Introduzione all'analisi statistica dei dati multivariati

Novembre 1997

Giovanni M. Marchetti

Dipartimento di Statistica — Università di Firenze

Indice

Introduzione: tecniche elementari

1

1.1 Analisi univariate e multivariate

Quasi nessun problema statistico e' caratterizzato da una sola variabile. I fenomeni oggetto di studio sono spesso il risultato di molteplici elementi concomitanti che non e' possibile controllare. Col termine *analisi multivariata* si indica quell'insieme di metodi statistici usati per analizzare simultaneamente piu' caratteri. L'esistenza di molte variabili interagenti l'una con l'altra complica alquanto l'analisi rispetto all'ideale caso univariato. Le procedure statistiche univariate possono essere generalizzate, ma la complessita' aumenta sempre piu' all'aumentare delle dimensioni del problema.

Fanno parte dell'analisi multivariata molte tecniche diverse, usate per risolvere problemi anche lontani fra loro. Pertanto e' utile all'inizio illustrare con degli esempi alcune fra le situazioni piu' comuni in cui e' opportuno ricorrere ai metodi statistici multivariati.

In tutte le analisi statistiche multivariate il materiale grezzo e' costituito da un certo numero di caratteri che si vogliono studiare simultaneamente. L'analisi e' detta multivariata perche' vi sono piu' variabili oggetto di studio e non una sola. Tuttavia, gli scopi possono differire alquanto.

In alcuni casi l'obbiettivo dell'analisi e' semplicemente quello di classificare le unita' statistiche sulla base di tutte le variabili considerate. L'intento e' in questo caso puramente descrittivo e volto a scoprire l'esistenza di eventuali gruppi di unita'.

In altri casi si e' interessati piuttosto a ridurre le dimensioni della variabile multipla considerata in modo da riuscire a semplificare l'interpretazione. Talvolta questo e' possibile perche' certe variabili sono fra loro correlate, altre volte perche' esse sono indipendenti una volta eliminato l'effetto di altre.

Usualmente i dati si presentano sotto forma di una tabella $n \times p$, dove n sono le unita' statistiche e p i caratteri studiati, contenente le determinazioni di ogni variabile su ogni unita'. Quando vi sono caratteri qualitativi, spesso le n unita' vengono classificate in tavole di contingenza multiple.

Naturalmente, e' importante distinguere i metodi di analisi per dati quantitativi (tabelle di misure) dai metodi di analisi per dati qualitativi.

Occorre inoltre tener presente che molto spesso e' possibile distinguere tra i caratteri quelli che possiamo considerare *dipendenti* e quelli che invece sono *esplicativi* nel senso che in qualche modo li consideriamo antecedenti logici degli altri.

Vi sono alcuni problemi in cui si isola un unico carattere oggetto di studio studiandone la dipendenza dagli altri caratteri considerati esplicativi. Questi possono essere semplicemente dei caratteri che stratificano la popolazione (come per esempio il sesso) oppure caratteri che si considerano potenzialmente responsabili delle variazioni del carattere dipendente, e quindi causali in senso lato. In questi casi lo strumento tipico di analisi statistica e' la *regressione.* Si e' soliti distinguere la *regressione semplice* (se vi e' una sola variabile esplicativa) dalla *regressione multipla* (se vi sono due o piu' variabili esplicative).

A rigore, essendo unica la variabile dipendente, la regressione e' un'analisi di tipo univariato, ma essa e' complicata dall'esistenza di variabili esplicative che possono essere anche numerose. Tuttavia e' possibile generalizzare quanto detto sopra considerando piu' di una variabile dipendente (si tratta della *regressione multipla multivariata*).

Si osservi infine che le tecniche di regressione cambiano radicalmente a seconda che il carattere dipendente sia quantitativo o qualitativo.

1.2 Analisi multivariata: alcuni esempi

Esempio 1.2 Nella tabella 1.1 sono riportati alcuni dati sulla delinquenza in 16 citta' americane nel 1970. Le variabili sono costituite dalle 7 tipologie di delinquenza e sono espresse come rapporti per 100000 abitanti. Si osservi che la tabella non e' una tavola di contingenza doppia, ma una matrice di misure che raccoglie le determinazioni di 7 varibili su 16 unita' statistiche. Uno dei possibili obbiettivi di un analisi di questi dati e' quello di studiare le associazioni tra le variabili e le somiglianze tra le citta'. La presenza di 7 variabili rende questa analisi relativamente complessa. L'analisi delle singole variabili separatamente e' del tutto insufficiente e le forme di associazione studiabili sono parecchie. Pertanto sono particolarmente utili quelle tecniche che riescono a semplificare l'analisi riducendo le dimensioni.

In questo esempio, le unita' statistiche sono essenzialmente uniche e lo studio delle somiglianze tra di esse diventa importante. Ci si puo' chiedere pertanto quali siano le citta' simili sotto il profilo di tutte le variabili considerate.

Esempio 1.2 Un altro tipico esempio di dati multivariati si incontra nell'analisi delle tabelle di contingenza: in questo caso si studiano simultaneamente piu' caratteri qualitativi. Nella tabella 1.2 e' riportata una tabella di contingenza riguardante il numero di furti secondo il sesso del ladro, l'eta' e il tipo di merce rubata, in un grande magazzino olandese, tra il 1978 e il 1979. Ne risulta una tabella di contingenza tripla $2 \times 13 \times 9$.

	Omicidi	Stupri	Rapine	Aggressioni	Furti	Truffe	Furti d'auto
Atalanta	16.5	24.8	106	147	1112	905	494
Boston	4.2	13.3	122	90	982	669	954
Chicago	11.6	24.7	340	242	808	609	645
Dallas	18.1	34.2	184	293	1668	901	602
Denver	6.9	41.5	173	191	1534	1368	780
Detroit	13.0	35.7	477	220	1566	1183	788
Hartford	2.5	8.8	68	103	1017	724	468
Honolulu	3.6	12.7	42	28	1457	1102	637
Houston	16.8	26.6	289	186	1509	787	697
Kansas City	10.8	43.2	255	226	1494	955	765
Los Angeles	9.7	51.8	286	355	1902	1386	862
New Orleans	10.3	39.7	266	283	1056	1036	776
New York	9.4	19.4	522	267	1674	1392	848
Portland	5.0	23.0	157	144	1530	1281	488
Tucson	5.1	22.9	85	148	1206	756	483
Washington	12.5	27.6	524	217	1496	1003	739

Fonte: Hartigan (1975)

Tabella 1.1: *Tassi di delinquenza in 16 citta' americane.*

I caratteri oggetto di studio sono due caratteri qualitativi e un carattere quantitativo raggruppato in classi.

Si puo' osservare che le tre variabili sono senz'altro associate fra loro, ma che e' difficile stabilire la struttura dell'associazione. Pertanto occorre ridurre la complessita' della tabella modellando per esempio gli scarti dalla situazione di indipendenza stocastica.

1.3 Analisi di regressione: alcuni esempi

Esempio 1.3 Su un campione di 24 bambini nati in un ospedale di cui 12 maschi e 12 femmine, si considerano le due variabili X, la durata stimata della gestazione (in settimane) e Y, il peso alla nascita (in grammi). I dati raccolti sono riportati nella tabella 1.3. L'esame diretto di queste osservazioni rivela un legame crescente tra peso e durata della gestazione. La questione di interesse e' se il tasso di crescita sia lo stesso per i maschi e per le femmine. A prima vista il problema sembra si possa risolvere con due regressioni semplici separate tra Y e X nei due gruppi di bambini. In realta' cosi' procedendo non si riesce a stabilire se i tassi di crescita sono eguali nei due gruppi e quale sia l'effetto del sesso sul peso alla nascita. Si osservi che la variabile dipendente, il peso, e' quantitativa, mentre vi sono due caratteri esplicativi di cui uno, il sesso, e' qualitativo.

Esempio 1.3 Talvolta e' la variabile dipendente ad essere qualitativa. Nella tabella 1.4 sono riportati i risultati di uno studio americano su 1329 individui maschi (Ku e Kullback, 1974). Per ogni individuo sono state rilevati tre caratteri: (a) se ha avuto un infarto alle coronarie, (b) il livello di colesterolo (in mg/100 cc) e (c) la pressione del sangue (in mm). Il primo carattere e' binario (presenza o assenza dell'infarto) ed e'

Maschi	Età								
	< 12	12–14	15–17	18–20	21–29	30–39	40–49	50–64	65+
Vestiti	81	138	304	384	942	359	178	137	45
Vestiario	66	204	193	149	297	109	53	68	28
Tabacco	150	340	229	151	313	136	121	171	145
Penne	667	1409	527	84	92	36	36	37	17
Libri	67	259	258	146	251	96	48	56	41
Dischi	24	272	368	141	167	67	29	27	7
Casalinghi	47	117	98	61	193	75	50	55	29
Dolci	430	637	246	40	30	11	5	17	28
Giochi	743	684	116	13	16	16	6	3	8
Gioielli	132	408	298	71	130	31	14	11	10
Profumi	32	57	61	52	111	54	41	50	28
Hobbies	197	547	402	138	280	200	152	211	111
Altro	209	550	454	252	624	195	88	90	34

Femmine	Età								
	< 12	12–14	15–17	18–20	21–29	30–39	40–49	50–64	65+
Vestiti	71	241	477	436	1180	1009	517	488	173
Vestiario	19	98	114	108	207	165	102	127	64
Tabacco	59	111	58	76	132	121	93	214	215
Penne	224	346	91	18	30	27	23	27	13
Libri	19	60	50	32	61	43	31	57	44
Dischi	7	32	27	12	21	9	7	13	0
Casalinghi	22	29	41	32	65	74	51	79	39
Dolci	137	240	80	12	16	14	10	23	42
Giochi	113	98	14	10	12	31	8	17	6
Gioielli	162	548	303	74	100	48	22	26	12
Profumi	70	178	141	70	104	81	46	69	41
Hobbies	15	29	9	14	30	36	24	35	11
Altro	24	58	72	67	157	107	66	64	55

Fonte: van der Heijden, Falguerolles e de Leeuw (1989)

Tabella 1.2: *Numero di furti in un grande magazzino.*

Maschi		Femmine	
Età	Peso	Età	Peso
40	2968	40	3317
38	2795	36	2729
40	3163	40	2935
35	2925	38	2754
36	2625	42	3210
37	2847	39	2817
41	3292	40	3126
40	3473	37	2539
37	2628	36	2412
38	3176	38	2991
40	3421	39	2875
38	2975	40	3231

Tabella 1.3: *Dati sul peso alla nascita di 24 bambini.*

quello oggetto di studio. Avendo raggruppato in classi gli altri due si ottiene una tavola 4×4 in cui in ogni cella vi e' il numero di individui che hanno subito l'infarto sul totale degli individui della cella. Come detto in precedenza, questo tipo di dati richiede l'ap-

	Pressione del sangue			
Colesterolo	< 127	127–146	147–166	> 166
< 200	2/119	3/124	3/50	4/26
200–219	3/88	2/100	0/43	3/23
220–259	8/127	11/220	6/74	6/49
> 259	7/74	12/111	11/57	11/44

Fonte: Ku e Kullback (1974)

Tabella 1.4: *Infarti rilevati su 1329 individui.*

plicazione di metodi statistici diversi da quella dell'esempio precedente, pur trattandosi sempre di metodi di regressione.

1.4 Notazioni

In questo capitolo prenderemo in esame alcuni metodi per l'analisi preliminare di dati multivariati. Cercheremo di mantenere le notazioni piu' semplici possibile. L'utilizzazione di vettori e matrici permettera' di scrivere certe relazioni in modo compatto. Delle operazioni fra vettori la piu' utilizzata sara' quella di prodotto scalare. Sporadicamente compariranno anche il prodotto di matrici, l'inversa di una matrice quadrata e gli autovalori e autovettori di una matrice simmetrica. Queste nozioni non sono comunque strettamente necessarie per capire la maggior parte degli argomenti trattati.

Tutti i problemi elencati in precedenza ammettono la seguente trattazione schematica. Essi infatti riguardano insiemi di variabili $(X_1, \ldots, X_p)$ o di mutabili (caratteri qualitativi)

$(A_1, \ldots, A_q)$ rilevati sulle stesse n unita'. Questi insiemi di caratteri possono a loro volta essere suddivisi separando i caratteri dipendenti da quelli esplicativi. La trattazione e' qui semplificata perche' talvolta quest'ultima distinzione non e' univoca.

Prendiamo in considerazione una generica variabile X. Le n osservazioni relative ad X

$$(x_1, x_2, \ldots, x_n)$$

sono spesso denotate con un vettore (colonna) $(n \times 1)$ $\mathbf{x}$.

Tali osservazioni a volte sono considerate come l'elenco completo di tutte le determinazioni assunte dalla variabile in una popolazione finita, oppure come un insieme di realizzazioni di una o piu' variabili aleatorie (v.a.) In questo secondo caso l'n-upla osservata deriva da un n-upla di v.a.

$$(X_1, X_2, \ldots, X_n)$$

denotata con il vettore aleatorio $(n \times 1)$ $\mathbf{X}$.

1.4.1 Variabili multiple

Supponiamo ora di avere p variabili $X_1, \ldots, X_j, \ldots, X_p$ osservate sulle stesse n unita'. Consideriamo i vettori $(n \times 1)$ ad esse associati e indichiamoli con $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(j)}, \ldots, \mathbf{x}_{(p)}$. Il generico vettore variabile e'

$$\mathbf{x}_{(j)} = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}.$$

Con tali vettori e' possibile costruire una matrice $\mathbf{X}$ formata da p colonne (i vettori delle variabili) e da n righe:

$$\mathbf{X} = \left[\mathbf{x}_{(1)}|\mathbf{x}_{(2)}|\cdots|\mathbf{x}_{(p)}\right] = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}.$$

Questa matrice (talvolta detta semplicisticamente "matrice dei dati" come se tutti i dati dovessero per forza assumere tale forma di matrice) e' utile perche' permette di esprimere in forma compatta certe formule.

Ogni riga della matrice $\mathbf{X}$, che come si vede e' di dimensioni $(n \times p)$, contiene le determinazioni di variabili diverse osservate sull'unita' corrispondente a quella riga. Il vettore (riga) corrispondente all'unita' i-esima sara' indicato con $\mathbf{x}_i{}'$. Pertanto,

$$\mathbf{x}_i{}' = (x_{i1}, x_{i2}, \cdots, x_{ip}).$$

Possimo riassumere quanto detto con la relazione seguente

$$\mathbf{X} = \left[\mathbf{x}_{(1)}|\mathbf{x}_{(2)}|\cdots|\mathbf{x}_{(p)}\right] = \begin{bmatrix} \mathbf{x}_1{}' \\ \mathbf{x}_2{}' \\ \vdots \\ \mathbf{x}_n{}' \end{bmatrix}.$$

1.4.2 Medie, varianze e covarianze

Medie e varianze delle variabili statistiche sopra definite saranno indicate come d'abitudine con $\bar{x}_j$ e s_j^2 per $j = 1, \ldots, p$.

Una variabile $\mathbf{x}_{(j)}$ espressa in scarti dalla media ha come componenti $x_{ij} - \bar{x}_j$ e pertanto puo' essere scritta come

$$\mathbf{x}_{(j)} - \bar{x}_j \mathbf{1}$$

espressione in cui $\mathbf{1}$ e' un vettore di n dimensioni tutto composto di uno.

Le medie di tutte le variabili possono essere raccolte in un vettore di dimensione p che prende il nome di vettore delle medie (o centroide).

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}.$$

1.5 Prodotto scalare e ortogonalita'

Una operazione fondamentale tra due vettori $\mathbf{x}$ e $\mathbf{y}$ e' il prodotto scalare $\mathbf{x}'\mathbf{y} = \sum x_i y_i$, dove x_i e y_i sono le componenti dei due vettori.

Un concetto collegato e' quello di *ortogonalita'* dei vettori: due vettori $\mathbf{x}$ e $\mathbf{y}$ si dicono ortogonali ($\mathbf{x} \perp \mathbf{y}$) se il loro prodotto scalare e' uguale a zero. Questa definizione corrisponde all'usuale concetto di perpendicolarita' della geometria Euclidea come si puo' verificare rappresentando i vettori come frecce uscenti dall'origine nel piano Cartesiano monometrico.

Il concetto di ortogonalita' si applica direttamente alle variabili statistiche e ha un gran numero di utilizzazioni importanti. Un primo esempio e' quello delle variabili espresse in scarti dalla media. Come e' noto la somma delle determinazioni e' sempre zero. Cio' significa che se la variabile $\mathbf{x}$ e' espressa in scostamenti dalla sua media allora e' sempre ortogonale al vettore unitario $\mathbf{1}$: $\mathbf{x}'\mathbf{1} = \sum x_i = 0$.

Un'altra applicazione comunissima si ha nella misura dell'associazione tra due variabili. La covarianze tra due variabili X_j e $X_{j'}$ e' definita da

$$s_{jj'} = 1/n \sum_i (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$$

e quindi usando la notazione vettoriale risulta

$$s_{jj'} = 1/n (\mathbf{x}_{(j)} - \bar{x}_j \mathbf{1})'(\mathbf{x}_{(j')} - \bar{x}_{j'} \mathbf{1}).$$

La covarianza e' dunque il prodotto scalare (a meno di un fattore $1/n$) tra i vettori variabile espressi in scarti dalla media.

Se due variabili sono incorrelate, cioe' la loro covarianza e' zero, cio' equivale a dire che i vettori degli scarti dalla media sono ortogonali.

1.5.1 Teorema di Pitagora

Il concetto di ortogonalita' si ritrovera' piu' volte nel seguito e contribuira' sempre a semplificare le analisi statistiche. Il motivo fondamentale e' costituito dalla possibilita' di applicare il teorema di Pitagora, per cui se un vettore $\mathbf{y}$ (l'ipotenusa) e' la somma di due vettori ortogonali $\mathbf{x}_1$ e $\mathbf{x}_2$ (i cateti), allora la lunghezza al quadrato di $\mathbf{y}$ e' eguale alla somma dei quadrati delle lunghezze di $\mathbf{x}_1$ ed $\mathbf{x}_2$.

Definiamo la lunghezza di un vettore $\mathbf{y}$ come la somma dei quadrati dei suoi elementi

$$S(\mathbf{y}) = \sum y_i^2 = \mathbf{y}'\mathbf{y}.$$

Allora l'enunciato del teorema di Pitagora e' il seguente: se $\mathbf{x}_1 \perp \mathbf{x}_2$, allora,

$$S(\mathbf{y}) = S(\mathbf{x}_1) + S(\mathbf{x}_2)$$

la cui verifica algebrica e' immediata.

A titolo di esempio si consideri l'identita' ben nota secondo la quale

$$\sum (x_i - \bar{x})^2 + n(\bar{x} - a)^2 = \sum (x_i - a)^2.$$

Questa si dimostra usando il teorema di Pitagora tenendo presente che il vettore di componenti $(x_i - \bar{x})$ e' ortogonale al vettore di componenti $(\bar{x} - a)$ per un valore a qualsiasi e osservando che la loro somma e' uguale a $x_i - a$.

Si osservi infin che anche la lunghezza di un vettore ha un significato statistico poiche' la varianza di una variabile $\mathbf{x}_{(j)}$ e' pari alla lunghezza al quadrato della variabile in scarti dalla media divisa per n.

1.6 Trasformazioni

Ogni carattere quantitativo X puo' essere trasformato mediante una funzione monotona $g(X)$ in modo da facilitare l'analisi successiva. Esistono classi di trasformazioni per approssimare la normalita' della distribuzione di un carattere, oppure per migliorare la linearita' dell'associazione tra due caratteri. E' in generale difficile determinare una trasformazione ottimale per piu' di uno scopo.

Ovviamente le trasformazioni lineari sono le piu' semplici e quelle maggiormente usate. Una di queste e' la standardizzazione che ha la caratteristica di trasformare una variabile X in modo tale che la media sia zero e la varianza uno. La standardizzazione e' definita dalla seguente trasformazione delle determinazioni x_i:

$$z_i = g(x_i) = \frac{x_i - \bar{x}}{s}$$

in modo tale che le z_i sono espresse in termini di scarti quadratici medi dalla media.

Molti utilizzano la standardizzazione per rendere omogenee delle variabili che sono espresse in unita' di misura diverse. Infatti le determinazioni z_i sono dei numeri puri e quindi confrontabili per variabili diverse e per questo il procedimento e' talvolta consigliabile. Tuttavia esso comporta delle conseguenze che vanno tenute presente nelle applicazioni.

1.7 Matrici di covarianza e di correlazione

L'associazione tra due variabili $\mathbf{x}_{(j)}$ e $\mathbf{x}_{(j')}$ e' misurata dalla covarianza $s_{jj'}$. Si osservi che l'esame di tutte le covarianze delle distribuzioni doppie non esaurisce lo studio dell'associazione multipla tra variabili. Tuttavia, e' utile avere un oggetto che riassuma tutte le covarianze. La matrice simmetrica $\mathbf{S}$ avente come elementi le covarianze $s_{jj'}$, e' detta matrice di varianze e covarianze

$$\mathbf{S} = \begin{bmatrix} s_{11} & \cdots & s_{1j'} & \cdots & s_{1p} \\ \vdots & & \vdots & & \vdots \\ s_{j1} & \cdots & s_{jj'} & \cdots & s_{jp} \\ \vdots & & \vdots & & \vdots \\ s_{p1} & \cdots & s_{pj'} & \cdots & s_{pp} \end{bmatrix}.$$

Sulla diagonale principale vi sono le covarianze di ciascuna variabile con se stessa, cioe' le varianze.

La matrice di varianza e covarianza verifica la seguente identita' (facilmente dimostrabile)

$$\mathbf{S} = 1/n \sum (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = 1/n \sum \mathbf{x}_i \mathbf{x}_i{}' - \bar{\mathbf{x}}\bar{\mathbf{x}}'.$$

Naturalmente, se le variabili sono espresse in scarti dalla media, la formula precedente si semplifica diventando

$$\mathbf{S} = 1/n \sum \mathbf{x}_i \mathbf{x}_i{}' = 1/n \mathbf{X}'\mathbf{X}.$$

Oltre alla matrice delle varianze e covarianze si puo' introdurre anche la matrice di correlazione contenente tutti i coefficienti di correlazione $r_{jj'} = s_{jj'}/s_j s_{j'}$ tra coppie di variabili.

$$\mathbf{R} = \begin{bmatrix} r_{11} & \cdots & r_{1j'} & \cdots & r_{1p} \\ \vdots & & \vdots & & \vdots \\ r_{j1} & \cdots & r_{jj'} & \cdots & r_{jp} \\ \vdots & & \vdots & & \vdots \\ r_{p1} & \cdots & r_{pj'} & \cdots & r_{pp} \end{bmatrix}.$$

Come e' noto, $\mathbf{R}$ e' uguale alla matrice di varianze e covarianze tra le variabili standardizzate. Sulla diagonale principale le correlazioni di ciascuna variabile con se stessa, identicamente uguali a 1.

