente dalle differenze elevate tra i valori (poiché essa è funzione del quadrato delle stesse), mentre la distanza della città a blocchi attua una compensazione, su un piano paritetico, tra differenze «grandi» e differenze «piccole» (Rizzi, 1985, p. 52; Everitt, 1993, pp. 46-47).

Per far comprendere chiaramente questo discorso consideriamo un semplice esempio numerico.

Caso A:

$$\mathbf{x}_1 = [10,\ 5]'$$

$$\mathbf{x}_2 = [12,\ 7]'$$

La distanza della città a blocchi è uguale a 4 e la distanza euclidea a $\sqrt{8}$.

Caso B:

$$\mathbf{x}_1 = [10,\ 5]'$$

$$\mathbf{x}_2 = [11,\ 8]'$$

La distanza della città a blocchi è ancora uguale a 4, mentre la distanza euclidea risulta uguale a $\sqrt{10}$.

In conclusione, la distanza della città a blocchi è la stessa nei due casi poiché in essa due differenze uguali a 2 equivalgono ad una differenza uguale a 1 e l'altra uguale a 3. Questo non accade con la distanza euclidea, poiché, considerando i quadrati, le differenze più grandi incidono maggiormente e non vengono compensate dalle differenze più piccole.

Si osservi che questo effetto verrebbe ulteriormente ampliato se si considerasse il quadrato della distanza euclidea (che è un indice di distanza, come vedremo in seguito).

Il ricercatore deve quindi decidere quale tipo di distanza (o eventualmente di indice di distanza) può ritenersi più appropriato per il particolare problema in esame.

3.2. *Proprietà delle distanze di Minkowski*

La classe delle distanze di Minkowski gode di importanti proprietà, di cui ci limiteremo ad enunciare quelle che ci paiono di maggior interesse per l'analisi dei dati statistici.

Proprietà I. La metrica di Minkowski è funzione non crescente dell'indice k, per cui valgono le seguenti disuguaglianze:

$$_1d_{ij} \geq {}_2d_{ij} \geq \cdots \geq {}_\infty d_{ij}$$

L'uguaglianza si verifica solo nel caso banale di $p = 1$ (dati unidimensionali). Le relazioni precedenti si ricavano dalla disuguaglianza di Jensen:

$$\left(\sum_{s=1}^{p} x_s^k\right)^{1/k} \geq \left(\sum_{s=1}^{p} x_s^h\right)^{1/h}, \quad h > k > 0; \quad x_s \geq 0; \qquad (4.11)$$

La proprietà I fa sì che, partendo da un'assegnata matrice dei dati, la matrice delle distanze della città a blocchi tra le unità abbia elementi ordinatamente non minori della matrice della distanze euclidee.

Proprietà II: La distanza di Minkowski è invariante per traslazione delle variabili:

$$_kd(\mathbf{x}_i + \mathbf{c};\ \mathbf{x}_j + \mathbf{c}) = {}_kd(\mathbf{x}_i;\ \mathbf{x}_j) \tag{4.12}$$

ove: $\mathbf{c}$ è un vettore p-dimensionale di costanti (non necessariamente uguali tra loro).

Questa proprietà è molto importante in statistica, poiché consente di affermare, tra l'altro, che la distanza rimane invariata quando essa viene calcolata, anziché sui valori originari delle variabili, sui rispettivi scostamenti dalla media.

La metrica di Minkowski non è però invariante se si trasformano linearmente una o più variabili, cioè se si sostituiscono i valori originari con i seguenti:

$$a_s x_{is} + c_s, \qquad i = 1, 2, ..., n; \qquad s = 1, 2, ..., p.$$

Dal punto di vista logico questo rappresenta un grave inconveniente, poiché un semplice cambiamento dell'unità di misura d'una variabile, cioè una modificazione di scala (ad esempio, lunghezze espresse in centimetri oppure in metri; temperature misurate in gradi centigradi o in gradi Farenheit) altera le distanze tra le unità (4). Vedremo nel successivo n. 5.1 i criteri adottati per superare questa limitazione.

