

Capitolo 4

L'analisi statistica multivariata applicata alle ricerche di mercato

4.1 Introduzione

I capitoli precedenti hanno discusso alcuni aspetti relativi alla raccolta delle informazioni rilevanti per un problema conoscitivo o decisionale che coinvolge la funzione di marketing.

Volgiamo ora la nostra attenzione ad alcuni strumenti di analisi delle informazioni raccolte che sono funzionali ad una varietà di problemi che sono illustrati più a fondo nelle sezioni 4.2-4.4.

L'informazione raccolta con una ricerca o indagine di mercato ha natura tipicamente multidimensionale, dal momento che, con riferimento ad un insieme di individui, prodotti, marche, opportunamente o casualmente selezionate, abbiamo rilevato una batteria di indicatori o attributi atti a descriverne il profilo. In altre circostanze chiediamo al rispondente un confronto su una pluralità di oggetti o marche, di modo che la raccolta dei dati si presenta sotto forma di una matrice di (dis)similarità o distanza.

In definitiva, i dati sottoposti all'analisi possono essere organizzati per riga e per colonna per formare una matrice. Questa può essere considerata come un'entità matematica che può essere manipolata in vari modi al fine di evidenziare alcuni aspetti di interesse. Le manipolazioni che possiamo operare formano una vera e propria *algebra*, i cui principi fondamentali vengono esposti nell'appendice B, che fornisce soltanto un riferimento sintetico ad alcune operazioni essenziali. Il nostro corso sarà interessato principalmente all'interpretazione geometrica di quelle operazioni e le richiamerà all'occorrenza.

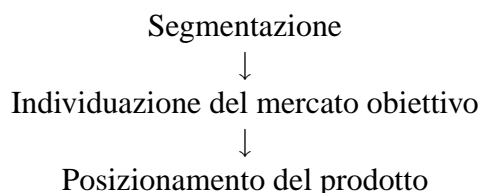
Dopo aver introdotto la matrice dei dati ed alcune sintesi elementari volte al calcolo delle medie e della matrice di varianze-covarianze e dell'associata matrice di correlazione (sezione 4.5), la nostra attenzione si concentrerà sulle misure della similarità o della distanza che possono essere calcolate a partire dai profili individuali misurati mediante una batteria di attributi.

4.2 La segmentazione del mercato

La segmentazione perviene ad una suddivisione del mercato in gruppi omogenei e distinti di consumatori che esprimono una domanda differenziata, richiedendo specifici prodotti e attributi. Ad essi vanno pertanto indirizzate specifiche politiche di marketing. Essa richiede la presenza di tre elementi essenziali:

1. Eterogeneità dei consumatori/utilizzatori
2. Differenziazione della domanda (l'eterogeneità si riflette sulla domanda di mercato che è differenziata)
3. Separazione (è possibile isolare segmenti di consumatori all'interno del mercato complessivo).

Perché una politica di segmentazione abbia successo occorre che vi sia uniformità di risposta alle variabili di marketing mix da parte degli acquirenti potenziali, la dimensione del segmento deve assicurarne la profittabilità ed inoltre il segmento deve essere accessibile (uniformità, profittabilità, accessibilità).



La segmentazione richiede che sia individuato il meccanismo che presiede alla formazione delle preferenze individuali ed al processo di scelta.

4.2.1 Le fasi operative

Dal punto di vista operativo si distinguono due fasi

1. Fase cognitiva, o analitica, consistente nella definizione dei segmenti mediante opportune tecniche statistiche.
2. Fase strategica, rivolta alla definizione degli strumenti di marketing mix.

La prima si articola nella scelta delle variabili (basi) e del modello di segmentazione. Le variabili di segmentazione possono riguardare aspetti demografici, economici e sociali dei consumatori e caratteristiche legate alla situazione specifica di consumo. La scelta è ovviamente condizionata dalle informazioni disponibili.

Variabili di segmentazione:

- *Geografiche*: Stato - Regione - Città - Densità (zona urbana, semiurbana, rurale) - Ripartizione - Clima - zona altimetrica.

- *Demografiche*: Sesso, età, Stato civile, dimensione del nucleo familiare, razza, religione. Si parla di *segmentazione geodemografica* quando la scelta delle basi ricade su variabili geografiche e demografiche. La base informativa è solitamente rappresentata dai dati censuari, l'unità territoriale minima essendo costituita dalla sezione di censimento (con riferimento al Censimento della popolazione del 1991, il territorio nazionale è stato suddiviso in 323.000 sezioni caratterizzate da una presenza media di 66 famiglie e 180 individui). Per l'individuazione di un numero limitato di profili si ricorre all'analisi dei grappoli (cfr. cap. 5).
- *Economiche*: reddito, attività economica, condizione professionale;
- *Sociali*: grado di istruzione, classe sociale se l'unità di segmentazione sono le famiglie o gli individui; la dotazione di infrastrutture sociali, come gli ospedali, gli asili nido, le scuole, se le unità sono i comuni o altre entità territoriali.
- *Psicografiche*: stile di vita, personalità. La segmentazione effettuata in base al profilo psicografico rileva la base informativa mediante questionari compilati dal rispondente che contengono una serie di domande attitudinali - ad esempio organizzate su una scala di Likert. Le informazioni vengono poi trattate con tecniche multivariate, come l'analisi dei fattori (cfr. cap. ??), per estrarre i profili latenti.
- *Comportamentali*: fedeltà di marca, intenzione d'acquisto, attributi richiesti al prodotto. Si parla in proposito di segmentazione in base alle preferenze. Queste possono presentarsi in tre tipologie fondamentali [3]: omogenee, diffuse e clusterizzate. Le prime richiedono una strategia indifferenziata - prezzo e disponibilità sono variabili cruciali - nel secondo caso ci si può collocare nel baricentro al fine di minimizzare l'insoddisfazione dei consumatori, anche se potrebbe essere opportuno concentrarsi su una dimensione per creare un segmento di mercato. Nel terzo caso è richiesta una strategia di marketing differenziata.

