
CAPITOLO 2

Dal semplice al multiplo Analisi della corrispondenza

Michael Greenacre

CONTENUTI

2.1 Introduzione.....	41	2.2 Analisi
delle correlazioni canoniche....	43	2.2.1 Due
variabili.....	43	2.2.2 Alcune
variabili	49	2.2.3 Analisi di
omogeneità	56	2.3 Approccio
geometrico.....	58	2.3.1 Scaling
della distanza chi-quadrato.....	59	2.3.2
Biplot	61	2.3.3 Analisi
congiunta delle corrispondenze....	65	2.3.4
Regolazione delle inerzie in MCA	67	2.4
Punti aggiuntivi.....	70	2.5 Discussione
e conclusioni	75	

2.1 Introduzione

L'analisi delle corrispondenze semplice (CA) è applicabile principalmente a una tabella di contingenza bidirezionale, che porta a una mappa che visualizza l'associazione tra due variabili categoriali. **L'analisi delle corrispondenze multiple (MCA) affronta il problema più generale delle associazioni tra un insieme di più di due variabili categoriali. Vedremo che la generalizzazione a più di due variabili non è né ovvia né ben definita.** In altre aree dell'analisi multivariata, come la regressione e la log-lineare

modellando, la situazione è meno complicata: ad esempio, il passaggio dalla regressione di una variabile di risposta su un singolo predittore al caso di più predittori è abbastanza semplice. Il problema principale che affrontiamo qui è che la nozione di associazione tra due variabili categoriali è un concetto complesso. Esistono diversi modi per generalizzare questo concetto a più di due variabili.

Tra i molti modi diversi che esistono per definire l'MCA, prenderemo in considerazione due approcci: il primo, la definizione forse più facile da comprendere, vale a dire quella di correlazione tra insiemi di variabili, nota come correlazione canonica, e il secondo, l'approccio geometrico, che è direttamente collegato alla visualizzazione dei dati e che presenta molte somiglianze con l'analisi delle componenti principali in stile Pearson. Nella spiegazione di ciascun approccio, considereremo il caso di due variabili e poi descriveremo le possibili generalizzazioni a più di due variabili.

Come illustrazione della teoria, utilizzeremo un set di dati dell'International Social Survey Program sull'ambiente (ISSP 1993), esaminando specificamente le domande sugli atteggiamenti nei confronti della scienza. Le domande del sondaggio che consideriamo sono le seguenti:

Quanto sei d'accordo o in disaccordo con ciascuno di questi stati
menti?

- R. Crediamo troppo spesso nella scienza e troppo poco nei sentimenti e fede.
- B. Nel complesso, la scienza moderna fa più male che bene.
- C. Qualsiasi cambiamento causato dall'uomo nella natura, non importa quanto sia scientifico efficace – rischia di peggiorare le cose.
- D. La scienza moderna risolverà i nostri problemi ambientali con pochi cambiamenti al nostro modo di vivere.

Ogni domanda ha cinque possibili categorie di risposta:

- 1. Fortemente d'accordo
- 2. D'accordo
- 3. Né d'accordo né in disaccordo
- 4. In disaccordo
- 5. Fortemente in disaccordo

Per evitare il problema delle differenze interculturali, utilizziamo i dati solo per il campione della Germania occidentale (le indagini ISSP distinguono ancora tra l'ex Germania occidentale e quella orientale). Mostriamo anche come mettere in relazione il

I risultati dell'MCA sono relativi alle variabili demografiche esterne di sesso, età e istruzione, anch'esse codificate come variabili categoriali come segue:

Sesso: maschio, femmina

Età (sei gruppi): 16–24, 25–34, 35–44, 45–54, 55–64, 65 e

più vecchio

Istruzione (sei gruppi): primaria incompleta, primaria completata, secondaria incompleta, secondaria completata, terziaria incompleta, terziaria completata

È stata eseguita una cancellazione per elenco degli intervistati con dati mancanti perché non vogliamo affrontare qui la questione ulteriormente complicata dei dati mancanti (vedi Capitolo 8). Ciò riduce il campione originale della Germania occidentale di circa il 14% per lasciare $n = 871$ intervistati con dati completi, che costituiscono il set di dati utilizzato in questo capitolo.

Nell'appendice di questo libro, la maggior parte dei risultati numerici di questo capitolo sono forniti insieme al codice nel linguaggio R per eseguire i calcoli (R Development Core Team 2005).

2.2 Analisi delle correlazioni canoniche

2.2.1 Due variabili

Iniziamo considerando solo le prime due variabili, A (riguardante la fede nella scienza) e B (riguardante il danno causato dalla scienza), entrambe formulate in modo sfavorevole nei confronti della scienza, in modo che il disaccordo indichi un atteggiamento favorevole nei confronti della scienza. Poiché ci sono solo due variabili, tutte le 871 risposte a queste domande possono essere codificate, senza perdita di informazioni, sotto forma di tabulazioni incrociate, fornite nella Tabella 2.1. **L'approccio correlazionale indaga come misurare l'associazione tra queste due variabili categoriali. Esistono già diverse misure di associazione per i dati categoriali, alcune delle quali dipendono dal fatto che le variabili siano misurate su una scala nominale o ordinale o se stiamo cercando di prevedere una variabile dall'altra. Nel seguito concentreremo il nostro interesse sul classico coefficiente di correlazione prodotto-momento applicabile ai dati metrici e sulla quantificazione delle categorie, ovvero su come ottenere valori numerici per le categorie di risposta per calcolare un coefficiente di correlazione tra le variabili. Poiché le categorie sono ordinate, una soluzione semplice sarebbe quella di utilizzare i valori esistenti da 1 a 5, così come sono codificati nel file di dati, presupponendo quindi che vi sia una differenza di intervallo uguale tra punti adiacenti su ciascuna scala.**

Tabella 2.1 Tabulazioni incrociate di 871 intervistati della Germania occidentale rispetto a due domande sull'atteggiamento nei confronti della scienza.

Crediamo troppo spesso nella scienza, non abbastanza nei sentimenti e nella fede	Nel complesso, la scienza moderna fa più male che bene					SOMMA
	B1	B1	B3	B4	B5	
	Essere d'accordo		nessuno dei due/		in forte	
	fortemente	Essere d'accordo	né	disaccordo	disaccordo	
A1-assolutamente d'accordo	27	28	30	22	12	119
A2-d'accordo	38	74	84	96	30	322
A3-né/né	3	48	63	73	17	204
A4-disagree	3	21	23	79	52	178
A5-fortemente in disaccordo	0	3	5	11	29	48
SOMMA	71	174	205	281	140	871

Ma si noti che tale scelta **sarebbe errata se una delle variabili erano nominali, ad esempio "provincia di residenza" o "religiosa". denominazione.**

Esistono due modi per calcolare il coefficiente di correlazione: uno è dai dati originali a livello di intervistato, che sono le 871 coppie di risposte alle due domande; l'altro approccio, più compatto, lo è direttamente dalla Tabella 2.1, poiché questa tabella fornisce le frequenze di occorrenza di tutte le coppie di categorie. Supponiamo che le risposte alle domande A e B sono codificati rispettivamente nelle matrici degli indicatori Z1 e Z2, le cui colonne sono zero-uno variabili dummy: cioè Z1 e Z2 lo sono entrambe le matrici 871 × 5. Allora la Tabella 2.1 è il prodotto incrociato Z1^TZ2 di le due matrici di indicatori. Inoltre, supponiamo che la proposta sono contenuti i valori di scala per le categorie delle due variabili i vettori s1 e s2, in modo che le risposte quantificate individuali siano nei vettori Z1s1 e Z2s2. Per semplificare notevolmente la notazione, lo è conveniente considerare le risposte quantificate come inizialmente medie centrate, 1^TZ1s1 = 1^TZ2s2 = 0, in modo che la covarianza s12 tra le due variabili e le loro varianze e possono essere scritte² come:

$$s_{12} = (1/n)s_1^T \overset{1}{Z_1^T} \overset{2}{Z_2^T} s_2 = s_1^T P_{12}s_2$$

$$\overset{1}{s_1^T} = (1/n)s_1^T \overset{1}{Z_1^T} \overset{1}{Z_1} s_1 = s_1^T \overset{1}{D_1} s_1 \text{ e } \overset{2}{s_2^T} = (1/n)s_2^T \overset{2}{Z_2^T} \overset{2}{Z_2} s_2 = s_2^T \overset{2}{D_2} s_2$$

dove $P_{12} = (1/n)Z_1^T Z_2$ è detta *matrice di corrispondenza*, contenente le frequenze relative, ovvero la Tabella 2.1 divisa per il totale complessivo di

$n = 871$. D_1 e D_2 sono matrici diagonali delle frequenze relative marginali, o masse, delle due variabili. (Nel Capitolo 1 questi sono indicati con D_r e D_c per "righe" e "colonne"; in questo capitolo utilizzeremo gli indici 1 e 2, poiché estenderemo i concetti a più di due variabili.)

Utilizzando la notazione precedente la correlazione può essere scritta come:

$$R = \frac{s_{12}}{\sqrt{s_1 s_2}} = \frac{s_{12}}{\sqrt{s_1 s_2}} \quad (2.1)$$

che può essere calcolato direttamente dalla Tabella 2.1 e dai suoi margini. Poiché questo calcolo coinvolge alcuni concetti importanti in CA, lo esamineremo in dettaglio, utilizzando i valori da 1 a 5 per le categorie di ciascuna variabile.

- Dalla Tabella 2.1 abbiamo i valori delle frequenze relative marginali (masse) per le categorie delle due variabili:

$$\begin{aligned} (1/n)1^T Z_1 &= (1/871)[119 \ 322 \ 204 \ 178 \ 48] \\ &= [0,137 \ 0,370 \ 0,234 \ 0,204 \ 0,055] \end{aligned}$$

$$\begin{aligned} (1/n)1^T Z_2 &= (1/871)[71 \ 174 \ 205 \ 281 \ 140] \\ &= [0,082 \ 0,200 \ 0,235 \ 0,323 \ 0,161] \end{aligned}$$

- Assumendo le scale di intervallo uguali 1, 2, 3, 4, 5 per le due variabili, le loro medie sono

$$(0,137 \times 1) + (0,370 \times 2) + \dots + (0,055 \times 5) = 2,672$$

$$(0,082 \times 1) + (0,200 \times 2) + \dots + (0,161 \times 5) = 3,281$$

e i vettori centrati s_1 e s_2 lo sono

$$\begin{aligned} s_1 &= \begin{matrix} 1.672 \\ 0.672 \\ 0.328 \\ 1.328 \\ 2.328 \end{matrix} \quad s_2 = \begin{matrix} 2.281 \\ 1.281 \\ 0.719 \\ 1.719 \end{matrix} \end{aligned}$$

- La **matrice di corrispondenza** è la matrice delle frequenze relative (diamo solo alcuni elementi della matrice):

$$P_{12} = \begin{pmatrix} \bar{y}_{12} & \bar{y}_{13} & \bar{y}_{14} \\ \bar{y}_{22} & \bar{y}_{23} & \bar{y}_{24} \\ \bar{y}_{32} & \bar{y}_{33} & \bar{y}_{34} \end{pmatrix} = \begin{pmatrix} 0.03100 & 0.01378 & 0.03330 \\ 0.04363 & 0.03444 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 \end{pmatrix}$$

e le matrici diagonali delle masse, D_1 e D_2 , contengono le frequenze relative marginali (masse) calcolate sopra.