1.8 Un Esempio

Per esemplificare i concetti finora esposti utilizzeremo i dati della tabella 1.5 in cui sono riportati per ciascuno dei 50 stati nordamericani le variabili seguenti:
X_1: Stima della popolazione al 1 luglio del 1975 (in migliaia)
X_2: Reddito pro capite al 1974
X_3: Percentuale di analfabeti sulla popolazione
X_4: Vita media in anni (69–71)
X_5: Percentuale di crimini per 100000 abitanti (1976)
X_6: Percentuale di diplomati (1970)

X_7: Numero medio di giorni con la temperatura minima sotto zero nella capitale (1931–1960)
X_8: Area dello stato in miglia quadrate. La trattazione di questo esempio merita una piccola discussione. Applicando brutalmente le definizioni si ottiene il seguente vettore delle medie (sotto cui sono riportati gli scostamenti quadratici medi).

$$\begin{array}{rcrrrrrrrr} \bar{\mathbf{x}}' & = & (4246.42 & 4435.8 & 1.17 & 70.88 & 7.38 & 53.1 & 104.46 & 70735.88) \\ & & 4464.49 & 614.5 & 0.61 & 1.34 & 3.69 & 8.1 & 51.98 & 85327.30 \end{array}$$

La diretta applicazione di indici statistici ai dati grezzi non e' mai consigliabile. E' fondamentale infatti uno studio preliminare del problema che chiarifichi gli obbiettivi dell'indagine, le assunzioni ammissibili ecc. Non ci si deve dimenticare che si sta lavorando con variabili aventi diversa unita' di misura e diversa natura. Per esempio, X_1 e X_7 sono conteggi (ma il secondo e' una media), X_2, X_4 e X_8 sono misure (espresse in dollari, anni, miglia quadre) e infine le altre sono percentuali.

Che significato ha la media di percentuali? Per esempio, si osservi che la media delle percentuali di analfabetismo non corrisponde alla percentuale media di analfabeti costruita come totale di analfabeti su totale di popolazione, a meno che non si calcoli una media ponderata con pesi uguali alla popolazione.

Una considerazione ulteriore meritano le unita' statistiche di questo esempio. Infatti, le unita' statistiche sono costituite da zone geografiche e le variabili sono riferite a regioni che non hanno una definizione assoluta, ma convenzionale. Questo e' conosciuto come problema dell'unita' areale modificabile e si presenta appunto nelle situazioni in cui le variabili sono misurate non per una unita' ben definita, ma per una unita' che puo' essere variata a piacere. Per esempio potremmo rilevare la percentuale di analfabetismo anche a livello piu' disaggregato, diciamo di contea, o, al contrario, piu' aggregato. Questo fatto ha delle conseguenze: fra l'altro risulta che gli indici di associazione tra variabili, dipendono in modo sistematico dal livello geografico scelto. Ossia, per esempio in questo caso, il coefficiente di correlazione dipende dal livello di aggregazione e certe correlazioni che possono apparire a livello piu' aggregato possono sparire a livello disaggregato (vedi Arbia (1989)).

Infine le variabili sono state raccolte da fonti statistiche ufficiali e si presentano in uno stato grezzo senza riferimento a una particolare indagine che si vuole intraprendere.

Alcune variabili non sono di diretto interesse, ma servono per calcolare degli indicatori standardizzati solitamente piu' utili. Per esempio, l'area probabilmente non sara' utile direttamente, quanto per calcolare la densita' di popolazione.

Valutiamo ora le associazioni tra variabili calcolando la matrice di correlazione. Introducendo la densita' di popolazione ed eliminando la popolazione e l'area, tale matrice si presenta come nella tabella 1.6 (siccome e' simmetrica si e' riportato solo il triangolo inferiore). Dal suo esame emergono alcune correlazioni piu' evidenti, come quella (negativa) tra vita media e tasso di delinquenza e quella positiva tra questo e il tasso di analfabetismo. Tuttavia, e' necessario ricordare che tali coefficienti di correlazione sono coefficienti "lordi" nel senso che contengono anche tutte le influenze delle restanti variabili da cui non sono depurati. Vedremo piu' avanti parlando della regressione multipla quali sono le tecniche per depurare le variabili dall'influenza lineare delle altre.

Stato	Popolaz.	Reddito	Analf.	Vita	Crim.	Diplom.	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	51945
California	21198	5114	1.1	71.71	10.3	62.6	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	103766
Connecticut	3100	5348	1.1	72.48	3.1	56.0	4862
Delaware	579	4809	0.9	70.06	6.2	54.6	1982
Florida	8277	4815	1.3	70.66	10.7	52.6	54090
Georgia	4931	4091	2.0	68.54	13.9	40.6	58073
Hawaii	868	4963	1.9	73.60	6.2	61.9	6425
Idaho	813	4119	0.6	71.87	5.3	59.5	82677
Illinois	11197	5107	0.9	70.14	10.3	52.6	55748
Indiana	5313	4458	0.7	70.88	7.1	52.9	36097
Iowa	2861	4628	0.5	72.56	2.3	59.0	55941
Kansas	2280	4669	0.6	72.58	4.5	59.9	81787
Kentucky	3387	3712	1.6	70.10	10.6	38.5	39650
Louisiana	3806	3545	2.8	68.76	13.2	42.2	44930
Maine	1058	3694	0.7	70.39	2.7	54.7	30920
Maryland	4122	5299	0.9	70.22	8.5	52.3	9891
Massachus.	5814	4755	1.1	71.83	3.3	58.5	7826
Michigan	9111	4751	0.9	70.63	11.1	52.8	56817
Minnesota	3921	4675	0.6	72.96	2.3	57.6	79289
Mississippi	2341	3098	2.4	68.09	12.5	41.0	47296
Missouri	4767	4254	0.8	70.69	9.3	48.8	68995
Montana	746	4347	0.6	70.56	5.0	59.2	145587
Nebraska	1544	4508	0.6	72.60	2.9	59.3	76483
Nevada	590	5149	0.5	69.03	11.5	65.2	109889
New Hamp.	812	4281	0.7	71.23	3.3	57.6	9027
New Jersey	7333	5237	1.1	70.93	5.2	52.5	7521
New Mexico	1144	3601	2.2	70.32	9.7	55.2	121412
New York	18076	4903	1.4	70.55	10.9	52.7	47831
North C.	5441	3875	1.8	69.21	11.1	38.5	48798
North D.	637	5087	0.8	72.78	1.4	50.3	69273
Ohio	10735	4561	0.8	70.82	7.4	53.2	40975
Oklahoma	2715	3983	1.1	71.42	6.4	51.6	68782
Oregon	2284	4660	0.6	72.13	4.2	60.0	96184
Pennsylv.	11860	4449	1.0	70.43	6.1	50.2	44966
Rhode I.	931	4558	1.3	71.90	2.4	46.4	1049
South C.	2816	3635	2.3	67.96	11.6	37.8	30225
South D.	681	4167	0.5	72.08	1.7	53.3	75955
Tennessee	4173	3821	1.7	70.11	11.0	41.8	41328
Texas	12237	4188	2.2	70.90	12.2	47.4	262134
Utah	1203	4022	0.6	72.90	4.5	67.3	82096
Vermont	472	3907	0.6	71.64	5.5	57.1	9267
Virginia	4981	4701	1.4	70.08	9.5	47.8	39780
Washington	3559	4864	0.6	71.72	4.3	63.5	66570
West Virginia	1799	3617	1.4	69.48	6.7	41.6	24070
Wisconsin	4589	4468	0.7	72.48	3.0	54.5	54464
Wyoming	376	4566	0.6	70.29	6.9	62.9	97203

Fonte: Statistical abstract of the United States (1977),
County and City Data Book (1977), Bureau of the Census

Tabella 1.5: *Alcuni indicatori rilevati sui 50 stati americani.*

	Densità	Reddito	Analf.	Vita	Crim.	Dipl.
Densità	1.00					
Reddito	0.32	1.00				
Analf.	0.00	−0.43	1.00			
Vita	0.09	0.34	−0.58	1.00		
Crim.	−0.18	−0.23	0.70	−0.78	1.00	
Dipl.	−0.08	0.61	−0.65	0.58	−0.48	1.00

Tabella 1.6: *Matrice di correlazione.*

1.9 Analisi grafiche

Lo studio dell'associazione tra variabili e' facilitato da semplici rappresentazioni grafiche. Ci limiteremo qui alle rappresentazioni grafiche per variabili doppie. Esistono rappresentazioni grafiche per variabili multiple, ma queste risultano molto piu' difficili da interpretare e presentare. Se ci si limita a due dimensioni le rappresentazioni grafiche sono molto intuitive e potenti grazie alle capacita' interpretative dell'occhio umano. Una delle tecniche piu' comuni

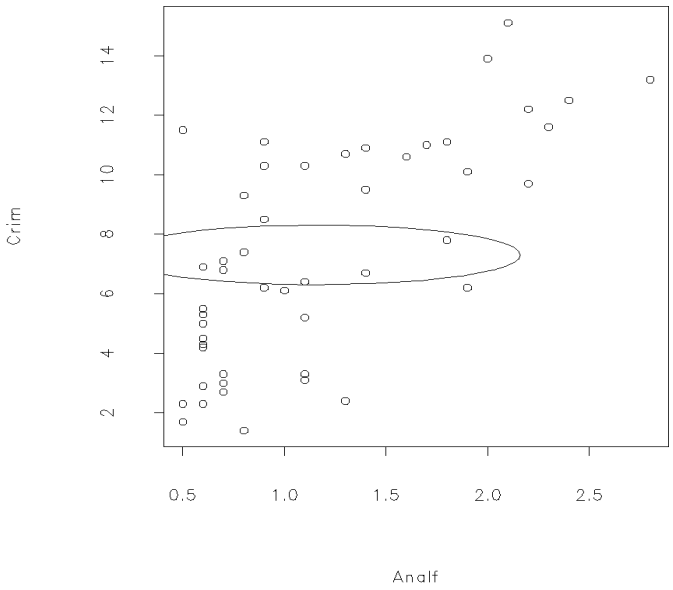

Figura 1.1: *Scatterplot sui dati grezzi*

e' quella dello *scatterplot* o grafico di dispersione dei punti $(x_{ij}, x_{ij'})$ relativi a due caratteri quantitativi j e j'. Nel grafico 1.1 e' riportato lo scatter relativo alle variabili Analfabetismo e Tasso di delinquenza (Crim) che mette in evidenza la correlazione positiva ($r = 0.7$) tra le due. Sul grafico abbiamo anche riportato una circonferenza centrata sul punto medio (il centroide) e avente raggio unitario. La circonferenza appare come un'ellisse perche' le scale delle ascisse e delle ordinate non sono le stesse. E' evidente che le distanze in verticale sono maggiori di quelle in orizzontale a causa della differenza delle scale e questo fatto e' sottolineato dalla

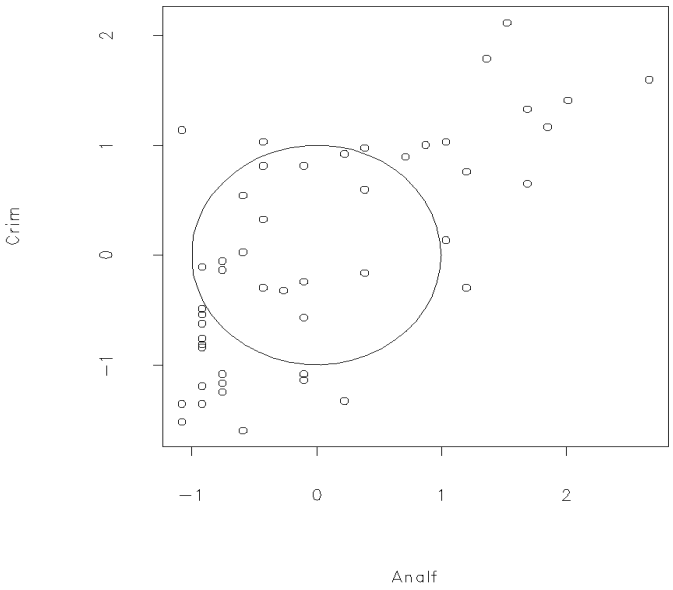

Figura 1.2: *Scatterplot sulle variabili standardizzate*

forma schiacciata della circonferenza.

Nel grafico 1.2 e' rappresentato invece lo scatter relativo alle stesse variabili standardizzate. Il grafico appare esattamente eguale a prima, eccezion fatta per le scale che sono cambiate. Anche su questo grafico si puo' sovrapporre una circonferenza centrata nelle medie e di raggio unitario. Le medie delle variabili standardizzate sono entrambe nulle e pertanto la circonferenza e' centrata nell'origine. Inoltre essa sul grafico non appare piu' schiacciata perche' le scale sono le stesse sia in ascisse che in ordinate. In altre parole, uno spostamento di un cm in orizzontale corrisponde allo spostamento di un cm in verticale.

Si osservi che il sistema di distanze tra i punti cambia anche se i grafici apparentemente sono uguali (a parte le scale). Dei punti che prima distavano relativamente nella scala verticale, dopo la standardizzazione (che riaggiusta tale scala a quella orizzontale) risultano piu' vicini. L'intero sistema di distanze e' cambiato anche se l'occhio coglie l'identica struttura delle unita'. L'apparente contraddizione si dissolve considerando il cambiamento delle scale del quale l'occhio non sa tener conto perche assume che lo spazio sia isotropo.

1.10 Caratteri qualitativi

Se vi sono dei caratteri qualitativi, questi non possono essere trattati nella forma vettoriale esattamente come se fossero caratteri quantitativi. Le diverse proprieta' dei caratteri si devono tradurre in una diversa struttura algebrica.

Un carattere qualitativo che assume un numero finito di modalita' non numeriche si dice mutabile. Le sue modalita' si dicono talvolta livelli della mutabile. Ci limiteremo al caso in cui i caratteri qualitativi siano di tipo sconnesso cioe' con modalita' prive di ordinamento.

Supponiamo che A sia una mutabile con 4 livelli e che essa sia rilevata su $n = 6$ unita'

fornendo la seguente successione di determinazioni

$$(A_2, A_1, A_1, A_2, A_4, A_3)$$

Questa mutabile puo' essere rappresentata con 4 vettori indicatori $\mathbf{a}_{(1)}$, $\mathbf{a}_{(2)}$, $\mathbf{a}_{(3)}$, $\mathbf{a}_{(4)}$, come segue

Unita'	$\mathbf{a}_{(1)}$	$\mathbf{a}_{(2)}$	$\mathbf{a}_{(3)}$	$\mathbf{a}_{(4)}$
1	0	1	0	0
2	1	0	0	0
3	1	0	0	0
4	0	1	0	0
5	0	0	0	1
6	0	0	1	0

Il vettore $\mathbf{a}_{(1)}$ e' un indicatore della modalita' A_1 della mutabile (che e' rilevata sulla seconda e terza unita'). Analogo e' il significato degli altri vettori di indicatori.

La procedura si generalizza in modo ovvio a un numero qualsiasi di mutabili e di modalita'. Una regola evidente e' che la somma dei vettori indicatori e' sempre eguale al vettore $\mathbf{1}$. Inoltre la somma degli elementi dell'indicatore e' eguale alla frequenza marginale della modalita' del carattere. Per esempio la somma degli elementi di $\mathbf{a}_{(1)}$ e' eguale a 2 che e' la frequenza associata a A_1.

> Talvolta, gli indicatori delle modalita' si riuniscono in una matrice di indicatori $\mathbf{A} = (\mathbf{a}_{(1)}, \mathbf{a}_{(2)}, \mathbf{a}_{(3)}, \mathbf{a}_{(4)})$. Questa notazione ha alcuni vantaggi. Supponiamo per esempio di voler calcolare le medie di una variable $\mathbf{y}$ per ogni classe di una mutabile caratterizzata dalla matrice di indicatori $\mathbf{A}$. L'espressione $\mathbf{A}'\mathbf{y}$ fornisce il vettore dei totali della variabile $\mathbf{y}$ per ogni livello della mutabile. D'altra parte il prodotto $\mathbf{A}'\mathbf{A}$ e' una matrice diagonale contenente sulla diagonale le frequenze marginali della mutabile.
>
> Pertanto il vettore delle medie parziali di $\mathbf{y}$ e' dato dalla divisione di ogni elemento di $\mathbf{A}'\mathbf{y}$ per le frequenze marginali, e dunque e' semplicemente
>
> $$\mathbf{m} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y}.$$

1.10.1 Tavole di contingenza

Nel caso in cui si debbano analizzare due o piu' mutabili queste spesso sono classificate in *tavole di contingenza.*

Consideriamo la tabella seguente contenente la distribuzione della popolazione italiana secondo la ripartizione territoriale ed il livello di istruzione (fonte: ISTAT, frequenze in migliaia):

	Laurea	Diploma	Media	Elementare	Senza Titolo	Analfabeti
Nord	66	282	629	1056	358	23
Centro	36	136	239	395	180	23
Mezzogiorno	47	184	380	676	416	114

Si tratta di una tavola di contingenza doppia che raccoglie le frequenze congiunte associate ad ogni modalita' di istruzione e di ripartizione territoriale. Indichiamo con n_{ij} le frequenze congiunte e con f_{ij} le frequenze congiunte relative uguali alle precedenti divise per il numero di unita'. Otteniamo la tabella seguente

	Laurea	Diploma	Media	Elementare	S.T.	Analf.	Tot. (f_{i+})
Nord	1.26	5.38	12	20.1	6.83	0.44	46.1
Centro	0.69	2.6	4.56	7.54	3.44	0.44	19.3
Mezz.	0.9	3.51	7.25	12.9	7.94	2.18	34.7
Tot. (f_{+j})	2.84	11.5	23.8	40.6	18.2	3.05	100

Si osservi che alla tavola delle frequenze congiunte (espresse in forma percentuale) abbiamo aggiunto le frequenze marginali ottenute calcolando i totali di riga e di colonna (indicate con f_{i+} e f_{+j}).

Molto utile anche le tabella delle frequenze condizionate, cioe' la tabella dei profili riga o colonna, ottenute scalando la tabella data con i totali di riga e di colonna. Per esempio la tabella dei profili riga e' la seguente

	Laurea	Diploma	Media	Elementare	S. T.	Analf.	
Nord	2.73	11.7	26.1	43.7	14.8	0.95	100
Centro	3.57	13.5	23.7	39.1	17.8	2.28	100
Mezz.	2.59	10.1	20.9	37.2	22.9	6.27	100
Tot. (f_{+j})	2.84	11.5	23.8	40.6	18.2	3.05	100

Ogni riga rappresenta una distribuzione del livello di istruzione condizionata alle tre ripartizioni territoriali.

Le distribuzioni condizionate sono collegate alla distribuzione marginale da una regola fondamentale: la frequenza marginale e' una media ponderata delle frequenze condizionate con pesi uguali alle *altre* frequenze marginali, cioe'

$$f_{+j} = \sum_i \frac{f_{ij}}{f_{i+}} f_{i+}.$$

Ad esempio, la frequenza marginale di laureati (0.0284) si puo' ottenere dalle frequenze condizionate $(0.0273, 0.0357, 0.0259)$ di laureati nelle tre ripartizioni, facendone la media ponderata:

$$0.0284 = 0.0273 \times 0.461 + 0.0357 \times 0.193 + 0.0259 \times 0.347.$$

Si puo' osservare che, essendo medie, le frequenze marginali sono sempre comprese nel campo di variazione delle corrispondenti frequenze condizionate.

Se la distribuzione marginale e' uguale alle distribuzioni condizionate i due caratteri studiati si dicono *indipendenti*. L'associazione tra i due caratteri qualitativi si studia infatti esaminando le differenze tra le distribuzioni condizionate e la distribuzione marginale.

1.10.2 Rappresentazioni grafiche

Esistono delle utili rappresentazioni grafiche anche per coppie di caratteri qualitativi, ma queste differiscono sensibilmente dalle corrispondenti rappresentazioni grafiche per caratteri quantitativi. Spesso si utilizza un grafico a barre come quello illustrato nel grafico 1.3 Il

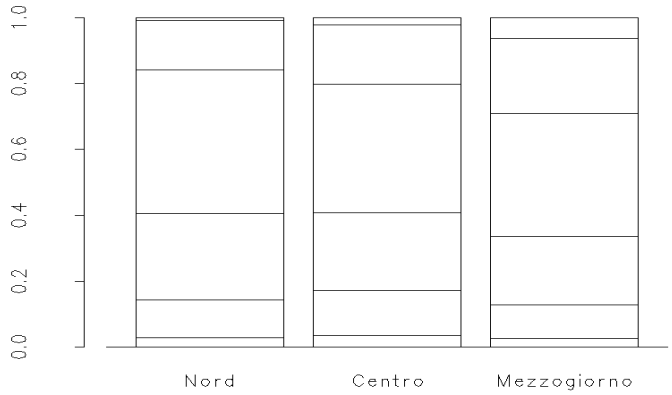

Figura 1.3: *Grafico a barre suddivise. Cattiva percezione*

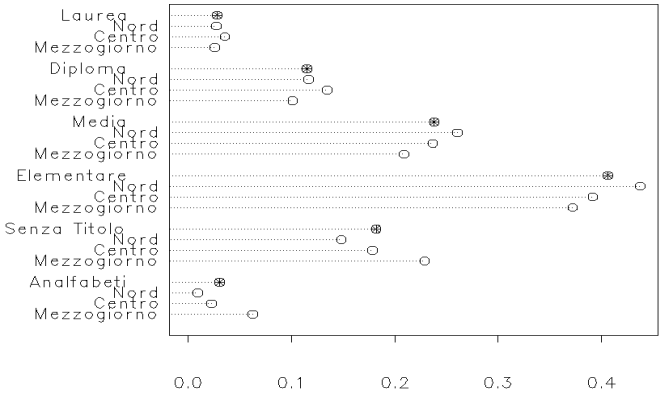

Figura 1.4: *Diagrammi affiancati. Buona percezione*

grafico e' costituito da barre di lunghezza unitaria divise in fasce di lunghezza proporzionale alle frequenze condizionate.