Con riferimento al modello di segmentazione, si possono individuare tre tipologie essenziali: i. segmentazione a priori ii. segmentazione a posteriori iii. segmentazione flessibile e composita.

Nel primo caso la definizione dei segmenti ed il loro numero sono stabiliti in via preliminare in base alle conoscenze teoriche e a studi precedenti. Il processo si riduce all'attribuzione dei soggetti a classi predisposte relative alle basi di segmentazione scelte. Nel caso di più basi si fa ricorso alla tabulazione incrociata. Si rende necessario un controllo a posteriori della capacità discriminante delle basi scelte mediante test χ^2 e modelli log-lineari. Successivamente alla fase di formazione dei segmenti si possono individuare i profili sottostanti mediante tecniche multivariate.

Nel secondo il numero e la tipologia dei segmenti non è prefissato, ma emerge dal raggruppamento degli intervistati mediante tecnica statistica. Si distinguono due casi: 1. viene individuata una variabile dipendente y , collegata alla preferenza o all'uso del

prodotto, rispetto alla quale viene valutata l'omogeneità del segmento ottenuto. 2. La formazione dei segmenti si fonda su una matrice di similarità calcolata sui profili dei rispondenti che risultano dalla rilevazione di alcuni caratteri inerenti il comportamento d'acquisto o le loro attitudini verso il prodotto.

Infine, la segmentazione flessibile risulta dalla integrazione dei risultati di un'analisi congiunta e di una simulazione sul comportamento di scelta dei consumatori. Si prende così in considerazione un numero elevato di ipotetici segmenti alternativi. Questi modelli si differenziano dai precedenti per la possibilità di costruire segmenti definiti in base alla risposta dei consumatori ad offerte alternative. Effettuata la scelta del segmento, si procede a valutarne l'ampiezza e le principali caratteristiche.

La base informativa richiesta dall'operazione di segmentazione è tipicamente multivariata. E' possibile organizzare l'informazione in una matrice le cui righe rappresentano il consumatore/cliente e le colonne le basi di segmentazione (cfr. sez. 4.5).

4.3 Il posizionamento del prodotto

Il prodotto ha nel marketing una dimensione multivariata poiché è considerato come una particolare combinazione di attributi nei confronti dei quali il consumatore esercita determinate preferenze.

La percezione di una marca da parte dell'acquirente (immagine di marca) dipende essenzialmente dai seguenti elementi:

1. le caratteristiche oggettive del prodotto/marca (servizio elementare offerto, componenti chimico-fisiche, organolettiche, etc.)
2. gli attributi del prodotto/marca (attributi di natura funzionale, percettiva, affettiva, estetica che danno origine a soddisfazione)
3. il grado di presenza degli attributi
4. il livello di importanza degli attributi e) il valore o utilità parziale degli attributi (tra gli attributi esiste un naturale trade-off, per cui il consumatore opera un processo di scelta).

Secondo la tecnica che prende il nome di analisi congiunta, l'aggregazione dei valori individuali (soggettivi) associati a ciascun attributo consente di pervenire ad una valutazione di sintesi sull'atteggiamento dei consumatori nei confronti del prodotto/marca. Secondo l'approccio compositivo, si perviene a tale valutazione di sintesi mediante media aritmetica ponderata dei punteggi sul grado di presenza con pesi pari all'importanza relativa di ciascun attributo. La conoscenza di questa valutazione ha importanza strategica per l'azienda ed orienta le strategie di marketing da adottare, ad es. modificando il

prodotto se non incontra il favore del mercato o spostando l'attenzione dei consumatori su particolari caratteristiche del prodotto mediante un'azione pubblicitaria.

In un mercato fortemente concorrenziale è di particolare importanza che l'azienda conosca la sua posizione di mercato. Questa coinvolge la segmentazione del mercato e l'analisi della quota di mercato. Un ulteriore aspetto è la posizione dei prodotti e delle marche offerti rispetto a quelli della concorrenza così come viene percepita dai consumatori. Lo strumento di analisi consiste nella costruzione di mappe di tipo percettivo a partire da indicazioni sul modo in cui i consumatori sentono simili o dissimili le varie alternative di prodotto offerte sul mercato.