- Quindi, la covarianza, le varianze e la correlazione lo sono

$$s_{12} = s_1^T P_{12} s_2 = (0.03100 \times \bar{y}_{12} + 0.01378 \times \bar{y}_{13} + 0.03330 \times \bar{y}_{14}) = 0.4988$$

$$s_{11}^2 = s_1^T D_1 s_1 = 0.137 \times (\bar{y}_{12})^2 + \dots + 0.055 \times (\bar{y}_{14})^2 = 1.233$$

$$s_{22}^2 = s_2^T D_2 s_2 = 0.082 \times (\bar{y}_{22})^2 + \dots + 0.161 \times (\bar{y}_{24})^2 = 1.412$$

$$r = \frac{s_{12}}{\sqrt{s_{11}^2 s_{22}^2}} = \frac{0.4988}{\sqrt{1.233 \times 1.412}} = 0.3780$$

Tutti i calcoli di cui sopra dipendono chiaramente dai valori della scala di intervallo uguale in s_1 e s_2 assunti all'inizio. Consideriamo ora questi valori di scala come incognite da determinare e poniamo la seguente domanda: quali valori di scala per s_1 e s_2 daranno la massima correlazione (Equazione 2.1) tra le due variabili? Questo è esattamente il problema della **correlazione canonica** tra le cinque variabili dummy in Z_1 e le cinque variabili dummy in Z_2 . Poiché la correlazione rimane la stessa anche se vengono effettuate trasformazioni lineari di s_1 e s_2 , è necessario introdurre **condizioni di identificazione** che fissino la scala di s_1 e s_2 per trovare la soluzione ottima. Le consuete condizioni di identificazione sono che le due variabili siano standardizzate, cioè che le medie siano zero, come in precedenza: $(1/n)1^T Z_1 s_1 = (1/n)1^T Z_2 s_2 = 0$ e,

inoltre che le varianze sono 1: $s_1^T D_1 s_1 = s_2^T D_2 s_2 = 1$. Sotto

queste condizioni, mostriamo ora che la soluzione ottima coincide

esattamente con le cosiddette coordinate standard delle categorie di risposta sulla prima dimensione principale di una semplice CA dell'originale tabulazioni incrociate (vedi Capitolo 1).

Considera la decomposizione in valori singolari (SVD) di quanto segue matrice normalizzata:

$$DPP_{UV}^{1/2} / \bar{y}_{12} = S \quad TTT \quad \text{Dove} \quad UU \quad VV \quad I = \quad (2.2)$$

dove S è la matrice diagonale dei valori singolari e U e V lo sono

le matrici dei vettori singolari sinistro e destro come colonne. Quindi, scrivendo l'equazione 2.2 per una coppia di vettori sinistro e destro, u e v, corrispondenti a un valore singolare \bar{y} , abbiamo, dopo aver moltiplicato per quello sinistro

da uT e a destra da v, e utilizzando l'ortogonalità del singolare vettori:

$$uDPD_{\bar{y}}^{1/2} / \bar{y}_{12} = P$$

Quindi se lasciamo s_1 e s_2 che è il $2 = \frac{1}{2}$, poi $sPs \quad T \quad 1 \quad 12 \quad 2 = \text{pag}$,

formula per la covarianza. Inoltre, le condizioni di identificazione

$T \quad P \quad D \quad s \quad D \quad s$ sono soddisfatti, poiché i vettori singolari hanno lunghezza

1: $uTu = vTv = 1$, quindi sembra che la correlazione sia data da the

valore singolare \bar{y} . Tuttavia, le condizioni di centratura non sono state

imposti e questi possono essere introdotti centrando prima la matrice su

essere scomposto come segue, sottraendo il prodotto della riga e

margini di colonna da ciascun elemento della matrice di corrispondenza:

$$DPP_{11}^{1/2} (PDU_{12} \quad 2_{12} \quad TTT \quad 12) / \bar{y}_{12} = \bar{y}V \quad T \quad (2.3)$$

dove P_{121} è il vettore (colonna) dei margini di riga di P_{12} , ovvero le masse di riga

(indicato con r nel Capitolo 1) e $1T \quad PT \quad 12$ è il vettore (riga) dei margini delle colonne,

le masse della colonna (indicate con cT). Nel gergo della CA, questo è noto come

"rimuovendo la soluzione banale" perché la matrice non centrata (Equazione 2.2)

ha una soluzione massimale banale con valore singolare pari a 1 per s_1 e s_2 uguali

a 1 (quindi U, V e S nell'equazione 2.2 hanno tutti questo singolare extra banale

componente, che viene eliminato dalla centratura nell'Equazione 2.3).

Abbiamo quindi il seguente risultato: ogni valore singolare è una correlazione

tra le variabili A e B, in base ai valori di scala dei vettori singolari trasformati s_1 e s_2 , e quindi la correlazione massima è

ottenuto per il primo (cioè il più grande) valore singolare dell'equazione 2.3 o,

equivalentemente, il secondo valore singolare più grande della matrice non centrata

(Equazione 2.2). Le soluzioni s_1 e s_2 sono esattamente i vettori dello standard coordinate in CA sul primo asse principale. Il valore singolare più grande γ_1 dell'equazione 2.3, chiamata anche prima correlazione canonica, è uguale a 0,4106 nel nostro esempio, rispetto al valore di 0,3780 ottenuto con le scale dell'intervallo uguale (da 1 a 5). I valori della scala sono:

$$s_1^T = [\gamma_1, 0,17 \quad \gamma_0,560 \quad \gamma_0,248 \quad 1,239 \quad 2,741]$$

$$s_2^T = [\gamma_1, 571 \quad \gamma_0,667 \quad \gamma_0,606 \quad \gamma_0,293 \quad 1,926]$$

Questi valori di scala sono standardizzati, ma poiché qualsiasi trasformazione lineare lascia invariata la correlazione, è conveniente effettuare la trasformazione in modo che anche gli endpoint abbiano i valori 1 e 5, con un intervallo di 4, al fine di effettuare un confronto con le precedenti scale equi-intervalli. Per ad esempio, per la prima variabile, l'intervallo di valori è $2.741 - \gamma_1$ (γ_1 0.17) = 3,758, quindi per rendere l'intervallo esattamente di quattro unità, dovremmo moltiplicare tutti i valori per $4/3.758 = 1.064$, in tal caso il più basso il valore è ora $\gamma_1, 017 \times 1,064 = \gamma_1, 083$. Poi la somma di 2.083 a tutto i valori porteranno la scala ad avere valori minimo e massimo uguali rispettivamente a 1 e 5. Questa procedura dà quanto segue riscalato valori per i due insiemi di categorie di risposta:

$$\text{valori delle righe riscalati} = [1 \quad 1,486 \quad 1,818 \quad 3,402 \quad 5]$$

$$\text{valori delle colonne riscalati} = [1 \quad 2.034 \quad 2.103 \quad 3.132 \quad 5]$$

I punti della scala emergono nell'ordine atteso in entrambi i casi, ma è interessante studiarne le distanze relative. Paragonato a i valori equi-intervalli considerati in precedenza, tali valori riscalati mostrano che le categorie "in disaccordo" e "fortemente in disaccordo" per la domanda A sono ulteriormente distanziate, con differenze relativamente piccole tra valori di scala assegnati alle categorie "fortemente d'accordo", "d'accordo" e "né né." Per la domanda B la differenza tra "non sono d'accordo" e "fortemente in disaccordo" è ancora più grande, quasi due unità intere. Per entrambi domande la categoria neutra "né/né" non è al centro delle scala ma vicino alla categoria dell'accordo.

Prima di passare al caso di più variabili, osserviamo che, in quanto sopra, è stato derivato solo un insieme di valori di scala per ciascuno variabile, corrispondente al primo valore singolare γ_1 . Ulteriori serie di i valori della scala possono essere determinati in modo graduale massimizzando la correlazione tra un'altra coppia di punteggi del soggetto basati su diversi valori di scala, indicando dove i punteggi dei soggetti non sono correlati con quelli già ottenuti, cioè $s_D^T s_D = 0$ $\begin{matrix} T_1 & 1 & 1 \\ & & 2 \end{matrix} = \begin{matrix} T \\ 2 & 2 & 2 \end{matrix}$. La soluzione è dato dal secondo insieme di vettori singolari dell'equazione 2.3, trasformato

come prima alle coordinate standard, corrispondenti alla seconda singolare valore, \tilde{y}_2 , che è la seconda correlazione canonica. Per un tavolo d'ordine $I \times J$, questo processo può essere continuato per ottenere un totale di $\min\{I - 1, J - 1\}$ correlazioni canoniche e valori di scala associati: nel nostro esempio 5×5 è possibile calcolare quattro serie di valori di scala e correlazioni canoniche. Le correlazioni canoniche sono le radici quadrate del *principale* *inerzie* solitamente riportate sugli assi della mappa (vedi Capitolo 1 e Sezione 2.3 di seguito).

2.2.2 Diverse variabili

Per effettuare la transizione al caso di più variabili, notare che the il problema è quasi identico se lo riformuliamo massimizzando il correlazione tra le due variabili e la loro media (o la loro somma). In generale, per due variabili z_1 e z_2 con correlazione \tilde{y} , la correlazione tra uno di essi e la loro media $1/2(z_1 + z_2)$ (o la loro somma $z_1 + z_2$) è uguale a $(1/2\sqrt{1 + \tilde{y}})/\sqrt{1 + \tilde{y}}$, quindi massimizzare \tilde{y} è equivalente a massimizzando la correlazione tra le variabili e la loro media (o somma). L'unica vera differenza è il valore del massimo trovato: in quest'ultima formulazione sarà $(1/2)\sqrt{1 + \tilde{y}}$ e non il valore di \tilde{y} si. La media di due variabili categoriali ci porta a considerare il matrice delle due matrici di indicatori $[Z_1 \ Z_2]$, dove la media di le due quantificazioni delle variabili, basate rispettivamente su s_1 e s_2 , è uguale a

$$\frac{1}{2}(\mathbf{Z}_s \mathbf{Z}_s' + \mathbf{Z}_2 \mathbf{Z}_2') = \frac{1}{2}[\mathbf{Z}_1 \ \mathbf{Z}_2] \begin{bmatrix} \tilde{y}_s & \tilde{y}_{s_2} \\ \tilde{y}_{s_2} & \tilde{y}_{\tilde{y}} \end{bmatrix}$$

Consideriamo ora cosa succede quando applichiamo l'algoritmo CA standard alla matrice dei superindicatori $\mathbf{Z} = [\mathbf{Z}_1 \ \mathbf{Z}_2]$. Perché \mathbf{Z} ha totale somma $2n$, con ciascuna delle n righe che si sommano a una costante 2 e colonna somme uguali alle frequenze marginali di ciascuna variabile, la matrice di corrispondenza è $[1/(2n)]\mathbf{Z}$, la matrice della massa di riga è $(1/n)\mathbf{I}$ e la matrice la matrice della massa della colonna è $\mathbf{D} = 1/2\text{diag}(\mathbf{D}_1, \mathbf{D}_2)$, dove **diag**($\mathbf{D}_1, \mathbf{D}_2$) è il matrice diagonale formata dalle due matrici diagonali \mathbf{D}_1 e \mathbf{D}_2 definito. Quindi, la SVD per calcolare la soluzione CA di \mathbf{Z} è (nella sua forma non centrata, vedere l'equazione 2.2):

$$\sqrt{\frac{1}{2N}} \mathbf{Z} \mathbf{D} \mathbf{U} \mathbf{V}' = \mathbf{U} \mathbf{U}' \mathbf{D} \mathbf{V} \mathbf{V}' \quad \text{Dove} \quad \mathbf{U} \mathbf{U}' = \mathbf{I}$$

che, in una delle sue formulazioni simmetriche agli autovalori, può essere scritta come:

$$\frac{1}{4N} \mathbf{D} \mathbf{Z} \mathbf{Z}^T \mathbf{D} = \mathbf{G}^2 \quad (2.4)$$