Questo tipo di grafico e' sconsigliabile nel caso in cui le modalita' del carattere oggetto di studio siano piu' di due, come in questo caso in cui si sono rappresentate le distribuzioni condizionate del livello di istruzione. Infatti, mentre e' possibile un confronto delle frequenze della prima ed ultima modalita' (perche' sono allineate), e' difficile invece confrontare i livelli intermedi.

Nel grafico occorrerebbe aggiungere un diverso tratteggio o diversi colori per le varie modalita', e una legenda per individuarle. Anche cosi' il grafico e' spesso di difficile interpretazione.

Un grafico migliore si ottiene invece affiancando dei diagrammi lineari costruiti come per le distribuzioni semplici. Il grafico 1.4 illustra come procedere correttamente. Si noti che per ogni livello di istruzione si rappresentano le frequenze condizionate (indicate con un cerchietto vuoto) e la frequenza marginale (indicata con un cerchietto pieno). Quindi si ottengono sei diagrammi sovrapposti. Non si tratta pero' dei diagrammi in serie delle sei distribuzioni

condizionate, infatti, si puo' notare che in ciascuno la somma delle frequenze non e' uguale a uno. Invece, le distribuzioni condizionate vengono rappresentate, per cosi' dire, *in parallelo*, in modo da agevolare il confronto tra le frequenze condizionate e la loro media marginale.

Pertanto, si percepisce subito il rapporto esistente tra la percentuale di individui che hanno il diploma della media inferiore nelle tre ripartizioni e il totale nazionale. Si vede ad esempio che si va dal 20% del Mezzogiorno al 26% del Nord, mentre il Centro ha un valore prossimo a quello medio nazionale. Queste informazioni erano ovviamente ricavabili anche dalla tabella dei profili riga, ma non si potevano percepire facilmente dal grafico 1.3.

1.11 Note bibliografiche

Gli argomenti trattati in questo e nei seguenti capitoli sono sviluppati con maggior dettaglio in tutti i libri di statistica multivariata: si veda, per esempio, Fabbris (1990). Tra i manuali in lingua inglese si possono indicare, tra gli altri, Mardia, Kent e Bibby (1979) e Seber (1984).

Una lettura fondamentale per approfondire le distinzioni fra vari tipi di indagine statistica e‘ Cox e Snell (1981) che presenta altresi' una ampia raccolta di esempi svolti e problemi reali.

L'assimilazione delle variabili statistiche con vettori ad n componenti e l'uso del formalismo dell'algebra lineare per descrivere l'associazione tra caratteri e' tipica della cosiddetta scuola francese di analisi dei dati. Si veda per esempio Lebart, Morineau e Warwick (1984).

Un testo introduttivo ai metodi grafici in statistica multivariata e' Chambers, Cleveland, Kleiner & Tukey (1983). Un'altra lettura interessante e' Barnett (1981), una raccolta di saggi (alcuni facili, altri piuttosto difficili) da cui si puo' avere un'idea degli sviluppi dei metodi grafici per rappresentare dati multidimensionali. Un campo di ricerca collegato e' quello dei grafici dinamici (cfr. Cleveland e McGill (1988)).

Per approfondire lo studio dei caratteri qualitativi, si puo' far riferimento a Zanella (1988).

Metodi di classificazione

2

2.1 Introduzione

Ci occuperemo ora di un problema fondamentale dell'analisi di dati multivariati, quello della classificazione delle unita' statistiche. In molti esempi introdotti in precedenza, uno degli obbiettivi principali dell'indagine e' quello del raggruppamento delle unita' in classi omogenee sulla base di *tutti* i caratteri considerati. L'attenzione e' puntata soprattutto sulle unita' statistiche che spesso non sono viste come elementi di un campione ma come essenzialmente uniche.

Se da un parte si tratta di una esigenza molto sentita da parte dei ricercatori, d'altro lato e' difficile formalizzare esattamente il problema che si presenta in modo alquanto indefinito. La difficolta' fondamentale e' che cosa si debba intendere come gruppo. Daremo pertanto alcune indicazioni generali prima di considerare in dettaglio alcune tecniche particolarmente utili.

1. I gruppi dovrebbero essere insiemi di unita' da un lato piu' omogenei possibile e, dall'altro piu' separati possibile. Si tratta di semplificare una realta' complessa costituendo gruppi di unita' vicine tra loro. Cio' suggerisce di introdurre degli *indici di distanza* in modo da precisare la nozione di vicinanza e di omogeneita'.

2. Non viene impiegata una classificazione *a priori*, ossia non si sa nulla sulle classi, neanche per una parte delle unita'. I gruppi sono incogniti sia dal punto di vista delle unita' in essi contenute, sia — nel caso piu' generale — quanto al loro numero.

3. Ogni unita' e' caratterizzata da p osservazioni su altrettante variabili o mutabili e, nella ricerca dei gruppi, si vuol tener conto di tutti i caratteri considerati.

Nonostante (o proprio per) le difficolta' di definizione del problema esistono moltissime procedure che consentono di raggruppare unita' e di formare classi e che pertanto vengono dette di *analisi dei gruppi*. Alcune di queste sono veramente utili anche come strumento generale di analisi descrittiva dei dati multivariati.

2.2 Le fasi dell'analisi dei gruppi

Data la grande varieta' delle procedure di analisi dei gruppi, e' importante saper individuare gli aspetti fondamentali di ogni metodo, tenendo presente che ciascuno ha delle caratteristiche che lo rendono opportuno in certe situazioni e non in altre. Ogni procedura in realta' e' il risultato di diverse scelte operate in relazione a 3 punti fondamentali.

1. *Operazioni preliminari.* Scelta delle unita' e dei caratteri. Loro trasformazione, omogeneizzazione. Ponderazione delle unita' e delle variabili.

2. *Indici di prossimita'.* Scelta di un indice di somiglianza o di distanza tra coppie di unita'.

3. *La costruzione dei gruppi.* Scelta dell'impostazione da adottare, della struttura delle classi, del criterio da ottimizzare.

I punti piu' importanti ai fini della caratterizzazione di un metodo di analisi dei gruppi sono il secondo e il terzo. Alcune scelte sono, ovviamente, collegate ad altre. Ad esempio, la scelta della ponderazione delle variabili, come vedremo, e' collegata alla scelta di una misura di prossimita', e le due scelte si influenzano reciprocamente.

2.3 Operazioni preliminari

2.3.1 Definizione e scelta delle unita'

La definizione e scelta delle unita' e' un problema fondamentale di ogni indagine statistica che condiziona ogni risultato seguente. E' importante distinguere due casi. Nel primo, il problema di base e' quello di scoprire qualche struttura in una popolazione completa, senza necessita' alcuna di estrapolare i risultati a una sovra popolazione.

Nel secondo caso, vi e' la necessita' di estendere i risultati e di effettuare delle inferenze, mentre l'analisi e' compiuta su un campione opportunamente scelto. Nell'esposizione seguente trascureremo questi problemi concentrandoci sulle analisi descrittive.

Collegato al problema della scelta delle unita' vi e' la possibilita' di pondererare le unita' stesse attribuendo un peso sulla base di varie considerazioni.

2.3.2 Scelta dei caratteri e ponderazione

Poiche' il raggruppamento si fonda sui caratteri presi in considerazione e quindi anche l'omogeneita' o diversita' dei gruppi e' definita in termini degli stessi caratteri, e' evidente l'importanza di questa scelta.

Le ponderazioni dei caratteri si possono distinguere in due tipi.

(a) Ponderazioni esplicite, quando cioe' le variabili vengono ponderate *a priori* per dare piu' importanza, per esempio, alle variabili fortemente collegate al fenomeno studiato.
(b) Ponderazioni implicite, quando i caratteri studiati risultano avere di per se' un peso diverso, per esempio perche' hanno varianze diverse o perche' essendo correlate essi in realta' misurano per cosi' dire la stessa cosa.

Quando non si vuole che vi siano ponderazioni implicite, ma che tutti i caratteri abbiano lo stesso peso, allora occorre riponderarli in modo da eliminare le disuguaglianze.

2.3.3 Omogeneizzazione delle scale

Spesso i caratteri rilevati sono su scale diverse. In questi casi taluni preferiscono rendere omogenee le scale prima di procedere alla classificazione, in modo da poter lavorare su dati tutti dello stesso tipo. Il problema della trasformazione delle scale e' tuttavia formidabile. Se e' relativamente facile passare da caratteri quantitativi a caratteri qualitativi, sacrificando informazione, il passaggio inverso e l'utilizzazione mista di caratteri qualitativi resi quantitativi con caratteri quantitativi originari sembra un operazione molto piu' discutibile. In seguito discuteremo un modo per trattare contemporaneamente caratteri qualitativi e quantitativi usando misure opportune di somiglianza tra caratteri.

2.4 Indici di distanza

L'omogeneita' dei gruppi puo' essere valutata tramite una misura del grado di vicinanza tra le unita' detta indice di prossimita'. Prossimita' e' un termine generico che serve per denotare indifferentemente o un *indice di somiglianza* o un *indice di distanza* tra unita'. Cominceremo con una trattazione astratta del concetto di somiglianza e distanza.

Date due unita' generiche i e i', si dice indice di somiglianza una funzione $s(i, i')$ a valori reali che gode delle seguenti proprieta':
(i) $0 \leq s(i, i') \leq 1$
(ii) $s(i, i') = s(i', i)$
(iii) $s(i, i) = 1$
Inoltre, $s(i, i') > s(i, i'')$, implica che i e' piu' vicina a i' che a i''. Quindi tanto maggiore e' l'indice di somiglianza e tanto piu' vicine sono le unita'.

Invece un indice di distanza e' una funzione $d(i, i')$ a valori reali tale che
(i) $d(i, i') \geq 0$
(ii) $d(i, i') = d(i', i)$
(iii) $d(i, i') = 0$ se e solo se le due unita' i e i' hanno le stesse determinazioni dei caratteri.
Inoltre $d(i, i') > d(i, i'')$ significa che l'unita' i e' piu' vicina a i'' che a i', cioe' tanto maggiore e' l'indice e tanto piu' lontane sono le unita'.

Un indice di distanza si dice poi una *metrica* se soddisfa alla disuguaglianza triangolare: date tre unita' qualsiasi i, i' e i'' risulta sempre che

$$d(i, i') \leq d(i', i'') + d(i'', i)$$

cioe' la distanza che intercorre tra due punti e' sempre minore della somma delle distanze tra tali punti e un terzo punto. Questa proprieta' che e' naturalissima nella nostra percezione delle distanze spaziali, non e' sempre verificata per certi indici di distanze in spazi astratti.

Infine talvolta una metrica $d(i,i')$ gode di una ulteriore proprieta' ancora piu' forte, la cosiddetta *disuguaglianza ultrametrica*: date tre unita' qualsiasi i, i' e i''

$$d(i,i') \leq \max\{d(i',i''), d(i'',i)\}.$$

In questo caso la distanza si dice distanza ultrametrica. La disuguaglianza ultrametrica richiede che la massima distanza tra l'unita' i'' e la coppia di unita' (i,i') non possa mai scendere al di sotto della distanza che separa i e i'.

Si osservi che se $d(i,i')$ e' una distanza ultrametrica allora e' automaticamente una metrica, perche' la disuguaglianza ultrametrica implica la disuguaglianza triangolare.

2.4.1 Equivalenze tra indici di distanza

Ad ogni indice di prossimita' e' associato un *ordinamento* delle coppie di unita'. Per chiarire questo importante concetto facciamo un esempio.

Supponiamo di avere 5 unita' a, b, c, d ed e e che le distanze tra di esse siano le seguenti

	a	b	c	d	e
a	0	0,1	0,2	0,5	0,6
b		0	0,3	0,2	0,9
c			0	0,1	0,8
d				0	0,7
e					0

Si noti che per la proprieta' di simmetria degli indici di distanza la matrice sopra riportata e' simmetrica (per questo si sono omesse le distanze nel triangolo inferiore). Naturalmente le distanze sulla diagonale principale, che corrispondono alle distanze tra ogni unita' e se stessa, sono nulle.

Allora, e' possibile ordinare le distanze dalle piu' piccole alle piu' grandi e in questo modo ordinare anche le coppie di unita' ad esse legate:

Ordinamento per le coppie	distanza
$\{a,a\}\{b,b\}\{c,c\}\{d,d\}\{e,e\}$	0
$\{a,b\}\{c,d\}$	0,1
$\{a,c\}\{b,d\}$	0,2
$\{b,c\}$	0,3
$\{a,d\}$	0,5
$\{a,e\}$	0,6
$\{d,e\}$	0,7
$\{c,e\}$	0,8
$\{b,e\}$	0,9

Si osservi che ad ogni distanza corrisponde un insieme di coppie caratterizzate da quella distanza (sono *ex-aequo*). A due distanze di cui la prima e' minore dell'altra, corrispondono due classi di coppie di cui la prima contiene coppie piu' simili dell'altra. E quindi queste classi sono ordinate.

Il concetto di ordinamento associato a un indice di distanza (o di somiglianza) e' utile perche' permette di confrontare due indici diversi. Infatti, diremo che due indici di prossimita' sono *equivalenti* se gli ordinamenti associati ad essi sono identici quali che siano le unita'. Percio' due indici equivalenti danno luogo allo stesso ordinamento delle coppie di unita'.

Per esempio, se sulle stesse 5 unita' utilizziamo un secondo indice di distanza equivalente esso potrebbe dar luogo alle seguenti distanze:

	a	b	c	d	e
a	0	0,3	0,4	0,7	0,8
b		0	0,5	0,4	1,1
c			0	0,3	1,0
d				0	0,9
e					0

Come si vede l'ordinamento delle coppie e' lo stesso anche se le distanze non sono le stesse. Si osservi anche che il secondo insieme di distanze non e' ottenibile dal primo mediante una semplice trasformazione lineare.

2.5 La distanza euclidea

In questa e nelle prossime sezioni daremo qualche esempio di indice di prossimita', dando maggiore spazio agli indici di distanza.

Supponiamo di avere misurato p variabili $X_1, \ldots, X_p$ su n unita' e di disporre quindi dei vettori unita' $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ per ogni unita' i. La distanza piu' comunemente usata in questo caso e' la *distanza Euclidea* definita da

$$d(i, i') = d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}.$$

Questa distanza corrisponde alla usuale distanza tra punti nello spazio fisico. Si osservi invece che facendone uso in campo statistico essa combina scarti tra grandezze che possono essere espresse in unita' di misura diverse. La somma non ha quindi nessun significato a meno che le unita' di misura siano le stesse.

Il quadrato della distanza Euclidea e' esprimibile con il prodotto scalare, come segue

$$d(\mathbf{x}_i, \mathbf{x}_{i'})^2 = (\mathbf{x}_i - \mathbf{x}_{i'})'(\mathbf{x}_i - \mathbf{x}_{i'}).$$

Una generalizzazione di questa distanza e' la *distanza Euclidea ponderata* con pesi $w_1, \ldots, w_p$ che ha la formula seguente

$$d_w(i, i') = d_w(\mathbf{x}_i, \mathbf{x}_{i'}) = \sqrt{\sum_{j=1}^{p} w_j (x_{ij} - x_{i'j})^2}.$$

Anche in questo caso si puo' utilizzare una notazione vettoriale. Basta definire una matrice diagonale di pesi $\mathbf{D}_w = \mathrm{diag}(w_1, \ldots, w_p)$ per cui risulta

$$d_w(\mathbf{x}_i, \mathbf{x}_{i'})^2 = (\mathbf{x}_i - \mathbf{x}_{i'})'\mathbf{D}_w(\mathbf{x}_i - \mathbf{x}_{i'}).$$

Si osservi che si possono utilizzare i pesi w_j per neutralizzare le unita' di misura delle variabili. Se dal punto di vista dimensionale w_j ha una unita' di misura che e' il reciproco del quadrato di quella di X_j, allora il singolo scarto $w_j(x_{ij} - x_{i'j})^2$ e' un numero puro.

2.5.1 Proprieta' della distanza Euclidea

La distanza Euclidea e' una metrica, cioe' soddisfa alla disuguaglianza triangolare, ed inoltre gode delle due proprieta' seguenti:
(i) Invarianza per traslazione: se $\mathbf{a}$ e' un vettore qualsiasi

$$d(\mathbf{x}_i + \mathbf{a}, \mathbf{x}_{i'} + \mathbf{a}) = d(\mathbf{x}_i, \mathbf{x}_{i'})$$

(ii) Omogeneita': se λ e' uno scalare qualsiasi

$$d(\lambda\mathbf{x}_i, \lambda\mathbf{x}_{i'}) = d(\mathbf{x}_i, \mathbf{x}_{i'})$$

(iii) Cambiamento di unita' di misura: se $\mathbf{D}$ e' una matrice diagonale tale che $\mathbf{y}_i = \mathbf{D}\mathbf{x}_i$ e' il vettore di osservazioni trasformato nelle nuove scale,

$$d(\mathbf{y}_i, \mathbf{y}_{i'}) = d_w(\mathbf{x}_i, \mathbf{x}_{i'})$$

dove i pesi w_j sono eguali al quadrato degli elementi sulla diagonale di $\mathbf{D}$.
(iv) Se $\mathbf{T}$ e' una trasformazione ortogonale, tale che $\mathbf{T}'\mathbf{T} = \mathbf{I}$ (cioe' una rotazione),

$$d(\mathbf{T}\mathbf{x}_i, \mathbf{T}\mathbf{x}_{i'}) = d(\mathbf{x}_i, \mathbf{x}_{i'}).$$

La proprieta' (ii) implica che la distanza Euclidea e' sensibile alla 'dimensione' delle unita'. Questa distanza spesso oppone gruppi di unita' di piccola dimensione (con un λ piccolo) a unita' di grandi dimensioni (con un λ grande). Supponiamo per esempio che un naturalista voglia classificare dei crani di uomini preistorici in base a misure antropometriche e che per controllo consideri anche crani di *homo sapiens* e di gorilla. Tuttavia se i crani appartengono a esemplari di eta' diversa per esempio vi sono anche dei piccoli, quest'ultimi avranno misure simili a quelle degli adulti, ma piu' piccole. Allora calcolando la distanza Euclidea tra questi crani risulteranno simili tra loro i crani dei piccoli di uomo e gorilla e degli adulti di uomo e gorilla, perche' la dimensione delle unita' finisce per oscurare le altre differenze presenti.

La proprieta' (iii) illustra il legame esistente tra ponderazione delle variabili e distanze. Infatti ogni ponderazione delle variabili equivale a cambiare la scala della variabile moltiplicandola per un peso p_j. Questo cambiamento di scala fa si' che la distanza euclidea si trasformi in distanza euclidea ponderata con pesi p_j^2.

Inoltre e' evidente che — se le variabili sono *incorrelate* — allora ciascuna variabile contribuisce alla distanza con gli scarti al quadrato $(x_{ij} - x_{i'j})^2$. Quindi una misura dell'importanza della variabile nella determinazione di tutte le distanze e' data dalla media di questi scarti. Tale media e' eguale al doppio della varianza della variabile e questo significa che le variabili che hanno una piccola varianza contribuiscono poco alla distanza, mentre le variabili che hanno maggior dispersione contribuiscono molto. Questo e' un esempio di ponderazione implicita delle variabili in proporzione alle varianze.

2.6 Standardizzazione

Supponiamo di avere la matrice $\mathbf{X}$ seguente

$$\mathbf{X} = \begin{bmatrix} 45 & 30000 \\ 43 & 35000 \\ 47 & 34000 \end{bmatrix};$$

Con matrice delle distanze euclidee

$$\mathbf{D} = \begin{bmatrix} 0 & 5000 & 4000 \\ & 0 & 1000 \\ & & 0 \end{bmatrix}.$$

Si osservi che il vettore delle medie e' $\overline{\mathbf{x}} = (45, 33000)'$ mentre le varianze sono $s_1^2 = 2,6$ e $s_2^2 = 4666666$. E' evidente che l'importanza della variabile X_1 sulle distanze e' trascurabile. L'ordinamento delle distanze e' il seguente:

$$d(1,2) > d(1,3) > d(2,3).$$

Siccome la varianza e' un indice che si puo' aumentare e diminuire semplicemente moltiplicando la variabile per una costante e' intuitivo che per dare un peso eguale alle variabili basta dividerle per lo scarto quadratico medio o per qualsiasi altro indice di variabilita'. Un operazione equivalente e' la standardizzazione. Cio' equivale a calcolare le distanze Euclidee ponderate con pesi eguali all'inverso della varianza.

La matrice dei dati standardizzati e' la seguente

$$\mathbf{Z} = \begin{bmatrix} 0,00 & -1.39 \\ -1,22 & 0,93 \\ 1,22 & 0,46 \end{bmatrix};$$

con matrice delle distanze Euclidee

$$\mathbf{D}_z = \begin{bmatrix} 0 & 2,62 & 2,22 \\ & 0 & 2,48 \\ & & 0 \end{bmatrix}.$$

Il fatto di aver standardizzato le variabili comporta che ognuna di esse abbia lo stesso peso e comporta altresi' che tutto il sistema di distanze venga sconvolto senza che l'ordinamento tra di esse sia conservato. Infatti l'ordinamento ora e' $d(1,2) > d(2,3) > d(1,3)$.

Si poteva dedurre che l'insieme delle distanze dovesse cambiare dopo aver standardizzato le variabili anche dall'esempio degli stati americani e dall'esame dei due grafici 1.1 e 1.2 relativi ai tassi di analfabetismo e di delinquenza. Standardizzare le variabili prima di calcolare le distanze Euclidee ha come risultato l'indubbio vantaggio di eliminare la dipendenza della distanza dalle unita' di misura, tuttavia ha uno svantaggio: quello di diluire le differenze tra gruppi, rispetto alle variabili maggiormente discriminanti. Cio' si puo' vedere dai grafici 2.1 e 2.2 in cui e' riportato lo scatter relativo a due variabili prima e dopo la standardizzazione. Il grafico e' costruito in modo da avere approssimativamente la stessa scala su entrambe le dimensioni. Mentre inizialmente si notano due gruppi separati, dopo la standardizzazione uno dei gruppi si schiaccia sull'altro attenuando la separazione tra i due.

