Carlo M. Porceddu Cilione

ANALISI DELLE CORRISPONDENZE

IL METODO E LE APPLICAZIONI

PREMESSA

Questo libro è frutto della passione che Carlo Maria Porceddu Cilione nutriva per l'analisi delle corrispondenze. Carlo ne ha redatto le pagine con la dedizione che profondeva in tutte le attività che hanno caratterizzato il suo percorso di ricercatore e studioso di statistica, ma purtroppo non è riuscito a portarne a termine la stesura. Nonostante questo, il contributo che ci lascia è estremamente ricco e foriero di spunti per sviluppi futuri. Per questa ragione, noi membri del Dipartimento di Scienze Aziendali dell'Alma Mater Studiorum - Università di Bologna, assieme ai familiari e ai cari di Carlo, abbiamo voluto portare alla luce queste pagine pubblicandole così come Carlo ce le ha lasciate, consci che l'incompletezza di alcune parti e qualche pagina mancante non vanno a discapito del valore scientifico dell'opera.

Molti di noi ritengono di avere un debito intellettuale nei confronti di Carlo e ritengono che la pubblicazione del suo elaborato sia un modo per ringraziarlo dell'impegno come docente e della disponibilità sempre mostrata nell'aiutarci nei nostri percorsi di ricerca. Questa premessa è anche l'occasione per ringraziare i familiari, che ci hanno messo a disposizione l'elaborato, e Mariachiara Colucci e Marco Visentin che hanno materialmente reso possibile la pubblicazione del libro.

Con grande stima, come colleghi, e con affetto, come compagni di un lungo viaggio, ringraziamo Carlo e ci pregiamo di dare in stampa il suo manoscritto.

I membri del Dipartimento di Scienze Aziendali
(Alma Mater Studiorum - Università di Bologna)

A mia madre e alla memoria di mio padre

Indice dei simboli: matrici e vettori

Simbolo	Ordine	Grandezza
$\mathbf{N}$	$I \times J$	matrice di contingenza
$\mathbf{r}_i$	$J \times 1$	profilo della i^{ma} riga e modalità di $\mathbf{N}$
$\mathbf{c}_j$	$I \times 1$	profilo della j^{ma} colonna e modalità di $\mathbf{N}$
$\mathbf{R}$	$I \times J$	matrice degli I profili delle righe
$\mathbf{C}$	$I \times J$	matrice dei J profili delle colonne
$\mathbf{O}$	$I \times J$	matrice omogenea con profili tutti eguali
$\mathbf{S}$	$I \times J$	matrice degli scarti relativi dal livello medio
$\overline{\mathbf{r}}$	$J \times 1$	profilo riga medio e delle masse dei pr. delle colonne
$\overline{\mathbf{c}}$	$I \times 1$	profilo colonna medio e delle masse dei pr. delle righe
$\mathbf{D}_{\overline{\mathbf{r}}}$	$J \times J$	matrice diagonale delle masse dei pr. delle colonne
$\mathbf{D}_{\overline{\mathbf{c}}}$	$J \times J$	matrice diagonale delle masse dei profili delle righe
$\overline{\mathbf{R}}$	$I \times J$	matrice con righe tutte eguali al profilo riga medio
$\overline{\mathbf{C}}$	$I \times J$	matrice con colonne eguali al profilo colonna medio
$\mathbf{D}_{\overline{\mathbf{r}}}^{-1}$	$J \times J$	metrica o matrice delle distanze in $\Re^J$
$\mathbf{D}_{\overline{\mathbf{c}}}^{-1}$	$I \times I$	metrica o matrice delle distanze in $\Re^I$
$\mathbf{e}_{\mathrm{k}}$	$n \times 1$	vettore unità: la componente k è 1, tutte le altre 0
$\mathbf{0}_{\mathrm{n}}$	$n \times 1$	vettore zero, o nullo, con componenti tutte 0
$\mathbf{1}_{\mathrm{n}}$	$n \times 1$	vettore somma con componenti tutte 1
$\mathbf{I}_{\mathrm{n}}$	$n \times n$	matrice diagonale identità
$\mathbf{D}_{\lambda}$	$A \times A$	matrice diagonale degli autovalori
$\mathbf{v}_a, \mathbf{v}_a^{\star}$	$J \times 1$	autovettori di rango a di $\Re^J$ con origine in $\mathbf{0}_{\mathrm{J}}$ e $\overline{\mathbf{r}}$
$\mathbf{u}_a, \mathbf{u}_a^{\star}$	$I \times 1$	autovettori di rango a di $\Re^I$ con origine in $\mathbf{0}_{\mathrm{I}}$ e $\overline{\mathbf{c}}$
$\mathbf{V}, \mathbf{V}^{\star}$	$J \times A$	matrici degli autovettori di $\Re^J$ con origine in $\mathbf{0}_{\mathrm{J}}$ e $\overline{\mathbf{r}}$
$\mathbf{U}, \mathbf{U}^{\star}$	$I \times A$	matrici degli autovettori di $\Re^I$ con origine in $\mathbf{0}_{\mathrm{I}}$ e $\overline{\mathbf{c}}$
$\mathbf{F}$	$I \times A$	matrice dei fattori (principali) dei profili delle righe
$\mathbf{G}$	$J \times A$	matrice dei fattori (principali) dei pr. delle colonne
$\hat{\mathbf{F}}$	$I \times A$	matrice dei fattori standard dei profili delle righe
$\hat{\mathbf{G}}$	$J \times A$	matrice dei fattori standard dei profili delle colonne
$\tilde{\mathbf{r}}$	$J \times 1$	profilo di una riga illustrativa
$\tilde{\mathbf{c}}$	$I \times 1$	profilo di una colonna illustrativa
$\mathbf{B}$	$J \times J$	matrice simmetrica di Burt
$\mathbf{P}_r, \mathbf{P}_c$	$J \times J$	matrici dei profili di righe e colonne di Burt
$\mathbf{M}$	$I \times J$	matrice reale generica, non di contingenza

Indice dei simboli: scalari

Simbolo	Grandezza
I, J	numero di righe e di colonne di una matrice
n_{i+} n_{+j}	totale marginale della riga i e della colonna j di $\mathbf{N}$
n_{++}	totale generale della matrice $\mathbf{N}$
Q	numero di variabili attive nell'Analisi delle Corrisp. Multiple
A	numero di autovalori positivi non banali
λ_a	autovalore di rango a e inerzia sull'asse fattoriale a
τ_a	tasso d'inerzia sull'asse fattoriale di rango a
$d^2_D(\mathbf{p}_k, \mathbf{p}_l)$	distanza distribuzionale tre due profili
$In_{\mathbf{0}}$	inerzia rispetto all'origine della nuvola di profili
$In_{\overline{\mathbf{r}}}$	inerzia rispetto al baricentro della nuvola dei profili delle righe
$In_{\overline{\mathbf{c}}}$	inerzia rispetto al baric. della nuvola dei profili delle colonne
$CTR_a(\mathbf{p})$	contributo del profilo $\mathbf{p}$ all'inerzia dell'asse fattor. di rango a
$COS^2_a(\mathbf{p})$	qualità della rappresentazione del profilo $\mathbf{p}$ sull'asse di rango a
$QLT_{A^*}(\mathbf{p})$	qualità della rappresent. di $\mathbf{p}$ nel sottospazio A^*-dimensionale
$INR_{A^*}(\mathbf{p})$	inerzia risp. al baricentro di $\mathbf{p}$ nel sottospazio A^*-dimensionale
$\Re^J$	spazio euclideo J-dimensionale degli I profili delle righe
$\Re^I$	spazio euclideo I-dimensionale dei J profili delle colonne

Convenzioni

Una lettera latina maiuscola in grassetto, come $\mathbf{N}$, è impiegata per indicare una matrice; una lettera latina minuscola in grassetto, come $\mathbf{u}$, per i vettori ed una lettera latina, minuscola o maiuscola, come a o J, per gli scalari. Quando lo scalare è un indice, come j, l'estremo superiore dell'insieme che può percorrere è indicato con la corrispondente lettera maiuscola, in questo caso J. Alcune lettere sono riservate per indicare grandezze specifiche, come gli indici i e j per indicare una riga ed una colonna generica di una matrice ed a per indicare il rango di un autovalore o di un asse fattoriale.

La notazione $\stackrel{\text{def}}{=}$ indica la definizione di un simbolo. Non si confonda la definizione con l'uguaglianza. Così, ad esempio, l'area S di un rettangolo è definita come prodotto della lunghezza a e b di due lati contigui: $S \stackrel{\text{def}}{=} a\,b$, mentre l'area del triangolo rettangolo che si ricava è eguale ad $A = \frac{1}{2}\,S$, ossia alla metà dell'area precedentemente definita.

Formulario dell'Analisi delle Corrispondenze semplici

(il simbolo è definito nella Sezione indicata)

Simbolo		Relazioni	Sez.
$\mathbf{N}$	=	$n_{++}\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{R} \;=\; n_{++}\,\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}$	1.3
$\mathbf{R}$	=	$1/n_{++}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-1}\,\mathbf{N} \;=\; \mathbf{D}_{\overline{\mathbf{c}}}^{-1}\,\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}$	1.5
$\mathbf{C}$	=	$1/n_{++}\,\mathbf{N}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-1} \;=\; \mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{R}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-1}$	1.7
$\overline{\mathbf{r}}$	=	$\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{1}_{\mathrm{J}} \;=\; \mathbf{R}^T\,\overline{\mathbf{c}} \;=\; \mathbf{R}^T\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{1}_{\mathrm{I}}$	1.6
$\overline{\mathbf{c}}$	=	$\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{1}_{\mathrm{I}} \;=\; \mathbf{C}\,\overline{\mathbf{r}} \;=\; \mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{1}_{\mathrm{J}}$	1.8
$\mathbf{O}$	=	$n_{++}\,\overline{\mathbf{c}}\,\overline{\mathbf{r}}^T$	1.9
$\overline{\mathbf{R}}$	=	$\mathbf{1}_{\mathrm{I}}\,\overline{\mathbf{r}}^T \;=\; \mathbf{1}_{\mathrm{I}}\,\mathbf{1}_{\mathrm{I}}^T\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{R} \;=\; \mathbf{1}_{\mathrm{I}}\,\mathbf{1}_{\mathrm{I}}^T\,\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}} \;=\; \mathbf{1}_{\mathrm{I}}\,\overline{\mathbf{c}}^T\,\mathbf{R}$	4.8
$\overline{\mathbf{C}}$	=	$\overline{\mathbf{c}}\,\mathbf{1}_{\mathrm{J}}^T \;=\; \mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{1}_{\mathrm{J}}\,\mathbf{1}_{\mathrm{J}}^T \;=\; \mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{R}\,\mathbf{1}_{\mathrm{J}}\,\mathbf{1}_{\mathrm{J}}^T \;=\; \mathbf{C}\,\overline{\mathbf{r}}\,\mathbf{1}_{\mathrm{J}}^T$	2.8
$\mathbf{U}^\star$	=	$\mathbf{C}\,\mathbf{V}^\star\,\mathbf{D}_\lambda^{-\frac{1}{2}}$	3.14
$\mathbf{V}^\star$	=	$\mathbf{R}^T\,\mathbf{U}^\star\,\mathbf{D}_\lambda^{-\frac{1}{2}}$	4.8
$\mathbf{F}$	=	$\mathbf{D}_{\overline{\mathbf{c}}}^{-1}\,\mathbf{U}^\star\,\mathbf{D}_\lambda^{\frac{1}{2}} = \mathbf{R}\,\mathbf{G}\,\mathbf{D}_\lambda^{-\frac{1}{2}} = \mathbf{R}\,\hat{\mathbf{G}}$	4.8
$\mathbf{G}$	=	$\mathbf{D}_{\overline{\mathbf{r}}}^{-1}\,\mathbf{V}^\star\,\mathbf{D}_\lambda^{\frac{1}{2}} = \mathbf{C}^T\,\mathbf{F}\,\mathbf{D}_\lambda^{-\frac{1}{2}} = \mathbf{C}^T\,\hat{\mathbf{F}}$	4.1
$\mathbf{D}_\lambda$	=	$\mathbf{F}^T\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{F} = \mathbf{G}^T\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{G}$	4.2
$\hat{\mathbf{F}}$	=	$\mathbf{F}\,\mathbf{D}_\lambda^{-\frac{1}{2}}$	4.8
$\hat{\mathbf{G}}$	=	$\mathbf{G}\,\mathbf{D}_\lambda^{-\frac{1}{2}}$	4.3
$\mathbf{I}$	=	$\hat{\mathbf{F}}^T\,\mathbf{D}_{\overline{\mathbf{c}}}\,\hat{\mathbf{F}} = \hat{\mathbf{G}}^T\,\mathbf{D}_{\overline{\mathbf{r}}}\,\hat{\mathbf{G}}$	
$\mathbf{1}_{\mathrm{I}}$	=	$\mathbf{D}_{\overline{\mathbf{c}}}^{-1}\,\overline{\mathbf{c}} \;=\; \mathbf{R}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-1}\,\overline{\mathbf{r}} \;=\; \mathbf{R}\,\mathbf{1}_{\mathrm{J}}$	2.2
$\mathbf{1}_{\mathrm{J}}$	=	$\mathbf{D}_{\overline{\mathbf{r}}}^{-1}\,\overline{\mathbf{r}} \;=\; \mathbf{C}^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-1}\,\overline{\mathbf{c}} \;=\; \mathbf{C}^T\,\mathbf{1}_{\mathrm{I}}$	2.2
1	=	$\mathbf{1}_{\mathrm{J}}^T\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{1}_{\mathrm{J}} \;=\; \mathbf{1}_{\mathrm{J}}^T\,\mathbf{R}^T\overline{\mathbf{c}} \;=\; \mathbf{1}_{\mathrm{I}}^T\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{1}_{\mathrm{I}} \;=\; \mathbf{1}_{\mathrm{I}}^T\,\mathbf{C}\,\overline{\mathbf{r}}$	
I	=	$\mathbf{1}_{\mathrm{I}}^T\,\mathbf{1}_{\mathrm{I}}$	1.3
J	=	$\mathbf{1}_{\mathrm{J}}^T\,\mathbf{1}_{\mathrm{J}}$	1.3
A	=	$\min(I,J) - 1$	4.8

e inoltre $\quad \mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}} = \mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{R} \qquad \overline{\mathbf{C}}\,\mathbf{R}^T = \mathbf{C}\,\overline{\mathbf{R}}^T$

Marchi registrati

Sono elencati qui sotto i nomi che a parere dell'autore sono marchio registrato. La loro presenza, od assenza, non implica alcuna valutazione del loro stato giuridico.

`Coca Cola` è marchio registrato di The Coca-Cola Company, Atlanta, GA, USA.

`SAS`, `SAS Stat`, `SAS Insight` e `JMP` sono marchi registrati del SAS Institute, Cary, NC, USA.

`SPAD` e `SPAD N` sono marchi registrati di Cisia-Ceresta, Saint-Mandé, France.

`BMDP` e `Ca` sono marchi registrati di .....

`SPSS` e `ANACOR` sono marchi registrati di .....

PREFAZIONE

L'Analisi delle Corrispondenze è un sofisticato metodo statistico per l'analisi di dati multidimensionali che trova applicazione in quasi tutte le discipline scientifiche. La sua capacità di tradurre in forma grafica praticamente ogni tipo di tabella di dati numerici, ne fa uno strumento d'analisi estremamente efficace e in rapida diffusione.

Questo libro costituisce una introduzione, per quanto possibile semplice, al metodo e alle sue applicazioni. Contemporaneamente esso offre al lettore un accesso ai più significativi risultati recentemente raggiunti in questo campo ed apparsi nella letteratura scientifica internazionale. Utenti, o potenziali utenti, ai quali questo testo è mirato sono:

- lo studente che per la prima volta si avvicina alle nuove idee dell'Analisi Multidimensionale dei Dati;
- il ricercatore che ha raccolto e organizzato i dati in forma di tabelle e ha necessità di analizzarli;
- lo statistico che desidera allargare i suoi interessi al campo multidimensionale.

Categorie con esigenze diverse e in qualche modo contrastanti, a cui si è cercato di far fronte ponendo l'accento sull'applicabilità; sfruttando a fondo l'aspetto geometrico e grafico; evidenziando tutti i passaggi matematici in modo da consentire anche al lettore meno addestrato di seguire passo passo lo svolgimento dei calcoli, ed infine eliminando ogni riferimento ad altre parti della Statistica non essenziale alla comprensione del metodo. Il libro non è esente da ripetizioni che, se allungano la stesura, aiutano però il lettore meno esperto a fissare l'attenzione sui punti cruciali, e contiene tutte le informazioni necessarie ad evitare al lettore l'incombenza di doverle reperire in altri testi. Si tratta quindi di un libro da leggere con attenzione, da seguire insomma "con la penna in mano" nel suo procedere sino alle conclusioni, dimostrate anziché semplicemente affermate. Queste, è noto, sono più a lungo ricordate

quando è spiegato come vi si perviene, mentre sono presto dimenticate quando sono soltanto affermate.

L'esclusiva concentrazione sull'Analisi delle Corrispondenze è finalizzata ad una trattazione esaustiva ed approfondita, e tuttavia contenuta entro un limitato numero di pagine. Lo sforzo compiuto è stato quello di rendere comprensibile la complessa teoria che sta alla base dell'Analisi delle Corrispondenze, per consentire una utilizzazione consapevole del software d'analisi ormai largamente disponibile, in un campo in cui la sola conoscenza teorica – per quanto aggiornata – non basta. Questo libro assegna, pertanto, un rilievo primario alla applicazione pratica del metodo illustrata, con molteplici esempi, nella seconda parte del primo volume.

Il formato adottato in questo libro è in due volumi. Le ragioni per tenere tabelle, mappe, grafici e figure separate dal testo sono eminentemente pratiche. La rappresentazione grafica dei risultati è alla base dell'Analisi delle Corrispondenze: l'inserimento di numerose mappe nel testo ne comporterebbe un indesiderabile spezzettamento. Di qui la scelta di collocare in un secondo volume i risultati grafici di quanto analizzato nel testo evitandone l'eccessiva frammentazione e consentendo al lettore di mantenere sott'occhio una mappa o una tabella di dati frequentemente ed in punti diversi richiamata nel testo, senza la necessità di rintracciarla ogni volta. Nel secondo volume, inoltre, hanno trovato collocazione quegli approfondimenti di specifici punti che, se inseriti nel testo, avrebbero interrotto il filo dell'esposizione.

La suddivisione del primo volume segue l'ovvia, ma logica, procedura di presentare il metodo nella prima parte e le applicazioni nella seconda. Tuttavia ai lettori nuovi all'argomento può risultare utile affrontare le due parti simultaneamente anziché successivamente: sono proprio le applicazioni – infatti – a rendere vive le equazioni. Nella prima parte sono state omesse le dimostrazioni concernenti le basi dell'Algebra lineare, che peraltro si possono trovare nei testi citati in bibliografia, e ciò per non appesantire la trattazione e, soprattutto, per mantenere il filo del ragionamento. I casi presentati nella seconda parte sono stati scelti in quanto tipici e simili, per analogia, a molti altri. Una rassegna esaustiva delle effettive e potenziali applicazioni di un metodo di analisi capace, come questo, della più ampia applicazione, sarebbe stata del tutto impossibile.

Gli eventuali errori, di stampa e non, restano – s'intende – responsabilità dell'Autore. La loro segnalazione sarà accolta con gratitudine.

Questo libro è la forma finale ed organica che hanno assunto gli ap-

punti, le note e il materiale didattico utilizzato nelle lezioni sull'Analisi Multidimensionale tenute dall'Autore agli studenti delle Facoltà di Economia e Commercio e di Statistica dell'Università di Bologna. Il loro entusiasmo, i commenti, le domande e le osservazioni sono state la spinta che hanno quasi obbligato l'Autore a scrivere questo libro.

L'Autore desidera esprimere la sua riconoscenza a tutti coloro che, in varia misura, gli sono stati prodighi di aiuto, leggendo e commentando il manoscritto: all'amico prof. Paolo Braghieri che ha rimaneggiato alcune Sezioni per rendere più chiara e incisiva l'esposizione; al dr. S. Dalmonte per aver realizzato le TAVV. 6 e 7;...

Un particolare ringraziamento va poi agli autori, agli editori ed ai proprietari di copyright per aver concesso di riprodurre molte tabelle di dati, in particolare alla Soc. XX per i dati nella TAV. X, ...

.....Motta, C. Boari, B. Maggi, G.L. Marzocchi

ai colleghi M. Galli e C. Petrella che sono sempre riusciti a recuperare i files che il distratto autore aveva smarrito ed a mantenere in funzione il computer ove, con TEX , questo libro è stato scritto.

Carlo M. Porceddu Cilione

Verona, ...
informazioni, consigli, suggerimenti
per la sua pazienza e competenza
nessuno più di me è convinto che un insegnante impara dai suoi allievi

PRESENTAZIONE GENERALE DEL METODO

Poiché l'Analisi delle Corrispondenze è ancora poco nota nel nostro Paese, può essere utile a quanti vi si accostano per la prima volta avere una presentazione schematica dei vari aspetti della metodologia.

Le *origini* dell'Analisi delle Corrispondenze risalgono lontano nel tempo, ma soltanto in anni recenti questa potente metodologia statistica ha trovato pratica applicazione grazie alla disponibilità degli elaboratori elettronici. La sua diffusione, lenta all'inizio e confinata principalmente in Francia e Giappone, è stata costante, ma la sua importanza è ora ovunque riconosciuta. Il recente inserimento di specifiche routines di calcolo in alcuni dei più diffusi sistemi di software statistico, ne è autorevole conferma.

Oggetto dell'analisi sono le matrici di contingenza, i cui elementi indicano il numero di volte che sono state rilevate congiuntamente le caratteristiche di due diverse grandezze. In matrici di questo tipo le righe e le colonne giocano ruoli analoghi, in quanto rappresentano una ripartizione dell'insieme dei dati secondo *due* grandezze di tipo qualitativo, ciascuna a sua volta ripartita in un gruppo di caratteristiche, o modalità. La metodologia può essere impiegata anche per l'analisi di matrici numeriche di altro tipo: specificamente per l'analisi di dati raccolti tramite inchieste, sondaggi e ricerche di mercato, campi questi ove l'Analisi delle Corrispondenze non ha rivali. Comunque, in tutti questi casi è spesso necessaria una preventiva ricodifica dei dati e occorrono opportuni adattamenti nell'interpretazione dei risultati. Un'unica metodologia è quindi in grado di risolvere problemi di tipo diverso e questo contrasta con quanto accade in Statistica, ove problemi specifici richiedono strumenti d'analisi specifici.

Fine dell'analisi è quello di spiegare perché la matrice dei dati si scosta da una situazione di *omogeneità* che si presenta quando le righe (o le colonne) sono proporzionali, portando alla luce l'intreccio di legami, le corrispondenze, tra le righe, tra le colonne e tra righe e colonne della matrice dei dati e perciò tra le diverse caratteristiche dell'insieme dei dati in esame. Questi legami non vengono espressi in forma numerica, ma in forma grafica, perché come scrive J. P. Benzécri, uno dei padri della metodologia, lo scopo *est de s'opposer*

à la traduction irréversible des choses en nombres. Conseguenza diretta di questo fatto è che questa metodologia richiede da parte dell'analista uno stile di analisi completamente nuovo.

Il *procedimento* che sta alla base del metodo consiste nel "geometrizzare il problema", nel senso che le righe e le colonne della matrice, opportunamente ricodificate, vengono intese come punti geometrici in due diversi spazi multidimensionali, nei quali è definita una distanza, dando vita quindi a due "nuvole" di punti. Per poterne decifrare la struttura, ciascuna nuvola viene proiettata in un sottospazio a due dimensioni: su un piano. Questi sono scelti in maniera ottimale, in modo tale che i punti proiettati diano una rappresentazione il più possibile fedele, della nube originaria. Grazie alle preventive trasformazioni operate simmetricamente sulle righe e sulle colonne della matrice dei dati, è possibile far coincidere i due piani, ottenendo così una "mappa" unica sulla quale le righe e le colonne della matrice vengono ad essere rappresentate dalle proiezioni dei loro punti rappresentativi. L'interpretazione delle prossimità tra proiezioni sulla mappa conduce l'analista a risalire alle prossimità tra punti delle nuvole nel loro spazio multidimensionale e perciò a riconoscere i legami tra le caratteristiche il cui l'insieme dei dati è ripartito.

In *conclusione*, l'Analisi delle Corrispondenze è un potente strumento per ripresentare i dati in modo grafico e comprensibile *senza ipotizzare modelli o strutture "a priori"*, ma in grado anche di mettere alla prova congetture e ipotetici modelli.

INDICE DEI CAPITOLI E DELLE SEZIONI

PARTE PRIMA: IL METODO

INDICE DELLE TAVOLE

PARTE PRIMA: IL METODO

CAPITOLO 1: Matrici e profili

Sommario

Questo primo capitolo indica come si possa procedere all'esame di una matrice di contingenza, senza fare ricorso all'Analisi delle Corrispondenze. L'obiettivo è quello di individuare eventuali associazioni tra le modalità delle variabili in gioco, cercando rassomiglianze e differenze tra righe e colonne della matrice. Si tratta di un procedimento piuttosto elementare che può giovarsi di un foglio elettronico, perché sostanzialmente si riduce a una comparazione di percentuali e al calcolo di medie e di indici.

Gli elementi di base presentati in questo capitolo porranno il lettore in grado di

- familiarizzarsi con le variabili categoriche e distinguere i diversi tipi delle loro modalità;
- rendersi conto di come si costruisce una matrice di contingenza e riconoscerla tra altri tipi di matrice;
- capire la nozione di profilo, e distinguere i profili delle modalità delle righe da quelli delle colonne della matrice;
- individuare, tramite un confronto tra profili, le associazioni tra modalità;
- comprendere il significato di profilo medio e di massa di un profilo;
- costruire degli indici per quantificare il grado di associazione tra modalità;
- rendersi conto dei limiti di questo tipo di analisi.

CAPITOLO 1

1.1 - Introduzione

La nostra attività quotidiana ci pone sovente sotto gli occhi dei numeri disposti in forma di tabella, ordinati cioè per righe e per colonne. Gli esempi sono molteplici: l'orario delle Ferrovie, le estrazioni del lotto sulle varie ruote, le tavole dei cambi delle valute e delle distanze chilometriche tra città, ecc. Quello che si va a cercare in queste tabelle è sempre *un* numero, o qualche numero: l'ora di arrivo o di partenza di un treno, i numeri estratti su una certa ruota, quanti yen occorrono per un dollaro, quanti chilometri ci sono tra due città, ecc. Non sempre però è un'informazione "puntuale" quella che interessa, molto spesso le relazioni tra numeri sono ben più interessanti dei singoli numeri.

Nella tabella *Spettacoli* di TAV. 2, i biglietti d'ingresso venduti nel nostro Paese nel 1991 per 8 tipi di spettacolo teatrale e musicale, sono ripartiti tra le 20 regioni italiane. Si tratta di una tabella, o in termini matematici, di una *matrice* con 20 righe e 8 colonne. Una matrice come questa viene costruita non solo per vedere, ad esempio, quanti biglietti si sono venduti per l'Operetta in Basilicata o per la Lirica nel Veneto, ma anche, e sopratutto, per confrontare la capacità delle regioni a "generare" spettatori ai diversi tipi di spettacolo, dal momento che ad ogni biglietto venduto è corrisposto uno spettatore pagante [1]. Allo stesso tempo la matrice permette di confrontare la capacità dei diversi tipi di spettacolo di attrarre pubblico nelle regioni, capacità che può risultare relativamente più alta in alcune regioni, piuttosto che in altre. Tutto ciò conduce quindi a cercare nella matrice quali relazioni vi siano tra regioni, tra tipi di spettacolo e tra regioni e tipi di spettacolo.

Questo può essere fatto comparando visivamente fra loro le righe della matrice (le regioni) o confrontandole singolarmente con una riga "media", e parimenti comparando le colonne (i tipi di spettacolo) tra loro o confrontandole con una colonna "media". Questo modo di procedere, praticabile per

[1] Nel numero di biglietti venduti sono compresi quelli a riduzione, ma esclusi gli ingressi gratuiti.

tabelle di ridotta dimensione, ma altrimenti lento e dispersivo, ha seri limiti ed è descritto in questo capitolo.

Un metodo particolarmente indicato per mettere in luce l'intreccio dei legami che intercorrono in matrici di questo tipo è l'Analisi delle Corrispondenze. La risposta che fornisce non è un indice numerico, dal momento che la struttura delle associazioni è di solito troppo complessa e articolata per essere ridotta a un numero, ma una o più mappe grafiche. Nella TAV. 37 è riportata la mappa principale ottenuta dall'analisi della matrice *Spettacoli* di TAV. 2. Le due rette ortogonali indicano gli assi di riferimento di un piano sul quale i punti che rappresentano i valori relativi delle righe (le regioni) sono messi in corrispondenza con quelli delle colonne (i tipi di spettacolo).

Non occorre essere uno statistico esperto per rendersi conto che le prossimità tra punti possono indicare, le associazioni tra le grandezze che rappresentano. Così raggruppamenti, opposizioni e tendenze appaiono evidenti sulla mappa, mentre sono spesso difficili da discernere nella matrice anche dopo prolungati confronti. Invece le relazioni tra punti colte sulla mappa possono essere facilmente riscontrate nella matrice. In questo senso l'Analisi delle Corrispondenze può essere visto come uno strumento grafico esplorativo della struttura delle relazioni nella matrice.

La prima parte di questo libro, a partire dal secondo capitolo, è dedicata a mostrare per quali vie da una matrice come quella di TAV. 2 si pervenga ad una mappa grafica e come questa vada interpretata. Ma prima di iniziare è necessario dare una precisa definizione alle grandezze che entreranno in gioco.

1.2 - Variabili categoriche

Il termine *variabile* è usato in Statistica per indicare una caratteristica o una proprietà che è possibile rilevare sugli elementi di un insieme. Gli elementi possono essere degli individui, degli animali, degli oggetti, degli eventi, ecc., tutti caratterizzati dal fatto di possedere la caratteristica o la proprietà che si vuole rilevare. Una variabile è detta *categorica*[1] quando è misurata con una *scala* costituita da un numero limitato (fino a qualche decina) di *modalità*, ossia di possibili stati esclusivi che la variabile può assumere. Per esempio la variabile *sesso* è misurata con una scala a due modalità: maschio e femmina; la *ripresa funzionale* di un arto dopo un intervento può essere misurata con una scala a tre modalità: nulla, parziale, completa. Esistono tre diversi tipi di variabili, differenziati in base al tipo di scala.

[1] Il nome deriva dall'inglese; modalità è detta *category*.

Quando le modalità della scala sono pure etichette, ossia dei semplici nomi, le variabili sono dette *nominali*. Sono variabili di tipo classificatorio, come il *sesso*, la *professione* esercitata, la *religione* d'appartenenza, la *marca* del prodotto, ecc. Tale è anche la variabile *mezzo di trasporto* cittadino, che può, ad esempio, essere catalogata con sei modalità: metrò, tram, bus, auto, scooter e bicicletta. Abitualmente le modalità vengono codificate con dei numeri perché sono meglio gestiti dal software d'analisi. Per esempio si può usare 1 per indicare il metrò, 2 per il tram, ecc. Si tratta comunque di semplici ricodifiche con codici convenzionali, privi di ogni caratteristica numerica, per cui l'ordine delle modalità resta irrilevante. Quando le modalità sono soltanto due, ad esempio sì/no, presente/assente, acceso/spento, ecc. la variabile nominale è chiamata *binaria* (o *logica*) e, di solito, l'equivalente codifica numerica è 1/0. Naturalmente, ogni operazione aritmetica su questi 'numeri' è priva di senso.

Esistono scale le cui modalità hanno invece un intrinseco ordine naturale di precedenza e le variabili così misurate sono dette *ordinali*. Per esempio la *durezza di un minerale* è classificata secondo i 10 livelli della scala di Mohs, la variabile *titolo di studio* è misurabile con cinque modalità: nessuno, licenza elementare, licenza media, diploma, laurea, che possono essere codificate con gli interi da 1 a 5 . I codici numerici che si scelgono sono comunque convenzionali. Le modalità della scala sono ordinate, ma le distanze tra modalità restano indefinite. Anche se si può dire che un laureato è più istruito di un diplomato non è possibile stabilire un numero che indichi *di quanto*.

Le variabili nominali e ordinali vengono indicate genericamente come *qualitative*, in opposizione alle variabili *quantitative*, dette anche *continue* o *numeriche*, per le quali, invece, è possibile quantificare la distanza tra le modalità della scala. Il numero di modalità può essere limitato, come per la variabile 'Frequenza di *shampoo* alla settimana', misurata con 7 modalità, per cui 6 shampoo alla settimana sono esattamente 2 volte 3 shampoo alla settimana. La variabile '*Livello di colesterolo* nel sangue' può invece assumere un numero potenzialmente infinito di valori, ma in pratica il potere risolutivo degli strumenti di misura e la precisione di quelli di elaborazione, riduce i possibili valori ad un numero finito, anche se elevato. Questo numero può essere drasticamente ridotto ricodificando i valori assunti in un numero limitato di intervalli esclusivi e contigui, detti *classi*, in base a una precisa regola che assegni ogni valore a una classe. Questa operazione comporta una perdita d'informazione, compensata però da alcuni vantaggi: facilità

d'interpretazione, adeguamento ad esigenze di ripartizione predefinite, uniformità della distribuzione risultante [1] . Così, ad esempio, l'*età* di un individuo, espressa in anni, può essere misurata con una scala ridotta a 4 modalità, individuate, ad esempio, da un codice numerico assegnato alle classi: $1 = [0, 21]$, $2 =]21, 40]$, $3 =]40, 65]$ e $4 =]65, 110]$.

Variabile *categorica* è il termine generico che si usa per indicare indifferentemente una variabile nominale, una ordinale oppure una continua ripartita in poche classi. L'Analisi delle Corrispondenze si può applicare a matrici ottenute incrociando le modalità di due o più variabili categoriche. Il metodo è comunque indifferente al *tipo* di variabile, perché le considera tutte come *nominali* ed è anche poco sensibile a come vengono stabiliti gli estremi delle classi (Sez. 2.8). Se poi tra le modalità di una variabile sussiste un ordine naturale, questo potrà eventualmente emergere dai risultati dell'analisi, ma non viene assunto "a priori".

1.3 - Matrici di contingenza

Quando gli elementi di un insieme vengono classificati in base alle modalità di *due* [2] variabili categoriche, i risultati possono essere presentati in una *matrice di contingenza*. Se I è il numero di diverse modalità esclusive della prima variabile, effettivamente utilizzate per classificare gli elementi, e J è quello della seconda variabile, la matrice è costruita computando, per ogni possibile coppia (i, j) di modalità – con $i = 1, 2, \ldots, I$ e con $j = 1, 2, \ldots, J$ – il numero di volte che un elemento è risultato possedere *congiuntamente* la modalità i della prima variabile e la modalità j della seconda. Questi conteggi sono detti *frequenze assolute* delle modalità (i, j). Le possibili combinazioni di modalità sono $I \times J$, per cui disponendo le $I \times J$ frequenze in una matrice con I righe e con J colonne si ottiene una matrice di contingenza. Tale è quindi la matrice *Spettacoli* di TAV. 2.

Spesso si confonde la frequenza con la quantità di elementi. Per esempio, ad un gruppo di scolari viene rilevato il *colore degli occhi*, classificato secondo le 4 modalità: 1 - azzurro, 2 - verde, 3 - nocciola e 4 - nero, e il *colore dei capelli*, classificato come: 1 - biondo, 2 - castano, 3 - bruno e 4 - fulvo. Nella matrice di contingenza 4×4 che si ottiene, all'incrocio della riga 3 con la colonna 2 si legge il numero di volte (frequenza assoluta) che

---

[1] Criteri d'identificazione del numero e delle ampiezze delle classi sono trattati nella Sez. 5.16.

[2] Il caso con più di due variabili è considerato nel Cap. 8.

si è rilevato uno scolaro con occhi nocciola *e* capelli castani, o anche quanti scolari del gruppo esaminato hanno occhi nocciola *e* capelli castani.

Le matrici di contingenza hanno caratteristiche peculiari. Sono sempre costituite da numeri interi mai negativi che non sono misure di un fenomeno, dipendenti da un'unità di misura, ma conteggi, semplici registrazioni di ciò che si è verificato. Le due variabili categoriche giocano nella matrice un ruolo *simmetrico* per cui gli elementi sono tutti omogenei, tutti della stessa natura: numero di individui, numero di oggetti, ecc. È così perfettamente lecito sommare gli elementi o accorpare righe e colonne: ciò che ne risulta ha ancora un significato.

L'Analisi delle Corrispondenze è particolarmente adatta allo studio di tabelle di contingenza, ma può essere applicata con successo anche a matrici di altro tipo: matrici di punteggi, di misure, di presenze/assenze, di distanze, ecc. In questi casi è spesso necessaria una ricodifica preliminare dei dati ed occorrono particolari adattamenti all'interpretazione dei risultati. A queste applicazioni è completamente dedicata la seconda parte di questo libro, a partire dal Capitolo 8.

L'esposizione generale della metodologia richiede un riferimento formale (TAV. 1), perciò da qui in avanti una matrice di contingenza verrà indicata col simbolo $\mathbf{N}$, mentre I indicherà il numero delle sue righe e J quello delle colonne. Quando $I = J$ la matrice è quadrata. Le righe verranno individuate non dalla loro modalità, come ad esempio da Piemonte, Basilicata, ecc. nella matrice *Spettacoli* di TAV. 2, perché queste variano da matrice a matrice, ma dal numero che indica il posto, o rango, che la riga, e quindi la corrispondente modalità, occupa a partire dall'alto e scendendo verso il basso. Una riga generica verrà indicata con i, per cui $i = 1, 2, \ldots, I$. Si parla così di riga numero i, di riga i^{ma} o, brevemente, di riga i. Allo stesso modo una colonna di $\mathbf{N}$ verrà individuata dal numero che indica il suo posto a partire da sinistra e procedendo verso destra. La colonna generica e la corrispondente modalità verrà indicata con j e così $j = 1, 2, \ldots, J$. Il numero che si trova all'incrocio della riga i con la colonna j indica la frequenza assoluta con cui, nell'insieme considerato, sono state rilevate congiuntamente le modalità i e j delle due variabili. Dal momento che questa frequenza dipende soltanto da i e da j verrà indicata con n_{ij}, dove il primo indice si riferisce sempre alla riga ed il secondo alla colonna. Il numero intero non negativo n_{ij} è detto *elemento* della matrice $\mathbf{N}$. Tutto questo vale per ogni riga $i = 1, 2, \ldots, I$ e per ogni colonna $j = 1, 2, \ldots, J$, per cui una matrice

di contingenza ha la forma

$$\mathbf{N} \stackrel{\text{def}}{=} \begin{pmatrix} n_{11} & n_{12} & \dots & n_{1j} & \dots & n_{1J} \\ n_{21} & n_{22} & \dots & n_{2j} & \dots & n_{2J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ n_{i1} & n_{i2} & \dots & n_{ij} & \dots & n_{iJ} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ n_{I1} & n_{I2} & \dots & n_{Ij} & \dots & n_{IJ} \end{pmatrix}.$$

1.4 - Totali marginali

Da una matrice di contingenza si possono ricavare altri dati di interesse per la comprensione del fenomeno che è stato rilevato: i totali marginali e il totale generale (TAV. 1).

Quando in ogni colonna si sommano gli elementi di tutte le righe, si ottiene la *riga marginale*. Il totale della j^{ma} colonna, indicato con n_{+j}, per mettere in rilievo che la somma è fatta sul primo indice, quello che indica la riga, è definito come

$$n_{+j} \stackrel{\text{def}}{=} \sum_{i=1}^{I} n_{ij} \qquad\qquad j = 1, 2, \dots, J.$$

Così nell'esempio della matrice *Spettacoli* di TAV. 2, l'indice della somma rappresenta via via le 20 regioni, ove allo spettacolo j hanno assistito n_{ij} spettatori. Ad esempio, per la prima colonna (Prosa) risulta che

$$n_{+1} = 639\,074 + 9\,219 + \ldots + 174\,260 = 10\,981\,793$$

spettatori hanno assistito a rappresentazioni di Prosa nel 1991. La riga marginale di questa matrice è una sequenza di $J = 8$ elementi

$$10\,981\,793 \quad 2\,867\,332 \quad \ldots \quad 312\,309 \quad 457\,542$$

che indicano quanti spettatori *in Italia*, hanno assistito nel 1991 a ciascuno degli 8 tipi di spettacolo. In altre parole, la riga marginale mostra come gli spettatori sono ripartiti tra i tipi di spettacolo teatrale e musicale, *a prescindere dalla regione*.

La *colonna marginale* si ottiene invece sommando, per ogni riga, gli elementi di tutte le colonne. Ne risulta una sequenza di I elementi, tanti quante sono le righe e il suo elemento generico è definito come

$$n_{i+} \stackrel{\text{def}}{=} \sum_{j=1}^{J} n_{ij} \qquad\qquad i = 1, 2, \dots, I.$$

Nell'esempio, la colonna marginale è una successione di $I = 20$ elementi, ciascuno dei quali indica quanti spettatori hanno assistito nella regione a degli spettacoli, *prescindendo dal tipo*: fornisce cioè la ripartizione degli spettatori per regione.

Ma quanti sono stati complessivamente gli spettatori ? Il loro numero si può ottenere sia sommando gli spettatori per regione e per tipo di spettacolo, ossia tutti gli elementi della matrice, oppure sommando tutti gli elementi della riga marginale o, anche, sommando quelli della colonna marginale. Questo numero, indicato con n_{++} per segnalare che la somma è fatta su entrambi gli indici, è detto *totale generale* ed è definito, in modo equivalente, come

$$n_{++} \stackrel{\text{def}}{=} \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = \sum_{j=1}^{J} n_{+j} = \sum_{i=1}^{I} n_{i+}.$$

Si viene così a conoscere che nel 1991 gli spettatori paganti che hanno assistito in Italia a spettacoli teatrali e musicali sono stati ben 26 196 957, quasi la metà della popolazione !

Il totale generale, la riga e la colonna marginale non diventano parte della matrice $\mathbf{N}$ che resta con I righe e J colonne. La riga e la colonna marginale di una matrice di contingenza sono sempre costituite da numeri interi *positivi*[1] . Questo fatto ha importanti implicazioni perché consentirà di definire i profili delle modalità (Sez. 1.5 e 1.7) e la distanza tra due profili (Sez. 2.8).

Nelle Sezioni che seguono verrà infatti mostrato come il confronto tra profili permetta di vedere se vi siano delle similarità tra regioni nella "generazione" di spettatori ai diversi spettacoli (Sez. 1.5), tra tipi di spettacolo nell'attrarre spettatori nell'ambito delle diverse regioni (Sez. 1.7) e infine tra regioni e tipi di spettacolo (Sez. 1.6, 1.8 e 1.9).

1.5 - Profili delle righe

Tornando alla la matrice *Spettacoli* di TAV. 2, si vede che in Piemonte ($i = 1$) gli spettatori paganti sono stati $n_{1+} = 1\,960\,491$, e di questi quelli che hanno assistito a rappresentazioni di Prosa ($j = 1$) sono stati

$$n_{Piemonte,\,Prosa} = n_{11} = 639\,074$$

---

[1] Una riga o una colonna di $\mathbf{N}$ tutta di zeri va rimossa perché indica soltanto l'assenza di una modalità.

pari alla quota

$$n_{Piemonte,\,Prosa}/n_{Piemonte,\,+} = n_{11}\,/\,n_{1+} = 639\,074\,/\,1\,960\,491 = 0.326.$$

Invece nella regione Marche ($i = 11$) gli spettatori sono stati complessivamente $n_{11,+} = 544\,992$, ossia quasi un terzo che nel Piemonte, e di questi hanno assistito a rappresentazioni di Prosa

$$n_{Marche,\,Prosa} = n_{11,1} = 193\,816$$

spettatori, pari alla quota

$$n_{Marche,\,Prosa}/n_{Marche,\,+} = n_{11,1}/n_{11,+} = 193\,816\,/\,544\,992 = 0.356.$$

In conclusione, anche se nelle Marche il numero complessivo di spettatori è circa un terzo di quello del Piemonte, in entrambe le regioni la Prosa ha la stessa quota di spettatori, ossia vi ha la stessa importanza. Il confronto tra regioni, ostacolato dal diverso numero complessivo di spettatori, può essere facilitato riconducendo le regioni ad avere tutte lo stesso numero di spettatori: uno solo. Queste considerazioni portano al concetto di *profilo* di una modalità [1] , e quindi a definire la matrice $\mathbf{R}$ dei profili delle righe di $\mathbf{N}$ di TAV. 3, ottenuta dividendo tutti i J elementi di ciascuna riga i per il loro totale n_{i+} , i^{ma} componente della colonna marginale,

$$\mathbf{R} \stackrel{\text{def}}{=} \begin{pmatrix} \frac{n_{11}}{n_{1+}} & \frac{n_{12}}{n_{1+}} & \dots & \frac{n_{1j}}{n_{1+}} & \dots & \frac{n_{1J}}{n_{1+}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{n_{i1}}{n_{i+}} & \frac{n_{i2}}{n_{i+}} & \dots & \frac{n_{ij}}{n_{i+}} & \dots & \frac{n_{iJ}}{n_{i+}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{n_{I1}}{n_{I+}} & \frac{n_{I2}}{n_{I+}} & \dots & \frac{n_{Ij}}{n_{I+}} & \dots & \frac{n_{IJ}}{n_{I+}} \end{pmatrix}. \qquad (1.5.1)$$

L'elemento generico di $\mathbf{R}$, per $i = 1, 2, \dots, I$ e per $j = 1, 2, \dots, J$, vale quindi

$$r_{ij} = \frac{n_{ij}}{n_{i+}}$$

ed è un numero reale non negativo che, nell'esempio, rappresenta la quota dello spettacolo j nella regione i . In Statistica prende il nome di *frequenza relativa di j condizionata da i* e viene tradizionalmente indicata con $f(j|i)$.

Dalla definizione di profilo consegue che se due righe di $\mathbf{N}$ sono proporzionali, i corrispondenti profili in $\mathbf{R}$ risultano *eguali*. Così il confronto fra

[1] Il concetto di profilo è più generale di quello di distribuzione, perché può estendersi anche a matrici che non sono di contingenza.

regioni, risulta facilitato perché è più semplice individuare delle uguaglianze tra le componenti dei profili in $\mathbf{R}$, che non delle proporzionalità tra quelle delle righe di $\mathbf{N}$. Mentre la riga marginale di $\mathbf{R}$ è priva di significato, la sua colonna marginale è costituita tutta da 1 perché

$$\sum_{j=1}^{J} r_{ij} = \sum_{j=1}^{J} \frac{n_{ij}}{n_{i+}} = \frac{n_{i+}}{n_{i+}} = 1 \qquad\qquad i = 1, 2, \ldots, I.$$

Con i profili è andata perduta la numerosità degli spettatori nelle diverse regioni, dal momento che tutte risultano averne ora uno solo, per cui, nota $\mathbf{R}$, soltanto l'ulteriore conoscenza della colonna marginale consente di risalire alla matrice di contingenza $\mathbf{N}$.

Il confronto tra profili nella TAV. 3 rivela una certa similarità tra le regioni Piemonte e Marche

	Tipi di Spettacolo							
Regioni	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Piemonte	0.326	0.105	0.157	0.018	0.047	0.296	0.011	0.040
Marche	0.356	0.100	0.141	0.020	0.024	0.326	0.007	0.026

nel senso che entrambe hanno quote piuttosto simili di spettatori agli 8 tipi di spettacolo, con netta preponderanza, un 30% circa, della Prosa ($j = 1$) e dei Concerti di Musica Leggera e Folkloristica ($j = 6$). Invece Valle d'Aosta e Sicilia appaiono in controtendenza

	Tipi di Spettacolo							
Regioni	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
V. d'Aosta	0.218	0.005	0.094	0.020	0.004	0.501	0.024	0.132
Sicilia	0.537	0.110	0.183	0.021	0.027	0.103	0.010	0.009

perché in Sicilia è preferita la Prosa ($j = 1$) e in Valle d'Aosta i Concerti di Musica Leggera e Folkloristica ($j = 6$), senza penalizzare i Saggi Coreografici e Folkloristici ($j = 8$). Tra queste due situazioni estreme, il confronto tra i profili delle regioni rivela tutta una sfumatura di situazioni che è lasciato al lettore di esplorare perché questo libro non si propone di presentare originali scoperte sullo spettacolo in Italia, ma soltanto di mostrare come sia possibile "far parlare" i dati raccolti.

Al contenuto numerico dei profili si può dare forma grafica mediante diagrammi a barre o a torta. Questi ultimi si possono impiegare quando le modalità di una delle due variabili sono poche, 5 o 6 al massimo. Negli I diagrammi circolari a torta, ciascuno con J "fette", l'area della "fetta" j è proporzionale a $r_{ij} = n_{ij}/n_{i+}$. Nella TAV. 6 i profili delle regioni sono rappresentati da diagrammi a barre. La barra rettangolare j della regione i ha altezza proporzionale a r_{ij}, mentre l'ampiezza, la stessa per tutte le barre, non è rilevante. Il confronto visivo dei diagrammi conferma la similarità dei profili di Piemonte e Marche e una difformità tra quelli di Valle d'Aosta e Sicilia.

1.6 - Profilo riga medio

La trasformazione in profili può essere ripetuta per la riga marginale, il cui totale è n_{++}, il totale generale. L'elemento generico j^{mo} del profilo marginale risulta quindi n_{+j}/n_{++}. Qual'è il significato di questo profilo ? La risposta risulta evidente se si scrive, per ogni colonna $j = 1, 2, \ldots, J$,

$$\frac{n_{+j}}{n_{++}} = \frac{\sum_{i=1}^{I} n_{ij}}{n_{++}} = \sum_{i=1}^{I} \frac{n_{i+}}{n_{++}} \frac{n_{ij}}{n_{i+}} = \sum_{i=1}^{I} \frac{n_{i+}}{n_{++}} r_{ij} = \frac{\sum_{i=1}^{I} \frac{n_{i+}}{n_{++}} r_{ij}}{\sum_{i=1}^{I} \frac{n_{i+}}{n_{++}}}$$

avendo tenuto conto al denominatore che la somma degli elementi della colonna marginale è eguale al totale generale

$$\sum_{i=1}^{I} \frac{n_{i+}}{n_{++}} = 1.$$

Dunque il profilo della riga marginale non è altro che la *media ponderata* dei profili delle righe, con peso n_{i+}/n_{++}, quota di spettatori nella regione i. I pesi, ottenuti dividendo la colonna marginale per il totale generale, non sono altro che le I componenti del profilo della colonna marginale e prendono il nome di *masse* dei profili riga.[1] Sono tutte *sempre positive* e la loro somma vale 1. La ponderazione dei profili è necessaria se si vuole tener conto del diverso numero complessivo di spettatori nelle 20 regioni.

Dal momento che il profilo della riga marginale è una media, viene chiamato *profilo riga medio*: è una regione fittizia, una sorta di 'media' nazionale. Come appare dalla TAV. 4, i suoi elementi sono le quote di spettatori *in Italia* in ciascuno degli 8 tipi di spettacolo. Questo profilo medio

[1] Nell'Analisi delle Corrispondenze si preferisce usare il termine *massa* piuttosto che *peso*. I due termini sono del tutto equivalenti.

è utile come riferimento perché il confronto tra un profilo e il profilo medio permette di indagare le relazioni tra regioni e tipi di spettacolo.

Se in una regione i risulta che per lo spettacolo j è $r_{ij} \simeq n_{+j}/n_{++}$, ossia che il rapporto $r_{ij} \,/\, (n_{+j}/n_{++}) \simeq 1$, significa che nella regione i la quota di spettatori allo spettacolo j non si distingue dalla quota media nazionale. Se il rapporto supera 1 significa che c'è stato un eccesso di spettatori, se risulta inferiore a 1 che c'è stata carenza di spettatori, sempre rispetto alla media nazionale. È preferibile, tuttavia, avere lo 0 come valore di riferimento, anziché 1. Si tratta di un semplice spostamento di scala che, per ogni spettacolo j, si ottiene sottraendo 1 al rapporto tra la quota di spettatori nella regione e la quota media nazionale. Il risultato ha anche un preciso significato perché

$$s_{ij} = \frac{r_{ij}}{\frac{n_{+j}}{n_{++}}} - 1 = \frac{r_{ij} - \frac{n_{+j}}{n_{++}}}{\frac{n_{+j}}{n_{++}}}$$

rappresenta lo *scarto relativo di* r_{ij} *dalla quota media* $\frac{n_{+j}}{n_{++}}$.

In conclusione, se in una regione i risulta che per lo spettacolo j è $s_{ij} \gg 0$ significa che la regione ha "generato" un numero di spettatori nettamente al di sopra della media nazionale: c'è stato un "surplus" di spettatori che hanno privilegiato lo spettacolo j. Si dice allora che tra regione i e spettacolo j c'è *attrazione*.

Quando invece $s_{ij} \ll 0$, la regione registra un "deficit" di spettatori allo spettacolo j. Vi è *repulsione* tra regione i e spettacolo j.

Se, infine, $s_{ij} \simeq 0$, tra i e j c'è *indifferenza*[1]: la quota di spettatori non si distingue dalla quota media nazionale e il rapporto $r_{ij}/(n_{+j}/n_{++})$, è vicino a 1.

Le situazioni di forte attrazione e di forte repulsione sono ovviamente quelle che gettano maggior luce sul fenomeno che si è rilevato. Così, ad esempio, dal confronto del profilo della Regione Valle d'Aosta ($i = 2$) col

---

[1] L'adozione dei termini *attrazione*, *indifferenza* e *repulsione* troverà giustificazione nella Sez. 4.13.

profilo medio (media nazionale),

	Tipi di Spettacolo								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Tot.*
Valle d'Aosta	0.218	0.005	0.094	0.020	0.004	0.501	0.024	0.132	*1.000*
media naz.	0.419	0.109	0.144	0.013	0.039	0.246	0.012	0.017	*1.000*
rapporto	*0.5*	*0.0*	*0.7*	*1.6*	*0.1*	*2.0*	*2.0*	*7.6*	
s_{2j}	-0.5	-1.0	-0.3	$+0.6$	-0.9	$+1.0$	$+1.0$	$+6.6$	
	rep.	*rep.*	*ind.*	*ind.*	*rep.*	*att.*	*att.*	*att.*	

balza evidente un fatto rimarchevole: la regione ha una quota di spettatori ai Saggi Coreografici e Folkloristici ($j = 8$) che supera di 7.6 volte la media nazionale, mentre risultano nettamente deficitari gli spettacoli di Lirica e Balletti ($j = 2$) e la Rivista e Commedia Musicale ($j = 5$).

In una situazione analoga si trova il Veneto ($i = 5$) per la Lirica ed i Balletti ($j = 2$), perché

	Tipi di Spettacolo								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Tot.*
Veneto	0.333	0.290	0.102	0.007	0.032	0.217	0.003	0.016	*1.000*
media naz.	0.419	0.109	0.144	0.013	0.039	0.246	0.012	0.017	*1.000*
rapporto	*0.8*	*2.7*	*0.7*	*0.5*	*0.8*	*0.9*	*0.2*	*0.9*	
s_{5j}	-0.2	$+1.7$	-0.3	-0.4	-0.2	-0.1	-0.8	-0.1	
	ind.	*att.*	*ind.*	*rep.*	*ind.*	*ind.*	*rep.*	*ind.*	

Qui $s_{52} = +1.7$, per cui la quota di spettatori a spettacoli di Lirica e Balletti ($j = 2$) risulta quasi tripla della media nazionale per questo tipo di spettacolo, mentre risultano trascurabili, sempre rispetto alla media nazionale, le quote dell'Operetta ($j = 4$) e sopratutto degli spettacoli di Burattini e Marionette ($j = 7$).

La TAV. 8 riporta la matrice **S** degli scarti relativi dalla media, per tutti i profili delle 20 regioni. Le considerazioni che si possono trarre dal suo esame sono lasciate al lettore.

Un fatto importante che va fin da ora sottolineato è che ogni elemento s_{ij} si limita ad indicare l'assenza o la presenza, più o meno marcata, di un'associazione tra modalità, ma non la sua importanza nell'ambito del fenomeno rilevato. Questo perché s_{ij} è basato unicamente sul confronto della *forma* dei profili, ma ignora le masse che tengono conto delle differenti affluenze di spettatori. Si vedrà invece nei prossimi capitoli, come l'Analisi

delle Corrispondenze sia in grado non solo di rilevare le associazioni, ma anche di dar loro la giusta importanza.

1.7 - Profili delle colonne

In modo simile e simmetrico a quanto fatto per le righe di $\mathbf{N}$, ogni colonna di $\mathbf{N}$ viene divisa per il suo totale. Viene così definita la matrice dei *profili delle colonne*

$$\mathbf{C} \stackrel{\text{def}}{=} \begin{pmatrix} \frac{n_{11}}{n_{+1}} & \frac{n_{12}}{n_{+2}} & \cdots & \frac{n_{1j}}{n_{+j}} & \cdots & \frac{n_{1J}}{n_{+J}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{n_{i1}}{n_{+1}} & \frac{n_{i2}}{n_{+2}} & \cdots & \frac{n_{ij}}{n_{+j}} & \cdots & \frac{n_{iJ}}{n_{+J}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{n_{I1}}{n_{+1}} & \frac{n_{I2}}{n_{+2}} & \cdots & \frac{n_{Ij}}{n_{+j}} & \cdots & \frac{n_{IJ}}{n_{+J}} \end{pmatrix}. \tag{1.7.1}$$

L'elemento generico

$$c_{ij} = \frac{n_{ij}}{n_{+j}}$$

è un numero reale non negativo che indica la quota della regione i nello spettacolo j. In Statistica è detta *frequenza relativa di i condizionata da j* e viene indicata con $f(i|j)$. La riga marginale di $\mathbf{C}$ è tutta costituita da 1 perché

$$\sum_{i=1}^{I} c_{ij} = \sum_{i=1}^{I} \frac{n_{ij}}{n_{+j}} = \frac{n_{+j}}{n_{+j}} = 1 \qquad\qquad j = 1, 2, \ldots, J.$$

Confrontando gli 8 profili colonna di TAV. 5, si nota, ad esempio, una certa similarità tra le rappresentazioni di Prosa ($j = 1$) e gli spettacoli di Commedie Musicali e di Rivista ($j = 5$), nel senso che entrambi hanno quote abbastanza simili di spettatori nelle 20 regioni.

Ad analoghe conclusioni si perviene esaminando la TAV. 7 ove i profili degli 8 tipi di spettacolo sono raffigurati da diagrammi a barre.

1.8 - Profilo colonna medio

Nella Sez. 1.6 si è visto che il profilo della colonna marginale di $\mathbf{N}$ è costituito dalle masse, ossia dai pesi, dei profili delle I righe. Ma che cosa rappresenta per i profili delle J colonne? Per ogni riga $i = 1, 2, \ldots, I$, le componenti del profilo sono

$$\frac{n_{i+}}{n_{++}} = \frac{\sum_{j=1}^{J} n_{ij}}{n_{++}} = \sum_{j=1}^{J} \frac{n_{+j}}{n_{++}} \frac{n_{ij}}{n_{+j}} = \sum_{j=1}^{J} \frac{n_{+j}}{n_{++}} c_{ij}.$$

Dunque il profilo della colonna marginale è il *profilo colonna medio*, media pesata dei profili colonna con peso, o massa, la corrispondente componente del profilo riga medio. Ogni massa è sempre positiva e la loro somma vale 1. La TAV. 5 riporta il profilo medio delle colonne della matrice *Spettacoli*.

Come conseguenza del fatto che su righe e colonne di $\mathbf{N}$ si è operato con trasformazioni dello stesso tipo, i profili della riga e della colonna marginale vengono ad avere un ruolo simmetrico. Quando si considera la matrice $\mathbf{R}$ dei profili delle righe, il profilo della *riga* marginale è il profilo riga medio, e il profilo della *colonna* marginale contiene le masse dei profili delle righe. Quando invece si considera la matrice $\mathbf{C}$ dei profili delle colonne, il profilo della *colonna* marginale è il profilo colonna medio, mentre il profilo della *riga* marginale contiene le masse dei profili delle colonne, come indicato in questo specchietto e come è facile verificare confrontando le TAV. 4 e 5.

	Contenuto del profilo	
Matrice	*della riga marginale*	*della col. marginale*
$\mathbf{R}$	profilo riga medio	masse dei profili riga
$\mathbf{C}$	masse dei profili colonna	profilo colonna medio.

Il profilo colonna medio rappresenta un tipo di spettacolo fittizio, desunto dal complesso degli 8 tipi di spettacolo che è utile come riferimento per evidenziare l'associazione, tra tipo di spettacolo e regione. Quando, per uno spettacolo j, risulta che nella regione i è

$$s_{ij} = \frac{c_{ij}}{\frac{n_{i+}}{n_{++}}} - 1 = \frac{c_{ij} - \frac{n_{i+}}{n_{++}}}{\frac{n_{i+}}{n_{++}}} \gg 0$$

significa che questo tipo di spettacolo ha nella regione i una quota di spettatori nettamente al di sopra della media degli 8 tipi di spettacolo: ha un "surplus" di spettatori. Tra spettacolo j e regione i c'è *attrazione*.

Quando invece $s_{ij} \ll 0$, lo spettacolo registra un "deficit" di spettatori nella regione i. Si parla di *repulsione* tra spettacolo j e regione i.

Se, infine, $s_{ij} \simeq 0$, tra j e i c'è *indifferenza*: la quota di spettatori non si distingue dalla media degli 8 tipi di spettacolo e il rapporto $c_{ij}/(n_{i+}/n_{++})$ è vicino a 1.

Questi scarti relativi dalla media sono stati indicati con s_{ij} perché coincidono con quelli ottenuti nella Sez. 1.6 dai rapporti tra profili delle

righe e profilo riga medio. Il fatto non deve sorprendere, perché, ad esempio, quando tra i e j c'è indifferenza

$$r_{ij} = \frac{n_{ij}}{n_{i+}} \simeq \frac{n_{+j}}{n_{++}} \tag{1.8.1}$$

e moltiplicando ambo i membri per n_{i+} e divedendo per n_{+j}, operazioni che lasciano inalterato il segno della diseguaglianza grazie alla positività delle componenti dei profili marginali, si ottiene

$$c_{ij} = \frac{n_{ij}}{n_{+j}} \simeq \frac{n_{i+}}{n_{++}} \tag{1.8.2}$$

per cui è indifferente effettuare il confronto tramite la (1.8.1) o la (1.8.2). Risultati analoghi si ottengono considerando una situazione di attrazione o di repulsione.

Come esempio si può verificare che per gli spettacoli di Lirica e Balletto ($j = 2$), s_{i2}, calcolato ora come $c_{i2}/(n_{i+}/n_{++}) - 1$, risulta identico a quello di TAV. 8 ottenuto dai profili delle righe

Regione	*Lirica*	*media*	*rapp.*	s_{i2}	
Piemonte	0.072	0.075	*1.0*	*0.0*	*ind.*
Valle Aosta	0.000	0.002	*0.0*	*−1.0*	*rep.*
Lombardia	0.143	0.178	*0.8*	*−0.2*	*ind.*
Trentino AA	0.011	0.026	*0.4*	*−0.6*	*rep.*
Veneto	0.239	0.090	*2.7*	*+1.7*	*att.*
...	...	...	...	...	...
Totale	*1.000*	*1.000*			

1.9 - Completa omogeneità

Nelle Sezioni precedenti si è visto che le associazioni tra le modalità i e j può essere rivelata o dal confronto tra l'elemento r_{ij} di $\mathbf{R}$ e il j^{mo} elemento del profilo riga medio oppure tra l'elemento c_{ij} di $\mathbf{C}$ e l'i^{mo} elemento del profilo colonna medio. Ci si può chiedere se allo stesso risultato si possa arrivare anche da un confronto che coinvolga l'elemento n_{ij} di $\mathbf{N}$, la matrice di contingenza. La risposta è affermativa perché moltiplicando ambo i membri della (1.8.1) per n_{i+} o della (1.8.2) per n_{+j}, risulta

$$n_{ij} \simeq \frac{n_{i+}\, n_{+j}}{n_{++}}. \tag{1.9.1}$$

Il verificarsi di questa situazione indica che l'associazione (i, j) tra le modalità delle due variabili è stata rilevata altrettanto frequentemente di quanto

fosse da attendersi in base ai dati complessivi. Il confronto è ora tra un numero intero ed uno reale che rappresenta un livello neutro di riferimento, una situazione di perfetta indifferenza tra i e j, dal momento che è ottenuto esclusivamente dalle masse n_{i+} di i e n_{+j} di j, senza che intervenga alcuna affinità particolare tra i e j. Questa situazione di perfetta indifferenza è detta di *omogeneità* tra i e j, quando tutti i profili, delle righe e delle colonne, risultano eguali ai loro profili medi.

Riprendendo l'esempio e considerando il Veneto ($i = 5$) e la Lirica e Balletti ($j = 2$), dalla TAV.2 si ottiene

$$n_{Veneto,\,Lirica} = n_{52} = 686\,236$$

$$\frac{n_{5+}\, n_{+2}}{n_{++}} = \frac{2\,365\,382 \times 2\,867\,332}{26\,196\,957} = 258\,897.836$$

Il rapporto tra la frequenza rilevata e quella che si avrebbe in condizioni di omogeneità, vale

$$\frac{n_{52}}{\frac{n_{5+}\, n_{+2}}{n_{++}}} = \frac{686\,236}{256\,897.836} = 2.7 \qquad \text{e quindi} \qquad s_{52} = 1.7$$

che è lo stesso valore trovato precedentemente nella Sez. 1.6 tramite i profili riga e nella Sez. 1.8 tramite i profili colonna. Perciò s_{ij} è anche una misura della distanza da una situazione di completa omogeneità.

1.10 - Riepilogo

Con l'introduzione delle matrici $\mathbf{R}$ e $\mathbf{C}$ dei profili, il problema di individuare le affinità tra le modalità delle righe e tra le modalità delle colonne, è stato ricondotto a un confronto visivo tra profili riga e tra profili colonna.

Inoltre, la costruzione della matrice degli scarti relativi dalla media, che per ogni riga $i = 1, 2, \ldots, I$ e per ogni colonna $j = 1, 2, \ldots, J$ indica

$$s_{ij} = \frac{r_{ij}}{\frac{n_{+j}}{n_{++}}} - 1 = \frac{c_{ij}}{\frac{n_{i+}}{n_{++}}} - 1 = \frac{n_{ij}}{\frac{n_{i+}\, n_{+j}}{n_{++}}} - 1 \begin{cases} \ll 0 & \text{repulsione} \\ \simeq 0 & \text{indifferenza} \\ \gg 0 & \text{attrazione} \end{cases}$$

consente anche di valutare le associazioni tra modalità delle righe e delle colonne. Queste tre matrici, $\mathbf{R}$, $\mathbf{C}$ e $\mathbf{S}$, sono tutte dello stesso ordine $I \times J$ della matrice di contingenza $\mathbf{N}$, per cui alla fin fine il problema di individuare la struttura dei dati scrutando queste matrici invece che $\mathbf{N}$, risulta semplificato solo in parte. In più, un confronto basato unicamente sui profili e che ne ignora le masse, non è in grado di stabilire l'importanza di un'eventuale associazione.

Un metodo alternativo, per mettere in luce le associazioni tra modalità e dar loro il giusto peso, è l'Analisi delle Corrispondenze. L'idea fondamentale di base è quella di scomporre gli scarti relativi dalla media nella forma

$$s_{ij} = \frac{r_{ij}}{\frac{n_{+j}}{n_{++}}} - 1 = \frac{c_{ij}}{\frac{n_{i+}}{n_{++}}} - 1 = \frac{n_{ij}}{\frac{n_{i+}\,n_{+j}}{n_{++}}} - 1 = \sum_{a=1}^{A} \frac{1}{\sqrt{\lambda_a}}\, f_{ia}\, g_{ja} \qquad (1.10.1)$$

che vale per ogni riga $i = 1, 2, \ldots, I$ e per ogni colonna $j = 1, 2, \ldots, J$ e dove $A = min(I, J) - 1$. Quando la sommatoria è nulla gli scarti relativi dalla media sono nulli: è la situazione di omogeneità, o di perfetta indifferenza, tra le modalità, situazione che risulta ovviamente di scarso interesse per capire il fenomeno rilevato. La sommatoria esprime invece le eventuali associazioni tra i e j. In ogni termine, λ_a è un numero reale compreso tra 0 e 1 e tale che $\lambda_a > \lambda_{a+1}$, mentre f_{ia} e g_{ja} sono due numeri reali, detti *fattori* di ordine a, legati rispettivamente al profilo i delle modalità delle righe e j delle colonne.

Ciascun termine della somma specifica un'associazione tra le modalità i e j: di attrazione quando il termine è positivo, perché eleva il rapporto sopra l'unità; di repulsione quando è negativo, ossia quando f_{ia} e g_{ja} sono di segno opposto; di indifferenza quando almeno uno dei due fattori è trascurabile. Delle associazioni che intercorrono tra le modalità, l'Analisi delle Corrispondenze fornisce dunque un quadro molto più sfumato e realistico del procedimento di confronto che è stato illustrato in questo capitolo.

Inoltre i termini della somma decrescono sostanzialmente come $\sqrt{\lambda_a}$, sicché molto spesso questa può essere limitata ai primi A^* termini, spesso due o tre. La rappresentazione dei profili tramite i fattori porta così a una utile semplificazione ("data reduction") perché sostituisce gli $I \times J$ numeri di $\mathbf{R}$, di $\mathbf{C}$ o di $\mathbf{S}$, con $A^* \times (I + J)$ numeri dovuti agli I valori di f_{ia} e ai J valori di g_{ja}.

L'aspetto più rilevante dell'Analisi delle Corrispondenze è la sua capacità di rappresentare congiuntamente i profili delle righe e delle colonne su mappe grafiche: i fattori essendo le coordinate dei profili. Così, invece di leggere direttamente la matrice $\mathbf{S}$ degli scarti relativi dal livello medio nella TAV. 8, la si legge tramite le risultanze grafiche, ordinate per importanza decrescente, che mettono in luce via via le associazioni più rimarchevoli.

Come l'Analisi delle Corrispondenze pervenga alla scomposizione (1. 10.1) e, sopratutto, alla sua rappresentazione grafica, è mostrato nei tre prossimi capitoli.

1.11 - Bibliografia essenziale

Per appprofondire gli argomenti trattati, al termine di ogni capitolo il lettore troverà uno smilzo elenco di riferimenti bibliografici. L'elenco, che non è né completo né sistematico, è limitato ai testi ed agli articoli più recenti e facilmemente reperibili. Molti di essi contengono bibliografie dettagliate. La soggettività della scelta è stata mitigata consultando alcuni colleghi. Beninteso, *choisir n'est pas exclure !*

Per questo primo Capitolo il lettore può consultare

Francesco Della Beffa (1992). *Come fare analisi descrittive*. Franco Angeli ed., Milano. 201 pg. ISBN n.i., ove l'esposizione chiara ed accessibile spazia dalla costruzione di tabelle di contingenza a partire dai dati, alla derivazione dei profili, al loro confronto, ai test per la significatività dei risultati. Gli esempi sono molto numerosi e fanno riferimento a casi reali nell'ambito del Marketing e delle Ricerche di Mercato.

Ai molteplici aspetti dell'analisi statistica descrittiva è interamente dedicato il ponderoso testo di

Giuseppe Leti (1983). *Statistica descrittiva*. il Mulino ed., Bologna. 941 pg. ISBN 88-15-00278-2, che è un testo di riferimento in questo campo: una "summa" organica dei concetti e dei metodi.

All'analisi statistica di dati strutturati in due o più dimensioni si rivolge il testo di

Angelo Zanella (1988). *Lezioni di Statistica. Parte seconda: strutture dei dati in due o più dimensioni*. Vita e Pensiero ed., Milano. 364 pg. ISBN 88-343-8656-6, la cui ultima parte esemplifica con casi concreti l'analisi strutturale di una matrice di dati.

PARTE PRIMA: IL METODO

CAPITOLO 2: Spazi e nuvole di profili

Sommario

Questo secondo capitolo fornisce gli elementi matematici di base, indispensabili per comprendere i fondamenti dell'Analisi delle Corrispondenze. La sua lettura non è quindi facoltativa, ma il lettore non deve lasciarsi intimidire perché molti dei concetti esposti hanno una interpretazione geometrica semplice e intuitiva. In più, numerosi esempi cercano di rendere accessibili gli argomenti presentati anche ai meno versati in matematica. L'intento è di far capire i concetti, piuttosto che fornire dimostrazioni rigorose, reperibili nei testi citati in bibliografia.

Dalla lettura di questo capitolo il lettore potrà

- acquisire il concetto di vettore, inteso come entità matematica, suscettibile di un'interpretazione geometrica;
- familiarizzarsi con le operazioni tra vettori: prodotto scalare, lunghezza, proiezione e distanza tra vettori;
- rendersi conto di come la nozione di vettore sia legata al concetto di spazio euclideo multi-dimensionale;
- vedere come un profilo sia un vettore con caratteristiche peculiari;
- immaginare i profili come una nuvola di punti dotati di massa, immersi in uno spazio euclideo;
- comprendere la nozione di distanza distribuzionale tra punti-profilo e rendersi conto dei motivi per cui viene adottata.

CAPITOLO 2

2.1 - La matrice d'esempio Spettacoli-3

L'Analisi delle Corrispondenze rivela tutte le sue potenzialità quando la matrice da analizzare è di grandi dimensioni, come nel caso, della matrice *Spettacoli* in TAV. 2 che è di ordine 20×8. Ma per meglio esemplificare le basi matematiche dell'Analisi delle Corrispondenze è conveniente riferirsi a una matrice di dimensioni più ridotte. Perciò dalla matrice *Spettacoli*, accorpando le righe delle regioni dell'Italia settentrionale (1 = Nord), dell'Italia centrale (2 = Centro) e dell'Italia meridionale ed insulare (3 = Sud) è stata ottenuta una matrice 3×8 i cui elementi sono stati poi divisi per $10\,000$, per ridurre il numero di cifre, e quindi arrotondati all'intero più prossimo. Il risultato è la matrice di contingenza *Spettacoli-3* di TAV. 14 qui riprodotta

$$\mathbf{N} = \begin{pmatrix} 576 & 175 & 198 & 21 & 55 & 370 & 14 & 28 \\ 269 & 63 & 88 & 5 & 30 & 145 & 11 & 7 \\ 254 & 49 & 90 & 7 & 18 & 131 & 6 & 10 \end{pmatrix}.$$

In essa, ad esempio, il primo elemento $n_{11} = 576$ indica che 576 decine di migliaia (circa) di biglietti sono state vendute nell'Italia Settentrionale ($i = 1$) per rappresentazioni di Prosa ($j = 1$) nel 1991, ovvero che 576 decine di migliaia (circa) di spettatori paganti hanno assistito nel Nord a rappresentazioni di Prosa nel 1991.

La matrice *Spettacoli-3* verrà impiegata d'ora innanzi per esemplificare l'Analisi delle Corrispondenze di matrici di contingenza aventi meno righe che colonne, ossia con $I \leq J$.