Dove $\mathbf{V} \mathbf{V}^T = \mathbf{I}$

questo è,

$$\frac{1}{4N} \mathbf{D} \mathbf{C} \mathbf{D} = \mathbf{G}^2 \quad (2.4)$$

dove $\mathbf{C} = \mathbf{Z}^T \mathbf{Z}$ e $\mathbf{L} = \mathbf{G}^2$. La matrice \mathbf{C} , detta *matrice di Burt*, lo è una struttura dati importante in MCA: è la matrice di tutti i dati a due vie tabulazioni incrociate delle variabili categoriali, che nel caso di specie di due variabili categoriali può essere scritta come:

$$\mathbf{C} = \begin{pmatrix} \mathbf{z}_{11} & \mathbf{z}_{12} \\ \mathbf{z}_{21} & \mathbf{z}_{22} \end{pmatrix} \quad \mathbf{D} \mathbf{P}_1 \mathbf{D} = \mathbf{G}^2$$

Usiamo la notazione $\mathbf{L} = \mathbf{G}^2$, cioè i quadrati λ^2 del singolare i valori, o inerzie principali di \mathbf{Z} che compaiono sulla diagonale di \mathbf{G}^2 , sono indicato con λ^2 sulla diagonale di \mathbf{L} . Scrivere l'equazione 2.4 per un singolo autovettore \mathbf{v} , suddiviso in due sottovettori \mathbf{v}_1 e \mathbf{v}_2 (uno corrispondente alle righe della tabella originale, l'altro alle colonne) e moltiplicando come prima a sinistra per \mathbf{v}^T e a destra per \mathbf{v} , definendo $\mathbf{s} = \mathbf{D} \mathbf{v}_1 / 2 \mathbf{v}_2$ analogamente suddiviso in \mathbf{s}_1 e \mathbf{s}_2 , otteniamo l'autoequazione:

$$\frac{1}{4} \mathbf{s}_1^T \mathbf{D} \mathbf{P}_1 \mathbf{s}_1 + \frac{1}{2} \mathbf{s}_1^T \mathbf{D} \mathbf{P}_2 \mathbf{s}_2 + \frac{1}{2} \mathbf{s}_2^T \mathbf{D} \mathbf{P}_1 \mathbf{s}_1 + \frac{1}{4} \mathbf{s}_2^T \mathbf{D} \mathbf{P}_2 \mathbf{s}_2 = \lambda^2 \quad (2.5)$$

questo è,

$$\frac{1}{4} (\mathbf{s}_1^T \mathbf{D} \mathbf{P}_1 \mathbf{s}_1 + \mathbf{s}_1^T \mathbf{D} \mathbf{P}_2 \mathbf{s}_2 + \mathbf{s}_2^T \mathbf{D} \mathbf{P}_1 \mathbf{s}_1 + \mathbf{s}_2^T \mathbf{D} \mathbf{P}_2 \mathbf{s}_2) = \lambda^2 \quad (2.5)$$

Il valore massimo dell'equazione 2.5, dato dal più grande non banale autovalore $\lambda^2 = \lambda_1^2$, coincide con la soluzione della semplice CA del singolo tabella a due vie con matrice di corrispondenza \mathbf{P}_{12} , tranne che il suo massimo è ora uguale a $1/4(1 + \lambda_1^2 + \lambda_1^2 + 1) = 1/2(1 + \lambda_1^2)$, dove λ_1^2 è il massimo correlazione canonica in CA semplice. Secondo le nostre osservazioni precedenti, $1/2(1 + \lambda_1^2)$ è esattamente il quadrato della correlazione tra uno dei due

due variabili categoriali (quantificate) e la loro media (o somma). Quindi s_1 e s_2 derivati sopra sono identici a s_1 e s_2 della semplice CA, e ciò che abbiamo derivato sono le coordinate standard delle colonne della matrice dell'indicatore Z . Notare che l'autovalore \bar{y}_1 sopra è anche il valore singolare di C perché C è simmetrico: nel linguaggio della CA geometrica (vedi Capitolo 1 e Sezione 2.3 sotto), \bar{y}_1 è la radice quadrata dell'inerzia principale della matrice di **Burt** C .

Il lettore attento avrà notato che nell'Equazione 2.4 la condizione di identificazione di s implicita nella standardizzazione di v nella SVD è che la sua somma ponderata dei quadrati sia pari a 1, cioè $1/2(s_1^T D_2 s_2) = s^T D s = 1$, e non che i due vettori s_1 e s_2 siano individualmente normalizzati a 1. Si può dimostrare, tuttavia, che se s_1 e s_2 costituiscono una soluzione corrispondente ad un autovalore $1/2(1 + \bar{y})$, allora s_1 e s_2 costituiscono un'altra soluzione corrispondente all'autovalore $1/2(1 - \bar{y})$ (vedi, ad esempio, Greenacre 1984: sezione 5.1). L'ortogonalità di questi eigen $D_1 s_1 - s_2$

vettori, $s_1^T D_2 s_2 = 0$, insieme alla normalizzazione complessiva vincolo, implicano le normalizzazioni individuali $s_1^T D_1 s_1 = s_2^T D_2 s_2 = 1$. Per quanto riguarda i singoli vincoli di centratura, questi sono automatici, poiché ogni insieme di variabili dummy (colonne di Z_1 e Z_2) ha la stessa somma, pari a 1, il vettore delle unità, per cui ogni insieme ha lo stesso baricentro, uguale al baricentro complessivo $(1/n)1$.

La scena è ora pronta per una possibile generalizzazione dell'AC al caso multivariabile, dove sono presenti Q variabili categoriali, codificate nelle matrici indicatori Z_1, Z_2, \dots, Z_Q . Il problema può essere definito come trovare un insieme di valori di scala s_1, s_2, \dots, s_Q per le variabili in modo che una misura complessiva di correlazione sia massimizzata. Per generalizzare il caso a due variabili, la misura scelta è la somma delle correlazioni al quadrato dei punteggi individuali $Z_1 s_1, Z_2 s_2, \dots, Z_Q s_Q$ con il punteggio sommato Zs , dove Z e s sono le concatenazioni di Z_q e s_q , rispettivamente. Specificiamo un vincolo di identificazione complessivo $s^T D s = 1$, dove $D = (1/Q)\text{diag}(D_1, D_2, \dots, D_Q)$. Questo vincolo complessivo non implica che le varianze individuali $s_q^T D_q s_q$ saranno 1 nella soluzione finale, a differenza del caso $Q = 2$ descritto nel paragrafo precedente.

Anche in questo caso ci sono due modi per ottenere la soluzione, in un modo eseguendo un CA della matrice del superindicatore $Z = [Z_1 Z_2, \dots, Z_Q]$, in alternativa un CA della matrice di Burt C , che ora è una matrice a blocchi con Q righe di blocchi -wise e colonna. Indichiamo il numero di categorie per la q -esima variabile categoriale con J_q e lasciamo che $J = J_q$ sia il numero totale di categorie. Allora Z è di ordine $n \times J$ e C è di ordine $J \times J$. Poiché Z ha somma totale nQ , con somme di riga pari a una costante Q e somme di colonna pari alle frequenze marginali di ciascuna variabile, la matrice di corrispondenza è $(1/Qn)Z$, la matrice della massa riga

è $(1/n)I$ e la matrice della massa della colonna è D . Quindi, la SVD da calcolare la soluzione CA di Z è (nella sua forma non centrata, vedere l'Equazione 2.2):

$$\sqrt{N} \frac{Z^{12}}{Q_n} D U V^T = T T^T \tilde{Y} \quad U U^T V V^T = I \quad (2.6)$$

Per eliminare la soluzione banale, la matrice da scomporre è (vedi Equazione 2.3):

$$\sqrt{N} \frac{\tilde{Y}^{CON}}{Q_n n} \tilde{Y}^{1\tilde{Y}} G G^T \tilde{Y} \quad \tilde{Y}^{1\tilde{Y} 2/}$$

dove $(1/n)1$ è il vettore delle masse delle righe e $1^T D$ è il vettore delle colonne masse della matrice dell'indicatore (indicate con c^T nella semplice CA). L'SVD per la CA della matrice di Burt C (non centrata) è

$$D^{1/2} \frac{C}{2} D V V^T = C^2 \quad T T^T \quad V \text{ dove} \quad V V^T = I \quad (2.7)$$

e $C = Z^T Z$. Ancora una volta, la forma centrata della matrice sul lato sinistro dell'Equazione 2.7 rimuove la soluzione banale sotto forma di le frequenze relative previste:

$$D^{1/2} \frac{\tilde{Y} C}{2} D^{1/2} \tilde{Y}^{1\tilde{Y}} G G^T \tilde{Y}^{1\tilde{Y} 2/}$$

I vettori singolari di destra, che ci forniscono i valori di scala per le variabili Q , sono identiche nei due problemi. Il valore massimo della correlazione quadrata media è data dal quadrato della prima valore singolare nell'analisi (centrata) di Z , cioè il primo singolare valore nell'analisi (centrata) di C . Si noti che i valori singolari Anche \tilde{Y} nell'analisi di C sono autovalori, poiché la matrice è decomposto è definito simmetrico positivo. Le coordinate standard x che forniscono i valori di scala, partizionati in x_1, x_2, \dots, x_Q per la Q variabili, sono date dalla consueta trasformazione del singolare vettori:

$$x D v = \tilde{Y}^{1\tilde{Y} 2/}$$

dove v è il primo vettore singolare di destra, cioè la prima colonna di V . Solo le coordinate principali sono leggermente diverse nei due problemi, poiché i valori singolari differiscono.

Applichiamo ora la teoria di cui sopra alle quattro variabili descritte in Sezione 2.2.1. Nei termini della nostra notazione: $Q = 4$, $J_q = 5$ per ogni q , $J = 20$, Z

è 871×20 e C è 20×20 . Alcune righe della matrice dei dati originale e della matrice degli indicatori sono riportate nella tabella A.1 e nella tabella A.2 dell'appendice computazionale alla fine di questo libro, che dettaglia tutti i delle fasi di calcolo per ottenere la soluzione in questo esempio. La matrice Burt completa è riprodotta nella Tabella 2.2. Nella Tabella 2.3 riproduciamo le coordinate standard per la prima e la seconda soluzione ottima, insieme alle loro corrispondenti misure di correlazione. Inoltre, viene fornita la correlazione quadrata di ciascuna variabile quantificata con il punteggio totale, dimostrando che la misura di correlazione è uguale alla loro media. Il primo e ottimale insieme di valori di scala, con una correlazione quadrata media di 0,457, aumenta monotonicamente per le domande A, B e C, ma la domanda D ha uno schema abbastanza diverso, con i poli estremi che si oppongono alle categorie intermedie. Ciò dimostra che esiste un possibile problema con le risposte alla domanda D, che è stata formulata in senso inverso rispetto alle altre domande. Il secondo insieme di valori di scala cattura un asse di "polarizzazione", dove tutte e quattro le domande hanno lo schema delle categorie estreme opposte a quelle intermedie, e qui la domanda D si adatta maggiormente alle altre. Questa interpretazione è supportata dalle correlazioni quadrate, che mostrano un valore basso per la domanda D nella prima soluzione. L'MCA agisce quindi effettivamente come un'analisi degli item e questo risultato ci mostra che la domanda D ha peggiorato l'affidabilità del punteggio totale basato sulla seconda soluzione ottimale e dovrebbe preferibilmente essere rimossa.