4.4 Le mappe percettive

Uno dei fondamentali problemi delle ricerche di mercato è analizzare e comprendere come il consumatore o cliente percepisca il prodotto (inteso in senso ampio, potendosi trattare di un servizio commerciale, turistico, finanziario, etc.) o la marca. Da tale conoscenza scaturisce, solitamente, un vantaggio competitivo differenziale.

Ciò implica la conoscenza de: 1. il numero di dimensioni o fattori latenti che il consumatore utilizza nel discriminare i prodotti o le marche 2. la natura dei fattori latenti (identificazione) 3. il posizionamento dei prodotti esistenti lungo queste dimensioni 4. la localizzazione del prodotto ideale lungo le medesime dimensioni.

Le mappe percettive costituiscono rappresentazioni di oggetti, marche, prodotti in uno spazio dimensionale. Evidenziano quali prodotti sono in diretta competizione nella percezione del consumatore e suggeriscono come posizionare il prodotto al fine di massimizzare le preferenze e le vendite. Esse pertanto sintetizzano in modo efficace la struttura del mercato e sono suscettibili di altri impieghi quali l'identificazione dei punti deboli di un prodotto, lo sviluppo e la valutazione della concezione di nuovi prodotti, l'identificazione delle differenze tra gruppi.

Fondamentalmente, esistono due approcci alla costruzione delle mappe percettive:

1. Approccio basato sugli attributi dei prodotti: fa affidamento sulla valutazione individuale delle singole caratteristiche degli oggetti, utilizzando una scala di Likert o del differenziale semantico. Tali valutazioni sono poi sottoposte ad una analisi fattoriale o all'analisi discriminante.
2. Approccio basato su misure di similarità o preferenza: al consumatore viene chiesto direttamente il giudizio sulla similarità tra oggetti. L'analisi di scaling multidimensionale colloca gli oggetti in uno spazio di dimensioni pari al numero dei fattori latenti utilizzati nella formazione del giudizio di similarità o preferenza.

Il vantaggio del secondo approccio costituisce anche il suo limite principale: da un lato non si richiede che il confronto avvenga sulla base di attributi prefissati e quindi consente che nel collocare gli oggetti il rispondente utilizzi le dimensioni che abitualmente usa

nella realtà. Dall'altro risulta problematico individuare quelle dimensioni proprio perché non si rileva l'attitudine verso determinati attributi.

4.5 La matrice dei dati

Supponiamo di aver misurato, su una scala ad intervallo o di rapporti, p caratteri relativi a n oggetti (marche, prodotti, individui). Le misurazioni vengono raccolte nella matrice di dati:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ik} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{np} \end{bmatrix},$$

il cui termine generico, x_{ik} , fornisce la misurazione k -esima per l'unità i . Solitamente, l'indice i contrassegna un individuo o un prodotto, mentre l'indice j contrassegna un attributo di i . Tuttavia, in alcuni casi (matrici di dissimilarità e distanza, correlazione) essi identificano il medesimo insieme di unità.

Se per una qualunque delle variabili la misurazione ha natura nominale a due categorie (dicotomica), essa viene rappresentata nella matrice dei dati utilizzando la codifica binaria $x_{ik} = 1$ se un evento si verifica, $x_{ik} = 0$ altrimenti.

Se l'analisi è prevalentemente indirizzata alle unità di riga, possiamo rappresentare la matrice dei dati come una matrice a blocchi il cui blocco generico è rappresentato da un vettore $1 \times p$,

$$\mathbf{x}'_i = [x_{i1}, x_{i2}, \dots, x_{ik}, \dots, x_{ip}],$$

che contiene il profilo dell'unità di riga:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n]'$$

Viceversa, se l'oggetto dell'analisi sono gli attributi o, in generale, le unità di colonna denoteremo con \mathbf{x}_k il vettore colonna $n \times 1$ e

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k, \dots, \mathbf{x}_p].$$

Il primo obiettivo che ci proponiamo è quello di conseguire delle sintesi che consentano di rappresentare con parsimonia le informazioni più importanti che sono contenute nella matrice dei dati. Le analisi che effettueremo riguardano essenzialmente: a) le relazioni esistenti tra le unità di riga (oggetti, individui, marche): similarità e distanza; b) le relazioni esistenti tra le unità di colonna (variabili, caratteristiche, attributi): correlazione; c) le relazioni che intercorrono tra le unità di riga e quelle di colonna.

Con riferimento al punto a), se p è inferiore a 3, è agevole rappresentare graficamente le unità di riga come punti nello spazio euclideo p -dimensionale; inoltre, quando $p > 3$, ed il numero delle le unità di riga è sufficientemente contenuto, si può percepire immediatamente la similarità mediante semplici strumenti grafici tra cui:

- *le facce di Chernoff*: ciascuna unità di riga viene rappresentata mediante una faccia costruita in modo da associare ad ognuna dei p attributi un particolare tratto somatico: forma del viso, occhi, naso, sopracciglia, bocca, etc. Un esempio viene fornito dalla figura 4.1, ottenuta utilizzando la funzione `faces()` del software S-plus a partire da una matrice contenente la percentuale di individui che concordano con 11 statement relativi ad altrettanti attributi di otto marche di cereali.
- *stelle e diamanti*; gli attributi delle unità di riga, invece di definire i tratti somatici delle facce, costituiscono la lunghezza di segmenti disposti a raggiera e le cui estremità sono unite da linee. Il risultato è una successione di figure poliedriche, ognuna associata ad un oggetto diverso, che consentono di evidenziare visivamente i diversi gradi di similarità tra i medesimi.