2.2 - Vettori

La matrice d'esempio ha 3 righe e 8 colonne. Si possono però immaginare delle matrici costituite da una sola colonna, per esempio dalla prima di $\mathbf{N}$: la matrice si riduce allora ad un *vettore colonna* di ordine 3×1 o, più brevemente, ad un vettore di ordine 3 che si indica abitualmente con una

lettera minuscola in grassetto, e con gli *elementi* racchiusi tra parentesi

$$\mathbf{x} = \begin{pmatrix} 576 \\ 269 \\ 254 \end{pmatrix}.$$

Analogamente una matrice costituita da una sola riga è detta *vettore riga*. Così la prima riga di $\mathbf{N}$ è un vettore riga di ordine 8

$$\mathbf{y}^T = (576 \;\; 175 \;\; 198 \;\; 21 \;\; 55 \;\; 370 \;\; 14 \;\; 28).$$

La notazione ^{T} indica l'operazione di trasposizione (APP. A). Per vettore *trasposto* si intende che un vettore colonna di ordine $n \times 1$ diventa un vettore riga di ordine $1 \times n$ e viceversa. Nel seguito col termine *vettore* si farà esclusivo riferimento a vettori colonna, per cui

$$\mathbf{x} = \begin{pmatrix} 576 \\ 269 \\ 254 \end{pmatrix} \qquad \text{oppure} \qquad \mathbf{x} = (576 \;\; 269 \;\; 254)^T$$

sono due modi equivalenti di indicare lo stesso vettore colonna, il secondo spesso usato per risparmiare spazio e carta. Invece

$$\mathbf{x}^T = (576 \;\; 269 \;\; 254) \quad \text{oppure} \quad \mathbf{y}^T = (576 \;\; 175 \;\; 198 \;\; 21 \;\; 55 \;\; 370 \;\; 14 \;\; 28)$$

sono due vettori riga: il primo ottenuto trasponendo $\mathbf{x}$, il secondo dalla prima riga di $\mathbf{N}$. In pratica: quando accanto al simbolo di vettore (ad esempio $\mathbf{z}$) non compare ^{T} si tratta di un vettore colonna, quando invece compare ($\mathbf{z}^T$) si tratta di un vettore riga.

Una matrice ridotta a un solo elemento è detta *scalare*. In questo libro scalare e numero reale sono considerati sinonimi.

Alcuni vettori avranno un ruolo importante nella trattazione che seguirà. Sono il vettore *zero* o *nullo* $\mathbf{0}_n$ con gli n elementi tutti nulli[1] e gli n vettori *unità* $\mathbf{e}_k$ con $k = 1, 2, \ldots, n$ i cui n elementi sono tutti nulli ad eccezione del k^{mo} che vale 1. Questi ultimi corrispondono alle n colonne della matrice identità $\mathbf{I}$ di ordine $n \times n$ che ha gli elementi della diagonale principale eguali ad 1 e tutti gli altri nulli (APP. A). Per esempio, per vettori di ordine 3 si ha

$$\mathbf{0}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \qquad \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

---

[1] L'indice apposto ai vettori $\mathbf{0}$ e $\mathbf{1}$ ne indica l'ordine. Così $\mathbf{1}_3 = (1 \; 1 \; 1)^T$.

2.3 - Spazi euclidei

C'è una profonda connessione tra l'algebra dei vettori [1] e la geometria. Connessione che risulta utile allo studio di entrambe, nel senso che idee provenienti da una delle due aree possono suscitare utili intuizioni nel campo dell'altra. L'idea di base è quella di rappresentare con un vettore la posizione di un punto relativamente alla posizione di un altro, una volta assegnati gli assi di riferimento, ossia delle rette orientate sulle quali è fissata l'origine e l'unità di misura delle coordinate, come in TAV. 9. Conviene scegliere un sistema di riferimento cartesiano ortogonale e far coincidere l'origine dei vettori con l'origine degli assi, corrispondente al vettore nullo $\mathbf{0}_n$. In tal modo viene a stabilirsi una corrispondenza uno-a-uno tra vettori e punti, come si vede in TAV. 10. Così un vettore di ordine 3 come $\mathbf{x} = (576 \quad 269 \quad 254)^T$ può essere pensato nel nostro spazio fisico come un *punto* X. Gli *elementi* di $\mathbf{x}$, legati alle *coordinate* o ascisse di X sui tre assi del sistema di riferimento, vengono detti *componenti* di $\mathbf{x}$. Dal punto di vista geometrico è comodo talvolta far corrispondere al vettore $\mathbf{x}$ il segmento orientato $\overrightarrow{OX}$ che ha inizio nell'origine e l'altra estremità nel punto X, anziché il semplice punto X sopra definito.

Si tende abitualmente ad identificare i concetti di vettore e di rappresentazione geometrica del vettore: si parla così di "punto $\mathbf{x}$" anziché di punto X corrispondente al vettore $\mathbf{x}$, di "coordinate" anziché di componenti o elementi, di "lunghezza del vettore" anziché di lunghezza del segmento $\overrightarrow{OX}$ corrispondente a $\mathbf{x}$, ecc.

In generale un vettore di ordine n, con origine in $\mathbf{0}_n$, individua un punto in uno spazio n-dimensionale, per cui a ogni vettore risulta associato *univocamente* un punto.

Come tutti i vettori di ordine n individuano tutti i punti dello spazio n-dimensionale, così, reciprocamente, si può pensare allo *spazio* n-dimensionale come all'insieme di *tutti* i vettori di ordine n e, dal momento che questo è una naturale estensione dello spazio uni-dimensionale (la retta), bi-dimensionale (il piano) e tri-dimensionale (lo spazio in cui viviamo), viene chiamato *spazio Euclideo* n-dimensionale ed indicato con $\Re^n$. Perciò $\Re^3$ indica il nostro spazio fisico.

Più precisamente, se $\mathbf{x}$ e $\mathbf{y}$ sono due generici vettori di ordine n

---

[1] Definizioni rigorose dei concetti di base presentati in questo capitolo si possono trovare nei testi citati nella Sez. 2.12.

dello spazio euclideo $\Re^n$, allora per *ogni* coppia $\mathbf{x}$ e $\mathbf{y}$, il vettore ottenuto dalla loro somma

$$\mathbf{x} + \mathbf{y}$$

è anch'esso un vettore di $\Re^n$. Inoltre se a è un numero reale, anche il vettore

$$a\,\mathbf{x}$$

appartiene ancora a $\Re^n$. Queste operazioni algebriche su vettori (APP. A) hanno una corrispondente interpretazione geometrica, come mostrato nella TAV. 11. Così se $\mathbf{x}$ e $\mathbf{y}$ sono intesi come due segmenti orientati di $\Re^n$, la regola del parallelogramma delle forze permette di individuare il punto che corrisponde a $\mathbf{x}+\mathbf{y}$. La moltiplicazione di un vettore per un numero $a > 1$, è una amplificazione delle componenti del vettore e quindi un allontanamento del punto $\mathbf{x}$ dall'origine, ma senza cambiare direzione. Quando $0 < a < 1$ si ha invece una compressione e il punto si avvicina all'origine. Se $a = 0$ il vettore risultante è $\mathbf{0}_n$, il vettore nullo, e il nuovo punto viene a coicidere con l'origine, mentre se $a < 0$ il punto $a\,\mathbf{x}$ risulta opposto, rispetto all'origine, al punto originale $\mathbf{x}$.

Le operazioni di somma di due vettori e di moltiplicazione per uno numero reale consentono di definire il concetto di *combinazione lineare* che, a sua volta porta a definire la *dimensione* di uno spazio euclideo e ad individuare il suo *sistema di riferimento.*

2.4 - Vettori di base e dimensione

Tutti noi siamo consci dell'esistenza di una dimensione spaziale nel nostro spazio. Così lo spazio fisico che percepiamo, si pensi alla stanza in cui ci troviamo, è tridimensionale. La superficie di una parete la percepiamo bidimensionale e la linea d'intersezione tra parete e soffitto come unidimensionale. Il concetto di dimensione di uno spazio è però molto più generale e per arrivare a comprenderlo è opportuno partire dalla familiare equazione di una retta.

Quando x e y rappresentano i due assi cartesiani ortogonali di un piano, l'equazione di una retta si scrive $ax + by = c$, con a, b e c costanti. L'idea di legame lineare può estendersi ai vettori, per cui $a\mathbf{x}+b\mathbf{y}$ è chiamata *combinazione lineare* dei vettori $\mathbf{x}$ e $\mathbf{y}$. Più in generale, disponendo di n vettori $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ dello stesso ordine e di n numeri $a_1, a_2, \ldots, a_n$, il vettore che risulta dalla somma di multipli dei vettori

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_n\mathbf{x}_n$$

è detto *combinazione lineare* degli n vettori. Questo porta alla definizione di vettori *linearmente indipendenti* quando $a_1 = a_2 = \ldots = a_n = 0$ è la *sola* condizione perché il vettore risultante sia il vettore $\mathbf{0}_n$

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_n\mathbf{x}_n = \mathbf{0}_n$$

purché, beninteso, nessuno dei vettori $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ sia nullo. L'indipendenza non è quindi una caratteristica del singolo vettore, ma dell'insieme dei vettori considerati e quello che conta non sono i valori che assumono i coefficienti a_i, ma se questi sono tutti nulli o meno. Per esempio i tre vettori di $\Re^3$

$$\mathbf{x}_1 = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{x}_2 = \begin{pmatrix} 0 \\ 3 \\ 0 \end{pmatrix} \quad \mathbf{x}_3 = \begin{pmatrix} 0 \\ 0 \\ 5 \end{pmatrix} \tag{2.4.1}$$

sono linearmente indipendenti perché se deve essere

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + a_3\mathbf{x}_3 = \mathbf{0}_3$$

$$a_1 \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 3 \\ 0 \end{pmatrix} + a_3 \begin{pmatrix} 0 \\ 0 \\ 5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

ne risulta che

$$\begin{pmatrix} 2a_1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 3a_2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 5a_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{da cui} \quad \begin{pmatrix} 2\,a_1 \\ 3\,a_2 \\ 5\,a_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

e quindi i tre coefficienti della combinazione lineare risultano tutti nulli

$$a_1 = a_2 = a_3 = 0.$$

I tre vettori linearmente indipendenti (2.4.1) sono detti *vettori di base* e costituiscono *una base* per $\Re^3$.

Ci sono tre punti importanti da sottolineare. Primo, una base di $\Re^3$ è sempre costituita da *terne* di vettori di ordine 3 . Questo significa che aggiungendo alla terna di base un qualunque altro vettore dello stesso ordine, si ottiene un insieme di (quattro) vettori linearmente dipendenti. Per esempio aggiungendo alla terna (2.4.1) il vettore ottenuto dalla prima colonna di $\mathbf{N}$ si hanno i quattro vettori

$$\mathbf{x}_1 = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{x}_2 = \begin{pmatrix} 0 \\ 3 \\ 0 \end{pmatrix} \quad \mathbf{x}_3 = \begin{pmatrix} 0 \\ 0 \\ 5 \end{pmatrix} \quad \text{e} \quad \mathbf{x} = \begin{pmatrix} 576 \\ 269 \\ 254 \end{pmatrix}$$

che risultano linearmente dipendenti perché i coefficienti della loro combinazione lineare non sono tutti nulli. Infatti se $a_1 = 576/2$, $a_2 = 269/3$, $a_3 = 254/5$, allora $\mathbf{x}$ è ottenibile come combinazione lineare dei tre vettori di base

$$\frac{576}{2}\mathbf{x}_1 + \frac{269}{3}\mathbf{x}_2 + \frac{254}{5}\mathbf{x}_3 = \mathbf{x}.$$

Più in generale, quattro o più vettori nello spazio $\Re^3$ sono *sempre* linearmente dipendenti.

Secondo, di basi se ne possono scegliere diverse in $\Re^3$. Per esempio i tre vettori unità definiti nella Sez. 2.2

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{2.4.2}$$

sono linearmente indipendenti e costituiscono quindi una altra base per $\Re^3$, detta *base canonica*, ossia conforme ai canoni, alle regole.

Terzo, *ogni* altro vettore di ordine 3 può essere rappresentato, in un *unico* modo, come combinazione lineare dei tre vettori di base prescelti. Quindi una base può immaginarsi come un riferimento per i vettori, non diverso da un sistema di coordinate per i punti. Così rispetto alla base (2.4.1) il vettore $\mathbf{x} = (576 \;\; 269 \;\; 254)^T$ viene rappresentato come

$$\begin{aligned} \mathbf{x} &= a_1\,\mathbf{x}_1 + a_2\,\mathbf{x}_2 + a_3\,\mathbf{x}_3 \\ \begin{pmatrix} 576 \\ 269 \\ 254 \end{pmatrix} &= \frac{576}{2}\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} + \frac{269}{3}\begin{pmatrix} 0 \\ 3 \\ 0 \end{pmatrix} + \frac{254}{5}\begin{pmatrix} 0 \\ 0 \\ 5 \end{pmatrix}. \end{aligned}$$

Questo mostra che potendo scegliere, è preferibile utilizzare la base canonica (2.4.2) costituita dalla terna di vettori unità, perché in tal caso l'espressione del vettore $\mathbf{x}$ risulta semplificata: i coefficienti della combinazione lineare sono proprio le componenti del vettore, e coincidono con le coordinate del punto che rappresenta

$$\begin{aligned} \mathbf{x} &= a_1\,\mathbf{e}_1 + a_2\,\mathbf{e}_2 + a_3\,\mathbf{e}_3 \\ &= 576\,\mathbf{e}_1 + 269\,\mathbf{e}_2 + 254\,\mathbf{e}_3 \\ \begin{pmatrix} 576 \\ 269 \\ 254 \end{pmatrix} &= 576\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 269\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 254\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \end{aligned}$$

Questa base privilegiata viene detta *canonica* o *standard*. Rispetto ad essa *ogni* vettore $\mathbf{z} = (a \;\; b \;\; c)^T$ può esprimersi in modo univoco come somma di

multipli dei vettori di base

$$\mathbf{z} = a\,\mathbf{e}_1 + b\,\mathbf{e}_2 + c\,\mathbf{e}_3.$$

Geometricamente $\mathbf{z}$ è il punto ottenuto spostandosi di a unità su $\mathbf{e}_1$, di b su $\mathbf{e}_2$ e di c su $\mathbf{e}_3$, come mostrato in TAV. 12.

Le considerazioni fatte finora possono essere estrapolate allo spazio $\Re^n$, per cui ogni vettore di ordine n può essere espresso come combinazione lineare di un qualunque insieme di n vettori indipendenti di ordine n. Il numero massimo di vettori indipendenti non nulli è quindi n, l'ordine dei vettori. Questo numero è detto *dimensione* dello spazio $\Re^n$.

Uno spazio euclideo è necessariamente dotato di una struttura *metrica* che lo caratterizza rendendolo unico per una specificata dimensionalità e ne arricchisce la geometria, permettendo di definire distanze, lunghezze ed angoli.

2.5 - Prodotto scalare e distanza

Quattro amici cenano in pizzeria: due di loro bevono birra e gli altri due cocacola. Una birra costa 4 000 lire, una coca 2 000 e una pizza 5 000, per cui, alla fine, il conto risulta

$$2 \times 4\,000 + 2 \times 2\,000 + 4 \times 5\,000 = 32\,000 \quad \text{lire.}$$

Se ora si indica con $\mathbf{c} = (2 \;\; 2 \;\; 4)^T$ il vettore delle consumazioni e con $\mathbf{p} = (4\,000 \;\; 2\,000 \;\; 5\,000)^T$ quello dei prezzi unitari, allora (APP. A)

$$\begin{aligned}\mathbf{c}^T\mathbf{p} &= (2 \;\; 2 \;\; 4) \begin{pmatrix} 4\,000 \\ 2\,000 \\ 5\,000 \end{pmatrix} \\ &= 2 \times 4\,000 + 2 \times 2\,000 + 4 \times 5000 = 32\,000.\end{aligned}$$

Questo esempio illustra il procedimento per ottenere $\mathbf{c}^T\mathbf{p}$: moltiplicare ogni elemento del primo vettore per il corrispondente elemento del secondo. La loro somma fornisce $\mathbf{c}^T\mathbf{p}$. Così, dati due vettori dello stesso ordine $\mathbf{x} = (x_1 \;\; x_2 \ldots x_n)^T$ e $\mathbf{y} = (y_1 \;\; y_2 \ldots y_n)^T$ il loro prodotto $\mathbf{x}^T\mathbf{y}$ è definito come

$$\mathbf{x}^T\mathbf{y} \stackrel{\text{def}}{=} x_1y_1 + x_2y_2 + \ldots + x_ny_n = \sum_{k=1}^{n} x_k\, y_k \tag{2.5.1}$$

ed è chiamato *prodotto scalare* dei vettori $\mathbf{x}$ e $\mathbf{y}$. Dal momento che il risultato è un numero, ne deriva che i due vettori si possono scambiare, perché

$$\mathbf{x}^T\mathbf{y} = (\mathbf{x}^T\mathbf{y})^T = \mathbf{y}^T\mathbf{x}.$$

Difatti, l'esempio della cena in pizzeria, oltre a rivelarci che i camerieri sono grandi esperti di prodotto scalare, mostra che il conto poteva essere calcolato alternativamente come

$$\mathbf{c}^T\mathbf{p} = (2\ \ 2\ \ 4)\begin{pmatrix}4\,000\\2\,000\\5\,000\end{pmatrix} = \mathbf{p}^T\mathbf{c} = (4\,000\ \ 2\,000\ \ 5\,000)\begin{pmatrix}2\\2\\4\end{pmatrix} = 32\,000.$$

Quando il prodotto scalare di due vettori è nullo, i due vettori sono detti *ortogonali* o *perpendicolari* tra loro. L'ortogonalità è quindi una caratteristica della coppia di vettori. Per esempio i tre vettori unità risultano mutuamente ortogonali. Infatti

$$\mathbf{e}_1^T\mathbf{e}_2 = (1\ \ 0\ \ 0)\begin{pmatrix}0\\1\\0\end{pmatrix} = 1\times 0 + 0\times 1 + 0\times 0 = 0$$

e così si può mostrare che sono ortogonali due a due

$$\mathbf{e}_1^T\mathbf{e}_2 = \mathbf{e}_1^T\mathbf{e}_3 = \mathbf{e}_2^T\mathbf{e}_3 = 0.$$

Il prodotto scalare permette di definire la *lunghezza*, al quadrato, di un vettore $\mathbf{x}$, ossia la distanza del punto $\mathbf{x}$ dall'origine

$$d^2(\mathbf{x},\mathbf{0}_n) \stackrel{\text{def}}{=} \mathbf{x}^T\mathbf{x} = x_1x_1 + x_2x_2 + \ldots + x_nx_n = \sum_{k=1}^{n} x_k^2. \qquad (2.5.2)$$

Quando la lunghezza di un vettore è 1, il vettore è detto *unitario*. Così i tre vettori unità $\mathbf{e}_1$, $\mathbf{e}_2$ ed $\mathbf{e}_3$ scelti come base dello spazio euclideo $\Re^3$, sono unitari perché,

$$d^2(\mathbf{e}_1,\mathbf{0}_3) = \mathbf{e}_1^T\mathbf{e}_1 = (1\ 0\ 0)\begin{pmatrix}1\\0\\0\end{pmatrix} = 1\times 1 + 0\times 0 + 0\times 0 = 1$$

e lo stesso vale per $\mathbf{e}_2$ ed $\mathbf{e}_3$. Immaginando che i tre vettori individuino tre rette orientate passanti per l'origine $\mathbf{0}_3$, i tre vettori unitari fissano su ciascuna di esse l'unità di lunghezza. I tre vettori di base, in quanto ortogonali tra loro e di lunghezza unitaria, costituiscono una base *ortonormale* di $\Re^3$.

Il prodotto scalare ha una semplice interpretazione geometrica, mostrata in TAV. 13. Se $\mathbf{x}$ e $\mathbf{y}$ sono due vettori non nulli di $\Re^3$, di lunghezza $\mathbf{x}^T\mathbf{x}$ e $\mathbf{y}^T\mathbf{y}$, il loro prodotto scalare può essere espresso in forma trigonometrica come

$$\mathbf{x}^T\mathbf{y} = \mathbf{x}^T\mathbf{x}\ \ \mathbf{y}^T\mathbf{y}\ \cos\theta$$

dove θ è l'angolo formato all'origine dai due vettori. La *lunghezza della proiezione* ortogonale di $\mathbf{x}$ sulla retta individuata da $\mathbf{y}$, che può immaginarsi come l'ombra che $\mathbf{x}$ proietta su $\mathbf{y}$, è il numero

$$\text{lungh. della proiezione di } \mathbf{x} \text{ su } \mathbf{y} \stackrel{\text{def}}{=} \mathbf{x}^T\mathbf{x}\cos\theta$$

per cui il prodotto scalare di $\mathbf{x}$ e $\mathbf{y}$ è eguale $\mathbf{y}^T\mathbf{y}$ volte la lunghezza della proiezione di $\mathbf{x}$ su $\mathbf{y}$. Il significato geometrico diventa ancora più chiaro se uno dei due vettori è un vettore $\mathbf{u}$ di lunghezza unitaria: $\mathbf{u}^T\mathbf{u} = 1$. In tal caso il vettore, e il corrispondente punto, $(\mathbf{x}^T\mathbf{u})\,\mathbf{u}$ sono detti *proiezione ortogonale* di $\mathbf{x}$ su $\mathbf{u}$, e

$$\mathbf{x}^T\mathbf{u} = \mathbf{x}^T\mathbf{x}\,\cos\theta$$

è la *componente* di $\mathbf{x}$ sulla retta individuata da $\mathbf{u}$, ossia la distanza della proiezione dall'origine. La lunghezza della proiezione definita qui sopra è consistente con quella presentata in modo più informale nelle Sezioni precedenti. Per esempio se il vettore $\mathbf{x}$ è la prima colonna di $\mathbf{N}$ ed i vettori unitari sono i vettori della base canonica $\mathbf{e}_1$, $\mathbf{e}_2$ e $\mathbf{e}_3$, la lunghezza della proiezione di $\mathbf{x}$ su $\mathbf{e}_1$ risulta essere di

$$\mathbf{x}^T\mathbf{e}_1 = (576\ \ 269\ \ 254)\begin{pmatrix}1\\0\\0\end{pmatrix} = 576$$

unità, e corrisponde proprio alla prima componente di $\mathbf{x}$. Parimenti risulta che 269, la seconda componente di $\mathbf{x}$, è la lunghezza della proiezione di $\mathbf{x}$ su $\mathbf{e}_2$ che può essere interpretata come la coordinata sul secondo asse del punto $\mathbf{x}$, o, più correttamente, del punto X corrispondente al vettore $\mathbf{x}$. Infine si vede che 254 è la lunghezza della proiezione del vettore $\mathbf{x}$ sul terzo asse individuato da $\mathbf{e}_3$ ed è la coordinata del punto $\mathbf{x}$ su questo asse. Nello spazio $\Re^n$ la lunghezza della proiezione ortogonale di $\mathbf{x}$ sull'asse individuato dal vettore unitario $\mathbf{e}_k$, con $k = 1, 2, \ldots, n$ è proprio la k^{ma} componente di $\mathbf{x}$.

Il prodotto scalare permette di definire una *distanza* $d(\mathbf{x}, \mathbf{y})$ tra due vettori $\mathbf{x}$ e $\mathbf{y}$ di $\Re^n$, come radice quadrata della somma dei quadrati degli scostamenti tra coordinate

$$d^2(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} (\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) = \sum_{k=1}^{n}(x_k - y_k)^2. \tag{2.5.3}$$

La distanza $d(\mathbf{x}, \mathbf{y})$ è detta *distanza euclidea canonica*: è un numero reale positivo o nullo quando i due vettori sono eguali. Per esempio la distanza tra

i due vettori di base $\mathbf{e}_1$ e $\mathbf{e}_2$ risulta

$$d^2(\mathbf{e}_1,\mathbf{e}_2) = (\mathbf{e}_1-\mathbf{e}_2)^T(\mathbf{e}_1-\mathbf{e}_2) = \left[\begin{pmatrix}1\\0\\0\end{pmatrix}-\begin{pmatrix}0\\1\\0\end{pmatrix}\right]^T\left[\begin{pmatrix}1\\0\\0\end{pmatrix}-\begin{pmatrix}0\\1\\0\end{pmatrix}\right]$$
$$= (1\ \ -1\ \ 0)\begin{pmatrix}1\\-1\\0\end{pmatrix} = 1+1+0 = 2.$$

e quindi $d(\mathbf{e}_1,\mathbf{e}_2) = \sqrt{2}$. È il teorema di Pitagora nello spazio $\Re^3$: in un triangolo con due lati ortogonali di lunghezza unitaria, la lunghezza del terzo si ricava da $1^2+1^2 = (\sqrt{2})^2$.

Un'importante osservazione che si può fare fin da ora, è che nello spazio euclideo $\Re^n$ dotato del prodotto scalare (2.5.1), e quindi della distanza (2.5.3), è possibile individuare una base ortonormale $\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n$: la base canonica di $\Re^n$. Analogamente, se in $\Re^n$ è individuabile una base ortonormale, allora e solo allora il prodotto scalare è come in (2.5.1) e la distanza è la distanza euclidea canonica (2.5.3). Questo tema importante per l'Analisi delle Corrispondenze, verrà approfondito più avanti nella Sez. 3.6 del prossimo capitolo.

2.6 - Profili e masse

Nelle Sezioni precedenti sono stati presentati gli strumenti matematici e geometrici necessari a sviluppare la metodologia dell'Analisi delle Corrispondenze. Il procedimento verrà illustrato passo passo utilizzando come esempio la matrice di contingenza *Spettacoli-3* di ordine 3×8 di TAV 14. Da essa, dividendo gli elementi di ciascun riga e di ciascuna colonna per il loro totale, si ricavano le matrici $\mathbf{R}$ e $\mathbf{C}$ dei profili delle righe e delle colonne [1]. A partire da questa Sezione il confronto tra modalità delle $J \geq I$ colonne avverrà tramite il confronto dei rispettivi profili, ossia delle colonne della matrice $\mathbf{C}$. Il confronto tra i profili delle righe sarà trattato nella Sez. 4.8, mentre le associazioni che possono intercorrere tra i profili delle righe e delle colonne verranno prese in esame a partire dalla Sez. 4.9.

All'inizio di questo capitolo si è visto che un vettore può immaginarsi come una matrice ridotta ad una sola colonna. Reciprocamente, una matrice di ordine $I\times J$ può ritenersi costituita da J vettori colonna di ordine I. La

[1] La somma delle componenti di un profilo vale 1. A causa degli arrotondamenti alcuni profili dell'esempio risultano avere somma 1.001.

matrice $\mathbf{C}$ dell'esempio è costituita da 8 vettori di ordine 3: $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_8$ così disposti

$$\mathbf{C} = \begin{array}{c} \begin{matrix} \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 & \mathbf{c}_4 & \mathbf{c}_5 & \mathbf{c}_6 & \mathbf{c}_7 & \mathbf{c}_8 \end{matrix} \\ \begin{pmatrix} 0.524 & 0.610 & 0.527 & 0.636 & 0.534 & 0.573 & 0.452 & 0.622 \\ 0.245 & 0.220 & 0.234 & 0.152 & 0.291 & 0.224 & 0.355 & 0.156 \\ 0.231 & 0.171 & 0.239 & 0.212 & 0.175 & 0.203 & 0.194 & 0.222 \end{pmatrix} \end{array}.$$

Anche il profilo della colonna marginale è un vettore di ordine 3 e, dal momento che è la media ponderata dei profili colonna, come si è visto nella Sez. 1.8, verrà indicato con $\overline{\mathbf{c}}$

$$\overline{\mathbf{c}} = \begin{pmatrix} \overline{c}_1 \\ \overline{c}_2 \\ \overline{c}_3 \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} n_{1+}/n_{++} \\ n_{2+}/n_{++} \\ n_{3+}/n_{++} \end{pmatrix} = \begin{pmatrix} 0.548 \\ 0.236 \\ 0.216 \end{pmatrix}. \tag{2.6.1}$$

Analogamente il profilo della riga marginale, media ponderata dei profili delle righe, come nella Sez. 1.6, è un vettore riga di ordine 8 e verrà indicato con $\overline{\mathbf{r}}^T$

$$\begin{aligned} \overline{\mathbf{r}}^T &= (\overline{r}_1 \quad \overline{r}_2 \quad \overline{r}_3 \quad \overline{r}_4 \quad \overline{r}_5 \quad \overline{r}_6 \quad \overline{r}_7 \quad \overline{r}_8) \\ &\stackrel{\text{def}}{=} (\frac{n_{+1}}{n_{++}} \quad \frac{n_{+2}}{n_{++}} \quad \frac{n_{+3}}{n_{++}} \quad \frac{n_{+4}}{n_{++}} \quad \frac{n_{+5}}{n_{++}} \quad \frac{n_{+6}}{n_{++}} \quad \frac{n_{+7}}{n_{++}} \quad \frac{n_{+8}}{n_{++}}) \\ &= (0.419 \quad 0.110 \quad 0.144 \quad 0.013 \quad 0.039 \quad 0.247 \quad 0.012 \quad 0.017) \end{aligned} \tag{2.6.2}$$

e dalla Sez. 1.8 è ormai noto che le sue componenti $\overline{r}_1 \ \overline{r}_2 \ \ldots \ \overline{r}_8$ sono le *masse* (o pesi relativi) dei profili colonna $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_8$, che tengono conto della diversa affluenza di pubblico ai tipi di spettacolo, informazione che va perduta passando dalla matrice di contingenza $\mathbf{N}$ a quella dei profili $\mathbf{C}$.

L'interpretazione geometrica dei vettori permette di affermare che nello spazio euclideo $\Re^3$ con origine nel punto $\mathbf{0}_3$ e con base costituita da $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, i profili $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_8$, individuano univocamente 8 punti le cui coordinate sono le tre componenti dei profili. Gli 8 punti costituiscono una *nuvola* di punti dotati di massa.

Profili e punti della nuvola non sono però entità astratte, ma hanno un preciso significato: rappresentano gli 8 tipi di spettacolo, e così ad esempio il punto $\mathbf{c}_1 = (0.524 \ 0.245 \ 0.231)^T$ indica le rappresentazioni di Prosa. Le sue tre coordinate sono le quote di spettatori a rappresentazioni di Prosa nelle tre aree: $c_{11} = 0.524$ quota di spettatori nel Nord, $c_{12} = 0.245$ quota al Centro e $c_{13} = 0.231$ quota al Sud. I tre vettori della terna di base sono associati alle modalità delle righe e quindi alle tre Aree Geografiche: il punto $\mathbf{e}_1 = (1 \ 0 \ 0)^T$

indica una situazione in cui ci sono spettatori soltanto al Nord, il punto $\mathbf{e}_2$ soltanto al Centro ed $\mathbf{e}_3$ soltanto al Sud. Questi punti individuano quindi tre situazioni estreme per l'affluenza di spettatori.

Più in generale, le J colonne di $\mathbf{C}$ individuano una nuvola di J punti nello spazio euclideo I-dimensionale $\Re^I$. Ciascun punto $\mathbf{c}_j$, con $j = 1, 2, \ldots, J$, è dotato della massa $\overline{r}_j$ e, riferito alla base canonica, è esprimibile in un unico modo rispetto a questa

$$\mathbf{c}_j = c_{1j}\,\mathbf{e}_1 + c_{2j}\,\mathbf{e}_2 + \ldots + c_{ij}\,\mathbf{e}_i \; + \ldots + c_{Ij}\,\mathbf{e}_I$$

dove il vettore $c_{ij}\,\mathbf{e}_i$ individua il punto proiezione ortogonale del profilo $\mathbf{c}_j$ su $\mathbf{e}_i$ e c_{ij} è la coordinata di questa proiezione, ossia la sua distanza dall'origine.

Il punto $\overline{\mathbf{c}}$ è il *baricentro* della nuvola

$$\overline{\mathbf{c}} \;=\; \sum_{j=1}^{J} \overline{r}_j\;\mathbf{c}_j \tag{2.6.3}$$

media ponderata dei profili o, in altri termini, combinazione lineare dei J profili colonna $\mathbf{c}_j$ con coefficienti le loro masse $\overline{r}_j$, come nella Sez. 1.8. Le J masse hanno somma 1. La sua i^{ma} componente vale

$$\overline{c}_i \;=\; \sum_{j=1}^{J} \overline{r}_j\; c_{ij} \qquad\qquad i = 1, 2, \ldots, I$$

mentre, un'altra espressione del baricentro in cui interviene la matrice dei profili colonna che verrà impiegata nel seguito, è

$$\overline{\mathbf{c}} = \mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{1}_{\mathrm{J}} \tag{2.6.4}$$

dove (APP. A)

$$\mathbf{D}_{\overline{\mathbf{r}}} = diag\,(\overline{r}_1\,\overline{r}_2\;\ldots\;\overline{r}_j\;\ldots\;\overline{r}_J) \tag{2.6.5}$$

è la matrice diagonale di ordine $J \times J$ delle masse dei J profili colonna e $\mathbf{1}_{\mathrm{J}}$ è un vettore di ordine J le cui componenti sono tutte 1. Il baricentro $\overline{\mathbf{c}}$ sarà il punto di riferimento per lo studio della configurazione della nuvola dei profili colonna, così come il profilo colonna medio lo è stato per il metodo dei confronti, illustrato nel primo Capitolo.

Se due profili sono eguali, se cioè le loro prime componenti sono eguali fra loro, le seconde eguali tra loro, ecc., i due punti coincidono. Se le componenti differiscono di poco i due punti sono vicini, se differiscono di molto lontani. Perciò confrontare i profili col profilo medio, come veniva fatto nel primo capitolo, equivale a esaminare la dispersione della nuvola di punti attorno al

suo baricentro. Ecco quindi che il problema del confronto dei profili risulta "geometrizzato" e ricondotto a quello di osservare, in una rappresentazione grafica opportuna, come sono disposti i punti della nuvola. Il problema risulta semplificato, perché è più agevole osservare delle distanze tra punti che raffrontare sequenze di numeri. Resta il problema, non trascurabile, che oltre le tre dimensioni ogni rappresentazione grafica è impossibile. L'Analisi delle Corrispondenze supera questa difficoltà proiettando i punti della nuvola in un sottospazio di inferiore e più accessibile dimensionalità, cercando che siano rispettate "al meglio" le distanze effettive tra i punti. Così, se per dare una rappresentazione accessibile alla nostra percezione, il sottospazio è bi-dimensionale (un piano), le proiezioni dei punti possono essere graficate su un foglio di carta. C'è un inconveniente nel far questo, perché se sul grafico due proiezioni risultano coincidenti (o quasi) non è detto che tali fossero i due punti della nuvola. Il procedimento di proiezione comporta perciò delle distorsioni nella riproduzione delle distanze, peraltro quantificabili, ma questo è un prezzo che conviene pagare per potersi rendere conto della configurazione della nuvola di punti e quindi della struttura delle relazioni che intercorrono tra i profili e tra le modalità che rappresentano.

2.7 - Simplesso dei profili

Nelle Sezioni precedenti si è visto che allo spazio euclideo $\Re^3$ con origine in $\mathbf{0}_3$ appartengono tutti e solo i vettori di ordine 3 esprimibili come combinazioni lineari dei 3 vettori della base canonica ortonormale $\mathbf{e}_1, \mathbf{e}_2$ e $\mathbf{e}_3$. Tra questi vettori ci sono anche i profili dello stesso ordine, le cui componenti hanno due vincoli notevoli. Il primo è che non sono negative, per cui i punti individuati restano confinati nel triedro positivo di $\Re^3$. Il secondo è che la loro somma è 1 e questo significa che i punti restano ulteriormente confinati nella porzione triangolare di piano che interseca i tre vettori di base $\mathbf{e}_1, \mathbf{e}_2$ e $\mathbf{e}_3$ ai loro estremi: un *triangolo equilatero* con vertici i punti di coordinate $(1, 0, 0)$, $(0, 1, 0)$ e $(0, 0, 1)$, come si vede nella TAV. 15. In altri termini, i due vincoli sulle componenti costringono i profili in in una regione triangolare di uno spazio bi-dimensionale $\Re^2$ (un piano) contenuto in $\Re^3$. Allo stesso modo i punti individuati da profili con due componenti sono confinati sul *segmento* che congiunge gli estremi dei vettori di base $\mathbf{e}_1 = (1 \quad 0)^T$ e $\mathbf{e}_2 = (0 \quad 1)^T$, mentre quelli individuati da profili con 4 componenti restano all'interno di un *tetraedro* con 4 vertici, regione dello spazio $\Re^3$, contenuto in $\Re^4$. La generalizzazione del segmento, del triangolo e del tetraedro porta al concetto di *simplesso* $I-1$-dimensionale con I vertici $(1, 0, \ldots, 0)$, $(0, 1, \ldots, 0)$, $\ldots$,

$(0, 0, \ldots, 1)$, regione dell'iperpiano $\Re^{I-1}$, contenuto in $\Re^I$, ove i profili con I componenti restano confinati[1]. Lo spazio $\Re^I$ è lo spazio dei *vettori* di ordine I, tra i quali ci sono *anche* i profili di ordine I. Il nome di *spazio dei profili* delle colonne che abitualmente si dà a questo spazio intende soltanto sottolineare il fatto che l'interesse è rivolto a questi ultimi, e non che lo spazio è 'popolato' soltanto da profili.

I profili a tre componenti, come quelli dell'esempio, si possono rappresentare direttamente sul simplesso piano a 3 vertici mediante un *diagramma ternario* o triangolare, diffuso particolarmente in geologia e in chimica, dove talvolta occorre graficare gruppi di campioni di cui sono state rilevate le *percentuali* di tre loro componenti. Sono in commercio dei fogli già predisposti per tali grafici, in cui gli assi coordinati sono i tre lati di un triangolo equilatero. Ogni lato è graduato da 0 a 1 e rappresenta una delle tre componenti. La TAV. 16 illustra il procedimento per posizionare sul simplesso il profilo-colonna $\mathbf{c}_1$ relativo alla quota d'affluenza di spettatori a rappresentazioni di Prosa.

Accade sovente che i punti-profilo risultino concentrati in una regione molto limitata del simplesso, perché le loro componenti variano in un intervallo ristretto dell'intervallo possibile $[0, 1]$. Ad esempio, nella TAV. 14 si vede che le componenti di $\mathbf{c}_2$ variano tra $c_{32} = 0.171$ e $c_{12} = 0.610$ e questo è l'intervallo di variazione più ampio per le componenti degli 8 profili.

2.8 - Distanza distribuzionale tra profili

Nel primo capitolo, la similarità tra due colonne di una matrice di contingenza $\mathbf{N}$ veniva valutata confrontando visivamente i rispettivi profili. Nell'Analisi delle Corrispondenze la similarità tra due colonne j e k di $\mathbf{N}$, è *misurata* dalla distanza tra i loro profili $\mathbf{c}_j$ e $\mathbf{c}_k$ che va sotto il nome di *distanza distribuzionale*, così definita

$$d_D^2(\mathbf{c}_j, \mathbf{c}_k) \stackrel{\text{def}}{=} \sum_{i=1}^{I} \frac{1}{\overline{c}_i} (c_{ij} - c_{ik})^2 \qquad (2.8.1)$$

$$= \sum_{i=1}^{I} \frac{1}{\frac{n_{i+}}{n_{++}}} \left(\frac{n_{ij}}{n_{+j}} - \frac{n_{ik}}{n_{+k}} \right)^2$$

[1] Un iperpiano è uno spazio euclideo con una dimensione in meno dello spazio ambiente $\Re^I$ dei profili, così come una retta è un iperpiano di $\Re^2$, un piano è un iperpiano di $\Re^3$, ecc.

che vale per tutti i profili $j, k = 1, 2, \ldots, J$. Questa distanza è simile alla distanza euclidea canonica (2.5.3), in quanto è ancora ottenuta come somma dei quadrati degli scarti tra le coordinate dei due profili su ciascuno degli I assi individuati dai vettori di base, ma, differenza importante, ciascuno di questi scarti è ora ponderato con un peso pari all'inverso della corrispondente componente del profilo-colonna medio che, fin dalla Sez. 1.8, è noto essere la massa, sempre positiva, del profilo-riga. L'effetto della ponderazione è quello di equilibrare l'influenza delle righe nel computo della distanza tra profili-colonna, aumentando l'importanza degli scarti che si riferiscono a righe con modalità complessivamente meno frequenti. Nel caso dell'esempio, si può dire che gli scarti hanno peso elevato per le regioni con bassa affluenza di spettatori, e peso esiguo quando l'affluenza vi è particolarmente alta.

Nella Sez. 2.5 la distanza euclidea canonica tra due generici vettori di $\Re^n$ è stata espressa tramite il loro prodotto scalare. Per esprimere in tale forma la distanza distribuzionale (2.8.1) tra due profili di $\Re^I$, occorre costruire la matrice diagonale che ha per elementi diagonali le componenti di $\overline{\mathbf{c}}$, e che nel caso dell'esempio, dalla (2.6.1) risulta essere

$$\overline{\mathbf{c}} = \begin{pmatrix} 0.548 \\ 0.236 \\ 0.216 \end{pmatrix} \quad \mathbf{D}_{\overline{\mathbf{c}}} = \begin{pmatrix} 0.548 & 0 & 0 \\ 0 & 0.236 & 0 \\ 0 & 0 & 0.216 \end{pmatrix}$$

$$\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} = \begin{pmatrix} \frac{1}{0.548} & 0 & 0 \\ 0 & \frac{1}{0.236} & 0 \\ 0 & 0 & \frac{1}{0.216} \end{pmatrix} = \begin{pmatrix} 1.825 & 0 & 0 \\ 0 & 4.237 & 0 \\ 0 & 0 & 4.630 \end{pmatrix}.$$

In forma matriciale la distanza distribuzionale (2.8.1) tra due profili colonna $\mathbf{c}_j$ e $\mathbf{c}_k$ di ordine I si esprime dunque

$$d_D^2(\mathbf{c}_j, \mathbf{c}_k) \stackrel{\text{def}}{=} (\mathbf{c}_j - \mathbf{c}_k)^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, (\mathbf{c}_j - \mathbf{c}_k) = \sum_{i=1}^{I} \frac{1}{\overline{c}_i} \, (c_{ij} - c_{ik})^2. \qquad (2.8.2)$$

dove $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$ è la matrice diagonale, di ordine $I \times I$, dei pesi da dare agli scarti

$$\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} = diag \left(\frac{1}{\overline{c}_1} \, \frac{1}{\overline{c}_2} \, \ldots \, \frac{1}{\overline{c}_i} \, \ldots \frac{1}{\overline{c}_I} \right). \qquad (2.8.3)$$

Nell'Analisi delle Corrispondenze, la distanza distribuzionale conferisce allo spazio dei profili una struttura euclidea. In questo spazio, che verrà indicato ancora con $\Re^I$, con l'intesa però che la distanza tra profili è la distanza distribuzionale, il *prodotto scalare* tra due profili $\mathbf{c}_j$ e $\mathbf{c}_k$ di ordine I, è

espresso come

$$\mathbf{c}_j^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{c}_k \stackrel{\text{def}}{=} \sum_{i=1}^{I} \frac{1}{\overline{c}_i} \, c_{ij} \, c_{ik} = \sum_{i=1}^{I} \frac{1}{\frac{n_{i+}}{n_{++}}} \, \frac{n_{ij}}{n_{+j}} \, \frac{n_{ik}}{n_{+k}}. \tag{2.8.4}$$

È detto prodotto scalare *associato* [1] *alla matrice* $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$. Quando risulta nullo, i due profili sono $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$ -*ortogonali*.

La *lunghezza* di un profilo è pari alla sua distanza dall'origine, e quindi per la (2.8.2) alla radice quadrata di

$$d_D^2(\mathbf{c}_j, \mathbf{0}_{\mathrm{I}}) = \mathbf{c}_j^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{c}_j = \sum_{i=1}^{I} \frac{1}{\overline{c}_i} \, c_{ij}^2 = \sum_{i=1}^{I} \frac{1}{\frac{n_{i+}}{n_{++}}} \left(\frac{n_{ij}}{n_{+j}} \right)^2. \tag{2.8.5}$$

In particolare, la lunghezza del profilo $\overline{\mathbf{c}}$ che individua il baricentro, risulta

$$d_D^2(\overline{\mathbf{c}}, \mathbf{0}_{\mathrm{I}}) = \overline{\mathbf{c}}^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \overline{\mathbf{c}} = \sum_{i=1}^{I} \frac{\overline{c}_i^2}{\overline{c}_i} = \sum_{i=1}^{I} \overline{c}_i = 1. \tag{2.8.6}$$

Il profilo $\overline{\mathbf{c}}$ è detto avere lunghezza $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$ -*unitaria*. Infine, la distanza distribuzionale tra un profilo $\mathbf{c}_j$ e il baricentro, per la (2.8.2), è la radice quadrata di

$$d_D^2(\mathbf{c}_j, \overline{\mathbf{c}}) \;=\; (\mathbf{c}_j - \overline{\mathbf{c}})^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, (\mathbf{c}_j - \overline{\mathbf{c}}) = \sum_{i=1}^{I} \frac{1}{\overline{c}_i} \, (c_{ij} - \overline{c}_i)^2. \tag{2.8.7}$$

La definizione di distanza distribuzionale permette di evidenziare alcune proprietà geometriche generali della nuvola dei J profili. Anzitutto, come anticipato nella Sez. 2.7, *tutti* i punti della nuvola sono contenuti nel simplesso a $I-1$ vertici dell'iperpiano di $\Re^I$ che contiene anche il baricentro $\overline{\mathbf{c}}$ della nuvola. Inoltre l'iperpiano del simplesso risulta $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale al profilo $\overline{\mathbf{c}}$, e quindi alla retta che passa per l'origine e per il baricentro. Infatti, tenendo conto della (2.8.7), per ogni $j = 1, 2, \ldots, J$ è

$$(\mathbf{c}_j - \overline{\mathbf{c}})^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \overline{\mathbf{c}} = \mathbf{c}_j^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \overline{\mathbf{c}} - \overline{\mathbf{c}}^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \overline{\mathbf{c}} = \sum_{i=1}^{I} c_{ij} \, \frac{1}{\overline{c}_i} \, \overline{c}_i - 1 = 1 - 1 = 0.$$

Perciò, *tutti* i J punti della nuvola ed il loro baricentro sono contenuti nel simplesso incluso in un iperpiano $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale a $\overline{\mathbf{c}}$. Tutto questo può

---

[1] La matrice diagonale definita positiva $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$ è detta *metrica* dello spazio $\Re^I$. Frequentemente però, con lo stesso termine si indica la distanza (2.8.2).

essere riassunto con l'espressione

$$\left(\mathbf{C} - \overline{\mathbf{C}}\right)^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \overline{\mathbf{c}} = \mathbf{0}_{\mathrm{I}} \tag{2.8.8}$$

dove $\overline{\mathbf{C}} = \overline{\mathbf{c}}\,\mathbf{1}_{\mathrm{J}}^T$ è una matrice di ordine $I \times J$ in cui ogni colonna è $\overline{\mathbf{c}}$, mentre $\mathbf{1}_{\mathrm{J}}$ è il vettore di ordine J le cui componenti sono costituite tutte da 1.

Un ultimo punto importante da sottolineare è che nello spazio dei profili $\Re^I$, gli I vettori della base canonica $\mathbf{e}_i, \mathbf{e}_2, \ldots, \mathbf{e}_I$ costituiscono una base $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale, *ma non* $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortonormale. Infatti sono $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonali due a due perché per ogni $i, k = 1, 2, \ldots, I$ con $j \neq k$ è

$$\mathbf{e}_i^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{e}_k = 0$$

ma non hanno lunghezza $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitaria perché

$$\mathbf{e}_i^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{e}_i = d_D^2(\mathbf{e}_i, \mathbf{0}_{\mathrm{I}}) = \frac{1}{\overline{c}_i}. \tag{2.8.9}$$

2.9 - Proprietà equidistributiva

Tornando all'esempio, è ora possibile, grazie a quanto stabilito nella Sez. precedente, quantificare ad esempio la distanza tra i Concerti di Musica Classica ($j = 3$) e quelli di Musica Leggera ($j = 6$) in base alle rispettive quote di spettatori nelle tre aree, ossia in base ai profili della matrice $\mathbf{C}$ (3×8), riportata nella TAV. 14.

$$d_D^2(\mathbf{c}_3, \mathbf{c}_6) = (\mathbf{c}_3 - \mathbf{c}_6)^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, (\mathbf{c}_3 - \mathbf{c}_6) =$$

$$\left[\begin{pmatrix} 0.527 \\ 0.234 \\ 0.239 \end{pmatrix} - \begin{pmatrix} 0.573 \\ 0.224 \\ 0.203 \end{pmatrix}\right]^T \begin{pmatrix} 1.825 & 0 & 0 \\ 0 & 4.237 & 0 \\ 0 & 0 & 4.630 \end{pmatrix} \left[\begin{pmatrix} 0.527 \\ 0.234 \\ 0.239 \end{pmatrix} - \begin{pmatrix} 0.573 \\ 0.224 \\ 0.203 \end{pmatrix}\right]$$

$$= (-0.046 \;\; 0.010 \;\; 0.036) \begin{pmatrix} 1.825 & 0 & 0 \\ 0 & 4.237 & 0 \\ 0 & 0 & 4.630 \end{pmatrix} \begin{pmatrix} -0.046 \\ 0.010 \\ 0.036 \end{pmatrix}$$

$$= 1.825 \times (-0.046)^2 + 4.237 \times 0.010^2 + 4.630 \times 0.036^2 = 0.0103.$$

Appare evidente l'effetto della ponderazione degli scarti tra coordinate. Mentre il peso relativo al Nord ($i = 1$) vale 1.825 perché il Nord ha una quota alta di spettatori ($\overline{c}_3 = 0.548$), i pesi relativi al Centro ($i = 2$) e al Sud ($i = 3$) sono più che doppi, perchè la loro quota di spettatori è meno della metà di quella del Nord. Il risultato è che

$$d_D(C.m.\,classica,\, C.m.\,leggera) = +\sqrt{0.0103} = 0.101.$$

Con questa espressione si intende che la distanza distribuzionale tra le due modalità, valutata tramite i loro profili, è 0.101 . La distanza dipende quindi dalla particolare matrice che si considera. Se venisse invece calcolata dagli omologhi profili colonna della matrice *Spettacoli* di ordine 20×8 di TAV. 2, si otterrebbe una distanza di 0.361. Questo risultato fa capire che l'accorpamento delle 20 righe della matrice *Spettacoli* in base alle 3 aree geografiche, procedimento seguito per costruire la matrice *Spettacoli-3*, ha alterato la distanza distibuzionale tra i profili colonna.

Questa osservazione porta a rispondere ora a una domanda che, legittimamente, il lettore si sarà posto. Perché usare la distanza distribuzionale (2.8.1) invece di quella euclidea canonica (2.5.3), che, oltre a esserci familiare, ha anche il pregio di potersi calcolare in modo più semplice? I motivi sono due. Il primo è che la strategia dell'Analisi delle Corrispondenze, che verrà delineandosi nel prossimo Capitolo, consiste nel rendere intellegibile 'al meglio' la configurazione della nuvola dei profili, proiettandola in sottospazi derivati dai vettori della base. Da questo punto di vista la base canonica non ha i requisiti ottimali e va quindi sostituita. Se si vuole restare però nell'ambito degli spazi euclidei, la distanza può essere una distanza euclidea pesata, del tipo di quella distribuzionale. La scelta poi dei pesi $1/\overline{c}_i$ deriva, ecco il secondo e principale motivo, dal fatto che la distanza distribuzionale gode di una proprietà fondamentale, detta *proprietà equidistributiva* che può enunciarsi così: se due *righe proporzionali* di $\mathbf{N}$ vengono cumulate in una sola, la distanza distribuzionale tra due qualunque profili-*colonna* resta immutata [1]. Vale anche il contrario, per cui se una riga viene proporzionalmente suddivisa in due nuove righe e sostituita da queste, le mutue distanze tra profili colonna, calcolate con la (2.8.1), non cambiano. La dimostrazione verrà data più avanti nella Sez. 4.8. Una perfetta proporzionalità si presenta ben di rado in pratica, ma costituisce tuttavia una situazione limite alla quale, talvolta, una situazione reale è prossima, per cui sostituire due righe di $\mathbf{N}$ *quasi* proporzionali con la loro somma, non altera *sensibilmente* i risultati dell'Analisi delle Corrispondenze. Ad esempio, la variabile *età* può essere suddivisa in classi più o meno ampie e quindi avere modalità più o meno dettagliate, ma la scelta tra le due suddivisioni non è critica perché l'Analisi delle Corrispondenze delle due matrici

[1] Questa proprietà non è valida per la distanza euclidea canonica, ma vale per alcuni altri tipi di distanza, come mostrato da B. Escofier in *Analyse factorielle et distances répondant au principe d'équivalence distributionelle.* Rev. Stat. Appl. (1978), vol. 26, n. 4, pg. 29 - 37.

porterebbe a risultati molto prossimi.

Riprendendo l'esempio all'inizio di questa Sezione, si spiega così perché la distanza tra i due tipi di Concerto risulti 0.101 quando i profili sono di ordine 3 e 0.361 quando sono di ordine 20: si sono cumulate righe che erano lontane dall'essere proporzionali.

Per verificare la proprietà equidistributiva si possono accorpare nella matrice di contingenza *Spettacoli-3* dell'esempio, le ultime due righe dal momento che sono quasi proporzionali. La distanza tra i due profili $\mathbf{c}_3$ e $\mathbf{c}_6$, che sono ora di ordine 2×1, non dovrebbe sostanzialmente mutare. La matrice di contingenza che si ottiene è

$$\begin{pmatrix} 576 & 175 & 198 & 21 & 55 & 370 & 14 & 28 \\ 523 & 112 & 178 & 12 & 48 & 276 & 17 & 17 \end{pmatrix}$$

dove la prima riga si riferisce al Nord e la seconda al Centro-Sud. La matrice dei profili colonna risulta

$$\begin{pmatrix} 0.524 & 0.610 & 0.527 & 0.636 & 0.534 & 0.573 & 0.452 & 0.622 \\ 0.476 & 0.390 & 0.473 & 0.364 & 0.466 & 0.427 & 0.548 & 0.378 \end{pmatrix}$$

Il baricentro e la matrice diagonale dei pesi (2.8.3) risultano

$$\overline{\mathbf{c}} = \begin{pmatrix} 0.548 \\ 0.452 \end{pmatrix} \qquad \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} = \begin{pmatrix} 1.825 & 0 \\ 0 & 2.212 \end{pmatrix}$$

per cui $d_D^2(\mathbf{c}_3, \mathbf{c}_6) =$

$$\left[\begin{pmatrix} 0.527 \\ 0.473 \end{pmatrix} - \begin{pmatrix} 0.534 \\ 0.466 \end{pmatrix} \right]^T \begin{pmatrix} 1.825 & 0 \\ 0 & 2.212 \end{pmatrix} \left[\begin{pmatrix} 0.527 \\ 0.473 \end{pmatrix} - \begin{pmatrix} 0.534 \\ 0.466 \end{pmatrix} \right]$$

$$= (-0.046 \quad 0.046) \begin{pmatrix} 1.825 & 0 \\ 0 & 2.212 \end{pmatrix} \begin{pmatrix} -0.046 \\ 0.046 \end{pmatrix} = 0.008.$$

Per cui $d_D(C.m.\,classica, C.m.\,leggera) = \sqrt{0.008} = 0.092$.

Buona parte della differenza tra 0.101 e 0.092 è conseguenza degli arrotondamenti per aver operato con solo tre decimali. Per rendersene conto si può sostituire l'ultima riga della matrice *Spettacoli-3* con due righe aventi ognuna la metà degli effettivi. Si ottiene così una nuova matrice di contingenza di ordine 4×8 con le due ultime righe eguali. Per la proprietà equidistributiva la distanza tra il terzo e il sesto profilo colonna di questa nuova matrice deve rimanere 0.101. In effetti, la distanza che risulta è ora 0.105. La differenza è imputabile unicamente agli arrotondamenti.

2.10 - Rappresentazione della distanza distribuzionale

Noi siamo abituati a valutare le distanze su mappe geografiche, stradali, ecc. ritenendo implicitamente che l'unità di scala sia la stessa per ascisse e ordinate. Ma cosí non è per una mappa che riporti le posizioni dei punti-profilo: ascisse e ordinate hanno scala diversa. Occorre quindi alterare le posizioni dei punti-profilo perché la mappa rappresenti le distanze in modo conforme alle nostre abitudini. Questo si può ottenere facilmente perché, riprendendo la (2.8.1), la distanza distribuzionale tra due profili di ordine I può scriversi

$$d_D^2(\mathbf{c}_j, \mathbf{c}_k) \stackrel{\text{def}}{=} \sum_{i=1}^{I} \frac{1}{\overline{c}_i} (c_{ij} - c_{ik})^2 = \sum_{i=1}^{I} \left(\frac{c_{ij}}{\sqrt{\overline{c}_i}} - \frac{c_{ik}}{\sqrt{\overline{c}_i}} \right)^2$$

che è l'espressione della distanza euclidea canonica (2.5.1). Basta quindi dilatare di un fattore $1/\sqrt{\overline{c}_i}$ ciascuna delle I coordinate di un profilo per ottenere una nuova rappresentazione in $\Re^I$ dei punti-profilo tale che le distanze euclidee canoniche tra di essi corrispondano *numericamente* alle effettive distanze distribuzionali.

Nella TAV. 17 gli 8 profili dell'esempio sono stati graficati usando come coordinate le loro effettive componenti. I punti giacciono in un triangolo equilatero che ha per vertici i punti $\mathbf{e}_1$, $\mathbf{e}_2$ e $\mathbf{e}_3$ individuati dai vettori unitari (2.4.2) della terna di base. Per rappresentare la distanza distribuzionale in modo conforme al nostro modo abituale, le lunghezze dei vettori di base vengono trasformate, dilatandole in base a quanto trovato nella Sez. 2.8

$$\sqrt{\mathbf{e}_1^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-1} \, \mathbf{e}_1} = \sqrt{1.825} \quad \sqrt{\mathbf{e}_2^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-1} \, \mathbf{e}_2} = \sqrt{4.237} \quad \sqrt{\mathbf{e}_3^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-1} \, \mathbf{e}_3} = \sqrt{4.630}$$

e, di conseguenza, il simplesso non è più un triangolo equilatero. Le coordinate dei punti vanno dilatate conformemente, per cui, ad esempio, le coordinate trasformate di $\mathbf{c}_1$ diventano

$$\left(\sqrt{1.825} \times 0.524 \quad \sqrt{4.237} \times 0.254 \quad \sqrt{4.630} \times 0.231 \right)^T .$$

La TAV. 17 mostra anche le posizioni assunte dagli 8 profili in modo che le relative distanze distribuzionali appaiano come distanze euclidee, conformemente alle nostre consuetudini.

Il lettore non deve preoccuparsi troppo della rappresentazione delle distanze distribuzionali tra profili, perché l'Analisi delle Corrispondenze si incaricherà di risolvere questo problema.

2.11 - Riepilogo

A questo punto il lettore ha acquisito le tre nozioni fondamentali dell'Analisi delle Corrispondenze: quella di profilo, di massa e di distanza tra profili. Ogni profilo $\mathbf{c}_j$, dove $j = 1, 2, \ldots, J$, è interpretato come un vettore che ha per componenti gli I elementi c_{ij} della j^{ma} colonna della matrice $\mathbf{C}$, come in (1.7.1), ed individua univocamente il punto $\mathbf{c}_j$ nello spazio euclideo I-dimensionale $\Re^I$ con base $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_I$ ed origine in $\mathbf{0}_{\mathrm{I}}$. A motivo dei vincoli sulle componenti, i J punti sono confinati in una regione dell'iperpiano di $\Re^I$, detta simplesso a I vertici.

Ciascun punto $\mathbf{c}_j$ è dotato di una massa (o peso) $\overline{r}_j$, normalizzata a somma uno come in (2.6.2), che tiene conto dell'importanza relativa di ciascun profilo nell'ambito dei J profili considerati. Il complesso dei J punti 'pesanti' prende il nome di nuvola dei punti profilo-colonna, o, più brevemente, di nuvola dei profili delle colonne.

La definizione di distanza tra profili gioca un ruolo cruciale perché influisce direttamente sui risultati dell'analisi. Tuttavia la scelta della distanza distribuzionale (2.8.1), se non obbligata, è per lo meno ampiamente giustificata dalla proprietà equidistributiva.

Nel capitolo che seguirà, coerentemente con le scelte fatte, verrà affrontato il problema di rendere intellegibile la nuvola di punti, proiettandola in un sottospazio accessibile ai nostri sensi, in modo tale che vi siano riprodotte "al meglio" le distanze distribuzionali che misurano le similarità tra i J profili e quindi il grado di associazione tra le modalità che rappresentano. In altri termini, si farà in modo che l'informazione sulla configurazione geometrica dei profili che può estrarsi da questa proiezione, possa ritenersi ottimale, in base ad un criterio prestabilito.

2.12 - Bibliografia essenziale

Molti degli argomenti trattati all'inizio di questo capitolo appartengono a quel settore della Matematica che va sotto il nome di Algebra Lineare. La sua importanza e sopratutto il vasto impiego che trova in molte Scienze applicate, fa sì che i testi disponibili siano innumerevoli. Il lettore interessato potrà approfondire gli argomenti consultando, ad esempio

Brunella Bruno (1992). *Lezioni di Algebra lineare 1: sistemi di equazioni lineari e spazi vettoriali euclidei.* Decibel ed. Padova. 134 pg. ISBN 88-08-20192-9.

Carlo Cercignani (1985). *Vettori, Matrici, Geometria.* Zanichelli ed. Bologna. 191 pg. ISBN n.i., ove l'Algebra lineare è presentata seguendo un approccio puramente matematico.

Altri testi che pongono l'accento anche sull'aspetto statistico sono

Renato Leoni, Natale Carlo Lauro (1984). *Introduzione all'Analisi statistica multidimensionale. Vol. 1: Algebra lineare.* Liguori ed. Napoli. 297 pg. ISBN 88-207-1288-1.

Alfredo Rizzi (1988). *Il linguaggio delle matrici: applicazioni in economia, in statistica e nelle scienze sociali.* NIS ed. Roma. 270 pg. ISBN n. i.

Alexander Basilewsky (1983). *Applied Matrix Algebra in the Statistical Sciences.* North Holland ed. 389 pg. ISBN 0-444-00756-3.

Shayle R. Searle (1982). *Matrix Algebra useful for Statistics.* Wiley ed. 438 pg. ISBN 0-471-86681-4.

Una introduzione all'Algebra lineare che privilegia invece l'aspetto geometrico e grafico è

Paul E. Green, J. Douglas Carrol (1976). *Mathematical tools for Applied multivariate analysis.* Academic Press ed. 376 pg. ISBN 0-12-297552-9. Il testo, ricco di illustrazioni, di esempi e di esercizi (con soluzione) che ne sono parte integrante, non richiede conoscenze matematiche preliminari troppo avanzate. È utile come testo di accompagnamento.

Un altro testo classico sull'Algebra lineare è riportato nella bibliografia al termine del prossimo Terzo capitolo.

PARTE PRIMA: IL METODO

CAPITOLO 3: Autovalori e autovettori

Sommario

Questo capitolo è interamente dedicato all'individuazione di un nuovo sistema di riferimento che sostituisca quello espresso dai vettori della base canonica per lo spazio dei profili $\Re^J$. È il sistema degli assi fattoriali d'inerzia, centrato nel baricentro ed individuato dagli autovettori. Gli assi fattoriali garantiscono la massima visibilità e la minima distorsione della configurazione della nuvola dei profili, quando questa venga proiettata in sottospazi, accessibili alla nostra percezione, da essi individuati.

Il procedimento d'individuazione dei nuovi assi è illustrato in dettaglio impiegando come esempio la matrice *Spettacoli-3*, sicché anche il lettore meno addestrato sarà in grado di seguirne lo sviluppo senza eccessiva difficoltà.

Un'attenta lettura di questo capitolo, da farsi possibilmente con una penna in mano, metterà il lettore in grado di

- acquisire il concetto di inerzia, che viene assunta come indicatore della dispersione geometrica della nuvola dei profili;
- rendersi conto del ruolo assegnato al baricentro dal teorema di Huygens sull'inerzia;
- capire in che modo il metodo di Lagrange permetta di individuare gli autovettori sui quali risulta massima l'inerzia delle proiezioni dei profili;
- comprendere il preciso significato che autovalori ed autovettori assumono nell'Analisi delle Corrispondenze.

CAPITOLO 3

3.1 - Dispersione ed inerzia

Nel primo capitolo è stato affrontato il problema del confronto tra profili per metterne in luce somiglianze e dissimilarità. Nel secondo capitolo ai profili è stata data una interpretazione vettoriale, e quindi geometrica, raffigurandoli come una nuvola di punti dotati di massa, immersi in uno spazio euclideo provvisto di distanza distribuzionale. Il confronto tra profili è stato così, in un certo senso, 'geometrizzato' in quanto le distanze tra punti riflettono differenze tra profili: punti lontani indicano profili dissimili, punti vicini profili simili e punti coincidenti profili eguali. In altri termini, la *configurazione* dei punti traduce la *struttura* dei profili della matrice di contingenza.

La configurazione dei punti ha quindi un interesse intrinseco, mentre la base di riferimento dello spazio euclideo ne ha uno del tutto secondario, perché traslando e/o ruotando comunque la base, la disposizione dei punti resta immutata. Per agevolare il confronto tra profili, l'Analisi delle Corrispondenze individua una nuova base di riferimento, che permetta di rendere visibile 'al meglio' la configurazione dei punti in un sottospazio di inferiore dimensionalità accessibile alla nostra percezione, di solito bi-dimensionale, e permetta al medesimo tempo di contenere l'inevitabile deformazione della configurazione e la conseguente perdita di informazione geometrica sulla struttura dei profili. È evidente che la base canonica non offre alcuna garanzia in tal senso e così pure i sottospazi che da essa si possono ricavare. È necessario perciò individuare una *nuova* base di riferimento che garantisca la migliore visibilità della configurazione col minimo di distorsione.

Per poter individuare questa nuova base, occorre che venga prima chiarito che cosa si intende con 'visibile al meglio' nell'Analisi delle Corrispondenze. Un'occhiata alla TAV. 18 mostra che se si potesse scattare soltanto una foto ricordo di una casetta, converrebbe scegliere l'inquadratura n. 3, perché è quella che ne mostra contemporaneamente il fronte e il lato, anche se un po' distorti. Del resto basta sfogliare una qualunque rivista specializzata o un catalogo di vendite per corrispondenza per verificare che vi abbondano

foto di questo tipo: in cui, cioè l'immagine bidimensionale dell'oggetto tridimensionale vi appare 'sparpagliata' il più possibile. Ecco quindi il criterio di identificazione del sottospazio di assegnata dimensionalità che rende 'visibile al meglio' la configurazione: è quello ove le proiezioni dei punti sono le più 'disperse' possibili perché maggiore è la visibilità della configurazione geometrica e minore la distorsione. In tali circostanze la dispersione delle *proiezioni* è più prossima alla dispersione dei *punti* nel loro spazio originale.

Occorre dunque quantificare la dispersione dei punti tramite un indice che tenga conto delle distanze (distribuzionali) dei profili da un punto di riferimento.

Nell'Analisi delle Corrispondenze la dispersione geometrica della nuvola dei profili intorno ad un punto generico $\mathbf{x}$ in $\Re^I$ è misurata dall'inerzia [1], che è funzione non dei valori assoluti delle semplici distanze, ma dei loro quadrati. Si definisce infatti *inerzia* rispetto ad $\mathbf{x}$ della nuvola dei J punti $\mathbf{c}_j$, dotati di massa $\overline{r}_j$, il numero ottenuto come somma ponderata dei quadrati delle distanze distribuzionali

$$In_{\mathbf{x}} \stackrel{\text{def}}{=} \sum_{j=1}^{J} \overline{r}_j \, d_D^2(\mathbf{c}_j, \mathbf{x}).$$

Il fatto che la somma delle masse sia 1 consente di interpretare l'inerzia come una distanza quadratica media ponderata. Ne deriva che una nuvola di punti molto compatta ha inerzia esigua, una molto dispersa ha inerzia grande. La scelta dell'inerzia, che è una grandezza mai negativa, rende i risultati sensibili agli effetti delle distanze più grandi, dal momento che intervengono al quadrato: due punti con eguali masse, disposti a distanze da $\mathbf{x}$ che stanno in rapporto 10 hanno inerzie in rapporto 100. Tuttavia, l'inerzia è scelta perché è il quadrato della distanza che si scompone su due assi ortogonali in base al teorema di Pitagora (Sez. 3.7), è il quadrato della distanza che entra nel metodo dei minimi quadrati (Sez. 3.14) ed è all'inerzia che si applica il teorema di Huygens (Sez. 3.4).

Viene naturale riferire la dispersione o all'origine della base di riferimento o al baricentro della nuvola che dipende esclusivamente dalla configurazione dei punti [2], anche perché studiare la dispersione dei punti rispetto al

---

[1] Il termine deriva dalla fisica ove un corpo puntiforme di massa m a distanza d da un punto, ha, rispetto a questo, un *momento d'inerzia* pari a $m\,d^2$.

[2] A una data configurazione di punti-profilo corrisponde un unico baricentro,

baricentro equivale a studiare gli scarti tra profili e profilo medio.

3.2 - Inerzia riferita all'origine

Conseguentemente alla definizione, l'espressione dell'inerzia calcolata rispetto all'origine $\mathbf{0}_{\mathrm{I}}$ della base canonica di riferimento $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_I$ di $\Re^I$, assume le forme equivalenti

$$
\begin{aligned}
In_{\mathbf{0}} &= \sum_{j=1}^{J} \overline{r}_j \, d_D^2(\mathbf{c}_j, \mathbf{0}_{\mathrm{I}}) = \sum_{j=1}^{J} \overline{r}_j \sum_{i=1}^{I} \frac{c_{ij}^2}{\overline{c}_i} \\
&= \sum_{j=1}^{J} \overline{r}_j \, \mathbf{c}_j{}^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{c}_j \qquad (3.2.1) \\
&= tr\,[\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}] \qquad (3.2.2) \\
&= tr\,[\mathbf{C}\,\mathbf{R}^T] \qquad (3.2.3)
\end{aligned}
$$

dove nella (3.2.1) si è utilizzata l'espressione vettoriale (2.8.5) della distanza distribuzionale, mentre nelle (3.2.2) e (3.2.3) tr indica la funzione *traccia* (APP. A), ossia la somma degli elementi diagonali di una matrice quadrata e $\mathbf{D}_{\overline{\mathbf{r}}}$ è la matrice diagonale (2.6.5) di ordine $J \times J$ delle masse dei J profili colonna.

L'equivalenza tra le espressioni (3.2.1) e (3.2.2) può mostrarsi tenendo presente le proprietà della funzione *traccia* (APP. A) ed il fatto che l'inerzia è un numero che può quindi essere considerato come una matrice di ordine 1×1, per cui

$$
\begin{aligned}
In_{\mathbf{0}} &= \sum_{j=1}^{J} \overline{r}_j \, \mathbf{c}_j^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{c}_j = tr \sum_{j=1}^{J} \overline{r}_j \, \mathbf{c}_j^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{c}_j \\
&= \sum_{j=1}^{J} \overline{r}_j \, tr\,[\mathbf{c}_j^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{c}_j] = \sum_{j=1}^{J} \overline{r}_j \, tr\,[\mathbf{c}_j \, \mathbf{c}_j^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}] \\
&= tr\Big[\Big(\sum_{j=1}^{J} \overline{r}_j \, \mathbf{c}_j \, \mathbf{c}_j^T\Big) \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\Big] = tr\,[\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}].
\end{aligned}
$$

L'equivalenza tra le espressioni (3.2.2) e (3.2.3) deriva dal fatto che la matrice $\mathbf{R}$ dei profili delle righe, introdotta nella Sez. 2.6, può esprimersi in

---

ma ad un dato baricentro corrisponde un numero finito di configurazioni diverse (Sez. X).

funzione dei profili delle colonne e delle loro masse

$$\mathbf{R} = \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}} \tag{3.2.4}$$

perché il suo generico elemento, per ogni $i = 1, 2, \ldots, I$ e ogni $j = 1, 2, \ldots, J$ è

$$r_{ij} = \frac{n_{ij}}{n_{i+}} = \frac{1}{\frac{n_{i+}}{n_{++}}}\,\frac{n_{ij}}{n_{+j}}\,\frac{n_{+j}}{n_{++}} = \frac{1}{\overline{c}_i}\,c_{ij}\,\overline{r}_j.$$

Analoga espressione può scriversi per i profili delle colonne

$$\mathbf{C} = \mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{R}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}} \tag{3.2.5}$$

dove $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$ è la matrice diagonale (2.6.5) di ordine $J \times J$ degli inversi delle masse $\overline{r}_j$ dei profili colonna. Dalle (3.2.4) e (3.2.5) si ricava che le due matrici dei profili sono legate tra loro e alla matrice di contingenza dalla relazione

$$\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{R} = \mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}} = \frac{1}{n_{++}}\,\mathbf{N} \tag{3.2.6}$$

ed esprime tre diversi modi di rappresentare la matrice delle frequenze *relative*, il cui elemento generico è n_{ij}/n_{++}.

L'inerzia può esprimersi anche tramite gli elementi della matrice di contingenza o dei profili. Queste due forme sono le più frequentemente utilizzate. Dalla definizione (3.2.1) si ottiene

$$In_{\mathbf{0}} = \sum_{j=1}^{J} \overline{r}_j \sum_{i=1}^{I} \frac{c_{ij}^2}{\overline{c}_i} = \sum_{j=1}^{J} \frac{n_{+j}}{n_{++}} \sum_{i=1}^{I} \frac{n_{ij}^2}{n_{+j}^2}\,\frac{n_{++}}{n_{i+}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}^2}{n_{i+}\,n_{+j}} \tag{3.2.7}$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}}{n_{i+}}\,\frac{n_{ij}}{n_{+j}} = \sum_{i=1}^{I}\sum_{j=1}^{J} r_{ij}\,c_{ij} \tag{3.2.8}$$

3.3 - Inerzia riferita al baricentro

L'inerzia della nuvola quando viene riferita al baricentro assume una espressione un po' più complicata, perché vale

$$In_{\overline{\mathbf{c}}} = \sum_{j=1}^{J} \overline{r}_j\, d_D^2(\mathbf{c}_j, \overline{\mathbf{c}}) = \sum_{j=1}^{J} \overline{r}_j \sum_{i=1}^{I} \frac{(c_{ij} - \overline{c}_i)^2}{\overline{c}_i} \tag{3.3.1}$$

$$= \sum_{j=1}^{J} \overline{r}_j\,(\mathbf{c}_j - \overline{\mathbf{c}})^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,(\mathbf{c}_j - \overline{\mathbf{c}}) \tag{3.3.2}$$

$$= tr\left[(\mathbf{C} - \overline{\mathbf{C}})\,\mathbf{D}_{\overline{\mathbf{r}}}\,(\mathbf{C} - \overline{\mathbf{C}})^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\right] \tag{3.3.3}$$

dove $\overline{\mathbf{C}}$ è la matrice di ordine $I \times J$ in cui ogni colonna è $\overline{\mathbf{c}}$.

L'inerzia riferita al baricentro, introdotta qui in un ambito puramente geometrico, coincide con la *varianza* in Statistica e, quando è intesa come indice di connessione di una matrice di contingenza, prende il nome di *contingenza quadratica media relativa*, indicata abitualmente con Φ^2.