Un'altro fattore che implicitamente pondera le variabili e' costituito dalla correlazione tra le variabili stesse. Se le variabili sono incorrelate e standardizzate, ciascuna ha lo stesso peso nella formazione della distanza, ma se le variabili pur standardizzate sono correlate fra loro allora avviene come se certe variabili fossero contate piu' di una volta, ossia la distanza risentira' maggiormente di certe variabili, anche non osservabili, che influiscono sulle variabili osservate.

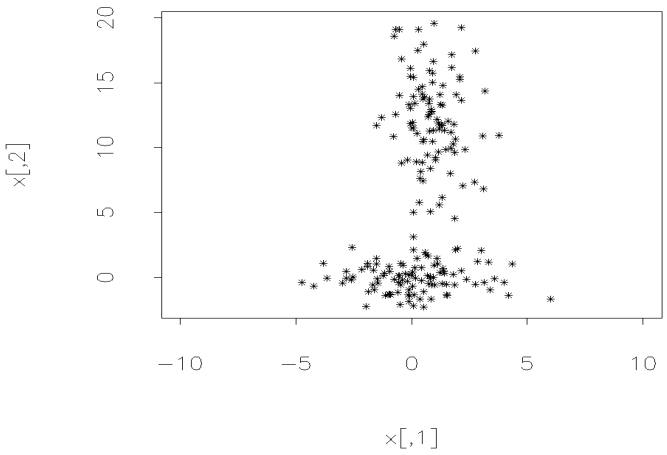

Figura 2.1: *Due gruppi evidenti*

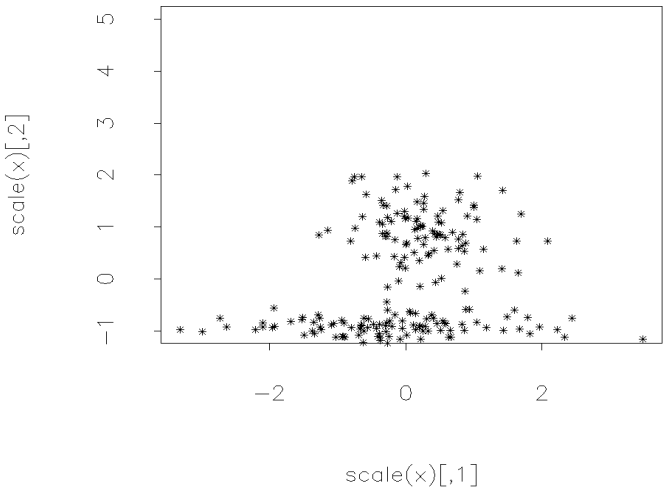

Figura 2.2: *Standardizzando i gruppi si diluiscono*

2.7 Altre distanze per matrici di misure

Finora abbiamo parlato di una distanza, quella Euclidea, che corrisponde al concetto intuitivo di distanza che tutti abbiamo. Tuttavia, e' possibile introdurre anche tipi di distanza diversi, del tutto comprensibili, ma che danno luogo a una geometria diversa dall'usuale. Per esempio, nella distanza Euclidea si sommano degli scarti al quadrato, mentre sembrerebbe piu' logico sommare gli scarti in valore assoluto. Difatti, si puo' anche introdurre un indice di distanza definito come segue:

$$d_1(i,i') = d_1(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} \mid x_{ij} - x_{i'j} \mid$$

detta *distanza* L_1 o distanza *city-block*. L'origine del nome e' dovuto al fatto che essa misura la distanza che c'e' tra due punti su un piano nel caso in cui si sia vincolati a muoversi solo parallelamente agli assi coordinati e non si possa andare in diagonale, come per l'appunto avviene per chi si sposta da un punto all'altro di una citta'. Come si vede, la geometria indotta da questa distanza e' del tutto particolare, cio' nonostante si tratta di una metrica esattamente come la metrica Euclidea. Anch'essa possiede proprieta' simili a quelle della metrica Euclidea, ma non la proprieta' di invarianza per rotazione (la proprieta' (iv)) che e' tipica della metrica Euclidea.

Una generalizzazione delle distanze Euclidea ed L_1 e' la *distanza di Minkowsky* definita come segue

$$d_\lambda(i,i') = d_\lambda(\mathbf{x}_i, \mathbf{x}_{i'}) = \left(\sum_{j=1}^{p} \mid x_{ij} - x_{i'j} \mid^\lambda \right)^{1/\lambda}.$$

La distanza di Minkowsky dipende da un parametro λ. Se $1 \leq \lambda < \infty$ l'indice di distanza e' una metrica, altrimenti non lo e'. Se $\lambda = 1$ si ottiene la distanza L_1, se $\lambda = 2$ si ottiene invece la distanza Euclidea.

2.8 Distanza chi-quadrato

Consideriamo ora una distanza particolarmente utile quando si analizzano tabelle di contingenza. Si considerino i dati dell'esempio 1.2: a ogni riga corrisponde una tipologia di furto e una distribuzione di frequenza secondo l'eta'. Per ogni colonna, cioe' per ogni eta' si ha una distribuzione condizionata di frequenza secondo il tipo di furto. E' interessante dunque sapere quali sono le distribuzioni condizionate simili fra loro e a tal fine si puo' usare una distanza detta distanza chi-quadrato.

Siano $f_{ij} = n_{ij}/n_{i+}$ le frequenze congiunte relative. Consideriamo due generici profili riga i e i' della tabella di contingenza. Essi hanno come elementi le frequenze condizionate relative f_{ij}/f_{i+} e $f_{i'j}/f_{i'+}$ (per $i = 1, \ldots, I$ e $j = 1, \ldots, J$). Allora la distanza chi-quadrato fra i due profili riga e' una distanza eulidea ponderata con gli inversi delle frequenze marginali di colonna:

$$d^2_\chi(i,i') = \sum_{j=1}^{J} 1/f_{+j}(f_{ij}/f_{i+} - f_{i'j}/f_{i'+})^2$$

Dunque quando i e i' hanno lo stesso profilo risulta $d^2_\chi(i,i') = 0$. La differenza tra i profili i e

i' per la colonna j e' divisa per f_{+j} in modo da dare meno importanza a quelle modalita' delle colonne che hanno i margini piu' alti. Ovviamente si potra' anche introdurre una distanza chi-quadrato tra le colonne della tabella di contingenza.

2.9 Indici di somiglianza

Gli indici di somiglianza sono stati utilizzati inizialmente nella tassonomia numerica degli animali e delle piante. Sono estremamente utili quando i caratteri considerati sono qualitativi.

Un'indice di somiglianza molto utilizzato e' *l'indice di Gower* che ha proposto in realta' un indice generale valido sia per dati quantitativi che per dati qualitativi. L'indice e' il seguente

$$s(i,i') = \frac{\sum_j c_{ii'j}}{\sum_j w_{ii'j}}$$

dove $c_{ii'j}$ e' una misura di somiglianza tra i e i' tenuto conto solo del carattere j, mentre $w_{ii'j}$ e' un peso che puo' assumere solo valori 1 e 0 e assume valori nulli solo quando non e' sensato un confronto tra i e i' per quel carattere.
(i) Nel caso in cui X_j sia un carattere quantitativo, si pone

$$c_{ii'j} = 1 - \mid x_{ij} - x_{i'j} \mid / R_j$$

dove R_j e' il campo di variazione della variabile j usato per eliminare il problema della scala (e' equivalente dividere per lo scostamento quadratico medio s_j o per R_j.

Nel caso di caratteri qualitativi, l'indice di Gower definisce diversamente i valori $c_{ii'j}$ e $w_{ii'j}$.
(ii) Se X_j e' un carattere dicotomico, i valori sono determinati secondo la tabella seguente.

Presenza/assenza del carattere dicotomico j

Unita' i	1	1	0	0
Unita' i'	1	0	1	0
$c_{ii'j}$	1	0	0	0
$w_{ii'j}$	1	1	1	0

Pertanto dai confronti vengono esclusi i casi in cui entrambe le unita' presentano l'assenza del carattere, mentre la somiglianza e' uno se vi e' co-presenza del carattere.
(iii) Se X_j e' un carattere qualitativo politomico i valori di $w_{ii'j}$ sono sempre uno (salvo in caso di dato mancante), mentre $c_{ii'j} = 1$ se le due unita' hanno la stessa modalita' del carattere, e zero altrimenti. Se questa seconda definizione viene applicata a dati dicotomici si ottiene un indice diverso, in quanto $c_{ii'j} = 1$ anche nella situazione di co-assenza del carattere nelle due unita'.

Se tutti i caratteri sono dicotomici $s(i,i')$ coincide con un indice di somiglianza detto di Jaccard. Se tutti i caratteri sono dicotomici, ma viene applicata la regola (iii), si ottiene un indice chiamato *simple matching coefficient*.

L'utilita' di questa famiglia di indici discende dal fatto che Gower ha dimostrato che la matrice delle somiglianze di elemento generico $s(i,i')$ e' semidefinita positiva e questa

proprieta' e' fondamentale laddove si vogliano utilizzare i metodi di *scaling* multidimensionale. Infatti si puo' dimostrare che la distanza definita da

$$d(i,i') = 2\sqrt{1 - s(i,i')}$$

e' una metrica (cioe' soddisfa la disuguaglianza triangolare) ed esiste una configurazione di punti per i quali essa e' una distanza Euclidea.

2.10 Strutture di classificazione

Una volta definito un indice di prossimita' e' necessario introdurre una definizione precisa del concetto di gruppo. Il miglior modo per farlo e' quello di stabilire delle strutture matematiche tali da poter essere utilizzate per la classificazione.

Le due strutture [1] piu' comunemente utilizzate sono le *partizioni* e le *gerarchie*.

2.10.1 Partizioni

Una partizione dell'insieme delle unita' statistiche U e' un insieme di parti $\{A_1, \ldots, A_G\}$ che siano disgiunte a due a due e la cui riunione sia eguale ad U.

Una partizione particolarmente importante quando tutti i caratteri sono quantitativi e i vettori unita' sono visti come punti di uno spazio Euclideo, e' la partizione *generata* da G punti $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_G$. Essa e' definita considerando in ciascuna classe A_g tutti quelle unita' che sono piu' vicine (rispetto alla distanza Euclidea) a $\mathbf{m}_g$ che agli altri punti. Precisamente se una unita' qualsiasi $\mathbf{x}$ appartiene a A_t allora

$$d(\mathbf{x}, \mathbf{m}_t) = \min_{g=1,\ldots,G} d(\mathbf{x}, \mathbf{m}_g).$$

2.10.2 Gerarchie

Un'altra struttura di classificazione la cui origine risale agli studi tassonomici e' la gerarchia.

Un insieme di parti $\mathcal{H}$ di U e' detto gerarchia, se dati due insiemi A e B appartenenti ad $\mathcal{H}$ si puo' verificare una ed una sola delle tre possibilita' seguenti
(i) $A \cap B = \emptyset$
(ii) $A \subset B$
(iii) $B \subset A$.

Una gerarchia e' detta totale se contiene tutti gli insiemi composti da una sola unita'. Per esempio, se

$$U = \{u, v, w, x, y, z\},$$

e si definisce

$$\mathcal{H} = \{u, v, w, x, y, z, uv, wx, wxyz, U\},$$

allora $\mathcal{H}$ e' una gerarchia.

La relazione di inclusione tra le classi della gerarchia puo' essere rappresentata con un grafo detto *albero*. Nel grafico 2.3 viene data una rappresentazione di questo tipo per la

[1] Tralasceremo in questa sede le strutture di tipo probabilistico, in cui si utilizzano ad esempio miscugli di distribuzioni.

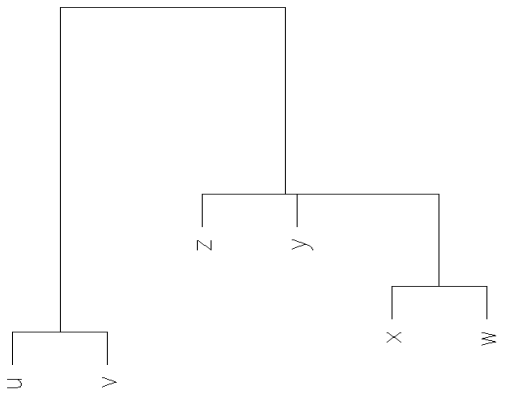

Figura 2.3: *Un albero gerarchico*

gerarchia sopra definita.

Per ogni classe A di una gerarchia si definisce l'insieme dei suoi successori immediati ossia l'insieme di quelle classi che sono incluse in A, e non sono incluse in altra classe di A. Per esempio l'insieme dei successori immediati di $wxyz$ nell'ultimo esempio e' $\{wx, y, z\}$.

Una gerarchia si dice *binaria* se ogni sua classe A ha o due successori immediati o nessuno. Pertanto, la gerarchia sopra definita non e' binaria.

Si osservi che la relazione "e' incluso in" definita tra le classi di una gerarchia non e' definita per tutte le classi.

2.10.3 Dendrogrammi

Le gerarchie vengono utilizzate nell'analisi dei gruppi associando a un albero (spesso binario) un indice di dispersione delle classi che permette di graduare la gerarchia.

Una gerarchia totale $\mathcal{H}$ si dice graduata se esiste una funzione reale $h(A)$ definita per ogni classe A della gerarchia che misuri la dispersione della classe e che conservi l'ordine di inclusione, ossia che goda della proprieta' seguente: se A e B sono due classi qualsiasi della gerarchia tali che $A \subset B$, allora $h(A) \leq h(B)$. Inoltre, la funzione $h(\cdot)$ e' tale che $h(i) = 0$ per ogni unita' singola i. Una gerarchia graduata si dice comunemente *dendrogramma*.

Ecco due esempi di funzioni di graduazione.
(a) $h(A) = \max_{i,i' \in A}\{d(i, i')\}$
(b) $h(A) = \text{dev}(A)$ dove

$$\text{dev}(A) = \sum_i d^2(\mathbf{x}_i, \overline{\mathbf{x}}_A)$$

e' chiamata *devianza* di A ed e' la somma dei quadrati delle distanze Euclidee al quadrato tra i vettori unita' compresi nella classe A e il vettore delle medie $\overline{\mathbf{x}}_A$ della classe. Ovviamente la devianza puo' essere interpretata come un indice di dispersione del gruppo (ma non essendo divisa per la numerosita' del gruppo, dipende dal numero di unita' che ne fanno parte). Nel grafico 2.4 e' riportato l'albero della gerarchia precedente con una scala che indica il livello dell'indice $h(A)$ per ogni classe. Si osservi che ad ogni dendrogramma corrispondono delle

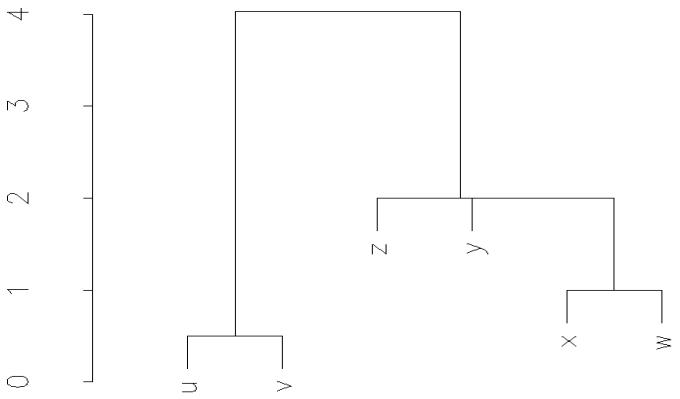

Figura 2.4: *Un dendrogramma*

partizioni indotte cioe' le partizioni che si ottengono tagliando l'albero a vari livelli. Cosi' facendo "cadono" i rami associati alle classi di una partizione. Inoltre le classi della partizione ottenuta tagliando l'albero al livello h_1 sono tutte contenute nelle classi della partizione ottenuta tagliando l'albero al livello $h_2 > h_1$.

Percio' la graduazione della gerarchia permette di *ordinare tutte le classi della gerarchia* e diremo che la classe A della gerarchia *precede* la classe B se $h(A) < h(B)$. Si possono cosi' confrontare anche le classi non incluse l'una nell'altra.

2.11 Ultrametrica associata a un dendrogramma

Ad ogni dendrogramma si puo' associare un indice di distanza $\delta(i, i')$ tra le unita' sfruttando la graduazione. Infatti per misurare la distanza tra due unita' i e i' si procede nel modo seguente: si cerca la piu' piccola classe della gerarchia che contenga entrambe le unita', diciamo $A_{ii'}$, e si va a vedere quanto vale l'indice $h(A)$. In formule,

$$\delta(i, i') = h(A_{ii'})$$

dove

$$A_{ii'} = \min_{A \in \mathcal{H}} \{h(A) \mid i \in A; i' \in A; A \subset \mathcal{H}\}.$$

Allora, si puo' dimostrare che l'indice di distanza $\delta(i, i')$ e' un'ultrametrica detta *ultrametrica associata al dendrogramma*.

Nell'esempio rappresentato nel grafico 2.4 si calcola facilmente che la matrice dell'ultrametrica e' la seguente

δ	u	v	w	x	y	z
u	0	0,5	4	4	4	4
v		0	4	4	4	4
w			0	1	2	2
x				0	2	2
y					0	2
z						0

Dunque ad ogni dendrogramma e' associata una ultrametrica. Si puo' dimostrare che l'ultrametrica associata *caratterizza* un dendrogramma, nel senso che si puo' passare dall'uno all'altra con una corrispondenza [2] essenzialmente biunivoca.

2.12 La costruzione dei gruppi

L'utilizzazione dell'indice di distanza e di una delle due strutture di classificazione sopra discusse, cioe' le partizioni e i dendrogrammi avviene utilizzando un procedimento di adattamento della struttura stessa ai dati.

E' opportuna fare una distinzione tra i metodi di classificazione sulla base della struttura di raggruppamento che impiegano. Distingueremo:
(a) *Metodi gerarchici* in cui la struttura e' il dendrogramma;
(b) *Metodi non gerarchici* in cui la struttura e' la partizione.

Il primo passo dell'analisi dei gruppi come detto e' quello della definizione di opportuni indici di prossimita'. Tuttavia l'esame diretto della matrice della prossimita' non da' in generale una comprensione maggiore di quella desumibile dall'esame diretto dei dati stessi. E' necessaria infatti un operazione preliminare di semplificazione dell'informazione contenuta nella matrice delle prossimita'.

La linea di attacco del problema potrebbe essere quella di definire — basandosi sulle prossimita' — degli indici che misurino l'omogeneita' e la separazione delle classi (delle partizioni o delle gerarchie). Tuttavia, una volta definiti, gli indici dovrebbero poi essere calcolati su *tutte* le partizioni possibili o tutte le gerarchie possibili delle n unita'e il numero di partizioni o di gerarchie da considerare e' elevatissimo, gia' per valori piccoli di n, come mostra la seguente tabella.

# unita'	# distanze	# partizioni	# gerarchie
4	6	15	18
10	45	115975	2571912000
15	105	1382958545	$6,96 \times 10^{18}$
20	190	$5,17 \times 10^{13}$	$5,64 \times 10^{29}$

Percio', non e' materialmente possibile ottimizzare un criterio in modo globale, ma occorre limitare la ricerca a un sottoinsieme (molto ridotto) delle partizioni o delle gerarchie.

2.13 Metodi gerarchici

I metodi gerarchici sono quelli in cui la struttura di classificazione e' il dendrogramma. Vista l'equivalenza tra dendrogrammi e ultrametriche discussa precedentemente, si possono definire anche come quei metodi che *trasformano l'indice di distanza iniziale in una distanza ultrametrica*.

Sio osservi che in un albero binario, il numero totale dei nodi e' uguale al numero dei nodi terminali (n) piu' il numero dei nodi non terminali($n-1$) e quindi e' pari a $2n-1$. Si osservi anche che ogni distanza ultrametrica al massimo puo' avere come valori solo quelli

[2] Esiste in realta' una classe di dendrogrammi ai quali e' associata la stessa ultrametrica. Tali dendrogrammi si dicono equivalenti.

corrispondenti a un nodo dell'albero. Quindi trasformare $n(n-1)/2$ indici di distanza iniziali $d(i,i')$ in un dendrogramma (binario), significa ottenere alla fine $n-1$ distanze ultrametriche $\delta(i,i')$.

Una prima classificazione dei metodi gerarchici e' tra
(a) *Metodi ordinali*, se utilizzano come informazione solo l'ordinamento associato all'indice di distanza. Pertanto questa classe di metodi e' invariante rispetto a qualsiasi trasformazione monotona delle distanze.
(b) *Metodi non ordinali*, se utilizzano i valori numerici delle distanze. In tal caso la proprieta' di invarianza e' perduta.

Un'altra classificazione dei metodi gerarchici e' fatta sulla base dell'algoritmo usato per la costruzione dei gruppi. Si distinguono
(c) *Metodi basati su un criterio locale.* Essi sono basati generalmente su un algoritmo detto *agglomerativo* che descriveremo fra breve e nel quale intervengono ad ogni passo solo una parte delle distanze.
(d) *Metodi basati su un criterio globale.* Essi per costruire i gruppi mettono in gioco tutte le distanze $d(i,i')$ cercando di minimizzare lo scarto tra di esse e le ultrametriche $\delta(i,i')$.

2.13.1 L'Algoritmo agglomerativo

Dovendo costruire una gerarchia sull'insieme di unita' U e' chiaro che si puo' scegliere tra due strategie: quella divisiva, che parte da U e procede suddividendolo via via, e quella agglomerativa che parte dalle unita' e forma i gruppi per fusioni successive.

Lo schema generale e' il seguente: esso presuppone che si sia definito un indice $D(A,B)$ di distanza tra classi che chiameremo *indice di aggregazione.* Per indice di aggregazione intendiamo cioe' una funzione reale positiva tale che $D(A,B) = D(B,A)$ e tale da misurare la distanza tra i due gruppi sulla base delle distanze tra unita'.
1. Si parte dalla partizione banale P_0 le cui classi sono ridotte ad un solo elemento.
2. Si costruisce una nuova partizione riunendo le due classi della partizione precedente, diciamo A e B, che rendono minimizzare l'indice di aggregazione $D(A,B)$.
3. Si ripete il passo precedente fino a riunire tutte le classi in una sola.

Al passo $t-1$ dell'algoritmo, le due classi $A^{(t-1)}$ e $B^{(t-1)}$ che minimizzano l'indice di aggregazione vengono fuse in una sola, diciamo $C^{(t)}$, e vanno a formare un nodo dell'albero binario, in corrispondenza del quale il valore della graduazione e' definito da

$$h(C^{(t)}) = h_t = D(A^{(t-1)}, B^{(t-1)})$$

Percio', poiche' t va da 0 a $n-1$, si ottengono n valori $h_0, h_1, \ldots, h_{n-1}$ che, affiancati alla gerarchia, danno luogo al dendrogramma finale.