---

(4) Si osservi che la metrica di Minkowski non conserva neppure il verso delle disuguaglianze tra coppie di vettori. Infatti, se tra due coppie di unità esiste la seguente relazione tra le distanze:

$$d(\mathbf{x}_1, \mathbf{x}_2) > d(\mathbf{x}_3, \mathbf{x}_4)$$

ma almeno con riferimento ad una variabile (poniamo la s-esima) si manifesta la disuguaglianza opposta:

$$|x_{1s} - x_{2s}| < |x_{3s} - x_{4s}|$$

è sempre possibile trovare una costante $h > 0$ tale che:

$$d[(x_{11}, \ldots, hx_{1s}, \ldots, x_{1p}), (x_{21}, \ldots, hx_{2s}, \ldots, x_{2p})] <$$
$$d[(x_{31}, \ldots, hx_{3s}, \ldots, x_{3p}), (x_{41}, \ldots, hx_{4s}, \ldots, x_{4p})]$$

Per la distanza euclidea vale inoltre la seguente:

Proprietà III. La distanza euclidea è invariante per trasformazioni ortogonali (rotazioni) delle variabili (Späth, 1980, p. 17):

$$_2d(\mathbf{T}\mathbf{x}_i, \mathbf{T}\mathbf{x}_j) = {}_2d(\mathbf{x}_i, \mathbf{x}_j) \tag{4.13}$$

ove $\mathbf{T}$ è una matrice $p \times p$ tale che $\mathbf{T}'\mathbf{T} = \mathbf{I}$.

Ne consegue che le distanze euclidee tra i punti in R^p non mutano sia quando si effettua una traslazione degli assi di riferimento (proprietà II), sia quando si opera una rotazione degli stessi. Quest'ultima proprietà costituisce anche una giustificazione sotto l'aspetto formale della rotazione utilizzata nell'analisi delle componenti principali e nell'analisi dei fattori.

In forza della proprietà II, anche la distanza della città a blocchi è invariante per traslazione, mentre — con riferimento a due sole variabili rappresentabili nel piano — essa risulta invariante per rotazione solo quando l'angolo della stessa è uguale a $\pi/2$, π, $(3/2)\pi$, 2π (Rizzi, 1987). Il discorso si estende alla rotazione in presenza d'un numero generico p di variabili.

Per una trattazione più approfondita delle proprietà delle distanze e più in generale degli indici di prossimità rinviamo a Gower and Legendre (1986) (5).

---

per cui la coppia delle unità originariamente «più distanti» diviene quella delle unità «meno distanti».

(5) In un precedente lavoro (Zani, 1975) avevamo suggerito una proprietà ulteriore che riteniamo debba essere soddisfatta — sotto l'aspetto logico — da una distanza (o da un indice di distanza). Si considerino due vettori $\mathbf{x}, \mathbf{y} \in R^P$ e si indichi con $I(\mathbf{x}, \mathbf{y})$ l'insieme dei vettori le cui componenti sono comprese tra le corrispondenti componenti di $\mathbf{x}$ e $\mathbf{y}$; denominiamo «vettori intermedi tra $\mathbf{x}$ e $\mathbf{y}$» tutti quelli appartenenti all'insieme $I(\mathbf{x}, \mathbf{y})$. Pare del tutto naturale richiedere che se $\mathbf{z}$ è un vettore intermedio tra $\mathbf{x}$ e $\mathbf{y}$, ciascuna delle distanze da questi ultimi non superi la distanza tra $\mathbf{x}$ e $\mathbf{y}$. Si definiscono *coerenti* i tipi di distanza (o di indici di distanza) che godono senza eccezioni di questa proprietà.

Nello studio citato si è dimostrato, fra l'altro, che la distanza euclidea è sempre coerente, mentre ad esempio l'espressione (4.9) non è coerente, in generale, ma lo diviene se si considerano variabili che assumono solo valori positivi.

3.3. *La distanza ultrametrica*

Definizione. Si dice distanza ultrametrica tra i vettori $\mathbf{x}, \mathbf{y} \in R^p$ una funzione che gode delle proprietà di non negatività, identità e simmetria come una distanza, mentre la disuguaglianza triangolare è sostituita dalla seguente:

$$d(\mathbf{x}, \mathbf{y}) \leq max[d(\mathbf{x}, \mathbf{z}), d(\mathbf{y}, \mathbf{z})] \qquad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in R^P \qquad (4.14)$$

Osservazione I. La proprietà ultrametrica implica la disuguaglianza triangolare. Infatti, se vale la (4.14) vale *a fortiori*:

$$d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z}) \qquad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in R^P \qquad (4.15)$$

essendo la distanza non negativa per definizione. Pertanto, le distanze ultrametriche rappresentano una sottoclasse (molto particolare) delle distanze.