L'inerzia rispetto al baricentro può essere espressa in forma equivalente tramite le distanze distribuzionali che intercorrono tra i J profili della nuvola. Infatti, dalla (3.3.2) si ottiene

$$\begin{aligned} In_{\overline{\mathbf{c}}} &= \sum_{j=1}^{J} \overline{r}_j \, (\mathbf{c}_j - \overline{\mathbf{c}})^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, (\mathbf{c}_j - \overline{\mathbf{c}}) \\ &= \sum_{j=1}^{J} \overline{r}_j \, (\mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_j + \overline{\mathbf{c}}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \overline{\mathbf{c}} - 2 \, \mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \overline{\mathbf{c}}) \end{aligned}$$

dove si è tenuto presente la proprietà (2.5.1) del prodotto scalare. Inoltre, per la (2.8.6) $\overline{\mathbf{c}}$ ha lunghezza $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitaria e, come in (2.6.3), è la media ponderata dei J profili, per cui

$$\overline{\mathbf{c}} = \sum_{k=1}^{J} \overline{r}_k \, \mathbf{c}_k$$

e quindi

$$\begin{aligned} In_{\overline{\mathbf{c}}} &= \sum_{j=1}^{J} \overline{r}_j \left(\mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_j + 1 - 2 \, \mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \sum_{k=1}^{J} \overline{r}_k \mathbf{c_k} \right) \\ &= \sum_{j=1}^{J} \overline{r}_j \left(\mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_j + 1 + \sum_{k=1}^{J} \overline{r}_k (- 2 \, \mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c_k}) \right) \end{aligned}$$

La distanza distribuzionale tra due punti $\mathbf{c}_j$ e $\mathbf{c}_k$ è data dalla (2.8.2), che può anche esprimersi come si è visto sopra,

$$d_D^2(\mathbf{c}_j, \mathbf{c}_k) = \mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_j + \mathbf{c}_k^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_k - 2 \, \mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_k.$$

Ricavando

$$-2 \, \mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_k = d_D^2(\mathbf{c}_j, \mathbf{c}_k) - \mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_j - \mathbf{c}_k^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} c_k$$

sostituendo, e tenendo presente che la massa complessiva $\sum_k \overline{r}_k = 1$, si

ottiene

$$In_{\overline{\mathbf{c}}} = \sum_{j=1}^{J} \overline{r}_j \left(\mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_j + 1 + \sum_{k=1}^{J} \overline{r}_k d_D^2(\mathbf{c}_j, \mathbf{c}_k) \right.$$
$$\left. - \mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_j \sum_{k=1}^{J} \overline{r}_k - \sum_{k=1}^{J} \overline{r}_k \, \mathbf{c}_k \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_k \right)$$
$$= \sum_{j=1}^{J} \overline{r}_j \left(1 + \sum_{k=1}^{J} \overline{r}_k d_D^2(\mathbf{c}_j, \mathbf{c}_k) - In_{\mathbf{0}} \right)$$

e, dal momento che nella prossima Sez. 3.4 verrà mostrato che $In_{\mathbf{0}} = In_{\overline{\mathbf{c}}} + 1$, risulta finalmente

$$In_{\overline{\mathbf{c}}} = \frac{1}{2} \sum_{j=1}^{J} \sum_{k=1}^{J} \overline{r}_j \, \overline{r}_k \, d_D^2(\mathbf{c}_j, \mathbf{c}_k). \tag{3.3.4}$$

L'inerzia può quindi essere espressa anche tramite i quadrati delle distanze tra punti della nuvola, ponderate con le masse dei due punti.

Le inerzie rispetto al baricentro e all'origine dei profili colonna della matrice $\mathbf{C}$ *Spettacoli-3* di ordine 3×8 riportata nella TAV. 14, risultano

$$In_{\overline{\mathbf{c}}} = \sum_{i=1}^{3} \sum_{j=1}^{8} \overline{r}_j \, \frac{(c_{ij} - \overline{c}_i)^2}{\overline{c}_i} = 0.007, \qquad In_{\mathbf{0}} = \sum_{i=1}^{3} \sum_{j=1}^{8} \overline{r}_j \, \frac{c_{ij}^2}{\overline{c}_i} = 1.007.$$

Come si vede, l'inerzia rispetto al baricentro risulta *inferiore* dell'unità a quella riferita all'origine. Questo risultato è conseguenza di un importante teorema sulle proprietà dell'inerzia, oggetto della Sezione che segue.

3.4 - Teorema di Huygens [1]

Nell'espressione (3.3.3) la matrice simmetrica di ordine $I \times I$

$$W_{\overline{\mathbf{c}}} = (\mathbf{C} - \overline{\mathbf{C}}) \, \mathbf{D}_{\overline{\mathbf{r}}} \, (\mathbf{C} - \overline{\mathbf{C}})^T \tag{3.4.1}$$

è detta *matrice d'inerzia* della nuvola di J punti rispetto al baricentro $\overline{\mathbf{c}}$. Il suo termine generale vale, per ogni $i, i' = 1, 2, \ldots, I$

$$w_{ii'} = \sum_{j=1}^{J} \overline{r}_j \, (c_{ij} - \overline{c}_i) \, (c_{i'j} - \overline{c}_{i'}).$$

In Statistica la matrice $W_{\overline{\mathbf{c}}}$, che è stata introdotta qui in un ambito puramente geometrico, prende il nome di matrice di *varianza e covarianza*.

---

[1] Christiaan Huygens (pron.: *hoichens*): L'Aia 1629, Leida, Parigi, Londra, L'Aia 1695.

Se $\mathbf{x}$ è un punto generico in $\Re^I$, si può sempre scrivere

$$\overline{\mathbf{c}} = \mathbf{x} + (\overline{\mathbf{c}} - \mathbf{x})$$

per cui, sostituendo nella (3.4.1) si ottiene

$$\begin{aligned} W_{\overline{\mathbf{c}}} &= [\mathbf{C} - (\mathbf{x} + (\overline{\mathbf{c}} - \mathbf{x}))\mathbf{1}_{\mathrm{J}}^T)]\, \mathbf{D}_{\overline{\mathbf{r}}}\, [\mathbf{C} - (\mathbf{x} + (\overline{\mathbf{c}} - \mathbf{x}))\mathbf{1}_{\mathrm{J}}^T)]^T \\ &= [(\mathbf{C} - \mathbf{x}\,\mathbf{1}_{\mathrm{J}}^T) - (\overline{\mathbf{c}} - \mathbf{x})\mathbf{1}_{\mathrm{J}}^T]\, \mathbf{D}_{\overline{\mathbf{r}}}\, [(\mathbf{C} - \mathbf{x}\,\mathbf{1}_{\mathrm{J}}^T)^T - \mathbf{1}_{\mathrm{J}}(\overline{\mathbf{c}} - \mathbf{x})^T] \\ &= (\mathbf{C} - \mathbf{x}\,\mathbf{1}_{\mathrm{J}}^T)\, \mathbf{D}_{\overline{\mathbf{r}}}\, (\mathbf{C} - \mathbf{x}\,\mathbf{1}_{\mathrm{J}}^T)^T \;-\; (\mathbf{C} - \mathbf{x}\,\mathbf{1}_{\mathrm{J}}^T)\, \mathbf{D}_{\overline{\mathbf{r}}}\, \mathbf{1}_{\mathrm{J}}\, (\overline{\mathbf{c}} - \mathbf{x})^T \\ &\quad - (\overline{\mathbf{c}} - \mathbf{x})\, \mathbf{1}_{\mathrm{J}}^T\, \mathbf{D}_{\overline{\mathbf{r}}}\, (\mathbf{C} - \mathbf{x}\,\mathbf{1}_{\mathrm{J}}^T)^T \;+\; (\overline{\mathbf{c}} - \mathbf{x})\, \mathbf{1}_{\mathrm{J}}^T\, \mathbf{D}_{\overline{\mathbf{r}}}\, \mathbf{1}_{\mathrm{J}} (\overline{\mathbf{c}} - \mathbf{x})^T. \end{aligned}$$

Il primo termine dell'ultima espressione è la matrice d'inerzia $W_{\mathbf{x}}$ riferita al punto $\mathbf{x}$, mentre gli altri tre termini valgono tutti $(\overline{\mathbf{c}} - \mathbf{x})(\overline{\mathbf{c}} - \mathbf{x})^T$ come risulta eliminando le parentesi e tenendo conto della definizione (2.6.4) di baricentro e del fatto, noto fin dalla Sez. 1.8, che le masse dei profili hanno somma 1

$$\mathbf{C}\, \mathbf{D}_{\overline{\mathbf{r}}}\, \mathbf{1}_{\mathrm{J}}^T = \overline{\mathbf{c}} \qquad \text{e} \qquad \mathbf{1}_{\mathrm{J}}^T\, \mathbf{D}_{\overline{\mathbf{r}}}\, \mathbf{1}_{\mathrm{J}} = \sum_{j=1}^{J} \overline{r}_j = 1.$$

Postmoltiplicando per $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$ e considerando la traccia, si ottiene

$$tr[W_{\overline{\mathbf{c}}}\, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}] = tr[W_{\mathbf{x}}\, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}] \;-\; tr[(\overline{\mathbf{c}} - \mathbf{x})^T\, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\, (\overline{\mathbf{c}} - \mathbf{x})]$$

per cui

$$In_{\mathbf{x}} \;=\; In_{\overline{\mathbf{c}}} \;+\; d_D^2(\overline{\mathbf{c}}, \mathbf{x}). \tag{3.4.2}$$

È questo il *teorema di Huygens*: l'inerzia della nuvola relativa ad un punto $\mathbf{x}$ di $\Re^I$ può essere scomposta in due parti: la prima $In_{\overline{\mathbf{c}}}$ che rappresenta l'inerzia relativa al baricentro $\overline{\mathbf{c}}$ e la seconda che rappresenta invece l'inerzia rispetto a $\mathbf{x}$ di tutta la massa (che per i profili vale 1) concentrata nel baricentro (TAV. 19). Di conseguenza $In_{\mathbf{x}}$ è minima quando $\mathbf{x} = \overline{\mathbf{c}}$, ossia quando l'inerzia è calcolata rispetto al baricentro. Allora l'inerzia può raggiungere il suo minimo assoluto, ossia 0, e questo quando i J punti collassano tutti nel baricentro. In tal caso, per la (3.3.1), deve essere $c_{ij} - \overline{c}_i = 0$ dal momento che le masse $\overline{r}_j$ e $\overline{c}_i$ non possono essere nulle (Sez. 1.4, nota 1) e per ogni $i = 1, 2, \ldots, I$ e per ogni $j = 1, 2, \ldots, J$, deve necessaraimente essere

$$\frac{c_{ij}}{\overline{c}_i} = 1.$$

Nella Sez. 1.10 si è visto che questa condizione indica una situazione di completa omogeneità: tutti i profili erano eguali ed eguali al profilo medio

$$\mathbf{c}_1 = \mathbf{c}_2 = \ldots = \mathbf{c}_j = \ldots = \mathbf{c}_J = \overline{\mathbf{c}}.$$

Ne consegue allora che alla situazione estrema di completa omogeneità tra profili colonna corrisponde, da un punto di vista geometrico, la configurazione degenere in cui tutti i J punti sono concentrati nel loro baricentro $\overline{\mathbf{c}}$, che risulta avere massa 1.

Conseguenza importante del teorema di Huygens è che quando il punto $\mathbf{x}$ è l'origine $\mathbf{0}_{\mathrm{I}}$ della base di riferimento, la cui distanza distribuzionale dal baricentro, per la (2.8.5), vale 1, si ha

$$In_{\mathbf{0}} = In_{\overline{\mathbf{c}}} + 1. \tag{3.4.3}$$

In particolare, se l'inerzia rispetto al baricentro è espressa, come in (3.3.4), tramite le distanze tra punti, ne deriva che

$$In_{\mathbf{0}} = \frac{1}{2} \sum_{j=1}^{J} \sum_{k=1}^{J} \overline{r}_j \, \overline{r}_k \, d_D^2(\mathbf{c}_j, \mathbf{c}_k) + 1. \tag{3.4.4}$$

3.5 - Riduzione della dimensionalità

Grazie alla definizione d'inerzia si è in grado ora di quantificare la dispersione geometrica dei J punti della nuvola in $\Re^I$, e, per il teorema di Huygens, è al baricentro che l'inerzia deve essere riferita in quanto l'inerzia rispetto a qualsiasi altro punto può da quella derivarsi. Si dispone adesso di un *criterio* obiettivo per identificare il sottospazio ottimale, visivamente comprensibile, nel quale la forma della nuvola appaia "al meglio". Per le considerazioni fatte nella Sez. 3.1, 'al meglio' va inteso ora come massima inerzia. Se, per iniziare, si prende in considerazione un sottospazio monodimensionale, la retta ottimale sarà quella che passa per il baricentro della nuvola e tale che su di essa le *proiezioni* dei J punti, forniscano un valore dell'inerzia, rispetto al baricentro, più vicino possibile a quello $In_{\overline{\mathbf{c}}}$ che ha la nuvola nel suo spazio ambiente $\Re^I$.

Tuttavia, come si è visto nella Sez. 3.2, l'inerzia assume un'espressione molto più semplice quando è calcolata rispetto all'origine. Ora, i risultati d'interesse che verranno ottenuti nelle Sezioni seguenti sono indifferenti al fatto che si usi come criterio l'inerzia riferita all'origine o al baricentro. Questa è una importante proprietà dell'Analisi delle Corrispondenze che verrà evidenziata nella Sez. 3.14 e che deriva dalla $\mathbf{D}_{\overline{\mathbf{c}}}^{-1}$-ortogonalità del vettore $\overline{\mathbf{c}}$ all'iperpiano del simplesso contenente i punti della nuvola, come si è visto nella Sez. 2.8. Di conseguenza, per agevolare il lettore nel seguire lo sviluppo dei calcoli, verrà utilizzato da ora e fino alla Sez. 3.14, il criterio formalmente più semplice, quello dell'inerzia riferita all'origine.

Sia dunque $\mathbf{u}$ un vettore generico di ordine I con origine in $\mathbf{0_I}$, origine della base canonica di $\Re^I$, che si vuole di lunghezza $\mathbf{D_{\overline{c}}^{-1}}$-unitaria, ossia tale che

$$d_D^2(\mathbf{u}, \mathbf{0_I}) = \mathbf{u}^T \mathbf{D_{\overline{c}}^{-1}} \mathbf{u} = \sum_{i=1}^{I} \frac{u_i^2}{\overline{c}_i} = 1 \tag{3.5.1}$$

perché quella che si sta iniziando a cercare è una nuova base, questa volta $\mathbf{D_{\overline{c}}^{-1}}$-*ortonormale* [1], relativamente alla quale la forma della nuvola appaia 'al meglio'. Inoltre il vincolo sulla lunghezza di $\mathbf{u}$ impedisce che le componenti possano assumere valori abnormi, tali da rendere assurdamente grande l'inerzia delle proiezioni della nuvola, come si vedrà nella Sez. 3.9. Ora, se il vettore $\mathbf{u}$ ha lunghezza $\mathbf{D_{\overline{c}}^{-1}}$-unitaria, allora la proiezione $\mathbf{D_{\overline{c}}^{-1}}$-ortogonale del punto $\mathbf{c}_j$ sulla retta individuata da $\mathbf{u}$ è, per definizione, il punto $h(\mathbf{c}_j)\,\mathbf{u}$ della retta più vicino a $\mathbf{c}_j$, vicino nel senso della distanza distribuzionale, e dove, Sez. 2.5,

$$h(\mathbf{c}_j) = \mathbf{c}_j^T \mathbf{D_{\overline{c}}^{-1}} \mathbf{u} = \sum_{i=1}^{I} c_{ij} \frac{1}{\overline{c}_i} u_i \tag{3.5.2}$$

è la sua coordinata su $\mathbf{u}$ (TAV. 20). Il valore assoluto $|h(\mathbf{c}_j)|$ è la distanza distribuzionale misurata lungo $\mathbf{u}$ tra la proiezione $h(\mathbf{c}_j)\,\mathbf{u}$ e l'origine $\mathbf{0_I}$. Prendendo in considerazione tutti i J punti della nuvola, le ascisse delle loro proiezioni su $\mathbf{u}$ si ottengono da un insieme di J prodotti scalari e costituiscono il vettore

$$\mathbf{h} = \mathbf{C}^T \mathbf{D_{\overline{c}}^{-1}} \mathbf{u} \tag{3.5.3}$$

dove

$$\mathbf{h} = \Big(h(\mathbf{c}_1) \quad h(\mathbf{c}_2) \quad \ldots \quad h(\mathbf{c}_j) \quad \ldots \quad h(\mathbf{c}_J) \Big)^T.$$

Il vettore $\mathbf{h}$ può essere immaginato come una "riduzione" monodimensionale della nuvola dei J profili. Per questo ad ogni proiezione $h(\mathbf{c}_j)\,\mathbf{u}$ viene assegnata la massa $\overline{r}_j$ che il punto $\mathbf{c}_j$ ha in $\Re^I$. La dispersione complessiva delle J proiezioni su $\mathbf{u}$, è misurata dall'inerzia che queste hanno rispetto all'origine, inerzia computata ora tramite le coordinate su $\mathbf{u}$

$$\begin{aligned} In_{\mathbf{0}}(\mathbf{u}) &= \sum_{j=1}^{J} \overline{r}_j \, d_D^2\Big(h(\mathbf{c}_j)\,\mathbf{u}, \mathbf{0_I}\Big) = \sum_{j=1}^{J} \overline{r}_j \, \Big(h(\mathbf{c}_j) - 0\Big)^2 \\ &= \mathbf{h}^T \mathbf{D_{\overline{r}}} \mathbf{h} \end{aligned} \tag{3.5.4}$$

---

[1] Un importante teorema dell'Algebra Lineare garantisce che ogni spazio euclideo possiede almeno una base ortonormale.

$$= \mathbf{u}^T \mathbf{D}_{\mathbf{c}}^{-\mathbf{1}} \mathbf{C} \mathbf{D}_{\overline{\mathbf{r}}} \mathbf{C}^T \mathbf{D}_{\mathbf{c}}^{-\mathbf{1}} \mathbf{u} \tag{3.5.5}$$

$$= \mathbf{u}^T \mathbf{D}_{\mathbf{c}}^{-\mathbf{1}} \mathbf{C} \mathbf{R}^T \mathbf{u} \tag{3.5.6}$$

grazie alla (3.5.3) e alla (3.2.4). Si tratta quindi di un'espressione del secondo ordine nelle componenti di $\mathbf{u}$ perché la (3.5.6) si può esplicitare in

$$In_{\mathbf{0}}(\mathbf{u}) = \sum_{i=1}^{I} \frac{1}{\overline{c}_i} \sum_{i'=1}^{I} u_i \, u_{i'} \sum_{j=1}^{J} c_{ij} \, r_{i'j}.$$

Il problema è adesso quello di individuare l'orientamento di $\mathbf{u}$, ossia di determinare le sue I componenti, in modo che l'inerzia delle proiezioni (3.5.6) sia la massima possibile e, di conseguenza la più prossima all'inerzia complessiva (3.2.1) della nuvola nel suo spazio ambiente $\Re^I$.

3.6 - Base ortogonale e ortonormale

Il lettore attento si sarà stupito vedendo che nel calcolo dell'inerzia delle proiezioni su $\mathbf{u}$ la distanza distribuzionale sia stata computata come semplice somma dei quadrati delle differenze tra coordinate, come per la distanza euclidea canonica, invece che come somma *pesata* dei quadrati delle differenze, come nella (2.8.5). Per comprenderne il motivo, occorre ritornare sul concetto di base di uno spazio euclideo. Limitando le osservazioni allo spazio dei profili $\Re^I$, si è visto nella Sez. 2.4, che *ogni* profilo $\mathbf{c}_j$ può essere espresso in modo univoco come combinazione lineare degli I vettori unitari $\mathbf{e}_i$ della base canonica con origine in $\mathbf{0}_{\mathrm{I}}$, ossia come somma di multipli dei vettori di base

$$\mathbf{c}_j = c_{1j} \, \mathbf{e}_1 + c_{2j} \, \mathbf{e}_2 + \ldots + c_{ij} \, \mathbf{e}_i + \ldots c_{Ij} \, \mathbf{e}_I = \sum_{i=1}^{I} c_{ij} \, \mathbf{e}_i$$

dove $c_{ij} \, \mathbf{e}_i$ è la proiezione ortogonale del profilo $\mathbf{c}_j$ sul vettore di base $\mathbf{e}_i$, c_{ij} la coordinata (ascissa) della proiezione su $\mathbf{e}_i$, e quindi anche la distanza della proiezione dall'origine. Nella Sez. 2.8, si è visto che questa base è $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale, ma non $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortonormale, perché, per $i, k = 1, 2, \ldots, I$ è

$$\mathbf{e}_i^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{e}_k = 0 \qquad i \neq k$$

$$\mathbf{e}_i^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{e}_i = 1/\overline{c}_i.$$

Ora, se $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_I$ è una altra base – sempre dello stesso spazio $\Re^I$ e sempre con origine in $\mathbf{0}_{\mathrm{I}}$ – questa però $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortonormale (in effetti è

la nuova base che verrà individuata nella Sez. 3.12) per la quale, quindi, per $i, k = 1, 2, \ldots, I$ risulta

$$\mathbf{u}_i^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{u}_k = 0 \qquad\qquad i \neq k \tag{3.6.1}$$

$$\mathbf{u}_i^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{u}_i = 1 \tag{3.6.2}$$

ogni profilo $\mathbf{c}_j$ di $\Re^I$, può anche esprimersi come combinazione lineare dei vettori di questa nuova base, per cui

$$\mathbf{c}_j = h_{1j} \, \mathbf{u}_1 + h_{2j} \, \mathbf{u}_2 + \ldots + h_{ij} \, \mathbf{u}_i + \ldots h_{Ij} \, \mathbf{u}_I = \sum_{i=1}^{I} h_{ij} \, \mathbf{u}_i$$

dove h_{ij} è la coordinata e $|h_{ij}|$ la distanza dall'origine, della proiezione $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale di $\mathbf{c}_j$ sul vettore di base $\mathbf{u}_i$. Esiste ovviamente un legame tra le coordinate c_{ij} nella base canonica e h_{ij} nella nuova, legame che verrà esplicitato nel prossimo Capitolo.

Ciò premesso, in $\Re^I$ la distanza del profilo $\mathbf{c}_j$ dall'origine è la distanza distribuzionale $d_D^2(\mathbf{c}_j, \mathbf{0}_\mathrm{I})$ espressa tramite le coordinate c_{ij} come nella (2.8.5), ma che, per quanto appena visto, può essere anche espressa tramite le coordinate h_{ij}

$$\begin{aligned} d_D^2(h(\mathbf{c}_j)\,\mathbf{u}, \mathbf{0}_\mathrm{I}) &= \sum_{i=1}^{I} \frac{1}{\overline{c}_i} \, c_{ij}^2 = \mathbf{c}_j^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{c}_j = \Big(\sum_{i=1}^{I} h_{ij} \, \mathbf{u}_i^T \Big) \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \Big(\sum_{k=1}^{I} h_{kj} \, \mathbf{u}_k \Big) \\ &= \sum_{i=1}^{I} \sum_{k=1}^{I} h_{ij} \, h_{kj} \mathbf{u}_i^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{u}_k = \sum_{i=1}^{I} h_{ij} \, h_{ij} \\ &= \sum_{i=1}^{I} h_{ij}^2 \end{aligned} \tag{3.6.3}$$

per la proprietà distributiva del prodotto di matrici e per la $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortonormalità della nuova base, come in (3.6.1) e (3.6.2). Pertanto, nello spazio dei profili $\Re^I$, se i profili sono riferiti agli I vettori $\mathbf{e}_i$ della base $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-*ortogonale*, la distanza distribuzionale computata tramite le coordinate c_{ij} si esprime come in (2.8.5), ma se, sempre nello *stesso* spazio, i profili sono riferiti agli I vettori $\mathbf{u}_i$ della base $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-*ortonormale*, la distanza, computata ora tramite le coordinate h_{ij}, assume l'espressione euclidea canonica (3.6.3).

Questo risultato è del tutto generale, per cui soltanto se la base di uno spazio euclideo è ortonormale, la distanza tra due punti è la distanza euclidea canonica.

3.7 - Aspetti geometrici e matematici

Da un punto di vista geometrico, il problema enunciato nella Sez. 3.5, si pone in questi termini: un vettore $\mathbf{u}$ di ordine $I \leq J$ è $\mathbf{D_c^{-1}}$-unitario ed ha per origine il punto $\mathbf{0}_\mathrm{I}$, ossia la stessa origine degli I vettori $\mathbf{e}_i$ della base canonica. Si può immaginare $\mathbf{u}$ come 'incernierato' in $\mathbf{0}_\mathrm{I}$, ma orientabile in tutte le direzioni. Ciascun orientamento di $\mathbf{u}$, individua una retta passante per $\mathbf{0}_\mathrm{I}$ sulla quale si proiettano i J punti della nuvola. Il valore che assume l'inerzia rispetto all'origine di queste J proiezioni, dipende soltanto dall'orientamento di $\mathbf{u}$. Si tratta allora di individuare quell'orientamento – ossia le I componenti di $\mathbf{u}$ – che rende l'inerzia delle proiezioni $In_\mathbf{0}(\mathbf{u})$, come nella Sez. 3.5, la *massima possibile*, ossia più prossima a $In_\mathbf{0}$, l'inerzia complessiva della nuvola, calcolata nella Sez. 3.2. Dalla TAV. 20, si vede che ciò significa massimizzare la somma dei J quadrati delle distanze $d_D^2(h(\mathbf{c}_j)\,\mathbf{u}, \mathbf{0}_\mathrm{I})$ delle proiezioni dall'origine, ciascuna proiezione dotata della massa $\overline{r}_j$. Peraltro, applicando il teorema di Pitagora a ciascuno dei J triangoli $\mathbf{D_c^{-1}}$-rettangoli, moltiplicando per la massa e sommando su tutti i triangoli, si ottiene la relazione

$$\sum_{j=1}^{J} \overline{r}_j\, d_D^2(\mathbf{c}_j, \mathbf{0}_\mathrm{I}) = \sum_{j=1}^{J} \overline{r}_j\, d_D^2(\mathbf{c}_j, h(\mathbf{c}_j)\mathbf{u}) \;+\; \sum_{j=1}^{J} \overline{r}_j\, d_D^2(h(\mathbf{c}_j)\mathbf{u}, \mathbf{0}_\mathrm{I}).$$

Le distanze sono quelle distribuzionali dal momento che le coordinate dei punti sono riferite alla base canonica, l'unica per il momento disponibile. Ora, la somma delle distanze al primo membro è una costante, perché queste dipendono esclusivamente dalla configurazione della nuvola, che è assegnata e non varia. Di conseguenza, rendere massima l'ultima somma al secondo membro equivale a rendere *minima* l' altra somma, ossia a ricercare quel vettore $\mathbf{u}$, $\mathbf{D_c^{-1}}$-unitario, per cui

$$\sum_{j=1}^{J} \overline{r}_j\, d_D^2(\mathbf{c}_j, h(\mathbf{c}_j)\mathbf{u}) = \mathit{minimo}. \tag{3.7.1}$$

Questo criterio è quello classico dei *minimi quadrati*, qui nella versione dei minimi quadrati *ponderati*. Il vettore $\mathbf{u}$ ottimale è dunque anche quello che individua la retta che passa *più vicina* ai J punti della nuvola. Più vicina nel senso della (3.7.1), ossia che rende minima la somma dei quadrati delle distanze tra i profili e le loro proiezioni $\mathbf{D_c^{-1}}$-ortogonali su $\mathbf{u}$, ciascuna distanza ponderata con la massa del profilo.

Dal punto di vista matematico si tratta invece di un problema di ot-

timizzazione non lineare vincolata [1], un classico problema di algebra lineare, che può essere risolto col metodo dei *moltiplicatori di Lagrange* [2]. Anche se non più utilizzato nei programmi di calcolo a causa della sua scarsa efficienza, questo metodo permette di comprendere in modo più semplice come si perviene all'individuazione della direzione ottimale. Ad esso è interamente dedicata la Sezione seguente. Il lettore non interessato ai dettagli matematici può passare direttamente alla Sez. 3.9.

3.8 - Metodo dei moltiplicatori di Lagrange

La ricerca di un massimo o di un minimo di una funzione di una sola variabile è un problema matematico ben noto. Se $f(x)$ è una funzione continua con derivata continua in un intervallo finito, i massimi e i minimi relativi di $f(x)$, all'*interno* dell'intervallo, sono raggiunti soltanto nei punti in cui la derivata $f'(x) = 0$. I massimi e i minimi sono detti *valori estremi* di $f(x)$ e i valori di x in cui la funzione raggiunge tali valori estremi sono detti *punti critici*.

Questa condizione si estende a funzioni di più variabili, per esempio di tre, e perché $f(x_1, x_2, x_3)$ abbia un estremo relativo nel punto critico (x_1, x_2, x_3), all'interno del dominio di definizione, è necessario [3] che ivi ne siano nulle le derivate parziali

$$\begin{cases} \dfrac{\partial f}{\partial x_1} = 0 \\ \dfrac{\partial f}{\partial x_2} = 0 \\ \dfrac{\partial f}{\partial x_3} = 0. \end{cases}$$

Queste condizioni costituiscono 3 equazioni nelle 3 incognite (x_1, x_2, x_3). Il metodo di sostituzione permette di trovare i valori critici e quindi l'estremo relativo, almeno quando dalle equazioni è possibile esplicitare formalmente le

---

[1] Il problema è non lineare perché nell'espressione di $In_{\mathbf{0}}(\mathbf{u})$ ottenuta nella Sez. 3.5, le incognite, le I componenti di $\mathbf{u}$, vi compaiono al secondo grado ed è vincolato perché $\mathbf{u}$ deve avere lunghezza $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitaria, come nella (3.5.1).

[2] Giuseppe Luigi Lagrange: Torino 1736, Berlino, Parigi 1813.

[3] Qui ci si limita a cercare le condizioni necessarie per un estremo, senza preoccuparsi se le condizioni sono sufficienti, né se forniscono massimi o minimi.

tre incognite. Si procede ricavando da una delle equazioni una delle incognite in funzione delle altre due, ad es. $x_3 = x_3(x_1, x_2)$, che viene sostituita nelle altre due equazioni. Restano così due equazioni in due incognite. Da una di queste si ricava $x_2 = x_2(x_1)$ che si sostituisce nella rimanente equazione. In questa è ora presente la sola variabile x_1, per cui è possibile ricavare il valore critico di x_1. Questo si sostituisce in $x_2 = x_2(x_1)$ che fornisce il valore critico di x_2. Infine i due valori di x_1 e x_2 si sostituiscono in $x_3 = x_3(x_1, x_2)$, il che consente di ricavare anche il valore critico di x_3. Si è determinato così il solo punto (o i soli punti) in cui la funzione può avere un estremo relativo, il cui valore si determina sostituendo in $f(x_1, x_2, x_3)$ i valori critici ottenuti.

Il problema posto nella Sez. 3.5 è però più complesso perché sul punto critico c'è un *vincolo* che ha l'effetto di diminuire il numero delle variabili e di restringere la regione in cui il punto critico può trovarsi. La funzione $In_{\mathbf{0}}(\mathbf{u})$ di cui si deve cercare un estremo è una forma quadratica con matrice simmetrica ($a_{ij} = a_{ji}$)

$$\begin{aligned} f(u_1, u_2, u_3) &= \sum_{i=1}^{3} \sum_{i'=1}^{3} a_{ii'}\, u_i\, u_{i'} \\ &= a_{11}\, u_1^2 + a_{22} u_2^2 + a_{33} u_3^2 + 2a_{12} u_1 u_2 + 2a_{13} u_1 u_3 + 2a_{23} u_2 u_3 \end{aligned}$$

e il vincolo è costituito da una forma quadratica con matrice diagonale in cui $d_{ii} = 1/\overline{c}_i$

$$\begin{aligned} v(u_1, u_2, u_3) &= \sum_{i=1}^{3} d_{ii}\, u_i^2 - 1 = 0 \\ &= d_{11}\, u_1^2 + d_{22} u_2^2 + d_{33} u_3^2 - 1 = 0. \end{aligned}$$

Ora, nell'equazione del vincolo, si può considerare una delle variabili, ad esempio u_3, come definita implicitamente in funzione delle altre, ossia $u_3 = u_3(u_1, u_2)$. La funzione diventa $f(u_1, u_2, u_3(u_1, u_2))$ e dipende ora soltanto da u_1 e da u_2, per cui i suoi estremi si devono cercare dove

$$\frac{\partial f}{\partial u_1} + \frac{\partial f}{\partial u_3}\,\frac{\partial u_3}{\partial u_1} = 0 \tag{3.8.1}$$

$$\frac{\partial f}{\partial u_2} + \frac{\partial f}{\partial u_3}\,\frac{\partial u_3}{\partial u_2} = 0 \tag{3.8.2}$$

D'altra parte $\partial u_3/\partial u_1$ e $\partial u_3/\partial u_2$ si possono anche ottenere derivando il vincolo

$$\frac{\partial v}{\partial u_1} + \frac{\partial v}{\partial u_3}\frac{\partial u_3}{\partial u_1} = 0 \tag{3.8.3}$$

$$\frac{\partial v}{\partial u_2} + \frac{\partial v}{\partial u_3}\frac{\partial u_3}{\partial u_2} = 0 \tag{3.8.4}$$

e, tenendo presente anche l'equazione del vincolo,

$$v(u_1, u_2, u_3) = 0 \tag{3.8.5}$$

si dispone di 5 equazioni, dalle quali si possono eliminare le due incognite aggiuntive $\partial u_3/\partial u_1$ e $\partial u_3/\partial u_2$ riconducendosi così a tre equazioni nelle tre incognite u_1, u_2 e u_3, per cui ci si riporta al caso senza vincolo presentato sopra, perché del vincolo si è già tenuto conto con la sostituzione eseguita.

Il metodo di Lagrange è in sostanza una notevole semplificazione del metodo di sostituzione per renderlo più pratico, particolarmente quando funzione e vincolo sono, come nel presente caso, forme quadratiche con matrice simmetrica. Infatti Lagrange osserva che se delle 5 equazioni si considerano la prima e la terza

$$\frac{\partial f}{\partial u_1} + \frac{\partial f}{\partial u_3}\frac{\partial u_3}{\partial u_1} = 0$$
$$\frac{\partial v}{\partial u_1} + \frac{\partial v}{\partial u_3}\frac{\partial u_3}{\partial u_1} = 0$$

è possibile eliminare l'incognita aggiuntiva $\partial u_3/\partial u_1$, perché queste due condizioni possono valere simultaneamente soltanto se c'è proporzionalità, ossia se esiste una relazione lineare tra i coefficienti nelle colonne. Deve quindi esistere un numero reale λ_1 tale che

$$\frac{\partial f}{\partial u_1} = \lambda_1 \frac{\partial v}{\partial u_1} \tag{3.8.6}$$
$$\frac{\partial f}{\partial u_3} = \lambda_1 \frac{\partial v}{\partial u_3}. \tag{3.8.7}$$

Passando ora a considerare la seconda e la quarta equazione

$$\frac{\partial f}{\partial u_2} + \frac{\partial f}{\partial u_3}\frac{\partial u_3}{\partial u_2} = 0$$
$$\frac{\partial v}{\partial u_2} + \frac{\partial v}{\partial u_3}\frac{\partial u_3}{\partial u_2} = 0$$

si vede ancora che perché queste due condizioni valgano simultaneamente per $\partial u_3/\partial u_2$, occorre che

$$\frac{\partial f}{\partial u_2} = \lambda_2 \frac{\partial v}{\partial u_2} \tag{3.8.8}$$
$$\frac{\partial f}{\partial u_3} = \lambda_2 \frac{\partial v}{\partial u_3}. \tag{3.8.9}$$

Le condizioni (3.8.7) e (3.8.9) mostrano che i due coefficienti di proporzionalità devono essere eguali, ossia che $\lambda_1 = \lambda_2 = \lambda$, per cui le condizioni necessarie (3.8.6), (3.8.8) e (3.8.9), insieme a quella (3.8.5) del vincolo forniscono

$$\begin{cases} \dfrac{\partial}{\partial u_1}(f(u_1,u_2,u_3) - \lambda\, v(u_1,u_2,u_3)) = 0 \\ \dfrac{\partial}{\partial u_2}(f(u_1,u_2,u_3) - \lambda\, v(u_1,u_2,u_3)) = 0 \\ \dfrac{\partial}{\partial u_3}(f(u_1,u_2,u_3) - \lambda\, v(u_1,u_2,u_3)) = 0 \\ v(u_1,u_2,u_3) = 0. \end{cases} \tag{3.8.10}$$

Si hanno così 4 equazioni, che permettono di ottenere il valore, o i valori, dell'incognita aggiuntiva λ, *in corrispondenza dei quali* si ricavano i valori critici di u_1, u_2, u_3 che forniscono il valore dell'estremo, o degli estremi, della funzione $f(u_1,u_2,u_3)$. Si hanno quindi valori estremi *soltanto* in corrispondenza di valori speciali del parametro λ, la cui determinazione è dunque pregiudiziale alla ricerca dei valori critici della funzione.

Il parametro λ è detto *moltiplicatore di Lagrange*, mentre la *funzione di Lagrange* è definita come

$$\mathcal{L}(u_1,u_2,u_3,\lambda) \stackrel{\text{def}}{=} f(u_1,u_2,u_3) - \lambda\, v(u_1,u_2,u_3).$$

Ora, le condizioni necessarie perché la funzione di Lagrange abbia un estremo, si ottengono, per quanto visto sopra, annullandone le derivate parziali rispetto ai 4 parametri, supposti indipendenti. Queste condizioni coincidono con le (3.8.10).

Perciò le condizioni necessarie (3.8.10) per un estremo *vincolato* della funzione $f(u_1,u_2,u_3)$ sono formalmente le stesse per un estremo *non vincolato* della funzione di Lagrange $\mathcal{L}(u_1,u_2,u_3,\lambda)$. Difatti, quando sono soddisfatte le condizioni di vincolo, il termine $\lambda\, v(u_1,u_2,u_3)$ è nullo e la funzione di Lagrange si riduce alla funzione originale. In pratica, quindi, per individuare gli estremi vincolati della funzione, basta costruire la funzione di Lagrange ed annullarne le 3 derivate parziali, rispetto alle 3 incognite u_1, u_2, u_3. È importante sottolineare comunque, che *non* è di questa funzione che si cerca un estremo: essa è costruita e derivata soltanto per rendere più agevole l'individuazione degli estremi vincolati di $f(u_1,u_2,u_3)$.

Se $\mathbf{u} = (u_1 \quad u_2 \quad u_3)^T$ è il vettore delle incognite, la funzione di

Lagrange in forma vettoriale si scrive

$$\mathcal{L}(\mathbf{u}, \lambda) = f(\mathbf{u}) - \lambda\, v(\mathbf{u})$$

e le condizioni necessarie (3.8.10) per l'esistenza di un estremo vincolato di $f(\mathbf{u})$

$$\frac{\partial}{\partial \mathbf{u}}\, \mathcal{L}(\mathbf{u}, \lambda) = \mathbf{0}_3. \tag{3.8.11}$$

Nella Sez. 3.12. verrà mostrato come questi risultati si possano estendere a funzioni di I variabili con più vincoli indipendenti.

3.9 - Autovalori ed autovettori

Nella Sezione precedente si è visto che valori estremi di una funzione, come l'inerzia (3.5.6) delle proiezioni su un vettore

$$f(\mathbf{u}) = In_{\mathbf{0}}(\mathbf{u}) = \mathbf{u}^T \mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\, \mathbf{C}\mathbf{R}^T\, \mathbf{u}, \tag{3.9.1}$$

soggetta a un vincolo, come quello (3.5.1) che tale vettore sia di lunghezza $\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}$-unitaria

$$v(\mathbf{u}) = \mathbf{u}^T \mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\, \mathbf{u} - 1 = 0, \tag{3.9.2}$$

si trovano costruendo la funzione di Lagrange

$$\mathcal{L}(\mathbf{u}, \lambda) = f(\mathbf{u}) - \lambda\, v(\mathbf{u}) = \mathbf{u}^T \mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\, \mathbf{C}\mathbf{R}^T\, \mathbf{u} - \lambda\, (\mathbf{u}^T \mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\, \mathbf{u} - 1)$$

ed annullandone le derivate parziali rispetto alle I incognite: le componenti del vettore $\mathbf{u}$. L'estremo vincolato di $In_{\mathbf{0}}(\mathbf{u})$ va quindi ricercato in corrispondenza del punto $(\mathbf{u}, \lambda)$ in cui grazie, alla simmetria (APP. A) della matrice $\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\, \mathbf{C}\, \mathbf{R}^T$, si ha

$$\frac{\partial \mathcal{L}(\mathbf{u}, \lambda)}{\partial \mathbf{u}} = 2\, \mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\, \mathbf{C}\ \mathbf{R}^T\, \mathbf{u} - 2\, \lambda\, \mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\, \mathbf{u} = \mathbf{0}_{\mathrm{I}}.$$

Premoltiplicando ambo i membri per $\mathbf{D}_{\mathbf{c}}$ e semplificando, le *condizioni necessarie*[1] per un estremo vincolato di $In_{\mathbf{0}}(\mathbf{u})$ risultano

$$\mathbf{C}\, \mathbf{R}^T\, \mathbf{u} - \lambda\, \mathbf{u} = \mathbf{0}_{\mathrm{I}}. \tag{3.9.3}$$

Queste si possono esplicitare per ciascuna delle I componenti del vettore $\mathbf{u}$

$$\sum_{i'=1}^{I} \sum_{j=1}^{J} c_{ij}\, r_{i'j}\, u_{i'} - \lambda\, u_i = 0$$

---

[1] Le condizioni sono anche *sufficienti* perché si sa che una soluzione deve esistere, in quanto l'inerzia delle proiezioni deve assumere un qualche valore finito.

ossia

$$\sum_{i'=1}^{I} q_{ii'}\, u_{i'} - \lambda\, u_i = 0 \quad \text{dove} \quad q_{ii'} = \sum_{j=1}^{J} c_{ij} r_{i'j}$$

è l'elemento generico, *mai negativo*, della matrice quadrata di ordine $I \times I$ ottenuta dal prodotto delle due matrici dei profili

$$\mathbf{Q} = \mathbf{C}\,\mathbf{R}^T.$$

Si tratta di una matrice $\mathbf{D_c^{-1}}$-*simmetrica*, nel senso che

$$\mathbf{D_c^{-1}}\,\mathbf{Q} = (\mathbf{D_c^{-1}}\,\mathbf{Q})^T.$$

Queste matrici hanno un ruolo importante nell'Analisi delle Corrispondenze e nell'APP. B è mostrato come le proprietà delle matrici $\mathbf{D_c^{-1}}$-simmetriche siano una generalizzazione delle proprietà delle matrici simmetriche, per cui hanno autovalori reali ed autovettori $\mathbf{D_c^{-1}}$-ortogonali due a due. In particolare gli autovalori della matrice $\mathbf{C}\,\mathbf{R}^T$ non possono essere negativi, come si vedrà nella Sez. 3.15.

Ci si chiede ora sotto quali condizioni possano esistere un numero λ ed un vettore $\mathbf{u}$ tali che l'equazione (3.9.3) sia vera. La risposta risulta evidente se, grazie all'identità $\mathbf{I}\,\mathbf{u} = \mathbf{u}$, questa viene scritta nella forma equivalente

$$(\mathbf{C}\,\mathbf{R}^T - \lambda\,\mathbf{I})\,\mathbf{u} = \mathbf{0}_{\mathrm{I}}$$

dove $\mathbf{I}$ è la matrice identità di ordine $I \times I$ e $\lambda\,\mathbf{I}$ una matrice diagonale col numero λ in tutti gli elementi della diagonale principale (APP. A). Si tratta di un sistema lineare omogeneo [1] di I equazioni, quante sono le righe della matrice

$$\mathbf{C}\,\mathbf{R}^T - \lambda\,\mathbf{I}$$

in I incognite, quante sono le componenti di $\mathbf{u}$.

Questo sistema può avere o *una* o *infinite* soluzioni. Infatti, se il determinante della matrice dei coefficienti

$$\det|\mathbf{C}\,\mathbf{R}^T - \lambda\,\mathbf{I}| \neq 0$$

l'*unica* soluzione possibile è il vettore $\mathbf{u} = \mathbf{0}_{\mathrm{I}}$, soluzione che risulta di scarso interesse perché non permette di rendere visibile la configurazione della nuvola perché tutte le proiezioni si concentrerebbero nell'origine. Ma, se il

[1] Il sistema è *lineare* perché le incognite vi compaiono al primo grado, ed è *omogeneo* perché il vettore dei termini noti è $\mathbf{0}_{\mathrm{I}}$.

determinante

$$\det|\mathbf{C}\,\mathbf{R}^T - \lambda\,\mathbf{I}| = 0 \tag{3.9.4}$$

si può ottenere come soluzione un vettore $\mathbf{u}$ non nullo, individuato però a meno di una costante moltiplicativa arbitraria. Ecco quindi la condizione perché l'equazione (3.9.3) sia vera: scegliere quel valore di λ che rende nullo il determinante della matrice $\mathbf{C}\,\mathbf{R}^T - \lambda\,\mathbf{I}$. La (3.9.4) è detta *equazione caratteristica* della matrice $\mathbf{C}\,\mathbf{R}^T$. Si tratta di un'equazione polinomiale di ordine $I \leq J$ nella incognita λ che, per il teorema fondamentale dell'algebra, ammette I soluzioni dette *autovalori* [1] della matrice $\mathbf{C}\,\mathbf{R}^T$ che sono *sempre* reali per la $\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}$-simmetria della matrice e *non negativi*, come verrà mostrato fra poco e nella prossima Sez. 3.15.

Ma, se gli autovalori sono tutti distinti, come sempre accade di trovare nelle situazioni reali, a quale di essi corrisponderà il vettore cercato, quello che rende massima l'inerzia delle proiezioni? Se si indica con λ_0 questo autovalore e si premoltiplica per $\mathbf{u}^T\,\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}$ la condizione (3.9.3), si ottiene

$$\mathbf{u}^T\,\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\,\mathbf{C}\,\mathbf{R}^T\,\mathbf{u} = \lambda_0\,\mathbf{u}^T\,\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\,\mathbf{u}$$
$$In_0(\mathbf{u}) = \lambda_0$$

per la (3.9.1) e per il vincolo (3.9.2). Quindi, gli I autovalori appena trovati corrispondono a I possibili valori che può assumere l'inerzia riferita all'origine delle J proiezioni dei profili su $\mathbf{u}$. Questo fatto importante verrà approfondito nella prossima Sez. 3.15. Dal momento che si sta cercando il *massimo* vincolato di $In_0(\mathbf{u})$, λ_0 sarà l'autovalore *più grande*. Inoltre λ_0 non può essere negativo, perché tale non può risultare l'inerzia.

In base all'equazione (3.9.3), a λ_0, corrisponde un vettore $\mathbf{u}_0$ tale che

$$\mathbf{C}\,\mathbf{R}^T\,\mathbf{u}_0 = \lambda_0\,\mathbf{u}_0. \tag{3.9.5}$$

Dal momento che $\det|\mathbf{C}\,\mathbf{R}^T - \lambda_0\,\mathbf{I}| = 0$, le soluzioni sono infinite, ossia le I componenti di $\mathbf{u}_0$ possono essere determinate a meno di una costante moltiplicativa arbitraria, ma non nulla. Si può allora scegliere quel valore della costante moltiplicativa che rende $\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}$-unitaria la lunghezza di $\mathbf{u}_0$. I valori possibili sono *due*, uno col segno $+$ e uno col segno $-$, perché la costante è ottenuta come radice della forma quadratica (3.5.1), per cui sia $\mathbf{u}_0$ che $-\,\mathbf{u}_0$ hanno lunghezza $\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}$-unitaria. In termini geometrici, resta *indeterminato* il

---

[1] Le soluzioni dell'equazione caratteristica sono dette talvolta *radici latenti* o *valori caratteristici.*

verso di $\mathbf{u}_0$. Abitualmente si prende il valore col segno $+$. In questo senso si considera $\mathbf{u}_0$ come *unica* soluzione dell'equazione (3.9.3), e viene chiamato *autovettore* [1] della matrice $\mathbf{C}\,\mathbf{R}^T$ *corrispondente all'autovalore* λ_0.

Si conclude dunque che il vettore $\mathbf{u}_0$ è il vettore cercato, quello che individua la retta sulla quale $In_{\mathbf{0}}(\mathbf{u}_0)$, l'inerzia rispetto l'origine delle J proiezioni dei punti della nuvola, risulta la massima possibile.

3.10 - Esempio: calcolo degli autovalori

Le condizioni di massimo vincolato della funzione $In_0(\mathbf{u})$, portano a scrivere l'equazione (3.9.3)

$$\mathbf{C}\,\mathbf{R}^T\,\mathbf{u} = \lambda\,\mathbf{u}$$

che, nel caso dell'esempio *Spettacoli-3* di ordine 3×8 di TAV. 14, risulta

$$\begin{pmatrix} 0.551 & 0.546 & 0.546 \\ 0.235 & 0.238 & 0.236 \\ 0.215 & 0.216 & 0.218 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} \lambda\,u_1 \\ \lambda\,u_2 \\ \lambda\,u_3 \end{pmatrix}$$

e che può scriversi come sistema lineare omogeneo

$$\begin{pmatrix} 0.551-\lambda & 0.546 & 0.546 \\ 0.235 & 0.238-\lambda & 0.236 \\ 0.215 & 0.216 & 0.218-\lambda \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \qquad (3.10.1)$$

Questo sistema può avere una soluzione $\mathbf{u} \neq 0$ soltanto in corrispondenza di speciali valori della incognita λ, che rendono nullo il determinante della matrice dei coefficienti, ossia, secondo la (3.9.4), quando

$$\det \begin{vmatrix} 0.551-\lambda & 0.546 & 0.546 \\ 0.235 & 0.238-\lambda & 0.236 \\ 0.215 & 0.216 & 0.218-\lambda \end{vmatrix} = 0.$$

L'*equazione caratteristica* di una matrice si presenta *sempre* in questa forma: è eguale a zero il determinante della matrice ai cui elementi diagonali è sottratto il parametro incognito λ. Sviluppando il determinante si ottiene

$$(0.551-\lambda)\det \begin{vmatrix} 0.238-\lambda & 0.236 \\ 0.216 & 0.218-\lambda \end{vmatrix} - 0.546\,\det \begin{vmatrix} 0.235 & 0.236 \\ 0.215 & 0.218-\lambda \end{vmatrix}$$
$$+\,0.546\,\det \begin{vmatrix} 0.235 & 0.238-\lambda \\ 0.215 & 0.216 \end{vmatrix} = 0$$

---

[1] Gli autovettori sono detti anche *vettori latenti* o *vettori caratteristici*.

che si riduce all'equazione di terzo grado

$$-\lambda^3 + 1.007\,\lambda^2 \; - 0.007\,\lambda \; + 0.0000107 = 0.$$

Le sue tre soluzioni, o radici, ordinate per valori decrescenti,

$$\lambda_0 = 1, \qquad \lambda_1 = 0.005, \qquad \lambda_2 = 0.002 \tag{3.10.2}$$

sono gli *autovalori*, reali e non negativi, di

$$\mathbf{C}\,\mathbf{R}^T = \begin{pmatrix} 0.551 & 0.546 & 0.546 \\ 0.235 & 0.238 & 0.236 \\ 0.215 & 0.216 & 0.218 \end{pmatrix}.$$

È possibile controllare la correttezza del calcolo degli autovalori, perché, come verrà mostrato nella Sez. 3.15, la traccia di questa matrice deve risultare eguale alla somma di tutti gli I autovalori

$$tr\,\left[\mathbf{C}\,\mathbf{R}^T\right] = 0.551 + 0.238 + 0.218 = 1.007$$

risulta eguale a

$$\sum_{a=0}^{2} \lambda_a = 1 + 0.005 + 0.002 = 1.007.$$

3.11 - Esempio: calcolo dell'autovettore $\mathbf{u_0}$

Dal momento che si sta cercando il *massimo* della funzione $In_{\mathbf{0}}(\mathbf{u})$, occorre considerare l'autovalore più grande, ossia $\lambda_0 = 1$ ed è in corrispondenza di questo che va cercato il vettore $\mathbf{u}_0$. Sostituendo $\lambda_0 = 1$ nella (3.10.1), si ottiene il sistema

$$\begin{cases} (0.551-1)\,u_{01} & + & 0.546\,u_{02} & + & 0.546\,u_{03} & = & 0 \\ 0.235\,u_{01} & + & (0.238-1)\,u_{02} & + & 0.236\,u_{03} & = & 0 \\ 0.215\,u_{01} & + & 0.216\,u_{02} & + & (0.218-1)\,u_{03} & = & 0. \end{cases} \tag{3.11.1}$$

Questo sistema è omogeneo e il determinante della matrice è nullo, per cui soltanto due delle sue equazioni sono indipendenti, per esempio le prime due

$$\begin{cases} -0.449\,u_{01} & + & 0.546\,u_{02} & + & 0.546\,u_{03} & = & 0 \\ 0.235\,u_{01} & + & -0.762\,u_{02} & + & 0.236\,u_{03} & = & 0 \end{cases}$$

e quindi la sua soluzione $\mathbf{u}_0$ può essere ottenuta soltanto a meno di una costante moltiplicativa non nulla. Risolvendo il sistema col metodo di sosti-

tuzione si ha

$$u_{01} = \frac{\det \begin{vmatrix} 0.546 & 0.546 \\ -0.762 & 0.236 \end{vmatrix}}{\det \begin{vmatrix} -0.449 & 0.546 \\ 0.235 & -0.762 \end{vmatrix}} u_{03} = \frac{0.545}{0.214} u_{03} = 2.543\, u_{03}.$$

$$u_{02} = \frac{\det \begin{vmatrix} -0.449 & 0.546 \\ -0.234 & 0.236 \end{vmatrix}}{\det \begin{vmatrix} -0.449 & 0.546 \\ 0.235 & -0.762 \end{vmatrix}} u_{03} = \frac{0.234}{0.214} u_{03} = 1.094\, u_{03}.$$

Le infinite soluzioni del sistema (3.11.1) sono dunque

$$\mathbf{u_0} = \begin{pmatrix} 2.543\, u_{03} \\ 1.094\, u_{03} \\ 1.000\, u_{03} \end{pmatrix} = u_{03} \begin{pmatrix} 2.543 \\ 1.094 \\ 1.000 \end{pmatrix}.$$

Imponendo il vincolo (3.9.2) che questo vettore abbia lunghezza $\mathbf{D_{\overline{c}}^{-1}}$-unitaria

$$\begin{aligned} \mathbf{u}_0^T\, \mathbf{D_{\overline{c}}^{-1}}\, \mathbf{u}_0 &= \sum_1^3 \frac{1}{\overline{c}_i}\, u_{0i}^2 \qquad (3.11.2) \\ &= \frac{1}{0.548}\, (2.543\, u_{03})^2 + \frac{1}{0.236}\, (1.094\, u_{03})^2 + \frac{1}{0.216}\, u_{03}^2 \\ &= 21.503\, u_{03}^2 = 1 \end{aligned}$$

si ricava

$$u_{03} = \sqrt{\frac{1}{21.503}} = \pm\, 0.216$$

La direzione di $\mathbf{u}_0$ è quindi *indeterminata*, ma, scegliendo per esempio il valore positivo $u_{03} = 0.216$, si ottiene finalmente

$$\mathbf{u}_0 = \begin{pmatrix} 0.548 \\ 0.236 \\ 0.216 \end{pmatrix}. \qquad (3.11.3)$$

Questo è dunque il vettore $\mathbf{D_{\overline{c}}^{-1}}$-*unitario* cercato, l'autovettore soluzione del sistema (3.11.1) corrispondente all'autovalore $\lambda_0 = 1$, sul quale è massima l'inerzia delle proiezioni dei profili degli 8 tipi di spettacolo. Tenendo presente la (3.5.6), per la (3.11.2) e la (3.9.5), l'inerzia risulta

$$In_{\mathbf{0}}(\mathbf{u}_0) = \mathbf{u}_0^T\, \mathbf{D_{\overline{c}}^{-1}}\, \mathbf{C}\, \mathbf{R}^T\, \mathbf{u}_0 = \mathbf{u}_0^T\, \mathbf{D_{\overline{c}}^{-1}}\, \lambda_0\, \mathbf{u}_0 = \lambda_0 = 1$$

per cui l'autovalore λ_0 indica anche l'inerzia massima che si può ottenere su $\mathbf{u}_0$, ossia in uno spazio uni-dimensionale. Questo valore è da confrontare con

l'inerzia complessiva $In_{\mathbf{0}} = 1.007$, nello spazio tri-dimensionale, calcolata nella Sez. 3.3.

Peraltro, confrontando la (3.11.3) con la (2.6.1) si scopre che

$$\mathbf{u}_0 = \overline{\mathbf{c}}$$

ossia che l'autovettore appena trovato coincide col vettore che individua il baricentro della nuvola dei J punti-colonna nello spazio $\Re^I$. Di conseguenza l'asse individuato dal vettore $\mathbf{u}_0$ passa per l'origine e per il baricentro della nuvola. Ora, nella Sez. 2.8 si è visto che questo asse è $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale all'iperpiano del simplesso che contiene i punti della nuvola, per cui le loro proiezioni finiscono tutte nel baricentro che si trova a distanza $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitaria dall'origine $\mathbf{0}_{\mathrm{I}}$. Qui viene a concentrarsi tutta la massa della nuvola, che vale 1. È per questo motivo che l'inerzia delle proiezioni su $\mathbf{u}_0 = \overline{\mathbf{c}}$ vale

$$\lambda_0 = \left(\sum_{i=1}^{I} \overline{r}_j \right) \times d_D^2(\overline{\mathbf{c}}, \mathbf{0}_{\mathrm{I}}) = 1 \times 1 = 1.$$

È evidente che la configurazione della nuvola risulta indecifrabile su questo asse. La soluzione $(\mathbf{u}_0 = \overline{\mathbf{c}}, \lambda_0 = 1)$ non è d'interesse e viene quindi *scartata*.[1]

3.12 - Gli autovettori $\mathbf{u}_1, \ldots, \mathbf{u}_{I-1}$

Messa da parte la soluzione $(\lambda_0 = 1, \mathbf{u}_0 = \overline{\mathbf{c}})$, occorre cercare un altro vettore $\mathbf{u}$, che oltre ad essere $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitario sia adesso anche $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale ad $\mathbf{u}_0$, dal momento che si sta costruendo una nuova base $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortonormale di riferimento di $\Re^I$, e tale che su di esso $In_{\mathbf{0}}(\mathbf{u})$, l'inerzia rispetto all'origine delle proiezioni dei J punti sia massima, compatibilmente con questi due vincoli. Si deve quindi trovare il

$$\begin{array}{lll} \text{massimo di} & \mathbf{u}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{C} \mathbf{R}^T \mathbf{u} & \\ \text{con i vincoli} & \mathbf{u}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{u} - 1 = 0 & \mathbf{u} \text{ è } \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\text{-unitario} \\ \text{e} & \mathbf{u}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{u}_0 = 0 & \mathbf{u} \text{ è } \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\text{-ortogonale a } \mathbf{u}_0 = \overline{\mathbf{c}} \end{array}$$

Il metodo dei moltiplicatori di Lagrange fornisce ancora la soluzione, costruendo la funzione

$$\mathcal{L}(\mathbf{u}, \lambda, \mu) = \mathbf{u}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{C} \mathbf{R}^T \mathbf{u} - \lambda (\mathbf{u}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{u} - 1) - \mu \mathbf{u}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{u}_0$$

---

[1] Questa soluzione è detta *banale*, ossia priva di rilievo, o anche *degenere* o *spuria* ed è diretta conseguenza del fatto che le componenti di ogni profilo sono legate da una relazione lineare: $\sum_i c_{ij} = 1$.

ove compaiono due moltiplicatori di Lagrange, λ e μ, dal momento che adesso sono due i vincoli sulle componenti di $\mathbf{u}$. Per quanto visto nella Sez. 3.8, il massimo va necessariamente cercato dove si annullano le derivate della funzione di Lagrange

$$\frac{\partial \mathcal{L}(\mathbf{u}, \lambda, \mu)}{\partial \mathbf{u}} = 2\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{C}\,\mathbf{R}^T\,\mathbf{u} - 2\,\lambda\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u} - \mu\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_0 = \mathbf{0}_{\mathrm{I}}.$$

Premoltiplicando per $\mathbf{u}_0^T$ ed esprimendo $\mathbf{R}^T$ come nella (3.2.4), si ottiene

$$2\,\mathbf{u}_0^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u} - 2\,\lambda\,\mathbf{u}_0^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u} - \mu\,\mathbf{u}_0^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_0 = 0$$

e trasponendo l'equazione dopo averne cambiato il segno

$$\begin{aligned} -2\,\mathbf{u}^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{C}\;\mathbf{R}^T\,\mathbf{u}_0 + 2\,\lambda\,\mathbf{u}^T\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_0 + \mu\,\mathbf{u}_0^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_0 = 0 \\ -2\,\mathbf{u}^T\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\lambda_0\,\mathbf{u}_0 + 2\,\lambda\,\mathbf{u}^T\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_0 + \mu = 0 \\ \mu = 0 \end{aligned}$$

in base alla (3.9.5) e al vincolo di $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonalità di $\mathbf{u}$ con $\mathbf{u}_0 = \overline{\mathbf{c}}$.

Questo risultato porta a concludere che cercare il massimo dell'inerzia con i due vincoli su $\mathbf{u}$, *equivale* a cercare il massimo non vincolato della funzione di Lagrange $\mathcal{L}(\mathbf{u}, \lambda)$ in quanto in essa la condizione che il nuovo vettore unitario $\mathbf{u}$ sia $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale a $\mathbf{u}_0 = \overline{\mathbf{c}}$ è *già implicita*. Le condizioni di massimo dell'inerzia $In_{\mathbf{0}}(\mathbf{u})$ sono già state trovate nella Sez. 3.9, ove è risultato che questo si può avere soltanto in corrispondenza degli autovalori della matrice $\mathbf{C}\,\mathbf{R}^T$. Avendo già scartato l'autovalore $\lambda_0 = 1$, il massimo cercato si avrà in corrispondenza dell'autovalore λ_1, il più grande dopo λ_0, che viene detto *primo autovalore* della matrice $\mathbf{C}\,\mathbf{R}^T$. Questo nell'ipotesi che gli autovalori trovati siano tutti distinti, come sempre accade nelle applicazioni reali. All'autovalore λ_1 di $\mathbf{C}\,\mathbf{R}^T$ corrisponde il vettore $\mathbf{u}_1$ che soddisfa l'equazione (3.9.3)

$$\mathbf{C}\,\mathbf{R}^T\,\mathbf{u}_1 = \lambda_1\,\mathbf{u}_1$$

ed è chiamato *primo autovettore* della matrice $\mathbf{C}\,\mathbf{R}^T$.

In modo analogo si procede a determinare i successivi autovettori: $\mathbf{u}_2, \ldots,\ \mathbf{u}_{I-1}$ corrispondenti agli autovalori $\lambda_2, \ldots, \lambda_{I-1}$, che per costruzione risultano ordinati per valori *decrescenti*

$$\lambda_1 > \lambda_2 > \ldots > \lambda_{I-1}.$$

Si può allora concludere che

1 - l'avere scartato la soluzione banale $(\lambda_0 = 1, \mathbf{u}_0 = \overline{\mathbf{c}})$ riduce a $I - 1$ gli autovalori e gli autovettori di interesse;

2 - ciascun autovettore $\mathbf{u}_a$, con $a = 1, 2, \ldots, I-1$, soddisfa la condizione (3.9.3) per un massimo dell'inerzia $In_{\mathbf{0}}(\mathbf{u}_a)$,

$$\mathbf{C}\,\mathbf{R}^T\,\mathbf{u}_a = \lambda_a\,\mathbf{u}_a \tag{3.12.1}$$

e di $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortonormalità

$$\begin{aligned} \mathbf{u}_a^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_b &= 0 \\ \mathbf{u}_a^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_a &= 1; \end{aligned} \qquad a \neq b$$

3 - la somma delle I componenti di ogni autovettore non banale è nulla perché, risultando questi $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonali a coppie per costruzione e quindi anche ad $\mathbf{u}_0 = \overline{\mathbf{c}}$,

$$0 = \mathbf{u}_a^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_0 = \mathbf{u}_a^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\overline{\mathbf{c}} = \mathbf{u}_a^T\,\mathbf{1}_{\mathrm{I}} = \sum_{i=1}^{I} u_{ai}. \tag{3.12.2}$$

Invece la somma delle componenti dell'autovettore banale vale 1, perché $\mathbf{u}_0 = \overline{\mathbf{c}}$ è un profilo.

Infine, dalla (3.12.1) si deduce che i risultati dell'Analisi delle Corrispondenze di due matrici *proporzionali* $\mathbf{N}$ e $c\,\mathbf{N}$, dove c è una costante positiva, sono i medesimi, perché due matrici proporzionali hanno matrici dei profili eguali.

3.13 - Esempio: gli autovettori $\mathbf{u}_1$ e $\mathbf{u}_2$

Sostituendo nel sistema omogeneo (3.10.1) il primo autovalore $\lambda_1 = 0.005$, il più grande nella (3.10.2) dopo quello scartato $\lambda_0 = 1$, si ottiene

$$\begin{cases} (0.551 - 0.005)\,u_{11} + 0.546\,u_{12} + 0.546\,u_{13} = 0 \\ 0.235\,u_{11} + (0.238 - 0.005)\,u_{12} + 0.236\,u_{13} = 0 \\ 0.215\,u_{11} + 0.216\,u_{12} + (0.218 - 0.005)\,u_{13} = 0 \end{cases}$$

che ha nullo il determinante della matrice dei coefficienti, per cui le sue infinite soluzioni si ottengono facendo variare il parametro u_{13} in

$$\mathbf{u}_1 = \begin{pmatrix} -2.148\,u_{13} \\ 1.148\,u_{13} \\ 1.000\,u_{13} \end{pmatrix} = u_{13} \begin{pmatrix} -2.148 \\ 1.148 \\ 1.000 \end{pmatrix}.$$

Imponendo il vincolo che questo vettore sia di lunghezza $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitaria e scegliendo il valore *positivo*[1] di $u_{13} = \pm 0.232$ si ottiene finalmente il vettore

[1] Per rendere unica la soluzione, molte routines di calcolo scelgono convenzionalmente, per ogni autovettore non banale, quel valore che ne rende *posi-*

$\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitario cercato

$$\mathbf{u}_1 = \begin{pmatrix} -0.498 \\ 0.266 \\ 0.232 \end{pmatrix} \tag{3.13.1}$$

che è l'autovettore soluzione del sistema omogeneo, corrispondente all'autovalore $\lambda_1 = 0.005$ e sul quale i profili degli 8 tipi di spettacolo si proiettano con dispersione massima.

Ripetendo il procedimento col secondo autovalore $\lambda_2 = 0.002$ si ottiene il suo corrispondente autovettore $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitario

$$\mathbf{u}_2 = \begin{pmatrix} -0.009 \\ -0.331 \\ 0.340 \end{pmatrix}. \tag{3.13.2}$$

Si può così concludere che le soluzioni al problema di trovare le direzioni di massima dispersione delle proiezioni, sono

$$\lambda_1 = 0.005,\ \mathbf{u}_1 = \begin{pmatrix} -0.498 \\ 0.266 \\ 0.232 \end{pmatrix} \qquad \text{e} \qquad \lambda_2 = 0.002,\ \mathbf{u}_2 = \begin{pmatrix} -0.009 \\ -0.331 \\ 0.340 \end{pmatrix}.$$

È facile verificare che la somma delle componenti dei due autovettori è nulla, in quanto, per la (3.12.2), entrambi $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonali a $\mathbf{u}_0 = \overline{\mathbf{c}}$, e che sono anche $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonali tra loro

$$\begin{aligned} \mathbf{u}_1^T\, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\, \mathbf{u}_2 &= \begin{pmatrix} -0.498 \\ 0.266 \\ 0.232 \end{pmatrix}^T \begin{pmatrix} \frac{1}{0.548} & 0 & 0 \\ 0 & \frac{1}{0.236} & 0 \\ 0 & 0 & \frac{1}{0.216} \end{pmatrix} \begin{pmatrix} -0.009 \\ -0.331 \\ 0.340 \end{pmatrix} \\ &= \frac{0.498 \times 0.009}{0.548} - \frac{0.266 \times 0.331}{0.236} + \frac{0.232 \times 0.340}{0.216} \\ &= 0.008 - 0.373 + 0.365 = 0. \end{aligned}$$

Il fatto che i due autovettori risultino $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonali al vettore $\overline{\mathbf{c}}$, rivela che essi giacciono in un piano parallelo a quello del simplesso, dal momento che questo è a sua volta $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale al vettore $\overline{\mathbf{c}}$, così come si è visto nella Sez. 2.8.

3.14 - Assi fattoriali d'inerzia

tiva la prima componente. Nel presente caso avrebbero scelto $u_{13} = -0.232$, sicché $\mathbf{u}_1 = (+0.498 \quad -0.266 \quad -0.232)^T$ avrebbe verso *opposto* a quello stabilito nella (3.13.1).

Gli autovettori $\mathbf{u}_a$, con $a = 1, 2, \ldots, I-1$, appena individuati, hanno tutti origine in $\mathbf{0}_\mathrm{I}$, l'origine degli I vettori unitari di base $\mathbf{e}_i$, ma, come si è visto nella Sez. 3.4, il punto privilegiato rispetto al quale valutare l'inerzia è il *baricentro* $\overline{\mathbf{c}}$ della nuvola di punti. Lo stesso procedimento che nella Sez. precedente ha portato ad individuare autovalori ed autovettori, può essere ripetuto per ricercare quel vettore $\mathbf{u}^\star$, che sia di lunghezza $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitaria e che abbia *origine nel baricentro*, sul quale l'inerzia rispetto al baricentro delle proiezioni dei profili della nuvola risulti massima. La raffigurazione geometrica è illustrata nella TAV. 23. Procedendo come nella Sez. 3.5, se $\mathbf{g}$ indica il vettore di ordine J delle coordinate su $\mathbf{u}^\star$ delle proiezioni dei punti della nuvola, essendo $\overline{\mathbf{C}} = \overline{\mathbf{c}}\,\mathbf{1}_\mathrm{J}^T$ la matrice con le colonne eguali a $\overline{\mathbf{c}}$,

$$\mathbf{g} = (\,\mathbf{C} - \overline{\mathbf{C}}\,)^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}^\star$$

l'inerzia di queste rispetto al baricentro risulta

$$\begin{aligned} In_{\overline{\mathbf{c}}}(\mathbf{u}^\star) &= \mathbf{g}^T\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{g} \\ &= \mathbf{u}^{\star T}\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,(\mathbf{C} - \overline{\mathbf{C}})\,\mathbf{D}_{\overline{\mathbf{r}}}\,(\mathbf{C} - \overline{\mathbf{C}})^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}^\star. \end{aligned} \qquad (3.14.1)$$

Ci si chiede allora quale sia l'orientamento di $\mathbf{u}^\star$ che rende massima questa espressione, compatibilmente con l'essere $\mathbf{u}$ di lunghezza $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitaria

$$\mathbf{u}^{\star T}\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\mathbf{u}^\star = 1. \qquad (3.14.2)$$

Si costruisce così la funzione di Lagrange

$$\mathcal{L}(\mathbf{u}^\star, \lambda^\star) = \mathbf{u}^{\star T}\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,(\mathbf{C} - \overline{\mathbf{C}})\,\mathbf{D}_{\overline{\mathbf{r}}}\,(\mathbf{C} - \overline{\mathbf{C}})^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}^\star - \lambda^\star\,(\mathbf{u}^{\star T}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}^\star - 1)$$

la cui derivata (APP. A) stabilisce le I condizioni che devono essere soddisfatte dalle componenti di $\mathbf{u}^\star$ perché la (3.14.1) abbia il suo massimo valore

$$(\mathbf{C} - \overline{\mathbf{C}})\,\mathbf{D}_{\overline{\mathbf{r}}}\,(\mathbf{C} - \overline{\mathbf{C}})^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}^\star - \lambda\,\mathbf{u}^\star = \mathbf{0}_\mathrm{I}. \qquad (3.14.3)$$

Sviluppando i prodotti dei termini entro le parentesi si ottiene

$$(\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T - \mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\overline{\mathbf{C}}^T - \overline{\mathbf{C}}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T + \overline{\mathbf{C}}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\overline{\mathbf{C}}^T)\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}^\star = \lambda^\star\,\mathbf{u}^\star.$$

Le tre ultime matrici entro la parentesi sono simmetriche ed eguali perché

$$\begin{aligned} \mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\overline{\mathbf{C}}^T &= \mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{1}_\mathrm{J}\,\overline{\mathbf{c}}^T = \overline{\mathbf{c}}\,\overline{\mathbf{c}}^T \\ \overline{\mathbf{C}}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T &= \overline{\mathbf{c}}\,\mathbf{1}_\mathrm{J}^T\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T = \overline{\mathbf{c}}\,\overline{\mathbf{c}}^T \\ \overline{\mathbf{C}}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\overline{\mathbf{C}}^T &= \overline{\mathbf{c}}\,\mathbf{1}_\mathrm{J}^T\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{1}_\mathrm{J}\,\overline{\mathbf{c}}^T = \overline{\mathbf{c}}\,\overline{\mathbf{c}}^T \end{aligned}$$

per cui la (3.14.3) assume la forma

$$(\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T - \overline{\mathbf{c}}\,\overline{\mathbf{c}}^T)\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}^\star = \lambda^\star\,\mathbf{u}^\star \qquad (3.14.4)$$

che mostra come la matrice da diagonalizzare di ordine $I \times I$ sia $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-simmetrica [1] e abbia quindi autovalori ed autovettori reali, come mostrato nelle Sezioni B.2 e B.3 dell'Appendice B.

Svolta la parentesi, il primo termine per la (3.2.4) vale $\mathbf{C}\,\mathbf{R}^T\,\mathbf{u}^\star$, mentre l'altro vale

$$\overline{\mathbf{c}}\,\overline{\mathbf{c}}^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} = \overline{\mathbf{c}}\,\mathbf{1}_{\mathrm{I}}^T\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} = \overline{\mathbf{c}}\,\mathbf{1}_{\mathrm{I}}^T = \overline{\mathbf{c}}\,\mathbf{1}_{\mathrm{J}}^T\,\mathbf{R}^T = \overline{\mathbf{C}}\,\mathbf{R}^T.$$

per cui la (3.14.3) equivale a

$$\left(\mathbf{C} - \overline{\mathbf{C}}\right)\,\mathbf{R}^T\,\mathbf{u}^\star = \lambda^\star\,\mathbf{u}^\star. \tag{3.14.5}$$

ove $\mathbf{C} - \overline{\mathbf{C}}$ è la matrice degli scarti dei J profili delle colonne dal loro profilo medio ponderato.

Verrà mostrato ora che

1 - $\mathbf{u}_I^\star = \overline{\mathbf{c}}$ è un autovettore banale anche della (3.14.5), corrispondente però all'autovalore $\lambda_I^\star = 0$, e non più all'autovalore di rango 0 come accadeva quando la funzione da massimizzare era l'inerzia $In_{\mathbf{0}}(\mathbf{u})$ riferita all'origine;

2 - tutti i *rimanenti* $I-1$ autovalori e gli I elementi dei corrispondenti autovettori risultano *eguali* a quelli ottenuti quando la funzione era $In_{\mathbf{0}}(\mathbf{u})$, in altri termini gli autovettori $\mathbf{u}_a^\star$ hanno gli stessi orientamenti degli $\mathbf{u}_a$, per cui, per ogni rango $a = 1, 2, \ldots, I-1$ e per ogni $i = 1, 2, \ldots, I$ risulta

$$\lambda_a^\star = \lambda_a \qquad \text{e} \qquad u_{ai}^\star = u_{ai} \tag{3.14.6}$$

con l'unica differenza che gli autovettori $\mathbf{u}_a^\star$ hanno origine in $\overline{\mathbf{c}}$ e quelli $\mathbf{u}_a$ in $\mathbf{0}_{\mathrm{I}}$.