Se risulta che

$$0 = h_0 \leq h_1 \leq \cdots \leq h_{n-1}$$

si dice che l'indice di aggregazione e' monotono. In caso contrario si dice che per qualche valore di t avviene un inversione. Ossia, risulta che la dispersione dei due gruppi che si fondono $h(A \cup B)$ e' minore della dispersione di due gruppi che si sono fusi un passo precedente. Questa eventualita' e' possibile per certi indici $h(\cdot)$ che per questo sono scarsamente utilizzati perche' poco interpretabili.

Si noti che ad ogni fusione di due classi, intervengono le classi ottenute fino a quel momento. Si tratta infatti di una procedura sequenziale, in cui ad ogni passo non si ridiscutono piu' le scelte fatte nei passi precedenti.

Talvolta vi possono essere piu' coppie di classi che minimizzano la dispersione e si possono stabilire regole per la fusione simultanea di piu' du due classi. In questi casi la gerarchia risultante non e' piu' binaria.

2.14 Metodi gerarchici con criterio locale

Esistono moltissimi metodi gerarchici locali diversi a seconda dell'indice di aggregazione che utilizzano. Citeremo solo i piu' usati.

2.14.1 Criterio del legame singolo

L'indice di aggregazione e' definito da

$$D(A,B) = \min_{i\in A, i'\in B}\{d(i,i')\}.$$

Si dimostra che l'indice di aggregazione e' monotono e che il metodo di classificazione che ne deriva e' ordinale.

La vicinanza di due classi e' misurata dalla distanza che separa le due unita' piu' vicine. Se le classi A e B sono formate da n_A ed n_B unita', delle $n_A n_B$ distanze possibili il criterio del legame singolo ne considera solo una, la piu' piccola.

Il dendrogramma del grafico 2.4 e' appunto ottenuto dal criterio del legame singolo applicato alla matrice di osservazioni

$$\mathbf{X} = \begin{bmatrix} 1 & 0.0 \\ 1 & 0.5 \\ 5 & 3.0 \\ 5 & 4.0 \\ 3 & 4.0 \\ 5 & 6.0 \end{bmatrix}$$

e utilizzando la distanza Euclidea.

E' un criterio che permette di individuare gruppi di qualsiasi forma, purche' ben separati. Due gruppi possono essere aggregati nei primi passi ed essere considerati poco dissimili anche solo perche' esiste una catena di unita' che unisce i due gruppi. Ad ogni fusione le unita' non ancora classificate tendono ad essere incorporate in gruppi gia' esistenti piuttosto che formare nuovi gruppi. Questa proprieta' si chiama *effetto di concatenamento.*

2.14.2 Criterio del legame completo

L'indice di aggregazione stavolta e'

$$D(A,B) = \max_{i\in A, i'\in B}\{d(i,i')\}$$

per cui la vicinanza tra due classi e' misurata dalla distanza tra le due unita' piu' lontane (il diametro di $A \cap B$).

Si dimostra che l'indice di aggregazione e' monotono e che il metodo di classificazione che ne deriva e' ordinale.

2.14.3 Criterio del legame medio

L'indice di aggregazione del criterio del legame medio e'

$$D(A,B) = 1/n_A n_B \sum_{i \in A} \sum_{i' \in B} d(i,i')$$

dove n_A ed n_B sono le numerosita' rispettivamente di A e B. L'indice — che e' monotono — e' basato sulla distanza media tra due gruppi. Il metodi che ne risulta non e' monotono.

2.14.4 Criterio dei centroidi

L'indice di aggregazione puo' essere usato solo se tutti i caratteri sono quantitativi. La sua definizione e' la seguente

$$D(A,B) = d^2(\overline{\mathbf{x}}_A, \overline{\mathbf{x}}_B)$$

dove d^2 e' il quadrato della distanza Euclidea e $\overline{\mathbf{x}}_A$ e $\overline{\mathbf{x}}_B$ sono, rispettivamente, i vettori delle medie di A e di B. La vicinanza tra i gruppi e' misurata dalla distanza tra i centroidi. Nonostante la sua intuitivita' l'indice e' poco utilizzato perche' non e' monotono.

2.14.5 Criterio di Ward

L'indice di aggregazione del criterio di Ward e' il seguente

$$D(A,B) = \frac{n_A n_B}{n_A + n_B} d^2(\overline{\mathbf{x}}_A, \overline{\mathbf{x}}_B)$$

con le notazioni precedenti. L'indice e' semplicemente la devianza tra i gruppi A e B. Infatti, risulta che

$$\mathrm{dev}(A \cup B) = \mathrm{dev}(A) + \mathrm{dev}(B) + \frac{n_A n_B}{n_A + n_B} d^2(\overline{\mathbf{x}}_A, \overline{\mathbf{x}}_B)$$

espressione in cui la somma dei primi due termini e' detta devianza entro i gruppi, mentre l'ultimo termine e' detto devianza tra gruppi.

L'indice di Ward misura percio' la parte della dispersione di $A \cup B$ dovuta alle differenze tra i gruppi. L'indice e' monotono e il metodo che ne deriva e', ovviamente, non ordinale.

2.15 Discussione

Una classificazione gerarchica produce come risultato una successione di partizioni di n classi, $n-1$ classi, e cosi' via fino a una classe sola. Il fatto di non produrre un solo raggruppamento e' un vantaggio dei metodi gerarchici perche' permette di studiare diverse strutture possibili per i dati, con un numero diverso di gruppi.

Spesso il numero dei gruppi e' incognito e lo studio del dendrogramma e' utile per fare delle congetture. Ai livelli in cui l'indice di aggregazione cresce vistosamente e' chiaro che la

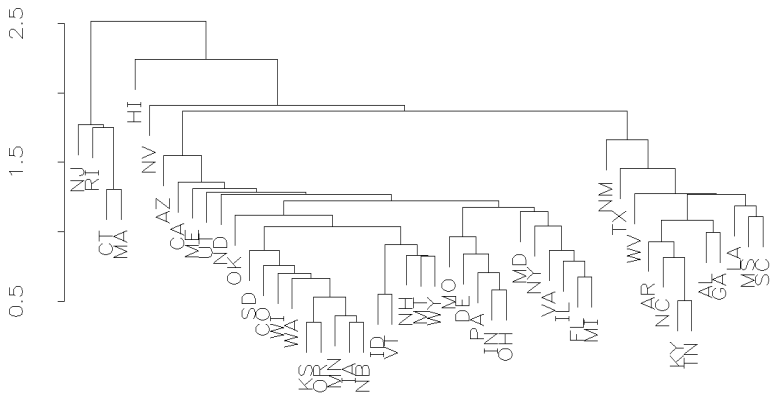

Figura 2.5: *Criterio del legame singolo*

fusione avviene a un costo elevato e quindi e' conveniente fermare il processo. Non esistono comunque dei criteri oggettivi per determinare il numero dei gruppi.

D'altra parte, i gruppi possono avere una dispersione diversa e se un criterio (come quello del legame singolo) e' sensibile alle piccole distanze, a volte non e' utile tagliare l'albero a un livello solo perche' produrrebbe un solo gruppo e una miriade di piccoli gruppi anche contenenti una sola unita'.

Ogni indice di aggregazione produce una gerarchia diversa e cio' talvolta puo' creare delle difficolta' di interpretazione. Se la diversita' dei risultati non e' rilevante, cioe' le partizioni indotte sono pressappoco le stesse, cio' e' ovviamente segno di una stabilita' dei risultati. Ma a volte criteri diversi forniscono delle descrizioni abbastanza diverse dei dati e quindi sono difficilmente accordabili. Per fare un esempio, riprendiamo i dati dell'esempio 1.8 e consideriamo

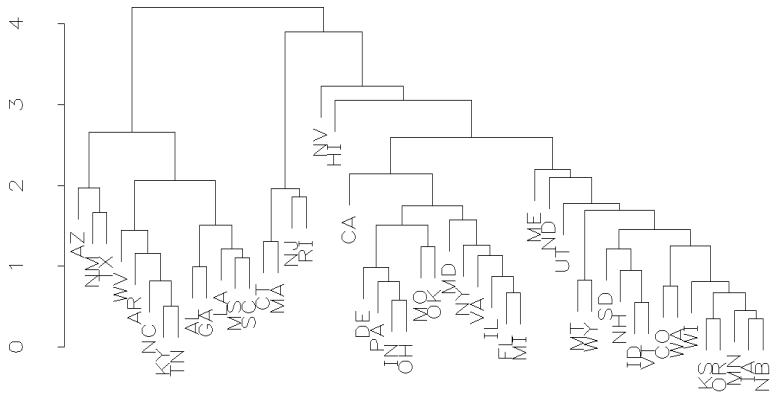

Figura 2.6: *Criterio del legame medio*

tutti gli stati americani, eccettuata l'Alaska che presenta a prima vista delle caratteristiche anomale. Si voglia costruire un dendrogramma sulla base delle variabili X_1^* (la densita' di

Alabama	1	Idaho	2	Pennsylvania	2
Arizona	1	Illinois	2	South Dakota	2
Arkansas	1	Indiana	2	Utah	2
Georgia	1	Iowa	2	Vermont	2
Kentucky	1	Kansas	2	Virginia	2
Louisiana	1	Maine	2	Washington	2
Mississippi	1	Maryland	2	Wisconsin	2
New Mexico	1	Michigan	2	Wyoming	2
North Carolina	1	Minnesota	2	—	
South Carolina	1	Missouri	2	Connecticut	3
Tennessee	1	Montana	2	Massachusetts	3
Texas	1	Nebraska	2	New Jersey	3
West Virginia	1	New Hampshire	2	Rhode Island	3
—		New York	2	—	
California	2	North Dakota	2	Hawaii	4
Colorado	2	Ohio	2	—	
Delaware	2	Oklahoma	2	Nevada	5
Florida	2	Oregon	2		

Tabella 2.1: *Gruppi ottenuti con il legame medio*

popolazione) e da X_2 a X_7 (escludendo l'area dello stato). Si stabilisca di scegliere la distanza Euclidea semplice sulle variabili standardizzate.

Nei grafici 2.5 e 2.6 sono riportati i due dendrogrammi relativi ai criteri del legame singolo e del legame medio. I due dendrogrammi sono abbastanza diversi e dimostrano le due strategie di raggruppamento dei due criteri.

L'esame del dendrogramma non suggerisce l'esistenza di gruppi naturali ben separati (altrimenti si sarebbero visti probabilmente anche con gli scatter delle variabli a due a due). Tuttavia, e' utili a fini operativi effettuare una prima classificazione degli stati e una buon raggruppamento sembra possibile in tre gruppi.

E' opportuno tagliare il dendorgramma in modo tale che i tre gruppi siano abbastanza "pieni". In questo caso questo si ottiene tagliando al livello dei 5 gruppi ed eliminando alcune unita' finite in gruppi di un unico elemento. Le numerosita' dei gruppi sono $(12, 31, 4, 1, 1)$ per il criterio del legame singolo e $(13, 30, 4, 1, 1)$ per il criterio del legame medio. Riportiamo nella tabella 2.1 il risultato del raggruppamento del legame medio (unita' e indice del gruppo). Le due partizioni del legame singolo e del legame medio non sono molto diverse. Lo si puo' verificare con una tabella di contingenza che incroci le due partizioni:

	1	2	3	4	5	*# singolo*
1	12	0	0	0	0	12
2	1	30	0	0	0	31
3	0	0	4	0	0	4
4	0	0	0	1	0	1
5	0	0	0	0	1	1
# medio	13	30	4	1	1	49

L'analisi completa dovrebbe ora proseguire con l'interpretazione dei gruppi e la determinazione delle variabili che maggiormente contribuiscono alla separazione dei gruppi.

2.15.1 Problemi di efficienza

I metodi gerarchici con criterio locale ammettono una formula di calcolo ricorsiva che consente di calcolare l'indice di aggregazione tra classi in funzione dell'indice calcolato al passo precedente. Inoltre si puo' usare un'unica formula dipendente da parametri per tutti i criteri precedenti.

Tuttavia, questo non e' il modo piu' efficiente per eseguire i calcoli. Per ogni criterio sono stati scoperti degli algoritmi ottimizzati che seguono procedure anche molto diverse dall'algoritmo generale agglomerativo. Sono stati ideati anche degli algoritmi che consentono di ridurre l'occupazione di memoria che normalmente e' proporzionale al numero di elementi della matrice delle distanze. Fino a poco tempo fa era impensabile classificare 10000 unita' con metodi gerarchici, cosa che attualmente e' perfettamente possibile.

2.16 Metodi con criterio globale

Si e' detto che i metodi gerarchici sono basati sulla trasformazione da $d(i, i')$ in un'ultrametrica $\delta(i, i')$. E' evidente che in questa trasformazione si vorrebbe minimizzare la distorsione in modo tale che la gerarchia finale si adatti il piu' possibile alla struttura di distanze di partenza.

Uno dei modi per definire una ultrametrica con uno scarto minimo da $d(i, i')$ e' il seguente che caratterizza la cosiddetta ultrametrica sottodominante. Essa e' definita come l'ultrametrica $\delta^-(i, i')$ che e' inferiore alla distanza di partenza — nel senso che $\delta^-(i, i') \leq d(i, i')$ per ogni i ed i' — e contemporaneamente e' la piu' vicina ad essa secondo il criterio

$$\min_{\delta} \sum_{i \in U} \sum_{i' \in U} \mid d(i, i') - \delta(i, i') \mid .$$

Come si vede il criterio e' un criterio globale perche' coinvolge tutte le distanze.

L'ultrametrica sottodominante ha un'interesse particolare perche' si dimostra che essa e' esattamente uguale all'ultrametrica associata al dendrogramma che si ottiene col criterio del legame singolo. Pertanto il criterio locale del legame singolo da' luogo a una gerarchia che ottimizza il criterio globale della sottodominante.

2.17 Albero di lunghezza minima

Introdurremo adesso un concetto, apparentemente molto lontano dai metodi gerarchici, ma in realta' ad essi strettamente collegato. Consideriamo n localita' e supponiamo di volerle collegare con un cavo telefonico in modo che il cavo non faccia cicli, colleghi tutte le localita' e abbia lunghezza minima. Questo classico problema di ricerca operativa e' stato risolto negli anni 50, fornendo anche un algoritmo (oggi notevolmente migliorato) per la determinazione del percorso ottimale.

Piu' in generale, il problema e' quello di determinare un albero di lunghezza minima che colleghi n punti. Come noto, un albero e' un grafo connesso, senza cicli. In un albero, per ogni coppia di unita' i e i', esiste un cammino ed uno solo $C_{ii'}$ che le unisca (altrimenti vi sarebbe un ciclo). Il numero degli spigoli dell'albero e' ovviamente $n - 1$. La lunghezza dell'albero e'

$$\sum_{(i,i') \in S} d(i, i')$$

dove S e' l'insieme delle coppie di unita' collegate dall'albero e $d(i, i')$ e' un qualsiasi indice di distanza tra i punti.

Nel grafico 2.7 e' riportato l'albero di lunghezza minima sul grafico di dispersione delle due variabili 'vita media' e 'percentuale di diplomati' dell'esempio 1.8 relativo ai 50 stati americani. E' estremamente utile aggiungere allo scatter l'albero di lunghezza minima che

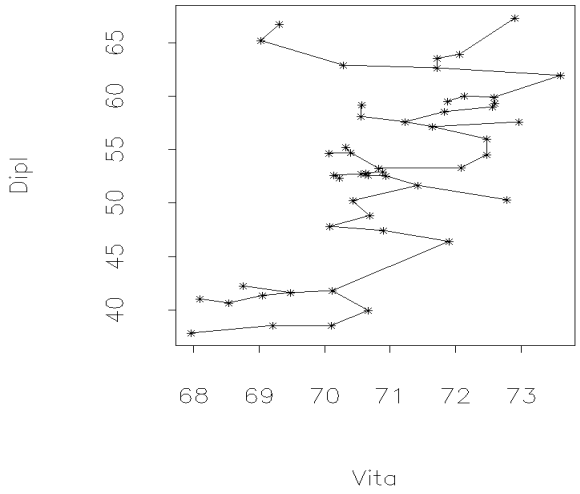

Figura 2.7: *Albero di lunghezza minima*

costituisce una sorta di "scheletro" dei dati. Gli spigoli piu' lunghi si possono tagliare e facendo cio' l'albero risulta scomposto in parti connesse che individuano altrettanti gruppi.

L'importanza dell'albero di lunghezza minima e il suo collegamento con i metodi gerarchici deriva dal risultato seguente: dato un albero di lunghezza minima su un insieme di unita' U, allora ad esso e' associata una gerarchia del legame singolo ossia la sua ultrametrica sottodominante. Le classi ottenute tagliando via via gli spigoli piu' lunghi dell'albero di lunghezza minima formano delle partizioni identiche a quelle della gerarchia del legame singolo. Date due unita' i e i' esiste solo un cammino che le unisca sull'albero di lunghezza minima. Allora, la lunghezza dello spigolo piu' lungo di questo cammino e' esattamente eguale alla distanza ultrametrica sottodominante tra i e i'.

2.18 Metodi non gerarchici

Ci occuperemo brevemente dei metodi di raggruppamento che determinano una sola partizione delle unita' e che chiameremo, negativamente, metodi non gerarchici. Essi sono meno flessibili dei metodi gerarchici — presuppongono infatti che il numero dei gruppi sia noto — e meno ricchi di informazioni, ma proprio per questo sono piu' veloci e relativamente poco costosi. In questa sede, per brevita', escluderemo metodi non gerarchici per classificare unita' su cui siano state rilevate mutabili.

Generalmente un metodo non gerarchico utilizza l'indice di distanza per calcolare un criterio di classificazione da ottimizzare e un algoritmo che consente di spostare le unita' da un

gruppo a un altro in modo da ottimizzare il criterio su una classe ristretta, ma presumibilmente utile, di partizioni.

L'algoritmo fondamentale e' detto *k-means* ed ha la seguente struttura semplificata.
1. Si parte con una partizione iniziale in G classi delle unita', dove G e' scelto *a priori*. E' consigliabile che la partizione sia determinata in modo ragionato.
2. Per ogni classe della partizione del passo corrente si calcola una opportuna *rappresentazione*. La rappresentazione ha lo scopo di precisare un modello di gruppo per la classe. Nei casi usuali, una rappresentazione di una classe e' semplicemente il vettore delle medie della classe. Si calcola inoltre un criterio che misuri la bonta' del raggruppamento.
3. Ottenute G rappresentazioni delle classi, le classi, cioe' i centroidi

$$\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_G,$$

le classi vengono ricalcolate, determinando la partizione di minima distanza (vedi 2.10.1) generata dalle rappresentazioni scelte. In tal modo la struttura di classificazione (la partizione) viene adeguata alle rappresentazioni delle classi.
4. Si ripete il processo, fino a che ricalcolando la partizione al punto 3. non si hanno piu' spostamenti di unita' da un gruppo a un altro.

Facciamo alcune precisazioni. Solitamente la distanza usata e' la distanza euclidea. Pertanto al punto 3. la partizione generata dai centroidi e' determinata partendo da questa distanza. In questo caso si puo' verificare che le frontiere della partizione di minima distanza sono lineari (iperpiani) ortogonali ai segmenti che uniscono i centroidi.

La bonta' di un raggruppamento $A_1, \ldots, A_G$ e' misurata con l'indice seguente

$$\sum_{g=1}^{G} \mathrm{dev}(A_g) = \sum_{g=1}^{G} \sum_{i \in A_g} d^2(\mathbf{x}_i, \overline{\mathbf{x}}_g)$$

che essendo la somma delle devianze interne ai gruppi e' (vedi 2.14.5) la devianza entro i gruppi. Si dimostra che l'algoritmo *k-means* converge, ossia che ad ogni passo fa decrescere il criterio della devianza entro i gruppi.

A rigore si tratta di un indice di dispersione e non di bonta' di raggruppamento. Tuttavia essendo la devianza entro i gruppi uguale alla devianza totale delle unita' meno la devianza tra gruppi che e' uguale a

$$\sum_{g} n_g d^2(\overline{\mathbf{x}}_g, \overline{\mathbf{x}})$$

minimizzare la devianza entro i gruppi equivale a massimizzare la devianza tra gruppi ossia a rendere massima la separazione dei gruppi.

A convergenza ottenuta, l'algoritmo si arresta a un punto di minimo, detto minimo locale, perche' non e' possibile essere certi che la partizione determinata che sia quella globalmente ottima (cioe' nell'ambito di tutte le partizioni in G classi).

Inoltre l'algoritmo puo' portare a soluzioni diverse a seconda della partizione iniziale scelta. Avendo eseguito vari tentativi con diverse partizioni iniziali, si scegliera, ovviamente, la soluzione che corrisponde al valore piu' basso dell'indice.

Esempio 2.18 Riprendiamo i dati analizzati precedentemente con metodi gerarchici e utilizziamo il metodo *k-means* nella versione sopra descritta, prendendo come partizione iniziale quella del metodo del legame medio. Il valore del criterio della devianza interna e' $70,9$ le numerosita' dei gruppi sono $(4, 8, 12, 11, 14)$. Il confronto col risultato del legame medio e' riportato nella tabella seguente.