Per chiarire il legame tra MCA e teoria dell'affidabilità, considerare le variabili Q come elementi che misurano un costrutto sottostante. Usare la correlazione quadrata media di 0,457, cioè 0,676 nella radice quadrata, come misura dell'affidabilità è una sovrastima perché anche per dati casuali troveremmo una correlazione positiva tra gli elementi e la loro somma (infatti, la correlazione quadrata media tra Q elementi non correlati e la loro somma è pari a $1/Q$). L'alfa di Cronbach è una misura di affidabilità che compensa questo ed è classicamente definito come:

$$\alpha = \frac{\sum_{j=1}^Q \bar{y}_j^2}{\sum_{j=1}^Q \bar{y}_j^2 + \frac{1}{Q} \sum_{j=1}^Q \sum_{k=1}^Q \bar{y}_j \bar{y}_k} = \frac{m \bar{q}^2}{m \bar{q}^2 + \frac{1}{Q} \sum_{j=1}^Q \sum_{k=1}^Q \bar{y}_j \bar{y}_k} \quad (2.8)$$

dove \bar{y}_j^2 è la varianza del punteggio dell'elemento j esimo e s^2 è la varianza del punteggio sommato. In MCA la somma delle varianze del punteggio dell'item ($\sum_{j=1}^Q \bar{y}_j^2$) è pari a $a^T D a$, che dalle condizioni di identificazione sopra descritte è un valore fisso, pari a Q . La varianza del punteggio sommato (s^2) è pari a Q^2 volte la varianza del punteggio medio \bar{z} , ovvero Q^2

[illegible]

Tabella 2.3 Risultati della CA della matrice di indicatori 871×20 Z (vedere Tabella A.2 in appendice) o, equivalentemente, della matrice di Burt C nella Tabella 2.2, mostrando le coordinate standard (valori di scala) per le quattro variabili sulle prime due dimensioni della soluzione (F1 e F2).

	F1	F2
A1	$\bar{y}1.837$	0,727
A2	$\bar{y}0.546$	$\bar{y}0.284$
A3	0,447	$\bar{y}1.199$
A4	1.166	0,737
A5	1.995	2.470
mq. corr.	0,510	0,382
B1	$\bar{y}2.924$	1.370
B2	$\bar{y}0.642$	$\bar{y}0.667$
B3	$\bar{y}0.346$	$\bar{y}0.964$
B4	0,714	$\bar{y}0.280$
B5	1.354	2.108
mq. corr	0,579	0,517
C1	$\bar{y}2.158$	0,909
C2	$\bar{y}0.247$	$\bar{y}0.592$
C3	0,619	$\bar{y}1.044$
C4	1,349	0,635
C5	1.468	3.017
mq. corr.	0,627	0,488
D1	$\bar{y}1.204$	1.822
D2	0,221	$\bar{y}0.007$
D3	0,385	$\bar{y}1.159$
D4	0,222	$\bar{y}0.211$
D5	$\bar{y}0.708$	1.152
mq. corr.	0,113	0,337
Rho	0,457	0,431
alpha di Cronbach	0,605	0,560

Nota: mq corr. è la correlazione al quadrato della variabile quantificata con il totale punto; rho è il corrispondente valore singolare di C, cioè il valore singolare al quadrato (o inerzia principale) di Z, che è la media aritmetica dei quattro corrispondenti correlazioni quadrate; L'alfa di Cronbach è la misura dell'affidabilità discussa in Sezione 2.2.2.

volte la \bar{y} che stiamo massimizzando. Possiamo quindi scrivere il massimo valore dell'equazione 2.8 come:

$$U_N = \frac{\bar{y}}{Q} \frac{\bar{y}\bar{y}}{1} \frac{Q}{Q^2} \frac{\bar{y}}{\bar{y}} \frac{Q}{Q} \frac{\bar{y}\bar{y}}{1} \frac{1}{Q} \frac{\bar{y}}{\bar{y}} \quad (2.9)$$

in modo che il massimo \bar{y} (il primo valore singolare di C nell'equazione 2.7, che è anch'esso un autovalore come abbiamo detto in precedenza) corrisponde alla massima affidabilità. Quindi, il valore massimo di Cronbach alfa per le prime due soluzioni è, rispettivamente (vedi Tabella 2.3):

$$A'1 = \frac{4}{3} \frac{\bar{y}_1}{\bar{y}} \frac{1}{4.0457 \times \bar{y}} \bar{y} = 0.605 \quad E \quad A'2 = -\frac{4}{3} \frac{\bar{y}_1}{\bar{y}} \frac{1}{4.0431 \times \bar{y}} \bar{y} = 0.560$$

Se si elimina la domanda D, come suggerisce la sua bassa correlazione tra gli elementi e il totale, un ricalcolo della soluzione fornisce un risultato molto più elevato. valore, 0,602, della correlazione quadrata media massima, e un aumento dell'alfa di Cronbach a 0,669. (Non riportiamo il completo risultati qui.)

La Tabella 2.4 mostra tutte le intercorrelazioni al quadrato nonché le varianze e covarianze delle quattro domande quantificate, secondo la prima soluzione ottimale. Questa tabella lo dimostra anche empiricamente la \bar{y} ottimale può essere calcolata come (a) la varianza del totale punteggio, o (b) la media delle quattro correlazioni quadrate dei rispettivi domande con il totale, oppure (c) la media di tutti gli elementi del totale matrice di varianza-covarianza tra le quattro domande.

2.2.3 Analisi di omogeneità

Una definizione alternativa ma equivalente della definizione correlazionale di MCA si basa sul criterio di "coerenza interna" di Guttman (vedi, ad esempio, Nishisato 1994). L'idea è cercare

valori della scala s_1, s_2, \dots, s_Q che danno punteggi individuali $Z_1s_1, Z_2s_2, \dots, Z_Qs_Q$ che sono il più vicini possibile tra loro (cioè omogenei). Mancanza di vicinanza può essere misurata dalla somma dei quadrati delle differenze dei valori della scala Q di ciascun individuo rispetto ai corrispondenti punteggio medio nel vettore $(1/Q)(Z_1s_1 + Z_2s_2 + \dots + Z_Qs_Q) = (1/Q)Zs$, che indicheremo ancora con z . L'obiettivo generale è quindi quello di

Tabella 2.4 Intercorrelazioni quadrate nonché varianze e covarianze delle quattro domande quantificate secondo la prima soluzione ottima.

	A a C D				Quadrato correlazione con totale
UN	1.1151	0,1396	0,1270	0,0059	0,5100
B	0,4440	1.2666	0,1868	0,0059	0,5793
C	0,4406	0,5697	1.3716	0,0480	0,6273
D	0,0403	0,0369	0,1274	0,2467	0,1129
Covarianza media	0,5100	0,5793	0,6273	0,1129	0,4574

Nota: le correlazioni al quadrato tra le quattro variabili A, B, C e D sono quantificate dai loro valori di scala sulla prima dimensione (triangolo in alto a destra della tabella, in corsivo) così come le loro correlazioni al quadrato con il punteggio totale (colonna di destra; cfr. *sq.corr.* nella colonna F1 della Tabella 2.3). Le varianze (diagonale della tabella) e le covarianze (triangolo in basso a sinistra della tabella) vengono quantificati, con la covarianza media di ciascuna variabile con se stessa e le altre mostrate nell'ultima riga in grassetto (ad esempio, 0,5793 = (0,4440 + 1,2666 + 0,5697 + 0,0369)/4). Si noti che queste covarianze medie sono identiche alle correlazioni al quadrato con il totale. Pertanto, la varianza del punteggio medio (the quantità massimizzata dall'MCA, sottolineata) è sia (a) la media dei quattro quadrati correlazioni dei punteggi delle domande con il punteggio totale e (b) la media dei quattro covarianze medie; in altre parole, è la media dell'intera varianza-covarianza 4 × 4 matrice. Notare inoltre la somma delle varianze delle quattro variabili, 1.1151 + 1.2666 + 1.3716 + 0.2467 = 4 che è la condizione di identificazione sui valori della scala. (Calcolare varianze e covarianze, dividere per $n = 871$, non $n - 1$.)

minimizzare, in questo caso, la seguente funzione di s , ovvero the media delle differenze Q al quadrato per ciascun individuo, mediata a turno su tutti *gli* n individui:

$$\frac{1}{n} \sum_{i=1}^n (\mathbf{z}_i - \mathbf{Z} \mathbf{s})^T (\mathbf{z}_i - \mathbf{Z} \mathbf{s}) + \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i^T \mathbf{z}_i - \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i^T \mathbf{Z} \mathbf{s} - \frac{1}{n} \sum_{i=1}^n \mathbf{s}^T \mathbf{Z}^T \mathbf{z}_i \quad (2.10)$$

Questo approccio è noto come *analisi di omogeneità* (Gifi 1990) e la funzione obiettivo (Equazione 2.10) è chiamata funzione di perdita. Qui La "perdita" si riferisce alla perdita di omogeneità, poiché lo sarebbe la perfetta omogeneità essere quando tutte le differenze $\mathbf{z}_i - \mathbf{Z} \mathbf{s}$ sono zero. Ancora una volta è richiesta una condizione di identificazione su \mathbf{s} , altrimenti la soluzione banale quando verranno trovati tutti gli elementi di \mathbf{s} costanti, con una perdita pari a zero. Con

lo stesso vincolo quadratico $sT \mathbf{Ds} = 1$ di prima, si può dimostrare che la perdita minima è ottenuta con gli stessi valori di scala ottimale descritti sopra, e il valore del minimo è pari a 1 meno il valore del corrispondente autovalore più grande del superindicatore matrice Z . Nel nostro esempio, gli autovalori successivamente massimizzati di 0,457 e 0,431 (vedi Tabella 2.4) corrispondono a perdite minime di 0,543 e 0,569, rispettivamente.

2.3 Approccio geometrico

L'approccio geometrico all'AC, introdotto nel Capitolo 1, risulta essere leggermente più problematico da generalizzare al caso multivariabile. Gran parte delle controversie su CA derivano da questa difficoltà, e qui chiariremo i problemi coinvolti in MCA come metodo grafico e proporremo una versione specifica di MCA che risolva in modo accettabile questi problemi. Affronteremo la geometria sia dalla prospettiva del ridimensionamento della distanza del chi quadrato che dalla prospettiva del biplot.

La Figura 2.1 mostra la consueta mappa CA della tabella di contingenza nella Tabella 2.1. La mappa viene stabilita utilizzando la teoria descritta nel Capitolo 1 e nella Sezione 2.2.1, vale a dire la SVD della matrice dei residui standardizzati, seguita dal calcolo delle coordinate principali per rappresentare i punti in una mappa. Le coordinate principali sono le coordinate standard moltiplicate per i rispettivi valori singolari (vedi Capitolo 1). Poiché le coordinate standard hanno una normalizzazione unitaria, le coordinate principali sono normalizzate per avere una somma (ponderata) dei quadrati uguale al rispettivo valore singolare quadrato della soluzione associata.