Per dimostrare il primo punto, bisogna tener presente quanto mostrato nella Sez. 3.9, ossia che $\overline{\mathbf{c}}$ è un autovettore della (3.9.3) corrispondente all'autovalore 1 e soddisfa quindi la condizione

$$\mathbf{C}\,\mathbf{R}^T\,\overline{\mathbf{c}} = 1\,\overline{\mathbf{c}}$$

ma, d'altra parte,

$$\overline{\mathbf{C}}\,\mathbf{R}^T\,\overline{\mathbf{c}} = \overline{\mathbf{c}}\,\mathbf{1}_{\mathrm{J}}^T\,\mathbf{R}^T\,\overline{\mathbf{c}} = \overline{\mathbf{c}}\,1$$

---

[1] La somma e la differenza di due matrici simmetriche $\mathbf{S}_1$ e $\mathbf{S}_2$ è una matrice simmetrica $\mathbf{S}$, per cui anche la somma e la differenza di due matrici $\mathbf{D}$-simmetriche è ancora una matrice $\mathbf{D}$ simmetrica: $(\mathbf{S}_1 \pm \mathbf{S}_2)\,\mathbf{D} = \mathbf{S}\,\mathbf{D}$.

per cui la (3.14.5) diventa

$$1\,\overline{\mathbf{c}} - \overline{\mathbf{c}}\,1 = \lambda_a^\star\,\overline{\mathbf{c}} \qquad \text{e quindi} \qquad \lambda_I^\star\ \overline{\mathbf{c}} = \mathbf{0}_{\mathrm{I}}.$$

Siccome le componenti di $\overline{\mathbf{c}}$ sono sempre positive, deve essere necessariamente $\lambda_I^\star = 0$. Si è posto $a = I$ perché l'autovettore banale $\overline{\mathbf{c}}$ corrisponde ora all'autovalore più *piccolo*, non potendo gli autovalori risultare negativi dal momento che rappresentano l'inerzia delle proiezioni sull'autovettore, come si vedrà meglio nella prossima Sezione.

Riassumendo, quando l'analisi in $\Re^I$ è fatta massimizzando l'inerzia riferita al baricentro, la soluzione banale è $(\lambda_I^\star = 0,\ \mathbf{u}_I^\star = \overline{\mathbf{c}})$, mentre, quando si massimizza quella riferita all'origine, risulta $(\lambda_0 = 1,\ \mathbf{u}_0 = \overline{\mathbf{c}})$. Questo risultato è conseguenza del *Teorema di Huygens*, che, come si è visto nell Sez. 3.4, lega l'inerzia riferita a un punto qualunque, qui l'origine $\mathbf{0}_{\mathrm{I}}$, all'inerzia riferita al baricentro $\overline{\mathbf{c}}$, come nella (3.4.3).

Dal punto di vista matematico, la soluzione banale $(\lambda_I^\star = 0,\ \mathbf{u}_I^\star = \overline{\mathbf{c}})$ è mera conseguenza del procedimento analitico adottato, mentre, dal punto di vista geometrico, la si interpreta tenendo presente che il profilo $\overline{\mathbf{c}}$ è $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale all'iperpiano del simplesso che contiene i J punti della nuvola. Perciò massimizzare l'inerzia delle proiezioni rispetto al baricentro equivale a considerare la nuvola dei profili nell'iperpiano del simplesso, eliminando così una dimensione di $\Re^I$.

La dimostrazione del secondo punto è immediata perché ogni altro autovettore della (3.14.5) deve risultare $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale a $\overline{\mathbf{c}}$, per cui questa, per $a = 1, 2, \ldots, I-1$, si riduce a

$$\mathbf{C}\,\mathbf{R}^T\mathbf{u}_a^\star = \lambda_a^\star\mathbf{u}_a^\star \tag{3.14.7}$$

che è la (3.9.3) già considerata nella Sez. 3.9 e seguenti. Gli autovalori $\lambda_a^\star$ risultano quindi identici ai corrispondenti λ_a del medesimo rango e sono quindi ordinabili per valori decrescenti

$$\lambda_1^\star > \lambda_2^\star > \ldots > \lambda_{I-1}^\star.$$

Gli autovettori $\mathbf{u}_a^\star$ hanno le medesime I componenti dei corrispondenti $\mathbf{u}_a$, con la differenza che hanno origine nel baricentro $\overline{\mathbf{c}}$, mentre gli autovettori $\mathbf{u}_a$ l'hanno in $\mathbf{0}_{\mathrm{I}}$. Restano così dimostrate le eguaglianze (3.14.6).

Gli autovettori $\mathbf{u}_a^\star$ risultano *linearmente indipendenti* perché corrispondono ad autovalori distinti e possono quindi costituire una *nuova base* dello spazio dei profili $\Re^I$, l'unica che verrà presa in considerazione d'ora

in avanti. La base canonica $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_I$ viene abbandonata perché non dà alcuna garanzia che l'inerzia delle proiezioni su assi di rango crescente abbia valori decrescenti. La nuova base risulta $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-*ortonormale* per costruzione, per cui *non* è ottenibile da una rotazione della base canonica che è soltanto $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale, dopo averne traslato rigidamente l'origine nel baricentro, mentre, come si è visto, può ottenersi dalla base $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_I$ con semplice traslazione dall'origine $\mathbf{0}_\mathrm{I}$ a $\overline{\mathbf{c}}$. Rispetto ad essa *ogni* profilo di $\Re^I$ può esprimersi come una combinazione lineare degli I autovettori. L'autovettore banale $\mathbf{u}_I^\star = \overline{\mathbf{c}}$ va aggiunto agli altri per descrivere l'intero spazio $\Re^I$, ma è inessenziale per l'analisi della configurazione della nuvola che, come si è visto, è contenuta nell'iperpiano del simplesso.

Nel caso dell'esempio *Spettacoli-3*, in base ai risultati della Sez. 3.10, le soluzioni, evidenziate nella TAV. 21 per le loro implicazioni geometriche, sono dunque

$$\begin{array}{lllll}
\left[\lambda_0 = 1\right] & \lambda_1 = 0.005 & \lambda_2 = 0.002 & & \\
 & \lambda_1^\star = 0.005 & \lambda_2^\star = 0.002 & \left[\lambda_3^\star = 0\right] & \\
\left[\mathbf{u}_0 = \overline{\mathbf{c}}\right] & \mathbf{u}_1 = \begin{pmatrix} -0.498 \\ 0.266 \\ 0.232 \end{pmatrix} & \mathbf{u}_2 = \begin{pmatrix} -0.009 \\ -0.331 \\ 0.340 \end{pmatrix} & & \\
 & \mathbf{u}_1^\star = \begin{pmatrix} -0.498 \\ 0.266 \\ 0.232 \end{pmatrix} & \mathbf{u}_2^\star = \begin{pmatrix} -0.009 \\ -0.331 \\ 0.340 \end{pmatrix} & \left[\mathbf{u}_3^\star = \overline{\mathbf{c}}\right] &
\end{array}$$

Dal momento che nel campo d'interesse gli autovalori coincidono, d'ora innanzi verrà lasciata cadere la *, mentre per i corrispondenti autovettori la distinzione verrà mantenuta per ben evidenziare la loro origine: l'origine della base canonica o il baricentro.

Alla retta passante per il baricentro della nuvola ed individuata da un autovettore $\mathbf{u}_a^\star$ viene dato il nome di *asse fattoriale di inerzia*[1]. In particolare di *primo* asse fattoriale d'inerzia a quella individuata da $\mathbf{u}_1^\star$, di *secondo* asse fattoriale d'inerzia a quella individuata da $\mathbf{u}_2^\star$ e così avanti. L'orientamento degli assi è fissato arbitrariamente, a causa dell'indeterminazione del segno dell'autovettore, come si è visto nella Sez. 3.9 e nelle seguenti.

---

[1] In fisica prende il nome di *asse principale d'inerzia*, o, più raramente, di *asse di massima elongazione*.

Gli $A = I - 1$ assi fattoriali sono tutti contenuti nell'iperpiano del simplesso ed il termine *asse* sta a significare che si tratta di un nuovo sistema di riferimento, centrato nel baricentro ed individuato dalla *nuova base* $\mathbf{D_c^{-1}}$ -*ortonormale* che garantisce la massima visibilità della configurazione col minimo di distorsione, come si andava cercando fin dalla Sez. 3.1.

Quindi, il primo asse fattoriale è quello che rende meglio visibile la nuvola in uno spazio mono-dimensionale, nel senso che su di esso la dispersione geometrica delle proiezioni vi è la massima possibile. La loro inerzia vale λ_1 . Il piano individuato dai primi due assi fattoriali è quello che dà la più fedele rappresentazione bi-dimensionale della configurazione della nuvola, perché l'inerzia di questo sottospazio, ossia $\lambda_1 + \lambda_2$, è superiore a quella di ogni altra coppia di assi fattoriali. Il sottospazio individuato dai primi tre assi fattoriali, fornisce infine la più fedele visione tridimensionale della configurazione della nuvola.

Può darsi il caso che con particolari strutture della matrice di contingenza la nuvola dei J profili colonna sia completamente contenuta in un sottospazio $\Re^{A*}$, con $A^* \leq A = I - 1$. D'ora innanzi A indicherà il numero degli autovalori non nulli.

I risultati della Sez. 3.3, permettono di aggiungere alle proprietà degli assi fattoriali quella di riprodurre 'al meglio' le *reciproche distanze* tra punti della nuvola, ogni distanza essendo ponderata con le masse dei due punti. Comunque, lo scopo è quello di evidenziare la nuvola di punti in sottospazi bi-dimensionali, o, quando si disponga del software adatto, anche tri-dimensionali. È chiaro che questa limitazione è dovuta a motivi pratici e non certo a vincoli teorici.

Per quanto visto nella Sez. 3.7, gli assi fattoriali sono anche le rette che passano più "vicine" ai punti della nuvola, nel senso che minimizzano la somma pesata dei quadrati delle distanze, ma diversamente da quanto avviene nella Regressione Lineare in Statistica. Limitando il paragone alla Regressione semplice, come è fatto nella TAV. 22, si può dire che sia la retta di Regressione sia il primo asse fattoriale passano per il baricentro della nuvola di punti, ma, mentre nell'Analisi delle Corrispondenze le distanze sono misurate $\mathbf{D_c^{-1}}$ -*ortogonalmente* all'asse fattoriale, nel caso della Regressione, le distanze sono misurate *parallelamente* alla direzione definita dalla variabile dipendente, perché una delle ipotesi alla base del metodo considera la variabile indipendente come deterministica ed i dati ad essa relativi 'noti senza errore'.

Infine, è importante rilevare fin da ora che gli assi fattoriali, tramite gli autovettori che li individuano, dipendono, a parità di dimensioni I e J, dai particolari elementi della matrice di contingenza in esame. In altri termini, da matrici diverse, anche se dello stesso ordine, si ottengono autovettori diversi con origine in baricentri diversi. Se però queste matrici hanno tutte gli stessi totali marginali, gli autovettori da esse derivati hanno tutti origine nello *stesso* baricentro. Questo punto importante, perché legato alla stabilità dei risultati dell'analisi, verrà sviluppato nel Cap. 7.

3.15 - Scomposizione dell'inerzia

L'inerzia riferita all'origine delle J proiezioni su uno degli autovettori $\mathbf{u}_a$, con $a = 0, 1, 2, \ldots, A$, per la (3.5.6), vale

$$In_{\mathbf{0}}(\mathbf{u}_a) = \mathbf{u}_a^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{C}\,\mathbf{R}^T\,\mathbf{u}_a = \mathbf{u}_a^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\lambda_a\,\mathbf{u}_a = \lambda_a$$

in quanto ogni autovettore soddisfa alle condizioni (3.12.1) ed è $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitario. Per le eguaglianze (3.14.6), che valgono per $a = 1, 2, \ldots, A$, anche l'inerzia riferita al baricentro sull'asse fattoriale individuato da $\mathbf{u}_a^\star$, risulta

$$In_{\overline{\mathbf{c}}}(\mathbf{u}_a^\star) = \lambda_a. \tag{3.15.1}$$

Come si vede, nell'Analisi delle Corrispondenze gli autovalori hanno un preciso significato: indicano il valore massimo possibile che può assumere la dispersione geometrica delle J proiezioni sul corrispondente asse fattoriale, dispersione misurata dall'inerzia, o varianza, riferita al baricentro. Conseguenza importante è che gli autovalori della matrice $\mathbf{C}\,\mathbf{R}^T$ *non possono essere negativi*, perché tale non può risultare l'inerzia, per come è stata definita nella Sez. 3.1. Inoltre, verrà mostrato nella Sez. 4.9 che nessun autovalore può risultare maggiore di 1.

Nel caso della matrice d'esempio di ordine 3×8 di TAV. 14, in base ai risultati della Sez. 3.10, l'inerzia riferita al baricentro delle proiezioni sugli $A = I - 1 = 2$ assi fattoriali risulta

$$In_{\overline{\mathbf{c}}}(\mathbf{u}_1^\star) = \lambda_1 = 0.005 \quad In_{\overline{\mathbf{c}}}(\mathbf{u}_2^\star) = \lambda_2 = 0.002 \quad \text{e} \quad In_{\overline{\mathbf{c}}} = \lambda_1 + \lambda_2 = 0.007$$

In generale, l'inerzia complessiva rispetto al baricentro risulta

$$In_{\overline{\mathbf{c}}} = \lambda_1 + \lambda_2 + \ldots + \lambda_A = \sum_{a=1}^{A} \lambda_a. \qquad A = I - 1$$

Questa espressione mostra come l'inerzia complessiva si ripartisce tra gli assi fattoriali. In questo senso si parla di *scomposizione dell'inerzia* sugli gli assi

fattoriali. La somma degli autovalori è legata alla matrice $\mathbf{C}\,\mathbf{R}^T$ in base alla (3.4.3) e alla (3.2.3)

$$\sum_{a=1}^{A} \lambda_a = tr\,[\mathbf{C}\,\mathbf{R}^T] - 1 \tag{3.15.2}$$

espressione che viene correntemente impiegata per controllare la correttezza del calcolo degli autovalori, come si è fatto nella Sez. 3.10.

L'inerzia delle proiezioni sull'asse a, rapportata all'inerzia complessiva

$$\tau_a = \frac{\lambda_a}{In_{\overline{\mathbf{c}}}(\mathbf{u}_a^\star)} = \frac{\lambda_a}{\sum_{a=1}^{A} \lambda_a}$$

è detta *tasso d'inerzia* o *inerzia relativa* dell'asse a ed è quindi un indicatore di quanto fedelmente le proiezioni sull'asse riflettono le vere posizioni dei punti della nuvola e quindi anche le loro reciproche distanze. Essendo gli autovalori distinti ed ordinati per valore decrescenti, i tassi d'inerzia decrescono proporzionalmente: $\tau_1 > \tau_2 > \ldots > \tau_A$. La loro somma vale 1. Nel caso della matrice d'esempio 3×8 di TAV. 14, questi risultano

$$\tau_1 = \frac{0.005}{0.005 + 0.002} = 0.71 \qquad \tau_2 = \frac{0.002}{0.005 + 0.002} = 0.29.$$

Abitualmente i tassi d'inerzia sono riportati come percentuali, per cui $\tau_1 = 71\%$ e $\tau_2 = 29\%$.

La frazione di inerzia totale nel sottospazio B-dimensionale individuato dai *primi* B assi fattoriali, ove $1 \leq B \leq A \leq I - 1$ è

$$\tau_1 + \tau_2 + \ldots + \tau_B = \sum_{a=1}^{B} \lambda_a / \sum_{a=1}^{A} \lambda_a$$

e quindi il suo complemento a 1 indica l'inerzia residua, ossia quanto dell'inerzia complessiva $In_{\overline{\mathbf{c}}}$ 'resta fuori' da questo sottospazio. Questa inerzia residua è la minima possibile, compatibilmente con i vincoli di $\mathbf{D}_{\overline{\mathbf{c}}}^{-1}$-ortogonalità degli assi fattoriali. Spesso si considerano sottospazi bi- o tri-dimensionali costituiti da assi non consecutivi: un esempio può essere il piano (1, 3), nel qual caso il suo tasso d'inerzia risulta

$$\tau_1 + \tau_3 = (\lambda_1 + \lambda_3) \,/\, \sum_{a=1}^{A} \lambda_a.$$

In definitiva, quindi, l'Analisi delle Corrispondenze può essere anche definita come la metodologia statistica che scompone l'inerzia complessiva di

una matrice di contingenza e quindi la sua struttura. Ciascun asse fattoriale, rende conto di una parte di questa struttura e il corrispondente autovalore ne misura in assoluto il potere esplicativo.

3.16 - Riepilogo

In questo capitolo è stato mostrato come viene risolto il problema di rendere visibile la configurazione della nuvola dei J profili-colonna in un sottospazio di ridotta dimensionalità, preservandone il più fedelmente la struttura. Per fare questo è stata individuata una nuova base $\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}$-ortonormale di $\Re^I$, costituita da $A \leq I-1$ autovettori $\mathbf{u}_a^\star$ con origine nel baricentro, che individuano un nuovo sistema di riferimento. Sono gli assi fattoriali, sui quali o in sottospazi da essi individuati, la configurazione della nuvola è resa visibile col minimo di distorsione. L'inerzia riferita al baricentro delle proiezioni sugli assi fattoriali è proprio l'autovalore al quale corrisponde l'autovettore che individua l'asse.

Per completare la soluzione del problema resta un ulteriore passo che verrà compiuto nel prossimo capitolo: calcolare le coordinate che i J profili assumono nel nuovo riferimento fattoriale e e costruire delle mappe bi- o tridimensionali che, rendendo visibile la configurazione della nuvola, rendano intelligibile la struttura dei profili della matrice di contingenza. Questo era proprio l'obiettivo che ci si era posto all'inizio di questo Capitolo.

3.17 - Bibliografia essenziale

La determinazione degli autovalori e degli autovettori di una matrice, oggetto di questo terzo capitolo, è un problema classico dell'Algebra Lineare, importante non solo per quest'ultima, ma per molte sue applicazioni: stabilità di sistemi meccanici, controllo delle vibrazioni, diagnostica di malfunzionamenti, ecc. L'argomento è trattato in tutti i testi citati nella Bibliografia del capitolo precedente, e in essi si possono trovare le dimostrazioni di molti risultati di base di cui qui si è dato soltanto cenno.

Chi desiderasse approfondire gli aspetti matematici e le questioni di calcolo numerico connesse al problema degli autovalori, può consultare il testo fondamentale

J. H. Wilkinson (1965). *The algebraic eigenvalue problem.* Clarendon Press, Oxford. 662 pg., reperibile in ogni biblioteca scientifica di rilievo e

Gene H. Golub, Charles F. Van Loan (1996). *Matrix Computations. 3d Ed.* The Johns Hopkins University Press. 694 pg. ISBN 0-8018-5414-8. È il

testo classico del calcolo matriciale, giunto alla terza edizione. Metodi, algoritmi e problemi di analisi numerica sono presentati in forma chiara e dettagliata. Il capitolo 8 è interamente dedicato alla diagonalizzazione di matrici simmetriche.

PARTE PRIMA: IL METODO

CAPITOLO 4: Fattori e mappe

Sommario

Col calcolo dei fattori si conclude in questo capitolo il processo di trasformazione della matrice di contingenza. I fattori permettono di evidenziare la configurazione dei profili in un contesto grafico che dovrà essere interpretato per cogliere la struttura delle associazioni tra profili. Il grado di fedeltà della rappresentazione grafica viene controllato tramite indicatori numerici. La proiezione di profili illustrativi sulle mappe permette di arricchire il contesto e mettere alla prova ipotesi e modelli.

L'attenta lettura di questo capitolo permetterà al lettore di

- acquisire il concetto di fattore di un profilo;
- distinguere i fattori principali da quelli standard;
- capire come viene costruita una mappa grafica;
- distinguere le mappe simmetriche da quelle asimmetriche;
- leggere l'output numerico di un programma d'Analisi delle Corrispondenze;
- impiegare correttamente gli indicatori diagnostici di supporto all'analisi;
- familiarizzarsi con le regole d'interpretazione dei risultati di un'analisi;
- rendersi conto dell'importante ruolo che possono svolgere i profili illustrativi;
- saper utilizzare le mappe asimmetriche sfruttando le loro caratteristiche peculiari.

CAPITOLO 4

4.1 - Fattori dei profili delle colonne

L'individuazione degli assi fattoriali, avvenuta alla fine del capitolo precedente, è soltanto una tappa intermedia, perché ciò che realmente interessa sono le *coordinate* dei profili sui nuovi assi. Una volta che queste siano disponibili, diventa possibile evidenziare la posizione dei profili in grafici mono, bi- o tri-dimensionali e rendendo così visibile la configurazione geometrica dei profili e decifrabile la struttura della matrice di contingenza.

Come si è mostrato nella Sez. 3.5 e nella 3.14, le coordinate di un profilo su un asse si ottengono tramite il prodotto scalare. Ad esempio, l'ascissa g_{j1} del profilo colonna $\mathbf{c}_j$ sul primo asse fattoriale d'inerzia, individuato dall'autovettore $\mathbf{u}_1^\star$ con origine nel baricentro $\overline{\mathbf{c}}$, è il numero [1]

$$g_{j1} = (\mathbf{c}_j - \overline{\mathbf{c}})^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{u}_1^\star = \mathbf{c}_j^T \, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \, \mathbf{u}_1^\star = \sum_{i=1}^{I} c_{ij} \, \frac{1}{\overline{c}_i} \, u_{i1}^\star \tag{4.1.1}$$

perché l'autovettore $\overline{\mathbf{c}} = \mathbf{u}_{\mathrm{I}}^\star$ è $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale ad $\mathbf{u}_1^\star$ per costruzione. Perciò, in base ai risultati della Sez. 3.14, si vede che un profilo ha la *medesima* ascissa su $\mathbf{u}_1^\star$ che ha origine in $\overline{\mathbf{c}}$, e su $\mathbf{u}_1$, che ha invece origine in $\mathbf{0}_{\mathrm{I}}$, come è mostrato nella TAV. 23 e che questa ascissa coincide con l'ascissa su $\mathbf{u}_1^\star$ del profilo $\mathbf{c}_j - \overline{\mathbf{c}}$ centrato rispetto al baricentro. In valore assoluto $|g_{j1}|$ indica la distanza distribuzionale, misurata lungo il primo asse, che intercorre tra la proiezione $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale $g_{j1} \, \mathbf{u}_1^\star$ e l'origine degli assi fattoriali, origine che coincide col baricentro della nuvola dei profili, come si è visto nella Sez. 3.14.

Ripetendo la proiezione per tutti i profili, ossia facendo variare $j = 1, 2, \ldots, J$ nella (4.1.1), sul primo asse si ottengono le coordinate

$$g_{11} \; g_{21} \; \ldots \; g_{j1} \; \ldots \; g_{J1}. \tag{4.1.2}$$

---

[1] Il primo indice di una coordinata indica il profilo, il secondo il rango dell'autovettore e del corrispondente asse fattoriale. Così, g_{31} è l'ascissa di $\mathbf{c}_3$ sul primo asse.

Viene quindi a stabilirsi una corrispondenza[1] tra i J profili e le loro J coordinate sul primo asse fattoriale che, in termini geometrici, è una operazione di proiezione, ottenuta tramite il vettore $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_1^\star$ di ordine I, come appare dalla (4.1.1). Esso proietta ciascun profilo $\mathbf{c}_j$, ove $j = 1, 2, \ldots, J$, nel punto di ascissa g_{j1}, detta *primo fattore principale*[2] *del profilo colonna* $\mathbf{c}_j$.

Torna utile considerare l'insieme ordinato (4.1.2) come un vettore di ordine J, ottenuto dai J prodotti scalari

$$\mathbf{g}_1 = (\mathbf{C} - \overline{\mathbf{C}})^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_1^\star = (g_{11}\ g_{21}\ \ldots\ g_{j1}\ \ldots\ g_{J1})^T. \tag{4.1.3}$$

Il vettore $\mathbf{g}_1$, vettore delle coordinate o dei primi fattori dei profili colonna, è detto vettore dei primi fattori dei profili delle colonne o, brevemente, *primo fattore delle colonne.*

Quanto fatto per il primo asse fattoriale, può essere ripetuto per gli assi successivi, ottenendo l'a^{mo} fattore delle colonne, ossia il vettore

$$\mathbf{g}_a = (\mathbf{C} - \overline{\mathbf{C}})^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\mathbf{u}_a^\star = (g_{1a}\ g_{2a}\ \ldots\ g_{ja}\ g_{Ja})^T, \tag{4.1.4}$$

delle coordinate dei profili colonna, sull'asse fattoriale a. Una differenza importante da sottolineare è che mentre l'ascissa di un profilo non poteva superare 1 su ogni vettore della base unitaria canonica, come si è rilevato nella Sez. 2.7, ora può capitare, anche se raramente, che l'ascissa di un profilo su un asse fattoriale superi 1. Dai vettori $\mathbf{g}_a$, per ogni $a = 1, 2, \ldots, A = I - 1$, si ottiene la *matrice dei fattori delle colonne* di ordine $J \times A$,

$$\mathbf{G} = \begin{array}{c} \begin{array}{cccccc} \mathbf{g}_1 & \mathbf{g}_2 & \cdots & \mathbf{g}_a & \cdots & \mathbf{g}_A \end{array} \\ \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1a} & \cdots & g_{1A} \\ g_{21} & g_{22} & \cdots & g_{2a} & \cdots & g_{2A} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ g_{j1} & g_{j2} & \cdots & g_{ja} & \cdots & g_{jA} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ g_{J1} & g_{J2} & \cdots & g_{Ja} & \cdots & g_{JA} \end{pmatrix} \end{array} \tag{4.1.5}$$

---

[1] Dal punto di vista matematico, la corrispondenza tra profili e coordinate fattoriali è una *funzione*, perché associa ad *ogni* profilo $\mathbf{c}_j$ dell'insieme dei J profili-colonna (il dominio della funzione) *il* numero g_{j1} dell'insieme reale (il codominio della funzione). Le coordinate fattoriali sono le immagini dei profili.

[2] L'aggettivo 'principale', è abitualmente sottinteso. Viene utilizzato qualora occorra distinguere queste coordinate dalle coordinate 'standard' che saranno definite nella prossima Sez. 4.2.

Nel caso specifico dell'esempio, la matrice $\mathbf{C}$ dei profili colonna è nella Tav. 14, la matrice $\mathbf{D}_{\overline{\mathbf{c}}}^{-1}$ è stata ottenutata nella Sez. 2.8 ed il primo autovettore $\mathbf{u}_1^\star$ nella Sez. 3.14, per cui il primo fattore, in base alla (4.1.3), risulta

$$\mathbf{g}_1 = (\mathbf{C} - \overline{\mathbf{C}})^T \mathbf{D}_{\overline{\mathbf{c}}}^{-1} \mathbf{u}_1^\star = \mathbf{C}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-1} \mathbf{u}_1^\star =$$

$$\begin{pmatrix} 0.524 & 0.245 & 0.231 \\ 0.610 & 0.220 & 0.171 \\ 0.527 & 0.234 & 0.239 \\ 0.636 & 0.152 & 0.212 \\ 0.534 & 0.291 & 0.175 \\ 0.573 & 0.224 & 0.203 \\ 0.452 & 0.355 & 0.194 \\ 0.622 & 0.156 & 0.222 \end{pmatrix} \begin{pmatrix} 1.825 & 0 & 0 \\ 0 & 4.237 & 0 \\ 0 & 0 & 4.630 \end{pmatrix} \begin{pmatrix} -0.498 \\ 0.266 \\ 0.232 \end{pmatrix} = \begin{pmatrix} 0.049 \\ -0.122 \\ 0.043 \\ -0.179 \\ 0.032 \\ -0.049 \\ 0.198 \\ -0.150 \end{pmatrix}$$

Perciò $g_{11} = 0.049$ è l'ascissa del primo profilo, relativo alle quote di spettatori a rappresentazioni di Prosa, $g_{21} = -0.122$ l'ascissa del secondo profilo, relativo delle quote di spettatori a spettacoli di Lirica e Balletto, e così via.

Proiettando gli otto profili sul secondo autovettore $\mathbf{u}_2^\star$, ottenuto nella Sez. 3.15, si ottiene il secondo fattore delle colonne per la matrice d'esempio

$$\begin{aligned} \mathbf{g}_2 &= (\mathbf{C} - \overline{\mathbf{C}})^T \mathbf{D}_{\overline{\mathbf{c}}}^{-1} \mathbf{u}_2^\star = \mathbf{C}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-1} \mathbf{u}_2^\star \\ &= (0.012 \ -0.049 \ 0.040 \ 0.111 \ -0.142 \ -0.005 \ -0.200 \ 0.122)^T \end{aligned}$$

qui scritto in forma trasposta per ragione di spazio.

La *distanza* distribuzionale tra due profili può calcolarsi ora anche tramite le coordinate fattoriali che essendo riferite ad una base $\mathbf{D}_{\overline{\mathbf{c}}}^{-1}$-ortonormale, per le considerazioni svolte nella Sez. 3.6, si esprimono così

$$d_D^2(\mathbf{c}_j, \mathbf{c}_k) = \sum_{a=1}^{2} (g_{ja} - g_{ka})^2.$$

Ad esempio la distanza tra Concerti di Musica Classica ($j = 3$) e di Musica Leggera ($j = 6$) risulta

$$d_D^2(\mathbf{c}_3, \mathbf{c}_6) = (0.043 + 0.049)^2 + (0.040 + 0.005)^2 = 0.0083 + 0.0020 = 0.0103$$

che ovviamente è il medesimo valore trovato nella Sez. 2.9 tramite l'espressione (2.8.2) della distanza distribuzionale.

4.2 - Proprietà dei fattori

I fattori delle colonne godono di tre notevoli proprietà, riassunte in questo schema, dove $\overline{\mathbf{r}}$ indica il vettore di ordine J delle masse dei profili colonna e $\mathbf{D}_{\overline{\mathbf{r}}}$ la matrice diagonale da esso ottenuta, definita nella Sez. 2.6,

$$\overline{\mathbf{r}}^T \mathbf{g}_a = \sum_{j=1}^{J} \overline{r}_j \, g_{ja} = 0 \qquad (a = 1, \ldots, A)$$

$$\mathbf{g}_a^T \mathbf{D}_{\overline{\mathbf{r}}} \mathbf{g}_a = \sum_{j=1}^{J} \overline{r}_j \, g_{ja}^2 = \lambda_a \qquad (a = 1, \ldots, A)$$

$$\mathbf{g}_a^T \mathbf{D}_{\overline{\mathbf{r}}} \mathbf{g}_b = \sum_{j=1}^{J} \overline{r}_j \, g_{ja} \, g_{jb} = 0. \quad (a, b = 1, \ldots, A; b \neq a)$$

La prima proprietà riguarda la media ponderata delle coordinate dei J profili su un asse fattoriale, e questa è nulla perché, tenendo presente la (4.1.4), la (2.6.4) e la (3.12.2) risulta

$$\overline{\mathbf{r}}^T \mathbf{g}_a = \mathbf{g}_a^T \overline{\mathbf{r}} = \mathbf{g}_a^T \mathbf{D}_{\overline{\mathbf{r}}} \mathbf{1}_{\mathrm{J}} = \mathbf{u}_a^{\star T} \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{C} \mathbf{D}_{\overline{\mathbf{r}}} \mathbf{1}_{\mathrm{J}} = \mathbf{u}_a^{\star T} \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \overline{\mathbf{c}} = 0.$$

Questa proprietà vale per ciascun fattore, per cui

$$\overline{\mathbf{r}}^T \mathbf{G} = \mathbf{G}^T \mathbf{D}_{\overline{\mathbf{r}}} \mathbf{1}_{\mathrm{J}} = \mathbf{0}_{\mathrm{A}}.$$

Perciò il baricentro di ciascun fattore coincide con l'origine dell'asse fattoriale, e quindi col baricentro $\overline{\mathbf{c}}$ della nuvola dei profili delle colonne.

La seconda proprietà esprime un risultato già acquisito fin dalla Sez. 3.15 e riguarda l'inerzia dei fattori: la somma ponderata dei quadrati delle coordinate su un asse è l'autovalore che corrisponde all'autovettore che individua l'asse. La proprietà deriva dalla (4.1.4), dalla (2.6.4) e da quanto ottenuto nella Sez. 3.15,

$$\begin{aligned} \mathbf{g}_a^T \mathbf{D}_{\overline{\mathbf{r}}} \mathbf{g}_a = \mathbf{u}_a^{\star T} \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{C} \mathbf{D}_{\overline{\mathbf{r}}} \mathbf{C}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{u}_a^{\star} &= \mathbf{u}_a^{\star T} \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{C} \mathbf{R}^T \mathbf{u}_a^{\star} \\ &= \mathbf{u}_a^{\star T} \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \lambda_a \mathbf{u}_a^{\star} = \lambda_a \end{aligned}$$

In conclusione ogni fattore ha valore medio nullo e varianza λ_a .

La terza importante proprietà non interessa più il singolo fattore, ma *coppie* di fattori e si dimostra tenendo presente la (3.2.4) e la (3.14.6)

$$\begin{aligned} \mathbf{g}_a^T \mathbf{D}_{\overline{\mathbf{r}}} \mathbf{g}_b = \mathbf{u}_a^{\star T} \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{C} \mathbf{D}_{\overline{\mathbf{r}}} \mathbf{C}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{u}_b^{\star} &= \mathbf{u}_a^{\star T} \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{C} \mathbf{R}^T \mathbf{u}_b^{\star} \\ = \mathbf{u}_a^{\star T} \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \lambda_b \mathbf{u}_b^{\star} = \lambda_b \, 0 = 0. \end{aligned}$$

è detta proprietà di *ortogonalità* dei fattori delle colonne: i fattori delle colonne risultano ortogonali due a due. Si noti: ortogonali, *non* $\mathbf{D}_{\mathbf{c}}^{-1}$-ortogonali. Questa proprietà avrà un'importante conseguenza, come si vedrà nella Sez. 4.4

Se con $\mathbf{D}_\lambda$ si indica la matrice diagonale, di ordine $A \times A$, ottenuta dai primi A autovalori non nulli

$$\mathbf{D}_\lambda \stackrel{\text{def}}{=} diag\,(\lambda_1\ \lambda_2\ \ldots\ \lambda_A) = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_a & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & \ldots & \lambda_A \end{pmatrix} \tag{4.2.1}$$

la seconda e la terza proprietà si possono riassumere così

$$\underset{A\times J}{\mathbf{G}^T}\ \underset{J\times J}{\mathbf{D}_{\overline{\mathbf{r}}}}\ \underset{J\times A}{\mathbf{G}} = \underset{A\times A}{\mathbf{D}_\lambda}\,. \tag{4.2.2}$$

La matrice di ordine $A \times A$ al primo membro è detta in Statistica matrice di *varianza e covarianza*. Se nell'espressione della terza proprietà dei fattori si dividono ambo i membri per le deviazioni standard $\sqrt{\lambda}_a$ e $\sqrt{\lambda}_b$ che sono quantità positive, si ottiene

$$\sum_{j=1}^{J} \overline{r}_j\, \frac{(g_{ja} - 0)}{\sqrt{\lambda}_a}\, \frac{(g_{jb} - 0)}{\sqrt{\lambda}_b} = 0 \qquad (b \neq a)$$

il che indica come i fattori delle colonne risultino, per costruzione, *non correlati* linearmente due a due. Attenzione però, non correlati linearmente significa che non sono legati da una relazione lineare, ma questo non esclude che la relazione possa essere *non* lineare, ad esempio quadratica, come si vedrà nella Sez. 8.2

4.3 - Fattori standard dei profili delle colonne

I fattori (principali) delle colonne sono normalizzati ad avere inerzia, o varianza, pari all'autovalore che corrisponde all'autovettore che individua l'asse. In alcune applicazioni, tuttavia, può tornar utile considerare anche i fattori normalizzati ad avere inerzia unitaria, per cui

$$\hat{\mathbf{g}}_a \stackrel{\text{def}}{=} \frac{1}{\sqrt{\lambda}_a}\, \mathbf{g}_a \qquad (a = 1, 2, \ldots, A)$$

è detto a^{mo} fattore *standardizzato*, o *standard*, dei profili delle colonne. La sua componente $\hat{g}_{aj}$ è l' a^{mo} fattore standard del profilo $\mathbf{c}_j$ e $\hat{\mathbf{G}}$, ottenuta come la (4.1.5), è la matrice di ordine $J \times A$ dei fattori standard delle colonne. Poiché gli autovalori sono sempre inferiori ad 1, come si vedrà nella Sez. 4.9, il coefficiente $1/\sqrt{\lambda_a}$ risulta maggiore di 1, per cui le coordinate standard $\hat{\mathbf{g}}_a$ risultano sempre *più disperse* delle omologhe $\mathbf{g}_a$. Così per la matrice d'esempio *Spettacoli-3* di TAV.14, tutte le ascisse degli 8 profili dei tipi di spettacolo risultano amplificate di 14.40 volte sul primo asse e di 24.45 sul secondo.

Per distinguerli dai fattori standard, i fattori $\mathbf{g}_a$, definiti nella (4.1.4) sono detti *fattori principali*, ma in tutti i casi in cui non vi sia ambiguità, l'aggettivo 'principale' viene abitualmente omesso.

Su ogni asse $a = 1, \ldots, A$ i fattori standard delle colonne hanno media ponderata nulla come i fattori principali, ma inerzia unitaria

$$\overline{\mathbf{r}}^T \hat{\mathbf{g}}_a = \sum_{j=1}^{J} \overline{r}_j \, \hat{g}_{ja} = 0 \qquad\qquad \underset{A\times J}{\hat{\mathbf{G}}^T} \; \underset{J\times J}{\mathbf{D}_{\overline{\mathbf{r}}}} \; \underset{J\times A}{\hat{\mathbf{G}}} \; = \; \underset{A\times A}{\mathbf{I}}.$$

La matrice di varianza e covarianza è ridotta ora alla matrice identità. Le dimostrazioni sono analoghe a quelle per i fattori principali, a meno del coefficiente $1/\sqrt{\lambda_a}$.

4.4 - Rappresentazione grafica dei profili delle colonne

Il lettore che fin dal secondo Capitolo ha pazientemente seguito la scomposizione fattoriale di una matrice di contingenza, si sarà spesso chiesto quale ne sia l'utilità pratica. Finalmente, in questa Sez. verrà mostrato come i fattori possano avere una rappresentazione grafica, peculiarità questa che ha fatto la fortuna dei metodi fattoriali. Si è in grado così di esaminare le posizioni dei profili, e quindi la struttura delle loro relazioni, in sottospazi accessibili alla nostra percezione, ad esempio su un singolo asse fattoriale individuato dall'autovettore $\mathbf{u}_a^\star$. La coordinata della proiezione del profilo $\mathbf{c}_j$ è allora g_{ja}. L'inerzia complessiva su questo asse è l'autovalore λ_a e τ_a il corrispondente tasso d'inerzia, definito nella Sez. 3.15. Una migliore cognizione della configurazione dei profili può ottenersi esaminando le proiezioni dei profili in uno spazio bi-dimensionale, per esempio sul piano individuato dalla coppia di autovettori $\mathbf{u}_a^\star$ e $\mathbf{u}_b^\star$. La proiezione del profilo $\mathbf{c}_j$ ha allora le coordinate $(\, g_{ja}\,,\; g_{jb}\,)$. L'inerzia sul piano è $\lambda_a + \lambda_b$ ed il tasso d'inerzia

$\tau_a + \tau_b$. Quando si disponga del software adatto è sempre utile osservare la struttura anche in spazi fattoriali tridimensionali.

Passando da una rappresentazione mono-dimensionale ad una bi- o tri-dimensionale, sorge il problema di come rappresentare gli assi coordinati di riferimento, perché gli autovettori che individuano gli assi fattoriali *non* sono ortogonali, ma $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonali a coppie, per costruzione. [1] D'altro canto, la rappresentazione grafica mira unicamente ad evidenziare le *distanze* distribuzionali tra profili, perché sono queste che traducono il loro grado di similarità. Nella Sez. 3.6 si è visto che se la base di riferimento è $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortonormale, la distanza distribuzionale tra profili è computata come semplice somma dei quadrati delle differenze tra coordinate, come si fa abitualmente nel nostro spazio ordinario. Tutto ciò autorizza a rappresentare sulle mappe gli assi fattoriali come rette *ortogonali* orientate. L'orientamento è arbitrario, perché tale è quello degli autovettori come si è visto nella Sez. 3.13. Il punto d'incrocio degli assi coordinati indica il baricentro della nuvola dei profili e l'unità di distanza su ogni asse è stabilita dalla lunghezza dell'autovettore che lo individua.

Tra i piani generati dalle coppie di assi fattoriali, quello individuato dai primi due ha la quota più rilevante dell'inerzia complessiva $In_{\overline{\mathbf{c}}}$ e riproduce quindi con minore distorsione le effettive distanze tra i punti della nuvola. La porzione di piano che contiene la rappresentazione dei profili è detta *mappa* principale.

Nel caso dell'esempio di TAV. 14 gli autovettori sono soltanto due, per cui $A = 2$ e la mappa principale riporta le distanze distribuzionali senza alcuna distorsione. I fattori degli 8 profili colonna sono stati calcolati nella Sez. 4.1, per cui le coordinate dei punti risultano

$$\mathbf{c}_1 = g_{11}\,\mathbf{u}_1^\star + g_{12}\,\mathbf{u}_2^\star = \quad 0.049\,\mathbf{u}_1^\star + 0.012\,\mathbf{u}_2^\star \qquad \mathbf{c}_1 \equiv (\ \ 0.049,\ \ 0.012)$$
$$\mathbf{c}_2 = g_{21}\,\mathbf{u}_1^\star + g_{22}\,\mathbf{u}_2^\star = -0.122\,\mathbf{u}_1^\star - 0.049\,\mathbf{u}_2^\star \qquad \mathbf{c}_2 \equiv (-0.122, -0.049)$$
$$\ldots = \ldots$$
$$\mathbf{c}_8 = g_{81}\,\mathbf{u}_1^\star + g_{82}\,\mathbf{u}_2^\star = -0.150\,\mathbf{u}_1^\star + 0.122\,\mathbf{u}_2^\star \qquad \mathbf{c}_8 \equiv (-0.150,\ \ 0.122)$$

Questi risultati sono riportati per esteso nella mappa di TAV. 24. I

---

[1] Per rendersene conto basta tracciare nella TAV. 19 il segmento che collega l'origine $\mathbf{0}_2$ col baricentro $\overline{\mathbf{c}}$. Questo segmento, che rappresenta graficamente il vettore $\overline{\mathbf{c}}$, *non* risulta ortogonale al segmento del simplesso, mentre, come si è mostrato nella Sez. 3.6, il vettore $\overline{\mathbf{c}}$ è $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale al segmento del simplesso.

punti occupano una regione molto ridotta del simplesso triangolare, come lasciava prevedere l'esiguo valore 0.07 dell'inerzia della nuvola. Le distanze tra due punti traducono senza distorsioni, ma solo in questo caso, le effettive distanze distribuzionali tra profili che rappresentano, ma, in generale, lo spazio A-dimensionale che li contiene non sarà un piano, per cui le distanze tra profili saranno raffigurate soltanto approssimativamente sulla mappa. Per valutare il grado di fedeltà della rappresentazione, e poter quindi valutare correttamente le associazioni tra modalità, è *indispensabile* giovarsi di strumenti diagnostici. Sono indicatori numerici del grado di distorsione e sono un po' come il bastone bianco del cieco per l'analista che dall'esame di mappe piane cerca di risalire all'effettiva configurazione dei profili nel loro oscuro spazio multidimensionale.

4.5 - Contributo relativo

La dispersione geometrica delle proiezioni dei J profili della nuvola sull'asse fattoriale individuato dall'autovettore $\mathbf{u}_a^\star$ è misurata dall'inerzia

$$\lambda_a = \sum_{j=1}^{J} \overline{r}_j \, g_{ja}^2. \tag{4.5.1}$$

alla quale ogni singolo profilo $\mathbf{c}_j$ contribuisce col termine $\overline{r}_j \, g_{ja}^2$, detto *contributo assoluto* del profilo a λ_a. Ciascun termine può essere espresso relativamente al totale λ_a, ottenendo il *contributo relativo* del profilo $\mathbf{c}_j$ all'inerzia dell'asse di rango a, e definito dal rapporto, sempre positivo e variabile tra 0 e 1,

$$CTR_a(\mathbf{c}_j) \stackrel{\text{def}}{=} \frac{\overline{r}_j \, g_{ja}^2}{\lambda_a} = \frac{\overline{r}_j \, g_{ja}^2}{\sum_{j=1}^{J} \overline{r}_j \, g_{ja}^2} \tag{4.5.2}$$

che indica la quota d'inerzia sull'asse che è riconducibile al profilo $\mathbf{c}_j$. Questo è il più prezioso degli indicatori, perché permette di ordinare i profili secondo l'importanza del ruolo avuto nell'orientare l'autovettore dell'asse. Si può immaginare ogni profilo $\mathbf{c}_j$ come dotato di un 'potere magnetico' d'attrazione degli assi, dipendente dalla posizione e dalla massa. Il contributo relativo indica quindi lo "sforzo", valutato in termini d'inerzia, compiuto da ciascun profilo per attirare verso di sè l'asse fattoriale. Ciò che conta è il *prodotto* della massa per il quadrato della coordinata, ma il contributo relativo è tanto più significativo, quanto più risulta superiore alla massa.

Grazie al contributo relativo, ci si può render conto se l'asse è stato orientato da pochi profili con elevato contributo, o da più profili con contributi

più o meno bilanciati.

Per ogni asse fattoriale, la somma dei contributi relativi di tutti i J profili, per la (4.5.1), vale

$$\sum_{j=1}^{J} CTR_a(\mathbf{c}_j) = \frac{\sum_{j=1}^{J} \overline{r}_j \, g_{ja}^2}{\lambda_a} = 1.$$

Per calcolare il contributo relativo $CTR_{(a,b)}(\mathbf{c}_j)$ del profilo $\mathbf{c}_j$ all'inerzia del *piano* individuato dagli assi fattoriali a e b, si può partire dalla scomposizione dell'inerzia

$$\overline{r}_j \, (g_{ja}^2 + g_{jb}^2) = \overline{r}_j \, g_{ja}^2 + \overline{r}_j \, g_{jb}^2$$

che divisa per l'inerzia sul piano, grazie alla (4.5.2), fornisce

$$CTR_{(a,b)}(\mathbf{c}_j) = \frac{\overline{r}_j \, (g_{ja}^2 + g_{jb}^2)}{\lambda_a + \lambda_b} = \frac{\lambda_a}{\lambda_a + \lambda_b} \, CTR_a(\mathbf{c}_j) + \frac{\lambda_b}{\lambda_a + \lambda_b} \, CTR_b(\mathbf{c}_j).$$

Espressioni analoghe valgono per spazi a tre o più dimensioni.

La TAV. 25 riporta i contributi relativi degli 8 profili delle colonne della matrice d'esempio *Spettacoli-3* di TAV. 14. Due degli 8 profili contribuiscono per oltre il 50% all'inerzia dei due assi e quindi a definire il loro orientamento: gli spettacoli di Lirica e Balletti, $CTR_1(\mathbf{c}_2) = 34\%$, e le rappresentazioni di Prosa, $CTR_1(\mathbf{c}_1) = 21\%$, per il primo asse e le Riviste e Commedie Musicali, $CTR_2(\mathbf{c}_5) = 35\%$, e gli spettacoli di Burattini e Marionette, $CTR_2(\mathbf{c}_7) = 21\%$, per il secondo. Dei due profili che maggiormente hanno contribuito ad orientare il primo asse, il primo, la Lirica, ha una massa piuttosto esigua, $\overline{r}_2 = 0.109$, ma si trova lontano dall'origine, $|\, g_{12} \,| = 0.122$ dalla parte negativa dell'asse, mentre il secondo, la Prosa, ha la massa più grande, $\overline{r}_1 = 0.419$, anche se è vicino, $|\, g_{11} \,| = 0.049$, all'origine. È piuttosto frequente il fatto che siano i punti più eccentrici ad orientare i primi assi fattoriali, particolarmente nel caso di matrici di ridotte dimensioni come questa.

4.6 - Coseno quadrato

La *qualità* della rappresentazione del profilo $\mathbf{c}_j$ sull'asse fattoriale a è misurata dal rapporto dei quadrati di due distanze: quella calcolata sull'asse tra la proiezione del profilo e l'origine e quella, calcolata in $\Re^I$, tra il profilo e il baricentro. Abitualmente questo rapporto viene indicato con $COS_a^2(\mathbf{c}_j)$ perché, come mostrato nella TAV. 26, può interpretarsi geometricamente come il coseno quadrato dell'angolo tra il vettore $\mathbf{c}_j$, inteso come

segmento orientatato che individua la posizione del profilo, e l'autovettore $\mathbf{u}_a^\star$ che individua l'asse

$$COS_a^2(\mathbf{c}_j) \stackrel{\text{def}}{=} \frac{(g_{ja}-0)^2}{d_D^2(\mathbf{c}_j,\overline{\mathbf{c}})} = \frac{g_{ja}^2}{\sum_{a=1}^A g_{ja}^2}. \tag{4.6.1}$$

Questo indicatore, sempre positivo, varia dunque tra 0 ed 1 e *non* dipende dalla massa del profilo. Quando assume il valore 0 significa che il profilo si trova nello spazio complementare $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonale all'asse, mentre il valore 1 indica che il profilo si trova proprio sull'asse. Un valore vicino ad 1 indica che il profilo è prossimo all'asse e che quindi $|g_{ja}|$ riproduce abbastanza bene l'effettiva distanza $d_D(\mathbf{c}_j,\overline{\mathbf{c}})$ del profilo dal baricentro.

Se nella (4.6.1) si moltiplica numeratore e denominatore per la massa del profilo, si ottiene

$$COS_a^2(\mathbf{c}_j) = \frac{\overline{r}_j\, g_{ja}^2}{\sum_{a=1}^A \overline{r}_j\, g_{ja}^2}. \tag{4.6.2}$$

ove il numeratore rappresenta l'inerzia del profilo $\mathbf{c}_j$ sull'asse di rango a, ossia in un sottospazio monodimensionale, mentre il denominatore rappresenta l'inerzia del profilo scomposta su tutti gli A assi fattoriali. Il loro rapporto indica quindi la quota d'inerzia del profilo dovuta al contributo dell'asse. Per questo $COS_a^2(\mathbf{c}_j)$ viene anche detto contributo relativo *dell'asse all'inerzia del profilo.* Oltre che per la massa, la Qualità della rappresentazione, espressa dalla (4.6.2), e il Contributo relativo, espresso dalla (4.5.1), differiscono per il campo d'azione della somma al denominatore: sugli A assi per il contributo relativo del profilo all'inerzia dell'asse e sui J profili per il Contributo Relativo del profilo all'inerzia dell'asse.

La scomposizione dell'inerzia della nuvola dei profili secondo i profili e secondo gli assi è schematizzata nella TAV. 27.

La Qualità della rappresentazione di un profilo sul *piano* individuato dagli assi fattoriali a e b si ottiene dalla somma dei contributi sui due assi

$$COS_{(a,b)}^2(\mathbf{c}_j) = \frac{g_{ja}^2+g_{jb}^2}{\sum_{a=1}^A g_{ja}^2} = COS_a^2(\mathbf{c}_j) + COS_b^2(\mathbf{c}_j) \tag{4.6.3}$$

e quindi vale 1 nello spazio $\Re^I$

$$\sum_{a=1}^A COS_a^2(\mathbf{c}_j) = \frac{\sum_{a=1}^A g_{ja}^2}{\sum_{a=1}^A g_{ja}^2} = 1.$$

Nella caratterizzazione di assi e piani andranno presi in considerazione soltanto i profili con una elevata qualità di rappresentazione, per garantirsi che le distanze tra proiezioni rappresentino con un grado di fedeltà accettabile le distanze reali.

La TAV. 25 riporta i valori dell'indicatore per i primi due assi. Sul primo sono ben rappresentati i profili dei Concerti di Musica Leggera che hanno un $COS_1^2(\mathbf{c}_6) = 0.991$, e della Prosa con $COS_1^2(\mathbf{c}_1) = 0.940$. Sul secondo il profilo Rivista e Commedia Musicale con $COS_2^2(\mathbf{c}_5) = 0.953$. Le coordinate delle proiezioni di questi profili sugli assi riproducono in modo eccellente le distanze del profilo dal baricentro nello spazio bidimensionale. È evidente che sulla mappa fattoriale gli indicatori $COS_{(1,2)}^2(\mathbf{c}_j) = 1.000$ per tutti i J profili perché tutti i punti giacciono sul piano individuato dai primi due assi fattoriali: la rappresentazione delle distanze reali è quindi perfetta.

4.7 - Qualità e inerzia di un profilo

Molto spesso è importante conoscere quanto un profilo sia ben rappresentato in un *sottospazio* individuato dai *primi* $A^* \leq A$ assi fattoriali, perché se risulta ad esempio, che un profilo è ben rappresentato nel sottospazio tridimensionale, è inutile indagarne la posizione in sottospazi che coinvolgano assi di ordine superiore al terzo. Il numero A^* di assi può essere imposto in quasi tutti i programmi d'analisi.

La *qualità della rappresentazione* di un profilo in un sottospazio si valuta mediante un indicatore aggregato che varia tra 0 ed 1 e che è una generalizzione della (4.6.3),

$$QLT_{A^*}(\mathbf{c}_j) \stackrel{\text{def}}{=} \sum_{a=1}^{A^*} COS_a^2(\mathbf{c}_j) = \frac{\sum_{a=1}^{A^*} g_{ja}^2}{\sum_{a=1}^{A} g_{ja}^2}.$$

Le distanze delle proiezioni dal baricentro risultano di solito inferiori a quelle reali, perché

$$d_D(g_{j1}\,\mathbf{u}_1^\star + g_{j2}\,\mathbf{u}_2^\star \ldots + g_{jA^*}\,\mathbf{u}_{A^*}^\star, \overline{\mathbf{c}}) \leq d_D(\mathbf{c}_j, \overline{\mathbf{c}})$$
$$g_{j1}^2 + g_{j2}^2 + \ldots + g_{jA^*}^2 \leq g_{j1}^2 + g_{j2}^2 + \ldots + g_{jA^*}^2 + \ldots + g_{jA}^2$$

e l'uguaglianza si ha solo quando il profilo coincide con la sua proiezione e si trova già nel sottospazio A^*-dimensionale. In tal caso $QLT_{A^*}(\mathbf{c}_j) = 1$ e la rappresentazione della distanza del profilo nel sottospazio generato dai primi A^* assi è perfetta. Se non è così, la rappresentazione sarà tanto più fedele quanto più $QLT_{A^*}(\mathbf{c}_j)$ si avvicina all'unità e, comunque, se $A^* = A$, allora

$QLT_A(\mathbf{c}_j) = 1$. Questo indicatore si rivelerà prezioso per i profili illustrativi, come si vedrà nella Sez. 4.12.

Un indizio sulla localizzazione di un profilo all'interno della nuvola si può ottenere confrontando la massa del profilo con la *quota d'inerzia* che questi ha nel sottospazio A^*-dimensionale. La sua inerzia viene riferita all'inerzia totale della nuvola nello spazio A-dimensionale, come calcolata nella Sez. 3.3, per cui si definisce

$$INR_{A^*}(\mathbf{c}_j) \stackrel{\text{def}}{=} \frac{\sum_{a=1}^{A^*} \overline{r}_j \, g_{ja}^2}{In_{\overline{\mathbf{c}}}} = \overline{r}_j \, \frac{\sum_{a=1}^{A^*} g_{ja}^2}{In_{\overline{\mathbf{c}}}}.$$

Quando la massa $\overline{r}_j$ è nettamente inferiore a $INR_{A^*}(\mathbf{c}_j)$ allora nel sottospazio A^*-dimensionale il profilo deve trovarsi lontano dal baricentro. Questo fornisce un mezzo per individuare tra i punti della nuvola quelli eccentrici e quelli centrali, per i quali la massa è nettamente superiore a $INR_{A^*}(\mathbf{c}_j)$.

Nel caso dell'esempio, quando $A^* = A = 2$ per i profili Prosa e Lirica risulta

j	*Profili* $\mathbf{c}_j$	*Masse* $\overline{r}_j$	*Inerzia* $INR_2(\mathbf{c}_j)$
1	*Prosa*	0.42	0.15
2	*Lirica*	0.11	0.27

Perciò Prosa è un punto centrale della nuvola dei profili e Lirica un punto periferico.

Gli indicatori presentati in queste ultime Sezioni sono strumenti indispensabili per una corretta interpretazione dei risultati dell'Analisi delle Corrispondenze. Ma perché il lettore possa rendersi conto pienamente della loro importanza e, sopratutto, possa vedere in pratica il loro impiego, occorre disporre anche dei risultati dell'analisi della matrice dei profili delle righe perché, come si vedrà nella Sez. 4.9, è peculiare dell'Analisi delle Corrispondenze il fatto che le proiezioni dei profili delle colonne e delle righe su un asse vadano interpretate *congiuntamente*.

4.8 - Analisi dei profili delle righe

A partire dal secondo Capitolo, tutte le Sezioni fino a questa sono state dedicate ad illustrare dettagliatamente l'analisi dei J profili delle colonne nella matrice $\mathbf{C}$. Un'esposizione analoga può farsi, in modo del tutto simmetrico, per gli I profili delle righe nella matrice $\mathbf{R}$ definita nella Sez. 1.5. Questa perfetta simmetria giustifica il nome di Analisi delle Corrispondenze.

Coerentemente con la definizione di vettore della Sez. 2.2, il generico i^{mo} profilo riga $\mathbf{r}_i$ è un *vettore colonna* di ordine J che si indica con

$$\mathbf{r}_i = \begin{pmatrix} r_{i1} \\ \vdots \\ r_{iJ} \end{pmatrix} \qquad \text{o, equivalentemente, con} \qquad \mathbf{r}_i = (r_{i1} \ \ldots \ r_{ij} \ \ldots r_{iJ})^T.$$

Le righe della matrice $\mathbf{R}$ sono perciò la forma trasposta di questi vettori, per cui

$$\mathbf{r}_i^T = (r_{i1} \ r_{i2} \ \ldots \ r_{ij} \ \ldots r_{iJ}) \qquad \text{e quindi} \qquad \mathbf{R} = \begin{pmatrix} \mathbf{r}_1^T \\ \vdots \\ \mathbf{r}_I^T \end{pmatrix}.$$

La loro media ponderata è il profilo $\overline{\mathbf{r}}$, vettore di ordine J delle masse dei profili delle colonne, definito nella Sez. 2.6,

$$\overline{\mathbf{r}} \stackrel{\text{def}}{=} \sum_{i=1}^{I} \overline{c}_i \, \mathbf{r}_i = \mathbf{R}^T \mathbf{D}_{\overline{\mathbf{c}}}^{-1} \mathbf{1}_{\mathrm{I}} \qquad \overline{r}_j = \sum_{i=1}^{I} \overline{c}_i \, r_{ij} = \frac{n_{+j}}{n_{++}}.$$

I profili delle righe possono interpretarsi geometricamente come una nuvola di I punti immersi in uno spazio euclideo J-dimensionale $\Re^J$ in cui il punto $\mathbf{0}_{\mathrm{J}}$ rappresenta l'origine della base canonica costituita dai J vettori unitari $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_J$, di ordine J, e r_{ij} è la coordinata del profilo $\mathbf{r}_i$ sul vettore $\mathbf{e}_j$. Le loro masse sono le I componenti $\overline{c}_i$ del profilo $\overline{\mathbf{c}}$ ed il loro baricentro è individuato dal profilo $\overline{\mathbf{r}}$. Così per la matrice d'esempio $\mathbf{R}$ di ordine 3×8 di TAV 14, la nuvola dei profili riga è costituita da 3 punti, relativi al Nord, al Centro e al Sud, immersi in uno spazio 8-dimensionale. Si tratta quindi di uno spazio completamente *diverso* da quello tri-dimensionale che contiene gli 8 profili colonna relativi ai tipi di spettacolo.

Anche tra i profili delle righe viene definita una distanza distribuzionale dello stesso *tipo* della (2.8.1)

$$d_D^2(\mathbf{r}_i, \mathbf{r}_k) \stackrel{\text{def}}{=} \sum_{j=1}^{J} \frac{1}{\overline{r}_j} (r_{ij} - r_{kj})^2 = \sum_{j=1}^{J} \frac{1}{\frac{n_{+j}}{n_{++}}} \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{kj}}{n_{k+}} \right)^2 \qquad (4.8.1)$$

che conferisce allo spazio $\Re^J$ una struttura euclidea. Quando due righe della matrice $\mathbf{N}$ sono proporzionali, i profili corrispondenti sono eguali, i due punti che li rappresentano coincidono in $\Re^J$ e la loro distanza è nulla.

La definizione di distanza (4.8.1) consente ora di dimostrare la proprietà *equidistributiva*, alla quale si è fatto cenno nella Sez. 2.9. Proprietà molto importante perché garantisce l'indipendenza dei risultati dalla codifica

delle modalità: la distanza distribuzionale tra righe (tra colonne) non cambia se si aggregano due colonne (due righe) con profili identici.

Si consideri una matrice di contingenza $\mathbf{N}$ di ordine $I \times J$ avente due righe i_1 e i_2 proporzionali. I corrispondenti profili $\mathbf{r}_{i_1}$ e $\mathbf{r}_{i_2}$ sono eguali e le loro massse siano $\overline{c}_{i_1}$ e $\overline{c}_{i_2}$. Di ogni elemento di $\mathbf{N}$ si può calcolare la frequenza relativa che può esprimersi in funzione delle componenti dei profili

$$\frac{n_{ij}}{n_{++}} = \frac{r_{ij}}{\overline{r}_j} = \frac{c_{ij}}{\overline{c}_i}.$$

In particolare, per le righe i_1 e i_2 di questa matrice, risulta che $r_{i_1 j} = r_{i_2 j}$ per qualunque colonna j, per cui

$$\frac{c_{i_1 j}}{\overline{c}_{i_1}} = \frac{r_{i_1 j}}{\overline{r}_j} = \frac{r_{i_2 j}}{\overline{r}_j} = \frac{c_{i_2 j}}{\overline{c}_{i_2}}.$$

Se ora le due righe proporzionali i_1 e i_2 di $\mathbf{N}$ vengono accorpate in una, i_0, la nuova matrice risulta di ordine $(I-1) \times J$, il corrispondente profilo $\mathbf{r}_0$ risulta eguale a $\mathbf{r}_{i_1}$ ed $\mathbf{r}_{i_2}$, mentre la massa $\overline{c}_{i_0}$ è pari alla somma delle due masse. Riassumendo, si ha questa situazione

	Matrice di ordine	
	$I \times J$	$(I-1) \times J$
profili	$\mathbf{r}_{i_1} = \mathbf{r}_{i_2}$	$\mathbf{r}_{i_0} = \mathbf{r}_{i_1} = \mathbf{r}_{i_2}$
masse	$\overline{c}_{i_1}$ e $\overline{c}_{i_2}$	$\overline{c}_{i_0} = \overline{c}_{i_1} + \overline{c}_{i_2}$
freq. relative	$\frac{c_{i_1 j}}{\overline{c}_{i_1}} = \frac{c_{i_2 j}}{\overline{c}_{i_2}}$	$\frac{c_{i_0 j}}{\overline{c}_{i_0}} = \frac{c_{i_1 j}}{\overline{c}_{i_1}} = \frac{c_{i_2 j}}{\overline{c}_{i_2}}$.

La distanza distribuzionale tra due profili colonna $\mathbf{c}_j$ e $\mathbf{c}_k$ della matrice di ordine $I \times J$ con due righe i_1 e i_2 eguali, è

$$d_D^2(\mathbf{c}_j, \mathbf{c}_k) = \sum_{i=1}^{I} \frac{(c_{ij} - c_{ik})^2}{\overline{c}_i} = \ldots + \frac{(c_{i_1 j} - c_{i_1 k})^2}{\overline{c}_{i_1}} + \frac{(c_{i_2 j} - c_{i_2 k})^2}{\overline{c}_{i_2}} + \ldots$$

e soltanto questi due termini della sommatoria interessano le righe i_1 e i_2. La loro somma, per le eguaglianze viste sopre, risulta

$$\begin{aligned} \frac{(c_{i_1 j} - c_{i_1 k})^2}{\overline{c}_{i_1}} + \frac{(c_{i_2 j} - c_{i_2 k})^2}{\overline{c}_{i_2}} &= \overline{c}_{i_1} \left(\frac{c_{i_1 j}}{\overline{c}_{i_1}} - \frac{c_{i_1 k}}{\overline{c}_{i_1}} \right)^2 + \overline{c}_{i_2} \left(\frac{c_{i_2 j}}{\overline{c}_{i_2}} - \frac{c_{i_2 k}}{\overline{c}_{i_2}} \right)^2 \\ &= (\overline{c}_{i_1} + \overline{c}_{i_2}) \left(\frac{c_{i_1 j}}{\overline{c}_{i_1}} - \frac{c_{i_1 k}}{\overline{c}_{i_1}} \right)^2 = \overline{c}_{i_0} \left(\frac{c_{i_0 j}}{\overline{c}_{i_0}} - \frac{c_{i_0 k}}{\overline{c}_{i_0}} \right)^2. \end{aligned}$$

La distanza distribuzionale tra due profili colonna $\mathbf{c}_j$ e $\mathbf{c}_k$ della nuova matrice di ordine $(I-1)\times J$ con la riga i_0 ottenuta dall'accorpamento, è

$$d_D^2(\mathbf{c}_j,\mathbf{c}_k)=\sum_{i=1}^{I-1}\frac{(c_{ij}-c_{ik})^2}{\overline{c}_i}=\ldots+\frac{(c_{i_0 j}-c_{i_0 k})^2}{\overline{c}_{i_0}}+\ldots$$

e questo è l'unico termine della sommatoria che interessa i_0. Ma, per le eguaglianze viste sopra questo termine risulta eguale a quello ottenuto nell'espressione della distanza distribuzionale per la matrice $I\times J$. Considerazioni analoghe si possono fare quando in $\mathbf{N}$ vengono accorpate due colonne proporzionali. Perciò la distanza distribuzionale tra due profili riga (colonna) non muta quando due profili colonna (riga) eguali vengono accorpati in uno con massa pari alla somma delle masse dei due profili. In altri termini, non c'è perdita d'informazione sulla configurazione geometrica della nuvola, quando si aggregano due profili eguali, né guadagno quando un profilo è egualmente suddiviso.

Il prodotto scalare tra due vettori nello spazio $\Re^J$ è definito come

$$\mathbf{r}_i^T\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\mathbf{r}_k\stackrel{\text{def}}{=}\sum_{j=1}^{J}\frac{1}{\overline{r}_j}\,r_{ij}\,r_{kj}$$

essendo

$$\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}=diag\left(\frac{1}{\overline{r}_1}\;\frac{1}{\overline{r}_2}\;\ldots\;\frac{1}{\overline{r}_j}\;\ldots\;\frac{1}{\overline{r}_J}\right)$$

la matrice [1] diagonale di ordine $J\times J$ degli inversi delle masse dei J profili delle colonne. Quando il prodotto scalare è nullo i due vettori $\mathbf{r}_i$ e $\mathbf{r}_k$ risultano $\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}$-ortogonali in $\Re^J$. Così la distanza distribuzionale (4.8.1) tra due profili, espressa tramite il prodotto scalare, è la radice quadrata di

$$d_D^2(\mathbf{r}_i,\mathbf{r}_k)=(\mathbf{r}_i-\mathbf{r}_k)^T\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}(\mathbf{r}_i-\mathbf{r}_k)$$

e quindi la *lunghezza* di un profilo riga, ossia la distanza del profilo, inteso come punto di $\Re^J$, dall'origine della base di riferimento, è la radice quadrata di

$$d_D^2(\mathbf{r}_i,\mathbf{0}_{\mathrm{J}})=\mathbf{r}_i^T\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,\mathbf{r}_i=\sum_{j=1}^{J}\frac{1}{\overline{r}_j}\,r_{ij}^2.$$

Perciò un profilo riga ha lunghezza $\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}$-unitaria quando

$$\mathbf{r}_i^T\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,\mathbf{r}_i=1. \tag{4.8.2}$$

---

[1] La matrice $\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}$ è detta *metrica* dello spazio $\Re^J$.

L'inerzia complessiva della nuvola degli I profili, riferita all'origine $\mathbf{0}_{\mathrm{J}}$ della base canonica, è ottenuta come somma dei prodotti delle masse dei profili per i quadrati delle loro distanze dall'origine

$$\begin{aligned} In_{\mathbf{0}} &= \sum_{i=1}^{I} \overline{c}_i \, d_D^2(\mathbf{r}_i, \mathbf{0}_{\mathrm{J}}) = \sum_{i=1}^{I} \overline{c}_i \sum_{j=1}^{J} \frac{r_{ij}^2}{\overline{r}_j} = \sum_{i=1}^{I} \overline{c}_i \, \mathbf{r}_i{}^T \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}} \, \mathbf{r}_i \\ &= tr\left[\mathbf{R}^T \mathbf{D}_{\overline{\mathbf{c}}} \mathbf{R} \, \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\right] = tr\left[\mathbf{R}^T \mathbf{C}\right] \end{aligned} \tag{4.8.3}$$

grazie alla relazione (3.2.6) tra le matrici dei profili. Per la proprietà della funzione traccia, $In_{\mathbf{0}} = tr\left[\mathbf{R}^T \mathbf{C}\right] = tr\left[\mathbf{C}\,\mathbf{R}^T\right]$, si vede che l'inerzia riferita all'origine dei profili delle righe in $\Re^J$ coincide con l'inerzia (3.2.3) riferita ad $\mathbf{0}_{\mathrm{I}}$ dei profili delle colonne in $\Re^I$. Analoga eguaglianza sussiste per le inerzie computate rispetto al baricentro. Infatti, l'inerzia complessiva riferita al baricentro $\overline{\mathbf{r}}$ vale

$$\begin{aligned} In_{\overline{\mathbf{r}}} &= \sum_{i=1}^{I} \overline{c}_i \, d_D^2(\mathbf{r}_i, \overline{\mathbf{r}}) = \sum_{i=1}^{I} \overline{c}_i \sum_{j=1}^{J} \frac{(r_{ij} - \overline{r}_j)^2}{\overline{r}_j} \\ &= \sum_{i=1}^{I} \overline{c}_i \, (\mathbf{r}_i - \overline{\mathbf{r}})^T \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}} \, (\mathbf{r}_i - \overline{\mathbf{r}}) \\ &= tr\left[(\mathbf{R} - \overline{\mathbf{R}}) \, \mathbf{D}_{\overline{\mathbf{c}}} \, (\mathbf{R} - \overline{\mathbf{R}})^T \, \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\right] \end{aligned} \tag{4.8.4}$$

ove $\overline{\mathbf{R}} = \mathbf{1}_{\mathrm{I}} \, \overline{\mathbf{r}}^T$ è la matrice con righe tutte eguali ad $\overline{\mathbf{r}}$. Ma per il teorema di Huygens

$$In_{\overline{\mathbf{r}}} = In_{\mathbf{0}} + 1.$$

Confrontando questa relazione con la (3.4.2) risulta che $In_{\overline{\mathbf{r}}} = In_{\overline{\mathbf{c}}}$, e quindi entrambe le nuvole, seppur contenute in spazi *diversi* hanno le stesse inerzie complessive, sia rispetto all'origine che al baricentro. In altri termini, nei rispettivi spazi, le due nuvole hanno la medesima dispersione geometrica, quando questa è valutata in termini di inerzia.

La ricerca degli assi fattoriali d'inerzia procede in modo analogo a quello seguito per i profili delle colonne, per cui, se $\mathbf{v}$ è un generico vettore di ordine J dello spazio $\Re^J$, con origine in $\mathbf{0}_{\mathrm{J}}$ e lunghezza $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$-unitaria, l'inerzia riferita all'origine delle proiezioni degli I profili su questo vettore, è il numero

$$In_{\mathbf{0}}(\mathbf{v}) = \mathbf{v}^T \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}} \, \mathbf{R}^T \mathbf{D}_{\overline{\mathbf{c}}} \mathbf{R} \, \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}} \, \mathbf{v} \tag{4.8.5}$$

$$= \mathbf{v}^T \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}} \, \mathbf{R}^T \mathbf{C} \, \mathbf{v} \tag{4.8.6}$$

che si può esplicitare in

$$In_{\mathbf{0}}(\mathbf{v}) = \sum_{j=1}^{J} \frac{1}{\overline{r}_j} \sum_{j'=1}^{J} v_j \, v_{j'} \sum_{i=1}^{I} r_{ij} \, c_{ij'}.$$

Tra tutti i possibili vettori $\mathbf{v}$, occorre trovare quello su cui l'inerzia $In_{\mathbf{0}}(\mathbf{v})$ delle proiezioni degli I profili $\mathbf{r}_i$ sia massima. Un procedimento analogo a quello seguito nella Sez. 3.9, basato sul metodo dei moltiplicatori di Lagrange, porta a esprimere la condizione per un estremo vincolato dell'inerzia su $\mathbf{v}$ mediante l'equazione agli autovalori

$$\mathbf{R}^T \mathbf{D}_{\overline{\mathbf{c}}} \mathbf{R} \, \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}} \mathbf{v} = \mu \, \mathbf{v} \tag{4.8.7}$$

ovvero

$$\mathbf{R}^T \mathbf{C} \, \mathbf{v} = \mu \, \mathbf{v} \tag{4.8.8}$$

per il legame (3.8.6) tra le matrici dei profili. Il moltiplicatore di Lagrange è ora indicato con μ. Confrontando l'equazione (4.8.8) con la corrispondente (3.9.3) si vede che la matrice da diagonalizzare è ora $\mathbf{R}^T\mathbf{C}$ di ordine $J \times J$ nell'analisi dei profili riga, mentre era $\mathbf{C}\,\mathbf{R}^T$ di ordine $I \times I$ in quella dei profili colonna. Le due matrici sono quadrate, ma di diverso ordine, e per quanto mostrato nell'APP. B, sono entrambe $\mathbf{D}$-simmetriche. In particolare $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$-simmetrica la prima e $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-simmetrica la seconda.

Il procedimento di calcolo di autovalori ed autovettori si sviluppa in modo perfettamente analogo a quello mostrato nelle Sez. 9.3 e seguenti, per cui si vede che l'equazione (4.8.8) ammette I autovalori reali e non negativi. Anche ora l'autovalore più grande, detto banale, vale $\mu_0 = 1$ e il suo corrispondente autovettore è $\mathbf{v}_0 = \overline{\mathbf{r}}$ che individua il baricentro della nuvola nel riferimento della base canonica. Il profilo $\overline{\mathbf{r}}$ risulta $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$-ortogonale al simplesso degli I profili, nel senso che su $\mathbf{v}_0 = \overline{\mathbf{r}}$ tutti i gli I profili si proiettano in un unico punto: il baricentro. Anche in questa analisi l'autovalore e l'autovettore banali sono lasciati da parte, riducendo così a $I-1$ il numero di autovalori e di autovettori che si ottengono dall'equazione (4.8.8).

Procedendo come nella Sez. 3.14, si dimostra facilmente che se invece l'analisi viene fatta ricercando quel vettore $\mathbf{v}^\star$ con origine nel *baricentro* e di lunghezza $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$-unitaria, sul quale l'inerzia delle proiezioni $In_{\overline{\mathbf{r}}}(\mathbf{v}^\star)$ sia massima, si trova che, tralasciando la soluzione banale, gli $I-1$ autovalori hanno i *medesimi* autovalori dell'analisi precedente e gli autovettori le *stesse* J componenti degli autovettori dell'analisi precedente, con l'unica differenza di avere origine nel baricentro $\overline{\mathbf{r}}$. Questi autovettori costituiscono una nuova

base $\mathbf{D}_{\overline{\mathbf{r}}}^{-1}$ -*ortonormale* per lo spazio $\Re^J$ ed individuano $J-1$ assi fattoriali d'inerzia dell'iperpiano che contiene la nuvola degli I profili. Gli autovalori esprimono l'inerzia delle proiezioni dei profili su questi assi. L'inerzia complessiva rispetto al baricentro è quindi la somma dei $J-1$ autovalori, e, per quanto visto poco sopra, tra le inerzie [1] nei due spazi sussiste l'eguaglianza

$$In_{\overline{\mathbf{r}}} = \sum_{a=1}^{J-1} \mu_a = tr\,[\mathbf{R}^T\mathbf{C}] - 1 = tr\,[\mathbf{C}\,\mathbf{R}^T] - 1 = \sum_{a=1}^{I-1} \lambda_a = In_{\overline{\mathbf{c}}}. \qquad (4.8.9)$$

Questa relazione tra inerzie complessive non è l'unica che sussiste tra i due spazi. Questo non deve sorprendere perché i profili delle righe e delle colonne sono stati ottenuti con procedimenti perfettamente simmetrici partendo dalla stessa matrice di contingenza. Verrà mostrato ora che:

1 - le due matrici $\mathbf{R}^T\mathbf{C}$ e $\mathbf{C}\,\mathbf{R}^T$ hanno in comune gli stessi autovalori positivi;
2 - escludendo l'autovalore banale, gli autovalori comuni positivi sono in numero di $A = min\,(I-1, J-1)$, mentre sono nulli i restanti $I-1-A$ quando $I > J$ o $J-1-A$ quando $J > I$;
3 - per ciascun autovalore λ_a, tra i corrispondenti autovettori $\mathbf{v}_a^\star$ in $\Re^J$ e $\mathbf{u}_a^\star$ in $\Re^I$, sussiste un'importante relazione.