	1	2	3	4	5	*# medio*
1	12	0	0	0	1	13
2	0	8	0	10	12	30
3	0	0	4	0	0	4
4	0	0	0	1	0	1
5	0	0	0	0	1	1
# k-means	12	8	4	11	14	49

Dunque il metodo *k-means* ha riallocato alcune unita' dei gruppi del legame medio (essenzialmente quelle del gruppo 2), mentre il gruppo 1 e' rimasto stabile. Alcune unita' isolate sono state riallocate all'interno di alcuni gruppi piu' numerosi. Vi e' una tendenza (che spesso si riscontra) a dare gruppi di uguale numerosita'. Un'altra

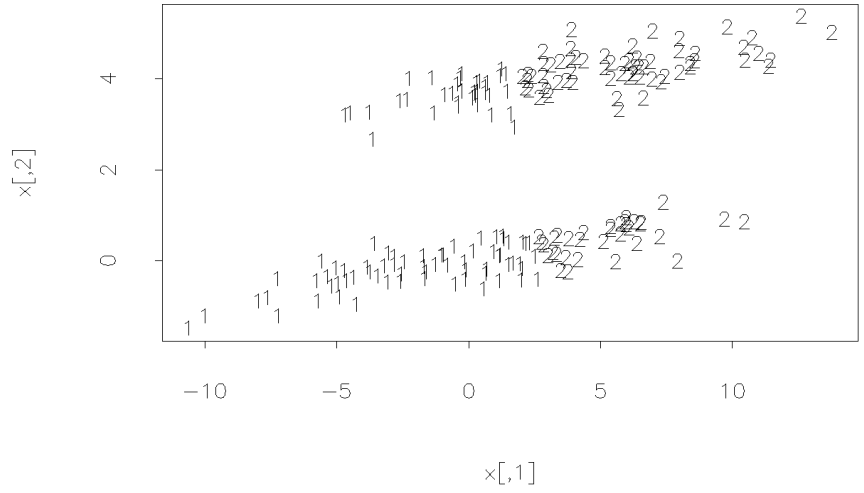

Figura 2.8: *Gruppi allungati*

tendenza tipica di questo metodo non gerarchico e' quella di formare gruppi sferici. Si consideri il grafico del grafico 2.8 in cui e' rappresentato lo scatter relativo a due variabili da cui si deduce visivamente l'esistenza di due gruppi abbastanza allungati.

I dati sono stati generati da due distribuzioni normali bivariate con medie $\boldsymbol{\mu}_1 = (0, 0)$ e $\boldsymbol{\mu}_2 = (4, 4)$ e matrice di varianze e covarianze

$$\mathbf{V} = \begin{bmatrix} 16.0 & 1.50 \\ 1.50 & 0.25 \end{bmatrix}.$$

I due gruppi sono abbastanza vicini lungo la dimensione di minor variabilita'. Cio' nonostante la separazione dei gruppi e' evidente.

Sul grafico i punti sono rappresentati con l'indice del gruppo assegnato dal metodo *k-means*. Il metodo ha formato due gruppi sferici mancando completamente i gruppi allungati. La causa e' da ricercarsi essenzialmente nel criterio della devianza interna ai gruppi basata sulla distanza Euclidea. Il metodo infatti "sente" piu' vicini i punti dell'altro gruppo, che non i punti sul bordo del suo gruppo.

Se avessimo utilizzato un metodo gerarchico, il criterio del legame singolo avrebbe ricostruito esattamente i due gruppi allungati, mentre il criterio del legame medio avrebbe fallito esattamente per lo stesso motivo del metodo *k-means*.

2.19 Note bibliografiche

L'analisi dei gruppi e' una tecnica nata fuori dell'ambito strettamente statistico. Gli statistici hanno sempre rilevato l'estrema indeterminatezza del concetto di gruppo e le difficolta' della scelta di un criterio obbiettivo.

Gordon (1981) fornisce un'introduzione completa e dal punto di vista statistico, mentre Hartigan (1975) presenta un'ampia raccolta di esempi stimolanti e di soluzioni, da un punto di vista piu' anticonformista. Una rassegna con una applicazioni a dati elettorali e' contenuta in Chiandotto (1978) e Chiandotto e Marchetti (1980).

Un campo che in queste pagine e' stato completamente trascurato e' quello dei modelli di classificazione e delle tecniche inferenziali ad essi collegate (cfr. tra gli altri McLachlan e Basford (1988)).

Riduzione di dimensionalità

3

3.1 Proiezioni ortogonali

Spesso dovendo rilevare dei caratteri su un gruppo di individui o di unità si e' tentati di raccogliere un gran numero di variabili senza pensare alla loro futura utilizzazione.

L'analisi preliminare di queste variabili e' estremamente difficoltosa se si vogliono studiare simultaneamente. Con l'analisi dei gruppi e' possibile classificare le unità, e ridurre la complessità dei dati, ma anche in tal caso la presenza di molte variabili e le correlazioni esistenti fra di esse creano molte difficoltà.

Ci occuperemo ora pertanto delle situazioni in cui si abbiano p variabili e non vi sia una variabile dipendente, ma si voglia in qualche modo condensare l'insieme dei dati riducendone le dimensioni.

Affronteremo questo problema da un punto di vista geometrico considerando prima i vettori delle unità $\mathbf{x}_i$. Volendo fare una rappresentazione grafica di questi vettori, cio' e' possibile fino a che la dimensione p e' minore o eguale a tre, come abbiamo visto nel capitolo 2, in caso contrario ci si deve accontentare di grafici di dimensione ridotta. Ad esempio possiamo rappresentare graficamente le p distribuzioni marginali unidimensionali (con un istogramma ad esempio) ed eventualmente le $p(p-1)/2$ distribuzioni marginali bivariate (con uno *scatterplot*).

Si noti che questo tipo di rappresentazioni e' ottenuta *proiettando* i vettori unità $\mathbf{x}_i$ sugli assi coordinati definiti dai vettori

$$\begin{aligned} \mathbf{e}_1 &= (1,0,0,\cdots,0)' \\ \mathbf{e}_1 &= (0,1,0,\cdots,0)' \\ &\vdots \end{aligned}$$

$$\mathbf{e}_1 \quad = \quad (0, 0, 0, \cdots, 1)'$$

D'altra parte le proiezioni ortogonali sugli assi non sempre permettono di capire esattamente le distribuzioni congiunte. Ci proponiamo allora di cercare delle proiezioni piu' "illuminanti" su degli assi diversi definiti da vettori inclinati rispetto a quelli canonici sopra definiti. Nel

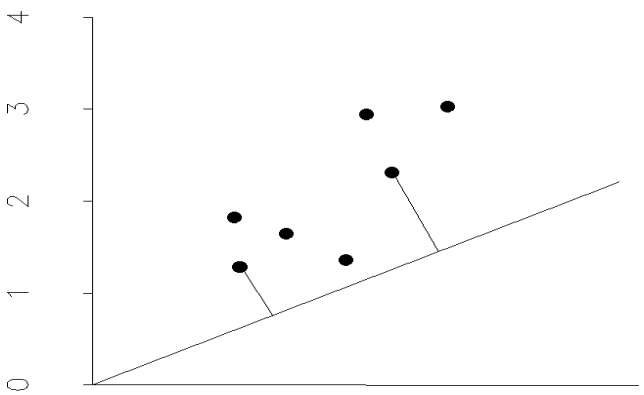

Figura 3.1: *Proiezione ortogonale su un vettore*

grafico 3.1 si sono rappresentate alcune unità bivariate e le loro proiezioni ortogonali su un asse.

Approfondiamo il concetto di proiezione ortogonale su un vettore $\mathbf{v}$. Supponiamo per semplicità che il vettore abbia lunghezza unitaria, cioe' che $\mathbf{u}'\mathbf{u} = 1$ e indichiamo con $c_i\mathbf{v}$ la proiezione ortogonale dell'unità $\mathbf{x}$ i sull'asse $\mathbf{v}$ (ovviamente la proiezione deve essere un multiplo del vettore che definisce l'asse).

Come si determina c_i? Semplicemente osservando che se la proiezione e' ortogonale, allora il vettore differenza $\mathbf{x}_i - c_i\mathbf{v}$ (che e' il segmento che scende dal punto sull'asse) deve essere ortogonale al vettore $\mathbf{v}$ stesso. Pertanto si avrà

$$(\mathbf{x}_i - c_i\mathbf{v})'\mathbf{v} = 0$$

da cui si ricava $c_i = \mathbf{x}_i{}'\mathbf{v}$. Ovviamente, le coordinate dei vettori unità sull'asse $\mathbf{v}$ sono date dagli scalari c_i.

Al termine dell'operazione ci ritroviamo con n determinazioni c_i che possono essere utilizzate come determinazioni di una nuova variabile e che permettono semplicemente di "vedere" i dati da quel particolare punto di vista definito dall'asse $\mathbf{v}$.

Le unità multivariate risultano proiettate dunque su una sola dimensione. La proiezione naturalmente non puo' conservare tutta l'informazione relativa alle variabili originali e quindi parte di essa viene perduta. Ridurre le dimensioni da p a una puo' sembrare piuttosto drastico, ma potremmo proiettare le unità su spazi a due, a tre, ecc. dimensioni.

Resta poi il problema della scelta dell'asse su cui proiettare i dati secondo un criterio che evidenzi la nostra necessità di partenza di condensare l'informazione e ridurre le dimensioni senza perdere molto.

Prima di affrontare i problemi sopra citati, osserviamo piu' da vicino la nuova variabile

$$\mathbf{c} = (c_1, \ldots, c_n)'$$

ora costruita per comprenderne meglio la natura. Risulta infatti che

$$\mathbf{c} = \begin{bmatrix} \mathbf{x}_1{}'\mathbf{v} \\ \mathbf{x}_2{}'\mathbf{v} \\ \vdots \\ \mathbf{x}_n{}'\mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1{}' \\ \mathbf{x}_2{}' \\ \vdots \\ \mathbf{x}_n{}' \end{bmatrix} \mathbf{v}.$$

Percio' $\mathbf{c} = \mathbf{X}\mathbf{v}$. Scriviamo ora la matrice delle osservazioni $\mathbf{X}$ usando i vettori delle variabili e otteniamo

$$\begin{aligned} \mathbf{c} &= (\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(p)}) \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_p \end{bmatrix} \\ &= v_1\mathbf{x}_{(1)} + v_2\mathbf{x}_{(2)} + \cdots + v_p\mathbf{x}_{(p)}. \end{aligned}$$

dove $v_1, \ldots, v_p$ sono le componenti del vettore $\mathbf{v}$ su cui si proietta. In conclusione le coordinate c_i sono ottenute mediante una *combinazione lineare delle variabili* originali [1].

3.2 La prima componente principale

Occupiamoci ora della definizione di un criterio che che permetta di misurare la perdita di informazione passando da p variabili a una sola.

Un criterio ragionevole e' basato sulla varianza della variabile $\mathbf{c}$. Risulta infatti che la varianza di $\mathbf{c}$ e' sempre minore della somma delle varianze delle variabili componenti. Dovendo riassumere le p variabili con l'unica variabile sintetica $\mathbf{c}$ vogliamo che la sua varianza sia la piu' grande possibile.

Se le variabili di partenza sono espresse in scarti dalla media, la varianza di $\mathbf{c}$ e' data dalla formula

$$s^2(\mathbf{c}) = 1/n\mathbf{c}'\mathbf{c} = 1/n\mathbf{v}'\mathbf{X}'\mathbf{X}\mathbf{v} = \mathbf{v}'\mathbf{S}\mathbf{v}.$$

Pertanto, la massimizzazione della varianza di $\mathbf{c}$ si traduce nel problema seguente

$$\max_{\mathbf{v}'\mathbf{v}=1} \mathbf{v}'\mathbf{S}\mathbf{v}$$

ossia nella massimizzazione della funzione (quadratica) $\mathbf{v}'\mathbf{S}\mathbf{v}$ rispetto a tutti i vettori $\mathbf{v}$ di lunghezza uno.

Il problema sopra enunciato puo' essere risolto in generale. L'asse ottimale, chiamiamolo $\mathbf{v}_1$, e' l'autovettore associato all'autovalore piu' grande della matrice di varianze e covarianze

[1] Analogamente alla regressione multipla (si veda il capitolo seguente). C'e' una differenza importante e cioe' che qui la variabile dipendente e' c_i, una variabile di sintesi di quelle date, mentre nella regressione multipla e' una variabile osservabile $\mathbf{y}$.

S. Esso si chiama *primo asse principale* [2]. Basta pertanto ricavare il primo autovalore λ_1 della matrice **S** e trovare un autovettore associato $\mathbf{v}_1$ di lunghezza 1.

La variabile $\mathbf{c}_1 = \mathbf{X}\ \mathbf{v}_1$ e' detta *prima componente principale* estratta dalle variabili. Essa non fa altro che raccogliere le coordinate delle unità su un nuovo sistema di riferimento unidimensionale in modo tale da massimizzarne la varianza.

Siccome $\mathbf{v}_1$ e' un autovettore di **S**, esso soddisfa all'identità

$$\mathbf{S}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$$

e dunque la varianza dei punti sulla prima componente principale e'

$$s^2(\mathbf{c}_1) = \mathbf{v}_1{}'\mathbf{S}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1{}'\mathbf{v}_1 = \lambda_1.$$

cioe' e' uguale all'autovalore piu' grande della matrice delle varianze e covarianze.

Abbiamo detto prima che questa varianza per quanto massimizzata e' sempre minore o uguale alla somma delle varianze delle variabili ed infatti quest'ultima e' semplicemente la somma degli elementi sulla diagonale di **S**. (Si ricordi che la somma di tutti gli autovalori $\lambda_1, \ldots, \lambda_p$ della matrice **S** e' uguale alla somma degli elementi sulla diagonale).

La somma delle varianze delle variabili e' un indice di variabilità globale che abbiamo già incontrato sotto altra forma. Infatti e' semplicemente la devianza dell'insieme U unità divisa per n:

$$\begin{aligned} \sum_j s_j^2 &= (1/n) \sum_i \sum_j x_{ij}^2 \\ &= (1/n) \sum \mathbf{x}_i{}'\mathbf{x}_i = (1/n)\mathrm{dev}(U). \end{aligned}$$

Percio' disponiamo anche di un indice relativo di bontà di rappresentazione: basta dividere la varianza della prima componente principale λ_1 per la somma delle varianze, ovvero per la somma degli autovalori di **S**

$$\tau_1 = \frac{\lambda_1}{\lambda_1 + \cdots + \lambda_p}.$$

L'indice τ_1 e' positivo e minore di uno e potrebbe anche essere uno nel caso in cui la matrice **S** abbia un solo autovalore diverso da zero. Cio' puo' capitare se le variabili sono tutte linearmente dipendenti.

Esempio 3.2 Per illlustrare la costruzione e il significato della prima componente principale si consideri l'esempio 1.2 riguardante 7 tipi di delinquenza in 16 città americane. Le variabili sono tutte dei rapporti su 100000 abitanti. Le medie e le varianze delle variabili sono le seguenti

	Omicidi	Stupri	Rapine	Aggress.	Furti	Truffe	F. d'auto
$\bar{x}_j$	9.7	28.1	243.5	196.2	1375.7	1003.6	689.1
s_j^2	24.10	145.4	24718.5	7131.7	93055.6	68427.6	22755.8

[2] Per chi non conosce cosa sia un autovalore di una matrice, e' sufficiente per il momento capire che il problema ha una soluzione determinabile $\mathbf{v}_1$.

Gli autovalori della matrice di varianze e covarianze sono i seguenti:

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
150714.6	30685.5	19274.6	11717.3	3814.1	42.1	10.1

Il rapporto percentuale $100 \times \tau_1 = 69.69$ tra la varianza della prima componente (il primo autovalore) e la somma delle varianze delle variabili (che e' uguale alla somma degli autovalori cioe' 216258.7) indica che la prima componente principale "spiega" circa il 70% della varianza complessiva. Pertanto, una sola variabile riassume il 70% della varianza delle variabili originali.

L'autovettore associato al primo autovalore ha le seguenti componenti

v_1	v_2	v_3	v_4	v_5	v_6	v_7
0.0019	0.017	0.18	0.10	0.74	0.61	0.15
Omicidi	Stupri	Rapine	Aggress.	Furti	Truffe	F. d'auto

che costituiscono i coefficienti della combinazione lineare delle variabili (sotto elencate) che vanno a formare la prima componente principale. La variabile cui viene attribuito il peso maggiore e' il tasso di furti che in effetti ha la varianza piu' elevata e che quindi domina la prima componente principale.

Per misurare l'entità della presenza di ciascuna variabile nella prima componente principale si possono calcolare i coefficienti di correlazione tra componente principale e variabili che risultano:

Omicidi	Stupri	Rapine	Aggress.	Furti	Truffe	F. d'auto
0.16	0.54	0.45	0.48	0.95	0.91	0.39

I coefficienti di correlazione consentono una interpretazione della prima componente principale che in questo caso e' correlata positivamente con tutte le variabili, ma essenzialmente ai furti e le truffe che dominano tutte le altre. Tuttavia, la prima componente e' influenzata pochissimo dal tasso di omicidi di cui praticamente non si tiene conto poiche' ha valori bassi (fortunatamente) e poco variabili rispetto agli altri. Questa sensibilità alle varianze ripropone il problema delle ponderazioni implicite delle variabili e la discussione sulla opportunità della standardizzazione.

Se vogliamo dare lo stesso peso a tutte le variabili standardizziamo i dati e calcoliamo la prima componente principale su di essi. La matrice delle varianze e covarianze diventa la matrice di correlazione e i suoi autovalori risultano stavolta i seguenti

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
3.45	1.33	0.94	0.62	0.36	0.17	0.11

La percentuale di varianza spiegata e' si ottiene rapportando 3.45 al totale delle varianze che stavolta e' 7 e risulta 49.3%. Percio' se tutte le variabili hanno lo stesso peso, e' piu' difficile che una sola le riassuma in buona percentuale e dunque la prima componente principale sulle variabili standardizzate spiega solo il 49.3%.

Inoltre anche i coefficienti della combinazione lineare cambiano diventando

v_1	v_2	v_3	v_4	v_5	v_6	v_7
0.28	0.43	0.38	0.46	0.38	0.34	0.31
Omicidi	Stupri	Rapine	Aggress.	Furti	Truffe	F. d'auto

(si noti il maggior equilibrio rispetto al caso non standardizzato).

E' chiaro pero' che in questo esempio, in cui tutte le variabili hanno sostanzialmente la stessa unità di misura, la scelta se standardizzare o meno equivale alla scelta di una ponderazione delle variabili e quindi e' in un certo senso questione di gusti del ricercatore. Dove questo aspetto soggettivo diventa piu' inquietante e' quando le variabili hanno *diverse* unità di misura e la scala (centimetri, metri) puo' essere scelta arbitrariamente). La prima componente principale, come abbiamo appena visto, non e' invariante al cambiamento di scala delle variabili e quindi dipende dall'unità di misura scelta. Questo fatto molto spiacevole consiglia in questi casi di standardizzare obbligatoriamente.

3.3 La seconda componente principale

La riduzione ad una sola componente principale spesso e' insufficiente. Nell'esempio precedente, mantenendo il peso implicito delle variabili, una sola componente spiega il 70% della variabilità, cioe' abbastanza (intuitivamente), ma, standardizzando, la percentuale di varianza spiegata scende al 49.3% e, stavolta, sembra (sempre intuitivamente) poco.

Ora, si dimostra anche che l'autovalore massimo di questa matrice e' uguale al *secondo* autovalore λ_2 (in ordine di grandezza) di $\mathbf{S}$, e cosi' pure l'autovettore associato $\mathbf{v}_2$ e' il secondo autovettore di $\mathbf{S}$. La seconda componente principale sarà dunque

$$\mathbf{c}_2 = \mathbf{X}\mathbf{v}_2.$$

ed essa per costruzione risulta ortogonale alla prima.

Dobbiamo dunque generalizzare il meccanismo di costruzione di una componente principale ad altre componenti. Una tecnica e' quella di generalizzare il procedimento di proiezione ortogonale dei punti unità $\mathbf{x}_i$ su una retta a proiezioni su un piano.

Come una retta e' definita da un asse $\mathbf{v}$ di lunghezza 1, e' comodo definire un piano qualsiasi mediante due vettori $\mathbf{v}_1$ e $\mathbf{v}_2$ di lunghezza 1 e ortogonali fra loro. Infatti, come si puo' facilmente verificare la proiezione di un vettore $\mathbf{x}_i$ sul piano e' il punto

$$c_{i1}\mathbf{v}_1 + c_{i2}\mathbf{v}_2$$

dove $c_{i1}\mathbf{v}_1$ e' la proiezione ortogonale di $\mathbf{x}_i$ su $\mathbf{v}_1$ e $c_{i2}\mathbf{v}_2$ e' la proiezione ortogonale di $\mathbf{x}_i$ su $\mathbf{v}_2$. Percio' le coordinate del vettore unità sul piano sono (c_{i1}, c_{i2}). Per quanto detto nella sezione 3.1 risultano definite due variabili $\mathbf{c}_h = \mathbf{X}\mathbf{v}_h (h = 1, 2)$ combinazioni lineari delle variabili originali e ortogonali fra loro (perché $\mathbf{v}_1 \perp \mathbf{v}_2$).

Per determinare le due variabili in modo ottimale occorre introdurre una generalizzazione del criterio della varianza. La scelta naturale e' la somma delle varianze delle due combinazioni lineari, ovvero della varianza multivariata dei punti proiettati che risulta

$$\sum_{h=1}^{2} \mathbf{v}_h{}'\mathbf{S}\mathbf{v}_h.$$

Questo criterio va massimizzato rispetto a ai vettori $\mathbf{v}_h$ di lunghezza unitaria e ortogonali fra loro.

Il problema di massimizzazione ammette una soluzione rappresentata dagli autovettori associati ai primi due autovalori della matrice di varianze e covarianze $\mathbf{S}$. In questo modo vengono costruite due componenti principali $\mathbf{c}_1$, identica a prima, e $\mathbf{c}_2$ ortogonale ad essa, chiamata ovviamente *seconda componente principale*. I due primi autovalori corrispondono alle varianze delle due componenti. L'indice

$$\tau_1 + \tau_2 = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \cdots + \lambda_p}$$

indica la quota di varianza complessiva spiegata dalle prime due componenti.

Esempio 3.3 Riprendendo l'esempio dei dati sulla delinquenza, e considerando i dati standardizzati, vediamo che la seconda componente principale ha una varianza $\lambda_2 = 1.33$ pari al 19.03% della varianza totale. L'insieme della prima e della seconda componente principale hanno una varianza complessiva che spiega il 68.34% della varianza totale. Le componenti del secondo autovettore della matrice di correlazione sono

v_1	v_2	v_3	v_4	v_5	v_6	v_7
-0.60	-0.06	-0.19	-0.26	0.39	0.59	0.11
Omicidi	Stupri	Rapine	Aggress.	Furti	Truffe	F. d'auto

da cui si calcola la seconda componente principale. Dalle correlazioni tra le due componenti e le variabili qui sotto riportate

	Omicidi	Stupri	Rapine	Aggress.	Furti	Truffe	F. d'auto
$\mathbf{c}_1$	0.53	0.81	0.7	0.85	0.72	0.64	0.57
$\mathbf{c}_2$	-0.776	-0.07	-0.2	-0.30	0.45	0.68	0.13

risulta che mentre la prima componente principale e' una sorta di media delle variabili correlata positivamente a tutte (una specie di indicatore dell'intensità della delinquenza), la seconda principale invece oppone le prime quattro variabili a cui e' correlata inversamente (cioe' Omicidi, stupri, rapine e aggressioni) alle altre tre con cui la correlazione e' positiva. Percio' all'aumentare della seconda componente aumentano in media i reati contro il patrimonio, mentre diminuiscono quelli contro la persona e viceversa.