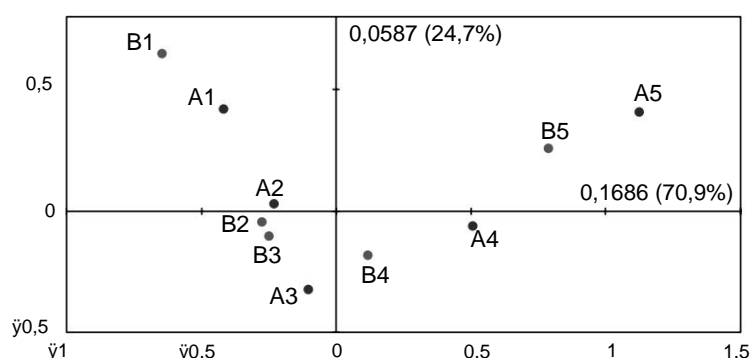


Figura 2.1 Mappa CA simmetrica della Tabella 2.1. La percentuale di inerzia visualizzata nella mappa è del 95,6%. Le categorie per le domande A e B vanno da 1 (fortemente d'accordo) a 5 (fortemente in disaccordo).

Il valore singolare quadrato λ^2 è chiamato inerzia principale, corrispondente a un asse principale, o dimensione (vedere Capitolo 1). La Figura 2.1 mostra i punti rappresentati dalle loro coordinate principali calcolate per i primi due assi principali.

2.3.1 Scaling della distanza chi-quadrato

La CA semplice è giustificata principalmente dall'uso della distanza chi-quadrato (χ^2) come misura della dissomiglianza tra i profili di riga e tra i profili di colonna di una tabella a due vie. Nella Figura 2.1, dove sia le righe che le colonne sono visualizzate in coordinate principali, le distanze tra i punti delle righe si avvicinano in modo ottimale alle distanze χ^2 tra i profili delle righe e le distanze tra i punti delle colonne si avvicinano in modo ottimale alle distanze χ^2 tra i profili delle colonne. Ricordiamo dal Capitolo 1 che la distanza χ^2 al quadrato tra i profili delle righe, ad esempio, ha questa forma:

$$d_{ij}^2 = \sum_{j=1}^J \frac{p_{ij}^2}{r_i c_j} \quad (2.11)$$

in modo che ogni j -esima differenza al quadrato tra gli elementi del profilo sia ponderata inversamente dal margine della colonna c_j .

MCA è l'applicazione di CA alla matrice del superindicatore Z o alla matrice di Burt C . Sebbene la distanza χ^2 abbia senso per una tabella di contingenza a due vie, ha meno giustificazione se applicata alle righe e alle colonne della matrice del superindicatore o della matrice del superindicatore. Matrice di Burt. Per illustrare questo problema, consideriamo lo stesso esempio di quattro variabili sull'atteggiamento nei confronti della scienza nel contesto ambientale.

Come mostrato nell'Equazione 2.11, le distanze χ^2 tra righe e tra colonne sono calcolate tra i loro profili: nel caso di distanze tra profili di riga, le masse delle colonne della matrice di corrispondenza vengono utilizzate inversamente come pesi nel calcolo della distanza. I profili riga della matrice dei superindicatori $Z = [Z_1 Z_2 Z_3 Z_4]$ sono vettori con elementi uguali a zero a parte quattro valori di $1/4$ nelle posizioni delle categorie selezionate dal caso corrispondente. Quando si calcola la distanza tra due righe, le differenze tra valori zero coincidenti e valori coincidenti di $1/4$ sono zero, quindi non danno alcun contributo alla misura della distanza, e quindi sono solo le differenze tra categorie non coincidenti che contano nella funzione distanza.

Queste differenze al quadrato diverse da zero (ciascuna pari a $1/16$ in questo caso), derivanti da disaccordi tra gli intervistati, vengono quindi ponderate

dagli inversi delle masse delle colonne corrispondenti, proporzionali a le frequenze marginali delle rispettive categorie e sommate danno le distanze χ^2 . Per le righe, questa misura di distanza appare equa ragionevole e la ponderazione è conforme al concetto di χ^2 che il contributo delle categorie a bassa frequenza deve essere potenziato perché la loro varianza è intrinsecamente inferiore. Tuttavia, Gower (Capitolo 3, questo volume) preferisce una versione non ponderata di questa distanza misurare.

Ben diversa è invece la situazione per i profili delle colonne di Z e difficile, se non impossibile, da giustificare. Qui facciamo la distinzione tra il calcolo (a) delle distanze tra due categorie della stessa variabile e (b) distanze tra due categorie di variabili diverse. Indichiamo la frequenza relativa della j -esima categoria di colonna da c_j (cioè, per una particolare variabile, la somma delle quantità c_j è 1). COME mostrate da Greenacre (1989), le distanze χ^2 al quadrato tra due Le categorie di colonna di una matrice di superindicatori sono, nei due casi:

1. $1/c_j + 1/c_j$ tra le categorie j e j della stessa variabile q
2. $1/c_j + 1/c_j - 2p_{jj} / (c_j c_j)$ tra le categorie j e j di diverse variabili q e q

dove p_{jj} è la frequenza relativa di occorrenza delle categorie j e j (in effetti, le formule di cui sopra sono le stesse, poiché la frequenza di cooccorrenza delle categorie j e j della stessa variabile è zero). IL la precedente distanza "entro variabile" ha poco senso, poiché dipende solo sulle frequenze marginali, indipendentemente dalla relazione con le altre variabili è. Quest'ultima distanza "tra variabili" ha almeno una leggera giustificazione in quanto la distanza diminuisce all'aumentare dell'associazione tra le categorie j e j , ma ancora una volta la dominante Il ruolo svolto dalle frequenze marginali è difficile da difendere.

La situazione migliora se si considerano le distanze intercategoriali calcolate sulla matrice di Burt piuttosto che su quella degli indicatori. Perché il La matrice di Burt è simmetrica, non fa differenza se la calcoliamo le distanze χ^2 tra righe o tra colonne. La distanza quadrata tra le categorie può essere descritto verbalmente come segue:

1. Tra le categorie j e j della stessa variabile q : Questo La distanza quadrata entro la variabile è la media di $(Q-1)$ distanze χ^2 al quadrato tra le categorie j e j calcolate in le tabulazioni incrociate della variabile q con tutte le altre variabili $\bar{q} \neq q$, ma includendo anche un termine non necessario dalla qtabulazione incrociata di q con se stesso. (Questo termine implica la distanza

tra due profili di unità in una sottomatrice sulla diagonale di C ed è quindi una componente importante della distanza complessiva, tendendo a gonfiare la distanza.)

2. Tra le categorie j e j' di diverse variabili q e q' :
Questa distanza quadrata tra variabili è una media di $(Q/2)$ distanze χ^2 al quadrato tra profili delle categorie j e variabili q cross non uguali a q o q' ma che ne includono due termini aggiuntivi che possono anche essere considerati non necessari. (Questi misurano le distanze tra un profilo e un profilo dell'unità nuovamente sulla diagonale di C , tendendo nuovamente a gonfiare la distanza tra categorie.)

Nonostante le difficoltà teoriche di cui sopra per giustificare l'intero spazio geometria chi-quadrato, MCA come applicato regolarmente, ovvero il CA di Z o C — con successo restaura modelli interessanti di associazione tra le variabili. Sembra che le proiezioni a bassa dimensione dei punti sono più validi delle loro controparti a dimensione intera, il che è un paradosso dal punto di vista del ridimensionamento multidimensionale. Un altro aspetto preoccupante è l'inflazione delle inerzie totali di Z e di C , che porta a tutte le percentuali di inerzia sugli assi principali essere artificialmente basso. Questa inflazione può essere compresa anche considerando il calcolo dell'inerzia totale per la matrice di Burt C e il elevati contributi forniti dalle matrici diagonali sulla diagonale del blocco. È chiaro che l'MCA di un set di dati a due variabili non darà lo stesso risultato risultati come CA; le coordinate standard saranno le stesse, ma le inerzie principali (e quindi le coordinate principali) e le loro percentuali di inerzia saranno diverse. Nella Sezione 2.3.3 ne definiamo un altro variante dell'MCA, chiamata analisi congiunta delle corrispondenze, che risolve tutti questi problemi in una certa misura. Mostriamo anche che un semplice la regolazione della scala nella soluzione MCA migliora notevolmente l'adattamento da un punto di vista di scala multidimensionale.

2.3.2 Biploma

Il biplot riguarda la ricostruzione dei dati in una mappa congiunta del righe e colonne, anziché la ricostruzione della distanza. Nel semplice Nel caso di una tabella a due vie, possiamo pensare di ricostruire diverse varianti della tabella a seconda del modo in cui pensiamo alla tabella: o come un insieme di righe, un insieme di colonne o semplicemente una tabella di voci a due vie dove righe e colonne sono entità simmetriche (vedi Capitolo 3). COME per illustrare questo approccio consideriamo un tavolo a due vie come un insieme di righe. Ad esempio, la Tabella 2.5a mostra i profili di fila del bidirezionale

Tabella 2.5 (a) Profili di riga della Tabella 2.1, inclusa la riga media profilo. (b) Profili di riga approssimativi stimati dal biplot di Figura 2.2 (il profilo medio è sempre rappresentato esattamente dall'origine della mappa).

	(a) Profili originali					Somma
	B1	B2	B3	B4	B5	
A1	0,227	0,235	0,252	0,185	0,101	1
A2	0,118	0,230	0,261	0,298	0,093	1
A3	0,115	0,235	0,309	0,358	0,083	1
A4	0,017	0,118	0,129	0,444	0,292	1
A5	0,000	0,063	0,104	0,229	0,604	1
Media	0,075	0,176	0,211	0,303	0,235	1

	(b) Profili stimati					Somma
	B1	B2	B3	B4	B5	
A1	0,226	0,239	0,253	0,181	0,102	1
A2	0,117	0,229	0,265	0,294	0,094	1
A3	0,024	0,226	0,283	0,393	0,074	1
A4	0,002	0,135	0,169	0,387	0,307	1
A5	0,026	0,034	0,034	0,329	0,578	1
Media	0,075	0,176	0,211	0,303	0,235	1

Nota: la differenza tra le due tabelle è l'errore di approssimazione biplot, misurato come $100 - 95,6\% = 4,4\%$ dell'inerzia totale della tabella.

tabella nella Tabella 2.1, cioè condizionata a ciascuna categoria di risposta di domanda A, la percentuale di intervistati che rientrano nella risposta categorie della domanda B. Il biplot può essere pensato come un modo per ricostruire questi profili di riga in una mappa. Greenacre e Hastie (1987) e Greenacre (1993a) mostrano come la mappa asimmetrica di CA, con i punti delle righe nelle coordinate principali e i punti delle colonne nelle coordinate standard, è un biplot di questi profili. Il vettore di direzione definito da ciascun punto della colonna, chiamato asse biplot, può essere calibrato in unità di profilo e il valore approssimativo del profilo può essere leggere la mappa semplicemente proiettando i punti della riga sulla colonna asse (Figura 2.2). Il successo della ricostruzione dei dati dal biplot in questo modo viene misurato dalla percentuale di inerzia spiegato dalla mappa: in questo caso è del 95,6%, così la ricostruzione ha un errore solo del 4,4%. La Tabella 2.5b riporta i valori stimati dal biplot di Figura 2.2, a testimonianza dell'elevata accuratezza della ricostruzione dei dati.