Il lettore non interessato alle dimostrazioni può passare direttamente al calcolo dei fattori nell'ultima parte di questa Sezione. I risultati, comunque, sono riassunti nella TAV. 28.

Per dimostrare il primo punto conviene esplicitare meglio il legame tra le due equazioni agli autovalori, esprimendole entrambe in funzione di una matrice comune. Ricordando che tra le matrici $\mathbf{R}$ e $\mathbf{C}$ dei profili sussiste la relazione (3.2.6), si può porre

$$\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{R} = \mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}} = \mathbf{X} \qquad (4.8.10)$$

per cui $\mathbf{X}$ è la matrice di ordine $I \times J$ delle frequenze relative $\mathbf{X} = 1/n_{++}\,\mathbf{N}$.

Nello spazio $\Re^I$, autovalori ed autovettori su ogni asse fattoriale $a = 1, 2, \ldots, I-1$ soddisfano la condizione (3.14.7)

$$\mathbf{C}\,\mathbf{R}^T\mathbf{u}_a^\star = \lambda_a\,\mathbf{u}_a^\star$$

---

[1] Le inerzie $In_{\overline{\mathbf{r}}}$ e $In_{\overline{\mathbf{c}}}$ vengono indicate talvolta col termine comune di *traccia*.

che, espressa in funzione di $\mathbf{X}$ grazie alla (4.8.10) o alla (3.2.5), diventa

$$\mathbf{X}\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,\mathbf{X}^T\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\,\mathbf{u}_a^\star = \lambda_a\,\mathbf{u}_a^\star. \tag{4.8.11}$$

Nello spazio $\Re^J$, autovalori ed autovettori soddisfano su ogni asse $a = 1, 2, \ldots, J-1$ la condizione

$$\mathbf{R}^T\mathbf{C}\,\mathbf{v}_a^\star = \mu_a\,\mathbf{v}_a^\star$$

che rappresentata in funzione di $\mathbf{X}$, diventa

$$\mathbf{X}^T\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\,\mathbf{X}\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,\mathbf{v}_a^\star = \mu_a\,\mathbf{v}_a^\star. \tag{4.8.12}$$

Occorre ora mostrare in primo luogo che se nell'equazione (4.8.11) per i profili colonna l'autovettore $\mathbf{u}_a^\star$, associato ad un autovalore λ_a *positivo*, non è nullo, tale è anche l'autovettore dello stesso rango $\mathbf{v}_a^\star$ della (4.8.12) per i profili riga. Infatti, se *cost* indica una costante reale positiva e si pone

$$\mathbf{v}_a^\star = cost\,\mathbf{X}^T\,\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\,\mathbf{u}_a^\star \tag{4.8.13}$$

l'equazione agli autovalori (4.8.11) per i profili colonna diventa

$$\frac{1}{cost}\,\mathbf{X}\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,\mathbf{v}_a^\star = \lambda_a\,\mathbf{u}_a^\star \neq \mathbf{0}_{\mathrm{I}} \tag{4.8.14}$$

per cui, essendo $\lambda_a > 0$, quando $\mathbf{u}_a^\star \neq \mathbf{0}_{\mathrm{I}}$ deve essere necessariamente anche $\mathbf{v}_a^\star \neq \mathbf{0}_{\mathrm{J}}$.

Ciò posto, il primo membro dell'equazione (4.8.12) per i profili riga, grazie alla relazione (4.8.14) appena dimostrata, e quindi grazie alla relazione (4.8.13), diventa

$$\mathbf{X}^T\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\,(\mathbf{X}\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\mathbf{v}_a^\star) = \lambda_a\,cost\,\mathbf{X}^T\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\,\mathbf{u}_a^\star = \lambda_a\,\mathbf{v}_a^\star.$$

Avendo appena visto che i due autovettori non sono nulli, il confronto tra quest'ultima espressione e l'equazione (4.8.12) porta a concludere che $\mu_a = \lambda_a > 0$ e che quindi le due matrici $\mathbf{C}\,\mathbf{R}^T$ e $\mathbf{R}^T\mathbf{C}$ hanno in comune gli autovalori positivi che, come mostrato nella Sez. B1 dell'Appendice B, sono quelli della matrice simmetrica che da esse è sempre possibile costruire, grazie al fatto che sono entrambe $\mathbf{D}$-simmetriche.

Per dimostrare il secondo punto si può costruire la matrice di ordine $J \times I$

$$\mathbf{Y} = \mathbf{D}_{\mathbf{r}}^{-\frac{1}{2}}\mathbf{X}^T\mathbf{D}_{\mathbf{c}}^{-\frac{1}{2}}$$

per cui, tenendo conto della posizione (4.8.10) si ha

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{D}_{\mathbf{c}}^{-\frac{1}{2}}\mathbf{X}\,\mathbf{D}_{\mathbf{r}}^{-\frac{1}{2}}\mathbf{D}_{\mathbf{r}}^{-\frac{1}{2}}\mathbf{X}^T\mathbf{D}_{\mathbf{c}}^{-\frac{1}{2}} = \mathbf{D}_{\mathbf{c}}^{-\frac{1}{2}}\mathbf{X}\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\mathbf{X}^T\mathbf{D}_{\mathbf{c}}^{-\frac{1}{2}} = \mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\mathbf{C}\,\mathbf{R}^T\mathbf{D}_{\mathbf{c}}^{-\frac{1}{2}}$$

e

$$\mathbf{Y}\,\mathbf{Y}^T = \mathbf{D}_{\mathbf{r}}^{-\frac{1}{2}}\mathbf{X}^T\mathbf{D}_{\mathbf{c}}^{-\frac{1}{2}}\mathbf{D}_{\mathbf{c}}^{-\frac{1}{2}}\mathbf{X}\,\mathbf{D}_{\mathbf{r}}^{-\frac{1}{2}} = \mathbf{D}_{\mathbf{r}}^{-\frac{1}{2}}\mathbf{X}^T\mathbf{D}_{\mathbf{c}}\mathbf{X}\,\mathbf{D}_{\mathbf{r}}^{-\frac{1}{2}} = \mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\mathbf{R}^T\mathbf{C}\,\mathbf{D}_{\mathbf{r}}^{-\frac{1}{2}}.$$

Avendo le due matrici simmetriche $\mathbf{Y}^T\mathbf{Y}$, di ordine $J \times J$, e $\mathbf{Y}\mathbf{Y}^T$ di ordine $I \times I$, lo stesso rango ed essendo diagonali le matrici $\mathbf{D}_{\mathbf{r}}$ e $\mathbf{D}_{\mathbf{c}}$ delle masse *positive* dei profili, anche le matrici quadrate $\mathbf{C}\,\mathbf{R}^T$ e $\mathbf{R}^T\mathbf{C}$ devono avere lo stesso rango e quindi lo stesso numero di autovalori positivi comuni. Se, come capita sempre nei casi reali, gli autovalori positivi sono tutti distinti, il loro numero risulta quindi

$$A = min\,(I-1, J-1) \qquad (4.8.14')$$

ove il -1 tiene conto dell'eliminazione in entrambe le analisi dell'autovalore banale. Perciò A è il numero di fattori utili per la rappresentazione grafica della configurazione dei profili e coincide con la dimensione dei sottospazi che contengono *effettivamente* i profili delle due nuvole, sia in $\Re^J$ che in $\Re^I$.

Gli autovalori *non comuni* alle due matrici devono essere tutti nulli perché nella Sez. 4.8 si è visto che l'inerzia complessiva, ossia la somma di tutti gli autovalori, è la medesima nei due spazi. Perciò, a seconda dell'ordine $I \times J$ della matrice di contingenza, ossia del fatto che la matrice abbia più colonne che righe o viceversa, si ha che

se $I < J$ è $A = I - 1$ $\lambda_1 = \mu_1 \ldots \lambda_{I-1} = \mu_{I-1}$ e $\mu_I = \ldots = \mu_J = 0$
se $I > J$ è $A = J - 1$ $\lambda_1 = \mu_1 \ldots \lambda_{J-1} = \mu_{J-1}$ e $\lambda_J = \ldots = \lambda_I = 0$

Gli autovettori invece non sono mai nulli, anche quando corrispondono ad un autovalore nullo, perché devono risultare linearmente indipendenti, come visto nella Sez. 2.4, in quanto costituiscono una base ortonormale.

Da un punto di vista geometrico questo significa che *entrambe* le nuvole di punti sono comunque sempre contenute in un sottospazio A-dimensionale, un simplesso ad A vertici, sia nello spazio $\Re^J$ che nello spazio $\Re^I$.

Così per la matrice d'esempio di TAV. 14 di ordine 3×8, si è visto nelle Sez. precedenti che l'analisi dei profili colonna in $\Re^3$ ha fornito $I - 1 = 3 - 1 = 2$ autovalori: $\lambda_1 = 0.005$ e $\lambda_2 = 0.002$. L'analisi dei profili delle righe in $\Re^8$ fornisce invece $J - 1 = 8 - 1 = 7$ autovalori, ma soltanto i *primi* $A = min\,(3-1, 8-1) = 2$ di questi sono positivi ed eguali a quelli ottenuti nell'altra analisi, mentre i rimanenti 5 sono tutti nulli, per cui le rispettive matrici risultano essere

$$\mathbf{D}_\lambda = \begin{pmatrix} 0.005 & 0 \\ 0 & 0.002 \end{pmatrix} \qquad \mathbf{D}_\mu = \begin{pmatrix} \begin{pmatrix} 0.005 & 0 \\ 0 & 0.002 \end{pmatrix} & \mathbf{0}_{2\times 5} \\ \mathbf{0}_{5\times 2} & \mathbf{0}_{5\times 5} \end{pmatrix}$$

dove $\mathbf{0}_{2\times 5}$, $\mathbf{0}_{5\times 2}$ e $\mathbf{0}_{5\times 5}$ sono matrici di diverso ordine costituite tutte da zeri. Perciò anche la nuvola dei 3 profili delle righe giace in realtà su un *piano* dello spazio 8-dimensionale.

Per dimostrare il terzo punto, basta imporre la condizione che l'autovettore $\mathbf{v}_a^\star$ sia di lunghezza $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$-unitaria. Perciò, tenendo conto della posizione (4.8.13) e successivamente dell'equazione (4.8.11), si ottiene

$$\mathbf{v}_a^{\star T}\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\mathbf{v}_a^\star = cost^2\mathbf{u}_a^{\star T}\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{X}\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\mathbf{X}^T\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\mathbf{u}_a^\star = cost^2\mathbf{u}_a^{\star T}\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\lambda_a\mathbf{u}_a^\star = cost^2\,\lambda_a$$

e per la condizione (4.8.4), deve essere $1 = cost^2\,\lambda_a$ e quindi $cost = 1/\sqrt{\lambda_a}$. Di conseguenza, per la (4.8.13) e la (4.8.10), il legame tra autovettori del medesimo rango nei due spazi risulta

$$\mathbf{v}_a^\star = \frac{1}{\sqrt{\lambda_a}}\,\mathbf{X}^T\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_a^\star = \frac{1}{\sqrt{\lambda_a}}\,\mathbf{D}_{\overline{\mathbf{r}}}\mathbf{C}^T\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_a^\star. \tag{4.8.15}$$

Con ciò i tre punti sono dimostrati.

In $\Re^J$ gli autovettori $\mathbf{v}_a^\star$ corrispondenti agli autovalori non nulli, individuano A assi fattoriali, la cui direzione resta comunque indeterminata. I fattori dei profili delle righe si ottengono proiettando gli I profili $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$-ortogonalmente sugli assi fattoriali, per cui

$$f_{ia} = \mathbf{r}_i^T\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\mathbf{v}_a^\star \qquad \text{e} \qquad \mathbf{f}_a \stackrel{\text{def}}{=} \mathbf{R}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\mathbf{v}_a^\star \qquad\qquad (a = 1, 2, \ldots, A)$$

indicano rispettivamente l'ascissa del profilo $\mathbf{r}_i$ sull'asse a ed il vettore delle ascisse di tutti gli I profili sull'asse, che viene detto a^{mo} *fattore* dei profili delle righe. Il fattore $\mathbf{f}_0$ corrispondente alla soluzione banale $(\lambda_0, \mathbf{v}_0 = \overline{\mathbf{r}})$ vale $\mathbf{f}_0 = \mathbf{R}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\,\overline{\mathbf{r}} = \mathbf{R}\,\mathbf{1}_{\mathrm{J}} = \mathbf{1}_{\mathrm{I}}$. I fattori delle righe godono delle stesse proprietà di quelli delle colonne, illustrate nella Sez. 4.2. La media ponderata delle coordinate dei profili su un asse è nulla, la loro inerzia è λ_a ed i fattori risultano non correlati due a due

$$\overline{\mathbf{c}}^T\,\mathbf{f}_a = \sum_{i=1}^{I}\overline{c}_i\,f_{ia} = 0 \qquad\qquad (a = 1, \ldots, A)$$

$$\mathbf{f}_a^T\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{f}_a = \sum_{i=1}^{I}\overline{c}_i\,f_{ia}^2 = \lambda_a \qquad\qquad (a = 1, \ldots, A)$$

$$\mathbf{f}_a^T\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{f}_b = \sum_{i=1}^{I}\overline{c}_i\,f_{ia}\,f_{ib} = 0. \quad (a, b = 1, \ldots, A; b \neq a)$$

Qui $\overline{c}_i$ è la massa del profilo $\mathbf{r}_i$ e $\mathbf{D}_{\overline{\mathbf{c}}}$ la matrice diagonale di ordine $I \times I$ delle masse dei profili delle righe. Se poi con $\mathbf{F}$ si indica la matrice di ordine

$I \times A$ dei fattori delle righe

$$\mathbf{F} = \begin{array}{c} \begin{array}{cccccc} \mathbf{f}_1 & \mathbf{f}_2 & & \mathbf{f}_a & & \mathbf{f}_A \end{array} \\ \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1a} & \cdots & f_{1A} \\ f_{21} & f_{22} & \cdots & f_{2a} & \cdots & f_{2A} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ f_{i1} & f_{i2} & \cdots & f_{ia} & \cdots & f_{iA} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ f_{I1} & f_{I2} & \cdots & f_{Ia} & \cdots & f_{IA} \end{pmatrix} \end{array}$$

le due ultime proprietà dei fattori dei profili riga si riassumono in

$$\underset{A\times I}{\mathbf{F}^T}\ \underset{I\times I}{\mathbf{D}_{\overline{\mathbf{c}}}}\ \underset{I\times A}{\mathbf{F}}\ = \underset{A\times A}{\mathbf{D}_\lambda}.$$

I risultati delle analisi dei profili delle righe e delle colonne sono messi a confronto nella TAV. 29.

Anche per i profili delle righe vengono definiti i fattori *standard*, normalizzati ad avere inerzia unitaria invece che λ_a su ogni asse $a = 1, \ldots, A$,

$$\hat{\mathbf{f}}_a \overset{\text{def}}{=} \frac{1}{\sqrt{\lambda_a}}\,\mathbf{f}_a \qquad \text{e quindi} \qquad \hat{\mathbf{F}} = \mathbf{F}\,\mathbf{D}_\lambda^{-\frac{1}{2}}. \tag{4.8.16}$$

I fattori standard hanno le stesse proprietà dei fattori principali: media ponderata nulla e matrice di varianza e covarianza unitaria

$$\overline{\mathbf{c}}^T\,\hat{\mathbf{f}}_a = \sum_{j=i}^{I} \overline{c}_i\,\hat{f}_{ia} = 0 \qquad\qquad (a = 1, \ldots, A)$$

$$\underset{A\times I}{\hat{\mathbf{F}}^T}\ \underset{I\times I}{\mathbf{D}_{\overline{\mathbf{c}}}}\ \underset{I\times A}{\hat{\mathbf{F}}}\ = \underset{A\times A}{\mathbf{I}}.$$

Nel caso della matrice d'esempio di TAV. 14 di ordine 3×8, i due fattori sono

$$\mathbf{f}_1 = \mathbf{R}\mathbf{D}_{\overline{\mathbf{r}}}^{-1}\mathbf{v}_1^\star = \begin{pmatrix} -0.063 \\ 0.078 \\ 0.074 \end{pmatrix} \quad \text{e} \quad \mathbf{f}_2 = \mathbf{R}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-1}\mathbf{v}_2^\star = \begin{pmatrix} -0.001 \\ -0.066 \\ 0.075 \end{pmatrix}$$

per cui le matrici dei fattori principali e standard risultano

$$\mathbf{F} = \begin{pmatrix} -0.063 & -0.001 \\ 0.078 & -0.066 \\ 0.074 & 0.075 \end{pmatrix} \qquad \hat{\mathbf{F}} = \mathbf{F}\,\mathbf{D}_\lambda^{-\frac{1}{2}} = \begin{pmatrix} -0.907 & -0.016 \\ 1.128 & -1.403 \\ 1.074 & 1.576 \end{pmatrix}.$$

Il primo fattore standard $\hat{\mathbf{f}}_1$ risulta amplificato di 14.40 volte rispetto a $\mathbf{f}_1$, il secondo, $\hat{\mathbf{f}}_2$, di 24.45 rispetto a $\mathbf{f}_2$. Invece, se si utilizzassero i valori arrotondati a tre cifre decimali riportati nella Sez.3.10,, si otterrebbe $1/\sqrt{\lambda_1} = \sqrt{0.005} = 14.14$ e $1/\sqrt{\lambda_2} = \sqrt{0.002} = 22.34$.

Analogamente a quanto fatto per i profili delle colonne, anche per i profili delle righe vengono introdotti degli indicatori numerici, strumenti indispensabili per una corretta interpretazione dei risultati dell'analisi. Così, il *contributo relativo* del profilo $\mathbf{r}_i$ all'inerzia sull'asse di rango a è definito come

$$CTR_a(\mathbf{r}_i) \stackrel{\text{def}}{=} \frac{\overline{c}_i \, f_{ia}^2}{\lambda_a}$$

mentre la *qualità della rappresentazione* sull'asse della distanza del profilo $\mathbf{r}_i$ dal baricentro $\overline{\mathbf{r}}$, è definita come

$$COS_a^2(\mathbf{r}_i) \stackrel{\text{def}}{=} \frac{f_{ia}^2}{\sum_{a=1}^{A} f_{ia}^2}.$$

Analogamente, considerando le proiezione di un profilo in un sottospazio A^*-dimensionale, si definiscono la qualità della rappresentazione della distanza tra proiezione e baricentro

$$QLT_{A^*}(\mathbf{r}_i) \stackrel{\text{def}}{=} \sum_{a=1}^{A^*} COS_a^2(\mathbf{r}_i) = \frac{\sum_{a=1}^{A^*} f_{ia}^2}{\sum_{a=1}^{A} f_{ia}^2}$$

e la quota d'inerzia del profilo nel sottospazio

$$INR_{A^*}(\mathbf{r}_i) \stackrel{\text{def}}{=} \overline{c}_i \, \frac{\sum_{a=1}^{A^*} f_{ia}^2}{In_{\overline{\mathbf{r}}}}.$$

Per significato e proprietà di questi indicatori, il lettore può far riferimento alle Sez. 4.6, 4.7 e 4.8.

4.9 - Relazioni di transizione

Nella Sez. precedente è stata stabilita l'importante relazione (4.8.15) che permette di ricavare gli autovettori $\mathbf{v}_a^\star$ di $\Re^J$ dagli autovettori $\mathbf{u}_a^\star$ del medesimo rango in $\Re^I$. Esiste anche la corrispondenza inversa. Infatti, tenendo conto delle relazioni (4.8.10) tra le matrici dei profili, e (4.8.14) e (4.8.15) tra autovettori, si ricavano le seguenti corrispondenze tra autovettori del *medesimo rango* a nei due spazi, valide per $a = 1, 2, \ldots, A = min\,(I -$

$1, J-1$)

$$\begin{aligned} \mathbf{u}_a^\star &= \frac{1}{\sqrt{\lambda_a}}\,\mathbf{C}\,\mathbf{v}_a^\star \qquad & u_{ia}^\star &= \frac{1}{\sqrt{\lambda_a}}\sum_{j=1}^{J} c_{ij}\,v_{ja}^\star \\ \mathbf{v}_a^\star &= \frac{1}{\sqrt{\lambda_a}}\,\mathbf{R}^T\mathbf{u}_a^\star \qquad & v_{ja}^\star &= \frac{1}{\sqrt{\lambda_a}}\sum_{i=1}^{I} r_{ij}\,u_{ia}^\star. \end{aligned} \tag{4.9.1}$$

Il rango è stabilito dal comune valore λ_a dell'inerzia delle proiezioni dei profili sui due autovettori. È bene rammentare che gli autovettori $\mathbf{u}_a^\star$ sono vettori di ordine I, di lunghezza $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitaria ed hanno origine nel baricentro $\overline{\mathbf{c}}$ della nuvola dei J profili colonna in $\Re^I$, e di questo spazio costituiscono una base $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortonormale, mentre gli autovettori $\mathbf{v}_a^\star$ sono vettori di ordine J, hanno lunghezza $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$-unitaria e origine nel baricentro $\overline{\mathbf{r}}$ della nuvola degli I profili riga in $\Re^J$, di cui costituiscono una base $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$-ortonormale. I due spazi $\Re^I$ ed $\Re^J$ sono dunque *diversi*, hanno differente dimensionalità e differenti unità di misura sugli assi, ma in ciascuno di essi la nuvola dei profili resta comunque confinata in un loro sottospazio A-dimensionale.

Le relazioni (4.9.1) tra autovettori del medesimo rango nei due spazi sono estremamente utili dal punto di vista computazionale, perché rendono superflua una delle due analisi. Delle due matrici quadrate $\mathbf{C}\,\mathbf{R}^T$ e $\mathbf{R}^T\mathbf{C}$ basta diagonalizzare soltanto quella di dimensionalità inferiore, ossia la prima se $I \leq J$, come nel caso della matrice d'esempio 3×8 di TAV. 14, altrimenti la seconda, e ricavare poi gli autovettori dell'altra dalle (4.9.1). L'orientamento degli assi $\mathbf{u}_a^\star$ e $\mathbf{v}_a^\star$, e quindi i segni delle loro componenti, concordano automaticamente. Se gli autovettori fossero ottenuti da analisi separate, in conseguenza dell'indeterminatezza del segno degli autovettori, potrebbe capitare che i segni non concordino. L'impiego delle relazioni di transizione ha quindi l'ulteriore vantaggio di evitare errori nell'orientamento degli assi perché il loro accordo diventa implicito.

Addirittura, è possibile ricavare direttamente i *fattori*, perché tenendo conto della (3.2.5) e della (3.2.6) si ottengono queste importanti relazioni tra fattori ed assi fattoriali, valide sempre per ogni asse $a = 1, 2, \ldots, A$

$$\begin{aligned} \mathbf{f}_a &= \sqrt{\lambda_a}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_a^\star \qquad & f_{ia} &= \left(\sqrt{\lambda_a}/\overline{c}_i\right)\,u_{ia}^\star \\ \mathbf{g}_a &= \sqrt{\lambda_a}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\,\mathbf{v}_a^\star \qquad & g_{ja} &= \left(\sqrt{\lambda_a}/\overline{r}_j\right)\,v_{ja}^\star \end{aligned} \tag{4.9.2}$$

per cui autovettori e fattori risultano essere vettori allineati nel medesimo spazio: $\Re^J$ per i profili delle righe e $\Re^I$ per i profili delle colonne.

Premoltiplicando la prima equazione per $\mathbf{C}^T\mathbf{D}_\mathbf{c}^{-\mathbf{1}}$ e tenendo conto della (3.2.5), la seconda per $\mathbf{R}\,\mathbf{D}_\mathbf{r}^{-\mathbf{1}}$ e tenendo conto della (3.2.4), si ottengono le importanti *relazioni di transizione* dai fattori dei profili delle righe a quelli delle colonne e viceversa, valide per ogni asse $a = 1, 2, \ldots, A$

$$\begin{aligned} \mathbf{f}_a &= \frac{1}{\sqrt{\lambda_a}}\,\mathbf{R}\,\mathbf{g}_a \qquad & f_{ia} &= \frac{1}{\sqrt{\lambda_a}}\sum_{j=1}^{J} r_{ij}\, g_{ja} \\ \mathbf{g}_a &= \frac{1}{\sqrt{\lambda_a}}\,\mathbf{C}^T\mathbf{f}_a \qquad & g_{ja} &= \frac{1}{\sqrt{\lambda_a}}\sum_{i=1}^{I} c_{ij}\, f_{ia}. \end{aligned} \tag{4.9.3}$$

A meno del coefficiente $1/\sqrt{\lambda_a}$, la coordinata f_{ia} del profilo riga $\mathbf{r}_i$ sull'asse a di $\Re^J$ è la media pesata (baricentro) delle coordinate g_{ja} dei J profili colonna $\mathbf{c}_j$ sul *corrispondente*, ossia dello stesso rango, asse a di $\Re^I$. Il peso attribuito ad ogni coordinata è la componente r_{ij} del profilo $\mathbf{r}_i$, che misura l'importanza della modalità j in questo profilo, come si è visto nella Sez. 1.5.

Analogamente, e sempre a meno del coefficiente, la coordinata di un profilo colonna è il 'baricentro' delle coordinate dei profili riga, i pesi sono le componenti del profilo colonna. Per questo motivo le relazioni di transizione sono anche dette *relazioni baricentriche* o relazioni di *rappresentazione baricentrica*, anche se a rigore sarebbe più corretto chiamarle '*quasi* baricentriche', a causa del coefficiente $1/\sqrt{\lambda_a}$. La rappresentazione dei profili tramite i fattori principali (4.9.3) porta alla costruzione di mappe cosiddette *simmetriche* che verranno discusse in dettaglio nella Sez. 4.11.

Le relazioni (4.9.3) possono esprimersi in funzione dei fattori *standard*, definiti nelle Sezioni 4.3 e 4.8, aventi inerzia unitaria su ogni asse,

$$\begin{aligned} \mathbf{f}_a &= \mathbf{R}\,\hat{\mathbf{g}}_a \qquad & f_{ia} &= \sum_{j=1}^{J} r_{ij}\,\hat{g}_{ja} \\ \mathbf{g}_a &= \mathbf{C}^T\hat{\mathbf{f}}_a \qquad & g_{ja} &= \sum_{i=1}^{I} c_{ij}\,\hat{f}_{ia}. \end{aligned} \tag{4.9.4}$$

La rappresentazione dei profili tramite i fattori principali per una nuvola e i fattori standard per quelli dell'altra, produce mappe *asimmetriche*. Le loro proprietà verranno discusse nella Sez. 4.13. Si vede comunque fin da ora che la coordinata di un profilo su un asse è la media delle coordinate dei profili dell'altra nuvola, ponderate con le componenti dello stesso profilo, per cui, ad

esempio un profilo riga tende a disporsi più vicino al profilo colonna j per il quale la componente r_{ij} è più elevata.

Le relazioni di transizione consentono di dimostrare, e questo verrà fatto tra poco in questa Sezione, che un autovalore non può mai risultare maggiore di 1. Di conseguenza il coefficiente moltiplicativo $1/\sqrt{\lambda_a}$ non è mai inferiore a 1. Si tratta quindi di un coefficiente di 'dilatazione' che agisce diversamente sui diversi assi, allontanando maggiormente le proiezioni sugli assi di rango elevato. Nella Sez. precedente si è calcolato che nel caso della matrice d'esempio di ordine 3×8 di TAV. 14, questo coefficiente vale 14.40 per le coordinate sul primo asse e quasi il doppio, 24.45, per quelle sul secondo.

Senza questo coefficiente i profili delle righe sarebbero il 'baricentro' di quelli delle colonne e viceversa. È chiaro che una rappresentazione doppiamente baricentrica non è possibile e quindi è necessario un coefficiente di dilatazione superiore ad 1 e diverso per ciascun asse fattoriale della mappa.

Le relazioni di transizione hanno un ruolo fondamentale nell'Analisi delle Corrispondenze perché rivelano la perfetta simmetria che sussiste tra i fattori dei profili delle righe e delle colonne, e giustificano la rappresentazione *simultanea* dei profili delle due nuvole, fatto questo che rende *unico* questo metodo tra gli altri metodi fattoriali ove non sussiste alcuna simmetria di ruoli tra fattori delle righe e delle colonne e ciò comporta un tipo di interpretazione *diversa* per gli elementi dell'uno e dell'altro insieme. Il lettore è invitato a meditare sulle relazioni di transizione perché senza una loro piena comprensione non sarà mai in grado di interpretare correttamente i risultati di un'Analisi delle Corrispondenze.

Le relazioni di transizione possono esprimersi anche in una forma che ripropone i confronti tra profili e profilo medio fatte nel primo Capitolo. Nelle Sez. 4.2 e 4.8 si è visto che i fattori hanno media ponderata nulla

$$\sum_{j=1}^{J} \overline{r}_j \, g_{ja} = 0 \qquad \text{e} \qquad \sum_{i=1}^{I} \overline{c}_i \, f_{ia} = 0$$

per cui è possibile sottrarre queste quantità nulle dalle relazioni di transizione (4.9.3), ottenendo

$$f_{ia} = \frac{1}{\sqrt{\lambda_a}} \sum_{j=1}^{J} (r_{ij} - \overline{r}_j) \, g_{ja}$$

$$g_{ja} = \frac{1}{\sqrt{\lambda_a}} \sum_{i=1}^{I} (c_{ij} - \overline{c}_i) \, f_{ia}.$$

Si vede quindi che su un asse fattoriale la proiezione di un profilo riga (o colonna) si situa dalla parte delle proiezioni di quei profili colonna (o riga) che maggiormente si discostano dalla loro media. Queste considerazioni, che sono alla base di una corretta interpretazione di una mappa fattoriale, troveranno pratica applicazione nella Sez. 4.11.

L'espressione matriciale delle relazioni di transizione tra autovettori e tra fattori dei due spazi è

$$\underset{I\times A}{\mathbf{U}^\star} = \underset{I\times J}{\mathbf{C}}\ \underset{J\times A}{\mathbf{V}^\star}\ \underset{A\times A}{\mathbf{D}_\lambda^{-\frac{1}{2}}} \qquad \underset{J\times A}{\mathbf{V}^\star} = \underset{J\times I}{\mathbf{R}^T}\ \underset{I\times A}{\mathbf{U}^\star}\ \underset{A\times A}{\mathbf{D}_\lambda^{-\frac{1}{2}}} \tag{4.9.5}$$

$$\underset{I\times A}{\mathbf{F}} = \underset{I\times J}{\mathbf{R}}\ \underset{J\times A}{\mathbf{G}}\ \underset{A\times A}{\mathbf{D}_\lambda^{-\frac{1}{2}}} = \underset{I\times J}{\mathbf{R}}\ \underset{J\times A}{\hat{\mathbf{G}}} \qquad \underset{J\times A}{\mathbf{G}} = \underset{J\times I}{\mathbf{C}^T}\ \underset{I\times A}{\mathbf{F}}\ \underset{A\times A}{\mathbf{D}_\lambda^{-\frac{1}{2}}} = \underset{J\times I}{\mathbf{C}^T}\ \underset{I\times A}{\hat{\mathbf{F}}}. \tag{4.9.6}$$

mentre le relazioni tra fattori ed autovettori del medesimo spazio sono

$$\underset{I\times A}{\mathbf{F}} = \underset{I\times I}{\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}}\ \underset{I\times A}{\mathbf{U}^\star}\ \underset{A\times A}{\mathbf{D}_\lambda^{\frac{1}{2}}} \qquad \underset{J\times A}{\mathbf{G}} = \underset{J\times J}{\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}}\ \underset{J\times A}{\mathbf{V}^\star}\ \underset{A\times A}{\mathbf{D}_\lambda^{\frac{1}{2}}}\ . \tag{4.9.7}$$

Le relazioni di transizione permettono di dimostrare che gli autovalori delle matrici $\mathbf{R}^T\mathbf{C}$ e $\mathbf{C}\,\mathbf{R}^T$ non possono essere maggiori di 1. Infatti la relazione di transizione (4.9.3) dai fattori delle righe a quelli delle colonne scritta nella forma

$$f_{ia}\,\sqrt{\lambda_a} = \sum_{j=1}^{J} r_{ij}\,g_{ja}$$

mostra che $f_{ia}\,\sqrt{\lambda_a}$ è la media ponderata delle coordinate g_{ja} con pesi r_{ij} e quindi è necessariamente superiore, od al più eguale, al più piccolo dei valori g_{ja} e, al medesimo tempo, inferiore o al massimo eguale al più grande dei valori g_{ja} nella sommatoria

$$\min_j\,[g_{ja}]\,\sum_{j=1}^{J} r_{ij} \le f_{ia}\,\sqrt{\lambda_a} \le \max_j\,[g_{ja}]\,\sum_{j=1}^{J} r_{ij}$$

ma, la somma delle J componenti del profilo $\mathbf{r}_i$ vale 1 e, di conseguenza, il più grande dei valori che può assumere $f_{ia}\,\sqrt{\lambda_a}$ sarà sempre inferiore o eguale al più grande dei valori g_{ja}

$$\max_i\,\left[f_{ia}\,\sqrt{\lambda_a}\right] \le \max_j\,[g_{ja}]. \tag{4.9.8}$$

D'altro canto, se si fosse partiti dall'altra formula di transizione dai fattori delle colonne a quelli delle righe e proceduto in modo analogo, si sarebbe

ottenuto

$$\max_j \left[g_{ja} \sqrt{\lambda_a} \right] \leq \max_i [f_{ia}].$$

Moltiplicando ambo i membri della (4.9.8) per $\sqrt{\lambda_a}$, che rappresentando un'inerzia non può essere negativa, e tenendo conto della diseguaglianza qui sopra, si può scrivere

$$\lambda_a \max_i [f_{ia}] \leq \max_j \left[g_{ja} \sqrt{\lambda_a} \right] \leq \max_i [f_{ia}]$$

che evidentemente può valere soltanto se $\lambda_a \leq 1$. Pertanto, escludendo l'autovalore banale, gli $A = min\,(I-1, J-1)$ autovalori che si ottengono diagonalizzando le matrici D-simmetriche $\mathbf{C}\,\mathbf{R}^T$ e $\mathbf{R}^T\mathbf{C}$ sono tutti compresi nell'intervallo

$$0 \leq \lambda_a \leq 1. \tag{4.9.9}$$

Da ultimo, le relazioni di transizione permettono anche di mostrare come gli stessi fattori $\mathbf{f}_a$ e $\mathbf{g}_a$ siano autovettori non banali delle matrici quadrate $\mathbf{R}\,\mathbf{C}^T$ e $\mathbf{C}^T\mathbf{R}$ rispettivamente. Gli autovalori di entrambe le matrici sono le inerzie λ_a. Infatti, sostituendo la seconda relazione delle (4.9.3) nella prima si ottiene per ogni asse $a = 1, 2, \ldots, A$,

$$\mathbf{f}_a = \frac{1}{\sqrt{\lambda_a}}\,\mathbf{R}\,\mathbf{g}_a = \frac{1}{\lambda_a}\,\mathbf{R}\,\mathbf{C}^T\mathbf{f}_a$$

da cui l'equazione agli autovalori

$$\mathbf{R}\,\mathbf{C}^T\mathbf{f}_a = \lambda_a\,\mathbf{f}_a \quad \text{col vincolo} \quad \mathbf{f}_a^T\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{f}_a = \lambda_a. \tag{4.9.10}$$

In modo analogo si ottiene l'altra equazione

$$\mathbf{C}^T\mathbf{R}\,\mathbf{g}_a = \lambda_a\,\mathbf{g}_a \quad \text{col vincolo} \quad \mathbf{g}_a^T\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{g}_a = \lambda_a. \tag{4.9.11}$$

Anche queste due equazioni non vanno impiegate separatamente per ottenere i fattori $\mathbf{f}_a$ e $\mathbf{g}_a$. Sarebbe un'inutile perdita di tempo che potrebbe provocare degli errori negli orientamenti dei corrispondenti autovettori a causa dell'indeterminatezza del loro segno, come si è visto nella Sez. 3.13.

4.10 - Formule di ricostruzione

Caratteristica importante dell'Analisi delle Corrispondenze è la possibilità di ricostruire la matrice di contingenza $\mathbf{N}$ con un grado di approssimazione che dipende dal numero di fattori impiegati. Infatti, la conoscenza degli A fattori $\mathbf{f}_a$ e $\mathbf{g}_a$, degli autovalori λ_a, dei vettori $\overline{\mathbf{r}}$ e $\overline{\mathbf{c}}$ delle masse

dei profili e della numerosità complessiva n_{++}, consente di ricostruire *esattamente* ogni elemento n_{ij} della matrice di contingenza $\mathbf{N}$, dalla quale tutte queste grandezze sono state ricavate

$$n_{ij} = \frac{n_{i+}\, n_{+j}}{n_{++}} \left(1 + \sum_{a=1}^{A} \frac{1}{\sqrt{\lambda}_a}\, f_{ia}\, g_{ja} \right). \qquad (4.10.1)$$

Il termine 1 è dovuto alla soluzione banale, come si vedrà tra poco. Portando al primo membro n_{i+} e n_{+j}, si ottengono le *formule di ricostruzione* dei profili delle righe e delle colonne

$$r_{ij} = \overline{r}_j \left(1 + \sum_{a=1}^{A} \frac{1}{\sqrt{\lambda}_a}\, f_{ia}\, g_{ja} \right) \qquad c_{ij} = \overline{c}_i \left(1 + \sum_{a=1}^{A} \frac{1}{\sqrt{\lambda}_a}\, f_{ia}\, g_{ja} \right).$$

Come queste formule vengano ricavate è mostrato alla fine di questa Sezione.

Abitualmente, la dispersione geometrica di un insieme di punti su un asse è misurata dalla deviazione standard, ossia dalla radice quadrata della loro varianza, o inerzia nel caso di profili. Quando si considerino le proiezioni di *tutti* i profili di una nuvola sull'asse di rango a, la deviazione standard è $\sqrt{\lambda_a}$ e, di conseguenza, nella (4.10.1), l'ordine di grandezza complessivo del termine di rango a, considerando tutti i punti delle due nuvole, risulta $(1/\sqrt{\lambda_a}) \times \sqrt{\lambda_a} \times \sqrt{\lambda_a} = \sqrt{\lambda_a}$. In effetti, nei termini della somma della formula di ricostruzione compare soltanto un profilo per ciascuna nuvola, e, se all'aumentare del rango è sicuro che gli autovalori decrescono e non superano 1, è anche però probabile che le due proiezioni f_{ia} e g_{ja} siano sempre meno disperse attorno ai loro baricentri. La concomitanza di questi due elementi permette di concludere che, sostanzialmente, nella formula di ricostruzione i termini della somma *decrescono* al crescere del rango.

La struttura cumulativa della formula di ricostruzione (4.10.1) permette di *approssimare* gli elementi della matrice di contingenza, limitando la somma ai termini di rango più basso, ossia ai primi, fornendo le approssimazioni

$$\begin{array}{lrcl}
\text{rango } a = 0: & n_{ij}^{(0)} & = & n_{i+}\, n_{+j}/n_{++}. \\
\text{rango } a = 1: & n_{ij}^{(1)} & = & (n_{i+}\, n_{+j}/n_{++}) + \\
 & & & (n_{i+}\, n_{+j}/n_{++}) \left(1/\sqrt{\lambda_1}\right) f_{i1}\, g_{j1}. \\
\text{rango } a = 2: & n_{ij}^{(2)} & = & (n_{i+}\, n_{+j}/n_{++}) + \\
 & & & (n_{i+}\, n_{+j}/n_{++}) \left(1/\sqrt{\lambda_1}\right) f_{i1}\, g_{j1} + \\
 & & & (n_{i+}\, n_{+j}/n_{++}) \left(1/\sqrt{\lambda_2}\right) f_{i2}\, g_{j2}. \\
\ldots\ldots & \ldots & = & \ldots\ldots\ldots
\end{array}$$

L'approssimazione di rango 0 genera una matrice che rispecchia una situazione di completa omogeneità, descritta nella Sez. 1.9, in cui la matrice è semplificata al massimo: tutte le righe e tutte le colonne hanno il medesimo profilo che coincide con quello medio. Geometricamente i punti delle due nuvole sono proiettati tutti in un unico punto: il loro baricentro, ove si concentra tutta la massa, che vale 1. Nel caso della matrice d'esempio di ordine 3×8 di TAV. 14, l'approssimazione di rango 0 del primo elemento, che indica in $n_{11} = 576$ le decine di migliaia di spettatori che nel 1991 hanno assistito a rappresentazioni di Prosa nel Nord, risulta

$$n_{11}^{(0)} = \frac{n_{1+}\, n_{+1}}{n_{++}} = \frac{1\,437 \times 1\,099}{2\,620} = 602.772$$

Nell'approssimazione di rango 1, i punti delle nuvole sono sostituiti dalle loro proiezioni sul primo asse fattoriale, ossia in un sottospazio monodimensionale, e nel caso dell'esempio, per quanto ottenuto nella Sez 4.4 e nella Sez. 4.8, risulta

$$n_{11}^{(1)} = 602.772 + 602.772 \times 14.459 \times (-0.063) \times 0.049 = 576.121.$$

Considerare poi l'approssimazione di rango 2 significa ridurre i punti delle nuvole alle loro proiezioni sul primo piano fattoriale, ossia in un sottospazio bidimensionale. Per la matrice d'esempio è

$$n_{11}^{(2)} = 576.121 + 602.772 \times 14.459 \times (-0.00077) \times 0.028 = 576.000 = n_{11}.$$

perché in questo caso il numero di fattori è $A = 2$ e quindi l'elemento viene ricostruito perfettamente.

Come si vede, l'aggiunta di termini rende gli elementi della matrice approssimata sempre più ricchi di dettagli finché, oltre un certo rango $A^* \leq A$ non si aggiungono che fluttuazioni statistiche. Perciò, arrestare la ricostruzione al rango A^* significa in pratica effettuare uno 'smoothing' multidimensionale sugli elementi della matrice di contingenza.

Nel primo Capitolo, nella Sez. 1.9, per quantificare l'associazione tra le modalità di una riga i e di una colonna j, e quindi del rapportarsi della componente di un profilo alla corrispondente componente del profilo medio, era stato introdotto lo scarto relativo dalla quota media

$$s_{ij} = \frac{n_{ij}}{\frac{n_{i+}\, n_{+j}}{n_{++}}} - 1 = \frac{r_{ij}}{\overline{r}_j} - 1 = \frac{c_{ij}}{\overline{c}_i} - 1.$$

Le formule di ricostruzione possono ora esprimersi in funzione di questo indice

$$s_{ij} = \sum_{a=1}^{A} \frac{1}{\sqrt{\lambda}_a} f_{ia}\, g_{ja} \tag{4.10.2}$$

che mostra come l'Analisi delle Corrispondenze permetta di graduare l'associazione tra modalità, scomponendo l'indice in una somma di elementi semplici. Semplici perché dipendono esclusivamente dalle proiezioni dei profili delle due modalità sull'asse di rango a. Se i fattori f_{ia} e g_{ja} hanno il medesimo segno, ossia le proiezioni sono entrambe dalla stessa parte rispetto all'origine dell'asse, il termine esprime un'*attrazione* tra le modalità i e j perché va ad incrementare l'indice rendendolo superiore ad 1. Al contrario, il termine è di repulsione se i due fattori sono di diverso segno e quindi le proiezioni sull'asse si trovano da parti opposte rispetto all'origine. Così le due modalità su alcuni assi si attraggono e su altri si respingono e questo tanto più intensamente quanto più elevati sono $|f_{ia}|$ e $|g_{ja}|$, ma il tutto in misura sempre più debole al crescere del rango dell'asse. Perciò l'Analisi delle Corrispondenze dà delle complesse associazioni tra modalità una immagine semplice ed ordinata che fa apparire primitivo ogni altro metodo basato su semplici confronti di profili, come quello illustrato nel primo Capitolo.

Nel caso dell'esempio la matrice degli scarti relativi risulta

$$\mathbf{S} = \begin{array}{c} \begin{array}{cccccccc} \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 & \mathbf{c}_4 & \mathbf{c}_5 & \mathbf{c}_6 & \mathbf{c}_7 & \mathbf{c}_8 \end{array} \\ \begin{pmatrix} \cdot & +0.1 & \cdot & +0.2 & \cdot & \cdot & -0.2 & +0.1 \\ \cdot & -0.1 & \cdot & -0.4 & +0.2 & \cdot & +0.5 & -0.3 \\ +0.1 & -0.2 & +0.1 & \cdot & -0.2 & -0.1 & -0.1 & \cdot \end{pmatrix} \end{array}$$

ove $\cdot$ indica un valore nullo o quasi. Non appaiono né forti attrazioni, né forti repulsioni, ma questo era da attendersi perché un basso valore dell'inerzia come $In_{\overline{\mathbf{c}}} = In_{\overline{\mathbf{r}}} = 0.007$ è indice di scarsa associazione tra i profili delle due nuvole.

Ecco infine come è ottenuta la formula di ricostruzione dell'intera matrice di contingenza $\mathbf{N}$. Supponendo che questa abbia più colonne che righe, come la matrice d'esempio di TAV. 14 che è di ordine 3×8, ne risulta $A = I - 1$ e l'equazione agli autovalori, come mostrato nella Sez. 5 dell'Appendice B, per la (B.5.2) e con le stesse notazioni, ma tenendo anche conto che $\overline{\mathbf{C}}\,\mathbf{R}^T = \mathbf{C}\,\overline{\mathbf{R}}^T$, si scrive

$$\mathbf{C}\,(\mathbf{R} - \overline{\mathbf{R}})^T = \mathbf{U}^\star\,\mathbf{D}_\lambda\,\mathbf{U}^{\star T}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$$

La formula di transizione (4.9.4) permette di esprimere gli autovettori $\mathbf{U}^\star$ in funzione dei corrispondenti $\mathbf{V}^\star$ nell'altro spazio, per cui

$$\mathbf{C}\,(\mathbf{R}-\overline{\mathbf{R}})^T = \mathbf{C}\,\mathbf{V}^\star\,\mathbf{D}_\lambda^{\frac{1}{2}}\,\mathbf{U}^{\star T}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$$

Poiché la matrice dei profili non è nulla, la sua inversa $\mathbf{C}^{-1}$ esiste e può premoltiplicare ambo i membri, fornendo poi, grazie alla (4.9.6),

$$\mathbf{R}^T-\overline{\mathbf{R}}^T = \mathbf{V}^\star\,\mathbf{D}_\lambda^{-\frac{1}{2}}\,\mathbf{U}^{\star T}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} = \mathbf{D}_{\overline{\mathbf{r}}}\left(\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\,\mathbf{V}^\star\,\mathbf{D}_\lambda^{-\frac{1}{2}}\right)\mathbf{D}_\lambda^{-\frac{1}{2}}\,\mathbf{D}_\lambda^{\frac{1}{2}}\,\mathbf{U}^{\star T}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}.$$

Tenendo ora conto della formula di transizione dei fattori delle colonne, si ha

$$\mathbf{R}^T-\overline{\mathbf{R}}^T = \mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{G}\,\mathbf{D}_\lambda^{-\frac{1}{2}}\,\mathbf{F}^T$$

che è la formula di ricostruzione della matrice delle righe in forma trasposta. Trasponendo l'espressione e premoltiplicando ambo i membri per $n_{++}\,\mathbf{D}_{\overline{\mathbf{c}}}$ e tenendo conto che $\mathbf{N}=n_{++}\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{R}$ e che $\overline{\mathbf{R}}=\mathbf{1}_{\mathrm{I}}\,\overline{\mathbf{r}}^T$ per cui

$$n_{++}\,\mathbf{D}_{\overline{\mathbf{c}}}\,\overline{\mathbf{R}} = n_{++}\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{1}_{\mathrm{I}}\,\overline{\mathbf{r}}^T = n_{++}\,\overline{\mathbf{c}}\,\overline{\mathbf{r}}^T$$

si ottiene finalmente l'espressione matriciale della formula di ricostruzione (4.10.1) della matrice di contingenza

$$\underset{I\times J}{\mathbf{N}} = n_{++}\;\underset{I\times 1}{\overline{\mathbf{c}}}\;\underset{1\times J}{\overline{\mathbf{r}}^T} + n_{++}\;\underset{I\times I}{\mathbf{D}_{\overline{\mathbf{c}}}}\;\underset{I\times A}{\mathbf{F}}\;\underset{A\times A}{\mathbf{D}_\lambda^{-1/2}}\;\underset{A\times J}{\mathbf{G}^T}\;\underset{J\times J}{\mathbf{D}_{\overline{\mathbf{r}}}} \tag{4.10.3}$$

mentre per le matrici dei profili le formule di ricostruzione sono

$$\mathbf{R} = \overline{\mathbf{R}} + \mathbf{F}\,\mathbf{D}_\lambda^{-\frac{1}{2}}\,\mathbf{G}^T\mathbf{D}_{\overline{\mathbf{r}}} \qquad \mathbf{C} = \overline{\mathbf{C}} + \mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{F}\,\mathbf{D}_\lambda^{-\frac{1}{2}}\,\mathbf{G}^T.$$

Limitando le matrici $\mathbf{F}$ e $\mathbf{G}$ alle prime $A^*\le A$ colonne e, conseguentemente la matrice $\mathbf{D}_\lambda$ ai primi A^* autovalori, si ottengono le approssimazioni di ordine A^* delle matrici $\mathbf{N}$, $\mathbf{R}$ e $\mathbf{C}$.

Se la matrice di contingenza avesse invece più righe che colonne, per cui $A=J-1$, l'equazione di partenza sarebbe la (B.5.4), che condurrebbe, ovviamente, al medesimo risultato. Le formule di ricostruzione possono ottenersi anche partendo dagli autovettori con origine in $\mathbf{0}_{\mathrm{I}}$ e $\mathbf{0}_{\mathrm{J}}$ invece che nei baricentri, purché si tenga conto della soluzione banale, priva d'interesse ai fini della rappresentazione grafica dei profili, ma che resta essenziale dal punto di vista matematico. Partendo dalla (B.5.1) o dalla (B.5.3) a seconda che la matrice di contingenza abbia più righe che colonne o viceversa, procedendo in modo analogo a quanto visto sopra, si arriva a

$$\mathbf{N} = n_{++}\;\mathbf{D}_{\overline{\mathbf{c}}}\;[\mathbf{f}_0\;\mathbf{F}]\begin{bmatrix}\lambda_0 & \mathbf{0}_{\mathrm{A}}^T\\ \mathbf{0}_{\mathrm{A}} & \mathbf{D}_\lambda^{-1/2}\end{bmatrix}[\mathbf{g}_0\;\mathbf{G}]^T\;\mathbf{D}_{\overline{\mathbf{r}}}$$

In questa espressione i fattori corrispondenti all'autovettore banale $\lambda_0 = 1$ valgono $\mathbf{f}_0 = \mathbf{R}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\,\overline{\mathbf{r}} = \mathbf{1}_{\mathrm{I}}$ e $\mathbf{g}_0 = \mathbf{C}^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\overline{\mathbf{c}} = \mathbf{1}_{\mathrm{J}}$, per cui

$$\mathbf{N} = n_{++}\;\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{1}_{\mathrm{I}}\;(\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{1}_{\mathrm{J}})^T + n_{++}\;\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{F}\,\mathbf{D}_{\lambda}^{-\frac{1}{2}}\mathbf{G}^T\,\mathbf{D}_{\overline{\mathbf{r}}}$$

da cui la (4.10.3). Ecco chiarita l'origine del termine 1 nelle formule di ricostruzione degli elementi (4.10.1): è dovuto a $(1/\sqrt{\lambda_0})\,f_{i0}\,g_{j0} = 1\times1\times1 = 1$. Questo secondo modo di ottenere le formule di ricostruzione chiarisce l'origine matematica del termine omogeneo.

4.11 - Interpretazione dei risultati

Una buona conoscenza dell'Analisi delle Corrispondenze e la sua corretta applicazione alla matrice da analizzare, sono le necessarie premesse per ottenere risultati corretti. La matrice merita sempre qualche attenzione perché può presentare delle insidie, per cui prima di procedere all'analisi, è sempre buona norma:

- identificare il tipo di matrice,
- verificare l'assenza di elementi mancanti,
- accertarsi che i suoi elementi nulli non siano zeri strutturali.

L'identificazione del *tipo* è essenziale perché alcuni tipi di matrice debbono essere pretrattati prima di essere sottoposti ad analisi. A partire dal Capitolo 8 verranno passati in rassegna i principali tipi di matrice e per ciascuno di essi verrà indicato l'eventuale pretrattamento e gli adattamenti che, in tali casi, l'interpretazione dei risultati richiede. In particolare, nel Capitolo 8 verrà mostrato come le formule di ricostruzione, ricavate nella precedente Sez. 4.10, permettano di ricostruire eventuali *elementi mancanti*, ossia elementi della matrice che per vari motivi non hanno un valore assegnato, e come tener conto degli *zeri strutturali*, ossia elementi $n_{ij} = 0$ che indicano l'*impossibilità logica* di un legame tra le modalità i e j della matrice.

Fatto questo, analizzare i risultati dell'Analisi delle Corrispondenze significa nell'ordine:

1 - commentare i due profili marginali che rappresentano l'origine degli assi fattoriali,

2 - esaminare la ripartizione dell'inerzia rispetto al baricentro sugli assi fattoriali, per farsi un'idea di quanti e quali assi siano potenzialmente caratterizzabili,

3 - caratterizzare gli assi, attribuendo loro un significato descrittivo in base alle prossimità e alle contrapposizioni che le proiezioni dei profili

hanno su di essi, ossia in base ai fattori. In questo senso si parla indifferentemente di caratterizzazione di un asse o di un fattore,

4 - interpretare le associazioni tra proiezioni su mappe bidimensionali e, possibilmente,

5 - in rappresentazioni tridimensionali.

L'output numerico di un programma di Analisi delle Corrispondenze precede sempre la stampa delle mappe fattoriali ed è sostanzialmente diviso in quattro parti. La prima, oltre a riprodurre, per controllo, la matrice dei dati con i suoi totali marginali ed il totale generale, riporta sempre le matrici dei profili con i profili marginali e, quando sono presenti, i profili illustrativi, che verranno discussi in dettaglio nella Sez. 4.13. La seconda, come si può vedere nella TAV. 30, contiene la tavola delle inerzie, e presenta per ciascun asse fattoriale fino ad un rango $A^* \leq A$ prefissato dall'utente: l'inerzia, il tasso d'inerzia, il tasso progressivo d'inerzia e il loro diagramma [1]. La terza parte riguarda i profili delle righe e la quarta quelli delle colonne, di solito, ma non sempre, in questo ordine, come mostrato nelle TAV. 31 e 32. Entrambe queste due parti riportano, per ogni asse e per ogni profilo: la modalità, la coordinata del corrispondente profilo, il contributo relativo del profilo all'inerzia dell'asse, la qualità della rappresentazione sull'asse e la massa del profilo. Viene spesso calcolata anche la quota d'inerzia del profilo nello spazio A^*-dimensionale. Limitare la stampa fino ad assi di questo rango - di solito A^* è dell'ordine di 3 o 4 - fa risparmiare carta ed evita di produrre informazioni su assi che si presume poco interpretabili. Quando presenti, i fattori di righe e/o colonne illustrative sono stampati in tavole separate che riportano, sempre per ogni asse e per ogni profilo: la modalità e la qualità della rappresentazione sull'asse. La presentazione di tutte queste informazioni può variare leggermente da un programma all'altro e può anche capitare che analizzando la stessa matrice con due programmi diversi ne risultino fattori con segni opposti: le ascisse che sono positive per uno sono negative per l'altro. Questo fatto è ininfluente sull'interpretazione dei risultati e dipende unicamente dalla convenzione sui segni adottata dalla routine di calcolo, come spiegato nelle Sez. 3.11 e 3.13.

L'ordine di stampa dei risultati presentato più sopra, corrisponde in linea di massima all'ordine con cui questi vanno esaminati, anche se il processo di interpretazione è difficilmente formalizzabile in una rigida successione di azioni, perché dipende molto dai dati, ma anche dall'esperienza e dall'abilità

[1] Talvolta detto impropriamente istogramma.

dell'analista, senza le quali la potenza esplorativa dell'Analisi delle Corrispondenze resta mortificata. Perciò l'interpretazione dei risultati non è univoca e solo l'attenta riflessione può garantirne la validità. Nei casi più complessi i migliori risultati si ottengono quando all'analista dei dati si affianca l'esperto del problema: medico, archeologo, economista, ecc. Infine un'avvertenza. Nelle pagine che seguono, verranno indicati dei valori di soglia per gli indicatori CTR_a e COS_a^2. Va subito precisato che qui il significato del termine 'soglia' non è quello dell'inferenza statistica. I valori di soglia indicati sono puramente indicativi perchè dipendono dalle dimensioni della matrice e variano, anche se non di molto, da caso a caso, particolarmente nel caso del contributo relativo. Si è preferito, però, indicare dei valori piuttosto che lasciare tutto nel vago col rischio che il lettore si faccia l'idea che nell'Analisi delle Corrispondenze tutto sia indefinito e nei suoi risultati si possa vedere ciò che si vuole. Con queste doverose premesse, quello che ora viene presentato è un tentativo di formalizzazione del modo di procedere all'interpretazione dei risultati di un'Analisi delle Corrispondenze. Serviranno da esempio i risultati dell'analisi della matrice *Spettacoli* di ordine 20×8 della TAV. 2.

1 - Considerazioni sui profili marginali

I profili marginali $\overline{\mathbf{r}}$ e $\overline{\mathbf{c}}$ sono la media ponderata degli I profili $\mathbf{r}_i$ e dei J profili $\mathbf{c}_j$. Geometricamente, rappresentano i baricentri delle due nuvole e l'origine degli assi fattoriali che indicano le principali direzioni di difformità dal profilo medio. Le coordinate dei profili sugli assi e sui piani sono sempre riferite al profilo medio.

Nel caso dell'esempio, l'esame della TAV. 4 e della TAV. 5 mostra che il 42 % degli spettatori ha preferito le rappresentazioni di Prosa che insieme agli spettacoli di Musica Leggera, hanno attratto i 2/3 degli spettatori italiani. L'esame della ripartizione geografica mostra il predominio di tre regioni - Lombardia, Lazio ed Emilia Romagna - che totalizzano il 40 % degli spettatori, il che suggerisce di effettuare un'analisi più dettagliata a livello provinciale in queste tre regioni, sfruttando la tecnica degli *elementi illustrativi* che verrà presentata nella Sez. 4.13.

2 - Esame delle inerzie

Molteplici sono le informazione che si possono trarre dall'esame della tavola delle inerzie, riportata nella TAV. 30: grado di fiducia con cui procedere alla caratterizzazione degli assi, numero di assi potenzialmente caratterizzabili, prevedibile facilità o difficoltà della loro caratterizzazione, ecc. Anche

se è non va mai dimenticato che l'inerzia è *uno* dei possibili indicatori della dispersione geometrica dei profili e l'esperienza mostra che in taluni casi può rivelarsi un indicatore "pessimistico" nel suggerire l'interpretabilità degli assi.

Inerzia rispetto all'origine

Prima di procedere all'esame dettagliato della Tavola delle inerzie, è sempre opportuno farsi un'idea del grado di fiducia, o di cautela, attribuibile alla caratterizzazione degli assi fattoriali. Anche se a questo delicato problema è dedicato l'intero Capitolo 7, è utile disporre fin dall'inizio di un indice che quantifichi questo grado di fiducia e che dovrà essere necessariamente legato al grado di dispersione geometrica dei profili, e quindi all'inerzia, perché tanto più i profili sono dispersi tanto più facile sarà caratterizzare gli assi e coglierne il significato. Questo indice di solito non viene stampato dai programmi d'analisi e va quindi calcolato a parte. Ecco come. Il teorema di Huygens lega l'inerzia complessiva rispetto all'origine a quella rispetto al baricentro $In_{\mathbf{0}} = 1 + In_{\overline{\mathbf{r}}} = 1 + In_{\overline{\mathbf{c}}}$ e nella Sez. precedente si è visto che inerzia 1 corrisponde alla situazione di completa omogeneità, ossia di indifferenza, tra le modalità delle righe e delle colonne. Così, nella TAV. 30, fatta 100 l'inerzia rispetto all'origine, risulta che il $92.6\,\% = 100/(1 + 0.079\,76)$ di questa è attribuibile ad una situazione di completa indifferenza tra le modalità, e che soltanto il rimanente 7.4 % è attribuibile ad effettive associazioni che si dovranno mettere in chiaro caratterizzando gli assi fattoriali. Quest'ultima percentuale, ossia $100 \times In_{\overline{\mathbf{r}}}/In_{\mathbf{0}}$, può assumersi in prima istanza come 'misura' dell'attendibilità da attribuire alla caratterizzazione degli assi prodotta dall'analista. Valori al di sotto dei pochi % dovrebbero indurre a cautela e spingere a un'analisi della stabilità della configurazione dei profili con i metodi del Capitolo 7.

Inerzie sugli assi

La TAV. 30 mostra come l'inerzia rispetto al baricentro $In_{\overline{\mathbf{r}}} = In_{\overline{\mathbf{c}}} = 0.079\,76$ sia stata scomposta lungo gli $A = min\,(20, 8) - 1 = 7$ assi fattoriali. Ogni riga si riferisce ad un asse e la colonna *Inerzia* ne elenca i valori, ordinati in modo decrescente. Inerzie e percentuali d'inerzia sull'asse dipendono molto dal tipo di matrice e dal caso analizzato. Nella Sez. 8.2 verrà mostrato che un'inerzia sull'asse prossima ad 1 rivela una associazione quasi perfetta tra un gruppo di righe da un lato ed un gruppo di colonne dall'altro. Comunque, inerzia grande indica proiezioni alquanto disperse sull'asse, piccola proiezioni molto ravvicinate all'origine. Per discriminare tra grande e piccolo di solito si tiene un valore di soglia intorno a 0.01. Così si ritiene che inerzie superiori alla soglia lascino prevedere nette opposizioni tra le modalità sui primi assi la cui

caratterizzazione sarà chiara, anche se spesso scontata, perché è da aspettarsi che gli aspetti più macroscopici del fenomeno rilevato si manifestino per primi. In base a questo criterio ciò è da attendersi per i primi due assi dell'esempio che hanno inerzie $\lambda_1 = 0.036\,97$ e $\lambda_2 = 0.025\,22$ rispettivamente. Inerzie sull'asse inferiori alla soglia lasciano invece presagire contrapposizioni più deboli e sfumate, e quindi meno ovvie, ma per questo potenzialmente interessanti, la cui interpretazione, di solito meno immediata, è bene controllare sugli assi di rango superiore e/o nelle matrici $\mathbf{R}$ e $\mathbf{C}$ dei profili o nella matrice $\mathbf{S}$ degli scarti. Massima cautela infine nell'interpretazione degli assi di rango elevato che, peraltro, mettono talvolta in luce aspetti interessanti del fenomeno rilevato e che l'attenzione e l'abilità dell'analista sapranno cogliere. Qui un puntuale riscontro nelle matrici dei profili è tassativo.

Percentuali d'inerzia sugli assi

L'importanza relativa di un asse fattoriale può essere meglio rilevata esaminando la sua percentuale d'inerzia, ossia il rapporto percentuale tra l'inerzia sull'asse e l'inerzia complessiva riferita al baricentro. Si tratta di un'informazione *aggiuntiva*, perché nell'analisi delle Corrispondenze inerzia e percentuale d'inerzia sono grandezze indipendenti, in quanto, per avere τ_a non basta conoscere l'inerzia λ_a, occorre conoscere anche tutte le inerzie, o perlomeno la loro somma [1]. Perciò due matrici di contingenza con la medesima inerzia complessiva possono avere percentuali d'inerzia diverse: per esempio, la prima avere τ_1 molto grande e le successive τ_a piccole, la seconda percentuali che decrescono tutte lentamente. La prima nuvola è aghiforme, la seconda praticamente (iper)sferica. In ogni caso, le percentuali d'inerzia vanno sempre consultate perché forniscono importante informazioni. Così, nel caso dell'esempio, l'inerzia $\lambda_1 = 0.036\,97$ sul primo asse rappresenta il $46.35\,\%$ dell'inerzia $0.079\,76$ rispetto al baricentro. Scorrendo dall'alto verso il basso la colonna delle percentuali si nota che queste decrescono bruscamente passando dal secondo asse: $\tau_2 = 31.62\,\%$, al terzo: $\tau_3 = 9.53\,\%$. Il loro andamento viene meglio apprezzato esaminando il loro diagramma.

Quando il numero A di assi fattoriali è piccolo, dell'ordine di 4 o 5, perché è piccola una delle dimensioni I o J della matrice, le percentuali d'inerzia risultano alte, dovendo l'inerzia complessiva ripartirsi tra pochi assi.

---

[1] Nell'Analisi delle Corrispondenze le percentuali d'inerzia sugli assi sono *indipendenti* dall'inerzia complessiva rispetto al baricentro. La dimostrazione si può trovare in Lebart e al. (1984), op. cit. nella bibliografia del Cap. 5.

Percentuali progressive d'inerzia

La colonna successiva *Cumulo* elenca la somma progressiva delle percentuali d'inerzia. Così il primo valore coincide con τ_1 e l'ultimo vale 100, essendo la nuvola di punti contenuta esattamente in uno spazio ad $A = 7$ dimensioni.

Osservando i valori di questa colonna, si vede che la percentuale d'inerzia relativa alla mappa principale, ossia al piano individuato dai primi due assi fattoriali, rende conto del $77.97\,\%$ dell'inerzia, ossia della dispersione dei profili rispetto al baricentro, mentre nello spazio principale, ossia nello spazio tridimensionale individuato dai primi 3 assi fattoriali, tale precentuale sale all' $87.51\,\%$. Perciò esaminare la configurazione dei profili in questo sottospazio, significa rinunciare al 12.49% dell'informazione sulla effettiva localizzazione dei profili.

Diagramma delle percentuali d'inerzia

La rappresentazione grafica delle percentuali d'inerzia mediante barre di lunghezza proporzionale ai loro valori, permette di cogliere le differenze *relative* tra percentuali al crescere del rango. Nel caso dell'esempio le prime due risultano

$$\begin{aligned}(\tau_1 - \tau_2)/\tau_1 &= (46.35 - 31.62)/46.35 &= 0.32\\ (\tau_2 - \tau_3)/\tau_2 &= (31.62 - 9.53)/31.62 &= 0.70\end{aligned}$$

e il diagramma dà proprio l'impressione che passando dal primo al secondo asse si abbia un calo d'inerzia $\tau_2 - \tau_1$ di circa un terzo di τ_1 del primo asse, e che passando dal secondo al terzo il calo sia il $70\,\%$ di τ_2, ecc. Pur tenendo conto del fatto che il diagramma traduce soltanto approssimativamente queste differenze, perché è prodotto sostanzialmente da una macchina per scrivere usata come plotter, tuttavia, l'esame delle diminuzioni evidenziate dal diagramma suggerisce quali assi fattoriali siano potenzialmente suscettibili di interpretazione. Semplificando al massimo si può dire che, in generale, un diagramma delle inerzie ha sostanzialmente due andamenti tipici.

Il primo andamento è quello dell'esempio, ove il diagramma evidenzia una diminuzione irregolare delle inerzie sui primi due assi e, successivamente, una diminuzione regolare. In gergo si dice che il diagramma mostra un 'ginocchio' tra il secondo e il terzo asse. Dato che le inerzie misurano la dispersione rispetto all'origine delle proiezioni dei profili sugli assi, un andamento di questo tipo con due inerzie ben separate, indica che i primi due assi corrispondono a forme geometriche irregolari della nuvola e traducono contrapposizioni particolari tra profili sugli assi che vanno esaminate con attenzione. Questo

tipo di diagramma è piuttosto frequente e suggerisce di limitare la caratterizzazione agli assi che precedono il brusco calo perché di solito i restanti fattori non rappresentano che fluttuazioni aleatorie non interpretabili, rumore inevitabile che accompagna ogni rilevazione statistica. Questo fatto va comunque accertato perché capita sovente che almeno il primo di questi, ossia il terzo nel caso dell'esempio, risulti interpretabile.

L'altra forma comune di diagramma presenta invece una diminuzione regolare delle inerzie fin dal primo asse. Questo indica nuvole di forma quasi 'sferica', e quindi poco strutturate, i cui assi sono poco esplicativi. Un diagramma di questo tipo fa presagire assi di interesse limitato e una caratterizzazione non semplice.

Può infine presentarsi il caso che su due assi consecutivi le inerzie risultino quasi eguali, pur risultando nettamente diverse da quelle immediatamente precedenti e seguenti. Di solito si tratta di assi instabili, nel senso che il loro orientamento può modificarsi radicalmente se una riga o una colonna vengono eliminate dall'analisi. In questi casi occorre considerare il piano individuato dai due assi, e non gli assi separatamente, trattandosi in pratica di un sottospazio autonomo in cui la posizione degli assi è poco significativa perché definita a meno di una rotazione. Il caso di tre inerzie molto vicine è piuttosto raro: in questo caso bisognerebbe esaminare le proiezioni dei profili nel sottospazio tridimensionale da essi generato.

Quanti assi caratterizzare

Questo importante quesito, legato alla significatività dei fattori e quindi all'attendibilità della rappresentazione grafica dei profili, verrà affrontato per esteso nel Capitolo 7, sulla base di ipotesi statistiche di lavoro più o meno giustificabili. Comunque, la regola migliore cui attenersi resta quella di considerare soltanto quegli assi che si riesce a caratterizzare in base all'ovvia considerazione che se le proiezioni dei profili su un asse fossero dovute ad un fenomeno puramente aleatorio, la caratterizzazione risulterebbe impossibile.

3 - Caratterizzazione degli assi fattoriali

Caratterizzare o interpretare un asse significa dare all'asse un significato specifico in base alle contrapposizioni ed alle associazioni che le proiezioni dei profili assumono su di esso, traducendo i risultati dell'analisi, precisi, ma chiari soltanto all'analista, in un commento discorsivo che risulti comprensibile anche ai non esperti. Per fare questo occorre avere sempre presente che ogni profilo, inteso come punto dotato di massa in uno spazio euclideo,

rappresenta una distribuzione condizionata di frequenze nel caso di una matrice di contingenza, ed è legato ad una specifica modalità della matrice. La sua massa ne misura l'importanza relativa. Il profilo medio, all'origine del sistema degli assi fattoriali, rappresenta invece la distribuzione marginale. La distanza distribuzionale è definita unicamente tra profili delle righe e tra profili delle colonne e la distanza tra le proiezioni di due profili su un asse è l'immagine della effettiva distanza tra profili nel loro spazio ambiente, e quindi del grado di similarità tra le due distribuzioni: immagine più o meno fedele a causa delle distorsioni dovute alla proiezione. Errori di prospettiva che si possono valutare per ogni singolo profilo grazie agli indicatori di supporto alla caratterizzazione: il contributo del profilo all'inerzia dell'asse e la qualità della sua rappresentazione sull'asse, strumenti *indispensabili* per una corretta caratterizzazione. Infine, è bene tenere sempre presente che nei due spazi $\Re^J$ ed $\Re^I$ le due nuvole sono contenute in sottospazi della medesima dimensione A e che le relazioni di transizione giocano un ruolo fondamentale nell'interpretazione dei risultati perché mettono in corrispondenza *biunivoca* assi fattoriali del medesimo rango dei due spazi, assi sui quali la percentuale d'inerzia è la medesima. È così giustificato far coincidere i baricentri delle due nuvole e sovrapporre le coppie di assi dello stesso rango: $\mathbf{u}_1^\star$ e $\mathbf{v}_1^\star$, $\mathbf{u}_2^\star$ e $\mathbf{v}_2^\star$, ecc. ottenendo una rappresentazione *simultanea* dei profili, come si vede nella TAV. 35.

Ogni asse fattoriale va caratterizzato singolarmente, a partire dal primo, esaminando separatamente le proiezioni dei profili delle righe e delle colonne. Si inizia col dividere i profili in due gruppi: da una parte quelli con ascissa negativa sull'asse e dall'altra quelli con ascissa positiva, come si vede nella TAV. 31 per i profili delle colonne sul primo asse e nella TAV. 32 per i profili delle righe. All'interno di questi due gruppi, i profili vanno ordinati per valori decrescenti del loro contributo all'inerzia dell'asse.

La caratterizzazione di un asse è basata essenzialmente sull'esame dei profili che, sia sulla parte negativa che positiva dell'asse, hanno:

- un contributo relativo elevato. Elevato significa che pochi profili totalizzano un contributo superiore alla media o ad una quota significativa ad es. 0.500 o 0.600. Sono i profili più significativi perché hanno maggiormente contribuito ad orientare l'asse. Per questo si parla di modalità che *fanno l'asse*, intendendo con ciò che se nella matrice si sopprime una riga o una colonna che *non* sia una di queste, i risultati dell'analisi mutano di poco.

- una coordinata estrema *ed* una elevata qualità di rappresentazione. Si tratta di profili ben caratterizzati perché molto diversi dal profilo medio e queste differenze sono anche ben rappresentate perché i profili vengono a trovarsi praticamente sull'asse che, peraltro, non hanno contribuito ad orientare a causa della loro massa ridotta. Mettono in luce quindi modalità che caratterizzano esclusivamente un asse e che non giocheranno più alcun ruolo importante per gli altri assi fattoriali.
- una coordinata estrema *ed* una sufficiente qualità di rappresentazione. Mettono in luce modalità che intervengono soltanto parzialmente nella caratterizzazione dell'asse, perché legate anche ad altri assi, il che rende talvolta problematica la loro attribuzione a questo o a quell'asse.

Le prossimità tra proiezioni di profili dello stesso tipo, righe *o* colonne, quando abbiano una buona qualità della rappresentazione sull'asse, rivelano profili di 'forma' simile, e quindi un legame tra le modalità alle quali sono associati.

L'interpretazione delle prossimità tra proiezioni di profili delle righe *e* delle colonne, deve necessariamente basarsi sulle relazioni di transizione viste nella Sez. 4.9, e quindi sul cosidetto 'principio baricentrico', *non* essendo definita una distanza tra profili delle righe e delle colonne, dato che si trovano in spazi diversi.

Con questi suggerimenti come guida, ecco come si caratterizzano i primi 3 assi ottenuti dall'analisi della matrice *Spettacoli* della TAV. 2, limitandosi quindi ad una approssimazione $\mathbf{N}^{(3)}$ della matrice di contingenza, come si è visto nella Sez. 4.10.

Asse 1: la quota d'inerzia su questo asse è del $46.35\,\%$. Ciò significa che limitare l'analisi a questo asse comporta la rinuncia al $53.65\,\%$ dell'informazione sulla effettiva localizzazione dei profili nel loro spazio.

Profili delle colonne nella TAV. 31: l'asse è caratterizzato dagli spettacoli di Lirica e Balletti che hanno un contributo molto elevato $[\,CTR_1(\mathbf{c}_2) = 0.870\,]$. Questo profilo nel suo spazio ambiente a 7 dimensioni risulta alquanto eccentrico perché la sua massa è di molto inferiore all'inerzia che ha rispetto al baricentro $[\,\overline{\mathbf{r}}_2 = 0.109 \ll INR_7(\mathbf{c}_2) = 0.404\,]$ e si trova praticamente sull'asse $[\,COS_1^2(\mathbf{c}_2) = 0.998\,]$, fatto del resto prevedibile, dato che da solo ha orientato l'asse. Dalla parte negativa, la Prosa, i Concerti e gli spettacoli di Burattini e Marionette hanno un contributo infimo - la somma dei loro contributi non arriva al $13\,\%$ ed inoltre sono tutti mal rappresentati sull'asse. Si noti come la Prosa si trovi in una situazione opposta a quella della Lirica, perché ora la massa è molto superiore all'inerzia $[\,\overline{\mathbf{r}}_2 = 0.419 \gg INR_7(\mathbf{c}_2) = 0.116\,]$: si

tratta quindi di un profilo che in $\Re^7$ si trova piuttosto vicino al baricentro.