Le due componenti principali si possono rappresentare con uno scatter riportato nel grafico 3.2. I punti sullo scatter (qui etichettati col nome della città) sono esattamente le proiezioni delle unità sul piano definito dai due assi principali.

Si noti che, a seconda del programma usato per estrarre autovalori e autovettori, il grafico puo' risultare anche ribaltato rispetto a uno o entrambi gli assi perche' gli autovettori sono definiti a meno del segno e la scelta del segno e' arbitraria.

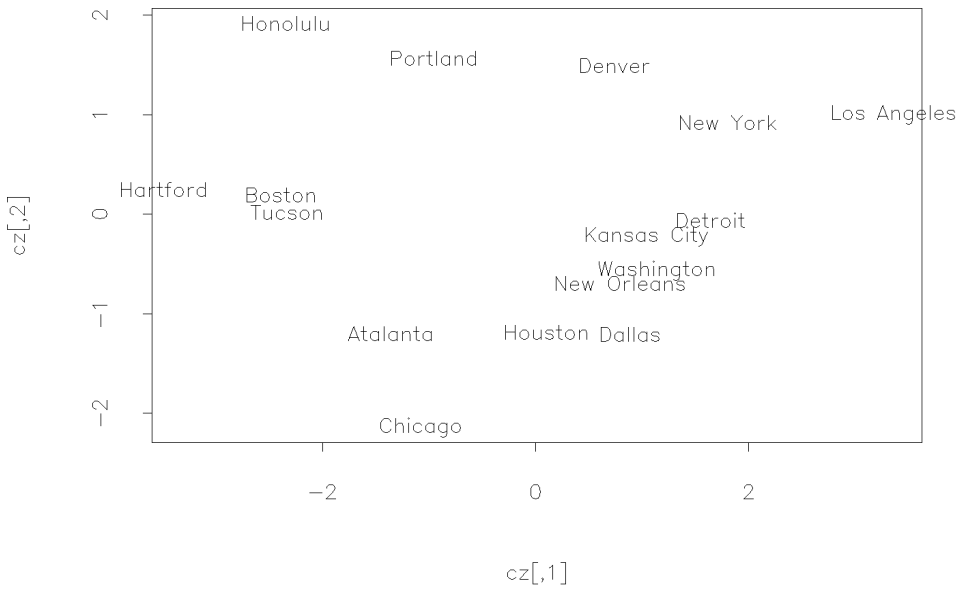

Figura 3.2: *Scatter delle componenti principali*

3.4 Scelta del numero di componenti

E' ovvio che il processo di estrazione delle componenti principali non e' per forza limitato alle prime due. L'estrazione della terza, della quarta, ecc. componente avviene senza difficolta' iterando lo schema che ormai dovrebbe essere familiare.

La terza componente principale e' una combinazione delle variabili con coefficienti uguali alle componenti dell'autovettore associato al terzo autovalore della matrice di varianze e covarianze, e cosi' via.

Il numero massimo di componenti che si possono estrarre e' esattamente uguale al numero di variabili. Non e' detto che tutti gli autovalori della matrice $\mathbf{S}$ che sono sempre ≥ 0) siano diversi da zero. Talvolta, alcuni sono nulli e cio' indica che la matrice di varianze e covarianze, e di riflesso anche l'insieme delle variabili, contengono delle dipendenze lineari. Un esempio tipico si ha quando le somme per riga della matrice $\mathbf{X}$ sono costanti: se le unita' sono i comuni di una regione e come variabili si rilevano le percentuali di voto ai partiti in occasione di una consultazione elettorale, la somma delle righe e' uguale a 100. In questo caso il rango della matrice $\mathbf{X}$ non e' piu' p, ma $p - 1$.

Siccome gli autovalori sono le varianze delle componenti, e' interessante studiare come la varianza complessiva, che e' la somma degli autovalori, si *concentra* nelle prime componenti. Si puo' usare un grafico speciale rappresentando la percentuale di varianza spiegata. Tale diagramma per l'esempio dei crimini e' riportato nel grafico 3.3. Solitamente si cerca sul diagramma il punto in cui vi e' una caduta brusca seguita da una curva con poca pendenza. Intuitivamente questo e' un criterio *ad hoc* per determinare il numero di componenti da estrarre.

Ovviamente, estraendo tutte le componenti si ottengono p componenti ortogonali la cui

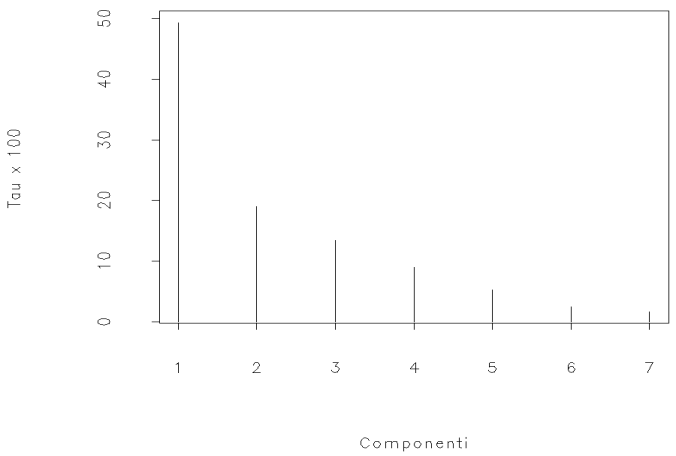

Figura 3.3: *Percentuale di varianza spiegata*

varianza totale e' uguale a quella delle variabili. Non solo, ma le componenti principali sono date da

$$(\mathbf{c}_1|\mathbf{c}_2|\cdots|\mathbf{c}_p) = \mathbf{X}(\mathbf{v}_1|\mathbf{v}_2|\cdots|\mathbf{v}_p) = \mathbf{XV}$$

dove $\mathbf{V}$ e' la matrice degli autovettori e, poiche' tale matrice e' una matrice di una rotazione, estrarre tutte le componenti principali corrisponde a ruotare il sistema di riferimento in modo che gli assi si trovino lungo le dimensioni a maggior variabilita'.

Un modo per controllare quanto si perde rinunciando a estrarre delle componenti e' quello di calcolare il potenziale di previsione delle componenti che e' uguale alla somma degli indici di determinazione lineare r^2_{hj} tra la componente h-esima e la variabile j-esima. Essi si ottengono elevando al quadrato i coefficienti di correlazione tra componenti e variabili.

	Omicidi	Stupri	Rapine	Aggress.	Furti	Truffe	F. d'auto
$\mathbf{c}_1$	0.28	0.66	0.51	0.73	0.52	0.42	0.33
$\mathbf{c}_2$	0.49	0.00	0.05	0.10	0.21	0.47	0.02
$\mathbf{c}_3$	0.07	0.05	0.18	0.00	0.10	0.02	0.51
$\mathbf{c}_4$	0.04	0.23	0.19	0.03	0.06	0.01	0.05
$\mathbf{c}_5$	0.08	0.00	0.05	0.08	0.09	0.02	0.09
$\mathbf{c}_6$	0.02	0.00	0.00	0.04	0.05	0.05	0.00
$\mathbf{c}_7$	0.00	0.03	0.02	0.03	0.00	0.02	0.00

Dalla tabella vediamo che la prima componente spiega il 28% della prima variabile, il 66% della seconda e cosi' via. Se decidiamo di considerare solo due componenti e di scartare tutte le altre, dalla tabella vediamo quale delle variabili "gettiamo via". Per esempio, scartando la terza componente che spiega solo una piccola frazione della varianza delle prime sette variabili, ma che spiega il 51% dell'ultima, sappiamo che stiamo scartando informazione essenzialmente dalla variabile 'furti d'auto'.

Si osservi anche che la somma per colonna degli indici di determinazione lineare deve essere eguale a 1 perche' le componenti sono ortogonali e tutte insieme chiaramente predicono

esattamente ciascuna delle variabili. Allora e' molto utile considerare anche la tabella seguente cumulata per colonna.

	Omicidi	Stupri	Rapine	Aggress.	Furti	Truffe	F. d'auto
$\mathbf{c}_1$	0.28	0.66	0.51	0.73	0.52	0.42	0.33
$\mathbf{c}_2$	0.77	0.66	0.56	0.83	0.73	0.88	0.35
$\mathbf{c}_3$	0.85	0.72	0.74	0.83	0.83	0.90	0.86
$\mathbf{c}_4$	0.89	0.95	0.93	0.86	0.89	0.92	0.91
$\mathbf{c}_5$	0.97	0.96	0.98	0.94	0.94	0.93	1.00
$\mathbf{c}_6$	0.99	0.96	0.98	0.97	1.00	0.98	1.00
$\mathbf{c}_7$	1.00	1.00	1.00	1.00	1.00	1.00	1.00

3.5 Componenti principali e analisi dei gruppi

La funzione delle componenti principali e' dunque quella della riduzione di dimensionalita' di una variabile multipla $X_1, \ldots, X_p$. Questo suggerisce che le componenti principali siano utili anche per rappresentare graficamente i dati multivariati al fine di determinare dei gruppi. Tuttavia, questo non e' vero basta pensare che il criterio dell'analisi in componenti principali e' quello della varianza totale e tale criterio non tiene conto in alcun modo della possibile esistenza di gruppi nei dati. Illustriamo questo punto con i dati dell'esempio 2.18. I dati

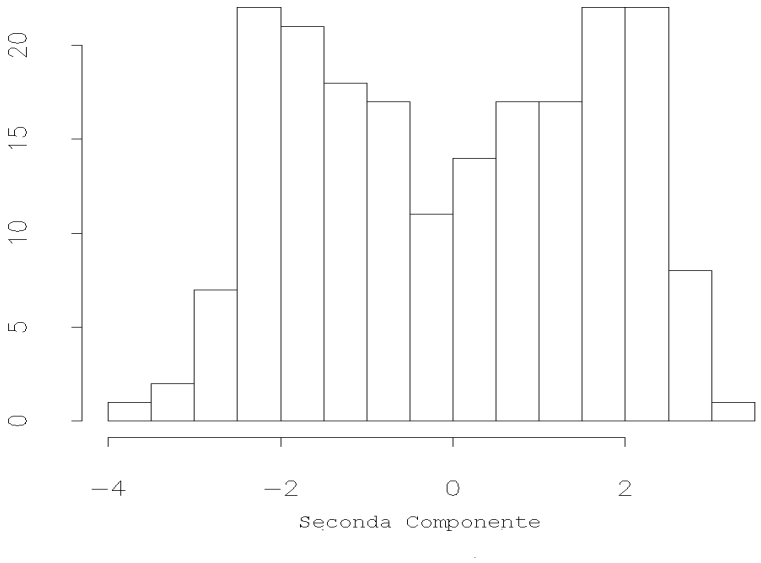

Figura 3.4: *Istogramma della seconda componente principale*

di questo esempio sono bidimensionali, percio' non ci sarebbe alcun problema di analisi in componenti principali, tuttavia supponiamo di essere costretti ad usare una sola componente e di vedere se i gruppi sono evidenti lungo questa componente. E' evidente che la prima componente non consente di vedere i gruppi perche' essi sono allungati nel senso della prima componente. Nel grafico 3.4 e' riportato l'istogramma relativo alla seconda componente

principale. Come si vede i gruppi non sono particolarmente evidenti. Si consideri ora un asse

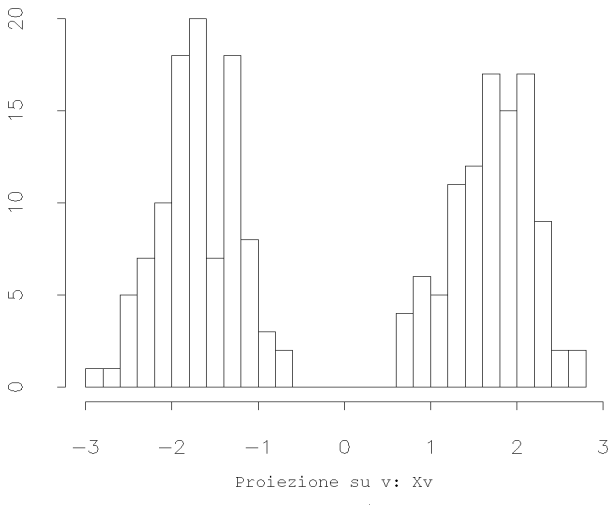

Figura 3.5: *Istogramma lungo la direzione* $\mathbf{v}$

diverso, definito dal vettore

$$\mathbf{v} = (-0, 168, 0.986)$$

La variabile ottenuta proiettando i punti su quest'asse e' rappresentata nell'istogramma del grafico 3.5 in cui i due gruppi emergono chiaramente. Dunque, anche in questo semplice esempio vediamo che l'analisi in componenti principali non e' particolarmente utile per *scoprire* i gruppi. Esistono infatti metodi migliori per proiettare i punti in modo da ottimizzare criteri di classificazione.

3.5.1 Distanza di Mahalanobis

Nel capitolo sulle distanze abbiamo parlato a lungo delle ponderazioni implicite e abbiamo concluso dicendo che le correlazioni fra variabili comportano a loro volta delle ponderazioni perche' piu' variabili misurano la stessa dimensione.

L'analisi in componenti principali trasforma le variabili di partenza in variabili ortogonali (e incorrelate perche' a media zero per costruzione). Percio' viene spontaneo pensare di ricalcolare le distanze *dopo* aver estratto (tutte) le componenti principali.

Allora, il risultato seguente e' degno di nota. La distanza Euclidea tra le unita', dopo aver estratto le componenti principali *e averle standardizzate* risulta la seguente:

$$D^2(i, i') = (\mathbf{x}_i - \mathbf{x}_{i'})' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_{i'}).$$

Essa prende il nome di *distanza generalizzata di Mahalanobis*. Si osservi che se le variabili sono incorrelate, la matrice di varianze e covarianze e' diagonale ed essa si riduce alla distanza tra le unita' dopo aver standardizzato le variabili. La distanza di Mahalanobis non risente delle correlazioni tra variabili, ma ha la tendenza negativa (gia' vista per la standardizzazione) ad attenuare le differenze tra gruppi, se esistono.

3.6 Approssimazioni di matrici

La soluzione del problema delle componenti principali e' strettamente collegato con l'approssimazione di matrici con matrici di rango minore. Come noto il rango di una matrice e' il numero massimo di colonne (o di righe) linearmente indipendenti. In una matrice $\mathbf{A}$ di dimensioni $I \times J$ il rango r non puo' superare il piu' piccolo dei due valori I, J.

Ora ogni matrice $\mathbf{A}$ puo' essere scomposta unicamente come segue

$$\mathbf{A} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1{}' + \sigma_2 \mathbf{u}_2 \mathbf{v}_2{}' + \cdots + \sigma_1 \mathbf{u}_r \mathbf{v}_r{}'$$

dove gli $\mathbf{u}_h$ sono di dimensione I e i $\mathbf{v}_h$ sono di dimensione J, $(h = 1, \ldots, r)$, tutti di lunghezza 1 e, separatamente, mutuamente ortogonali e $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, detti *valori singolari* della matrice $\mathbf{A}$.

Ad esempio la matrice

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

si scompone nella somma

$$\begin{aligned} \mathbf{A} &= \sqrt{3} \begin{bmatrix} -1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \\ 0 \end{bmatrix} \left[-\sqrt{2/3}, -1/\sqrt{6}, -1/\sqrt{6} \right] + \\ &+ \sqrt{3} \begin{bmatrix} -1/\sqrt{6} \\ 1/\sqrt{6} \\ -\sqrt{2/3} \\ 0 \end{bmatrix} \left[0, 1/\sqrt{2}, -1/\sqrt{2} \right]. \end{aligned}$$

L'interesse di questa scomposizione detta *scomposizione di Housholder-Young* o *scomposizione in valori singolari*, risiede nel fatto che se si scartano gli ultimi addendi della somma e si mantengono diciamo i primi r^* si ottiene una matrice $\mathbf{A}_{[r^*]}$ che approssima la matrice data $\mathbf{A}$ nel senso dei minimi quadrati tra tutte le matrici di rango r^*. Ossia, $\mathbf{A}_{[r^*]}$ rende minima la somma dei quadrati

$$\sum_i \sum_j (a_{ij} - b_{ij})^2$$

tra tutte le matrici $\mathbf{B}$ di rango almeno uguale a r^*. Inoltre, la somma dei quadrati di tutti valori singolari e' uguale alla somma dei quadrati degli elementi di $\mathbf{A}$.

3.6.1 Collegamento con le componenti principali

La tecnica dell'approssimazione di una matrice con un'altra di rango inferiore puo' essere adottata per una matrice di osservazioni quantitative $\mathbf{X}$ in cui supporremo che le colonne siano espresse in scarti dalle medie. Sia $X_{[1]}$ l'approssimazione di rango 1 di $\mathbf{X}$

$$\mathbf{X}_{[1]} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1{}'.$$

Allora e' semplice dimostrare che $\sigma_1 \mathbf{u}_1$ e' esattamente uguale alla prima componente principale $\mathbf{c}_1$, mentre $\mathbf{v}_1$ e' il primo asse principale. Inoltre la varianza della prima componente principale e'

$$\lambda_1 = \frac{\sigma_h^2}{n}.$$

Analogamente la h-esima componente principale e' semplicemente $\sigma_1 \mathbf{u}_1$. Percio' la scomposizione di in valori singoilari della matrice $\mathbf{X}$ ha la seguente struttura

$$\mathbf{X} = \mathbf{c}_1 \mathbf{v}_1{}' + \mathbf{c}_2 \mathbf{v}_2{}' + \cdots + \mathbf{c}_r \mathbf{v}_r{}'$$

e la matrice viene ricostruita sommando matrici ciascuna di rango 1 aventi la forma di un prodotto esterno (cioe' del tipo $\mathbf{x}\mathbf{y}'$) tra h-esima componente principale e h-esimo asse principale.

Questa tecnica dell'approssimazione di matrici ci sara' molto utile per spiegare l'analisi delle corrispondenze. Per il momento osserviamo che la scomposizione e' essenzialmente unica, anche se trasponiamo la matrice $\mathbf{X}$. Se i vettori $\mathbf{c}_h = \sigma_h \mathbf{u}_h$ di dimensione $(n \times 1)$ consentono di rappresentare le righe della matrice, i vettori $\sigma_h \mathbf{v}_h$ consentono di rappresentare le righe della matrice trasposta cioe' le colonne di $\mathbf{X}$.

3.7 Analisi delle corrispondenze

L'analisi delle corrispondenze e' un metodo di analisi delle tabelle di contingenza. La maggior differenza tra l'analisi delle corrispondenze ed altri metodi per l'analisi di dati categorici (come i modelli log-lineari) sta nell'impostazione tipicamente descrittiva della prima. Tuttavia la differenza non deve essere esagerata troppo perche' da una parte esistono delle versioni *model based* dell'analisi delle corrispondenze e dall'altra la pretesa dell'analisi delle corrispondenze di non fare assunzioni e di "far parlare i dati da soli" non corrisponde al vero.

L'analisi delle corrispondenze e' una tecnica con cui e' possibile rappresentare graficamente le distribuzioni parziali delle righe e delle colonne di una tabella doppia di contingenza. Le righe e le colonne della tabella di contingenza possono essere rappresentate come punti in uno spazio a due (o piu') dimensioni. Pertanto, le coordinate di questi punti vanno a costiture dei punteggi assegnati alle modalita' dei due caratteri incrociati. Inoltre, le coordinate sono costruite in modo tale da approssimare sul grafico le distanze chi-quadrato (vedi 2.8) tra profili riga o profili colonna.

Consideriamo una tabella di contingenza $\mathbf{F}$ per due caratteri categorici A e B rispettivamente di I e J modalita' e siano f_{ij} le frequenze relative congiunte. Conosciamo la definizione di distanza chi-quadrato tra i profili riga f_{ij}/f_{i+} della tabella. Analoga distanza si definisce tra i profili colonna f_{ij}/f_{+j}.

Osserviamo che i profili riga sono vincolati ad avere somma 1 e percio' vi sono delle dipendenze nella matrice $\mathbf{F}$. Ora e' possibile assegnare $I - 1$ coordinate $\mathbf{r}_i$ ad ogni modalita' di riga in modo tale che le distanze Euclidee tra questi vettori riga sia uguale alle distanze chi-quadrato tra le distribuzioni corrispondenti

$$d^2(\mathbf{r}_i, \mathbf{r}_i{}') = d_\chi^2(i, i').$$

Si osservi che la distribuzione marginale $\{f_{+j}\}$ e' la media delle distribuzioni parziali di riga ponderate con le frequenze marginali di colonna f_{i+}

$$f_{+j} = \sum_i \frac{f_{ij}}{f_{i+}} f_{i+}$$

A questa distribuzione marginale vengono assegnate coordinate nulle e localizzata nell'origine.

Una volta costruita la rappresentazione delle righe della tabella come punti in uno spazio a $I-1$ dimensioni si usa la distanza chi-quadato per interpretare la configurazione dei punti. Quando due punti riga sono vicini tra loro, i corrispondenti profili riga devono essere molto simili fra loro e dunque devono avere una struttura per colonna uguale.

Se due punti riga sono lontani, i profili avranno una struttura per colonna diversa. Se un punto riga e' vicino all'origine, ha un profilo riga simile al profilo marginale. Se due punti riga stanno da parti opposte rispetto all'origine, cio' significa che deviano dal profilo marginale in relazione a colonne diverse.

Cio' che abbiamo detto per le righe puo' essere ripetuto per i profili colonna. L'analisi delle corrispondenze e' infatti simmetrica ed e' possibile una rappresentazione delle colonne come punti in uno spazio a $J-1$ dimensioni in cui l'origine e' situata nella media ponderata dei profili colonna e in cui le distanze Euclidee si interpretano come distanze chi-quadrato nel modo prima visto.

Le coordinate dei punti riga e colonna si determinano con una procedura molto simile a quella delle componenti principali per una matrice di misure. Useremo la tecnica della scomposizione in valori singolari.