Per quanto concerne il punto b), la sintesi delle informazioni relative alla distribuzione congiunta delle variabili può essere effettuata attraverso i momenti. Il momento primo (la media aritmetica semplice) delle p variabili viene raccolto nel vettore

$$\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k, \dots, \bar{x}_p]', \quad \bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik}$$

I momenti secondi (centrati) sono raccolti nella matrice di covarianza \mathbf{S} , di dimensione $p \times p$:

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} & \dots & s_{1k} & \dots & s_{1p} \\ s_{21} & s_2^2 & \dots & s_{2k} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \dots & \vdots & \vdots \\ s_{k1} & s_{k2} & \dots & s_k^2 & \dots & s_{kp} \\ \vdots & \vdots & \dots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pk} & \dots & s_k^2 \end{bmatrix},$$

con

$$s_k^2 = \frac{1}{n} \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2, \quad s_{hk} = \frac{1}{n} \sum_{i=1}^n (x_{ih} - \bar{x}_h)(x_{ik} - \bar{x}_k);$$

tale matrice è simmetrica ($\mathbf{S} = \mathbf{S}'$) in virtù del fatto che $s_{hk} = s_{kh}$, e semidefinita positiva.

La descrizione effettuata a partire dai momenti fino al secondo ordine risulta sufficiente a descrivere il fenomeno solo sotto l'ipotesi di normalità. Altrimenti si dovrebbero

Figura 4.1: Facce di Chernoff per il data set `cereal.attitude`.

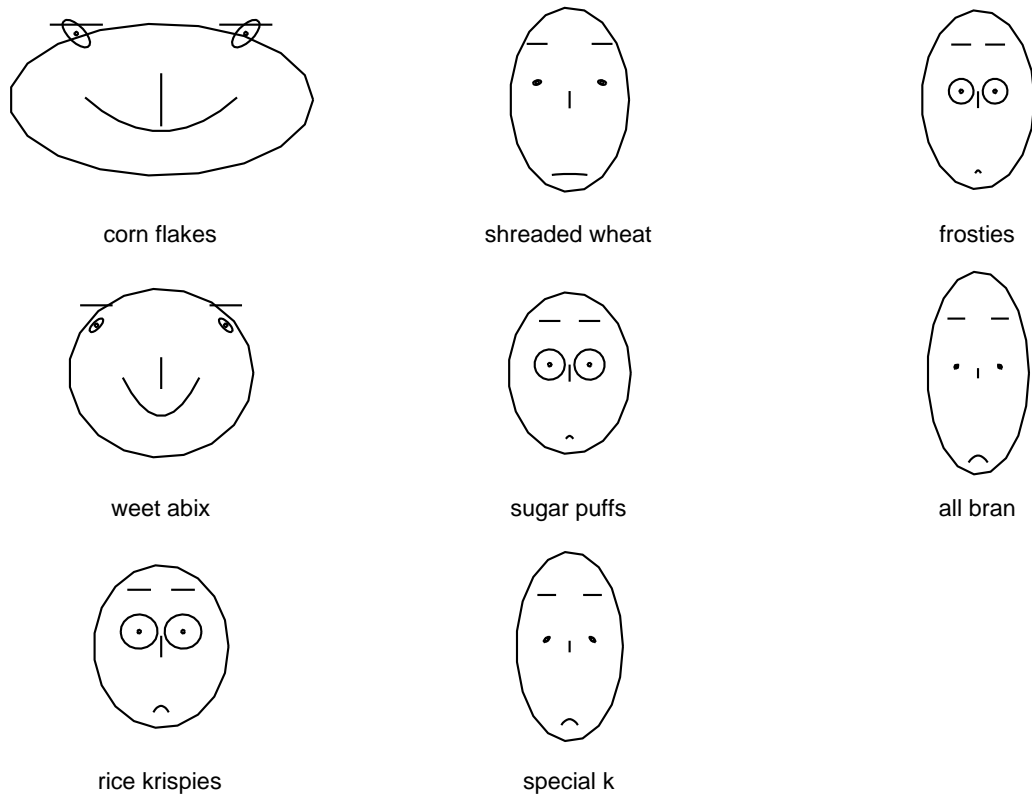
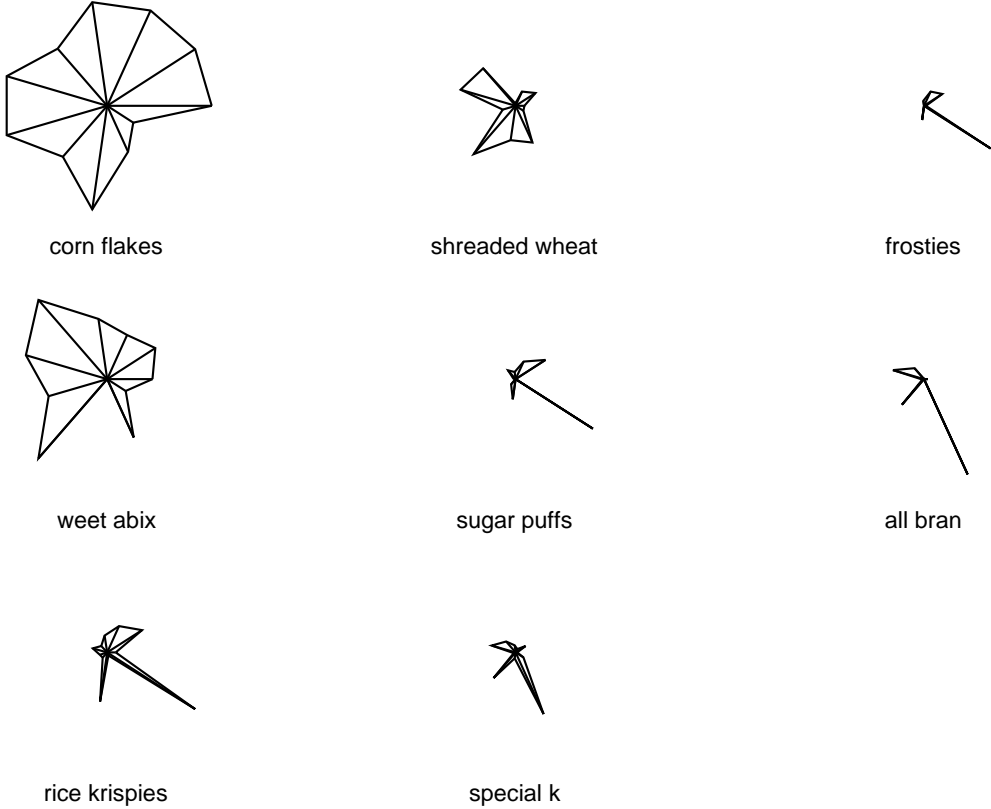