Profili delle righe nella TAV. 32: l'asse è caratterizzato dal Veneto che ha un contributo piuttosto elevato $[\, CTR_1(\mathbf{r}_5) = 0.832 \,]$. Il profilo si trova praticamente sull'asse $[\, COS_1^2(\mathbf{r}_5) = 0.995 \,]$ ed è molto eccentrico $[\, \overline{\mathbf{c}}_5 = 0.090 \ll INR_7(\mathbf{r}_5) = 0.388 \,]$. Nella parte negativa dell'asse i contributi sono esigui per tutti i profili. Scorrendo poi la colonna della qualità della rappresentazione si nota che questa è alta soltanto per la Lombardia $[\, COS_1^2(\mathbf{r}_3) = 0.711 \,]$, ma si tratta di un profilo che si proietta vicino all'origine $[\, f_{31} = -0.069 \,]$ perché nel suo spazio ambiente non è lontano dal baricentro $[\, \overline{\mathbf{c}}_3 = 0.178 \gg INR_7(\mathbf{r}_3) = 0.015 \,]$. Si tratta quindi di un profilo poco caratterizzato, perché poco difforme dal profilo medio.

In definitiva quindi il primo asse mette in rilievo la netta opposizione tra gli spettacoli di Lirica e Balletto e gli altri tipi di spettacoli e la stretta associazione con la regione Veneto. Ciò indica che questo è il tipo di spettacolo predominante nella regione e, reciprocamente, che Lirica e Balletti hanno nel Veneto la loro maggior quota di spettatori. Si tratta del fenomeno più importante, già rilevato nella matrice degli scarti relativi dalla media di TAV. 8, nella quale l'indice risulta $s_{52} = 1.7$. La stessa tavola mostra però che lo scarto relativo per le rappresentazioni di Saggi Coreografici e Folkloristici e la Valle d'Aosta è quasi triplo perché vale ben 6.6. Perchè allora il primo asse non mette in luce per primo questo legame, che appare ben più stretto? La risposta sta nel fatto che questi due profili sono sì lontani dal baricentro, ma dotati di masse irrisorie per cui il loro contributo all'inerzia di un asse risulta limitato, come mostra questo specchietto

	Profili	*Masse*	*Inerzia*
$j = 8$	Saggi Coreog. Fol.	0.017	0.066
$i = 2$	Valle d'Aosta	0.002	0.026
$j = 2$	Lirica e Balletti	0.109	0.404
$i = 5$	Veneto	0.090	0.388

Così l'Analisi delle Corrispondenze assegna ad ogni associazione tra modalità la giusta importanza.

Asse 2: la quota d'inerzia su quest'asse è del 31.62 %

Profili delle colonne nella TAV. 33: mentre il primo asse è stato nettamente orientato dai profili più eccentrici e con massa elevata, è da attendersi che l'orientamento degli assi successivi sarà più influenzato dalle masse, dal momento che ora i profili si trovano tutti a distanze compara-

bili. Ecco infatti che su questo secondo asse il profilo delle rappresentazioni di Prosa [$CTR_2(\mathbf{c}_1) = 0.264$] si oppone ai Concerti di Musica Leggera [$CTR_2(\mathbf{c}_6) = 0.600$] e [$COS_2^2(\mathbf{c}_6) = 0.911$]. Sono i due profili che hanno massa più elevata, e che si trovano non troppo lontani dal baricentro. Nessun altro profilo è ben rappresentato sull'asse.

Profili delle righe nella TAV. 33: Sicilia e Lazio ad una estremità e Sardegna e Piemonte dall'altra hanno contributi significativi per questo asse. Si noti come le due isole si oppongano sull'asse. Inoltre, scorrendo la colonna della qualità di rappresentazione, si trovano due profili interessanti. La regione Marche è molto ben rappresentata sull'asse [$COS_2^2(\mathbf{r}_{11}) = 0.874$], trovandosi anche sufficientemente lontana dall'origine [$f_{11,2} = 0.212$]. Si tratta quindi di un profilo che è tipico di questo asse, pur non avendo contribuito granché all'inerzia dell'asse [$CTR_2(\mathbf{r}_{11}) = 0.036$]. Anche il profilo della Valle d'Aosta ha una buona qualità di rappresentazione [$COS_2^2(\mathbf{r}_2) = 0.588$] ed è molto eccentrico: un profilo dunque ben caratterizzato e ben rappresentato sull'asse, anche se, a causa della massa esigua [$\overline{\mathbf{c}}_2 = 0.002$] non ha contribuito molto all'inerzia dell'asse.

In conclusione questo asse oppone le rappresentazioni di Prosa, associate con Lazio e Sicilia, ai concerti di Musica Leggera associati con Sardegna e Marche. Queste associazioni sono più deboli di quelle che sul primo asse legavano il Veneto agli spettacoli di Musica Lirica e ai Balletti. L'asse mette in luce anche il fatto, già notato nella Sez. 1.5, che i profili di Sicilia e Valle d'Aosta sono in netta controtendenza per cui si posizionano alle estremità opposte dell'asse.

Asse 3: la quota d'inerzia su quest'asse è del 9.53 %

Profili delle colonne nella TAV. 34: questo asse oppone gli spettacoli di Rivista [$CTR_3(\mathbf{c}_5) = 0.323$] a quelli di Operetta [$CTR_3(\mathbf{c}_4) = 0.258$] che sono ben rappresentati sull'asse [$COS_3^2(\mathbf{c}_4) = 0.629$].

Profili delle righe nella TAV. 33: qui Lazio e Campania si oppongono a Sicilia, Friuli ed alla Puglia che, pur con contributo più debole, è molto ben rappresentata sull'asse [$COS_3^2(\mathbf{r}_4) = 0.833$].

Si può concludere che questo terzo asse associa Lazio e Campania alla Rivista e Sicilia, Friuli e Puglia all'Operetta.

Contributi abnormi all'inerzia di un asse

Poco sopra sono stati elencati tre tipi di profili che concorrono a caratterizzare un asse. Di questi, il primo tipo – profili con contributo relativo elevato – è decisamente il più importante perchè i contributi tengono conto

anche della massa del profilo. Ora, il grado di generalità di un asse, e quindi il suo interesse ai fini dell'interpretazione, dipende molto dal *numero* di profili che hanno contribuito a definirlo. Accade però sovente che si trovi *un solo* profilo con contributo all'inerzia molto elevato, indice inequivocabile che la sua 'forma' è nettamente diversa dalle altre. Il fatto che un solo profilo orienti un asse, di solito il primo, ostacola lo studio degli altri profili perché le loro proiezioni sull'asse non si troveranno nelle posizioni più adatte a rivelare *al meglio* le distanze effettive. Anche gli assi successivi ne vengono influenzati, essendo legati al primo dal vincolo di ortogonalità.

Questo fatto, peraltro normale quando il numero A di fattori è piccolo, dell'ordine di 4 o 5, può creare problemi di interpretazione quando si presenta con valori di A più grandi. È il caso del primo asse dell'esempio. La TAV. 31 mostra che degli 8 profili delle colonne, soltanto quello relativo alla modalità Spettacoli Lirici e Balletti ha ascissa positiva ed un contributo elevatissimo, pari a 0.870, mentre la media degli 8 contributi vale $1/8 = 0.125$, dato che i contributi relativi hanno somma 1, come si è visto nella Sez. 4.5.

In casi come questo, la prima possibilità da considerare è quella di rifare l'analisi togliendo il profilo da quelli attivi e trattandolo come illustrativo, procedimento che verrà mostrato nella prossima Sez. 4.13. L'asse in questione sparisce e il resto dell'analisi resta sostanzialmente immutato, mentre gli assi vengono scalati di un posto. L'asse $a = 1$ della nuova analisi corrrisponde all'asse $a = 2$ della precedente, e così via. Ad ogni modo, questa scelta va fatta sempre con giudizio perché può spostare il campo di studio che nel presente caso concerne le attività teatrali e musicali. Se si considerasse come illustrativo il profilo della colonna Lirica e Balletti, l'analisi verrebbe sbilanciata verso gli spettacoli teatrali. Un'operazione del genere va fatta quindi dopo attento esame degli obiettivi dell'analisi e comunque, una volta fatta, devono essere ben precisati gli ambiti del *nuovo* campo di studio che essa implica.

Una seconda possibilità è l'accorpamento con un altro profilo. Però, in questi casi verrebbero accorpati profili molto diversi con conseguente difficoltà d'interpretazione dei risultati.

In taluni casi si possono operare delle correzioni 'ad hoc' che dipendono strettamente dal particolare caso analizzato. Nel caso dell'esempio si potrebbe rifare l'analisi dopo aver sottratto ai $686\,236$ biglietti venduti nel Veneto per spettacoli di Lirica e Balletti, i $568\,300$ biglietti venduti all'Arena di Verona, teatro piuttosto anomalo nel panorama dei teatri lirici italiani e

che da solo ha fatto vendere nel 1991 l'83% dei biglietti di questo tipo di spettacolo nel Veneto.

Se si decide invece di lasciare tutto come è, è sempre bene esaminare le proiezioni dei profili in mappe bidimensionali che non contengano il primo asse, o, ancora meglio, in spazi tridimensionali, perché gli assi successivi al primo cercano di recuperare l'effettiva struttura della nuvola.

4 - Interpretazione delle rappresentazioni bidimensionali

Rispetto all'esame degli assi, l'esame dei piani unisce al potere sintetico ed espressivo del grafico quello di un superiore potere informativo, che della configurazione dei profili rende un'immagine più vicina al reale, fonte sovente di ulteriori spunti all'interpretazione. La qualità della rappresentazione dei profili *sul piano* gioca ora un ruolo centrale, perchè permette di filtrarli, eliminando quelli mal rappresentati. Mappe così sfoltite impediscono errori di prospettiva, e perciò d'interpretazione, spesso imputabili all'eccessiva compattezza con cui i programmi d'analisi presentano i risultati. È importante quindi che il software d'analisi permetta di accedere a *tutti* i risultati, in modo che si possano costruire dei programmi in grado di leggerli, rielaborarli e presentarli nella forma più adatta all'interpretazione, ad esempio come nelle TAV. 36, 38 e 40. Va da sé che in seguito il lettore divenuto esperto sarà in grado di leggere i risultati, per quanto compatti, senza bisogno di ulteriori elaborazioni. La soglia del filtro va fissata ad un valore tale che permetta di far apparire sulla mappa tutti i profili che hanno concorso a caratterizzare i due singoli assi, ma non deve comunque scendere sotto 0.400 o 0.500 per garantire una rappresentazione accettabile. Perciò per fissare la soglia occorre aver prima caratterizzato i due assi, e questo costringe a seguire la regola aurea: 'passare dagli assi per interpretare la mappa'. La mappa diviene così lo strumento che permette una 'corretta' lettura delle TAV. 4 e 5 dei profili e della TAV. 8 degli scarti relativi dalla media, alle quali il lettore è invitato a fare *continuo* riferimento. In particolare, per due profili *ben rappresentati* sul piano (a, b) ed *eccentrici*, e perciò ben diversi dal profilo medio, lo scarto relativo di (i, j) dalla media, espresso dalla (4.10.2) si riduce a

$$s_{ij} = \sum_{a=1}^{A} \frac{1}{\sqrt{\lambda_a}} f_{ia}\, g_{ja} \simeq \frac{1}{\sqrt{\lambda_a}} f_{ia}\, g_{ja} + \frac{1}{\sqrt{\lambda_b}} f_{ib}\, g_{jb}.$$

dato che la qualità della loro rappresentazione è elevata praticamente solo su questo piano (a, b). Di conseguenza, se le proiezioni dei due profili si trovano

vicine sulla mappa, per cui $f_{ia} \simeq g_{ja}$ e $f_{ib} \simeq g_{jb}$, i due prodotti dei fattori sono positivi, e tra le due modalità i e j c'è 'attrazione', perché la quota di spettatori è in eccesso rispetto alla quota media. Viceversa se si trovano lontane, tra le due modalità c'è 'repulsione' perché la quota di spettatori risulta deficitaria rispetto alla quota media.

Per poter apprezzare correttamente sulla mappa le distanze, gli angoli e le proiezioni ortogonali è *necessario* che l'unità di scala sia la *stessa* sui due assi, in modo che un quadrato reale non sia rappresentato come un rettangolo.

Mappa 1,2: la percentuale d'inerzia su questo piano è del $77.97\,\%$. Ciò significa che si rinuncia al $22\,\%$ dell'informazione sull'effettiva localizzazione geometrica dei profili.

La TAV. 36 riporta in forma facilmente leggibile le informazioni essenziali per interpretare correttamente la mappa principale che è mostrata nella TAV. 37. Sono presi in considerazione soltanto i profili con qualità della rappresentazione sul piano $COS^2_{(1,2)} > 0.580$, per far apparire il profilo del Piemonte, intervenuto nella caratterizzazione del secondo asse, che ha $COS^2_{(1,2)} = 0.588$. Questa soglia lascia apparire altri profili che non erano intervenuti nella caratterizzazione del primo e del secondo asse, ma che risultano ben rappresentati sul piano. Sono Valle d'Aosta con $COS^2_{(1,2)} = 0.661$, Lombardia con $COS^2_{(1,2)} = 0.743$ ed Emilia Romagna con $COS^2_{(1,2)} = 0.711$, anche se solo la prima regione si trova a sufficiente distanza dal profilo medio.

Sulla mappa appaiono tre poli associativi distinti. Il primo, ben correlato al primo asse, mostra che, rispetto alla media nazionale, nel Veneto sono nettamente privilegiati gli spettacoli di Lirica e Balletti ($s_{52} = 1.7$). Questo fatto, il più rilevante, respinge ogni altra associazione nella parte sinistra della mappa, con i profili allineati lungo il secondo asse.

In basso, Lazio e Sicilia si trovano vicine a Prosa. Mentre si può dire che nelle due regioni gli spettatori sono principalmente interessati a questo tipo di rappresentazione, è difficile capire se Prosa è stata attirata da un singolo profilo o da entrambi dato che non sono abbastanza eccentrici per cui non si può dire che le rappresentazioni di Prosa sono maggiormente seguite nelle due regioni. Per vederlo accorre esaminare le posizioni dei profili sul terzo asse fattoriale.

Il terzo polo riguarda i concerti di Musica Leggera e Folkloristica che, in diversa misura, vengono preferiti in Piemonte, Marche, Sardegna e Valle d'Aosta. La posizione eccentrica della Valle d'Aosta è dovuta alla sua massa

esigua [1] e alla repulsione della Prosa ($s_{21} = -0.5$) ma, sopratutto, di Lirica e Balletti ($s_{22} = -1.0$). Ciò significa che nella regione le quote di spettatori a questi due tipi di spettacolo risultano inferiori alla media nazionale.

Sul secondo asse, Piemonte e Marche sono molto vicine. Ciò indica che i loro profili hanno 'forma' simile e, di conseguenza, che in queste due regioni le affluenze di spettatori agli 8 tipi di spettacolo sono proporzionali. Questo fatto era già stato rilevato nella Sez. 1.5, ove si era anche visto come le 'forme' dei profili di Valle d'Aosta e Sicilia fossero 'in controtendenza' e difatti le loro rappresentazioni sulla mappa risultano lontane. Infine, Emilia Romagna e Lombardia sono prossime al centro della mappa, che rappresenta il profilo medio. Infatti, come risulta dalla TAV. 6, il profilo medio nazionale è molto simile a questi due profili. Le due regioni possono pertanto prendersi come rappresentative del comportamento degli spettatori italiani nel 1991.

Rimarchevole è il fatto che le regioni italiane siano mescolate: dal punto di vista degli spettacoli non risultano differenze tra spettatori del Nord, Centro e Sud.

Mappa 2,3: la percentuale d'inerzia sul piano è del $41.15\,\%$.

La TAV. 38 elenca i profili che hanno superato un livello di filtraggio pari a 0.470, poco superiore alla più bassa qualità di rappresentazione che è della Campania con $COS^2_{(2,3)} = 0.471$.

Tradizionalmente l'asse fattoriale di rango più basso è quello orizzontale, ma, per motivi di spazio, nella mappa riportata nella TAV. 39 il secondo asse è quello verticale ed il terzo quello orizzontale. La mappa inizia a rivelare fenomeni che restavano offuscati nella mappa precedente – e in quella $(1, 3)$ che non è stata mostrata – a causa dell'orientamento del primo asse, forzato dalla stretta associazione tra Lirica e Veneto. Ai due poli associativi del secondo asse, Prosa e Concerti di Musica Leggera, e del terzo, Rivista ed Operetta, viene ora ad aggiungersi quello dei Saggi Coreografici e Culturali, un profilo ben correlato al piano, anche se di massa esigua.

In basso a sinistra, la posizione di Lazio e Liguria, intermedia tra Prosa e Rivista, rivela che le due regioni hanno quote di spettatori superiori alla media nazionale, per questi due tipi di spettacolo.

Le rappresentazioni di Operette risultano preferite in Sicilia, in Puglia e nel Friuli V.G., mentre la Valle d'Aosta privilegia nettamente i Saggi Coreografici e Folkloristici. Questa mappa ne rivela lo stretto legame ($s_{28} = 6.6$).

---

[1] Il baricentro è l'equivalente geometrico di una media ponderata e tende quindi a collocarsi verso i profili con massa maggiore.

Mappa 3,4: la percentuale d'inerzia è del 15.75 % .

Anche se la percentuale è alquanto bassa, l'esame di questa mappa potrebbe rivelare qualche fenomeno interessante, costretto a mostrarsi soltanto ora a causa dell'orientamento forzato del primo asse. Come si vede nella TAV. 40, con un filtro posto a 0.440 superano la soglia soltanto 4 regioni, peraltro tutte piuttosto centrali sulla mappa, e 3 tipi di spettacolo. Questo è dovuto al fatto che masse e distanze dal baricentro dei profili filtrati tendono a diminuire, come si vede nelle TAV. 36, 38 e 40, mano a mano che cresce il rango degli assi che individuano i piani, dato che l'inerzia sui piani diminuisce.

La mappa della TAV. 41 non mostra associazioni che non siano già state rilevate nella mappa precedente e anche il nuovo profilo Concerti di Musica Classica non ne mostra alcuna, perché l'unica sua 'affinità' è con l'Abruzzo ($s_{13,3} = 0.7$) e questa appariva nella mappa (1,3), che non è stata mostrata. In compenso, questa mappa mette in luce due 'repulsioni': la Campania ha una quota di spettatori deficitaria per l'Operetta ($s_{15,4} = -0.8$) ed il Friuli V. G. per la Rivista ($s_{6,5} = -0.6$). Questo spiega le loro posizioni sulla Mappa. Trattandosi comunque di regioni con profili non molto discosti dal profilo medio, la cautela è d'obbligo.

Insidie e cautele

Le pagine precedenti hanno evidenziato che per interpretare correttamente una mappa fattoriale, è bene tener presente questi criteri:

- osservare bene il rango dell'asse orizzontale e di quello verticale,
- controllare che i due assi abbiano la stessa unità di scala, come su una carta geografica,
- valutare la percentuale d'inerzia del piano, per farsi un'idea della dispersione geometrica delle proiezioni,
- non dimenticare che si confrontano *profili* e quindi vettori di frequenze relative, e *non* di frequenze assolute.
- ricordare che i profili sono raffigurati *relativamente* al profilo medio, che si trova all'origine degli assi. Evitare perciò affermazioni assolute del tipo 'una gran / piccola parte di ...', usando invece espressioni come 'sopra / sotto la media ...', ecc.,
- prendere in esame i profili con contributo relativo elevato, e quelli ben rappresentati *ed* eccentrici,
- filtrare i profili, rappresentando sulla mappa soltanto i profili potenzialmente interessanti,
- rammentare che le distanze tra proiezioni traducono abbastanza fedel-

mente le differenze di 'forma' tra profili delle righe, tra profili delle colonne e tra questi e il profilo medio,

- ricordarsi che non è *mai* stata definita una distanza tra un profilo riga e uno colonna, perché si trovano in spazi diversi.
- interpretare le prossimità tra profilo riga e profilo colonna come 'tendenza', perché in base alle relazioni di transizione, un profilo riga è il 'baricentro' di *tutti* i profili delle colonne e viceversa,
- fare sempre riferimento alle matrici **R** e **C** dei profili e **S** degli scarti relativi di (i, j) dalla media,
- ricordarsi che l'importanza delle 'attrazioni' e delle 'repulsioni' tra profili è data dall'inerzia dell'asse o del piano fattoriale che le rivela.

Attenzione particolare merita l'interpretazione delle prossimità tra proiezioni di profili appartenenti a nuvole *diverse*. Dato che un solo esempio non ha permesso di mostrare tutta la casistica, è opportuno ricorrere al caso fittizio di una matrice che ripartisca i biglietti venduti per regione e per attività sportiva. A questo si riferiscono le sei situazioni schematizzate nella TAV. 42. È opportuno ricordare che alla base dell'interpretazione c' il principio baricentrico delle relazioni di transizione: quanto più nel profilo colonna dell'attività sportiva j è elevata la quota c_{ij} di spettatori, tanto più la regione i 'attira' lo sport j ed i due profili saranno vicini. Se poi sono ben rappresentati su una mappa, la prossimità delle proiezioni traduce l'effettiva prossimità dei due profili. Non sempre però è vero il contrario, perché anche per profili ben rappresentati sulla mappa, la prossimità tra proiezioni di un profilo riga e colonna può risultare da attrazioni diverse, particolarmente quando i profili sono nella zona centrale della mappa. Invece questo non si verifica quando le proiezioni si trovano nella zona *periferica*, sempre che i profili siano ben rappresentati sul piano.

La TAV. 42 si riferisce al caso fittizio e schematizza sei casi che possono presentarsi frequentemente:

1 La prossimità tra due profili *centrali* come Calcio e Campania è scarsamente indicativa. Il Calcio potrebbe essere stato attirato nella posizione in cui è, da un lato dal Lazio e dall'altro e da Valle d'Aosta e Trentino. Lo stesso potrebbe dirsi per la Campania, che potrebbe essere stata spinta in quella posizione da Nuoto, Tennis e Concorsi Ippici. Per assicurarsi che la prossimità traduca una quota particolarmente elevata di spettatori del Calcio in Campania, occorre controllare nelle mappe di rango superiore: se i due profili vi risultano ancora

vicini, però eccentrici, è presumibile un'attrazione tra le due modalità. Per eliminare ogni dubbio è opportuno anche esaminare le matrici dei profili e degli scarti relativi e accertarsi che r_{ij}, quota di spettatori campani alle partite di calcio, sia superiore a $\overline{r}_j$, quota di spettatori italiani alle partite di calcio, o che c_{ij}, quota di spettatori del calcio in Campania, sia superiore a $\overline{c}_i$, quota di spettatori campani, oppure infine che s_{ij} sia positivo.

2 Questa prossimità tra due profili eccentrici traduce chiaramente una effettiva attrazione tra Pallacanestro e Marche. Se ne deduce con sicurezza che per gli spettatori marchigiani la Pallacanestro è lo sport più seguito, e anche che gli spettatori della pallacanestro sono essenzialmente concentrati nelle Marche.

3 Questo caso è analogo al precedente, con la differenza che ora i due profili orientano l'asse, rendendo più difficile l'interpretazione delle prossimità tra i rimanenti profili.

4 Dei 3 profili, Valle d'Aosta e Trentino sono senza dubbio attratti dallo Sci, per cui si può concludere che gli spettatori trentini e valdostani sono prevalentemente interessati alle discese Sciistiche. Invece è difficile capire se il profilo dello Sci è stato attirato da una singola regione o da tutte e due in modo equilibrato. Si può comunque affermare che nelle altre regioni lo Sci è meno seguito che in Trentino e Valle d'Aosta. Quanto detto vale anche per il caso di due sport e 1 regione. Oltre che esaminare la posizione dei 3 profili nelle mappe di rango superiore, è opportuno cercare un riscontro nelle matrici dei profili o degli scarti relativi.

5 Quattro profili sono coinvolti. Ciascuno dei due sport, Golf e Tennis, è stato certamente attratto dalla coppia di regioni Veneto e Friuli, però non è possibile capire in quale rapporto. Si desume, comunque, che nessun'altra regione ha una quota rilevante nei profili di questi due sport. Analogamente, nei profili delle due regioni, gli sport diversi da Golf e Tennis hanno quote minime di spettatori, ma nelle due regioni è impossibile sapere se uno dei due sport è preponderante. L'informazione mancante va cercata esaminando la posizione dei profili in mappe di rango superiore, o controllando nelle matrici dei profili o degli scarti.

6 La posizione della regione Lazio, dalla parte di Pugilato e Concorsi Ippici, traduce senza ambiguità il fatto che gli spettatori laziali seguono prevalentemente questi due sport, e in modo equilibrato. Al contrario,

siccome nessuno dei due sport si trova vicino alla regione Lazio, non si può affermare che i loro spettatori sono prevalentemente laziali perché, ad esempio, i Concorsi Ippici sono chiaramente attratti anche dalla regione Marche ed i suoi spettatori si ripartiscono prevalentemente tra queste due regioni.

5 - Interpretazione delle rappresentazioni 3-D

La rappresentazione dei profili su una mappa è facilmente ottenibile perché si utilizza la stampante come fosse un plotter. Tuttavia oggi sono disponibili dei programmi specifici che consentono di visualizzare la nuvola dei profili in spazi tridimensionali, dando sullo schermo grafico a colori del computer un'immagine della configurazione ancora più vicina a quella reale. La rappresentazione può esser fatta ruotare in tutte le direzioni per controllare che gli eventuali gruppi di profili restino tali quando la configurazione viene osservata da ogni lato. In più, la possibilità di colorare diversamente i profili e di simulare l'effetto profondità rendono prezioso questo strumento per tutti coloro che si occupano di analisi multidimensionale. Programmi con queste caratteristiche e altamente interattivi cominciano ad essere disponibili su molte piattaforme.

4.12 - Profili illustrativi

Il lettore che ha avuto la costanza di arrivare a questa Sezione si sarà fatto l'idea che l'Analisi delle Corrispondenze sia sostanzialmente un metodo per convertire una matrice di numeri non negativi in una particolare raffigurazione grafica ove le righe e le colonne della matrice vi sono rappresentate come punti. L'idea quindi di un metodo puramente descrittivo, che consentirebbe soltanto di esplorare la struttura dei dati. Quest'idea è corretta, ma parzialmente vera e questa Sezione mostrerà come grazie all'impiego di *profili illustrativi*[1] , l'Analisi delle Corrispondenze possa trasformarsi in un efficace strumento di indagine per mettere alla prova congetture e modelli concettuali, allargando dunque il contesto dell'interpretazione. Il lettore è invitato a riflettere attentamente sul contenuto di questa Sezione, perché non è possibile sfruttare pienamente le potenzialità dell'Analisi delle Corrispondenze, senza padroneggiare la tecnica dei profili illustrativi, che hanno un ruolo cruciale nell'Analisi delle Corrispondenze Multiple, come si vedrà nel prossimo Capitolo 5.

[1] Sono chiamati spesso *elementi supplementari* o *elementi passivi*.

I profili illustrativi sono in sostanza dei profili addizionali che non intervengono nella creazione degli assi fattoriali, ma che vengono proiettati su assi e su piani fattoriali *predeterminati*, per arricchirne o illustrarne meglio il contenuto. Ovviamente le righe e le colonne illustrative, da cui vengono ottenuti i profili dividendo ogni loro elemento per il totale degli elementi, devono essere dello stesso ordine I e J di quelle che intervengono nella costruzione degli assi e che prendono il nome di righe e colonne *attive*. Queste definiscono la struttura informativa di base, alla quale possono essere riferite altre informazioni, costruendo così un ipotetico modello concettuale. L'Analisi delle Corrispondenze permette allora di indagare la struttura empirica delle relazioni tra elementi di base ed elementi illustrativi e di avvalorare o meno il modello. La distinzione tra profili attivi ed illustrativi ha grandi analogie con quella che sussiste nella regressione multipla tra variabili da esplicare, o dipendenti, ed esplicative, o indipendenti.

Perché le distanze tra profili attivi abbiano un senso, è necessario che questi ultimi costituiscano un insieme omogeneo, e che si rifeririscano tutti al medesimo tema, al medesimo punto di vista, mentre i profili illustrativi possono riferirsi a variabili eterogenee. In generale vengono considerate come righe o colonne illustrative:

- variabili incerte, ossia con grado di affidabilità inferiore alle altre perché per esempio rilevate in condizioni diverse o più difficili. Si supponga ad esempio che sia stata fatta un'indagine campionaria per rilevare nelle 20 regioni quante volte nel 1991 gli intervistati sono andati al cinema: mai, una volta all'anno, una al mese, ecc. Ciascuna di queste modalità andrebbe senz'altro trattata come colonna illustrativa della matrice *Spettacoli* di TAV. 2, dal momento che non è stata ottenuta come le altre dal conteggio dei biglietti venduti.
- elementi aberranti, ossia con un profilo completamente diverso dagli altri. Sono elementi che perturbano l'analisi in quanto orientando da soli un asse alterano le rappresentazioni degli altri profili. Questo caso si è presentato nella matrice *Spettacoli* con i profili del Veneto e di Lirica e Balletti. Considerare elemento illustrativo Lirica e Balletti – perché avrebbe poco senso una matrice attiva che escludesse il Veneto – sposta però il campo dell'indagine che era lo studio delle attività teatrali e musicali in Italia nel 1991. Si vedano le considerazioni fatte al riguardo nella Sezione precedente. Frequente è il caso di profili eccentrici con massa debole che indicano frequenze assolute ridotte

nella matrice $\mathbf{N}$. Questi profili vanno considerati senz'altro come illustrativi.

- nuove variabili che si rendono disponibili in ritardo, *dopo* che un'analisi è già stata effettuata e che per vari motivi non si intende rifare. Per esempio potrebbero rendersi disponibili con grande ritardo i dati relativi alle Rappresentazioni di Prosa Dialettale nelle 20 regioni. Invece di considerare questa colonna come attiva ed analizzare la nuova matrice di ordine 20×9, la si può proiettare come illustrativa sulle mappe fattoriali di TAV. 37 e di TAV. 39 per vedere come la Prosa Dialettale si relaziona alle altre attività teatrali.
- variabili di *natura* diversa da quella delle variabili attive: ad esempio i biglietti venduti nelle 20 regioni per spettacoli Cinematografici. Considerare attiva questa colonna sposterebbe il campo d'indagine da quello delle attività che si svolgono in un teatro a quello degli spettacoli *tout court*. La separazione tra gruppi di variabili ottenute da inchieste e da ricerche di mercato è sempre ben netta. Da un lato le variabili socio-demografiche: sesso, età, professione, ecc. e dall'altro quelle legate a opinioni, aspirazioni o ai consumi. Questi diversi aspetti non vanno mai mescolati nell'analisi e lo studio delle loro interrelazioni fornisce sempre risultati di estremo interesse.
- elementi raggruppati. I profili possono essere riuniti in gruppi tramite i metodi della Sez. X o in base a criteri definiti 'a priori', per esempio accorpando le regioni del Nord, del Centro e del Sud. Il profilo medio ponderato di ogni gruppo può essere proiettato come elemento illustrativo sulle mappe fattoriali.
- elementi ottenuti disaggregando i gruppi. È il caso opposto del precedente. Per esempio, alla matrice *Spettacoli* si possono aggiungere come righe illustrative le vendite di biglietti, per gli 8 tipi di spettacolo, nelle Province italiane. Come esempio, nella mappa di TAV. 45 sono raffigurati i profili delle 9 province siciliane.
- matrici a 3 indici ove di solito il terzo indice è il tempo. Per esempio, se sono disponibili le matrici *Spettacoli* relative agli anni $1991 - 2000$, si può considerare come attiva la matrice 20×8 ottenuta dal *cumulo* delle 10 matrici ed affiancarle come illustrative le 80 colonne delle 10 matrici, per studiare l'evoluzione temporale delle attività teatrali e musicali nel decennio $1991 - 2000$. Considerando invece come illustrative le 200 righe che si ottengono impilando le 10 matrici sotto alla matrice cumulata, si può studiare invece come nelle regioni siano

cambiate le scelte degli spettatori nell'arco del decennio. Questo importante campo d'analisi è trattato nella Sez. 8.9 e nel Capitolo 15.

La proiezione di un profilo illustrativo $\tilde{\mathbf{r}}$ o $\tilde{\mathbf{c}}$ su un asse fattoriale predeterminato si ottiene, come per un profilo attivo, tramite il prodotto scalare e le relazioni di transizione (4.9.2), legando così l'ascissa $\tilde{f}_a$ del profilo di una *riga illustrativa* a quelle g_{ja} dei J profili delle colonne attive, e l'ascissa $\tilde{g}_a$ del profilo di una *colonna illustrativa* a quelle f_{ia} delle I righe attive, e questo per ogni asse $a = 1, 2, \ldots, A$, per cui

$$\begin{aligned} \tilde{f}_a &= \tilde{\mathbf{r}}^T \, \mathbf{D}_{\mathbf{r}}^{-\mathbf{1}} \, \mathbf{v}_a^\star = \frac{1}{\sqrt{\lambda_a}} \, \tilde{\mathbf{r}}^T \, \mathbf{g}_a = \frac{1}{\sqrt{\lambda_a}} \sum_{j=1}^{J} \tilde{r}_j \, g_{ja} \\ \tilde{g}_a &= \tilde{\mathbf{c}}^T \, \mathbf{D}_{\mathbf{c}}^{-\mathbf{1}} \, \mathbf{u}_a^\star = \frac{1}{\sqrt{\lambda_a}} \, \tilde{\mathbf{c}}^T \, \mathbf{f}_a = \frac{1}{\sqrt{\lambda_a}} \sum_{i=1}^{I} \tilde{c}_i \, f_{ia}. \end{aligned} \tag{4.12.1}$$

Questa è la versione del 'principio baricentrico' della Sez. 4.9 per i profili illustrativi: a meno del coefficiente di amplificazione $1/\sqrt{\lambda_a}$, la posizione $\tilde{f}_a$ (o $\tilde{g}_a$) su un asse di un profilo illustrativo $\tilde{\mathbf{r}}$ (o $\tilde{\mathbf{c}}$) è la media delle J coordinate dei profili delle colonne (delle righe) attive, ottenute dall'analisi della matrice attiva $\mathbf{N}$, ponderata con le J componenti $\tilde{r}_j$ (o $\tilde{c}_i$) del profilo illustrativo. Così ad esempio, se la colonna supplementare fosse il numero di biglietti venduti nelle 20 regioni nel 1991 per spettacolo cinematografici, la proiezione del suo profilo su un asse tenderebbe a collocarsi più vicina a quelle regioni con le quote più alte di cinefili. Si noti infine come le (4.12.1) proiettino i profili *singolarmente*, per cui anche una intera matrice illustrativa avrebbe ogni suo profilo proiettato indipendentemente da tutti gli altri. Spesso un profilo illustrativo deve essere ricodificato per poterlo proiettare correttamente su un asse fattoriale. I casi più importanti sono passati in rassegna nella Sez. 8.8.

È conveniente considerare i profili illustrativi come punti dotati di *massa nulla*, e quindi anche di inerzia nulla sia rispetto all'origine che al baricentro dei punti attivi, ed incapaci di 'attirare' verso di sè gli assi fattoriali.

Dato che la massa non interviene, si possono calcolare anche per i profili illustrativi gli scarti relativi dalla media dei profili attivi, ottenendo

$$\tilde{s}_j = \frac{\tilde{r}_j - \overline{r}_j}{\overline{r}_j} \qquad \text{e} \qquad \tilde{s}_i = \frac{\tilde{c}_i - \overline{c}_i}{\overline{c}_i}.$$

Essendo nulla la massa, è nullo anche il contributo relativo CTR_a di

un profilo illustrativo all'inerzia di un asse, mentre non lo sono

$$COS_a^2(\tilde{\mathbf{r}}) \stackrel{\text{def}}{=} \frac{(\tilde{f}_a - 0)^2}{d_D^2(\tilde{\mathbf{r}}, \overline{\mathbf{r}})} = \frac{\tilde{f}_a^2}{\sum_{a=1}^{A} \tilde{f}_a^2} \quad \text{e} \quad COS_a^2(\tilde{\mathbf{c}}) \stackrel{\text{def}}{=} \frac{(\tilde{g}_a - 0)^2}{d_D^2(\tilde{\mathbf{c}}, \overline{\mathbf{c}})} = \frac{\tilde{g}_a^2}{\sum_{a=1}^{A} \tilde{g}_a^2}$$

che permettono di giudicare la qualità della rappresentazione dei profili su un asse. Sommando i COS^2 per due o più assi fino ad $A^* \leq A$, si ottiene la qualità della rappresentazione di un profilo su un piano e in un sottospazio

$$QLT_{A^*}(\tilde{\mathbf{r}}) \stackrel{\text{def}}{=} \sum_{a=1}^{A^*} COS_a^2(\tilde{\mathbf{r}}) \qquad QLT_{A^*}(\tilde{\mathbf{c}}) \stackrel{\text{def}}{=} \sum_{a=1}^{A^*} COS_a^2(\tilde{\mathbf{c}}).$$

A questo proposito occorre tenere presente che un profilo illustrativo può *non* appartenere al sottospazio A-dimensionale che contiene la nuvola dei profili attivi. Per rendersene conto si consideri la matrice di contingenza $\mathbf{N}$ di ordine 5×3 riportata in alto nella TAV. 43 insieme a due colonne illustrative. Al centro della Tavola sono riportate la matrice $\mathbf{C}$ dei profili di $\mathbf{N}$, il profilo medio $\overline{\mathbf{c}}$ delle colonne attive ed i profili $\tilde{\mathbf{c}}_1$ e $\tilde{\mathbf{c}}_2$ delle due colonne illustrative, ottenuti dividendone ogni elemento per i rispettivi totali che sono 3 e 2. L'Analisi delle Corrispondenze di $\mathbf{N}$ fornisce $A = min\,(5, 3) - 1 = 2$ assi fattoriali e le coordinate g_{ja} delle tre colonne attive su di essi che, a loro volta, grazie alle (4.12.1), consentono di calcolare le coordinate di $\tilde{\mathbf{c}}_1$ e $\tilde{\mathbf{c}}_2$.

Il primo profilo illustrativo risulta essere una combinazione lineare dei 3 profili attivi riferiti alla base unitaria di $\Re^5$

$$\tilde{\mathbf{c}}_1 = a \times \mathbf{c}_1 + b \times \mathbf{c}_2 + c \times \mathbf{c}_3$$

$$\begin{pmatrix} 0.333 \\ 0.333 \\ 0.333 \\ 0.000 \\ 0.000 \end{pmatrix} = a \begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} + b \begin{pmatrix} 0.0 \\ 0.5 \\ 0.5 \\ 0.0 \\ 0.0 \end{pmatrix} + c \begin{pmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.5 \\ 0.5 \end{pmatrix}$$

dato che i coefficienti *non* sono tutti nulli in quanto

$$a = 1/3 \qquad b = 2/3 \qquad c = 0.$$

Riferiti al sistema di assi fattoriali, i profili attivi sono per costruzione una combinazione lineare dei due autovettori $\mathbf{u}_1^\star$ e $\mathbf{u}_2^\star$

$$\begin{aligned} \mathbf{c}_1 &= -0.500 \times \mathbf{u}_1^\star + 1.936 \times \mathbf{u}_2^\star \\ \mathbf{c}_2 &= -0.500 \times \mathbf{u}_1^\star - 0.646 \times \mathbf{u}_2^\star \\ \mathbf{c}_3 &= +2.000 \times \mathbf{u}_1^\star + 0.000 \times \mathbf{u}_2^\star \end{aligned}$$

e perciò $\tilde{\mathbf{c}}_1$ è anche una combinzione lineare dei due autovettori

$$\tilde{\mathbf{c}}_1 = \frac{1}{3}\,\mathbf{c}_1 + \frac{2}{3}\,\mathbf{c}_2 = \frac{1}{3}\,(-0.500\,\mathbf{u}_1^\star + 1.936\,\mathbf{u}_2^\star) + \frac{2}{3}\,(-0.500\,\mathbf{u}_1^\star - 0.646\,\mathbf{u}_2^\star)$$
$$= -0.500\,\mathbf{u}_1^\star + 0.215\,\mathbf{u}_2^\star.$$

e giace dunque sul piano fattoriale che questi individuano. Si ha quindi che $QLT_A(\tilde{\mathbf{c}}_1) = 1$.

Invece, il secondo profilo illustrativo $\tilde{\mathbf{c}}_2$ *non* risulta essere una combinazione lineare dei 3 profili attivi, perché $a = b = c = 0$, e quindi neppure dei 2 autovettori $\mathbf{u}_1^\star$ e $\mathbf{u}_2^\star$ che individuano il piano, per cui deve trovarsi in uno spazio di superiore dimensionalità. Infatti, come si vede dalla figura nella TAV. 43, la distanza dal baricentro della proiezione di $\tilde{\mathbf{c}}_2$ sul piano risulta

$$d_D^2(\overline{\mathbf{c}},\, \tilde{g}_1\,\mathbf{u}_1^\star + \tilde{g}_2\,\mathbf{u}_2^\star) = \tilde{g}_1^2 + \tilde{g}_2^2 = (0.750)^2 + (-0.323)^2 = 0.667$$

molto *inferiore* all'effettiva distanza del profilo dal baricentro che va calcolata con la (2.8.7)

$$d_D^2(\overline{\mathbf{c}}, \tilde{\mathbf{c}}_2) = \sum_{i=1}^{I} \frac{(\overline{c}_i - \tilde{c}_i)^2}{\overline{c}_i} = \frac{(0.2-0.0)^2}{0.2} + \frac{(0.3-0.0)^2}{0.3} + \frac{(0.3-0.5)^2}{0.3}$$
$$+ \frac{(0.1-0.5)^2}{0.1} + \frac{(0.1-0.0)^2}{0.1} = 2.333$$

e di conseguenza, come risulta dalla TAV. 42, la qualità della rappresentazione sul piano della effettiva distanza del profilo dal baricentro, è

$$QLT_2(\tilde{\mathbf{c}}_2) = \frac{\tilde{g}_1^2 + \tilde{g}_2^2}{d_D^2(\overline{\mathbf{c}}, \tilde{\mathbf{c}}_2)} = \frac{0.667}{2.333} = 0.286 \;<\; 1.$$

Perciò $\tilde{\mathbf{c}}_2$ è contenuto in uno spazio $\Re^4$, simplesso a 4 vertici di $\Re^5$ e non in $\Re^2$ come i profili attivi. Da tutto ciò emerge che $QLT_A(\tilde{\mathbf{r}})$ e $QLT_A(\tilde{\mathbf{c}})$ sono preziosi indicatori dell'appartenenza, se $QLT_A = 1$, o meno, se $QLT_A < 1$, di un profilo illustrativo allo spazio A-dimensionale in cui è contenuta la nuvola dei profili attivi. La non appartenenza di un profilo illustrativo al sottospazio dei profili attivi si verifica frequentemente quando una delle dimensione della matrice attiva è preponderante sull'altra. Così per la matrice *Spettacoli* di ordine 20×8 è da aspettarsi che il profilo di una riga illustrativa, che deve essere un vettore di ordine 8, sia contenuto nel sottospazio ad $A = 7$ dimensioni ed abbia quindi $QLT_7(\tilde{\mathbf{r}}) = 1$, mentre quello di una colonna illustrativa, un vettore di ordine 20, difficilmente vi apparterrà, per cui $QLT_7(\tilde{\mathbf{c}}) < 1$.

Anche i profili illustrativi possono essere ricostruiti partendo dai fattori, in modo analogo a quanto fatto nella Sez. 4.10 per quelli attivi. Più precisamente risulta, per $j = 1, 2, \ldots, J$ e per $i = 1, 2, \ldots, I$,

$$\tilde{r}_j = \overline{r}_j \left(1 + \sum_{a=1}^{A} \frac{1}{\sqrt{\lambda_a}} \tilde{f}_a \, g_{ja}\right) \quad \tilde{c}_i = \overline{c}_i \left(1 + \sum_{a=1}^{A} \frac{1}{\sqrt{\lambda_a}} \tilde{g}_a \, f_{ia}\right) \qquad (4.12.2)$$

dove $\tilde{r}_j$ e $\tilde{c}_i$ sono le componenti ricostruite del profilo illustrativo riga o colonna, $\overline{r}_j$ e $\overline{c}_i$ sono le masse o componenti del profilo medio dei profili attivi, $\tilde{f}_a$ e $\tilde{g}_a$ sono le ascisse del profilo illustrativo sull'asse di rango a ed infine f_{ia} e g_{ja} le ascisse dei profili attivi $\mathbf{r}_i$ e $\mathbf{c}_j$. La ricostruzione è perfetta quando la somma è spinta fino ad A termini per tener conto di tutti i fattori, *purché* il profilo illustrativo appartenga al sottospazio A-dimensionale dei profili attivi, come $\tilde{\mathbf{c}}_1$ nell'esempio, altrimenti, come per $\tilde{\mathbf{c}}_2$, si riesce a ricostruire soltanto la sua *proiezione* in questo sottospazio.

Le espressioni qui sopra possono interpretarsi come delle formule di regressione. Prendendo in considerazione quella di sinistra, se si indicano con $\tilde{s}_j = (\tilde{r}_j - \overline{r}_j)/\overline{r}_j$ gli scarti relativi dalla media del profilo attivo, dato che $\hat{g}_{ja} = g_{ja}/\sqrt{\lambda_a}$ sono i fattori standard delle colonne attive ed operando analogamente sulla seconda espressione, si ottiene per $j = 1, 2, \ldots, J$ e per $i = 1, 2, \ldots, I$,

$$\tilde{s}_j = \sum_{a=1}^{A} \tilde{f}_a \, \hat{g}_{ja} \qquad \tilde{s}_i = \sum_{a=1}^{A} \tilde{g}_a \, \hat{f}_{ia}$$

Si cerca quindi di spiegare $\tilde{s}_j$ o $\tilde{s}_i$ in funzione dei fattori standard $\hat{g}_{ja}$ o di $\hat{f}_{ia}$. Le coordinate fattoriali del profilo supplementare sono i coefficienti della regressione. L'operazione geometrica di proiezione di un profilo illustrativo su una mappa fattoriale equivale quindi ad una regressione sui fattori generati dall'analisi. L'equivalenza è evidente perché una regressione si può sempre interpretare come una proiezione.

Come esempio di righe illustrative della matrice *Spettacoli*, si possono considerare le vendite di biglietti per attività teatrali e musicali nelle 9 province siciliane. I loro profili si ottengono dividendo ciascun elemento della riga per il totale della riga, mentre i 9 totali, divisi per la loro somma, forniscono le masse dei 9 profili, ossia il peso relativo. Il profilo ottenuto dalla media dei 9 profili, ponderata con le le masse, non è altro che il profilo attivo della Sicilia. Perciò, *se* i 9 profili sono contenuti nel sottospazio a 7 dimensioni dei profili attivi, e questo è il caso perché dalla TAV. X per tutti e 9 i profili risulta $QLT_7(\tilde{\mathbf{r}}_i) = 1$, qui costituiscono una nuvola di punti il

cui baricentro è il punto individuato dal profilo attivo Sicilia. Anche i profili illustrativi si possono filtrare, di solito ponendo al loro valore di $COS^2_{(a,b)}$ una soglia intorno a 0.500 o 0.400, ma questo valore, oltre a variare da caso a caso in base alle dimensioni della matrice attiva, viene a dipendere ora anche dall'appartenenza o meno del profilo illustrativo allo spazio di quelli attivi.

La TAV. 44 riporta le coordinate fattoriali e la qualità della rappresentazione nei piani $(1, 2)$ e $(2, 3)$ e nello spazio ad $A = 7$ dimensioni dei 9 profili illustrativi. L'interpretazione delle prossimità delle proiezioni per i profili illustrativi si fa con le stesse regole dei profili attivi. Soltanto 5 delle 9 province sono ben rappresantate sulla mappa (1, 2), come si vede nella TAV. 45. I loro profili sono prossimi a quello della Prosa, perché le loro prime componenti hanno i valori più elevati, mentre per Siracusa e Ragusa non è trascurabile la quota dei Concerti di Musica Leggera, il che spiega la loro posizione meno eccentrica. In modo del tutto analogo, anche i profili delle altre province italiane possono essere proiettati sulle mappe precedentemente ottenute, rendendo così l'analisi un processo iterativo di confronto tra regioni ed una più dettagliata visione della situazione dello spettacolo in Italia.

4.13 - Mappe asimmetriche

Le coordinate dei punti rappresentati sulle mappe fattoriali della Sez. 4.11 erano i fattori *principali*, e quindi (f_{ia}, f_{ib}) per i profili delle righe e (g_{ja}, g_{jb}) per le colonne. Questo tipo di mappa, sovrapposizione di due mappe ottenute da spazi distinti, è detto *simmetrico* perché i fattori sono dello stesso tipo per entrambe le nuvole ed è il più frequentemente utilizzato per evidenziare le configurazioni dei profili, anche se, come si è visto con l'esempio, non sempre permette di riconoscere immediatamente le associazioni tra profili appartenenti a nuvole diverse.

Queste associazioni possono essere meglio osservate graficamente costruendo delle mappe *asimmetriche*, ove le coordinate dei profili di una nuvola sono i fattori *principali* e quelli dell'altra i fattori *standard*. In tutto sono tre i tipi di mappa che si possono ottenere, come mostra questo specchietto

	Coordinate dei Profili	
Mappa	$\mathbf{r}_i$	$\mathbf{c}_j$
Simmetrica	$(f_{ia},\, f_{ib})$	$(g_{ja},\, g_{jb})$
Asimmetrica	$(f_{ia},\, f_{ib})$	$(\hat{g}_{ja},\, \hat{g}_{jb})$
Asimmetrica	$(\hat{f}_{ia},\, \hat{f}_{ib})$	$(g_{ja},\, g_{jb})$

Le mappe asimmetriche hanno caratteristiche peculiari che le rendono *sostanzialmente* diverse dalle mappe simmetriche e che possono essere meglio illustrate riprendendo la matrice d'esempio *Spettacoli-3*, di ordine 3×8, di TAV. 14. Nella Sez. 2.7 si è visto che gli 8 profili delle colonne - i tipi di spettacolo - sono contenuti in una regione triangolare equilatera di un sottospazio bidimensionale, detta simplesso, i cui 3 vertici sono i punti unitari sugli assi individuati dalla terna di vettori unità $\mathbf{e}_1$, $\mathbf{e}_2$ ed $\mathbf{e}_3$ con origine in $\mathbf{0}_3$. I *vertici* del simplesso stabiliscono quindi i 'confini ultimi' della regione di confinamento degli 8 profili, ma il fatto rilevante è che profili e vertici appartengono al *medesimo* sottospazio bidimensionale, come appare chiaramente nelle TAV. 15 e 21. L'Analisi delle Corrispondenze produce una nuova base di riferimento per il sottospazio, costituita da due autovettori aventi origine nel baricentro, che individuano gli assi fattoriali ai quali i profili vengono riferiti. Anche i tre vertici, possono venir riferiti ai nuovi assi, sfruttando la tecnica di proiezione dei profili illustrativi, illustrata nella Sez. 4.12. Nel caso dell'esempio, la matrice dei profili delle colonne, tratta dalla TAV. 14, ed i tre profili unità sono

$$\mathbf{C} = \begin{array}{c} \begin{matrix} \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 & & \mathbf{c}_8 \end{matrix} \\ \begin{pmatrix} 0.524 & 0.610 & 0.527 & \dots & 0.622 \\ 0.245 & 0.220 & 0.234 & \dots & 0.156 \\ 0.231 & 0.171 & 0.239 & \dots & 0.222 \end{pmatrix} \end{array} \quad \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Per proiettare un profilo colonna supplementare, ad esempio $\tilde{\mathbf{c}} = \mathbf{e}_1 = (1\ 0\ 0)^T$ sul piano fattoriale (1, 2), si sfrutta la (4.12.1) e si utilizzano le coordinate principali dei 3 profili delle righe sul primo asse, ricavate nella Sez. 4.9 e riprese nella parte centrale del prospetto qui sotto. Si ottiene così

la coordinata del primo vertice del simplesso sul primo asse fattoriale

$$\tilde{g}_1(\tilde{\mathbf{c}} = \mathbf{e}_1) = \frac{1}{\sqrt{\lambda_1}} \sum_{i=1}^{3} \tilde{c}_i \, f_{i1} = \frac{1}{\sqrt{\lambda_1}} \sum_{i=1}^{3} (\mathbf{e}_1)_i \, f_{i1}$$
$$= \frac{1}{\sqrt{0.005}} (1 \times -0.063 + 0 \times 0.078 + 0 \times 0.074) = -0.907$$

dove $(\mathbf{e}_1)_i$ indica la i^{ma} componente del vettore $\mathbf{e}_1$. Un calcolo analogo fornisce la coordinata sul secondo asse $\tilde{g}_2(\tilde{\mathbf{c}} = \mathbf{e}_1) = -0.016$. Queste due coordinate indicano la posizione del primo vertice $\mathbf{e}_1$ del simplesso sul piano fattoriale. Le coordinate dei tre vertici sono elencate nella parte sinistra del prospetto.

Vertici	$\tilde{\mathbf{g}}_1$	$\tilde{\mathbf{g}}_2$	*Righe*	$\mathbf{f}_1$	$\mathbf{f}_2$	$\hat{\mathbf{f}}_1$	$\hat{\mathbf{f}}_2$
$\mathbf{e}_1$	-0.907	-0.016	Nord	-0.063	-0.001	-0.907	-0.016
$\mathbf{e}_2$	$+1.128$	-1.403	Centro	$+0.078$	-0.066	$+1.128$	-1.403
$\mathbf{e}_3$	$+1.074$	$+1.576$	Sud	$+0.074$	$+0.075$	$+1.074$	$+1.576$

I vertici indicano l'estrema situazione osservabile, il tipo di profilo più polarizzato. Così $\mathbf{e}_1$ è il profilo colonna ove la quota di spettatori è concentrata esclusivamente nel Nord Italia, $\mathbf{e}_2$ nel Centro ed $\mathbf{e}_3$ nel Sud e nelle Isole. È interessante vedere come il primo asse riveli il contrasto tra il Nord ($\tilde{g}_{11} = -0.907$) da una parte e il Centro ($\tilde{g}_{21} = +1.128$) e il Sud ($\tilde{g}_{31} = +1.074$) dall'altra, mentre il secondo asse mette in luce le diversità tra Centro e Sud in termini di spettacoli teatrali e musicali.

La (4.12.1) appena utilizzata per proiettare i profili illustrativi, può essere espressa in termini di fattori standard, diventando

$$\tilde{f}_a = \sum_{j=1}^{J} \tilde{r}_j \, \hat{g}_{ja} \qquad\qquad \tilde{g}_a = \sum_{i=1}^{I} \tilde{c}_i \, \hat{f}_{ia} \tag{4.13.1}$$

per cui la coordinata del primo vertice può esprimersi come

$$\tilde{g}_1(\tilde{\mathbf{c}} = \mathbf{e}_1) = \sum_{i=1}^{3} \tilde{c}_i \, \hat{f}_{i1} = \sum_{i=1}^{3} (\mathbf{e}_1)_i \, \hat{f}_{i1} = (1 \times \hat{f}_{11} + 0 \times \hat{f}_{21} + 0 \times \hat{f}_{31}) = \hat{f}_{11}.$$

Questa relazione è di grande interesse perché mostra come la coordinata del primo vertice coincida con la coordinata *standard* del primo profilo $\mathbf{r}_1$ sul primo asse. Un calcolo analogo per gli altri vertici e per l'altro asse, conferma che le coordinate dei 3 vertici coincidono con quelle standard dei

profili delle 3 righe, ottenute nella Sez. 4.8 e riportate nella parte destra del prospetto qui sopra.

Questi risultati autorizzano a denominare *vertici delle righe* le righe rappresentate in coordinate standard e ad indicarli con le modalità Nord, Centro e Sud, come si è fatto nella mappa di TAV. 46 che è la semplice trasposizione su un piano della configurazione della nuvola dei profili delle colonne di TAV. 21, ove era mostrata in un riferimento tridimensionale. In entrambe le rappresentazioni, per esempio, il vertice $\mathbf{e}_1$, o Nord è sulla parte negativa del primo asse. L'interpretazione della mappa si basa sulle relazioni di transizione (4.9.4)

$$f_{ia} = \sum_{j=1}^{J} r_{ij}\, \hat{g}_{ja} \qquad\qquad g_{ja} = \sum_{i=1}^{I} c_{ij}\, \hat{f}_{ia}. \tag{4.13.2}$$

secondo le quali la posizione di ogni profilo colonna è *esattamente* la media, o baricentro, dei vertici delle righe ponderata con le componenti del profilo e tende quindi a disporsi sulla mappa verso quel vertice che corrisponde alla componente c_{ij} più alta. Ad esempio, nella mappa di TAV. 46 il profilo colonna di Burattini e Marionette $\mathbf{c}_7 = (0.010 \;\; 0.018 \;\; 0.011)^T$ è più vicino al vertice Centro che non agli altri due, mentre quello di Operette $\mathbf{c}_4 = (0.636 \;\; 0.152 \;\; 0.212)^T$ si trova più vicino al vertice Nord.

Dato che vertici e profili appartengono al medesimo spazio, è lecito calcolarne le distanze. Fissato quindi un vertice $\mathbf{e}_i$, per ogni profilo colonna $\mathbf{c}_j$, ove $j = 1, 2, \ldots, 8$, grazie alla (4.13.2) risulta

$$\begin{aligned} d^2_D(\mathbf{e}_i, \mathbf{c}_j) &= \sum_{a=1}^{2} \left(\hat{f}_{ia} - g_{ja}\right)^2 = \sum_{a=1}^{2} \hat{f}^2_{ia} - 2 \sum_{a=1}^{2} \hat{f}_{ia}\, g_{ja} + \sum_{a=1}^{2} g^2_{ja} \\ &= d^2_D(\mathbf{e}_i, \overline{\mathbf{c}}) - 2 \sum_{a=1}^{2} \hat{f}_{ia}\, g_{ja} + d^2_D(\mathbf{c}_j, \overline{\mathbf{c}}). \end{aligned}$$

Nella Sez. 4.10 si è visto che le formule di ricostruzione dei profili possono esprimersi in termini di scarto relativo dalla quota media come nella (4.10.2) che espressa tramite i fattori standard diviene

$$s_{ij} = \sum_{a=1}^{A} \frac{1}{\sqrt{\lambda_a}}\, f_{ia}\, g_{ja} = \sum_{a=1}^{A} \hat{f}_{ia}\, g_{ja} = \sum_{a=1}^{A} f_{ia}\, \hat{g}_{ja}. \tag{4.13.3}$$

Tenendo conto che nel caso dell'esempio $A = 2$, si ottiene l'importante relazione

$$d^2_D(\mathbf{e}_i, \mathbf{c}_j) = d^2_D(\mathbf{e}_i, \overline{\mathbf{c}}) - 2\, s_{ij} + d^2_D(\mathbf{c}_j, \overline{\mathbf{c}}). \tag{4.13.4}$$

Al secondo membro le due distanze sono fisse perché fissi sono profili e vertici rispetto al baricentro, per cui c'è proporzionalità tra gli scarti relativi e le distanze dei profili da un vertice

$$d_D^2(\mathbf{e}_i, \mathbf{c}_j) = -2\, s_{ij} + cost_i \tag{4.13.5}$$

dove $cost_i$ è una costante pari alla somma delle due distanze e che dipende perciò dal vertice considerato. La TAV. 47 riporta il grafico delle distanze tra gli 8 profili delle colonne ed i 3 vertici delle righe, ottenute sostituendo nella (4.13.5) i valori di s_{ij} della matrice *Spettacoli-3*, calcolati nella Sez. 4.10. Ciascun punto è indicato dalla coppia di indici $i\,j$ delle modalità interessate. I tre gruppi di punti, uno per ciascun vertice, si allineano con la medesima pendenza negativa [1]. Nella parte sinistra del grafico, ove $s_{ij} \ll 0$, si trovano le coppie di modalità distanti, come se tra esse vi fosse *repulsione*, rivelando che la regione i è sottorappresentata, rispetto alla quota media, nel profilo $\mathbf{c}_j$ dello spettacolo j, o, in altri termini, che $c_{ij} \ll \overline{c}_i$. Nella parte destra invece, ove $s_{ij} \gg 0$, tra le modalità vi è *attrazione* per cui profilo e vertice sono vicini indicando che la regione i è sovrarappresentata nel profilo $\mathbf{c}_j$. Infine, nella fascia centrale ove $s_{ij} \simeq 0$ la regione i è rappresentata nello spettacolo j a un livello non lontano da quello medio $\overline{c}_i$: è la zona d'*indifferenza*. Si può quindi concludere che la mappa asimmetrica mette sostanzialmente in luce gli scarti relativi da una situazione di indifferenza o di completa omogeneità, definita nella Sez. 1.10, che si verifica quando i profili non si distinguono dal loro profilo medio.

La (4.13.3) ha anche una interessante interpretazione geometrica. Si consideri sulla mappa asimmetrica di TAV. 46 il triangolo avente per vertici il baricentro $\overline{\mathbf{c}}$, un profilo $\mathbf{c}_j$ e il vertice $\mathbf{e}_i$. Ora, la geometria insegna che per *ogni* triangolo con lati a, b e c, il quadrato di un lato è eguale alla somma dei quadrati degli altri due, diminuita di due volte il prodotto di questi e del coseno dell'angolo che formano, per cui, ad esempio

$$a^2 = b^2 + c^2 - 2\, b\, c\, \cos\theta_{bc}.$$

---

[1] Le distanze si potrebbero graficare anche in funzione degli elementi c_{ij} dei profili $\mathbf{c}_j$, ma in tal caso le pendenze dei tre gruppi risulterebbero diverse perchè il coefficiente viene a dipendere dalla massa del vertice: maggiore la massa, minore la pendenza

$$d_D^2(\mathbf{e}_i, \mathbf{c}_j) = -2\,\frac{c_{ij} - \overline{c}_i}{\overline{c}_i} + cost_i = -\frac{2}{\overline{c}_i}\, c_{ij} + cost_i + 2.$$

Per il triangolo preso in considerazione, si ha quindi

$$d_D^2(\mathbf{e}_i, \mathbf{c}_j) = d_D^2(\mathbf{e}_i, \overline{\mathbf{c}}) + d_D^2(\mathbf{c}_j, \overline{\mathbf{c}}) - 2\, d_D(\mathbf{e}_i, \overline{\mathbf{c}})\, d_D(\mathbf{c}_j, \overline{\mathbf{c}})\, \cos\theta$$

dove θ è l'angolo al vertice $\overline{\mathbf{c}}$. Confrontando questa espressione con la (4.13.4) si vede subito che

$$s_{ij} = d_D(\mathbf{e}_i, \overline{\mathbf{c}})\, d_D(\mathbf{c}_j, \overline{\mathbf{c}})\, \cos\theta.$$

e quindi s_{ij} è dipende dall'angolo tra profilo e vertice. Di consequenza, quando i profili sono ben rappresentati sulla mappa asimmetrica, se l'angolo risulta inferiore a quello retto allora $s_{ij} > 0$ e la modalità i della prima variabile è rappresentata in eccesso nel profilo $\mathbf{c}_j$, se l'angolo è retto $s_{ij} = 0$ e vi è totale indifferenza tra le modalità i e j, mentre, se l'angolo è superiore a quello retto, è sottorappresentata.

Invece della mappa asimmetrica fin qui utilizzata, si può costruire quella in cui i profili sono le righe, in coordinate principali, ed i vertici le colonne, in coordinate standard. Valgono per questa espressioni analoghe a quelle appena trovate. Così la coordinata fattoriale di un vertice del simplesso di $\Re^J$ coincide con la coordinata standard della colonna corrispondente

$$\tilde{f}_a(\tilde{\mathbf{r}} = \mathbf{e}_j) = \hat{g}_{ja}$$

e la distanza di un profilo da un vertice colonna è

$$d_D^2(\mathbf{r}_i, \mathbf{e}_j) = -2\, s_{ij} + cost_j.$$

La scelta di una mappa o dell'altra è guidata dal tipo e dalla natura delle due variabili. Nel caso della matrice *Spettacoli* di ordine 20×8 di TAV. 2, sembra ragionevole costruire la mappa asimmetrica di TAV. 48 rappresentando le righe, le regioni, come profili e le colonne, i tipi di spettacolo, come vertici. Per garantire una adeguata rappresentazione delle distanze reali, vertici e profili sono stati filtrati come per la mappa simmetrica di TAV. 37, la sua equivalente simmetrica, mentre gli assi sono stati ruotati di 90 gradi, per poter mantenere su entrambi la stessa unità di scala. La posizione dei profili delle righe - le regioni - sono le stesse su entrambe le mappe, mentre i tre vertici - Prosa, Musica Leggera e Lirica - appaiono più 'sparpagliati' sulla mappa asimmetrica, essendo stati dilatati di un fattore $1/\sqrt{\lambda_1} = 5.20$ sul primo asse e $1/\sqrt{\lambda_2} = 6.30$ sul secondo.

Le mappe asimmetriche, pur presentando sostanziali vantaggi, non sempre sono utilizzabili perché quando le inerzie delle proiezioni sui due assi della mappa sono piccole - segnale questo di scarsa associazione tra le due

nuvole - i profili tendono ad accalcarsi intorno al profilo medio, lontani dai vertici, rendendo impercettibili le differenze tra distanze. In questi casi conviene utilizzare mappe simmetriche sulle quali le due nuvole di profili occupano approssimativamente lo stesso spazio.

4.14 - Come vengono calcolati inerzie e dei fattori

L'obiettivo finale dell'Analisi delle Corrispondenze è il calcolo delle inerzie sugli assi e dei fattori dei profili. Dal punto di vista matematico si tratta di ricavare gli autovalori $\mathbf{D}_\lambda$ e gli autovettori $\mathbf{V}^\star$ dalla matrice $(\mathbf{R} - \overline{\mathbf{R}})^T\,\mathbf{C}$ di ordine $I \times I$ e gli autovettori $\mathbf{U}^\star$ da $(\mathbf{C} - \overline{\mathbf{C}})\,\mathbf{R}^T$ di ordine $J \times J$. Entrambe le matrici sono quadrate, ma non simmetriche e soddisfano le condizioni

$$\mathbf{V}^{\star T}\,(\mathbf{R} - \overline{\mathbf{R}})^T\,\mathbf{C}\,\mathbf{V}^\star = \mathbf{D}_\lambda \qquad \text{con} \qquad \mathbf{V}^{\star T}\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,\mathbf{V}^\star = \mathbf{I}$$
$$\mathbf{U}^{\star T}\,(\mathbf{C} - \overline{\mathbf{C}})\,\mathbf{R}^T\,\mathbf{U}^\star = \mathbf{D}_\lambda \qquad \text{con} \qquad \mathbf{U}^{\star T}\,\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\,\mathbf{U}^\star = \mathbf{I}.$$

Il tempo di calcolo può ridursi notevolmente grazie ad alcune considerazioni, elencate qui di seguito.

Intanto, nella Sez. 4.8 e nella Sez. 3.14 si è visto che le equazioni

$$\mathbf{V}^T\,\mathbf{R}^T\,\mathbf{C}\,\mathbf{V} = \mathbf{D}_\lambda \qquad \text{con} \qquad \mathbf{V}^T\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,\mathbf{V} = \mathbf{I}$$
$$\mathbf{U}^T\,\mathbf{C}\,\mathbf{R}^T\,\mathbf{U} = \mathbf{D}_\lambda \qquad \text{con} \qquad \mathbf{U}^T\,\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\,\mathbf{U} = \mathbf{I}.$$

che coinvolgono sempre matrici quadrate, ma più semplici da calcolare, forniscono gli stessi autovalori non banali e gli stessi autovettori delle precedenti.

Il tempo di calcolo può essere poi dimezzato grazie alle relazioni di transizione (4.9.3) che consentono di ottenere gli autovettori delle righe da quelli delle colonne e viceversa, per cui basta diagonalizzare una sola delle due matrici: quella di dimensioni più ridotte e quindi $\mathbf{R}^T\,\mathbf{C}$ che è di ordine $J \times J$ quando il numero di righe supera quello delle colonne o $\mathbf{C}\,\mathbf{R}^T$ di ordine $I \times I$ nell'altro caso. Con l'ulteriore vantaggio che gli orientamenti degli autovettori di pari rango delle due nuvole concordano automaticamente.

Un altro sostanziale risparmio di tempo ed un incremento della precisione numerica dei risultati si ottiene simmetrizzando la matrice da diagonalizzare come mostrato nella Sez. B.5 dell'Appendice B, perché le routines di diagonalizzazione per matrici simmetriche sono molto più veloci e precise di quelle per matrici quadrate generiche. La sequenza delle trasformazioni e delle operazioni di calcolo è illustrata nella TAV. 49.

Per quanto riguarda gli algoritmi di diagonalizzazione, il lettore può consultare il testo classico di Golub e Van Loan (1996), citato nella bibliografia

del Capitolo 3. Il metodo di Lagrange, che ha il vantaggio di chiarire in modo didatticamente efficace come si perviene agli autovalori e agli autovettori di una matrice di piccole dimensioni, non viene più utilizzato da tempo.

Un cenno infine ai programmi per l'Analisi delle Corrispondenze. Tutti e tre i principali ambienti d'analisi statistica, SPSS, BMDP e SAS [1], permettono di effettuare sia l'Analisi delle Corrispondenze semplici, o binarie, presentata in questi primi capitoli, sia l'Analisi delle Corrispondenze Multiple che verrà illustrata nel prossimo capitolo. Escludendo per ora quest'ultima, occorre dire che il modulo ANACOR di SPSS ha due gravi svantaggi: non è provvisto di un'opzione per trattare i profili illustrativi ed è stranamente incapace di produrre mappe simmetriche. Il modulo CA di BMDP ha invece la limitazione di non produrre mappe asimmetriche. La procedura CORRESP di SAS è completa, nel senso che permette di ottenere entrambi i tipi di mappa. I tre programmi forniscono comunque tutti gli indicatori necessari per la corretta lettura di una mappa e sono disponibili per le piattaforme hardware più diffuse. Altri prodotti sono disponibili per applicazioni in campi specifici: biologia, archeologia, scienze sociali, ecc. Una comparazione dei prodotti software reperibili sul mercato, con i loro punti di forza e di debolezza e gli indirizzi dei fornitori si può trovare in Greenacre (1993), citato nella bibliografia della Sez. 4.17.

4.15 - Cenni storici

L'Analisi delle Corrispondenze, nella forma di approccio geometrico all'analisi multidimensionale dei dati che è stata esposta in questi capitoli, ha avuto origine in Francia negli anni '60 grazie al lavoro di Jean Paul Benzécri e dei suoi collaboratori. Nelle sue linee generali il metodo non era del tutto nuovo, ma con Benzécri viene posto in un coerente e rigoroso schema geometrico e matematico. La scoperta delle relazioni di transizione e l'introduzione dei contributi relativi e del concetto di qualità della rappresentazione per guidare l'interpretazione delle mappe grafiche, hanno reso l'Analisi delle Corrispondenze un metodo generale, o quasi, per analizzare dati multidimensionali, particolarmente quelli ottenuti da indagini e ricerche di mercato [2]. Il metodo si diffondeva rapidamente in Francia durante gli anni '70, diven-

---

[1] Dettagli e costi dei programmi si possono trovare nei siti `www.spss.com` e `www.sas.com`. BMDP è stato recentemente acquisito da SPSS.

[2] Questa importante estensione, nota come Analisi delle Corrispondenze Multiple, ò presentata in dettaglio nel Capitolo 5.

tando praticamente *il* metodo per l'analisi di dati multidimensionali. Occorre però arrivare agli anni '80 perché esso inizi a diffondersi anche all'esterno. In Olanda viene incorporato in un sistema d'analisi rivolto alle scienze sociali e verso la fine del decennio viene incluso nei grandi sistemi statistici anglosassoni: SPSS, SAS, BMDP, ecc.

4.16 - Riepilogo

Giunti a questo punto, è già possibile fare un bilancio dei principali punti di forza e di debolezza dell'Analisi delle Corrispondenze. Tra i primi va messo senz'altro al primo posto il fatto che il metodo non si limita a rivelare l'esistenza di un legame tra due variabili categoriche, ma che, evidenziando graficamente le associazioni tra modalità delle due variabili, mostra *perché* questa sussista. Tenendo conto *congiuntamente* dei legami multipli di ogni modalità permette poi di rivelare delle associazioni che potrebbero non apparire anche con una serie di ripetuti confronti a coppie tra profili. In più, le regole d'interpretazione delle mappe simmetriche sono le medesime per i profili delle righe e delle colonne il che facilita l'analisi e l'interpretazione. In questo è unico tra i metodi d'analisi statistica multidimensionale che portano alla rappresentazione grafica di una configurazione geometrica. La possibilità poi di incorporare nell'analisi informazioni aggiuntive lo rende, nelle mani di un analista esperto, uno strumento estremamente duttile. Infine, come verrà esemplificato ampiamente nella seconda parte, è uno strumento molto "portabile" perché può applicarsi a diversi tipi di matrici di dati. L'unica limitazione è la non-negatività degli elementi, anche se da un punto di vista puramente matematico, è sufficiente la positività dei totali marginali della matrice.

L'Analisi delle Corrispondenze ha anche alcuni punti di debolezza. Il più evidente è che nelle mappe simmetriche le prossimità tra proiezioni di profili appartenenti a nuvole diverse non possono interpretarsi direttamente, ma soltanto tramite le relazioni di transizione, il che spesso costringe a ricercare la conferma delle prossimità in piani fattoriali di rango superiore o direttamente nella matrice dei profili. Inoltre capita sovente che le mappe asimmetriche si rivelino di scarso aiuto nell'evidenziare le prossimità tra vertici e profili, a causa dell'affollarsi di questi intorno al loro baricentro. Esiste infine un certo margine di soggettività nell'interpretazione delle mappe, la cui ampiezza è però inversamente legata all'esperienza dell'analista e al suo grado di conoscenza del metodo.

Ulteriori dettagli sugli aspetti applicativi dell'Analisi delle Corrispondenze di matrici di contingenza sono presentati nel Capitolo 8, mentre nel prossimo verrà presentata la sua prima e più importante estensione.

4.17 - Bibliografia essenziale

Alcuni testi sull'Analisi delle Corrispondenze che il lettore può consultare:

Jean P. Benzécri (1992). *Correspondence Analysis Handbook.* Marcel Dekker. 665 pg. ISBN 0-8247-8437-5. Quest'opera monumentale - e fondamentale - è la traduzione in lingua inglese dell'edizione francese del 1980, scritta dal principale autore del metodo. Rispecchia fedelmente il punto di vista e le idee dell'autore sull' analisi dei dati e sulla Statistica, e presenta le basi geometriche e matematiche del metodo con svariati e dettagliati esempi di applicazioni nella tassonomia, in sociologia, in linguistica e medicina.

Il testo più citato e che completa, per dire, quello precedente è

Michael J. Greenacre (1984). *Theory and Applications of Correspondence Analysis.* Academic Press. 365 pg. ISBN 0-12-299050-1. Qui è presentata anche l'Analisi delle Corrispondenze Multiple ed è affrontato per la prima volta il problema della stabilità delle configurazioni di profili, temi che verranno presentati in dettaglio nei prossimi due Capitoli.

Chi fosse maggiormente interessato agli aspetti applicativi e ad una descrizione del metodo più verbale e grafica che matematica, può consultare

Michael J. Greenacre (1993). *Correspondence Analysis in Practice.* Academic Press. 195 pg. ISBN 0-12-299052-8. Il testo è corredato da numerose tavole e mappe.