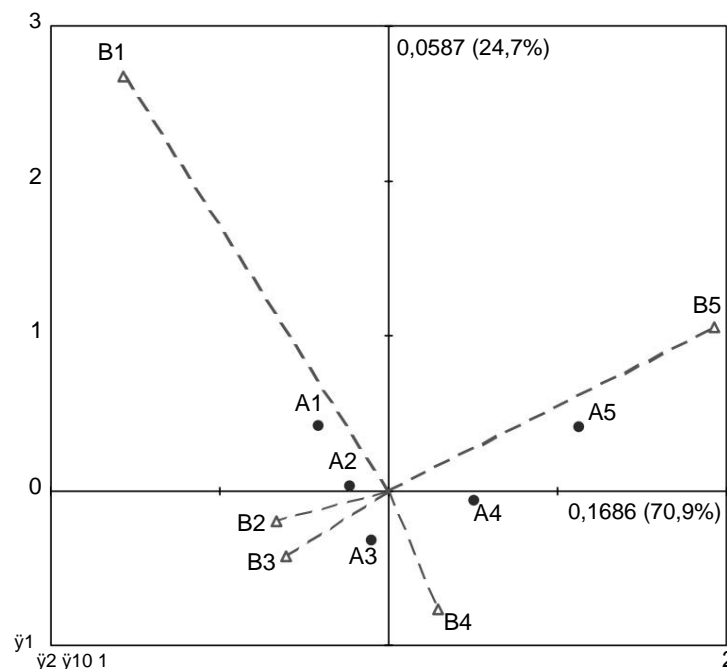


Figura 2.2 Mappa CA asimmetrica della Tabella 2.1 che mostra gli assi biplot. Le categorie dei punti della riga da A1 ad A5 possono essere proiettate su ciascun asse del biplot per leggere approssimazioni dei corrispondenti valori del profilo. La precisione di questa approssimazione è pari alla percentuale di inerzia visualizzata, che è pari al 95,6%; quindi è eccellente.

Interpretando la Figura 2.2 possiamo vedere, ad esempio, una direzione opposta alle categorie B2 e B3 che punta in basso a sinistra e la categoria B5 in alto a destra. Se proiettiamo A1, A2 e A3 su questa "dimensione" diagonale, è chiaro che si proiettano più o meno nella stessa posizione, dimostrando che i loro valori di profilo su B2, B3 e B5 sono simili, con valori di profilo su B2 e B3 sopra la media e quelli su B5 sotto la media (l'origine del biplot rappresenta sempre esattamente i valori medi del profilo). Questa deduzione dalla mappa può essere confermata nella tabella 2.5a, ed è solo per le categorie A4 e A5 che si registrano cambiamenti netti in questa direzione, aumentando in percentuale la risposta a B5 e diminuendo su B2 e B3.

Allo stesso modo, rispetto all'altra "dimensione" diagonale dall'alto a sinistra al basso a destra, che si oppone a B1 e B4, vediamo che A3, A4 e A5 si proiettano nelle stesse posizioni e quindi si stima che abbiano valori di profilo simili su B1 e B4. Ciò può essere verificato principalmente nella Tabella 2.5a, con l'unica eccezione del profilo di A5 su B4, che ha una frequenza osservata molto inferiore ai corrispondenti valori di A3 e A4. Questo errore di approssimazione rientrerebbe nell'inspiegabile inerzia del 4,4%.

Pensare in questo modo alla mappa congiunta mette in luce gli aspetti problematici dell'AC della matrice dell'indicatore Z o della matrice di Burt C. Nel caso di Z è inutile aspettarsi una buona approssimazione di una matrice di zeri e uno in una mappa bidimensionale di punti. È necessaria un'altra misura di qualità; ad esempio, si potrebbe dedurre da una mappa congiunta l'insieme più probabile di risposte per ciascun caso (riga) e poi contare quante di queste sono previsioni corrette (vedi Gower 1993; Greenacre 1994). La situazione è simile per la matrice di Burt C: qualsiasi tentativo di approssimare le matrici diagonali lungo la diagonale della matrice di Burt è chiaramente in conflitto con l'approssimazione delle tavole di contingenza più interessanti e rilevanti nel resto della matrice. In entrambi i casi le percentuali di inerzia saranno artificialmente basse a causa della natura strutturalmente altamente dimensionale delle matrici analizzate.

La Figura 2.3 mostra l'AC della matrice di Burt della Tabella 2.3, che rappresenta solo il 35,1% dell'inerzia totale; tuttavia la sua interpretazione è chiara: possiamo vedere lo stesso modello di associazione per le domande A e B già visto nella Figura 2.1, insieme a un modello simile di associazione con la domanda C. Ma le categorie di risposta per la domanda D non sono

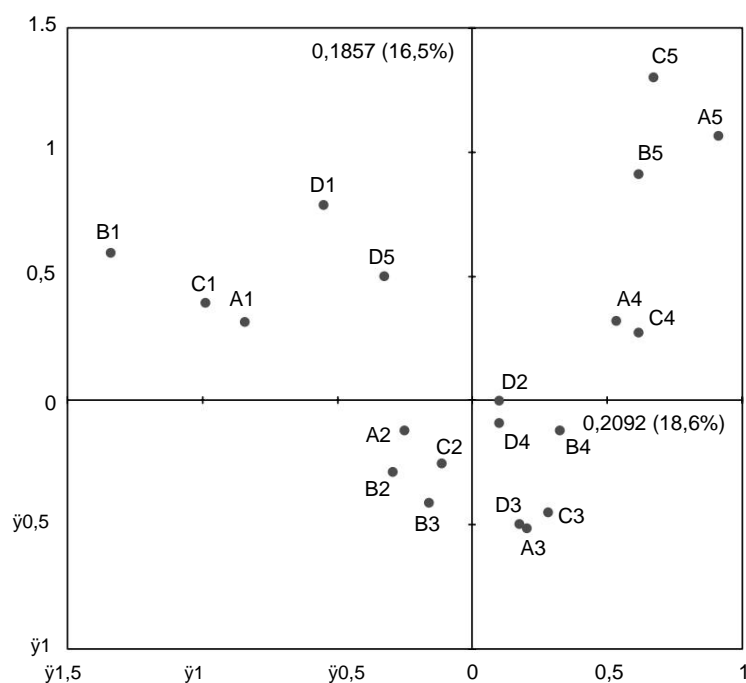


Figura 2.3 Mappa di analisi delle corrispondenze multiple della matrice di Burt della Tabella 2.3. La percentuale di inerzia spiegata è del 35,1%.

del tutto in linea con gli altri tre. Le categorie D1 e D5 di forte accordo e forte disaccordo si collocano all'interno dell'arco formato dalle altre domande, abbastanza vicine tra loro anche se agli estremi opposti della scala. Ciò dimostra chiaramente l'incompatibilità di questa domanda con le altre.

Notate quanto è diversa la scala della Figura 2.3 rispetto alla Figura 2.1, e come i punti sono stati spostati verso l'esterno nell'analisi della matrice di Burt. La maggior parte del problema delle basse percentuali di inerzia è dovuto a questo cambiamento di scala e questo può essere risolto con un semplice aggiustamento della scala della soluzione. Ciò sarà meglio spiegato dopo un resoconto dell'analisi congiunta della corrispondenza.

2.3.3 Analisi congiunta della corrispondenza

Come abbiamo appena visto, quando si applica CA alla matrice di Burt, le sottomatrici diagonali sulla "diagonale" della matrice a blocchi C gonfiano sia le distanze chi-quadrato tra i profili che l'inerzia totale di quantità artificiali. Nel tentativo di generalizzare la CA semplice in modo più naturale a più di due variabili categoriali, l'analisi delle corrispondenze congiunte (JCA) tiene conto della variazione solo nelle tabelle "fuori diagonale" di C, ignorando le matrici sulla diagonale del blocco. Quindi, nel caso a due variabili ($Q = 2$) quando esiste una sola tabella fuori diagonale, JCA è identico in tutto e per tutto al semplice CA (che non è il caso di MCA di Z o C, che danno inerzie principali diverse).

La soluzione non può più essere ottenuta con una singola applicazione della SVD e vari algoritmi sono stati proposti da Greenacre (1988), Boik (1996) e Tateneni e Browne (2000). Ad esempio, Greenacre (1988) descrive un algoritmo dei minimi quadrati alternati che tratta le matrici sulla diagonale del blocco come valori mancanti. L'algoritmo procede nei seguenti passi:

1. Eseguire MCA applicando CA alla matrice di Burt C e scegliere la dimensionalità S^* della soluzione (ad esempio, $S^* = 2$ è la più tipica).
2. Facoltativamente, eseguire una regolazione della soluzione lungo ciascuna delle dimensioni S^* per migliorare l'approssimazione alle matrici dei blocchi fuori diagonale (vedere la Sezione 2.3.4 di seguito).
3. Dalla mappa risultante, ricostruire i valori nei blocchi diagonali di C nello stesso modo in cui i dati sono stati ricostruiti nel biplot, utilizzando la formula di ricostruzione (vedi appendice, Equazione A.7). Sostituisci i valori originali nei blocchi diagonali con queste stime, chiamandola matrice di Burt modificata C^* .

4. Eseguire un'altra CA sulla matrice risultante C^* con modificata diagonale del blocco.
5. Ripetere i passaggi 3 e 4, ovvero sostituire i blocchi diagonali dei valori ricostruiti nella nuova soluzione, eseguendo nuovamente l'AC per ottenere un'altra soluzione, e così via, finché il processo non converge. La convergenza può essere misurata dalla massima differenza assoluta tra i valori sostituiti nei blocchi diagonali nell'iterazione attuale e i loro valori corrispondenti sostituiti durante l'iterazione precedente.

La Figura 2.4 mostra i risultati di una JCA bidimensionale applicata alla matrice di Burt della Tabella 2.3, dove abbiamo intenzionalmente lasciato la scala esattamente come nella Figura 2.3. Confrontando queste due figure si nota l'elevato grado di somiglianza nello schema delle categorie di risposta, ma soprattutto un cambiamento nella scala, con la mappa JCA ridotta in scala su entrambi gli assi, ma soprattutto sul secondo. La maggior parte delle proprietà della CA semplice si trasferiscono alla JCA, soprattutto la ricostruzione dei profili rispetto agli assi biplot (Greenacre 1993a: capitolo 16).

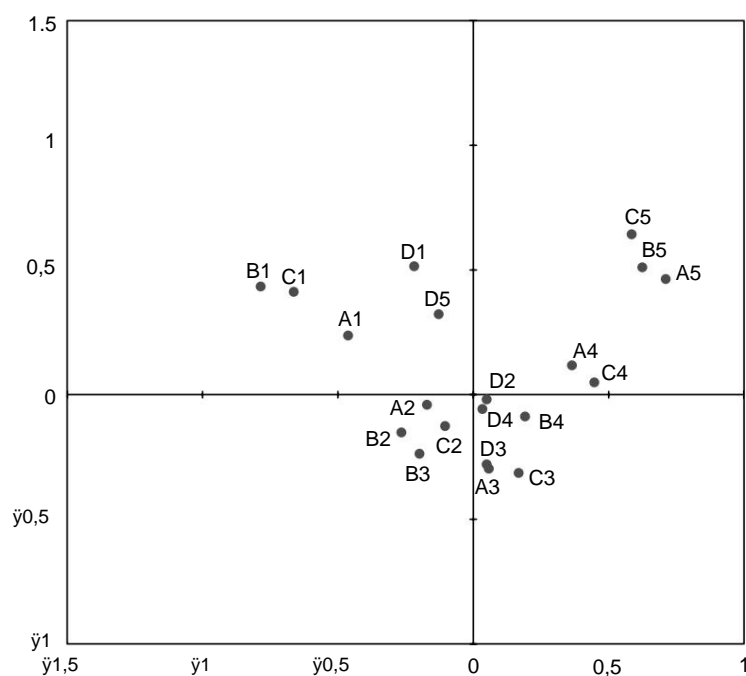


Figura 2.4 Mappa di analisi congiunta delle corrispondenze della matrice di Burt della Tabella 2.3. La percentuale di inerzia spiegata è dell'85,7%.