Figura 4.2: Grafico a diamante per il data set cereal.attitude



considerare anche i momenti di ordine superiore al secondo. La matrice \mathbf{S} può essere riscritta:

$$\mathbf{S} = \frac{1}{n}(\mathbf{X} - \mathbf{i}_n \bar{\mathbf{x}})'(\mathbf{X} - \mathbf{i}_n \bar{\mathbf{x}}) = \frac{1}{n}\mathbf{X}'\mathbf{X} - \bar{\mathbf{x}}\bar{\mathbf{x}}',$$

dove \mathbf{i}_n rappresenta un vettore $n \times 1$ di termini tutti unitari; ovvero può essere espressa in termini dei vettori riga della matrice \mathbf{X} :

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

Poiché la matrice di covarianza \mathbf{S} risente dell'unità di misura in cui sono espresse le variabili, introduciamo la matrice di correlazione \mathbf{R} : definendo $\mathbf{D} = \text{diag}\{s_k^2, k = 1, \dots, p\}$,

$$\mathbf{R} = \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2}$$

il generico elemento è il coefficiente di correlazione (lineare) di Bravais tra l' h -esima variabile e la k -esima variabile:

$$r_{hk} = \frac{s_{hk}}{s_h s_k}, \quad |r_{hk}| \leq 1.$$

4.6 Misura della similarità e della distanza

Data una coppia di unità statistiche, i e j , desideriamo confrontare i rispettivi profili, pervenendo a misure di distanza, d_{ij} , e similarità, c_{ij} . Dette misure dipendono dalla scala di misurazione degli attributi.

4.6.1 Similarità e distanza per caratteri quantitativi

Una misura di distanza deve godere delle seguenti proprietà:

1. $d_{ij} \geq 0$ (non negatività)
2. $d_{ii} = 0$
3. $d_{ij} = d_{ji}$ (simmetria)
4. $d_{ij} \leq d_{ir} + d_{rj}$ (diseguaglianza triangolare)

Se una misura di distanza soddisfa tutte le quattro proprietà, si dice che lo spazio di riferimento è metrico.

Distanza euclidea Siano \mathbf{x}_i e \mathbf{x}_j due vettori contenenti il profilo di due unità, misurato su p attributi. La distanza euclidea è definita dalla norma della differenza tra i vettori rappresentativi delle unità:

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\| = [(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)]^{1/2} = \left[\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right]^{1/2}.$$

La distanza euclidea gode delle quattro proprietà elencate sopra ed è definita in uno spazio uniforme e per così dire lineare, dove, euristicamente parlando, ci si può muovere da un punto all'altro in linea d'aria. Questa misura può essere estesa al fine di pervenire a misure della distanza dalla connotazione più statistica. In primo luogo si osserva che il contributo alla distanza complessiva fornito dalle p variabili che definiscono il profilo dipende dalla scala di misurazione di ciascuna di esse; ad esempio, il numero dei componenti la famiglia contribuirà di meno rispetto alla spesa per spettacoli nell'ultimo mese espressa in euro. Inoltre, la distanza euclidea non tiene conto della relazione statistica che intercorre tra le variabili.