Osservazione II (Zani, 1975). Per una generica terna di unità si considerino le distanze ultrametriche tra ciascuna delle tre coppie e si supponga di ordinarle in senso non crescente. Siano d_1, d_2, d_3 tali distanze con $d_1 \geq d_2 \geq d_3$; per la proprietà ultrametrica deve essere:

$$\max(d_2, d_3) = d_2 \geq d_1$$

Ma, per l'ordinamento assunto, deve valere allora la relazione:

$$d_1 = d_2 \geq d_3.$$

In conclusione, per la distanza ultrametrica, comunque si scelgano tre unità statistiche, delle tre distanze relative alle tre coppie possibili ve ne sono almeno due uguali e la terza non le supera. In altri termini, in uno spazio ultrametrico tre punti formano sempre tra loro un triangolo equilatero oppure isoscele con la base più piccola dei due lati uguali (6).

Esempio. Consideriamo 6 consumatori d'un certo tipo di beni e supponiamo che vi siano due sole marche di quel bene. Definiamo la seguente distanza tra due generici consumatori:

0 se la marca è la stessa per i due consumatori;

1 se la marca è diversa.

---

(6) Per approfondimenti sulle distanze ultrametriche si veda Scozzafava (1995).

Ipotizziamo che i primi quattro individui siano consumatori della marca A e gli altri due della marca B. La matrice delle distanze tra gli individui è riporta nella tab. 4.4. Si verifica facilmente che i valori numerici che compaiono nella suddetta matrice sono distanze ultrametriche, poiché per ogni terna di individui essi soddisfano la proprietà corrispondente.

TAB. 4.4. *Matrice di distanze ultrametriche tra 6 consumatori, con riferimento a due marche A e B (0 = stessa marca; 1 = marca diversa).*

	A	A	A	A	B	B
A	0	0	0	0	1	1
A		0	0	0	1	1
A			0	0	1	1
A				0	1	1
B					0	0
B						0

4. *Indici di distanza ed indici di dissimilarità*

Rinunciando ad alcune tra le quattro proprietà che definiscono una distanza si ottengono famiglie più ampie di indici di prossimità, che risultano però meno rigorosi ed in talune circostanze possono dare origine ad incongruenze.

Definizione. Si dice *indice di distanza* tra due vettori $\mathbf{x}, \mathbf{y} \in R^p$ una funzione che soddisfa le proprietà di non negatività, identità e simmetria.

L'esempio più noto di indice di distanza è il quadrato della distanza euclidea:

$$_2d^2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2 \tag{4.16}$$

Con riferimento ai vettori corrispondenti a due generiche unità statistiche, esso risulta:

$$_2d_{ij}^2 = \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \sum_{s=1}^{p}(x_{is} - x_{js})^2 \tag{4.17}$$

Tale indice non soddisfa la disuguaglianza triangolare, come si evince dal seguente semplice esempio. Si considerino i vettori:

$$\mathbf{x}_1 = [10, \ 5]'$$
$$\mathbf{x}_2 = [11, \ 7]'$$
$$\mathbf{x}_3 = [13, \ 9]'$$

La matrice dei quadrati delle distanze euclidee risulta:

$$\left[{}_2d_{ij}^2\right] = \begin{bmatrix} 0 & 5 & 25 \\ & 0 & 8 \\ & & 0 \end{bmatrix}$$

Pertanto:

$${}_2d_{13}^2 \geq {}_2d_{12}^2 + {}_2d_{23}^2$$

in contrasto con la disuguaglianza triangolare (7).

L'indice di distanza definito come quadrato della distanza euclidea gode però dell'importante *proprietà di additività*: la somma del valore dell'indice calcolato su un primo sottoinsieme di p' variabili e del valore dell'indice calcolato su un secondo sottoinsieme di p'' variabili $(p' + p'' = p)$ è uguale al valore dell'indice calcolato direttamente su tutte le p variabili:

$$\sum_{s=1}^{p'}(x_{is} - x_{js})^2 + \sum_{s=p'+1}^{p}(x_{is} - x_{js})^2 = \sum_{s=1}^{p}(x_{is} - x_{js})^2 \qquad (4.18)$$

(7) Con riferimento ad un'assegnata matrice dei dati, è sempre possibile ricondurre un generico indice di distanza I_{ij} ad una distanza (cioè far sì che esso soddisfi anche la disuguaglianza triangolare per l'insieme considerato) mediante la seguente trasformazione (Chandon et Pinson, 1981, p. 53):

$$d_{ij} = I_{ij} + c, \qquad \text{ove} \qquad c = max_{i,j}(I_{ij})$$

Tale trasformazione è però di scarso interesse applicativo, poiché quando si ritiene essenziale la proprietà di disuguaglianza triangolare si sceglie direttamente una distanza, e non un indice di distanza.

Si verifica facilmente che la distanza euclidea non soddisfa invece tale importante proprietà.