Altro testo d'interesse è

Michel Jambu (1991). *Exploratory and Multivariate Data Analysis.* Academic Press Inc. 475 pg. ISBN 0-12-380090. Questa traduzione dal francese parte dalla Statistica descrittiva per arrivare a presentare in dettaglio i metodi fattoriali e di raggruppamento. Purtroppo non è esente da molteplici errori di stampa nel testo, nelle formule e nelle mappe. Inoltre, la terminologia non standard adottata dall'autore ne consiglia la lettura a chi è già provvisto di buone basi.

Una preziosa 'palestra' per tutti coloro che si interessano di Analisi Multidimensionale è la rivista trimestrale

Les Cahiers de l'Analyse des Donnés. Ideata nel 1976 da J. P. Benzécri, che ne mantiene ancora la direzione scientifica, è dedicata in gran parte

alle applicazioni in quasi tutti i campi scientifici, presentando di queste le mappe e le interpretazioni e sovente anche le matrici dei dati. È stampata in Francia dall'editore Dunod.

Per approfondire l'impiego dei profili illustrativi nell'Analisi delle Corrispondenze è utile consultare

Pierre Cazes (1982) Note sur les éléments supplémentaires en Analyse des Correspondances. **I**. Pratique et utilisation. *Les Cahiers de l'Analyse des donnés* Vol. **VII**, n^o 1, pag. 9 - 23. È il primo di due importanti articoli sull'impiego dei profili illustrativi. Il secondo è citato nella Bibliografia del Cap. 8, alla Sez. 8.10.

PARTE PRIMA: IL METODO

CAPITOLO 5: Analisi delle Corrispondenze Multiple

Sommario

La metodologia dell'Analisi delle Corrispondenze può essere estesa ai profili ottenuti da una tabella di indicatori del tipo *individui* $\times$ *modalità*, nella quale, per ogni individuo statistico, le modalità di più variabili categoriche sono indicate in forma disgiuntiva completa. Il metodo è particolarmente adatto all'analisi delle risposte raccolte nei sondaggi ed ha proprietà matematiche specifiche che comportano particolari adattamenti alle regole d'interpretazione dei risultati. L'Analisi delle Corrispondenze che osserva la nuova interpretazione costituisce un metodo del tutto nuovo che prende il nome di Analisi delle Corrispondenze Multiple.

Dalla lettura di questo capitolo, il lettore verrà a conoscere, tra l'altro,

- la definizione di Analisi delle Corrispondenze Multiple;
- le analogie e le differenze tra Corrispondenze Multiple e Corrispondenze semplici;
- come è strutturata una tabella di indicatori delle modalità con codifica disgiuntiva completa;
- come viene ottenuta una matrice di Burt;
- come siano legate le analisi dei profili ricavati dalla tabella di indicatori e dalla matrice di Burt;
- l'impiego delle Corrispondenze Multiple nell'analisi dei questionari: sondaggi d'opinione, ricerche di mercato, ecc.;
- il ruolo cruciale dei profili illustrativi.

CAPITOLO 5

5.1 - Introduzione

Al termine 'Analisi delle Corrispondenze' si fa abitualmente seguire l'aggettivo *semplici* quando si intende riferirsi al metodo presentato nei capitoli precedenti, per distinguerlo dalla sua estensione più immediata e importante che va sotto il nome di 'Analisi delle Corrispondenze Multiple' (ACM), presentata in questo Capitolo. L'aggettivo 'multiple' si riferisce al fatto che *più di due* variabili categoriche possono venire analizzate congiuntamente. È il metodo più fecondo e di successo perché consente, tra l'altro, l'analisi multidimensionale dei dati di inchieste rilevati mediante sondaggio.

Il legame tra i due tipi di analisi è molto stretto. Per esempio, i risultati ottenuti analizzando con le Corrispondenze semplici la matrice di contingenza *Spettacoli-3* di ordine 3×8 della TAV. 14, sono equivalenti [1] all'Analisi delle Corrispondenze Multiple dei profili desunti da una tabella di ordine 576×11. Le 576 righe corrispondono ai biglietti venduti e le 11 colonne alle modalità delle due variabili categoriche, 3 per l'"Area Geografica' più 8 per il 'Tipo di Spettacolo'. Di ogni biglietto è indicato in quale Area è stato acquistato e per quale tipo di spettacolo. Gli indicatori delle modalità sono codificati in forma disgiuntiva completa, un tipo di codifica che verrà introdotto nella Sez. 5.6. L'equivalenza dei risultati induce a generalizzare [2] l'analisi estendendola al caso in cui *più* di due variabili categoriche vengano rilevate congiuntamente sul medesimo insieme.

Anche le variabili numeriche possono venire incluse nell'analisi, previa trasformazione in variabili categoriche mediante suddivisione in classi

---

[1] Si veda ad esempio la Sez. 1.4.6 in L. Lebart *et al.* (1995), op. cit. nella Sez. 5.24 e Greenacre (1984), pag. 130, op. cit. nella Sez. 4.17.

[2] L'estensione è legittimata anche dal fatto che l'ACM è un caso particolare dell'Analisi Canonica Generalizzata con più di due gruppi di variabili. Si veda la Sez. 3.1.3 in Lebart *et al.* (1995), op. cit. nella Sez. 5.24, e Bouroche e Saporta (1980), *L'Analyse des Donnèes*, Presses Universitaires de France, alle pagg. 80-91.

dell'intervallo di variazione dei loro valori e successiva codifica degli indicatori delle classi d'appartenenza in forma disgiuntiva completa, come verrà mostrato nella Sez 5.16. La capacità di analizzare variabili di tipo eterogeneo rende estremamente ampio il campo delle potenziali applicazioni dell'ACM.

5.2 - Notazioni

L'applicazione più frequente dell'ACM riguarda l'analisi delle risposte ottenute in una indagine effettuata tramite sondaggio con questionario *chiuso*, in cui cioè le modalità di risposta alle singole domande sono già predisposte e mutuamente esclusive: l'intervistato deve limitarsi a indicarne una [1].

La terminologia dell'ACM differisce da quella dei sondaggi, come mostra lo specchietto qui sotto che fa riferimento al caso di un sondaggio con I intervistati ai quali sono poste delle domande. Di queste, le Q che vengono considerate attive nell'analisi hanno complessivamente J modalità di risposta.

Sondaggio	*A. C. M.*	*Simbolo*	*Totale*
intervistato / questionario	individuo	i	I
domanda / quesito	variabile	q	Q
risposta possibile	modalità	j	J

Ogni domanda del questionario diventa una variabile per l'ACM, mentre per 'individuo' si intende qui un *individuo statistico* che può essere un intervistato, e quindi un essere umano nel caso di sondaggi, ma che in particolari applicazioni può essere un animale, una pianta, un prodotto, un evento, ecc.

Come mostra lo specchietto, anche in questo Capitolo i indicherà la riga generica della matrice da analizzare e I il numero di righe complessivo. Con Q verrà indicato il numero complessivo di variabili categoriche *attive* prese in esame e con q la generica variabile attiva avente J_q modalità esclusive. Con j verrà indicata una colonna attiva e quindi una qualunque modalità, il cui numero complessivo J è la somma delle modalità che ha

[1] L'Analisi Multidimensionale dei questionari *aperti* con risposte libere è trattata in L. Lebart e A. Salem (1994), *Statistique textuelle* Dunod ed., Paris. Un suo compendio in italiano si trova in S. Bolasco (1999), citato nella bibliografia al termine di questo Capitolo.

ogni variabile

$$J = \sum_{q=1}^{Q} J_q. \tag{5.2.1}$$

5.3 - Un esempio: l'ascolto radiofonico

Commissionata da alcune radio locali, nel 1995 è stata condotta una estesa indagine telefonica sull'ascolto delle trasmissioni radiofoniche in Emilia Romagna. Lo scopo dell'indagine era quello di 'conoscere meglio' gli ascoltatori. Si voleva sopratutto indagare come e da chi le trasmissioni venivano ascoltate nel corso della giornata e quali erano gli atteggiamenti degli ascoltatori nei riguardi delle interruzioni pubblicitarie. Per esemplificare la metodologia dell'ACM, dall'insieme dei dati raccolti, è stato estratto un campione casuale di 400 intervistati e sono state selezionate 14 delle domande originali, in buona parte modificate per maggiore semplicità e chiarezza espositiva. L'esempio ha sopratutto fini pedagogici.

Anche il questionario meno complesso comporta sempre dei gruppi omogenei di domande che indagano su aspetti diversi del problema e che, nel gergo dell'ACM, sono detti *temi* dell'indagine. Nell'esempio, le 14 variabili appartengono a 4 temi distinti:

A - programmi che vengono ascoltati (3 domande),
C - attività svolta durante l'ascolto e durata giornaliera dell'ascolto (2 domande),
D - atteggiamento verso la pubblicità radiofonica (5 domande),
E - profilo socio-demografico dell'ascoltatore (4 domande).

Le domande elencate nelle TAV. 5.1, 5.2 e 5.3 si riferiscono tutte a quesiti con scelta multipla: tra le possibili modalità proposte, l'intervistato ne poteva scegliere *una ed una sola*. Così, per rispondere al quesito: 'D4 - La musica facilita il ricordo degli spot radiofonici?', si poteva scegliere una soltanto di queste due possibili risposte:

1 - 'Sì, facilita.', oppure
2 - 'No, non facilita.'.

Per l'ACM i 14 quesiti individuano 14 variabili *categoriche*. La domanda E2 'Anno di nascita dell'intervistato' prevedeva originariamente come risposta un numero. Per trasformare la corrispondente variabile numerica in categorica, l'anno di nascita è stato convertito prima in 'Anni di età dell'intervistato' che, ripartiti poi in 7 classi, hanno fornito la nuova variabile 'E2 - Fascia di età dell'intervistato', categorica a 7 modalità che è quella riportata nella TAV.

5.3. La riduzione di variabili numeriche in categoriche è trattata nella Sez. 5.16.

In base al gruppo di variabili, o *tema*, che si sceglie come attivo [1], sono possibili diversi tipi di analisi. In linea di principio si potrebbero considerare attive tutte le variabili rilevate, ma ciò vorrebbe dire confrontare gli intervistati tenendo conto simultaneamente del tipo di programmi che ascoltano, di quello che fanno durante l'ascolto, del loro atteggiamento verso la pubblicità, oltre che del loro profilo socio-demografico. Diventerebbe arduo interpretare eventuali somiglianze o differenze tra intervistati, perché le cause potrebbero essere di tipo diverso. Perciò, è più sensato selezionare un gruppo di variabili che sia omogeneo rispetto a un tema ben definito e coerente con l'obiettivo dell'indagine. Il tema scelto, ossia il gruppo di variabili che si considerano attive, definisce il *punto di vista* secondo cui confrontare gli intervistati, confronto che risulterà così più facile da interpretare. Le variabili degli altri temi verranno considerate illustrative, nel senso che eventuali somiglianze o diversità tra intervistati potranno essere poi illustrate, ossia 'spiegate', dalle modalità di queste variabili. L'importanza e la ricchezza dell'ACM sta proprio in questo: far affiorare eventuali connessioni tra temi diversi che il loro studio separato non sarebbe in grado di far rivelare.

Di solito, ma non sempre, le variabili attive sono quelle che descrivono più o meno obiettivamente gli individui ed illustrative le domande che sono la ragione stessa dell'indagine. In questo esempio, sono considerate attive le $Q = 4$ variabili del tema E [2], che descrivono il profilo socio-demografico dell'ascoltatore intervistato, ed illustrative le altre 8 variabili, appartenenti ai temi A, C e D, per un totale di 50 modalità illustrative. Si è interessati perciò a studiare le somiglianze e le diversità socio-demografiche dei 400 intervistati, i cui profili saranno confrontati in base ai loro descrittori demo-sociali, somiglianze e diversità che verranno poi 'spiegate' dal tipo di programma ascoltato (tema A), dalle modalità di ascolto (tema C) e dall'atteggiamento verso la comunicazione pubblicitaria (tema D).

Si potrebbe anche considerare come attivo quest'ultimo tema e illus-

---

[1] E' il gruppo di variabili che servirà a calcolare gli assi fattoriali. Devono essere tutte di tipo categorico.

[2] Si vedrà nel Capitolo 7 che una analisi con solo 4 variabili attive può fornire risultati poco stabili. Nell'indagine originale le variabili attive comprendevano anche quelle socio-culturali, qui trascurate, perché le scelte sono spesso influenzate dal vissuto dell'intervistato.

trativi gli altri. I risultati non cambierebbero di molto, come capita nella maggior parte dei casi, anche se le due analisi focalizzerebbero l'attenzione su aspetti diversi del problema. Per questo è consigliabile effettuare più analisi, variando il tema attivo. In tutti i casi, la dicotomia tra variabili attive e illustrative ha molte analogie con quella tra variabili 'da spiegare' ed 'esplicative' di una regressione multipla.

Nei casi reali le modalità attive sono spesso alcune decine. In un'estesa e approfondita ricerca di mercato si può facilmente arrivare a una quarantina di modalità attive e a 150-200 illustrative.

5.4 - Codifica compatta

Esistono vari modi per organizzare i dati raccolti in un sondaggio. Per esempio, si possono riunire in una tabella del tipo *individui* × *variabili* di ordine $I \times Q$ avente tante righe quante sono gli individui intervistati e tante colonne quante sono le *variabili categoriche* attive, come nella TAV. 5.4. All'incrocio della riga i con la colonna q vi è il numero d'ordine della modalità scelta dall'intervistato i per rispondere alla domanda q: 1 se ha scelto la prima, 2 la seconda e infine J_q se ha scelto l'ultima modalità possibile per quella domanda. Nella TAV. 5.4 si vede che il primo intervistato ($i = 1$) ha scelto la seconda modalità per rispondere alla prima domanda ($q = 1$) e poi la terza e la prima modalità per rispondere alle due domande successive. Questa codifica ha il vantaggio di essere compatta, tanto da venire correntemente utilizzata per trasferire i dati tra computer o tra programmi, ma non è direttamente utilizzabile perché i totali marginali, ossia le somme per riga e per colonna, non avrebbero significato.

5.5 - Ipermatrice di contingenza

Con le risposte a un sondaggio che preveda due sole domande, $Q = 2$, si può costruire una matrice di contingenza mettendo in corrispondenza i due insiemi J_1 e J_2 di modalità, come si è visto nel primo Capitolo. Allo stesso modo, nel caso dell'esempio della TAV. 5.4 in cui le variabili attive sono $Q = 3$, si può pensare di costruire una 'ipermatrice di contingenza' a 3 dimensioni, una per ogni variabile, incrociando *tutte* le modalità. Complessivamente, gli elementi dell'ipermatrice risultano essere $J_1 \times J_2 \times J_3 = 2 \times 3 \times 3 = 18$, ma molti di essi saranno nulli, dato che gli intervistati sono soltanto $I = 15$. Il concetto può essere generalizzato, ma il numero di elementi cresce così rapidamente all'aumentare del numero di variabili, che in pratica pressoché tutti gli elementi dell'ipermatrice sono nulli. L'interesse per questo tipo di

codifica è perciò limitato, anche perché l'ipermatrice è difficile da gestire. Soltanto il caso con $Q = 3$ merita attenzione, particolarmente quando una delle variabili è il tempo. All'analisi di ipermatrici di questo tipo è dedicata la Sez. 8.9 e l'intero Cap. 15.

5.6 - Codifica disgiuntiva completa

Un modo alternativo di organizzare i dati raccolti è quello di ordinarli in una tabella di indicatori del tipo *individui* × *modalità* , con I righe, una per ogni individuo, e con J colonne, quante sono complessivamente le *modalità* attive, come si vede nella TAV 5.4. All'incrocio della riga i con la colonna j un simbolo qualunque, per esempio un '*sì*' o un '+', può indicare che l'intervistato ha scelto *quella* modalità di risposta, mentre un '*no*' o un '−' può indicare invece che l'ha rifiutata. Di solito si preferisce usare delle cifre, per esempio 1 per indicare le modalità scelte e 0 per quelle rifiutate. Questi 0 e 1 *non* sono numeri, ma semplici indicatori, come ad esempio quelli impiegati in elettronica per indicare lo stato di un circuito: 0 per circuito aperto e 1 per circuito chiuso. Questo tipo di codifica viene detto *disgiuntivo e completo*, disgiuntivo perché le modalità di ogni variabile sono esclusive, in quanto *soltanto una* può essere scelta e completo perché *necessariamente una* modalità è scelta. Per conservare tale carattere, talvolta si rende necessario prevedere, o aggiungere successivamente, a qualche variabile la modalità 'Nessuna risposta'. E' il caso delle risposte mancanti, o non risposte, trattate nella Sez. 5.21.

Come rivela chiaramente la TAV. 5.4, la tabella 15×8 è formata da $Q = 3$ sottotabelle di indicatori affiancate, una per ogni variabile attiva, con $J_1 = 2, J_2 = 3$ e $J_3 = 3$ colonne, entro le quali, in ciascuna riga, l'1 compare una e una sola volta dato che le risposte sono esclusive. Ne consegue che ognuna delle 3 sottotabelle ha la colonna marginale costituita da 1 . Questa peculiarità della codifica disgiuntiva completa ha importanti conseguenze, come si vedrà nelle Sez. 5.13 e 5.14.

5.7 - Matrice di Burt

La matrice di contingenza di Burt prende il nome dello psicologo britannico [1] che la introdusse nel 1950 e si ottiene incrociando due a due tutte le J modalità delle Q variabili attive, come si vede nella TAV. 5.4. La matrice di Burt, indicata abitualmente con $\mathbf{B}$, è simmetrica, di ordine $J \times J$

[1] Cyril Lodowic Burt: Londra 1883, Würzburg, Londra 1971.

ed assomiglia a una matrice di covarianza, nel senso che riassume i legami tra le modalità, prese *due a due*. Dato che i suoi elementi indicano il numero di individui che possiedono entrambe le modalità, ogni individuo vi compare Q^2 volte. In realtà, non si tratta di una vera matrice di contingenza, ma di un *patchwork* di blocchi, di una composizione di $Q \times Q$ matrici di contingenza, ciascuna ottenuta incrociando le J_q modalità di una variabile con le $J_{q'}$ delle altre e anche con le J_q di sè stessa. Nel primo caso il blocco è, in generale, rettangolare, mentre nel secondo i blocchi diagonali sono matrici quadrate, non nulle e diagonali, dato che le modalità di una stessa domanda sono esclusive. Gli elementi diagonali riportano il numero di individui che hanno scelto ogni singola modalità di risposta.

La matrice di Burt si può ottenere indifferentemente dalla tabella *individui*×*variabili* o da quella *individui*×*modalità* . In questo caso, la riga, o la colonna, j della matrice di Burt sono il conteggio delle righe della tabella di indicatori della Sez. 5.6 in cui è presente la modalità j . Ad esempio, la prima riga della matrice di Burt nella TAV. 5.4 ha $b_{11} = 8$ perché tanti sono gli individui che hanno scelto la prima modalità della prima variabile; $b_{12} = 0$ perché avendo essi scelto la prima modalità non potevano scegliere anche la seconda e ultima modalità della prima variabile; $b_{13} = 2$ perché due sono gli individui (il 3^o e il 10^o) che hanno scelto contemporaneamente la prima modalità della prima variabile e la prima della seconda, e così via.

La matrice di Burt ha il grande vantaggio di avere dimensioni ridotte, ma è meno informativa, sia perché si perde l'identità degli individui, sia perché non permette di risalire agli altri tipi di codifica. In particolare, mentre è sempre possibile costruire una matrice di Burt partendo da una tabella di indicatori, l'inverso non è possibile dato che la matrice di Burt si limita a riportare le associazioni soltanto tra *coppie* di modalità.

Lo specchietto riassume l'ordine delle matrici nei tre principali tipi di codifica, con riferimento al sondaggio sull'ascolto delle trasmissioni radio della Sez. 5.3.

Codifica	*Ordine*	*Esempio*
Compatta	$I \times Q$	400×4
Disg. Compl.	$I \times J$	400×21
Burt	$J \times J$	21×21

5.8 - Obiettivi dell'analisi

In sostanza, l'ACM è l'Analisi delle Corrispondenze dei profili ottenuti da una tabella di indicatori, codificati in forma disgiuntiva completa. Al solito, il suo fine è quello di rendere graficamente evidenti le relazioni tra modalità, tra individui e tra individui e modalità, proiettando i loro profili in sottospazi di ridotta dimensionalità e tali da mostrare la configurazione geometrica dei profili con la minore distorsione.

Mentre l'Analisi delle Corrispondenze semplici era incentrata esclusivamente sulle modalità (di due sole variabili), l'ACM ha a che fare anche con individui e con variabili. Lo studio di questi tre elementi comporta esigenze e problemi diversi che l'ACM cerca di contemperare nell'analisi delle sole modalità, perché questa permette di studiare implicitamente i legami tra coppie di variabili e al contempo, di esaminare il comportamento di interi segmenti di individui. In altri termini l'analisi delle modalità permette di effettuare in gran parte lo studio delle variabili e degli individui.

Studio degli individui

L'obiettivo è la ricerca di individui con profili simili, cioò col maggior numero di modalità in comune. L'obiettivo è quindi analogo a quello dell'analisi dei profili delle righe nelle Corrispondenze semplici; la differenza è che ora gli individui sono generalmente anonimi e possono essere molto numerosi: centinaia e talvolta migliaia nel caso di vaste indagini, per cui la rappresentazione individuale dei loro profili renderebbe qualunque mappa fattoriale eccessivamente affollata e illeggibile. Si agisce allora per due vie: per segmentazione e per raggruppamento. Nel primo caso gli individui sono studiati tramite i segmenti stabiliti dalle modalità perché si vedrà nella Sez. 5.15 che una modalità è il baricentro di tutti gli individui che la possiedono. Ad esempio, nello spazio dei profili, la posizione della modalità 'laureati' indica il baricentro del segmento di intervistati con questo titolo di studio.

L'altra via è quella di utilizzare le coordinate fattoriali degli individui ottenute con l'ACM per creare dei gruppi o 'cluster' di individui con profili di risposta il più possibile simili e di proiettare poi sulla mappa soltanto i baricentri di questi cluster come rappresentativi dei gruppi. L'analisi dei gruppi, ossia la costruzione di cluster di individui omogenei dal punto di vista del profilo delle risposte al tema attivo, sarà oggetto del prossimo Capitolo 6.

Studio delle variabili

Come si è detto, nell'ACM le variabili non compaiono esplicitamente, ma restano in secondo piano perché similitudini o difformità tra variabili si desumono dal confronto dei profili delle modalità che le costituiscono. Per

questo motivo l'analisi viene effettuata al livello di maggior dettaglio: quello delle modalità. Tuttavia, nell'interpretazione degli assi fattoriali è utile tener conto anche delle variabili che maggiormente hanno contribuito al loro orientamento.

Studio delle modalità

Le somiglianze tra modalità si possono indagare confrontando sia le colonne della matrice $\mathbf{C}$ dei profili di ordine $I \times J$, definita nella prossima Sez. 5.11, sia i profili della matrice di $\mathbf{B}$ di Burt. Nel primo caso due modalità sono simili se sono state scelte o rifiutate sempre, o quasi sempre, dai *medesimi* individui. Le altre modalità non intervengono nel confronto. Nel secondo caso la somiglianza è invece analoga a quella che si ha nelle Corrispondenze semplici, perché ogni profilo tiene conto dell'associazione della modalità con tutte le altre. Di conseguenza due profili di $\mathbf{B}$ risultano simili se le due modalità si associano sempre alle *medesime* modalità. Si vedrà comunque nella Sez. 5.19 che i risultati delle due analisi sono comparabili.

5.9 - Profili marginali

L'Analisi delle Corrispondenze Multiple (ACM) segue sostanzialmente lo stesso percorso di analisi delle Corrispondenze semplici, effettuato sulle matrici dei profili delle righe e delle colonne. Punto di partenza è ora la tabella degli indicatori delle modalità descritta nella Sez. 5.6, organizzati in forma disgiuntiva completa su I righe, gli individui attivi, e J colonne, le modalità attive. Per ogni individuo $i = 1, 2, \ldots, I$, si *conta* il numero di risposte fornite, numero che per ogni individuo risulta sempre eguale a Q, il numero di variabili attive, perché è possibile scegliere una e una sola modalità per variabile. Per simmetria col totale di colonna, è opportuno indicare questa frequenza con z_{i+}, per cui

$$z_{i+} = \overset{J}{\underset{j=1}{\#}} 1 = Q \qquad\qquad i = 1, ..., I$$

dove il simbolo $\#$ indica il *conteggio*, effettuato su tutta la riga i, degli indicatori delle modalità scelte: degli '1' come qui sopra, dei 'sì', ecc. secondo l'indicatore adottato. In ogni caso, z_{i+} *non* dipende dall'indicatore scelto.

Allo stesso modo, per ogni modalità di risposta e quindi per ogni colonna $j = 1, 2, \ldots, J$, si contano gli individui che l'hanno scelta, e quindi, per esempio, gli 1 nella colonna, ottenendo

$$z_{+j} = \overset{I}{\underset{i=1}{\#}} 1. \qquad\qquad j = 1, ..., J$$

Perciò, mentre il totale di ogni riga è fisso e pari al numero di domande che vengono considerate come variabili attive, il totale z_{+j} di una colonna può invece variare da 1, quando tutti gli intervistati tranne uno hanno rifiutato quella modalità di risposta, a I, nel caso che tutti l'abbiano scelta. Come nell'Analisi delle Corrispondenze semplici, una modalità rifiutata da tutti gli intervistati va preventivamente rimossa dall'analisi. I totali marginali sono perciò *sempre* positivi. I J totali z_{+j} costituiscono la diagonale principale della matrice $\mathbf{B}$ di Burt della Sez. 5.7, come si vede nella TAV. 5.4.

Il numero di individui che nel corso del sondaggio hanno risposto alla domanda q con J_q modalità, è necessariamente eguale al numero di intervistati

$$\sum_{j'=1}^{J_q} z_{+j'} = I \tag{5.9.1}$$

perché ciascuno di essi ha scelto necessariamente *una* delle modalità esclusive della domanda q.

Infine, il numero totale di risposte ottenute nel sondaggio, limitatamente alle variabili attive, risulta

$$z_{++} = \sum_{i=1}^{I} z_{i+} = \sum_{j=1}^{J} z_{+j} = I\,Q \tag{5.9.2}$$

perché ogni individuo ha fornito Q risposte.

Procedendo come nell'Analisi delle Corrispondenze semplici, vengono calcolati i profili marginali che costituiscono le masse con cui ponderare i profili. Poiché le frequenze (assolute) marginali z_{i+} e z_{+j} sono numeri reali positivi, è lecito dividerle per la frequenza complessiva z_{++}, ottenendo le J masse (o pesi relativi) dei profili delle colonne, e delle I righe, rispettivamente così definite

$$\overline{r}_j \stackrel{\text{def}}{=} \frac{z_{+j}}{z_{++}} = \frac{z_{+j}}{I\,Q} \qquad \overline{c}_i \stackrel{\text{def}}{=} \frac{z_{i+}}{z_{++}} = \frac{Q}{I\,Q} = \frac{1}{I}. \tag{5.9.3}$$

Entrambe le somme delle masse sono pari a 1

$$\sum_{j=1}^{J} \overline{r}_j = 1 \qquad \text{e} \qquad \sum_{i=1}^{I} \overline{c}_i = 1 \tag{5.9.4}$$

ma, mentre tutti gli individui hanno la stessa massa, questa può variare da $1/Q$ a 1 per le singole modalità. Il rapporto z_{+j}/I nella (5.9.3) rappresenta la frazione di individui che ha scelto la modalità j.

La massa $\overline{r}_q$ di una variabile $q = 1, 2, \ldots, Q$ è la somma delle J_q masse delle sue modalità e, per la (5.9.3) e la (5.9.1), vale

$$\overline{r}_q = \sum_{j'=1}^{J_q} \overline{r}_{j'} = \sum_{j'=1}^{J_q} \frac{z_{+j'}}{I\,Q} = \frac{1}{Q}. \tag{5.9.5}$$

Perciò, le variabili attive hanno tutte la stessa massa. Massa che si ripartisce tra le modalità proporzionalmente alle frequenze di risposta. Riassumendo, nell'ACM sono eguali le masse $\overline{c}_i$ degli individui, quelle $\overline{r}_q$ delle variabili, ma non quelle $\overline{r}_j$ delle modalità.

Anche nell'ACM vengono introdotti i vettori delle masse dei profili

$$\begin{aligned} \overline{\mathbf{r}} &\stackrel{\text{def}}{=} (\overline{r}_1\ \overline{r}_2\ \ldots\ \overline{r}_j\ \ldots\ \overline{r}_J)^T \\ \overline{\mathbf{c}} &\stackrel{\text{def}}{=} (\overline{c}_1\ \overline{c}_2\ \ldots\ \overline{c}_i\ \ldots\ \overline{c}_I)^T \end{aligned} \tag{5.9.6}$$

e le relative matrici diagonali $\mathbf{D}_{\overline{\mathbf{r}}}$ e $\mathbf{D}_{\overline{\mathbf{c}}}$ delle masse aventi per elementi diagonali le componenti di $\overline{\mathbf{r}}$ e di $\overline{\mathbf{c}}$ rispettivamente

$$\begin{aligned} \mathbf{D}_{\overline{\mathbf{r}}} &\stackrel{\text{def}}{=} diag\,(\overline{r}_1\ \overline{r}_2\ \ldots\ \overline{r}_j\ \ldots\ \overline{r}_J) \\ \mathbf{D}_{\overline{\mathbf{c}}} &\stackrel{\text{def}}{=} diag\,(\overline{c}_1\ \overline{c}_2\ \ldots\ \overline{c}_i\ \ldots\ \overline{c}_I) \end{aligned} \tag{5.9.7}$$

La prima matrice è di ordine $J \times J$ e la seconda di ordine $I \times I$, come mostra la TAV. 5.5. Va rilevato che per la (5.9.3) risulta

$$I\,\mathbf{D}_{\overline{\mathbf{c}}} = \mathbf{I} \tag{5.9.8}$$

dove $\mathbf{I}$ è la matrice diagonale identità del medesimo ordine di $\mathbf{D}_{\overline{\mathbf{c}}}$. Le matrici inverse $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$ e $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$ hanno per elementi diagonali i reciproci, ossia $1/\overline{r}_j$ e $1/\overline{c}_i$ rispettivamente.

5.10 - Profili delle righe

Nell'Analisi delle Corrispondenze semplici i profili delle righe (delle colonne) sono calcolati dividendo ciascun elemento della matrice di contingenza per il totale della riga (della colonna). Questa operazione non è possibile nell'ACM, perché la tabella degli indicatori delle modalità non è numerica. I profili vanno perciò costruiti con un procedimento di *assegnazione*.

Nel caso dei profili delle righe, come è indicato nella TAV. 5.5, la frequenza assoluta $z_{i+} = Q$ di ogni riga viene equiripartita tra le modalità scelte come risposta dall'individuo i, in modo che il totale della riga risulti 1, come deve essere per un profilo. Si assegna perciò valore $1/Q$ alle Q modalità che l'individuo i ha scelto e valore nullo alle altre $J - Q$ che ha

rifiutato, costruendo così la matrice $\mathbf{R}$, di ordine $I \times J$, il cui elemento generico vale quindi

$$r_{ij} \stackrel{\text{def}}{=} \begin{cases} 0 & \text{se } i \text{ non ha scelto la mod. } j\ ; \\ 1/Q & \text{se l'ha scelta.} \end{cases} \tag{5.10.1}$$

Si ha dunque

$$\sum_{i=1}^{I} r_{ij} = 1 \qquad \text{per cui} \qquad \mathbf{r}_i = (r_{i1}\ \ r_{i2}\ \ \cdots\ \ r_{ij}\ \ \cdots\ \ r_{iJ})^T$$

è un profilo riga, e dunque un vettore colonna di ordine J, le cui componenti sono i J elementi della riga i di $\mathbf{R}$ e al quale viene attribuita la massa $\overline{c}_i$, i^{ma} componente del profilo colonna marginale $\overline{\mathbf{c}}$, che per la (5.9.3) vale

$$\overline{c}_i = \frac{1}{I}.$$

I profili delle righe sono tutti ponderati in egual misura, il che significa che nell'ACM gli individui sono tenuti tutti nella medesima considerazione.

La matrice dei profili delle righe è una matrice reale non negativa su cui è possibile effettuare operazioni algebriche. Un esempio è nella TAV. 5.6.

In quanto desunte dalla medesima tavola, la matrice $\mathbf{R}$ e quella $\mathbf{B}$ di Burt della Sez. 5.7 sono legate dalla relazione

$$Q^2\, \mathbf{R}^T\, \mathbf{R} = \mathbf{B} \tag{5.10.2}$$

che tornerà utile in seguito.

5.11 - Profili delle colonne

In modo analogo, come mostra la TAV. 5.6, è costruita la matrice $\mathbf{C}$ dei profili delle colonne, di ordine $I \times J$, assegnando in ogni colonna $j = 1, 2, \ldots, J$ il valore $1/z_{+j}$, se la modalità è stata scelta dall'individuo i e valore nullo se è stata ignorata. Il generico elemento di $\mathbf{C}$ è quindi

$$c_{ij} \stackrel{\text{def}}{=} \begin{cases} 0 & \text{se } j \text{ non è stata scelta da } i\ ; \\ 1/z_{+j} & \text{se è stata scelta.} \end{cases} \tag{5.11.1}$$

Il profilo colonna $\mathbf{c}_j$ è un vettore colonna di ordine I le cui componenti hanno somma 1 e sono gli elementi della colonna j di $\mathbf{C}$

$$\mathbf{c}_j = (c_{1j}\ \ c_{2j}\ \ \cdots\ \ c_{ij}\ \ \cdots\ \ c_{Ij})^T$$

e la sua massa è $\overline{r}_j$, la j^{ma} componente del profilo riga marginale $\overline{\mathbf{r}}$ che per la (5.9.3) vale

$$\overline{r}_j = \frac{z_{+j}}{I\,Q}.$$

Perciò la massa del profilo $\mathbf{c}_j$ è proporzionale alla frequenza assoluta z_{+j}: le modalità scelte più frequentemente ricevono un peso maggiore nell'analisi, quelle scelte più raramente un peso minore.

Una dimostrazione analoga a quella della Sez. 2.6 porta a concludere che il profilo colonna medio $\overline{\mathbf{c}}$ è la media ponderata dei J profili delle colonne di $\mathbf{C}$, perché in ogni riga $i = 1, 2, \ldots, I$, si ha

$$\sum_{j=1}^{J} \overline{r}_j\, c_{ij} = \sum_{j=1}^{J} \frac{z_{+j}}{I\,Q}\, c_{ij} = \frac{1}{I\,Q}\, Q = \overline{c}_i$$

dato che nella riga i ci sono soltanto Q valori non nulli di c_{ij} e ciascuno di essi vale $1/z_{+j}$ per la (5.11.1).

La codifica disgiuntiva completa introduce, però, una ortogonalità artificiale tra i profili $\mathbf{c}_j$ delle modalità di una *medesima* variabile. Infatti, è nullo il prodotto scalare tra due di questi profili $\mathbf{c}_j$ e $\mathbf{c}_{j'}$, relativi alle modalità $j, j' = 1, 2, \ldots, J_q$, ma con $j \neq j'$,

$$\mathbf{c}_j^T\, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_{j'} = \sum_{i=1}^{I} c_{ij}\, \frac{1}{\overline{c}_i}\, c_{ij'} = 0 \tag{5.11.2}$$

perché per ogni individuo i le J_q modalità sono esclusive. Così, ad esempio, il prodotto scalare delle due prime colonne di $\mathbf{C}$ nella TAV. 5.6 è

$$\mathbf{c}_1^T\, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{c}_2 = 0 \times \frac{45}{3} \times \frac{1}{7} + \frac{1}{8} \times \frac{45}{3} \times 0 + \ldots + 0 \times \frac{45}{3} \times \frac{1}{7} = 0.$$

Da questi dati si vede che i profili sono anche semplicemente *ortogonali*, perché per essi è $\mathbf{c}_1^T\, \mathbf{c}_2 = 0$.

Inoltre, il profilo medio ponderato delle modalità della stessa variabile coincide con quello di tutti i J profili, perché

$$\sum_{j'=1}^{J_q} \overline{r}_{j'}\, c_{ij'} = \sum_{j'=1}^{J_q} \frac{z_{+j'}}{I}\, \frac{1}{z_{+j'}} = \frac{1}{I} = \overline{c}_i \tag{5.11.3}$$

Ogni sub-nuvola costituita dai profili delle modalità di una stessa variabile ha quindi come baricentro $\overline{\mathbf{c}}$, quello dell'intera nuvola. La nuvola dei J profili consiste dunque nella sovrapposizione di Q sub-nuvole, tante quante sono le variabili, tutte centrate nello stesso punto $\overline{\mathbf{c}}$ e con i profili $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonali.

Quando però, manca la risposta a una o più domande, la codifica non è più completa e $\overline{\mathbf{c}}$ non è più il profilo medio (baricentro) dei J_q profili. Di qui l'importanza di un accurato controllo dei dati raccolti prima di ogni analisi.

Anche il profilo riga medio $\overline{\mathbf{r}}$ è la media ponderata dei profili delle righe di $\mathbf{R}$, perché in ogni colonna $j = 1, 2, \ldots, J$ è

$$\sum_{i=1}^{I} \overline{c}_i \, r_{ij} = \frac{1}{I} \sum_{i=1}^{I} r_{ij} = \frac{1}{I} \, z_{+j} \, \frac{1}{Q} = \overline{r}_j$$

dato che in una colonna di $\mathbf{R}$ ci sono soltanto z_{+j} valori non nulli, ciascuno dei quali vale $1/Q$ per la (5.10.1).

Per concludere, grazie al modo con cui sono state costruite, le matrici dei profili e delle loro masse sono legate da una relazione analoga alla (3.2.6)

$$\mathbf{C} = \mathbf{D}_{\overline{\mathbf{c}}} \mathbf{R} \, \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}} \tag{5.11.4}$$

perché per le (5.9.3), la (5.10.1) e la (5.11.1), è proprio

$$\overline{c}_i \, r_{ij} \, \frac{1}{\overline{r}_j} = \frac{1}{I} \, \frac{1}{Q} \, \frac{I \, Q}{z_{+j}} = \frac{1}{z_{+j}} = c_{ij}.$$

5.12 - Distanza distribuzionale tra profili

Interpretati geometricamente, i profili $\mathbf{r}_i$ degli individui costituiscono una nuvola di I punti dotati di massa uniforme $\overline{c}_i$ nello spazio J-dimensionale $\Re^J$ con baricentro il profilo medio $\overline{\mathbf{r}}$. Analogamente, i profili $\mathbf{c}_j$ delle modalità costituiscono una nuvola di J punti con massa $\overline{r}_j$ nello spazio I-dimensionale $\Re^I$. Il loro baricentro è $\overline{\mathbf{c}}$. In entrambi gli spazi la configurazione geometrica della nuvola traduce la struttura dei profili, nel senso che profili simili sono rappresentati da punti vicini, anche se in $\Re^I$ si fa sentire il carattere disgiuntivo della codifica adottata.

I due spazi sono dotati di un sistema di riferimento ortogonale, individuato dai vettori della base canonica. Fine dell'ACM è di agevolare il confronto tra profili, evidenziando 'al meglio' la configurazione geometrica delle nuvole in sottospazi di ridotta dimensionalità. La base canonica non dà garanzie in tal senso, perciò, assunta come distanza quella distribuzionale tra profili della medesima nuvola, assunzione giustificata dalla proprietà equidistributiva che rende il metodo 'robusto', per ciascuna nuvola, separatamente, si ricerca un nuovo sistema di riferimento ortonormale, con origine nel baricentro della nuvola, e tale da rendere massima la dispersione delle proiezioni

dei profili su ciascun asse. La dispersione è misurata in termini di inerzia, o varianza. Gli assi individuati sono detti assi fattoriali d'inerzia ed a ciascuno corrisponde un preciso valore dell'inerzia. I fattori sono le ascisse dei profili nel nuovo sistema di riferimento degli assi fattoriali. Le relazioni di transizione tra fattori permettono di ricavare i fattori di una nuvola, noti i corrispondenti fattori dell'altra, e sono estremamente utili dal punto di vista computazionale perché, tra l'altro, rendono superflua una delle due analisi.

Come nell'Analisi delle Corrispondenze semplici, nello spazio J-dimensionale $\Re^J$ la distanza *distribuzionale* tra i profili $\mathbf{r}_i$ e $\mathbf{r}_{i'}$ di due individui e quindi tra due righe della matrice $\mathbf{R}$, è definita come

$$d_D^2(\mathbf{r}_i, \mathbf{r}_{i'}) \stackrel{\text{def}}{=} \sum_{j=1}^{J} \frac{(r_{ij} - r_{i'j})^2}{\overline{r}_j} = I\,Q \sum_{j=1}^{J} \frac{(r_{ij} - r_{i'j})^2}{z_{+j}} \tag{5.12.1}$$

grazie alla (5.9.3).

Questa distanza conferisce allo spazio $\Re^J$ una struttura euclidea. Essa è nulla se entrambi gli individui hanno scelto le stesse modalità e cresce all'aumentare del numero di modalità che i due individui *non* hanno in comune, ossia quanto più diversi sono i loro profili di risposta. Inoltre, ciascuna della J modalità influisce sulla distanza inversamente alla sua frequenza assoluta z_{+j}, per cui se uno dei due individui ha scelto delle modalità raramente scelte da tutti gli altri, si trova allontanato dall'altro individuo.

Analogamente, nello spazio I-dimensionale $\Re^I$ la distanza *distribuzionale* tra i profili $\mathbf{c}_j$ e $\mathbf{c}_{j'}$ di due modalità è definita

$$d_D^2(\mathbf{c}_j, \mathbf{c}_{j'}) \stackrel{\text{def}}{=} \sum_{i=1}^{I} \frac{(c_{ij} - c_{ij'})^2}{\overline{c}_i} = I \sum_{i=1}^{I} (c_{ij} - c_{ij'})^2 \tag{5.12.2}$$

Di conseguenza, i profili di due modalità coincidono quando queste sono scelte dai medesimi individui, e la loro distanza aumenta col numero di individui che scelgono una e non l'altra delle modalità. Inoltre, per la (5.11.1), c_{ij} e $c_{ij'}$ risultano inversamente proporzionali al numero di individui z_{+j} e $z_{+j'}$ che le hanno scelte, per cui la distanza diminuisce all'aumentare di ciascuna di queste frequenze: una modalità scelta da pochi individui viene a trovarsi lontana dalle altre. Ciò vale obbligatoriamente per modalità appartenenti alla stessa variabile.

La distanza di una modalità dall'origine della base ortogonale di rifer-

imento di $\Re^I$ è

$$d_D^2(\mathbf{c}_j, \mathbf{0}_\mathrm{I}) = \sum_{i=1}^{I} \frac{(c_{ij} - 0)^2}{\overline{c}_i} = I \sum_{i=1}^{I} c_{ij}^2 = I\, z_{+j} \left(\frac{1}{z_{+j}}\right)^2 = \frac{I}{z_{+j}} \qquad (5.12.3)$$

perché degli I termini della somma, soltanto z_{+j} non sono nulli e per la (5.11.1) valgono ciascuno $1/z_{+j}$.

La distanza del profilo di una modalità dal profilo medio, che è piatto e rappresenta il baricentro della nuvola di tutti i profili, per la (5.9.3) prima e per la (5.12.3) e (5.12.2) poi, risulta

$$\begin{aligned} d_D^2(\mathbf{c}_j, \overline{\mathbf{c}}) &= \sum_{i=1}^{I} \frac{(c_{ij} - \overline{c}_i)^2}{\overline{c}_i} = I \sum_{i=1}^{I} \left(c_{ij} - \frac{1}{I}\right)^2 \\ &= I \left(\sum_{i=1}^{I} c_{ij}^2 - \frac{2}{I} \sum_{i=1}^{I} c_{ij} + \frac{1}{I^2} \sum_{i=1}^{I} 1\right) = \frac{I}{z_{+j}} - 1. \end{aligned} \qquad (5.12.4)$$

Ne consegue che il profilo $\mathbf{c}_j$ tanto più assomiglia al profilo medio quanto più la sua frequenza assoluta è grande. In altri termini, nello spazio $\Re^I$ i profili di modalità scelte con maggior frequenza si collocano vicini al baricentro, quelli di modalità scelte raramente ne sono lontani.

5.13 - Inerzia delle modalità e delle variabili

Nella Sez. 5.8 si è detto che nell'ACM l'analisi viene effettuata al livello più basso, quello delle modalità in quanto il loro studio permette, indirettamente, quello delle variabili e dei segmenti di individui. È importante quindi valutare l'inerzia dei loro profili.

L'inerzia del profilo $\mathbf{c}_j$, riferita all'origine $\mathbf{0}_\mathrm{I}$ della base ortogonale di riferimento introdotta nella Sez. 3.1, per la (5.9.3) e (5.12.3), vale

$$In_{\mathbf{0}}(\mathbf{c}_j) = \overline{r}_j\, d_D^2(\mathbf{c}_j, \mathbf{0}_\mathrm{I}) = \frac{z_{+j}}{I\,Q} \frac{I}{z_{+j}} = \frac{1}{Q}. \qquad (5.13.1)$$

L'inerzia complessiva dei J profili della nuvola risulta

$$In_{\mathbf{0}} = \sum_{j=1}^{J} In_{\mathbf{0}}(\mathbf{c}_j) = \sum_{j=1}^{J} \frac{1}{Q} = \frac{J}{Q}. \qquad (5.13.2)$$

e dipende *unicamente* dal numero di modalità e di variabili.

Invece, l'inerzia riferita al baricentro $\overline{\mathbf{c}}$, che nella Sez. 3.4 si è visto essere il punto privilegiato per valutare l'inerzia, per la (5.9.3) e (5.12.4),

risulta

$$In_{\overline{\mathbf{c}}}(\mathbf{c}_j) = \overline{r}_j \, d_D^2(\mathbf{c}_j, \overline{\mathbf{c}}) = \frac{z_{+j}}{I\,Q} \left(\frac{I}{z_{+j}} - 1 \right) = \frac{1}{Q} - \overline{r}_j \qquad (5.13.3)$$

e quindi aumenta se diminuiscono gli individui che l'hanno scelta, fino al limite superiore $1/Q$. La (5.13.3) rivela come una modalità possa influenzare l'orientamento degli assi fattoriali: la struttura del profilo non interviene minimamente, ciò che conta è la sua massa. Per esempio, se la massa di una modalità è di $1/100$, mentre quella di una seconda è di $50/100$, le due inerzie sono rispettivamente $1 - 1/100 \simeq 1$ e $1 - 1/2 = 1/2$, e quindi le capacità di orientare un asse stanno nel rapporto di 2 a 1 . Questo significa che i primi assi fattoriali sono influenzati quasi esclusivamente dalle modalità rare, scelte da pochi individui. Perciò, prima di iniziare l'analisi è opportuno eliminare le modalità troppo rare che riportano fenomeni poco generali e quindi poco interessanti. Se tutte le modalità di una variabile, salvo una, hanno massa molto piccola, tutta la variabile va trasferita tra le illustrative, altrimenti, si può sopprimere la modalità e ai pochi individui che l'hanno scelta attribuire a caso una delle altre modalità purché abbiano massa sufficiente. Questa operazione, detta di *ventilazione* delle modalità rare, permette di conservare la struttura disgiuntiva della matrice dei profili.

Come detto, le variabili non compaiono esplicitamente nell'ACM, ma solo indirettamente tramite le loro modalità. Le nuvole dei profili delle modalità di una *medesima* variabile hanno però proprietà interessanti che occorre aver presente quando si interpretano le mappe o si vuol trasformare una variabile quantitativa in qualitativa, come si vedrà nella Sez. 5.16 . Si è visto che la nuvola consiste nella sovrapposizione di Q sub-nuvole, quante sono le variabili, tutte con lo stesso baricentro $\overline{\mathbf{c}}$ in comune. Questa proprietà deriva dal fatto che i profili delle modalità non sono indipendenti perché nella Sez. 5.6 si è visto che per via della codifica disgiuntiva completa il totale marginale delle modalità di *una* variabile vale sempre 1. Questa proprietà si conserva nelle proiezioni, per cui in tutte le mappe l'insieme delle modalità di una stessa variabile ha ancora per baricentro il baricentro della nuvola. Ne deriva, in particolare, che le ascisse delle proiezioni di queste modalità su un asse fattoriale non possono avere tutte lo stesso segno. Di conseguenza, gli assi fattoriali oppongono sia l'insieme di tutte le J modalità che l'insieme delle J_q modalità di ogni variabile.

L'inerzia delle modalità di una stessa variabile, per la (5.13.3) e la

(5.9.1) è

$$In_{\overline{\mathbf{c}}}(q) \stackrel{\text{def}}{=} \sum_{j'=1}^{J_q} In_{\overline{\mathbf{c}}}(\mathbf{c}_{j'}) = \frac{1}{Q} \sum_{j'=1}^{J_q} \left(1 - \frac{z_{+j'}}{I}\right) = \frac{1}{Q} \left(J_q - \frac{1}{I} \sum_{j'=1}^{J_q} z_{+j'}\right)$$
$$= \frac{1}{Q} (J_q - 1) . \qquad (5.13.4)$$

Così, il contributo di una variabile all'inerzia totale della nuvola, è minimo quando le modalità sono soltanto due e cresce linearmente al crescere delle sue modalità. Per equilibrare i contributi e far sì che tutte le variabili attive abbiano un ruolo paragonabile, bisogna fare in modo che abbiano tutte più o meno lo stesso numero di modalità, accorpandole, se possibile, o trasferendo intere variabili tra le illustrative. Bisogna evitare, ad esempio, di avere contemporaneamente come variabili attive il 'Sesso dell'intervistato' con 2 modalità e la 'Regione di residenza' con 20 modalità.

Nel caso delle 4 variabili attive dell'esempio, $Q = 4$, le inerzie risultano

q	*Variabile*	J_q	$In_{\overline{\mathbf{c}}}(q)$
1	E_1 : Sesso	2	$(2-1)/4 = 0.25$
2	E_2 : Età	7	$(7-1)/4 = 1.50$
3	E_3 : Titolo	4	$(4-1)/4 = 0.43$
4	E_4 : Profess.	8	$(8-1)/4 = 1.75$

L'inerzia totale della nuvola di profili delle modalità, per la (5.13.4) e per la (5.2.1), vale

$$In_{\overline{\mathbf{c}}} = \sum_{q=1}^{Q} In_{\overline{\mathbf{c}}}(q) = \frac{1}{Q} \left(\sum_{q=1}^{Q} J_q - Q\right) = \frac{J}{Q} - 1. \qquad (5.13.5)$$

e dipende soltanto dal *numero* di variabili e di modalità, in altre parole dalla particolare codifica adottata. Non è più quindi un indicatore statistico della dispersione delle nuvole di profili, né dell'intensità dei loro legami come nelle Corrispondenze semplici. In particolare, se ogni variabile ha $J_q = 2$ modalità l'inerzia totale riferita al baricentro è $In_{\overline{\mathbf{c}}} = 1$.

Anche nell'ACM le inerzie totali delle due nuvole di profili delle colonne e delle righe sono identiche e infatti, nel caso dell'esempio, si ha: $In_{\overline{\mathbf{c}}} = In_{\overline{\mathbf{r}}} = 21/4 - 1 = 4.25$.

Le inerzie totali riferite all'origine della base canonica (5.13.2) e al baricentro (5.13.6), verificano comunque il teorema di di Huygens della Sez. 3.4, perché risulta

$$In_{\overline{\mathbf{c}}} + 1 = \frac{J}{Q} - 1 + 1 = \frac{J}{Q} = In_{\mathbf{0}}.$$

5.14 - Autovalori, autovettori e fattori

L'ACM è in sostanza l'Analisi delle Corrispondenze semplici delle matrici dei profili $\mathbf{R}$ e $\mathbf{C}$ ricavate dalla tabella di indicatori organizzata in *individui* $\times$ *modalità*, con gli indicatori codificati in forma disgiuntiva completa. La ricerca degli assi fattoriali si sviluppa in modo analogo a quella delle Corrispondenze semplici descritta nella Sez. 3.9 per la nuvola dei profili delle colonne in $\Re^I$ e nella Sez. 4.8 per quella delle righe in $\Re^J$. In entrambi i casi porta a risolvere le due equazioni agli autovalori

$$\mathbf{C}\,\mathbf{R}^T\mathbf{u}_a^\star = \lambda_a\,\mathbf{u}_a^\star \qquad \text{con} \qquad \mathbf{u}_a^{\star T}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_a^\star = 1 \tag{5.14.1}$$

$$\mathbf{R}^T\mathbf{C}\,\mathbf{v}_a^\star = \mu_a\,\mathbf{v}_a^\star \qquad \text{con} \qquad \mathbf{v}_a^{\star T}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\,\mathbf{v}_a^\star = 1. \tag{5.14.2}$$

dove λ_a e μ_a sono gli autovalori e $\mathbf{u}_a^\star$ e $\mathbf{v}_a^\star$ i corrispondenti autovettori di rango $a = 1, 2, \ldots, A$ con origine nel baricentro delle due nuvole.

Ma quanti sono gli autovalori non nulli nell'ACM, o, in altri termini, quanto vale A? Per cominciare si può esaminare il problema da un punto di vista geometrico, considerando la nuvola dei profili colonna $\mathbf{c}_j$ nello spazio $\Re^I$ con origine in $\mathbf{0}_{\mathrm{I}}$ e, in particolare, i J_q profili delle modalità di una *stessa* variabile. Nella Sez. 2.3 si è visto che i profili possono immaginarsi anche come vettori che connettono i punti $\mathbf{c}_j$ all'origine $\mathbf{0}_{\mathrm{I}}$. I J_q vettori di questo fascio sono $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonali tra loro, come risulta dalla (5.11.2), a causa della codifica disgiuntiva ed individuano quindi un sottospazio di dimensione J_q ma, poiché la codifica è anche completa, uno dei J_q profili è $\overline{\mathbf{c}}$, il vettore che collega l'origine $\mathbf{0}_{\mathrm{I}}$ al baricentro della sub-nuvola che, come risulta dalla (5.11.3), coincide con quello dell'intera nuvola. Di conseguenza i profili delle modalità di una stessa variabile, intesi ora come punti, si trovano in un sottospazio a $J_q - 1$ dimensioni, *ortogonale* al profilo $\overline{\mathbf{c}}$ del baricentro. Così è per tutte le Q variabili attive i cui Q sottospazi generati hanno tutti in comune il profilo $\overline{\mathbf{c}}$. Di conseguenza la dimensione del sottospazio che contiene *tutta* la nuvola di J punti, è al più

$$J_1 + (J_2 - 1) + \ldots + (J_q - 1) + \ldots + (J_Q - 1) = J - Q + 1.$$

Altrettanti saranno quindi gli autovettori e gli autovalori. Di questi il più grande, quello corrispondente a $\overline{\mathbf{c}}$, vale 1.

Invece, nell'analisi rispetto al baricentro $\overline{\mathbf{c}}$ della nuvola, l'autovalore corrispondente a $\overline{\mathbf{c}}$ vale 0, per cui si troveranno al più

$$A = J - Q \tag{5.14.3}$$

autovalori non nulli. Geometricamente, ciò significa che la nuvola dei profili delle modalità è contenuta in un sottospazio di $\Re^I$ che ha al massimo dimensionalità $J-Q$ e che contiene anche il baricentro $\overline{\mathbf{c}}$. Tenendo conto che per la (5.13.5) l'inerzia complessiva dei profili è $(J-Q)/Q$, ne consegue che il valore medio degli autovalori è

$$\overline{\lambda}_a = \frac{J-Q}{Q\,(J-Q)} = 1/Q. \tag{5.14.4}$$

Dal punto di vista matematico, l'identificazione di una base in $\Re^I$ nel sottospazio ortogonale a $\overline{\mathbf{c}}$, si riconduce alla ricerca di autovalori e autovettori della matrice $\mathbf{R}^T\,\mathbf{C}$ di ordine $J \times J$ e di rango $A = J-Q$. Se, infatti, questo è il rango massimo di $\mathbf{C}$ per il ragionamento appena fatto, $\mathbf{C}$ ed $\mathbf{R}$ sono legate dalla (5.11.4) tramite le due matrici delle masse che sono diagonali con valori tutti positivi e quindi non singolari, e dunque anche il rango di $\mathbf{R}$ è quello di $\mathbf{C}$. Di conseguenza, il rango della matrice prodotto $\mathbf{C}\,\mathbf{R}^T$ non supera A.

Se $A = J-Q$ è il rango di $\mathbf{R}^T\,\mathbf{C}$, tale è anche il numero di autovalori non nulli λ_a e μ_a, ai quali, rispettivamente, corrispondono gli autovalori $\mathbf{u}_a^\star$ e $\mathbf{v}_a^\star$ negli spazi $\Re^I$ e $\Re^J$. Gli autovalori non nulli hanno valori inferiori a 1 e, nei casi pratici, mai coincidenti.

Le coordinate dei profili sugli assi fattoriali individuati dagli autovettori si ottengono, come nella Sez. 4.1 e 4.8, proiettando i profili sugli assi. Su ogni asse di rango $a = 1, 2, \ldots, A$, la coordinata di un profilo colonna $\mathbf{c}_j$ e riga $\mathbf{r}_i$, sono rispettivamente

$$g_{ja} \stackrel{\text{def}}{=} \mathbf{c}_j^T \mathbf{D}_{\overline{\mathbf{c}}}^{-1} \mathbf{u}_a^\star \qquad \text{e} \qquad f_{ia} \stackrel{\text{def}}{=} \mathbf{r}_i^T \mathbf{D}_{\overline{\mathbf{r}}}^{-1} \mathbf{v}_a^\star \tag{5.14.5}$$

elementi delle matrici

$$\mathbf{G} = \mathbf{C}^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-1}\,\mathbf{U}^* \qquad \text{e} \qquad \mathbf{F} = \mathbf{R}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-1}\,\mathbf{V}^* \tag{5.14.6}$$

di ordine $J \times A$ e $I \times A$ rispettivamente. Limitando le matrici delle coordinate alle prime $A^* < A$ colonne, si opera uno 'smoothing' della configurazione geometrica della nuvola, per cui è lecito affermare che l'ACM permette di

ottenere un ridotto numero A^* di fattori $\mathbf{g}_a$, ossia di variabili sintetiche di tipo numerico legate alle Q variabili attive. L'intensità del legame tra fattore (numerico) e variabili (categoriche) è dato dal rapporto di correlazione che verrà introdotto nella Sez. 5.18.

5.15 - Relazioni di transizione

Anche nell'ACM sussistono delle relazioni di transizione, analoghe alle (4.9.2), sia tra autovettori e fattori del medesimo rango

$$\begin{aligned} \mathbf{f}_a &= \sqrt{\lambda_a}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\mathbf{u}_a^\star \qquad & f_{ia} &= \left(\sqrt{\lambda_a}/\overline{c}_i\right)\, u_{ia}^\star \\ \mathbf{g}_a &= \sqrt{\lambda_a}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\mathbf{v}_a^\star \qquad & g_{ia} &= \left(\sqrt{\lambda_a}/\overline{r}_j\right)\, u_{ia}^\star. \end{aligned} \tag{5.15.1}$$

sia tra fattori del medesimo rango delle due nuvole, analoghe alle (4.9.3), le quali, tenendo conto delle (5.10.1) e (5.11.1), diventano

$$\begin{aligned} f_{ia} &= \frac{1}{\sqrt{\lambda_a}} \sum_j r_{ij}\, g_{ja} = \frac{1}{\sqrt{\lambda_a}}\,\frac{1}{Q} \sum_j g_{ja} \\ g_{ja} &= \frac{1}{\sqrt{\lambda_a}} \sum_i c_{ij}\, f_{ia} = \frac{1}{\sqrt{\lambda_a}}\,\frac{1}{z_{+j}} \sum_i f_{ia}. \end{aligned} \tag{5.15.2}$$

Nella prima relazione la somma va effettuata sulle sole Q modalità scelte dall'individuo i, mentre nella seconda limitatamente agli z_{+j} individui che hanno scelto la modalità j. In sostanza, la prima relazione indica che su un asse fattoriale, la posizione f_{ia} del profilo $\mathbf{r}_i$ è, a meno del coefficiente $1/\sqrt{\lambda_a}$ di dilatazione assiale, il baricentro (media *aritmetica*) delle modalità indicate dall'individuo i. La seconda, che il profilo $\mathbf{c}_j$ di una modalità si posiziona sull'asse, sempre a meno del coefficiente $1/\sqrt{\lambda_a}$, nel baricentro (media *aritmetica*) degli individui che l'hanno scelta. Di conseguenza, la posizione di una modalità su una mappa fattoriale rappresenta l'individuo 'medio' di un *segmento* di individui: dei maschi, delle femmine, dei laureati, ecc. Il fattore di dilatazione non è d'impaccio quando l'interpretazione viene fatta asse per asse.

I profili degli individui e delle modalità si trovano comunque in due spazi diversi, e quindi la distanza tra due modalità, quando queste sono intese come tali, misura la loro reciproca associazione, mentre quando sono intese come individui 'medii' è quella definita tra individui: due segmenti di individui j e j' risultano tanto più vicini quante più modalità simili hanno in comune. Nell'ACM, a meno del coefficiente di dilatazione, i due concetti

di distanza, anche se sono diversamente definiti, conducono alla medesima rappresentazione grafica.

Per concludere, anche se le relazioni di transizione (5.15.1) e (5.15.2) permettono una rappresentazione simultanea sugli assi fattoriali dei profili delle modalità *e* degli individui, fatto fondamentale per l'interpretazione dei risultati, in pratica, nell'ACM non vengono utilizzate perché i profili degli individui sono rappresentati o come segmenti, tramite le modalità, o come gruppi, tramite i baricentri dei gruppi. Se si dispone del software adatto, può essere interessante esaminare nello spazio dei primi tre fattori la configurazione della nuvola degli I punti che rappresentano i profili degli individui, per evidenziare eventuali zone di maggiore addensamento.

Le relazioni (5.15.2) non vengono utilizzate neppure per evitare l'analisi di una delle due nuvole di profili, perché si vedrà nella Sez. 5.19 che in pratica nell'ACM si analizza la matrice di Burt che ha dimensioni più ridotte.

5.16 - Codifica di variabili numeriche

Nell'Analisi delle Corrispondenze, sia semplici che multiple, ogni procedura che organizzi o rielabori i dati originali per adattarli al procedimento di analisi, è detta codifica. Vi rientra perciò anche la trasformazione di una variabile numerica in categorica mediante la suddivisione dell'intervallo di variazione dei suoi valori in classi, o sotto-intervalli contigui di valori, che definiscono altrettante modalità. Il risultato è una variabile categorica *ordinale* perché tra le modalità create sussiste un ordine di precedenza. Le variabili categoriche ordinali sono codificabili in forma disgiuntiva completa, come mostra la TAV. 5.8.

Il motivo che induce a codificare le variabili numeriche in classi contigue non vuote di valori è l'esigenza di avere tutte le variabili *attive*[1] in una forma comune, quella categorica, l'unica che l'ACM è in grado di analizzare. Ciò comporta da un lato una perdita di informazione sul valore della variabile, perché lo si sostituisce con la sua appartenenza a una classe di valori, dall'altro l'introduzione di una distanza artificiale tra due individui con valori della variabile che cadono da parti opposte dell'estremo di una classe. Un modo per limitare questo degrado dell'informazione è la codifica *sfumata*[2] che consiste essenzialmente nello 'spalmare' l'appartenenza di un

[1] La codifica non è necessaria se la variabile è illustrativa. Si veda la Sez. 5.17.

[2] È detta *fuzzy coding* in inglese e *codage flou* in francese.

valore ad una modalità su più modalità, invece di lasciarla concentrata in una sola. Il lettore interessato può consultare l'articolo di J. F. Gallego (1982) o il contributo di Yvette Grelet in Grangè e Lebart (1994), ma deve tener presente che tanto più si manipolano i dati iniziali, tanto più cauti si deve essere nell'interpretazione dei risultati della loro analisi.

Un altro aspetto importante è che l'ACM permette di evidenziare eventuali legami *non lineari* tra variabili numeriche trasformate in categoriche. Questo tipo di legame appare sovente quando le variabili rappresentano fenomeni che presentano una soglia intrinseca, per cui i profili delle modalità relative ai due estremi dell'intervallo di variazione risultano più vicini tra loro che non ai valori mediani [2] .

Per procedere alla codifica di una variabile numerica occorre definire sia il numero delle classi, sia i loro estremi. Questa separazione è piuttosto schematica perché le due scelte sono spesso effettuate simultaneamente.

Scelta del numero di classi di valori

Il numero di classi dipende molto dall'essere la variabile numerica attiva o illustrativa. Nel primo caso, se si opta per poche classi si corre il rischio di riunire nella stessa classe individui con valori molto diversi della variabile numerica, perdendo così gran parte dell'informazione di dettaglio. In tal caso le modalità raggruppano situazioni molto varie e la loro analisi non può far altro che mettere in luce fenomeni molto generici.

D'altro canto, se si decide per molte classi c'è il rischio che alcune contengano pochi individui e siano quindi poco rappresentative perché legate a fenomeni puntuali. Questo rischio si riduce se il numero I di individui è elevato. Possono sorgere però altri inconvenienti. Intanto, come si è visto nella Sez. 5.13, affinché le inerzie delle variabili attive non risultino squilibrate, occorre che queste ultime abbiano un numero paragonabile di modalità. Poi, nella Sez. 5.14 si è visto che ogni variabile interviene nell'analisi con sottospazi di dimensione $J_q - 1$, generati dalle sue J_q modalità. Aumentando J_q , cresce il numero di fattori su cui la variabile può influire, peggiorando l'aspetto

[2] Eventuali legami non lineari tra varibili numeriche restano invisibili in una Analisi nelle Componenti Principali, la quale tiene conto esclusivamente dei legami lineari. La giustificazione di questo fatto paradossale che riducendo l'informazione disponibile, come sempre avviene passando dal continuo al discreto, si aumenta la ricchezza del risultato può trovarsi nella Sez. 4.5.1 di Escofier e Pagès, (1998), *Analyses factorielles simples et multiples*, Dunod ed., 3a ediz.

sintetico dell'ACM, ossia la sua capacità di creare poche variabili numeriche sintetiche, riassuntive di quelle originali.

A titolo indicativo, il numero di classi di una variabile attiva dovrebbe collocarsi tra 4 e 5 quando gli individui sono un centinaio, per arrivare eventualmente a 7 e 8 con qualche migliaio.

Quando invece, la variabile numerica viene impiegata come illustrativa, questi vincoli sono meno stringenti, dato che la variabile non interviene attivamente nella costruzione degli assi fattoriali. può essere allora utile mantenere una suddivisione in molte classi.

Scelta degli estremi delle classi

Per prima cosa occorre chiedersi se la variabile da codificare ha delle soglie naturali, o tradizionali, che ne scandiscano le variazioni, come può essere l'età di pensionamento in una indagine socio-demografica in cui 'Età dell'intervistato' sia una variabile attiva.

Dopo di ciò, è assolutamente *indispensabile* costruire un istogramma piuttosto fine dei valori della variabile. Le irregolarità e le 'valli' della ripartizione suggeriscono possibili sezionamenti dell'intervallo di variazione.

Se poi i due criteri precedenti non suggeriscono alcuna soglia, conviene suddividere l'intervallo di variazione in classi comprendenti un *egual numero* z_{+j} di individui, piuttosto che in classi di eguale (o quasi) ampiezza. Questa suddivisione evita la presenza di modalità scelte da pochi individui, ossia con z_{+j} piccolo, di cui si è visto, per la (5.13.3), il grave effetto perturbativo dovuto essenzialmente al fatto che la distanza di una modalità dal baricentro cresce se la numerosità z_{+j} diminuisce. Non va poi dimenticato che le modalità rappresentano interi segmenti che meglio si confrontano se hanno numerosità comparabili.

In conclusione, occorre sempre trovare un compromesso tra una ripartizione sensata ed una tecnicamente ottimale.

5.17 - Profili illustrativi

I profili illustrativi giocano un ruolo molto importante nell'ACM. I motivi che inducono ad utilizzarli sono sostanzialmente gli stessi elencati nella Sez. 4.12: arricchire il contenuto delle mappe fattoriali, studiare le connessioni tra i diversi temi rilevati nell'indagine, 'spiegare' il comportamento di segmenti o di gruppi omogenei di intervistati, ecc. I profili illustrativi di individui o di modalità di una variabile vengono semplicemente proiettati

sulle mappe fattoriali già determinate dai profili attivi. Così nell'indagine sull'ascolto delle radio della Sez. 5.3, interessava indagare le relazioni tra le 4 variabili attive del tema E e le 11 variabili illustrative dei temi A, C e D. Si voleva, ad esempio, vedere quali erano le relazioni empiriche tra la struttura socio-demografica degli ascoltatori intervistati ed i loro atteggiamenti verso la comunicazione pubblicitaria (tema D).

Come ogni individuo attivo, anche l'individuo illustrativo deve aver risposto a tutte le Q domande del tema attivo, per cui il suo profilo $\tilde{\mathbf{r}}$ è un vettore di ordine J che viene costruito assegnando peso $\tilde{r}_j = 1/Q$ alle Q modalità che ha scelto per rispondere alle Q domande e peso nullo alle altre $J - Q$: esattamente come si è fatto per i profili attivi. Per esempio, nell'indagine sull'ascolto radiofonico il profilo illustrativo di un individuo è un vettore con $J = 21$ componenti, ed essendo $Q = 4$ le domande del tema attivo, le 4 componenti corrispondenti alle modalità scelte hanno valore $1/4$, mentre le altre 17 sono nulle.

Dato che di solito gli individui sono anonimi, è raro che siano considerati illustrativi singolarmente. più spesso, sono interi segmenti di individui che vengono considerati illustrativi tramite la modalità che li caratterizza e che ne rappresenta l'individuo medio.

Una variabile illustrativa deve essere stata proposta come domanda a tutti gli I individui attivi. Il profilo $\tilde{\mathbf{c}}$ di una sua modalità è un vettore di ordine I che, come per le modalità attive, si costruisce assegnando peso $\tilde{c}_i = 1/\tilde{z}_{+j}$ alle componenti che corrispondono agli $\tilde{z}_{+j}$ individui che la hanno scelta e peso nullo alle restanti. Per esempio, degli $I = 400$ intervistati $\tilde{z}_{+j} = 345$ hanno indicato la modalità illustrativa 'D4.1 - Sì, la musica facilita il ricordo degli spot'.

Le coordinate fattoriali $\tilde{f}_a$ e $\tilde{g}_a$ del profilo illustrativo di una riga e di una colonna sull'asse di rango a, ove $a = 1, 2, \ldots, A$, sono legate ai fattori delle righe f_{ia} e delle colonne g_{ja}, ottenuti dall'analisi dei profili attivi, tramite relazioni analoghe alle (4.12.1),

$$\tilde{f}_a = \frac{1}{\sqrt{\lambda_a}} \sum_{j=1}^{J} \tilde{r}_j \, g_{ja} = \frac{1}{\sqrt{\lambda_a}} \frac{1}{Q} \sum_{j} g_{ja} \tag{5.17.1}$$

$$\tilde{g}_a = \frac{1}{\sqrt{\lambda_a}} \sum_{i=1}^{I} \tilde{c}_i \, f_{ia} = \frac{1}{\sqrt{\lambda_a}} \frac{1}{\tilde{z}_{+j}} \sum_{i} f_{ia} \tag{5.17.2}$$

dove, l'ultima somma è effettuata nella (5.17.1) sulle sole modalità indicate dall'individuo illustrativo e nella (5.17.2) sui soli individui che hanno scelto la modalità illustrativa.

Le espressioni qui sopra sono analoghe alle relazioni di transizione (5.15.2) per i fattori dei profili attivi: la coordinata fattoriale di un individuo (di una modalità) attiva o illustrativa si ottiene dal prodotto del coefficiente di dilatazione $1/\sqrt{\lambda_a}$ per la media *aritmetica* delle coordinate fattoriali delle modalità di risposta scelte dall'individuo (degli individui che hanno indicato quella modalità).

Le variabili *numeriche* devono essere trasformate necessariamente in categoriche per partecipare come attive all'analsi, ma possono essere impiegate direttamente, ossia *senza* trasformazione, se vi partecipano come illustrative. In tal caso, della variabile vanno calcolati valor medio $\tilde{c}$ e deviazione standard campionaria $\tilde{s} = +\sqrt{s^2}$

$$\tilde{c} = \frac{1}{I}\sum_{i=1}^{I} \tilde{c}_i \qquad\qquad \tilde{s}^2 = \frac{1}{I}\sum_{i=1}^{I} (\tilde{c}_i - \tilde{c})^2.$$

Ad ogni valore $\tilde{c}_i$ della variabile va sottratta la media e il risultato va poi diviso per la deviazione standard, ottenendo una variabile $\hat{\tilde{\mathbf{c}}}$ con i valori standardizzati: valor medio nullo e varianza unitaria. L'ascissa della variabile $\hat{\tilde{\mathbf{c}}}$ sull'asse fattoriale $\mathbf{u}_a^\star$ si ottiene per proiezione tramite la (5.14.5), effettuando cioè il prodotto scalare tra tra variabile e l'asse. Grazie alla (5.15.1) si ottiene

$$\tilde{g}_{ja} = \hat{\tilde{\mathbf{c}}}^T\, \mathbf{D}_{\tilde{\mathbf{c}}}^{-1}\, \mathbf{u}_a^\star = \hat{\tilde{\mathbf{c}}}^T\, \frac{\mathbf{f}_a}{\sqrt{\lambda_a}} = \sum_{i=1}^{I} \left(\frac{\tilde{c}_i - \tilde{c}}{\tilde{s}}\right) \left(\frac{f_{ia} - 0}{\sqrt{\lambda_a}}\right) = I\, corr\, \left(\hat{\tilde{\mathbf{c}}},\, \hat{\mathbf{f}}_a\right).$$

L'ascissa della variabile numerica illustrativa standardizzata è quindi il coefficiente di correlazione lineare tra essa e il fattore standard di rango a, moltiplicato per I, il numero complessivo di individui attivi. Si ricordi che i fattori di qualunque rango hanno valore medio nullo per costruzione.

Su una mappa fattoriale la posizione di una variabile numerica illustrativa indica la *direzione* ove si trovano i valori più elevati della variabile. Questo è tanto più vero quanto più la proiezione si avvicina al cerchio delle correlazioni di raggio unitario. In tal caso esiste uno stretto legame tra variabile illustrativa e fattori dei profili.

Di solito, tuttavia, una variabile numerica illustrativa viene trasformata in categorica ordinale per poter evidenziare eventuali legami di tipo non lineare con altre variabili, sia attive che illustrative.

5.18 - Contributi, qualità e valori test

Il *contributo relativo* di una modalità attiva j all'inerzia dell'asse di rango a è definito, come nella (4.5.2), come rapporto tra l'inerzia riferita all'origine della proiezione g_{ja} del suo profilo sull'asse e l'inerzia complessiva λ_a sull'asse. Per la (5.9.3) si ha quindi

$$CTR_a(\mathbf{c}_j) \stackrel{\text{def}}{=} \frac{\overline{r}_j\, g_{ja}^2}{\lambda_a} = \frac{z_{+j}}{I\,Q}\,\frac{g_{ja}^2}{\lambda_a}. \tag{5.18.1}$$

Si tratta dell'indicatore più importante perché misura lo 'sforzo' fatto dalla modalità per attirare a sé l'asse e permette di evidenziare le modalità più efficaci da questo punto di vista. Per esso valgono tutte le proprietà elencate nella Sez. 4.5 nel caso delle Corrispondenze semplici.