Distanza euclidea ponderata Sia \mathbf{W} una matrice diagonale contenente i coefficienti di ponderazione, $\mathbf{W} = \text{diag}(w_1, \dots, w_p)$:

$$d_{ij} = [(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{W} (\mathbf{x}_i - \mathbf{x}_j)]^{1/2} = \left[\sum_{k=1}^p (x_{ik} - x_{jk})^2 w_k \right]^{1/2}.$$

La distanza euclidea ponderata è una forma quadratica della matrice \mathbf{W} . Può essere generalizzata al caso in cui \mathbf{W} sia una matrice simmetrica piena, la quale tuttavia deve essere semi definita positiva affinché si abbia $d_{ij} \geq 0$.

La necessità di ricorrere a tale distanza sorge in due contesti:

- Standardizzazione delle variabili: $\mathbf{W} = \text{diag}(1/s_1^2, \dots, 1/s_p^2)$; in tale caso la d.e. ponderata equivale alla distanza euclidea calcolata sui profili standardizzati:

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}.$$

L'obiettivo è quello di neutralizzare l'effetto della scala di misurazione delle variabili.

- Distanza di Mahalanobis. La matrice di ponderazione è l'inversa della matrice di covarianza: $\mathbf{W} = \mathbf{S}^{-1}$. Costituisce una misura statistica della distanza tra le unità, che viene calcolata al netto della correlazione esistente tra le variabili.

$${}_M d_{ij} = [(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)]^{1/2}.$$

Si dimostra facilmente che essa può anche essere calcolata come distanza euclidea ponderata applicata ai profili standardizzati, con $\mathbf{W} = \mathbf{R}^{-1}$:

$${}_M d_{ij} = [(\mathbf{z}_i - \mathbf{z}_j)' \mathbf{R}^{-1} (\mathbf{z}_i - \mathbf{z}_j)]^{1/2},$$

dove \mathbf{z}_i è il vettore $p \times 1$ che contiene i valori standardizzati

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}, \quad k = 1, \dots, p.$$

In notazione matriciale, $\mathbf{z}_i = \mathbf{D}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$. Dal punto di vista computazionale, mostreremo nel capitolo 6 che la distanza di Mahalanobis coincide con la distanza euclidea calcolata sulle componenti principali standardizzate.

Esempio Consideriamo i profili standardizzati di due unità calcolati su due variabili:

$$\mathbf{z}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \mathbf{z}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

$${}_M d_{ij}^2 = (-2 \quad -2) \mathbf{R}^{-1} \begin{pmatrix} -2 \\ -2 \end{pmatrix}$$

Nel caso in cui le variabili sono incorrelate, $\mathbf{R} = \mathbf{I}_2$, dove \mathbf{I}_2 è la matrice identità di ordine $p = 2$, la distanza di Mahalanobis coincide con la distanza euclidea: ${}_M d_{12}^2 = 8$. In presenza di debole correlazione positiva,

$$\mathbf{R} = \begin{bmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{bmatrix},$$

la distanza di Mahalanobis è più piccola rispetto alla distanza euclidea (${}_M d_{12}^2 = 6.67$); In presenza di forte correlazione positiva,

$$\mathbf{R} = \begin{bmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{bmatrix},$$

la distanza di Mahalanobis diminuisce ulteriormente (${}_M d_{12}^2 = 4.21$). Se infine le variabili sono fortemente correlate negativamente),

$$\mathbf{R} = \begin{bmatrix} 1.0 & -0.9 \\ -0.9 & 1.0 \end{bmatrix},$$

il quadrato della distanza è 10 volte più grande (${}_M d_{12}^2 = 80$).

Si osservi che se le p variabili sono incorrelate e omoschedastiche, che equivale a dire che presentano la stessa varianza, allora $\mathbf{S} = s^2 \mathbf{I}$ e $\mathbf{R} = \mathbf{I}$, dove \mathbf{I} è la matrice identità di ordine p , e la distanza di Mahalanobis risulta uguale a quella standardizzata ed è proporzionale a quella euclidea semplice (${}_M d_{ij} = d_{ij}/s$); se invece le variabili sono incorrelate e eteroschedastiche (varianza diversa), ovvero $\mathbf{S} = \text{diag}(s_1^2, \dots, s_k^2, \dots, s_p^2)$ e $\mathbf{R} = \mathbf{I}$, la distanza di Mahalanobis risulta uguale a quella standardizzata ed entrambe differiscono dalla distanza euclidea semplice.

Distanza di Minkowski Una famiglia di misure di distanza si ottiene dall'espressione

$$d_{ij} = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^\lambda \right]^{1/\lambda},$$

al variare di λ ($\lambda > 0$). Casi particolari:

1. $\lambda = 2 \Rightarrow$ distanza euclidea;
2. $\lambda = 1 \Rightarrow$ distanza della città a blocchi (o distanza di Manhattan);
3. $\lambda = \infty \Rightarrow$ distanza di Lagrange, $d_{ij} = \max_k \{|x_{ik} - x_{jk}|\}$.