Inoltre, il quadrato della distanza euclidea calcolata su vettori standardizzati, $\mathbf{z}_x$ e $\mathbf{z}_y$, presenta la seguente relazione con il coefficiente di correlazione lineare, r_{xy} (per la dimostrazione rinviamo al vol. I, pp. 160-161):

$$_2d_{ij}^2 = 2n(1 - r_{xy}) \tag{4.19}$$

Se i vettori sono normalizzati, cioè tali che $\|\mathbf{x}\| = 1$ e $\|\mathbf{y}\| = 1$, la relazione è la seguente (Späth, 1980, p. 20):

$$_2d_{ij}^2 = 2(1 - r_{xy}) \tag{4.20}$$

Le espressioni precedenti mostrano come sia possibile passare da una misura della relazione tra due vettori ad un indice di distanza tra essi (8).

Una categoria ulteriore di indici è fornita dalla seguente:

Definizione. Si dice indice di dissimilarità (ovvero indice di diversità nel senso di G. Leti) tra i vettori $\mathbf{x}, \mathbf{y} \in R^P$ una funzione che soddisfa le proprietà di non negatività e di simmetria ed anche la seguente proprietà:

$$\mathbf{x} = \mathbf{y} \Rightarrow d(\mathbf{x}, \mathbf{y}) = 0 \tag{4.21}$$

che è chiaramente più debole della proprietà di identità.

La classe degli indici di diversità è più ampia di quella degli indici di distanza, ma contiene misure meno rigorose della dissimilarità tra

---

(8) Dal punto di vista puramente algebrico, il coefficiente di correlazione potrebbe essere utilizzato come indice di prossimità tra i vettori riga corrispondenti alle unità u_i e u_j. Sotto l'aspetto dell'interpretazione statistica, tale coefficiente si rivela però inadatto, poiché i vettori suddetti contengono i valori di variabili *differenti*, in generale non direttamente confrontabili. Si osservi che neppure la consueta traduzione in termini di scostamenti standardizzati risolve il problema, poiché essa avviene con riferimento ai vettori colonna della matrice dei dati (che corrispondono ai valori delle singole variabili), mentre in un indice di prossimità tra due unità statistiche si considerano coppie di vettori riga. Per questi ultimi il calcolo d'un coefficiente di correlazione — definito, lo ricordiamo, come covarianza tra vettori standardizzati — richiederebbe un'ulteriore operazione di standardizzazione dei dati di ciascuna riga.

due vettori (corrispondenti a due unità statistiche), poiché un indice di diversità può risultare uguale a 0 anche quando i due vettori posti a confronto non sono identici.

Alcuni impieghi degli indici di dissimilarità saranno illustrati nel capitolo dedicato allo scaling multidimensionale.

5. *Impiego ed interpretazione delle distanze in statistica*

5.1. *La comparabilità delle variabili ed il problema dei dati mancanti*

I tipi di distanza considerati in precedenza sono tutti funzione delle differenze in modulo tra i valori che le p variabili presentano in due unità statistiche. La somma di tali differenze (o di una potenza delle stesse) ha però significato solo se tutte le variabili sono espresse nella stessa unità di misura (9).

Anche in tale circostanza, tuttavia, le distanze dei tipi precedenti non risultano del tutto appropriate per misurare la dissimilarità tra due generici elementi, poiché esse risultano fortemente influenzate dai caratteri con più elevato ordine di grandezza e con maggiore variabilità (che presentano quindi differenze più marcate tra i valori).

Supponiamo, ad esempio, di aver rilevato per un insieme di n aziende le variabili: X_1= fatturato; X_2= spesa per pubblicità televisiva; X_3= spesa per pubblicità a mezzo stampa. Le tre variabili sono tutte espresse in lire, per cui sarebbe lecito impiegare una distanza, ad esempio euclidea, per misurare la dissimilarità tra coppie di aziende, con riferimento agli aspetti considerati. Avendo però il fatturato un ordine di grandezza ed una variabilità molto maggiori degli altri due caratteri, la distanza sarebbe determinata in maniera preponderante da tale variabile.

Si possono superare entrambe le limitazioni suddette se le distanze tra le unità vengono calcolate, anziché sulle variabili originarie, su opportune trasformazioni delle stesse. Una scelta molto naturale è quella di considerare gli scostamenti standardizzati di ciascuna delle variabili (10).

---

(9) Fa eccezione, come s'è detto, la metrica di Canberra, che utilizza scostamenti relativi.

(10) Il tipo suddetto di trasformazione — pur essendo quello di più comune