3.7.1 Indipendenza

Come noto due mutabili (casuali) A e B si dicono indipendenti se la probabilita' che un'unita' sia classificata contemporaneamente nella modalita' i di A e j di B e' uguale al prodotto delle probabilita'. Lo scostamento dalla situazione di indipendenza e' spesso misurato tramite le *contingenze* relativizzate

$$e_{ij} = \frac{f_{ij} - \hat{f}_{ij}}{\hat{f}_{ij}^{1/2}}$$

dove $\hat{f}_{ij} = f_{i+} f_{+j}$ sono le frequenze relative stimate sotto l'ipotesi di indipendenza. Questi rapporti misurano gli scostamenti tra le frequenze osservate e quelle attese in caso di indipendenza tra i caratteri. Quanto piu' piccoli sono e tanto piu' vicini si e' alla situazione di indipendenza.

La somma dei quadrati di questi valori e' uguale al coefficiente di contingenza quadratica di Pearson, ϕ^2 che come e' noto e' una misura dell'associazione tra A e B. Inoltre l'indice $X^2 = n\phi^2$ dove n e' il numeor totale di unita' classificate, e' il cosiddetto indice chi-quadro.

> *Esempio* 3.7 Consideriamo i dati della tabella 3.1 che riguarda un'indagine svolta nel 1971 tramite questionario su 1554 israeliani classificati secondo due mutabili, la prima riguardante 'la principale preoccupazione', e la seconda 'la residenza propria e del padre'. Il primo carattere ha una modalita' aggiuntiva:'piu' di una preoccupazione'. Nella

Tabella le sigle per le colonne indicano la residenza: ASIAF = Asia o africa, EUSA = Europa o Stati Uniti, IS-AA = Israele, padre in Asia o Africa, IS-EU = Israele, padre in Europa o America, IS-IS = Israele, padre in Israele. Le contingenze relativizzate sono le seguenti

0.0021	0.0072	−0.0043	−0.0048	−0.0198
0.0046	0.0052	−0.0052	−0.0100	−0.0093
−0.0345	0.0582	−0.0235	−0.0557	0.0066
−0.0857	0.0552	−0.0188	0.0332	0.0023
−0.0184	0.0114	0.0055	−0.0030	0.0101
−0.0153	−0.0361	0.0132	0.0846	0.0147
−0.0062	0.0281	−0.0143	−0.0181	−0.0399
0.1549	−0.1254	0.0524	−0.0319	0.0219

che evidentemente sono molto piccole a parte quella corrispondente alla cella della situazione economica personale per gli israeliani che stanno in Africa o in Asia. L'indice $\phi^2 = 0.077$ e' a sua volta molto basso tuttavia il valore dell'indice X^2 e' 120.4 con 28 gradi di liberta' e quindi significativo.

In questa situazione dunque la tavola ha una struttura molto vicina all'indipendenza, tranne che per qualche frequenza che rende l'indice chi-quadro (in questo campione abbastanza grande) significativo.

In questi casi l'analisi delle corrispondenze puo' contribuire meglio a individuare le attrazioni tra modalita'. Infatti il punto di partenza e' proprio la tabella dei residui relativizzati rispetto al modello d'indipendenza $\mathbf{E} = (e_{ij})$ che viene rappresentata con la scomposizione in valori singolari

$$\mathbf{E} = \sum_{h=1}^{r} \sigma_h \mathbf{u}_h \mathbf{v}_h{}'$$

dove r e' il rango della matrice dei residui che e' minore o uguale a $\min\{I-1, J-1\}$.

Le coordinate dei punti riga e dei punti colonna si ottengono come nell'analisi in componenti principali rispettivamente mediante i vettori $\mathbf{u}_h$ e $\mathbf{v}_h$, ma introducendo una normalizzazione con l'inverso della radice delle frequenze marginali. Indichiamo con $\mathbf{r}_h$ e $\mathbf{c}_h$ i vettori

	ASIAF	EUSA	IS-AA	IS-EU	IS-IS
Arruolamento	61	104	8	22	5
Sabotaggio	70	117	9	24	7
Situazione militare	97	218	12	28	14
Situazione politica	32	118	6	28	7
Situazione economica	4	11	1	2	1
Altro	81	128	14	52	12
Piu' di una	20	42	2	6	0
Ristrettezze	104	48	14	16	9

Tabella 3.1: *Dati sui principali problemi degli Israeliani*

delle coordinate delle righe e delle colonne della tabella rispettivamente di dimensione $(I \times 1)$ e $(J \times 1)$. Le loro formule collegate alla scomposizione in valori singolari sono le seguenti

$$\begin{aligned} \mathbf{r}_h &= \text{diag}(f_{i+}^{-1/2})\sigma_h \mathbf{u}_h \\ \mathbf{c}_h &= \text{diag}(f_{+j}^{-1/2})\sigma_h \mathbf{v}_h \end{aligned}$$

Da queste definizioni risulta che i vettori delle coordinate delle righe (delle colonne) hanno medie ponderate con pesi f_{i+} (f_{+j}) nulle. Inoltre tali vettori (che corrispondono in questo senso alle componenti principali) hanno varianze uguali a σ_h^2.

I punteggi delle righe e delle colonne sono collegati fra di loro dalle formule seguenti dette formule di transizione

$$\begin{aligned} \mathbf{r}_h &= \frac{1}{\sigma_h}\mathbf{F}_r \mathbf{c}_h \\ \mathbf{c}_h &= \frac{1}{\sigma_h}\mathbf{F}_c' \mathbf{r}_h \end{aligned}$$

in cui le matrici $\mathbf{F}_r$ e $\mathbf{F}_c$ sono le matrici rispettivamente dei profili riga (distribuzioni condizionate per riga) e dei profili colonna (distribuzioni condizionate per colonna).

Queste formule si interpretano dicendo che, a meno del fattore moltiplicativo $1/\sigma_h$, la coordinata di una modalita' i di un carattere e' la media ponderata delle coordinate delle categorie dell'altro carattere con pesi uguali alle frequenze condizionate relative di i. Dunque le coordinate dei punti riga sono medie ponderate delle coordinate dei punti colonna e viceversa.

Infine le distanze euclidee tra punti riga o tra punti colonna sono le distanze chi-quadrato tra profili riga o tra profili colonna. Per questo, le rappresentazioni grafiche delle coordinate vengono spesso sovrapposte, anche se in realta' i punti riga e i punti colonna stanno in due spazi diversi.

Le formule di transizione sono usate per interpretare le distanze tra punti riga oppure tra punti colonna. Se un profilo riga e' eguale al profilo marginale, la prima formula di transizione dice che il punto riga deve essere la media ponderata delle colonne, cioe' deve stare nell'origine.

Se un profilo riga ha la frequenza parziale della colonna j piu' alta di quella marginale, questa colonna attrarra' il punto riga in quella direzione. Questo avviene se

$$\frac{f_{ij}}{f_{i+}} > f_{+j}$$

ovvero (moltiplicando sopra e sotto per f_{i+}) se

$$f_{ij} > \hat{f}_{ij}.$$

Quindi se il residuo dal modello di indipendenza e' positivo la riga i risultera' attratta dalla colonna j e viceversa: in generale quanto maggiore e' la frequenza osservata rispetto a quella attesa e tanto maggiore sara' la vicinanza dei punti i e occorre stare bene attenti a queste interpretazioni perche' il criterio della analisi delle corrispondenze e' definito in termini di distanze entro le righe o entro le colonne e non in termini di distanze tra righe e colonne.

Dato che la somma dei quadrati di tutti valori singolari σ_h e' uguale alla somma dei quadrati dei residui contenuti in $\mathbf{E}$, essa e' uguale all'indice ϕ^2. Pertanto ciascun vettore di coordinate $\mathbf{r}_h$ ed $\mathbf{c}_h$, (aventi varianza eguale a σ_h^2) contribuisce al coefficiente ϕ^2 per una parte uguale a

$$\frac{\sigma_h^2}{\sum \sigma_h^2}.$$

Si osservi infine che data la costruzione precedente, l'analisi delle corrispondenze e' utile se la tavola dei residui contiene degli elementi sistematici. Se vi e' indipendenza tra i due caratteri, la matrice $\mathbf{E}$ dovrebbe contenere solo elementi accidentali e quindi l'analisi delle corrispondenze non dovrebbe essere utilizzata.

3.8 Contributi assoluti e relativi

Esempio 3.8 Riprendiamo l'esempio delle preoccupazioni degli Israeliani a seconda della residenza e scomponiamo la matrice dei residui con l'analisi delle corrispondenze. Le quote di ϕ^2 spiegate dalle componenti estratte sono riportate nella tavola seguente.

	1	2	3	4
σ_h	0.059	0.015	0.0024	0.0001
%	77	19.8	3.1	0.1

Pertanto, la rappresentazione bidimensionale dell'analisi delle corrispondenze che spiega il 96% circa del ϕ^2 e' quasi esatta. Tale rappresentazione e' riportata nel grafico 3.6. Il

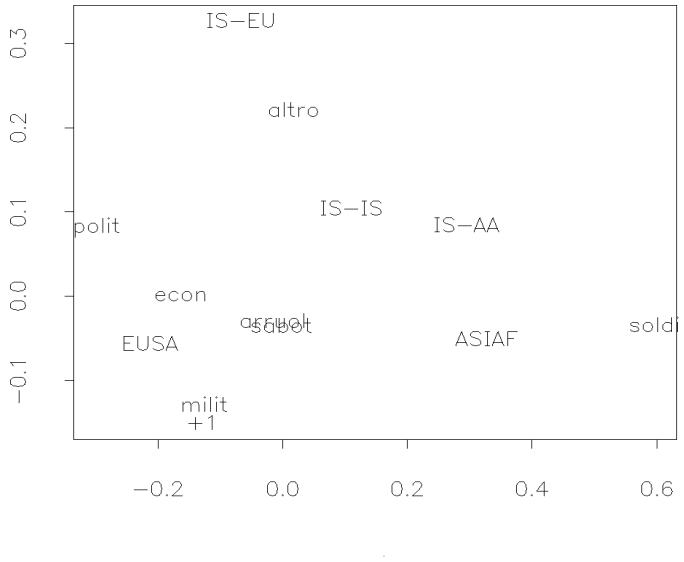

Figura 3.6: *Grafico della prime due componenti dell'analisi delle corrispondenze.*

grafico mette in evidenza che il primo asse e' determinato dall'opposizione tra le preoccupazioni relative alle ristrettezze personali ('soldi') e quelle relative alle situazioni

politiche e militari. A queste corrispondono l'opposizione tra gli israeliani che risiedono in Asia o Africa e quelli che risiedono in Europa o America. L'interpretazione sembra chiara: i problemi di natura piu' ampia sono sentiti dagli israeliani dei paesi occidentali, mentre quelli che vivono nei paesi in via di sviluppo hanno piuttosto problemi di situazione economica personale.

La seconda dimensione separa chi vive in Israele da chi vive fuori Israele. La causa e' essenzialmente la risposta 'altre preoccupazioni'. Quasi che i reali problemi di degli israeliani che vivono in Israele (con genitori in America o Europa) fossero difficilmente riconducibili alla classificazione prevista nell'indagine.

L'interpretazione degli assi di un'analisi delle corrispondenze e' facilitata introducendo certi indici chiamati *contributi* dei punti (riga o colonna) all'asse. Si vuole cioe' misurare qual e' il contributo del punto riga i alla varianza della componente $\mathbf{r}_h$ cioe' al quadrato del valore singolare σ_h^2. E, analogamente, qual e' il contributo del punto colonna alla varianza di $\mathbf{c}_h$ che e' sempre σ_h^2. Basta allora usare la relazione

$$\sigma_h^2 = \sum_{i=1}^{I} r_{ih}^2 f_{i+} = \sum_{j=1}^{J} c_{jh} f_{+j}$$

cioe' semplicemente la formula della varianza e definire i contributi riga all'asse h come

$$\text{CTR}(i) = \frac{1}{\sigma_h^2} r_{ih}^2 f_{i+}$$

e i contributi colonna all'asse h come

$$\text{CTR}(j) = \frac{1}{\sigma_h^2} c_{jh}^2 f_{+j}.$$

Le categorie con i contributi piu' forti saranno considerate come costitutive dell'asse h. Una buona regola e' quella di mettere in evidenza le righe o le colonne in cui i contributi sono piu' grandi della frequenza marginale. Al contributo, che e' sempre positivo, bisogna applicare il segno della coordinata per avere il senso dello stesso.

Un altro indice da prendere in esame e' la bonta' di approssimazione del punto sul grafico. Talvolta, usando delle rappresentazioni bidimensionali, certi punti sono mal rappresentati, perche' in realta' sono distanti dal piano su cui vengono proiettati (La misura della bonta' di approssimazione e' il coseno al quadrato tra il punto (riga o colonna) e il piano definito dagli assi principali). L'indice di qualita' di rappresentazione, che e' analogo alla somma dei coefficenti di correlazione al quadrato nell' analisi in componenti principali, e' compreso tra 0 (pessima qualita') a 1 (rappresentazione esatta).

Vediamo quali sono i contributi dei punti riga e colonna ai primi due assi. I risultati sono spesso piu' leggibili nella forma seguente, cioe' moltiplicati per 1000 e arrotondati.

Righe	f_{i+}	$\text{CTR}_1(i)$	$\text{CTR}_2(i)$	QLT
Arruolamento	129	0	8	295
Sabotaggio	146	0	12	738
Situazione militare	237	64	259	938
Situazione politica	123	184	55	995
Situazione economica	12	5	0	535
Altro	185	1	589	1000
Piu' di una	45	12	68	602
Ristrettezze	123	734	10	999
Colonne	f_{+j}	$\text{CTR}_1(j)$	$\text{CTR}_2(j)$	QLT
ASIAF	302	540	53	996
EUSA	506	383	108	1000
IS-AA	42	62	19	966
IS-EU	115	8	795	988
IS-IS	35	7	25	277

Si nota il forte contributo dei punti associati a 'ristrettezze' e a 'situazione politica' al primo asse e della riga 'altro' al secondo asse. La qualita' (bidimensionale) della rappresentazione non e' molto buona per le modalita' 'arruolamento' tra le righe e per la colonna'Israele, padre in Israele'. L'instabilita' di questa colonna e' dovuta al fatto di avere le frequenze molto basse.

3.9 Un esempio finale

Diamo ora un'illustrazione piu' consistente dell'analisi delle corrispondenze. L'analisi si limitera' alla rappresentazione grafica con qualche commento.

L'esempio 1.2 fornisce una tipica tavola di contingenza analizzabile con l'analisi delle corrispondenze. Si tratta di una tavola di contingenza molto ampia relativa a un problema abbastanza generico su cui si hanno poche informazioni *a priori*. L'esame diretto della tabella e' difficoltoso, data la mole dei dati, ed e' del tutto evidente che non si e' interessati a studiare se vi sia indipendenza tra i caratteri perche' questa e' senz'altro da rifiutare. I dati formano in realta' una tabella tripla $2 \times 13 \times 9$ che qui analizzeremo come una tabella doppia $(2 \times 9) \times 13$ considerando come righe tutte le combinazioni di eta' e sesso.

L'analisi delle corrispondenze ha come primo valori singolari: 0.59, 0.35, e 0.27 con percentuali spiegate del ϕ^2 pari rispettivamente al 58, 20 e 12%. Pertanto le prime due dimensioni spiegano insieme il 78% dello scostamento dall'indipendenza.

La rappresentazione grafica e' presentata nel grafico 3.7. Le classi d'eta' sono state unite da linee e questo evidenzia un comportamento parallelo dei maschi e delle femmine in situazioni diverse. Le linee hanno un brusco cambiamento all'eta' di 29 anni. I profili alle eta' maggiori sono diversi da quelli dei piu' giovani, ma sono piu' omogenei. Questi cambiano maggiormente per le classi di eta' piu' basse. Il secondo asse oppone il comportamento dei maschi da quello delle femmine che appare ben differenziato. I giovanissimi rubano giocattoli, dolci e materiale per scrivere piu' della media. Le ragazze rubano gioielli piu' della media, le piu' grandi profumi piu' della media, le donne adulte vestiti.

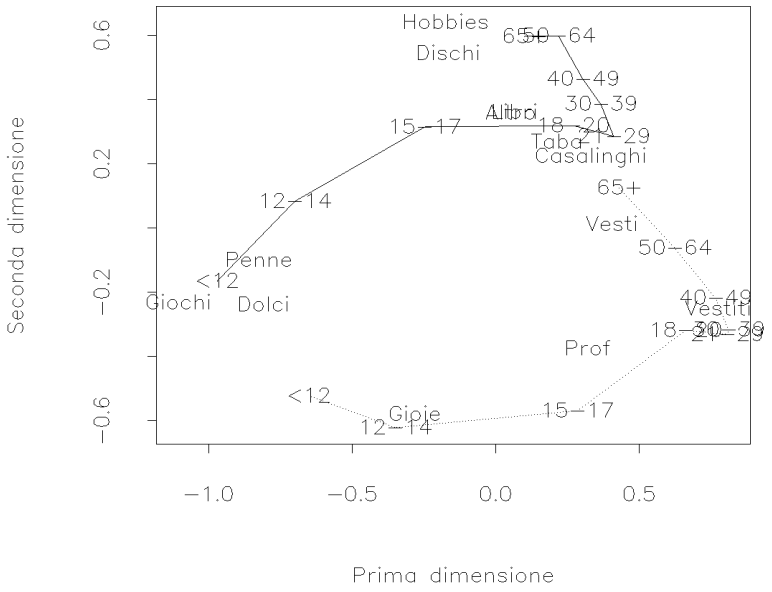

Figura 3.7: *Analisi delle corrispondenze sui dati dei furti. Maschi a tratto unito, femmine a tratteggio.*

Si osservi che l'analisi precedente usa le formule di transizione per interpretare la relazione tra punti riga e punti colonna e non le distanze tra di essi. Tutte le considerazioni che legano fra loro righe e colonne debbono essere valutate attentamente eventualmente ricorrendo ai profili riga e colonna originali, perche' a volte queste 'corrispondenze' possono essere fallaci anche a causa della rappresentazione approssimata.

3.10 Note bibliografiche

L'analisi in componenti principali e' un metodo tipicamente descrittivo. Quasi tutti i manuali di analisi multivariata comprendono un capitolo destinato a questa tecnica. Essa puo' essere derivata in molti modi diversi, collegati fra loro. Anche noi abbiamo parlato di proiezioni di unita' su piani fattoriali, di combinazioni lineari di variabili e, infine, di approssimazioni di matrici.

Spesso nei manuali si fa una certa confusione tra l'analisi in componenti principali e l'analisi dei fattori, che e' invece un modello probabilistico.

L'analisi delle corrispondenze e' lo strumento principale di molti statistici francesi fra cui Benzecri il quale ha contribuito al suo grande sviluppo in questo paese. Il metodo e' stato piu' volte scoperto e riproposto anche dagli anglosassoni.

Quello che qui e' stato detto in modo estremamente sintetico (e approssimato) si puo' ritrovare in modo piu' dettagliato in molti testi dedicati esclusivamente all'analisi delle corrispondenze. Oltre a Lebart, Morineau e Warwick (1984) e' consigliabile Greenacre (1984).

Goodman ha portato importanti contributi all'analisi delle tavole di contingenza con modelli ispirati all'analisi delle corrispondenze. Goodman (1991) presenta una rassegna di que-

sti sviluppi che oggi consentono di adattare e sottoporre a test questi modelli nell'ambito dell'inferenza classica.

Lauro e D'Ambra (1984) hanno proposto una versione non simmetrica dell'analisi delle corrispondenze.

L'esempio dei furti nel grande magazzino, con l'analisi relativa e' stato ripreso da van der Heijden, Falguerolles e de Leeuw (1989) i quali si sono occupati, fra gli altri, dell'uso combinato dell'analisi delle corrispondenze e dei modelli log-lineari.

Bibliografia

Arbia G. (1989). *Spatial data configuration in statistical analysis of regional economic and related problems*. Dordrecht: Kluwer Academic Publishers.

Barnett V. (ed.) (1981). *Interpreting multivariate data*. Chichester: John Wiley.

Chambers J. M., Cleveland W. S., Kleiner B, Tukey P. A. (1983). *Graphical methods for data analysis*. Monterey, California: Wadsworth.

Chiandotto B. (1978). *L'analisi dei gruppi: una metodologia per lo studio del comportamento elettorale*, parte prima. *Quaderni dell'Osservatorio Elettorale*, **4**.

Chiandotto B., Marchetti G. (1980). *L'analisi dei gruppi: una metodologia per lo studio del comportamento elettorale*, parte seconda. *Quaderni dell'Osservatorio Elettorale*, **7**.

Cleveland W. S., McGill M. E. (1988) *Dynamic graphics for statistics*. Belmont, California: Wadsworth.

Fabbris L. (1997). *Statistica multivariata*. Milano: McGraw-Hill Libri Italia.

Goodman L. A. (1991). Measures, models and graphical displays in cross-classified data. *J. of the American Statistical Society*. **86**, 1085–1138.

Gordon A. D. (1981). *Classification*. London: Chapman & Hall.

Greenacre M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Hartigan J. A. (1975). *Clustering algorithms*. New York: John Wiley.

Ku H. H., Kullback S. (1974). Loglinear models in contingency table analysis. *The American Statistician*, **28** 115–122.

Lauro N., d'Ambra L. (1984). L'analyse non symetrique des correspondences. In *Data analysis and informatics* (vol. 3), a cura di: Diday E., Jambu M., Lebart L., Pages J., Tomassone R. Amsterdam: Elsevier Science Publishers (North-Holland). 433–446.

Lebart L, Morineau A., Warwick K. M. (1984), *Multivariate descriptive statistics*. New York: John Wiley.

Mardia K. V., Kent J. T., Bibby J. M. (1979). *Multivariate analysis*. London: Academic Press.

McLachlan G. J., Basford K. E. (1988). *Mixture models: inference and applications to clustering*. New York: Marcel Dekker.

Seber G. A. F. (1984). *Multivariate observations*. New York: John Wiley.

Statistical Abstract of the United States, 1977 and *County and City Data Book*, 1977, U.S. Department of Commerce, Bureau of the Census.

van der Heijden P. G. M., de Falguerolles A., de Leuuw J. (1989). A combined approach to contingency table analysis using correspondence analysis and log-linear analysis (with discussion). *Applied Statistics* **38**, 249–292.

Zanella A. (1988). *Lezioni di statistica*, parte seconda. Milano: Vita e Pensiero.