Rispetto ai normali CA e MCA, ci sono due aspetti da considerare ricordarsi nel calcolare la percentuale di inerzia spiegata dal carta geografica. Innanzitutto occorre calcolare la percentuale per le due dimensioni della soluzione insieme, non separatamente, poiché le dimensioni in JCA non sono nidificati. In secondo luogo, nella soluzione finale (la CA del file modificato Matrice di Burt all'iterazione finale), il modo consueto di calcolare la la proporzione dell'inerzia spiegata coinvolge il rapporto tra la somma di le prime due inerzie principali e l'inerzia totale, ma entrambe le numeratore e denominatore di questa somma includono un importo dovuto al blocchi diagonali modificati, che si adattano esattamente alla soluzione. Questo l'importo, che può essere calcolato in vari modi (cfr. l'appendice computazionale, sezione A.5), deve essere scontato sia dal numeratore e denominatore per ottenere la percentuale di (fuori diagonale) spiegata l'inerzia. In questo esempio, viene considerata la percentuale di inerzia secondo la mappa JCA è pari all'85,7%, molto superiore al 35,1% spiegato nella mappa MCA basata sulla matrice di Burt. Il valore dell'85,7% misura adeguatamente il successo dell'approssimazione dei blocchi fuori diagonale relativo all'inerzia totale solo di questi blocchi, non influenzati dal blocchi diagonali. Questa sarebbe la qualità della mappa considerata come anche un biplot MCA: ovvero esprimere tutti e sei i blocchi fuori diagonale come profili (profili di righe o colonne, nel triangolo superiore o inferiore del Matrice di Burt), quindi la qualità di ricostruire questi profili come descritto nella sezione 2.3.2 sarebbe dell'85,7% e l'errore di ricostruzione, o residuo, sarebbe del 14,3%.

2.3.4 Regolazione delle inerzie in MCA

Poiché la differenza principale tra MCA e JCA nella Figura 2.3 e La Figura 2.4 mostra il cambiamento di scala, è possibile rimediare parzialmente problema della percentuale di inerzia in un MCA regolare mediante un compromesso tra la soluzione MCA e l'obiettivo JCA utilizzando una scala semplice riadattamenti della soluzione MCA. In questo approccio l'inerzia totale è misurato (come in JCA) dall'inerzia media di tutti i blocchi fuori diagonale di \mathbf{C} , calcolato direttamente dalle tabelle stesse oppure adeguando l'inerzia totale di \mathbf{C} eliminando i contributi fissi del blocchi diagonali come segue:

$$\text{inerzia media fuori diagonale} = \frac{Q}{Q-1} \frac{\sum \tilde{y}^2}{\sum y^2} \text{inerzia}(\mathbf{C}) - \frac{JQ}{Q^2} \frac{\sum \tilde{y}^2}{\sum y^2} \quad (2.12)$$

Le parti dell'inerzia vengono quindi calcolate dalle inerzie principali \tilde{y}_s ² di C (o dalle inerzie principali \tilde{y}_s di Z) come segue: per ogni $\tilde{y}_s \tilde{y} 1/Q$, calcolare le inerzie corrette:

$$I_s^{agg} = \frac{\tilde{y}_s Q}{\tilde{y}_s Q} \frac{\tilde{y}_s^2}{1} \frac{\tilde{y}_s}{\tilde{y}_s} \frac{1}{Q} \frac{\tilde{y}_s^2}{\tilde{y}_s} \quad (2.13)$$



e poi esprimerli come percentuali dell'equazione 2.12. Sebbene questi percentuali sottostimano quelle di un JCA, migliorano notevolmente i risultati di un MCA e sono consigliati in tutte le applicazioni dell'MCA.




Un'ulteriore proprietà delle inerzie principali corrette \tilde{y}_s ^{agg} è che lo sono identiche alle inerzie principali dell'AC semplice \tilde{y}_s ² nel caso di due variabili categoriali, dove $Q = 2: \tilde{y}_s$ in $\tilde{y}_s^{agg} = 4(\tilde{y}_s - 1/2)^2$, da quando abbiamo mostrato Sezione 2.2.2 la relazione $= 1/2(1 +)$. precedenza nella \tilde{y}_s

Nel nostro esempio l'inerzia totale di C è pari a 1.1138, e il ² $\tilde{y} 1/Q^2$ le prime sette inerzie principali sono tali che $\tilde{y}_s \tilde{y} 1/Q$, cioè $\tilde{y}_s = 1/16$. L'inerzia media fuori diagonale è pari a 0,17024, come mostrato nella Tabella 2.6 insieme alle diverse possibilità di inerzia e percentuali di inerzia. Quindi quella che sembra essere una percentuale spiegata in due dimensioni del 22,2% (= 11,4 + 10,8) nell'analisi dell'indicatore matrice Z, ovvero il 35,1% (= 18,6 + 16,5) nell'analisi della matrice di Burt C, viene mostrato avere un limite inferiore del 79,1% (= 44,9 + 34,2) quando le inerzie principali vengono adeguate. Rispetto alla soluzione MCA modificata, la soluzione JCA per questo esempio (figura 2.4) fornisce un ulteriore beneficio di 6,4 punti percentuali nell'inerzia spiegata, con un'età percentuale spiegata dell'85,7%. Sottolineiamo ancora che in JCA le soluzioni ci sono non nidificati, quindi le percentuali sono riportate per l'intera soluzione (in questo caso bidimensionale), non per dimensioni individuali.

Proponiamo la soluzione aggiustata come quella da segnalare di routine; non solo migliora notevolmente la misura di adattamento, ma anche rimuove l'incoerenza su quale delle due matrici analizzare, indicatore o Burt. La soluzione corretta è riportata nella Figura 2.5 e ha le stesse coordinate standard della Figura 2.3, ma utilizza quelle aggiustate inerzie principali per calcolare le coordinate principali, che portano al migliore qualità di visualizzazione. Ancora una volta abbiamo lasciato la scala identica a Figura 2.3 e Figura 2.4 a scopo di confronto.

Benzécri (1979) ha proposto le stesse inerzie corrette (Equazione 2.13), ma li esprime come percentuali della propria somma rispetto al dimensioni s per le quali $\tilde{y}_s \tilde{y} 1/Q$ (vedi un esempio nel Capitolo 5). Questo approccio va all'estremo opposto nel dare un'espressione eccessivamente ottimistica dell'inerzia spiegata, poiché spiega il 100% spazio delle dimensioni per cui $\tilde{y}_s \tilde{y} 1/Q$ (ci sono sei dimensioni

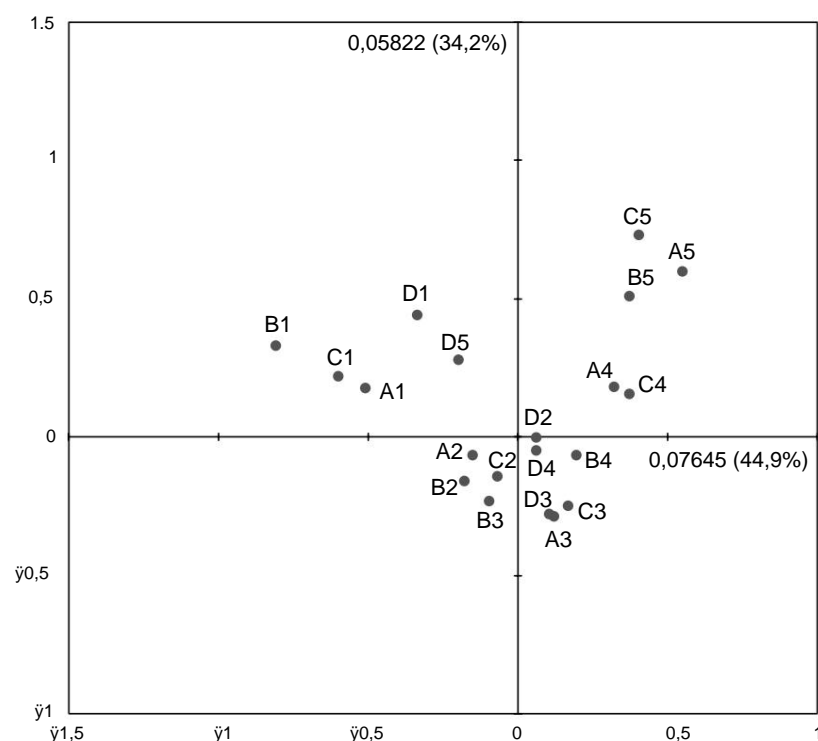


Figura 2.5 Mappa delle corrispondenze multiple della Tabella 2.3 con correzione delle inerzie principali (e quindi della scala delle coordinate principali) lungo ciascuna dimensione. La percentuale di inerzia spiegata è almeno del 79,1%. Si noti il cambiamento di scala rispetto alla Figura 2.3.

nel nostro esempio, come visto nella Tabella 2.6) quando in realtà i dati non sono ricostruiti esattamente nella mappa.

2.4 Punti supplementari

Finora abbiamo descritto tre diverse modalità di esecuzione dell'MCA:

Variante 1: CA della matrice degli indicatori

Variante 2: CA della matrice di Burt

Variante 3: una versione modificata delle varianti 1 o 2 che unifica e rettifica il problema del ridimensionamento, fornendo un'unica soluzione indipendentemente dalla matrice considerata, con misure notevolmente migliorate dell'inerzia spiegata

In tutte queste variazioni, le coordinate standard dei punti di categoria rimangono le stesse; cambiano solo le inerzie principali. Inoltre, abbiamo introdotto un metodo alternativo, JCA, che ha una soluzione diversa dalle varianti precedenti ed è analogo ai minimi quadrati

analisi fattoriale in quanto si concentra solo sulle associazioni tra variabili. In tutti questi casi è possibile visualizzare punti supplementari nel modo consueto per arricchire l'interpretazione (vedi Capitolo 5, dove questo aspetto è discusso in modo approfondito per il caso MCA). Qui definiamo un modo di visualizzare i punti supplementari che non dipende dalla variante del metodo utilizzato. Nel nostro esempio, prenderemo in considerazione tre variabili demografiche supplementari: genere, età e istruzione (le descrizioni complete delle categorie sono fornite nella Sezione 2.1).