4.6.2 Distanza e similarità tra variabili o attributi

Il coefficiente di correlazione di Bravais tra due variabili, r_{hk} , fornisce una misura del legame associativo di natura lineare tra le medesime. Come è noto, il suo campo di definizione è rappresentato dall'intervallo $[-1, 1]$. Al fine di pervenire ad una misura della distanza basata su r_{hk} si possono utilizzare due definizioni:

$$d_{hk} = 1 - r_{hk},$$

che assume valori in $[0, 2]$, ovvero,

$$d_{hk} = 1 - r_{hk}^2,$$

che a sua volta varia in $[0, 1]$. Va osservato che nel primo caso la distanza assume valore massimo in presenza di una perfetta correlazione negativa tra le variabili, mentre nel secondo si prescinde dal segno della correlazione, e pertanto la distanza è massima in assenza di correlazione. La scelta tra le due alternative dipende dal contesto e contiene elementi di arbitrarietà.

Se \mathbf{z}_h e \mathbf{z}_k costituiscono due vettori contenenti n realizzazioni delle due variabili (ad esempio, $z_{ik} = (x_{ik} - \bar{x}_k)/s_k$), la distanza euclidea,

$$d_{hk}^2 = (\mathbf{z}_h - \mathbf{z}_k)'(\mathbf{z}_h - \mathbf{z}_k) = \|\mathbf{z}_h\|^2 + \|\mathbf{z}_k\|^2 - 2\mathbf{z}_h'\mathbf{z}_k = 2n(1 - r_{hk}),$$

assume valori in $[0, 4n]$ e adotta una logica prossima alla prima definizione.

4.6.3 Misure di distanza per variabili qualitative dicotomiche

Supponiamo che la matrice \mathbf{X} contenga p misurazioni nominali effettuate su n individui; in particolare, si valuta la presenza (1) o l'assenza (0) di p attributi:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Profilo unità i	1	1	0	1	1	0	1	0	0
Profilo unità j	1	0	1	1	0	1	1	1	0

Con riferimento alle due unità, possiamo sintetizzare le due righe della matrice dei dati mediante la seguente tabella di contingenza:

		unità i	
		1	0
unità j	1	a	b
	0	c	d

dove a rappresenta il numero dei caratteri presenti in entrambe le unità; b il numero dei caratteri presenti nell'unità j , ma assenti nell'unità i ; c il numero dei caratteri presenti nell'unità i , ma assenti nell'unità j ; d il numero dei caratteri assenti in entrambe le unità. Ovviamente, si avrà:

$$a + b + c + d = p.$$

In letteratura sono presenti diversi modi di calcolare la similarità che differiscono principalmente per il trattamento riservato all'aggregato d .

1. Simple matching: una misura di similarità è fornita dalla frequenza relativa degli attributi presenti o assenti da entrambe le unità (coefficiente di simple matching):

$$c_{ij} = \frac{a + d}{p}$$

In corrispondenza si definisce la misura di distanza:

$$d_{ij} = \frac{b + c}{p}.$$

Nell'esempio precedente: $a = 3, b = 2, c = 3, d = 1$, per cui, $c_{ij} = 4/9, d_{ij} = 5/9$. Si osservi che la distanza euclidea fornisce $d_{ij} = [\sum_k (x_{ik} - x_{jk})^2]^{1/2} = \sqrt{b + c}$.

2. Coefficiente di similarità di Jaccard:

$$c_{ij} = \frac{a}{a + b + c}$$

A differenza del precedente, esclude dal confronto il numero di attributi che sono assenti da entrambe le unità. Per contro il coefficiente di distanza sarà,

$$d_{ij} = \frac{b + c}{a + b + c} = 1 - c_{ij}.$$

3. Coefficiente di similarità di Czekanowski:

$$c_{ij} = \frac{2a}{2a + b + c}, d_{ij} = \frac{b + c}{2a + b + c}$$

Assegna peso doppio al numero di attributi presenti in entrambe le unità e peso nullo agli attributi assenti in entrambe.

4.6.4 Similarità e distanza tra attributi dicotomici

Una misura di similarità per attributi dicotomici è fornita dalla statistica χ^2 calcolata con riferimento alla seguente tabella di contingenza:

		attributo h	
		1	0
attributo k	1	a	b
	0	c	d

dove a rappresenta il numero delle unità che presentano entrambi gli attributi; b il numero delle unità che presentano l'attributo k e non l'attributo h ; c il numero delle unità che presentano l'attributo h e non l'attributo k ; d il numero delle unità che non presentano entrambi gli attributi; si noti che

$$a + b + c + d = n.$$

$$\chi_{hk}^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + c)(b + d)(a + b)(c + d)}.$$

4.6.5 Misura della similarità per variabili qualitative politomiche

Il coefficiente di matching di Sneath misura la similarità c_{ij} mediante la frazione o la percentuale di attributi per i quali gli individui i e j presentano la stessa modalità.

4.6.6 Misura della distanza per misurazioni ordinali

Una soluzione consiste nell'attribuire un punteggio alle categorie ed utilizzare una delle misure di distanza o similarità introdotte per i caratteri quantitativi. L'operazione contiene ovvi elementi di arbitrarietà. Altrimenti si potrebbe declassare la misurazione al livello nominale, applicando il coefficiente di matching di Sneath.