Per motivare il nostro approccio consideriamo il caso della matrice degli indicatori, dove le categorie supplementari possono essere pensate come punti di riga o di colonna. Ad esempio, le categorie maschili e femminili possono essere aggiunte come due variabili fittizie di colonne supplementari o come due righe supplementari contenenti le frequenze per maschi e femmine nelle categorie di risposta. Queste due alternative sono equivalenti fino ai fattori di scala, come spiegheremo ora. Per posizionare una categoria di colonna supplementare utilizzando la cosiddetta relazione di transizione, o baricentrica, tra righe e colonne (vedi, ad esempio, Greenacre 1984 e il Capitolo 1 di questo volume), dobbiamo considerare tutti i punti rispondenti (righe) in standard coordinare le posizioni sulla mappa. Quindi ogni categoria di colonna, attiva o supplementare, è situata (nelle coordinate principali) alla media degli intervistati che ricadono in quella categoria. In alternativa, per posizionare una riga supplementare, ad esempio "maschio", dobbiamo considerare tutte le categorie di colonne attive in posizioni di coordinate standard; quindi il punto della riga "maschio" sarà pari alla media ponderata dei punti della colonna, utilizzando il profilo "maschio" tra le colonne attive. Ricorda che il profilo di frequenza "maschile" ha la somma di 1 nelle domande Q , quindi la sua posizione è una media delle medie; per ogni domanda, il gruppo "maschi" ha una posizione media in base alle frequenze maschili nelle categorie di quella particolare domanda, e la posizione finale di "maschi" è una media semplice di queste medie. Possiamo mostrare che la posizione di un punto supplementare come riga è la stessa della colonna supplementare fittizia, ma è ridotta su ciascuna dimensione del corrispondente valore singolare, cioè dello stesso fattore di scala che collega il principale alle coordinate standard su ciascuna dimensione (vedi Greenacre 1984: capitolo 5.1 per una dimostrazione di questo risultato). Pertanto un modo semplice per unificare la rappresentazione dei punti supplementari in tutte le situazioni sarebbe quello di pensare alle categorie supplementari sempre come medie delle principali posizioni coordinate degli intervistati, nel qual caso entrambi gli approcci daranno esattamente gli stessi risultati.

La nostra proposta è quindi la seguente: utilizzando le coordinate principali dei punti degli intervistati, calcolare le posizioni medie per le categorie supplementari, ad esempio la posizione media per i punti maschili, i punti femminili. Poiché è solo per la matrice dell'indicatore (variante 1 sopra elencata) che noi

(automaticamente) hanno punti rispondenti nell'AC, dobbiamo precisare cosa intendiamo per punti rispondenti nel caso della matrice di Burt e dell'analisi aggiustata (rispettivamente varianti 2 e 3 sopra). In questi ultimi casi, i punti degli intervistati vengono visualizzati, almeno teoricamente, come punti supplementari, cioè come medie delle categorie di colonne, in coordinate standard, secondo i rispettivi modelli di risposta. Poiché in tutte e tre le varianti dell'MCA queste coordinate standard sono identiche, i punti rispondenti avranno esattamente le stesse posizioni in tutti e tre i casi.

Pertanto, quando calcoliamo la media delle loro posizioni in base a una variabile supplementare, mostrando ad esempio la media dei punti maschili e medi femminili, anche i risultati saranno identici. Ma si noti che, per ottenere queste posizioni medie, non dobbiamo effettivamente calcolare tutti i punti originali degli intervistati. I calcoli possono essere eseguiti in modo molto più efficiente, grazie alle relazioni di transizione, semplicemente aggiungendo tabulazioni incrociate come righe o colonne supplementari. Di seguito sono riepilogati i calcoli in ciascun caso, presupponendo che una variabile supplementare sia codificata nell'indicatore Z_T denota l'insieme concatenato di tabulazioni incrociate si formano come Z , in modo che Z della variabile supplementare con le Q variabili attive:

1. Nel caso della matrice degli indicatori Z , avremmo già le coordinate principali degli intervistati, quindi possiamo fare il calcolo delle medie direttamente o aggiungere come flessibile punti della riga mentale le tabulazioni incrociate $Z_s Z_T$ della variabile supplementare con le variabili attive.
2. Nel caso della matrice di Burt C , non è necessario calcolare le posizioni degli intervistati (se richiesto per altri motivi, ciò potrebbe essere fatto aggiungendo Z come righe supplementari alla matrice di Burt C). Invece, possiamo semplicemente aggiungere la croce Z come righe supplementari (o $Z_T Z$ come tabulazioni Z_s flessibili colonne mentali), che porta alle stesse posizioni per le categorie supplementari della variante 1.
3. Nel caso dell'analisi aggiustata, facciamo come nella variante 2, poiché è solo a posteriori che aggiustiamo gli autovalori, e questo aggiustamento influenza solo le posizioni delle coordinate principali dei punti della categoria attiva, non le categorie supplementari che, ripetiamo, sono definite come medie dei punti degli intervistati;
4. Nel caso di JCA, si tratta ancora una volta di una semplice aggiunta di righe o colonne supplementari, come nelle varianti 2 e 3, alla matrice di Burt modificata C^* durante l'iterazione finale dell'algoritmo.

La tabella 2.7 mostra le tavole incrociate $Z_s Z_T$, e la figura 2.6 mostra le posizioni dei punti supplementari nell'MCA corretto (variante 3,

[illegible]

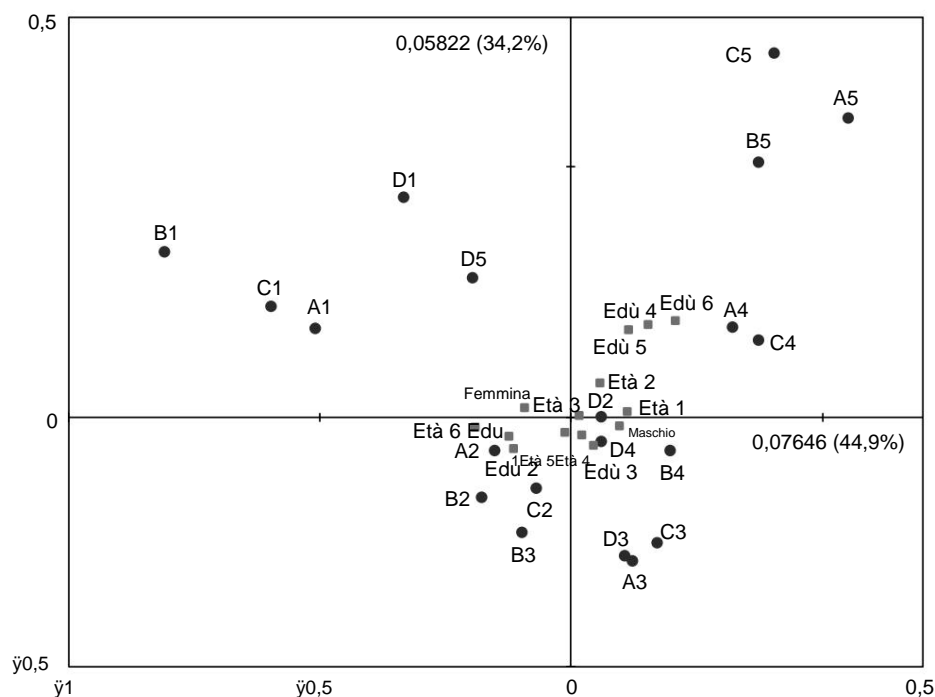


Figura 2.6 Soluzione MCA aggiustata, che mostra le posizioni delle categorie supplementari di sesso, età e istruzione.

cioè i punti supplementari sono sovrapposti alla Figura 2.5).

Qui possiamo vedere che le fasce di età e i gruppi di istruzione mostrano una tendenza orizzontale, con gli intervistati più giovani a destra che si spostano verso gli intervistati più anziani a sinistra, e i gruppi di istruzione inferiore a sinistra che si spostano verso i gruppi di istruzione superiore a destra. Inoltre si nota che i tre gruppi dell'istruzione superiore (dalla scuola secondaria in su) si separano a destra verso i poli di forte disaccordo delle domande, indicando che sono particolarmente favorevoli alla scienza. Troviamo anche il punto medio maschile sul lato destro e il punto medio femminile sul lato sinistro. Ricordare che queste variabili supplementari sono state aggiunte separatamente alla mappa e non in combinazione, ovvero il fatto che i maschi siano a destra e i gruppi con istruzione superiore siano a destra non implica che siano solo i maschi con istruzione superiore ad essere più favorevole alla scienza. Per vedere la posizione delle donne con un livello di istruzione più elevato, ad esempio, sarebbe necessario eseguire una codifica interattiva delle variabili demografiche. Ciò può essere fatto, ad esempio, codificando separatamente i sei gruppi educativi per maschi e femmine, fornendo 12 combinazioni di genere e istruzione, ciascuna rappresentata come un punto supplementare separato.

2.5 Discussione e conclusioni

Abbiamo dimostrato che estendere l'AC di due variabili al caso di più variabili non è una questione semplice, soprattutto nel caso geometrico.

Come spiegato nel Capitolo 1, l'AC semplice viene generalmente applicata a situazioni in cui vengono incrociate due diverse tipologie di variabili, ad esempio il paese di residenza e la risposta a una domanda del sondaggio. L'MCA viene applicato a un insieme di variabili, preferibilmente tutte con le stesse scale di risposta, che ruotano attorno a un problema particolare e dove siamo interessati ai modelli di associazione tra le variabili. Detto in altro modo, nell'AC semplice siamo interessati alle associazioni tra due variabili o tra due insiemi di variabili, mentre nell'MCA siamo interessati alle associazioni all'interno di un insieme di variabili.

Sebbene i concetti geometrici dell'AC semplice non si trasferiscano facilmente al caso a variabili multiple, l'aggiustamento delle inerzie principali e i metodi alternativi, come JCA, risolvono parzialmente la situazione. Poiché MCA ha interessanti proprietà di ottimalità dei valori di scala (grazie al raggiungimento della massima intercorrelazione e quindi della massima affidabilità in termini di alfa di Cronbach), il compromesso offerto dalla soluzione MCA aggiustata è quello più sensato e quello che raccomandiamo. L'aggiustamento, descritto nella Sezione 2.3.4, è facile da calcolare e modifica semplicemente la scala su ciascuna dimensione della mappa per approssimare al meglio le due tabelle di associazione tra coppie di variabili, lasciando intatto tutto il resto nella soluzione. Grazie a questo aggiustamento si ottengono stime delle inerzie spiegate molto più vicine ai valori reali rispetto ai valori pessimistici ottenuti in MCA.

Per questo motivo proponiamo la soluzione MCA adattata come quella da utilizzare normalmente in tutte le applicazioni MCA (grafiche).

La soluzione corretta ha anche la bella proprietà di essere identica alla semplice CA di una tavola incrociata nel caso di due variabili. JCA riproduce perfettamente anche la semplice CA nel caso a due variabili, poiché anch'essa si concentra esclusivamente sulla singola tavola incrociata fuori diagonale. JCA presenta l'ulteriore vantaggio nel caso multivariabile di ottimizzare l'adattamento a tutte le tabulazioni incrociate fuori diagonale di interesse.

È possibile aggiungere punti supplementari categoriali a qualsiasi variante dell'MCA, nonché alla JCA, come media degli intervistati che rientrano nelle categorie corrispondenti. Ciò equivale semplicemente ad aggiungere le tabulazioni incrociate delle variabili supplementari con le variabili attive come righe o colonne supplementari. Le soluzioni JCA e MCA aggiustato presentano il vantaggio che i punti attivi sono ridotti in scala rispetto alle soluzioni per l'indicatore e le matrici di Burt, quindi

portando ad una maggiore dispersione relativa dei punti supplementari nella mappa congiunta dei punti attivi e supplementari.

Non abbiamo affrontato l'importante argomento di imporre vincoli alla soluzione MCA, come vincoli lineari o di ordine. Il capitolo 4 fornisce una trattazione completa e aggiornata di questi problemi.

Nota sul software

Le analisi di questo capitolo sono state eseguite utilizzando XLSTAT (www.xlstat.com), distribuito da Addinsoft, e le funzioni R per CA, MCA e JCA, scritte da Oleg Nenadić. Entrambe le implementazioni, descritte nell'appendice di questo libro, includono la regolazione delle inerzie descritta in questo capitolo.