4.6.7 Misura della similarità per dati misti

In generale la matrice X contiene misurazioni effettuate su tutte le scale prese in considerazione. Un giudizio complessivo circa la similarità tra gli oggetti di riga si ottiene dall'indice di similarità di Gower:

$$c_{ij} = \frac{\sum_{k=1}^p c_{ij,k}}{\sum_{k=1}^p \delta_{ij,k}}$$

dove $c_{ij,k}$ è una misura di similarità fra le unità i e j calcolata con riferimento al k -esimo attributo, mentre $\delta_{ij,k}$ è una variabile nominale che assume valore unitario se le

unità possono essere confrontate con riferimento all'attributo k e zero altrimenti. In altre parole, essa denota l'ammissibilità del confronto.

La definizione di queste quantità varia a seconda della tipologia delle variabili:

- variabili quantitative:

$$c_{ij,k} = 1 - d_{ij,k} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}, \quad \delta_{ij,k} = 1$$

dove R_k rappresenta il campo di variazione (*range*) della variabile k

- variabili qualitative politomiche: $c_{ij,k}$ assume valore unitario se le unità presentano la stessa modalità e zero altrimenti, mentre $\delta_{ij,k} = 1$
- variabili qualitative dicotomiche: i valori della similarità e della variabile indicatrice dell'ammissibilità del confronto si ottengono dalla tabella seguente:

$c_{ij,k}$	Unità i		$\delta_{ij,k}$	Unità i	
Unità j	1	0	Unità j	1	0
1	1	0	1	1	1
0	0	0	0	1	0

Esempio Confronto tra tre modelli di automobile

Automobile	Cilindrata	Airbag	Stereo di serie	Paese di fabbr.
i	1000	1	1	D
j	1500	0	1	I
r	750	0	0	I

$$c_{ij} = \left[1 - \frac{|1000 - 1500|}{1500 - 750} + 0 + 1 + 0 \right] / (1 + 1 + 1 + 1) \approx 0.33$$

$$c_{jr} = \left[1 - \frac{|1500 - 750|}{1500 - 750} + 0 + 0 + 1 \right] / (1 + 0 + 1 + 1) \approx 0.33$$

4.7 Calcolo della matrice di distanza in R

Per il calcolo della matrice di distanze in R si impiega la funzione

```
dist(x, method = "euclidean", diag = FALSE, upper = FALSE)
```

appartenente alla libreria *mva*, dove

- `x` rappresenta una matrice di dati o un data frame
- `method` seleziona la misura di distanza. Le opzioni disponibili sono: la distanza euclidea (`euclidean`) che costituisce l'opzione di default, la distanza di Lagrange (`maximum`), della città a blocchi (`manhattan`), e la distanza di Canberra (`canberra`), definita $d_{ij} = \sum_h (|x_{ih} - y_{jh}| / |x_{ih} + y_{jh}|)$. Per variabili dicotomiche è disponibile soltanto la distanza di Jaccard (`binary`).
- `diag` e `upper` sono poste uguale a `TRUE` se si desidera la matrice di distanza contenga anche i valori nulli sulla diagonale e i valori del triangolo superiore. Per default il risultato è una matrice triangolare inferiore senza la diagonale principale.

Per la standardizzazione delle variabili originarie è disponibile la funzione

```
scale(x, center = TRUE, scale = TRUE)
```

Con riferimento al data set `mtcars` di R costruiamo la matrice delle distanze euclidee tra le prime 5 unità sulla base dei primi 7 attributi:

```
>library(mva) # accesso alla libreria mva
>data(mtcars) # accesso al data set mtcars
>help(mtcars) # descrizione del data set
>x <- mtcars[1:5,1:7] # selezione delle unita' e delle variabili
>d <- dist(scale(x))
>d
```

	Mazda RX4	Mazda RX4 Wag	Datsun 710	Hornet 4 Drive
Mazda RX4	0.0000000	0.7219648	2.7264138	3.4904700
Mazda RX4 Wag	0.7219648	0.0000000	2.681264	3.014230
Datsun 710	2.7264138	2.6812637	0.000000	3.628464
Hornet 4 Drive	3.4904700	3.0142299	3.628464	0.000000
Hornet Sportabout	4.3906444	4.1625262	6.266534	3.814147

```
> as.matrix.dist(d)
```

	Mazda RX4	Mazda RX4 Wag	Datsun 710	Hornet 4 Drive	H. Sportabout
Mazda RX4	0.0000000	0.7219648	2.726414	3.490470	4.390644
Mazda RX4 Wag	0.7219648	0.0000000	2.681264	3.014230	4.162526
Datsun 710	2.7264138	2.6812637	0.000000	3.628464	6.266534
Hornet 4 Drive	3.4904700	3.0142299	3.628464	0.000000	3.814147
Hornet Sportabout	4.3906444	4.1625262	6.266534	3.814147	0.000000

Con riferimento al calcolo della distanza per caratteri nominali dicotomici, si consideri il seguente esempio:

```
x.i <- c(1,1,0,1,1,0,1,0,0)
x.j <- c(1,0,1,1,0,1,1,1,0)
table(x.j,x.i)
d.jacc <- dist(rbind(x.i,x.j), method="binary")
```

Per il calcolo della matrice di dissimilarità per caratteri misti, basata sull'indice di Gower, è disponibile la funzione `daisy` nella libreria `cluster`.