Utile per interpretare un asse fattoriale è anche il contributo relativo di una variabile q, calcolato come somma dei contributi relativi (5.18.1) delle sue J_q modalità, dato che sono esclusive,

$$CTR_a(q) \stackrel{\text{def}}{=} \sum_{j'=1}^{J_q} CTR_a(\mathbf{c}_{j'}) = \frac{1}{Q\,\lambda_a} \sum_{j'=1}^{J_q} \frac{z_{+j'}}{I}\, g_{j'a}^2. \tag{5.18.2}$$

Il contributo relativo di tutte le Q variabili attive vale

$$\sum_{q=1}^{Q} CTR_a(q) = \frac{\lambda_a}{\lambda_a} = 1$$

come si verifica subito per le $Q = 4$ variabili della TAV. 5.10, ove i contributi relativi sono moltiplicati per 100, che, per il primo asse forniscono

$$CTR_1(E_1) + CTR_1(E_2) + CTR_1(E_3) + CTR_1(E_4) = \\ 2.2 + 33.0 + 31.2 + 33.6 = 100.$$

Dalla (5.18.2) si ricava che il contributo di una variabile varia tra 0 e $1/Q$, per cui la capacità di una variabile ad orientare un asse fattoriale è abbastanza limitata, dato che le variabili attive raramente sono meno di sette o otto. Ciò nonostante, il confronto dei contributi permette di individuare le variabili che hanno contribuito maggiormente a definire l'asse e di dare quindi ad esso un significato.

Nell'ACM il contributo relativo di una variabile attiva è legato ad una importante grandezza. Infatti la (5.18.2) si può scrivere

$$Q\,\lambda_a\, CTR_a(q) = \frac{\sum_{j'=1}^{J_q} \frac{z_{+j'}}{I}\,(\sqrt{\lambda_a}\, g_{j'a} - 0)^2}{\lambda_a}. \tag{5.18.3}$$

Ora, per la relazione di transizione (5.15.2), $\sqrt{\lambda_a}\, g_{j'a}$ è la coordinata media ponderata (baricentro) delle coordinate del segmento di individui che ha scelto la modalità di risposta j' tra le J_q possibili, e di conseguenza il numeratore del secondo membro nella (5.18.3) non è altro che l'inerzia complessiva di questi J_q baricentri locali rispetto al baricentro comune, l'origine degli assi fattoriali, ponderata con $z_{+j'}/I$, frazione di individui che hanno scelto proprio la modalità j' o importanza relativa del segmento j'. In Statistica questa inerzia è detta *inerzia tra i segmenti* stabiliti dalle J_q modalità esclusive della variabile q. Invece, il denominatore è la *inerzia totale* delle coordinate sull'asse di *tutti* gli I individui, indipendentemente dal segmento di appartenenza. Il rapporto al secondo membro della (5.18.3) è detto *rapporto di correlazione* di Pearson [1] tra variabile categorica q e fattore $\mathbf{f}_a$ che è una variabile numerica. Per traslato, si parla di rapporto di correlazione tra la variabile e l'asse fattoriale $\mathbf{u}_a^\star$. Il rapporto di correlazione viene indicato tradizionalmente con $\eta_a^2(q)$ e varia tra 0 ed 1. Quando è vicino a 0, caso di perfetta ininfluenza di q su $\mathbf{u}_a^\star$, i baricentri dei segmenti sull'asse sono praticamente coincidenti col baricentro comune che è all'origine dell'asse, mentre le coordinate in ogni segmento sono molto disperse. Questo significa che il legame tra le J_q modalità e le coordinate, o in altri termini tra la variabile q e l'asse $\mathbf{u}_a^\star$, è debole. Al contrario, quando $\eta_a^2(q)$ è vicino ad 1, caso di perfetta influenza di q su $\mathbf{u}_a^\star$, le coordinate degli individui di uno stesso segmento sono piuttosto addensate sull'asse ed i segmenti sono ben separati: il legame tra variabile ed asse fattoriale è allora piuttosto stretto.

Tenendo conto che per la (5.9.5) $\overline{r}_q = 1/Q$, il contributo relativo della variabile attiva q all'inerzia dell'asse di rango a dalla (5.18.3) prende la forma più familiare, analoga alla (5.18.1)

$$CTR_a(q) = \frac{\overline{r}_q\, \eta_a^2(q)}{\lambda_a} \tag{5.18.4}$$

per cui $\eta_a^2(q)$ può essere intesa come la 'coordinata' della variabile q sull'asse fattoriale di rango a. Questo suggerisce di rappresentare le variabili su una mappa fattoriale, utilizzando i valori di $\eta_a^2(q)$ come coordinate. Per esempio, dalla TAV. 5.9 e dalla TAV. 5.10 si vede che la prima variabile attiva dell'esempio, 'E1 = Sesso dell'intervistato', ha come coordinate sui primi due

---

[1] Karl Pearson: Londra 1857, Cambridge, Londra, Coldharbour nel Surrey 1936.

assi

$$\eta_1^2(E_1) = \lambda_1\, Q\, CRT_1(E_1) = 0.5675 \times 4 \times 2.2/100 = 0.05$$
$$\eta_2^2(E_1) = \lambda_2\, Q\, CRT_2(E_1) = 0.4636 \times 4 \times 5.9/100 = 0.11.$$

La TAV. 5.14 riporta la posizione delle 4 variabili attive dell'esempio sulla mappa principale e ne evidenzia i legami coi due assi.

Se il contributo della (5.18.4) viene sommato su tutte le variabili attive, si ottiene

$$\sum_{q=1}^{Q} \lambda_a\, CTR_a(q) = \sum_{q=1}^{Q} \overline{r}_q\, \eta_a^2(q)$$
$$\lambda_a = \sum_{q=1}^{Q} \overline{r}_q\, \eta_a^2(q) \qquad (5.18.5)$$

per cui l'inerzia sull'asse è la media ponderata dei rapporti di correlazione delle variabili. Ma, poiché l'ACM individua gli assi di massima inerzia, si può dire che questi sono anche quelli maggiormente legati alle variabili attive. In altri termini, i fattori, ossia l'insieme delle coordinate fattoriali degli individui sull'asse, sono le variabili numeriche più legate all'insieme delle Q variabili categoriche attive e quindi quelle che in questo senso meglio le sintetizzano. Questa importante proprietà dell'ACM verrà sfruttata quando si tratterà di creare gruppi di individui con carattristiche simili, come si vedrà nel prossimo Capitolo 6.

Infine, un'ultima osservazione che si può trarre dalle considerazioni appena fatte. Si è visto che per la (5.13.4) l'inerzia di una variabile q cresce linearmente col numero J_q delle sue modalità. Di conseguenza, cresce anche il numero di assi fattoriali sul cui orientamento la variabile può influire. Perciò, se due o più variabili hanno parecchie modalità e sono legate tra loro, i primi assi fattoriali esprimono principalmente questo legame ed è necessario esaminare assi e mappe di rango superiore per evidenziare eventuali legami tra altre variabili. Si vedrà che questo è il caso delle variabili E_2 'Età dell'intervistato' con 7 modalità ed E_4 'Qualifica professionale' con 8 modalità, che dalla mappa nella TAV. 5.14 risultano legate fra loro. La mappa principale di TAV. 5.15 mette in luce quasi esclusivamente il loro legame.

Nell'ACM è utile considerare anche la *percentuale d'inerzia* su un asse di rango a di una *variabile* q, rapportata a quella delle sue J_q modalità. La (5.13.1) indica che l'inerzia riferita all'origine di una qualunque di queste vale $In_{\mathbf{0}}(\mathbf{c}_j) = 1/Q$. Dato che i J_q profili sono ortogonali due a due, come

mostra la (5.11.2), l'inerzia delle *proiezioni* dei profili delle modalità su uno qualunque dei J_q vettori è sempre eguale a $1/Q$. Si ricordi che tra i J_q vettori c'è anche quello che collega l'origine $\mathbf{0}_{\mathrm{I}}$ col baricentro $\overline{\mathbf{c}}$, come si è visto nella Sez. 5.14. Di conseguenza, l'inerzia delle proiezioni dei J_q profili su un autovettore $\mathbf{u}_a^\star$ di $\Re^I$ sarà eguale al prodotto di $1/Q$ per il coseno quadrato $\cos^2\theta_a$ dell'angolo tra l'autovettore ed il sottospazio individuato dai J_q profili. L'angolo θ_a dipende sia dall'orientamento dell'asse fattoriale, sia da quello del sottospazio che contiene i J_q profili. Si può così definire la percentuale d'inerzia $\tau_a(q)$ di una *variabile* q sull'asse fattoriale di rango a, rapportata all'inerzia $In_{\overline{\mathbf{c}}}(q)$ nella (5.13.4) dei *soli* J_q profili, come

$$\tau_a(q) \stackrel{\text{def}}{=} \cos^2\theta_a \, \frac{1}{Q} \Big/ \frac{J_q - 1}{Q} \times 100 = \frac{\cos^2\theta_a}{J_q - 1} \times 100. \tag{5.18.6}$$

Questa espressione mostra che anche $\tau_a(q)$ non dipende dalla configurazione geometrica dei profili e che quando una variabile è molto legata a un asse, per cui l'angolo θ_a nella (5.18.6) è piccolo, se le sue modalità sono numerose, la percentuale di inerzia su un asse è piccola, in particolare sui primi assi. Per questo motivo non bisogna mai eccedere col numero di modalità, anche quando gli individui sono numerosi. Si può poi facilmente mostrare che $\cos^2\theta_a = \eta_a^2(q)$, il rapporto di correlazione incontrato poco sopra[1]. L'eguaglianza, oltre a chiarire ulteriormente il significato del rapporto di correlazione, permette di calcolare rapidamente $\tau_a(q)$.

Anche le variabili illustrative possono essere indicate su una mappa, calcolandone la coordinata su un asse tramite un'espressione analoga alla (5.18.3)

$$\tilde{\eta}_a^2(\tilde{q}) = \frac{\sum_{j'=1}^{J_q} \frac{\tilde{z}_{+j'}}{I} \left(\sqrt{\lambda_a}\, \tilde{g}_{j'a} - 0\right)^2}{\lambda_a} = \frac{1}{I} \sum_{j'=1}^{J_q} \tilde{z}_{+j'}\, \tilde{g}_{j'a}^2$$

dove $\tilde{q}$ è la variabile illustrativa a J_q modalità, $\tilde{z}_{+j'}$ sono gli individui che hanno scelto la modalità illustrativa j' tra le J_q e $\tilde{g}_{j'a}$ il baricentro del segmento illustrativo, come nella (5.17.2). Il calcolo è ora più laborioso dato che non esiste un contributo relativo per le variabili illustrative.

Anche la *qualità* della rappresentazione di un profilo su un asse è

---

[1] Basta confrontare il contributo relativo della variabile all'inerzia dell'asse, espresso in funzione dell'angolo, $CTR_a(q) = (cos^2\theta_a/Q)/\lambda_a$ con la (5.18.4).

definita in modo analogo alla (4.6.1) come

$$COS^2_a(\mathbf{c}_j) \stackrel{\text{def}}{=} \frac{g^2_{ja}}{\sum_{a=1}^{A} g^2_{ja}} \tag{5.18.7}$$

e non dipende dalla massa del profilo. Per questo indicatore valgono tutte le proprietà riportate nella Sez. 4.6 e può anche essere calcolato per ogni singola variabile, sia attiva che illustrativa, anche se in questo caso non è interpretabile geometricamente come quadrato di un coseno.

Valori test delle modalità illustrative

Il software d'analisi SPADn, specifico per l'analisi dei dati raccolti tramite sondaggio e che verrà presentato nella Sez. 5.22, utilizza estesamente un indicatore alternativo per valutare se il profilo di una modalità illustrativa occupa una posizione non casuale su un asse fattoriale. È il *Valore-test*, un indice che sostanzialmente misura la 'significtività' della distanza tra la coordinata del profilo sull'asse e l'origine dell'asse. Lo schema è quello dei classici test di ipotesi in cui l'ipotesi di base (o nulla) traduce la situazione 'priva di interesse'. più la situazione osservata (la coordinata del profilo sull'asse) è lontana dalla situazione 'priva di interesse' (la coordinata che si avrebbe sotto l'ipotesi di base), maggiore è la possibilità che sia interessante per illustrare l'asse. Il Valore-test misura questa distanza, mentre la compatibilità si valuta in termini di 'probabilità critica', ossia della probabilità di ottenere, sotto l'ipotesi di base, una coordinata così estrema come quella che si è osservata. Tanto più la probabilità critica è piccola (per esempio inferiore al 5%) tanto più la posizione del profilo illustrativo è interessante, nel senso che è poco probabile che sia dovuta al caso.

Punto di partenza è la relazione di transizione (5.17.2) che mostra come la coordinata fattoriale $\tilde{g}_a$ di una modalità illustrativa sull'asse fattoriale di rango a , sia , a meno di un coefficiente, la media aritmetica $\overline{f}_a$ delle coordinate fattoriali f_{ia} dei *soli* $\tilde{z}_{+j}$ individui che, tra gli I , hanno scelto proprio quella modalità, per cui si può scrivere

$$\tilde{g}_a = \frac{1}{\sqrt{\lambda_a}}\,\overline{f}_a \qquad \text{dove} \qquad \overline{f}_a = \frac{1}{\tilde{z}_{+j}} \sum_i f_{ia}. \tag{5.18.8}$$

Si procede poi *come* in un test statistico, prendendo ora in considerazione le coordinate fattoriali f_{ia} di *tutti* gli I individui attivi come popolazione di riferimento. Il valore medio e la varianza di questa popolazione di coordinate,

in base ai risultati della Sez. 4.8, sono

$$\sum_{i=1}^{I} \overline{c}_i \, f_{ia} = 0 \qquad\qquad \sum_{i=1}^{I} \overline{c}_i \, (f_{ia} - 0)^2 = \lambda_a.$$

Si ricordi che il profilo $\overline{\mathbf{c}}$ delle masse è piatto e che la somma delle sue I componenti $\overline{c}_i$ è 1.

Da questa popolazione finita di I coordinate si estraggono casualmente e senza reimmissione, $\tilde{z}_{+j}$ coordinate, indicate con F_{ia} , e se ne calcola la media aritmetica

$$\overline{F}_a = \frac{1}{\tilde{z}_{+j}} \sum_{i=1}^{\tilde{z}_{+j}} F_{ia}. \tag{5.18.9}$$

Si noti la differenza sostanziale tra questa media e quella della (5.18.8): anche se in entrambe il numero di elementi è il medesimo. Nella prima somma compaiono esclusivamente coordinate di individui che hanno scelto la modalità j , nella seconda possono essere presenti *anche* coordinate di individui che *non* hanno scelto la modalità. Per questo si sono utilizzati simboli differenti per indicare sia le coordinate che le medie.

Nelle estrazioni senza reimmissione, la probabilità di inclusione di ogni elemento nel campione f_{ia} resta costante in ogni estrazione, Kish (1965), pag. 40 [1] . Il valore atteso di ogni coordinata estratta F_{ia} coincide col valore medio della popolazione, fatto questo non molto evidente a prima vista, per cui $P(f_{ia}) = 1/I$ e quindi

$$E\,(F_{ia}) = \sum_{i=1}^{I} P(f_{ia})\, \overline{c}_i \, f_{ia} = \frac{1}{I} \sum_{i=1}^{I} \overline{c}_i \, f_{ia} = 0. \tag{5.18.10}$$

Di conseguenza, è nullo anche il valore atteso della media delle coordinate estratte casualmente nella (5.18.9)

$$E\left(\overline{F}_a\right) = \frac{1}{\tilde{z}_{+j}} \, E\left(\sum_{i=1}^{\tilde{z}_{+j}} F_{ia}\right) = \frac{1}{\tilde{z}_{+j}} \sum_{i=1}^{\tilde{z}_{+j}} E\,(F_{ia}) = 0. \tag{5.18.11}$$

Si può adesso calcolare, tramite la (5.18.8), la posizione $\tilde{G}_a$ che avrebbe sull'asse la modalità supplementare *se* le $\tilde{z}_{+j}$ coordinate degli individui fos-

---

[1] Il testo di Leslie Kish, *Survey sampling*, (1965) Wiley ed., resta un classico sull'argomento. Si veda anche L. Fabris *L'indagine campionaria* (1989), NIS, Roma, a pag. 55.

sero state estratte casualmente. Il suo valore atteso è nullo per la (5.18.11)

$$\tilde{G}_a = \frac{1}{\sqrt{\lambda_a}}\,\overline{F}_a \qquad E\left(\tilde{G}_a\right) = \frac{1}{\sqrt{\lambda_a}}\,E\left(\overline{F}_a\right) = 0. \tag{5.18.12}$$

Il Valore-test misura lo scarto tra la coordinata osservata $\tilde{g}_a$ e il valore che ci si deve attendere nell'ipotesi che le coordinate degli individui fossero disposte casualmente sull'asse, il tutto misurato in unità di deviazione standard, come una statistica di Student

$$\text{Valore} - \text{test} \stackrel{\text{def}}{=} \frac{\tilde{g}_a - E\left(\tilde{G}_a\right)}{\sqrt{VAR\left(\tilde{G}_a\right)}} \tag{5.18.13}$$

La varianza $VAR\left(\overline{F}_a\right)$ della media delle coordinate estratte casualmente senza reimmissione, è legata alla varianza λ_a della popolazione di riferimento, Kish (1965), tramite la cosidetta 'correzione per popolazione finita'

$$VAR\left(\overline{F}_a\right) = \frac{I - \tilde{z}_{+j}}{I - 1}\,\frac{\lambda_a}{\tilde{z}_{+j}}.$$

Le varianza della coordinata $\tilde{G}_a$ della modalità, sempre nell'ipotesi che le coordinate degli individui siano estratte casualmente, si deduce allora dalla prima delle (5.18.12)

$$VAR\left(\tilde{G}_a\right) = VAR\left(\frac{1}{\sqrt{\lambda_a}}\,\overline{F}_a\right) = \frac{1}{\lambda_a}\,VAR\left(\overline{F}_a\right) = \frac{I - \tilde{z}_{+j}}{I - 1}\,\frac{1}{\tilde{z}_{+j}}.$$

Finalmente, il Valore-test della (5.18.13) prende la forma

$$\text{Valore} - \text{test} = \tilde{g}_a \sqrt{\frac{I - 1}{I - \tilde{z}_{+j}}\,\tilde{z}_{+j}} \tag{5.18.14}$$

e misura, in unità di deviazioni standard, la distanza tra l'effettiva proiezione della modalità illustrativa sull'asse e l'origine [1]. Per il Teorema Centrale Limite, si può assumere che la distribuzione di probabilità del Valore-test sia la Normale standard, con valore atteso nullo e varianza unitaria. Di conseguenza, una modalità illustrativa è ritenuta 'significativa' se il suo Valore-test è superiore in valore assoluto a $1.96 \simeq 2$, corrispondente ad una probabilità critica inferiore al 5%.

[1] Se la variabile illustrativa ha due modalità, come la 'D3 - Qualità tecnica degli spot' nella TAV. 5.13, i due valori-test sono uno l'opposto dell'altro. Ciò è conseguenza del fatto che l'origine degli assi fattoriali è il baricentro anche delle modalità di una variabile illustrativa.

Si è lontani comunque dall'uso tradizionale del test statistico perché essendo le modalità illustrative abitualmente numerose, i loro Valori-test possono, in parte, risultare 'significativi' anche se in realtà non lo sono. È il problema che si incontra nei *test multipli*. Si veda al riguardo, per esempio, Hsu (1996) [2] . Comunque, grazie alla loro semplicità di calcolo, i Valori-test sono estesamente impiegati per produrre rapidamente un *elenco* delle modalità illustrative, ordinate per importanza decrescente nell'illustrare l'asse.

Il Valore-test ha senso soltanto per i profili delle modalità illustrative. Non sarebbe corretto costruire il Valore-test del profilo di una modalità attiva, perché i suoi z_{+j} individui hanno contribuito fattivamente all'orientamento dell'asse fattoriale: l'ipotesi di una loro estrazione casuale non avrebbe senso.

5.19 - Analisi dei profili di B.

La matrice simmetrica di Burt, definita nella Sez. 5.7 ed esemplificata nella TAV. 5.4, a ragione delle sue proprietà e delle ridotte dimensioni, gioca un ruolo importante nel calcolo dei fattori. In questa Sez. verrà mostrato come i risultati ottenuti nella Sez. 5.14. siano legati a quelli che si possono ottenere dall'analisi dei profili della matrice $\mathbf{B}$.

Si è visto nella Sez. 5.7 che l'elemento diagonale j , per una riga o colonna $j = 1, 2, \ldots, J$ di $\mathbf{B}$, vale $b_{jj} = z_{+j}$, per cui, tenendo presente la struttura a blocchi della matrice, il totale marginale risulta

$$b_{j+} = \sum_{j'=1}^{J} b_{jj'} = Q\, b_{jj} = Q\, z_{+j}$$

e, a causa della simmetria di $\mathbf{B}$, i totali marginali delle righe e delle colonne sono eguali

$$b_{j+} = b_{+j} = Q\, z_{+j}. \tag{5.19.1}$$

Sempre per il fatto che $\mathbf{B}$ è costituita da Q^2 blocchi, in ciascuno dei quali sono ripartiti gli I individui, il totale generale risulta

$$b_{++} = I\, Q^2. \tag{5.19.2}$$

È possibile ora definire il profilo marginale, la cui j -ma componente è definita come

$$\frac{b_{j+}}{b_{++}} \stackrel{\text{def}}{=} \frac{Q\, z_{+j}}{I\, Q^2} = \frac{z_{+j}}{I\, Q} = \overline{r}_j \tag{5.19.3}$$

---

[2] J. C. Hsu, *Multiple Comparison, Theory and Methods*, (1996) Chapman & Hall ed.

per cui la matrice diagonale delle masse

$$diag\ (\overline{r}_1, \overline{r}_2, \ldots, \overline{r}_J) = \mathbf{D}_{\overline{\mathbf{r}}}$$

coincide numericamente con quella definita nella Sez. 5.9 per le masse della matrice **C** dei profili delle colonne desunta dalla tabella di indicatori delle modalità codificati in forma disgiuntiva completa.

I profili di **B** sono definiti come nelle Corrispondenze semplici, ossia come rapporto tra ogni elemento e il totale della riga o della colonna, totali che per la (5.19.1) coincidono. Perciò le matrici dei profili delle righe $\mathbf{P}_r$ e delle colonne $\mathbf{P}_c$, come nella TAV. 5.7, sono

$$\mathbf{P}_r \stackrel{\text{def}}{=} \frac{1}{I\,Q^2}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\,\mathbf{B} \qquad\qquad \mathbf{P}_c \stackrel{\text{def}}{=} \frac{1}{I\,Q^2}\,\mathbf{B}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}. \tag{5.19.4}$$

Entrambe sono di ordine $J \times J$ e *non* sono simmetriche. Poiché invece **B** è simmetrica e $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$ diagonale, trasponendo le (5.19.4) si vede che tra le matrici dei profili sussiste la relazione

$$\mathbf{P}_r = \mathbf{P}_c^T \qquad \text{e quindi anche} \qquad \mathbf{P}_c = \mathbf{P}_r^T \tag{5.19.5}$$

risultato facilmente controllabile nella TAV. 5.7.

Nella Sez. 5.7 si è visto che gli elementi di una riga (o colonna) j della matrice di Burt sono il numero di righe con la modalità j della tabella di indicatori codificati in forma disgiuntiva completa. Dal punto di vista geometrico, questo significa che nello spazio $\Re^J$ ogni profilo $\mathbf{p}_j$ della matrice di Burt, è il baricentro dei profili $\mathbf{r}_i$ degli individui che l'hanno indicata. Inoltre, nello spazio $\Re^J$ la metrica è $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$ per i profili $\mathbf{r}_i$ e $(1/Q)\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$ per quelli $\mathbf{p}_j$ di Burt. Ne deriva che i profili degli individui e dei loro baricentri si trovano nel *medesimo* spazio $\Re^J$.

Dalle matrici $\mathbf{P}_r$ e $\mathbf{P}_c$ si ottengono i medesimi autovalori ed autovettori che si potrebbero ottenere dalle matrici dei profili **R** e **C** desunte dalla tavola degli indicatori delle modalità, codificati in forma disgiuntiva completa. Infatti, prendendo in esame l'equazione agli autovalori [1] (5.14.2),

$$\mathbf{R}^T\mathbf{C}\,\mathbf{v}_a = \mu_a\,\mathbf{v}_a \qquad \text{con} \qquad \mathbf{v}_a^T\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\,\mathbf{v}_a = 1 \tag{5.19.6}$$

ove la matrice $\mathbf{R}^T\mathbf{C}$ è di ordine $J \times J$, lo stesso di $\mathbf{P}_r$ e $\mathbf{P}_c$, e tenendo conto della relazione (3.11.2) $\mathbf{C} = \mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{R}\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$, la matrice da diagonalizzare si

---

[1] In questa Sezione verrà omesso l'* che identifica gli autovettori con origine nel baricentro della nuvola di profili.

può esprimere come

$$\mathbf{R}^T\mathbf{C} = \mathbf{R}^T\mathbf{D}_{\mathbf{c}}\mathbf{R}\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}} = \frac{1}{Q}\,\left[Q\,\mathbf{R}^T\right]\,\frac{1}{I}\,\left[I\,\mathbf{D}_{\mathbf{c}}\right]\,\frac{1}{Q}\,\left[Q\,\mathbf{R}\right]\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}$$
$$= \frac{1}{I\,Q^2}\,\left[Q^2\,\mathbf{R}^T\mathbf{R}\right]\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}} = \frac{1}{I\,Q^2}\,\mathbf{B}\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}} = \mathbf{P_c}$$

avendo tenuto conto della (5.9.8), della (5.10.2) e infine della (5.19.4). Le matrici $\mathbf{R}^T\mathbf{C}$ e $\mathbf{P}_c$ sono dunque equivalenti per cui la (5.19.6) si può scrivere

$$\mathbf{P}_c\,\mathbf{v}_a = \mu_a\,\mathbf{v}_a \qquad \text{col vincolo} \qquad \mathbf{v}_a^T\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,\mathbf{v}_a = 1. \tag{5.19.7}$$

Così la diagonalizzazione della matrice $\mathbf{P_c}$ dei profili delle colonne di Burt fornisce i medesimi autovettori ed autovalori della diagonalizzazione di $\mathbf{R}^T\mathbf{C}$ dei profili costruiti a partire dalla tabella di indicatori in forma disgiuntiva completa. La matrice $\mathbf{P_c}$ è semplice da calcolare a partire da $\mathbf{B}$ che, a sua volta si ottiene facilmente dalla codifica compatta di Sez. 5.4 ed è anche $\mathbf{D}$-simmetrica, il che permette l'impiego di routine di diagonalizzazione per matrici simmetriche, come indicato nella Sez. B.2 dell'Appendice B.

L'altro approccio che si può seguire per arrivare agli autovalori μ_a e agli autovettori $\mathbf{v}_a$ è quello di considerare la matrice di Burt come una matrice di contingenza procedendo direttamente all'Analisi delle Corrispondenze semplici dei suoi profili $\mathbf{P}_r$ e $\mathbf{P}_c$. In tal caso si può facilmente mostrare con un ragionamento analogo a quello fatto nella Sez. 3.9, che si perviene all'equazione, analoga formalmente alla (5.14.2),

$$\mathbf{P}_r^T\,\mathbf{P}_c\,{}_B\mathbf{v}_a = {}_B\mu_a\,{}_B\mathbf{v}_a \quad \text{ossia} \quad \mathbf{P}_c^2\,{}_B\mathbf{v}_a = {}_B\mu_a\,{}_B\mathbf{v}_a \tag{5.19.8}$$

grazie alla (5.19.5) e dove ${}_B\mu_a$ e ${}_B\mathbf{v}_a$ sono l'autovalore ed il corrispondente autovettore di rango a. Ma, quali sono i loro legami con i μ_a e $\mathbf{v}_a$ della (5.19.7), o, equivalentemente, della (5.19.6)? La risposta si ottiene subito premoltiplicando la (5.19.7) per $\mathbf{P}_c$ ottenendo, sempre per la (5.19.7),

$$\mathbf{P}_c\,\mathbf{P}_c\,\mathbf{v}_a = \mu_a\,\mathbf{P}_c\,\mathbf{v}_a = \mu_a\,\mu_a\,\mathbf{v}_a = \mu_a^2\,\mathbf{v}_a$$

per cui l'equazione diventa

$$\mathbf{P}_c^2\,\mathbf{v}_a = \mu_a^2\,\mathbf{v}_a$$

che, confrontata con la (5.19.8) rivela che le due diagonalizzazioni producono autovalori che stanno nelle relazioni

$${}_B\mu_a = \mu_a^2 \qquad \text{ossia} \qquad \mu_a = \sqrt{{}_B\mu_a}. \tag{5.19.9}$$

Gli autovettori ${}_B\mathbf{v}_a$ e $\mathbf{v}_a$ sono colineari, perché $\mathbf{P}_c^2$ non può non avere autovettori che non siano anche autovettori di $\mathbf{P}_c$ e quindi gli autovettori $\mathbf{v}_a$ e ${}_B\mathbf{v}_a$ sono identici. Non così i fattori dei profili, ossia le loro J ascisse sugli assi fattoriali di ogni rango $a = 1, 2, \ldots, A$, perché, per la (5.15.1) e per le relazioni appena trovate, si ha

$$ {}_B\mathbf{g}_a = \sqrt{{}_B\mu_a}\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,{}_B\mathbf{v}_a = \mu_a\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,\mathbf{v}_a = \sqrt{\mu_a}\,\sqrt{\mu_a}\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,\mathbf{v}_a = \sqrt{\mu_a}\,\mathbf{g}_a $$

mentre i fattori standard sono eguali in entrambe le analisi: ${}_B\hat{\mathbf{g}}_a = \hat{\mathbf{g}}_a$. Questi collegamenti sono di grande importanza perché a questo punto l'ACM, può essere definita indifferentemente come l'Analisi delle Corrispondenze semplici sia dei profili di $\mathbf{R}$ e di $\mathbf{C}$ ricavati dalla tabella di indicatori in forma disgiuntiva completa, sia dei profili $\mathbf{P}_r$ e $\mathbf{P}_c$ della matrice di Burt. Tutto ciò rivela che l'ACM non è pienamente multidimensionale, ma *congiuntamente bivariata*, nel senso che vengono prese in esame congiuntamente soltanto le associazioni tra le *coppie* di modalità delle variabili attive, ma non quelle di ordine superiore.

Nell'analisi di $\mathbf{B}$ sono rimasti esclusi i profili degli individui. Le loro coordinate fattoriali si possono ottenere considerando la matrice dei profili $\mathbf{R}$ come illustrativa. L'ascissa ${}_B\tilde{f}_a$ di un individuo illustrativo sull'asse fattoriale individuato dall'autovettore ${}_B\mathbf{v}_a$ si ottiene per proiezione, come nella Sez. 4.12, e, tenendo conto dei risultati appena ottenuti, risulta

$$ {}_B\tilde{f}_a = \frac{1}{\sqrt{{}_B\mu_a}}\,\frac{1}{Q}\,\sum_j {}_Bg_{ja} = \frac{1}{\mu_a}\,\frac{1}{Q}\,\sum_j \sqrt{\mu_a}\,g_{ja} = \frac{1}{\sqrt{\mu_a}}\,\frac{1}{Q}\,\sum_j g_{ja} = \tilde{f}_a $$

ove la somma è limitata alle Q modalità scelte dall'individuo. La coordinata ${}_B\tilde{f}_a$ è quindi identica a $\tilde{f}_a$ della (5.17.1) che si otterrebbe considerando il profilo $\mathbf{r}_i$ come illustrativo nell'analisi dei profili ricavati dalla tabella di indicatori.

5.20 - Interpretazione dei risultati

In questa Sezione verrà mostrato come leggere e interpretare i risultati di una ACM, prendendo come esempio pedagogico il sondaggio sull'ascolto delle trasmissioni radiofoniche della Sez. 5.3. La sequenza delle operazioni ricalca per gran parte quella dell'Analisi delle Corrispondenze semplici: esame delle inerzie per stabilire quanti assi caratterizzare; individuazione, tramite le coordinate e i contributi relativi, delle variabili e delle modalità che meglio descrivono gli assi; esame della collocazione sulle mappe dei profili delle modalità

attive e poi delle modalità illustrative e infine costruzione delle mappe dei profili degli individui per evidenziare la forma della loro nuvola.

1 - Esame delle inerzie

Inerzie sugli assi

La ripartizione delle inerzie (autovalori) sugli assi fattoriali è tradizionalmente il primo risultato stampato da un programma d'analisi. Il fatto che le inerzie sugli assi si possano ottenere, pur con fattori di dilatazione diversi, sia dall'analisi dei profili ricavati dalla tabella di indicatori sia dalla matrice di Burt, sta ad indicare che inerzie e percentuali d'inerzia sono elementi poco indicativi nell'ACM. La TAV. 5.9 si riferisce all'indagine sull'ascolto delle trasmissioni radio in cui sono considerate attive le variabili del tema E che delinea il profilo socio-demografico dell'ascoltatore. In questo caso il numero A di autovalori non nulli, differenza tra numero di modalità e di variabili, è $A = J - Q = 21 - 4 = 17$. Questa è dunque la dimensionalità degli iperspazi che contengono le due nuvole di profili.

L'inerzia complessiva non è legata alla struttura dei dati, come nelle Corrispondenze semplici, ma dipende unicamente dal numero di variabili e di modalità attive: $In_{\overline{\mathbf{r}}} = In_{\overline{\mathbf{c}}} = J/Q - 1 = 4.25$. Questa inerzia si ripartisce sugli assi fattoriali o, come anche si dice, viene 'estratta' dagli assi fattoriali. Il diagramma di TAV. 5.9 mostra come all'aumentare del rango a dell'asse, le inerzie λ_a sull'asse si attenuino regolarmente e lentamente con il caratteristico andamento 'a mensola' che è peculiare dell'ACM. E' raro che le inerzie non siano piccole. L'inerzia media su un asse, per la (5.14.4), è $\overline{\lambda} = 1/Q$. Spesso questo valore è preso come 'soglia' empirica per stabilire il numero A^* di assi interpretabili. Nel caso dell' esempio si dovrebbero esaminare i primi $A^* = 8$ perché hanno un'inerzia superiore a $\overline{\lambda} = 1/4 = 0,25$, ma in questa sezione l'esame sarà limitato ai primi due.

Percentuali d'inerzia sugli assi

La percentuale d'inerzia sull'asse fattoriale di rango a è definita come rapporto tra l'inerzia delle proiezioni dei J profili delle modalità sull'asse e l'inerzia totale della nuvola,

$$\tau_a \stackrel{\text{def}}{=} \frac{\lambda_a}{In_{\overline{\mathbf{c}}}} \times 100 = \frac{\lambda_a}{\sum_{a=1}^{A} \lambda_a} \times 100. \qquad (5.20.1)$$

L'espressione è la stessa delle Corrispondenze semplici, ove però le percentuali d'inerzia erano preziosi indicatori. Invece qui, per quanto visto poco sopra,

le percentuali d'inerzia sono degli indicatori pessimistici della quota di informazione sulla struttura geometrica estratta dagli assi. In altri termini, l'importanza dei primi assi è certamente superiore a quanto lasciano intendere i valori di τ_a, e quindi superiore al $68.53\,\%$ 'estratto' nell'esempio dai primi 8 assi fattoriali.

Stabilito il numero di assi da considerare e controllato con cura l'andamento delle inerzie sugli assi, si passa ad esaminare due indicatori relativi alle *variabili:* la percentuale d'inerzia $\tau_a(q)$ sugli assi fattoriali, che è riferita all'inerzia totale di *tutti* i J profili e il rapporto di correlazione $\eta^2_a(q)$, che è invece riferito all'inerzia dei *soli* J_q profili delle modalità della variabile. Entrambi sono stati descritti nella Sez. 5.18.

Il prospetto sottostante presenta la situazione nel caso dell'esempio

q	*Variabile*	J_q	$In_{\overline{c}}(q)$	$\eta^2_1(q)$	$\tau_1(q)\,\%$	$\eta^2_2(q)$	$\tau_2(q)\,\%$
1	E_1 : Sesso	2	0.25	0.05	5.0	0.11	11.0
2	E_2 : Età	7	1.50	0.75	12.5	0.74	12.3
3	E_3 : Titolo	4	0.75	0.70	23.3	0.20	6.7
4	E_4 : Profess.	8	1.75	0.76	10.9	0.80	11.4

Scorrendo i rapporti di correlazione, si vede che i primi due assi non sono legati alla variabile 'E1 - Sesso dell'intervistato' ed il secondo neppure alla 'E3 - Titolo di studio'. La percentuale d'inerzia che un asse riesce ad estrarre da una variabile *ben legata* ad esso è inversamente dipendente dal numero di modalità. Così, pur con valori molto simili di $\eta^2_1(q)$, il primo asse fattoriale riesce ad 'estrarre' il $23.3\,\%$ dell'inerzia della variabile 'E3=Titolo di studio' con 4 modalità, ma soltanto il $10.9\,\%$ della variabile 'E4=Professione' che ha invece 8 modalità. Occorrerà quindi esaminare assi di rango elevato per avere un quadro completo degli eventuali legami tra le professioni degli intervistati e le modalità delle altre variabili [1]. Per maggiore chiarezza, è bene costruire un grafico come quello di TAV. 5.14, posizionando le variabili su una mappa ottenuta incrociando i loro rapporti di correlazione $\eta^2_a(q)$ e $\eta^2_b(q)$ su due assi fattoriali a e b. Il grafico va interpretato come la proiezione sul piano (a, b) di una nuvola di Q variabili, per cui la distanza tra due proiezioni traduce la

[1] Si è mostrato che per molte buone ragioni il numero di modalità di una variabile *attiva* non dovrebbe superare 4 o 5. Questa regola è stata intenzionalmente violata nell'esempio degli ascolti radiofonici, dato il fine puramente didattico.

similitudine tra i due insiemi di modalità e facilita la selezione delle variabili maggiormente legate [1] ad entrambi gli assi e quindi al piano fattoriale, come è il caso delle due variabili 'E2 - Età dell'intervistato' e 'E4 - Qualifica professionale' col piano principale (1, 2). Per questi motivi è preferibile esaminare le variabili tramite i loro rapporti di correlazione, piuttosto che attraverso i contributi relativi.

2 - Caratterizzazione degli assi fattoriali

Nell'ACM, 'descrivere' un asse fattoriale significa dargli un significato in base alle coordinate e ai contributi relativi delle variabili, delle modalità e degli individui. Come nelle Corrispondenze semplici, è buona regola iniziare con l'esame asse per asse. Qui l'esame sarà limitato ai primi due.

Contributi delle variabili

I contributi relativi $CTR_a(q)$ delle variabili sono degli indicatori peculiari dell'ACM e vengono sempre stampati dai programmi di calcolo, spesso moltiplicati per 100 , come nella TAV. 5.10 e nelle seguenti. Il loro esame porta alle stesse conclusioni viste sopra perché i contributi sono legati ai rapporti di correlazione grazie alla (5.18.4).

Coordinate e contributi delle modalità

I risultati di una ACM sono stampati abitualmente come nella TAV. 5.10 e seguenti e risultano quindi organizzati diversamente da quelli delle Corrispondenze semplici: le modalità sono adesso raggruppate per variabile e non elencate in base a coordinate e contributi. L'elenco va redatto quindi manualmente ordinando le modalità attive, limitatamente a quelle con contributo rilevante, in base alla loro coordinata sull'asse. Ciò facilita l'individuazione delle modalità che hanno maggiormente contribuito ad orientare l'asse fattoriale e gli danno quindi un significato. Lo specchietto che segue riassume la situazione sul primo asse fattoriale

j	*Modalità*	*massa*	g_{j1}	$CTR_1(\mathbf{c}_j)$
3	E21 : Età = meno di 18	2.81	−1.75	15.1
11	E31 : Titolo = licen. elem.	2.19	−1.72	11.4
14	E41 : Qualifica = studente	7.25	−1.04	13.8
13	E31 : Titolo = laurea	4.00	+1.20	10.2

[1] ossia che hanno maggiormente contribuito.

Si vede che questo primo asse traduce il livello di istruzione, perché oppone gli studenti intervistati con meno di 18 anni e licenza elementare ai laureati.

La situazione sul secondo asse è invece la seguente

j	*Modalità*	*massa*	g_{j2}	$CTR_2(\mathbf{c}_j)$
4	E22 : Età = tra 18 e 25	4.06	−1.22	13.1
14	E41 : Qualifica = studente	7.25	−0.93	13.6
9	E27 : Età = oltre 60	3.44	+1.58	18.5
20	E47 : Qualifica = casalinga	2.62	+2.05	23.9

Il secondo asse traduce dunque l'età degli intervistati, in quanto oppone i giovani studenti tra 18 e 25 anni agli ultra sessantenni, particolarmente alle casalinghe.

C'è un punto importante da tener presente in questa fase. Nell'ACM può capitare che il primo asse, o i primi, siano orientati da modalità rare, scelte da pochi individui, dato che la distanza di un profilo $\mathbf{c}_j$ dal baricentro è inversamente proporzionale a z_{+j}, come indica la (5.12.4). E' importante quindi controllare anche la massa. Nei due specchietti qui sopra è moltiplicata per 100. Se questa risultasse piccola, per fissare le idee inferiore a 0.20, nell'espressione (5.18.1) del contributo, la distanza g_{ja}^2 sull'asse sarebbe grande e il profilo si troverebbe ben lontano dal baricentro. Per evitare che l'orientamento dei primi assi sia alterato da queste modalità, è necessario raggrupparle con altre della stessa variabile o eliminarle del tutto 'ventilando' su altre modalità i pochi individui che le hanno scelte, nel modo indicato nella Sez. 5.13.

Qualità della rappresentazione

Questo indicatore è poco utilizzato perché i profili delle modalità di una stessa variabile sono ortogonali due a due, e non possono quindi essere ben rappresentati simultaneamente sullo stesso asse fattoriale.

3 - Interpretazione delle mappe delle modalità

Una volta caratterizzati i principali assi, si passa a esaminare le proiezioni delle modalità sui piani fattoriali. Qui verrà esaminata soltanto la mappa principale, dato l'interesse relativamente limitato dell'esempio.

Le mappe si interpretano come nell'Analisi delle Corrispondenze semplici, tenendo anzitutto presente che nell'ACM l'origine degli assi fattoriali è il baricentro delle J_q modalità di ogni variabile. più una modalità ha

massa elevata più essa è 'attirata' verso l'origine. La mappa principale di TAV. 5.15 rivela subito che nell'indagine sugli ascolti radiofonici c'è equilibrio tra il numero di intervistati maschi e di intervistati femmine, dato che i loro punti rappresentativi sono praticamente equidistanti dall'origine. Infatti, dalla TAV. 5.10 risulta che le masse dei due punti sono rispettivamente 13.38 e 11.62 .

Nell'ACM gli assi fattoriali oppongono tra loro *simultaneamente* sia le J_q modalità di una stessa variabile sia il complesso di tutte le J modalità. Perciò, se i profili di due modalità di una *stessa* variabile sono vicini, significa che i due segmenti di individui hanno fatto scelte abbastanza simili, relativamente alle modalità delle altre variabili attive. Invece, se sono vicini i profili di due modalità di variabili *diverse*, significa che le due modalità sono state quasi sempre dagli stessi individui.

In pratica i due concetti di distanza si utilizzano congiuntamente, interpretando la vicinanza di modalità di variabili diverse come *associazione* tra modalità perché gli stessi individui le hanno frequentemente associate nelle loro scelte e la vicinanza tra modalità di una stessa variabile come *somiglianza* di comportamento di segmenti di individui diversi. Così, nella mappa fattoriale di TAV. 5.15, si interpreta la vicinanza della modalità 'E41 - Studenti' a 'E21 - Età inferiore a 18 anni' e a 'E22 - Età tra 18 e 25 anni' in termini di associazione perché sono sostanzialmente gli stessi intervistati che hanno scelto la prima ed entrambe le altre due modalità, e la prossimità tra 'E21 - Età inferiore a 18 anni' e 'E22 - Età tra 18 e 25 anni' in termini di somiglianza perché gli individui di questi due segmenti hanno evidentemente scelto le stesse modalità delle altre tre variabili.

Se le modalità di una variabile sono ordinate, per esempio perché ottenute suddividendo in classi di valori una variabile numerica, si può rendere la mappa più leggibile, e facilitarne l'interpretazione, collegando con una linea poligonale i punti rappresentativi delle modalità, rispettando l'ordine. Nella mappa principale della TAV. 5.15 sono state collegate le modalità delle variabili 'E2 - Fascia d'età' ed 'E3 - Titolo di studio'. Le modalità si trovano ai vertici delle poligonali convesse, dette *spezzate*, con baricentro l'origine degli assi [1] . L'evolversi delle 'spezzate' di variabili diverse permette di meglio apprezzare l'eventuale esistenza di associazioni tra variabili. Anche la direzione

[1] Quando le modalità sono due sole, il segmento che le collega passa per l'origine degli assi che è il loro baricentro. La proprietà si conserva nelle proiezioni in sottospazi.

dell'evoluzione è importante. Per esempio quella della variabile 'E3 - Titolo di studio' concorda con quella del primo asse fattoriale e quella di 'E2 - Età' sostanzialmente con quella del secondo.

L'esame di una mappa inizia di solito controllando l'andamento delle spezzate che spesso evidenziano importanti tendenze. Le principali tipologie dei loro andamenti sono schematizzate nella TAV. 5.17, supponendo che a ogni classe appartenga un numero abbastanza simile di individui. In caso contrario, oltre alla distorsione dovuta alla proiezione sul piano fattoriale, occorre anche tenere conto del fatto che i profili di classi più 'leggere' si collocano più lontani dal baricentro, ossia dall'origine degli assi, in accordo con la (5.12.4).

Interruzione della spezzata

L'andamento di una spezzata può talvolta evidenziare dei legami del tutto ovvi tra le modalità di una variabile. Per questo, la presenza di eventuali anomalie nella regolarità evolutiva di una spezzata deve mettere in allerta. Così, nel caso in alto a sinistra della TAV. 5.17 il profilo di una modalità rompe nettamente l'evoluzione regolare della spezzata. Ciò significa che il corrispondente segmento di individui ha effettuato delle scelte originali nei riguardi delle modalità delle altre variabili. Il fatto va accuratamente investigato.

Spezzata ad anello

Nella spezzata in alto a destra la vicinanza del primo punto all'ultimo indica che gli individui che hanno scelto la prima modalità e quelli (necessariamente diversi) che hanno invece scelto l'ultima hanno poi scelto quasi sempre le stesse, modalità delle altre variabili. In altri termini, i due segmenti estremi hanno tenuto un comportamento simile.

Fascio di spezzate

Sono i legami tra due o più spezzate che spesso mettono in luce i fenomeni più interessanti. Per esempio, le due spezzate in basso a sinistra mostrano un andamento più o meno parallelo. Ciò indica che se le due variabili erano originariamente numeriche, queste erano *correlate*. Positivamente, se gli andamenti si sviluppano nello stesso senso, negativamente se sono in controtendenza. Nel primo caso gli stessi individui hanno scelto simultaneamente le modalità d'egual rango di entrambe le variabili, nel secondo, di rango opposto.

L'andamento parallelo delle spezzate può anche presentarsi in forma non lineare, per esempio più o meno parabolica, come nel caso in basso a

destra. La correlazione tra le due variabili numeriche sussiste ancora. In ogni caso, che l'andamento sia lineare o meno, nella matrice di contingenza che incrocia le J_q modalità, o classi, della prima variabile con le $J_{q'}$ della seconda q', si osserva che gli elementi vicini alla diagonale principale hanno valori più elevati degli altri quando la correlazione è positiva. Sono invece preponderanti quelli vicini all'altra diagonale quando la correlazione fosse negativa.

Un'altra importante relazione non lineare tra due spezzate appare nella TAV. 5.15, e precisamente tra la variabile 'E2 - Età dell'intervistato' e 'E3 - Titolo di studio'. La spezzata di quest'ultima ha un andamento abbastanza lineare, con i meno istruiti posti tra i due estremi dell'andamento pressappoco parabolico dell'altra. Così, ad esempio, gli intervistati con licenza elementare e media o sono molto giovani perché ancora studenti, o molto anziani. L'analisi non ne spiega il motivo, ma solo evidenzia il fatto. Se le variabili originali fossero state entrambe numeriche, per esempio Età e Reddito, queste sarebbero legate da una relazione di secondo grado del tipo $\text{Età} \simeq A \times \text{Reddito}^2 + B$ con A e B costanti. Questo tipo di relazione implica che il coefficiente di correlazione lineare tra le due variabili originali sia vicino a zero.

Il grande pregio dell'ACM è la sua capacità di evidenziare l'eventuale presenza, o assenza, di associazioni tra modalità attive ed illustrative. Queste sono valutate in base al loro Valore-test, ritenendo la loro posizione su un asse 'significativa' quando l'indicatore è, in valore assoluto, superiore a 2. Le TAV. 5.11, 5.12 e 5.13 elencano le modalità illustrative dell'esempio, raggruppate per variabile. Per ogni modalità è data la numerosità $\tilde{z}_{+j}$, la coordinata fattoriale del suo profilo $\tilde{g}_{j1}$ sul primo e $\tilde{g}_{j2}$ sul secondo asse ed il Valore-test.

Le modalità illustrative sono spesso assai numerose, per cui è opportuno proiettare sulla mappa oltre alle modalità *del* tema attivo, quelle di *un solo* tema illustrativo per volta, limitando eventualmente queste ultime alle sole modalità con Valore-test significativo. Le prossimità tra modalità attive e illustrative si interpretano con i criteri di *associazione* e di *somiglianza* citati più sopra. Conclusi questi esami, si lasciano sulla mappa soltanto le modalità, di temi diversi, che hanno mostrato le associazioni più interessanti. Così, per esempio, nella TAV. 5.16 sono raffigurate tutte le modalità del tema attivo ed il 60% di quelle del tema illustrativo 'A - Programmi che vengono ascoltati': quelle con il Valore-test più elevato sul primo asse. Se ci fossero due modalità

illustrative con lo stesso Valore-test e soltanto una di esse dovesse essere selezionata, viene effettuata una scelta casuale tra le due. Si nota chiaramente il mutare delle preferenze d'ascolto con l'avanzare dell'età: dal rock contemporaneo dei più giovani, al jazz, al pop e al folk revival per le fascie intermedie fino alla musica italiana revival dei più anziani. Invece l'ascolto dei servizi sportivi sembra ben ripartito tra gli intervistati, anche se con prevalenza nelle fasce intermedie di età, data la sua leggera eccentricità verso questi profili. L'esame delle mappe fattoriali non verrà spinto oltre, dato l'interesse limitato dell'esempio.

4 - Interpretazione delle mappe degli individui

Nell'ACM gli individui sono quasi sempre anonimi e molto numerosi. Se venissero proiettati tutti sulle mappe delle modalità, come nelle Corrispondenze semplici, finirebbero per renderle illeggibili. Per questo ci si limita a esaminare la *forma* della nuvola degli individui su mappe di ordine via via crescente, dove ogni individuo è rappresentato da un punto. La mappa principale nel caso dell'esempio è nella TAV. 5.18 e mostra una nuvola di forma paraboloide con l'apertura orientata verso il quadrante negativo-positivo. Questa forma si incontra di frequente quando le variabili rilevate sono ordinali, come sono in questo caso, l'Età e il Titolo di Studio. All'estremo in basso del paraboloide ci sono gli individui giovani, in quello superiore gli anziani e, al centro, la modalità che i due segmenti hanno in comune, in questo caso la Licenza elementare. Si ricordi che per il principio baricentrico una modalità è il baricentro del segmento di individui che la possiedono: in questo caso i giovani e gli anziani sono i principali costituenti del segmento col più basso titolo di studio.

Una mappa più informativa si può ottenere rappresentando ogni individuo con una lettera che richiami una delle modalità possedute, per esempio una 'M' se è maschio e una 'F' se femmina. Su queste mappe si possono poi proiettare le due modalità 'Maschio' e 'Femmina' che, per la (5.15.2), sono i baricentri dei punti contraddistinti con 'M' e 'F'. Così, oltre alla forma della nuvola complessiva si evidenziano anche quella delle due sub-nuvole dei due sessi e si può controllare se esse sono ben separate, sfilacciate, compenetrate, ecc. Questa rappresentazione si può ripetere per le variabili più strutturalmente interessanti, sia attive che illustrative.

5.21 - Risposte mancanti

Si è in presenza di risposte mancanti, o non-risposte, quando l'intervistato ha fornito risposte non corrette o si è astenuto dal rispondere a una o più domande. Di conseguenza, venendo a mancare nella tabella dei dati l'indicazione di una o più modalità, il profilo della colonna marginale non risulta più piatto. E' opportuno ripristinare l'uniformità del profilo marginale prima di procedere a ogni analisi [1] . I modi per farlo sono diversi.

Quando la variabile fa parte di un tema *illustrativo* la soluzione consiste nel creare un'ulteriore modalità: 'Non risponde'. I risultati non ne sono influenzati, dato che derivano dalle variabili attive.

Quando invece la variabile appartiene al tema *attivo*, il problema è più delicato e occorre tener conto di altri elementi. Se il numero di individui che ha fornito risposte incomplete è piccolo rispetto al totale degli intervistati, la prima soluzione, e la più drastica, è quella di eliminare dall'analisi tutti gli individui con risposte incomplete. La seconda è quella di attribuire a caso all'individuo una delle altre modalità della variabile, previste nel questionario: è il procedimento di 'ventilazione', già descritto nella Sez. 5.13 per le modalità rare. La perdita di informazione è trascurabile con entrambe le soluzioni. Quando, invece, il numero di individui interessati *non* è trascurabile, si può ancora aggiungere alle modalità della variabile la nuova modalità 'Non risponde'. Dato però che ora il suo profilo influisce attivamente sui risultati, questa soluzione va scelta quando 'Non risponde' ha un significato ben preciso e non generico. Per esempio quando equivale a 'Non so' perché traduce l'ignoranza dell'intervistato relativamente alla materia indagata, oppure a 'Non saprei' perché raccoglie l'indecisione nella scelta della risposta, oppure a 'Preferisco non rispondere' perché la domanda era delicata, oppure ancora a 'Nessuna di queste risposte' per riunire altre modalità di risposta meno interessanti ai fini dell'indagine di quelle previste nel questionario, ecc. Al di fuori di casi di questo tipo, resta la possibilità di trasferire la variabile tra le illustrative, sempre che ciò non alteri sostanzialmente il tema attivo e sposti, di conseguenza, l'obiettivo dello studio.

5.22 - Programmi di analisi

L'analisi dei dati raccolti tramite sondaggio ha grandemente benefi-

---

[1] Non ripristinare l'uniformità del profilo marginale significa dare meno importanza agli astenuti. Greenacre (1984), op. cit. nella Sez. 4.17, riprende nelle pagg. 146 - 157 alcuni articoli di Benzàcri (1976) su diversi modi possibili di codificare le risposte: Favorevole, Contrario e Astenuto.

ciato della capillare diffusione dell'elaborazione elettronica. La potenza di calcolo e la capacità di memoria ora disponibili permettono di analizzare rapidamente e a un prezzo accessibile il cospicuo volume di dati raccolti che contempla sovente migliaia di intervistati e decine e decine di domande con centinaia di modalità di risposta. Programmi per l'Analisi delle Corrispondenze Multiple sono presenti in forma più o meno completa in tutti i pacchetti software elencati nella Sez. 4.14 e nei molti altri citati dai testi in bibliografia nella Sez. 4.17 e nella Sez. 5.24. Molti si possono scaricare gratuitamente dalla rete. Comunque, quale che sia il programma impiegato, l'utente dovrebbe trovare in questo Capitolo tutto l'aiuto teorico e pratico per una lettura ottimale dei risultati.

Per l'analisi dei questionari sull'ascolto delle trasmissioni radio si è impiegato il programma SPADN 3.5 per Windows [1] che, al momento, è il più evoluto e completo e rappresenta la 'dottrina ufficiale' dell'ACM, essendo stato messo a punto dai fondatori del metodo. Il programma, disponibile per diverse piattaforme, è modulare, nel senso che ogni procedura statistica prevista costituisce un modulo, trattato tramite un'icona. Le icone vanno concatenate tra loro con la tecnica del 'drag and drop' tra due finestre: quella dei moduli e quella della catena. Le prodedure statistiche sono poi eseguite in successione nell'ordine in cui appaiono nella catena, utilizzando i risultati dell'elaborazione della procedura precedente. La prima icona indica il file dei dati.

È possibile anche scrivere un programma, ossia una sequenza di comandi, da far eseguire come 'script'. Quest'ultimo modo permette agli analisti più esperti di avere il controllo completo dei passi dell'analisi. Il programma permette sofisticate realizzazioni grafiche. È una caratteristica importante perché si è visto che l'analisi dei questionari con l'ACM è sì ricca di risultati, ma procede per continue variazione dell'analisi precedente: un processo iterativo che converge lentamente alla soluzione definitiva, ossia ai risultati che verranno poi pubblicati o consegnati al committente.

5.23 - Conclusioni

L'ACM è l'estensione più immediata delle Corrispondenze semplici, la cui caratteristica principale, grazie alle relazioni di transizione, è la capacità di evidenziare graficamente le associazioni tra le righe e le colonne di una

[1] Le caratteristiche dettagliate del programma si possono trovare nel sito `www.cisia.com`.

matrice. In questo Capitolo si è visto come l'ACM permetta sì di esplorare le associazioni tra le modalità di un gruppo (tema attivo) di variabili categoriche e di metterle in relazione con quelle di altri gruppi (temi illustrativi), ma le relazioni di transizione non vengono sfruttate perché i singoli individui (le righe), che di solito sono anonimi e molto numerosi, non vengono messi singolarmente in relazione con le modalità (le colonne). Il confronto avviene indirettamente perché le modalità rappresentano anche l'individuo medio degli appartenenti a un segmento. La strada che si segue è quella di riunire in gruppi gli individui con profilo di risposta il più possibile simile, relativamente alle variabili del tema attivo, e di proiettare poi l'individuo medio (baricentro) di ogni gruppo sulle mappe fattoriali delle modalità.

Nel prossimo Capitolo sono illustrati quegli aspetti dell'Analisi dei gruppi direttamente attinenti alla costruzione di gruppi di profili e verrà mostrata la complementarità tra l'ACM e l'Analisi dei gruppi che porta a una più approfondita comprensione della struttura dei dati.

5.24 - Bibliografia essenziale

Per approfondire i concetti e le metodiche dell'ACM, il lettore può consultare

Ludovic Lebart, Alain Morineau, Marie Piron (1995). *Statistique exploratoire multidimensionelle*. Dunod ed., Paris., 440 pg., ISBN 2-10-002886-3, ove oltre ai metodi fattoriali esposti in forma concisa e rigorosa, vengono anche presentati metodi di analisi statistica più o meno recenti, basati su modelli probabilistici e non. Contiene una amplissima ed aggiornata bibliografia.

Un recente testo italiano sui metodi dell'Analisi Multidimensionale è

Sergio Bolasco (1999). *Analisi Multidimensionale dei Dati: metodi, strategie e criteri di interpretazione*. Carocci ed., Roma., 358 pg., ISBN 88-430-1401-3, che contiene anche una parte dedicata all'analisi dei dati testuali, che si possono ottenere, ad esempio, con questionari aperti a risposta libera.

Un esauriente articolo sulla codifica sfumata di variabili numeriche è questo di

F. J. Gallego (1982). *Codage flou en Analyse des Correspondances*. Les Cahiers de l'Analyse des données, Vol. VII, n.o 4, pag 413 - 430.

L'attuale stato dell'arte sull'analisi dei questionari è illustrato in

D. Grangè, Lodovic Lebart, ed. (1994). *Traitements statistiques des enquêtes*. Dunod ed., Paris, 255 pag., ISBN 2-10-002008-0. Senza dimostrazioni matematiche e di facile lettura, contiene i contributi di 10 autori di varia estrazione sulle diverse fasi di un'inchiesta: preparazione del questionario, codifica dei risultati, metodi tradizionali ed avanzati di analisi, ecc. Contiene anche un interessante cenno storico sulle inchieste ed un capitolo finale sull'analisi delle risposte a domande 'aperte' (alle quali l'intervistato risponde liberamente con una frase). Entrambi i contributi sono di Lodovic Lebart.

PARTE PRIMA: IL METODO

CAPITOLO 6: Analisi dei gruppi

Sommario

Nel Capitolo precedente dedicato alla Analisi delle Corrispondenze Multiple si è visto come l'elevato numero di individui (profili delle righe) coinvolti nell'analisi, impedisca la loro rappresentazione sulle mappe grafiche. I metodi dell'Analisi dei Gruppi, impiegati successivamente e complementarmente all'Analisi delle Corrispondenze, sono un mezzo semplice e non eccessivamente dispendioso in termini di calcolo, per aggregare i profili simili in termini di coordinate. Ciò permette di acquisire informazioni sintetiche sulla configurazione geometrica della nuvola di profili nel suo spazio ambiente. La proiezione sulle mappe fattoriali dei baricentri dei gruppi, ossia dei profili medi, permette poi di evidenziare i principali legami tra gruppi di profili e modalità.

La lettura di questo capitolo, porrò il lettore in grado di

- distinguere i diversi metodi di aggregazione;
- conoscere ed utilizzare al meglio le strategie di aggregazione;
- sfruttare la complementarietà tra l'Analisi delle Corrispondenze ed alcuni metodi dell'Analisi dei Gruppi;
- scomporre l'inerzia rispetto ai nodi dell'albero dei gruppi;
- 'tagliare' correttamente l'albero dei gruppi per stabilire la partizione finale;
- utilizzare i risultati dei metodi di aggregazione per interpretare i gruppi.

CAPITOLO 6

6.1 - Introduzione

Nei Capitoli precedenti si è visto come l'Analisi delle Corrispondenze, semplici o multiple, consenta di investigare la struttura nascosta di una matrice di dati rappresentandone i profili su mappe fattoriali bi- e tri-dimensionali. In questo Capitolo verrà mostrato come tale struttura si possa investigare anche con metodi appartenenti all'Analisi dei Gruppi, usati successivamente, e complementariamente, a una Analisi delle Corrispondenze. Tali metodi presuppongono infatti che tra profili, quasi sempre quelli delle righe che di solito sono i più numerosi, sia definita una *distanza* e che quindi le variabili siano di tipo *quantitativo*. La preventiva applicazione dell'Analisi delle Corrispondenze effettua di fatto una ricodifica delle variabili categoriche originali, trasformando le coordinate dei loro profili in coordinate fattoriali di tipo quantitativo. I risultati finali possono venire poi riportati sulle mappe fattoriali.

Analisi dei Gruppi è il nome generico che riunisce un ampio spettro di metodi matematici, statistici e algoritmici per aggregare in gruppi [1] 'omogenei' le righe o le colonne di una matrice. In questo Capitolo sarà considerato unicamente il caso dei profili delle righe. In ciascun gruppo tali profili devono avere coordinate il più possibile 'simili' e al tempo stesso il più possibile 'diverse' da quelle dei profili degli altri gruppi. I gruppi così ottenuti costituiscono una *partizione* [2] che può essere vista come un 'istogramma multidimensionale' per ridurre un ampio spazio costellato di profili in porzioni più piccole e più semplici da descrivere. Bisogna dire che al momento, l'individuazione di gruppi con caratteristiche ottimali è un problema della Statistica ancora non risolto perché quasi nessun metodo passa in rassegna tutte le possibili

---

[1] Gruppo è detto *cluster* in inglese e *classe* in francese. L'Analisi dei Gruppi è detta rispettivamente *Cluster Analysis* e *Classification*.

[2] Una partizione è la suddivisione dell'insieme dei profili in un numero ridotto di sottoinsiemi disgiunti e non vuoti, detti gruppi (di profili). I gruppi possono essere costitituiti anche da un solo profilo. Una definizione più precisa è nella Sez. 6.9.

partizioni per individuarne la 'migliore'. Ci si deve perciò accontentare di partizioni con caratteristiche sub-ottimali, ma comunque più che soddisfacenti, ottenute ricorrendo a tecniche miste, dette *strategie* di aggregazione, consistenti in più metodi impiegati in successione.

In questo Capitolo ci si limiterà a presentarne due che l'esperienza ha fatto riconoscere come le più adatte a fornire risultati prossimi a quelli ottimali quando si opera con profili: la strategia dei *Gruppi Stabili* e la strategia *Mista*. La prima è basata su un metodo di aggregazione a Centri Mobili, la seconda fa seguire alla prima un metodo di Aggregazione Gerarchica Ascendente basato sull'inerzia tra gruppi della partizione. La differenza tra i due metodi è sostanziale. I metodi *non gerarchici*, come l'aggregazione a Centri Mobili, costruiscono un'unica partizione, quelli *gerarchici* costruiscono invece un'intera successione di partizioni, ognuna delle quali è organizzata gerarchicamente ed ottenuta dalla precedente per aggregazione dei *due* gruppi più 'prossimi'. I metodi non gerarchici sono estremamente veloci e quindi ideali quando il numero di profili è elevato, ma presuppongono noto il numero dei gruppi, mentre i metodi gerarchici sono meno veloci e richiedono una maggiore disponibilità di memoria di calcolo. In compenso permettono di seguire la successione delle partizioni e di individuare più facilmente il numero dei gruppi.

6.2 - Obiettivi dell'Analisi dei Gruppi

In questo capitolo l'interesse è limitato all'applicazione di due metodi dell'Analisi dei Gruppi a una nuvola di *profili* delle righe [1], individuati dalle loro coordinate fattoriali e dotati di massa. Il fine è quello di individuare la 'migliore' partizione, e quindi il *numero* e la *composizione* dei gruppi, in modo che questi risultino:

1 - il più possibile 'omogenei' al loro interno, nel senso che i profili che li costituiscono devono avere coordinate fattoriali il più possibile simili;

2 - il più possibile 'differenti' gli uni dagli altri, sempre relativamente alle coordinate fattoriali.

Per soddisfare queste due esigenze occorre anzitutto introdurre un qualche criterio che permetta di valutare la 'somiglianza' dei profili in uno stesso gruppo e, al tempo stesso, la 'diversità' tra profili di gruppi diversi. In

[1] Per raggruppare i profili delle colonne è spesso impiegata la procedura *ACECLUS* di SAS, descritta nel manuale *SASSTAT Users' Guide*, Vol. 3, SAS Institute Inc., Cary, NC (USA), Cap. 68, pag. 3591.

secondo luogo, occorre individuare un algoritmo che permetta di trovare la 'migliore' partizione in termini delle condizioni 1) e 2). Un tale algoritmo, per ogni fissato numero H di gruppi della partizione, dovrebbe esaminare *tutte* le suddivisioni possibili dell'insieme di profili in H sottoinsiemi e individuarne la 'migliore', in base al criterio di omogeneità prestabilito. Questo modo di procedere sarebbe troppo lento, se non impraticabile, con i mezzi di calcolo attualmente disponibili perché il numero di partizioni da esaminare cresce esponenzialmente col numero di gruppi e, al momento, non si conosce alcun modo che eviti l'esame di tutte le possibili partizioni. Si ripiega così su *strategie* di aggregazione che forniscano partizioni ragionevolmente soddisfacenti per le applicazioni pratiche. In altri termini, si cercano delle soluzioni approssimate, e quindi sub-ottimali dal punto di vista dell'algoritmo di aggregazione, che siano tuttavia 'migliori' secondo criteri meno restrittivi di 1) e 2).

Per valutare l'omogeneità dei profili nei gruppi, o alternativamente la disomogeneità tra profili di gruppi diversi, la letteratura propone molti criteri, ma per i *profili* si adotta abitualmente il criterio basato sull' inerzia *nei* gruppi e *tra* gruppi, come verrà precisato nella Sez. 6.5.

6.3 - Analisi delle Corrispondenze e dei Gruppi

Un'importante proprietà, comune all'Analisi delle Corrispondenze sia semplici che Multiple e che viene qui anticipata informalmente perché sarà esposta in dettaglio nalla Sez. 8.7, è che le coordinate standard $\hat{g}_{j1}$ delle J modalità sul primo asse fattoriale, forniscono una graduazione ottimale *(optimal scaling)* dell'asse. E' ottimale nel senso che in base a questa scala, le coordinate f_{i1} degli I individui sul primo asse, o primo fattore delle righe, hanno la massima varianza, ossia la massima dispersione possibile. Anche sugli assi di rango successivo gli individui hanno la massima dispersione, compatibilmente con i vincoli imposti dall'ortogonalità con gli assi di rango precedente.

Rimandando il lettore alla Sez. 8.7 per maggiori dettagli ed esempi e limitando qui il caso ai profili delle righe, la proprietà di optimal scaling dell'Analisi delle Corrispondenze si può tradurre così: non è possibile trovare alcun vettore $\mathbf{w}$ di lunghezza unitaria in $\Re^J$ per la metrica di questo spazio, per il quale cioè $\mathbf{w}^T\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{w} = 1$, tale che sull'asse da esso individuato l'inerzia (varianza) delle proiezioni dei profili di $\mathbf{R}$ possa superare λ_1 , il primo autovalore non nullo, valore che si ottiene quando $\mathbf{w}_1 = \mathbf{v}_1$, ossia quando $\mathbf{w}_1$

coincide col primo autovettore $\mathbf{v}_1$, perché

$$\lambda_1 = Max\left(VAR\left(\mathbf{RD}_{\mathbf{r}}^{-\mathbf{1}}\mathbf{w}\right)\right) = VAR\left(\mathbf{RD}_{\mathbf{r}}^{-\mathbf{1}}\mathbf{v}_1\right) = VAR\ (\mathbf{f}_1)\,.$$

Ancora, se $\mathbf{w}$ è anche vincolato ad essere ortogonale ai primi $a-1$ autovettori $\mathbf{v}_1, \mathbf{v}_1, \ldots, \mathbf{v}_{a-1}$, allora l'inerzia ha come massimo il valore λ_a, il successivo a^{mo} autovalore. E ciò vale per ogni asse fattoriale $a = 1, 2, \ldots, A$. Tutto ciò, tradotto in termini più semplici, significa che gli autovettori individuano il sistema di riferimento più stabile perché sui suoi assi la nuvola degli I individui si proietta con la massima dispersione, se questa è misurata in termini di varianza.

Il fatto che l'Analisi delle Corrispondenze, sia semplici che multiple, produca la massima discriminazione tra individui, è uno dei motivi che inducono ad effettuare preventivamente un'analisi fattoriale sulla matrice dei dati ed impiegare poi le coordinate fattoriali dei profili per costruire i gruppi. Ma, di motivi validi ve ne sono altri. Intanto, grazie ai risultati nella Sez. 3.6, viene risolto una volta per tutte il problema della scelta della distanza tra profili: per profili espressi in coordinate fattoriali la distanza è quella euclidea canonica [1]. Ne consegue l'ulteriore vantaggio che qualunque sia il metodo di aggregazione, l'algoritmo da programmare è uno solo: quello basato sulla distanza euclidea canonica. L'unicità del tipo di distanza permette anche di interpretare congiuntamente sulla medesima mappa i risultati dell'Analisi delle Corrispondenze e dei Gruppi. Partendo da punti di vista diversi e con risultati di diverso tipo, entrambi i metodi concorrono così alla comprensione della matrice dei profili. Inoltre, dato che l'Analisi dei gruppi produce comunque una partizione, anche nel caso di totale assenza di struttura tra profili, il preventivo impiego dell'Analisi delle Corrispondenze mette al riparo da questo rischio.

Nell'Analisi delle Corrispondenze un profilo eccentrico può influenzare il primo asse fattoriale, e anche i successivi a causa dei vincoli di ortogonalità, come si è visto col profilo Veneto nell'analisi della matrice *Spettacoli*. Invece, i due metodi di aggregazione dei profili, presentati in questo Capitolo, sono relativamente poco sensibili ai profili eccentrici e quello di aggregazione ger-

---

[1] Se si opera direttamente sui profili $\mathbf{r}_i$, non espressi quindi in coordinate fattoriali, si devono usare le distanze distribuzionali (4.8.1) o (5.12.1). I risultati sono eguali a quelli ottenuti conservando tutte le A coordinate fattoriali. Geometricamente, si tratta di un cambiamento del sistema di riferimento della nuvola.

archica lo è sicuramente nei primi passi di aggregazione.

L'Analisi delle Corrispondenze Multiple si avvantaggia particolarmente dei metodi dell'Analisi dei Gruppi perché nell'ACM va persa la visualizzazione delle relazioni righe - colonne, ossia individui - modalità, punto di forza delle Corrispondenze semplici, per il fatto che gli individui sono numerosi, spesso migliaia, e anonimi. Le mappe diventerebbero illeggibili se vi venissero proiettati tutti singolarmente. Inoltre, l'approccio grafico di rado può fornire una visione esaustiva della struttura dei profili, perché la visualizzazione delle mappe è limitata a due o tre assi per volta, mentre di solito quelli interpretabili nell'ACM sono molti di più. Ad esempio, erano $A^* = 8$ nel caso dell'indagine sull'ascolto radiofonico. La strada che si segue è allora quella di riassumere ulteriormente la configurazione dei profili riunendoli in pochi gruppi omogenei dal punto di vista di *tutte* le coordinate nel loro spazio fattoriale, perché l'algoritmo di aggregazione funziona indifferentemente con 2 o con 8 coordinate. I gruppi vengono poi descritti in base alle modalità che più concorrono a caratterizzarli e soltanto il loro baricentro, l'individuo medio rappresentativo di ogni gruppo, è proiettato sulle mappe precedentemente ottenute dall'ACM, permettendo così di evidenziare graficamente le associazioni tra gruppi di individui e modalità. Infatti, i gruppi riuniscono gli individui con modalità attive relativamente simili, mentre, come si è visto nella Sez. 5.15, i segmenti riuniscono quelli che hanno in comune *una sola* specifica modalità.

Il fatto di conservare soltanto le prime coordinate fattoriali che danno il massimo dell'informazione geometrica sulla configurazione della nuvola di profili elimina le fluttuazioni dei dati che possono mascherare fenomeni interessanti, facilitando l'individuazione dei gruppi. Le coordinate fattoriali risultano così meno numerose e questo permette di accelerare l'elaborazione e di utilizzare metodi di aggregazione di tipo gerachico. Nell'ACM, anche con parecchie decine di modalità attive raramente il numero di fattori conservati supera 10. Se il numero di individui non è eccessivo, la matrice $\mathbf{F}$ delle coordinate fattoriali dei profili può essere conservata nella memoria centrale del computer. Ciò comporta una semplificazione della programmazione e una maggiore velocità i calcolo come si vedrà nella Sez. 6.12.

L'inconveniente di questo modo di procedere è che occorre indicare il numero di assi fattoriali da conservare. Utili indicazioni si possono trarre dall'esame del diagramma dei tassi d'inerzia e dai criteri statistici che verranno presentati nel prossimo Capitolo 7.

6.4 - Coordinate fattoriali di un profilo

I metodi dell'Analisi dei Gruppi presentati in questo capitolo verranno applicati esclusivamente a profili e in particolare a quelli delle *righe* perché abitualmente sono l'insieme più numeroso. Per formalizzare correttamente metodi, algoritmi e strategie occorre ora introdurre delle specifiche notazioni.

Nei capitoli precedenti si è visto che nella nuvola degli I profili delle righe, l'i^{mo} profilo è la i^{ma} *riga* $\mathbf{r}_i$ della matrice $\mathbf{R}$, di ordine $I \times J$, della (1.5.1) e (5.10.1)

$$\mathbf{r}_i = (r_{i1} \;\; r_{i2} \;\; \dots r_{ij} \;\; \dots r_{iJ})^T .$$

Le sue J componenti sono le coordinate del profilo sui J assi del sistema di riferimento ortogonale dello spazio $\Re^J$ che ha l'origine in $\mathbf{0}_\mathrm{J}$.

Invece, nel riferimento fattoriale ortonormale dello spazio $\Re^A$, la cui origine $\mathbf{0}_\mathrm{A}$ coincide col baricentro $\overline{\mathbf{r}}$ della nuvola, il medesimo i^{mo} profilo è individuato dalla i^{ma} *riga* della matrice dei fattori $\mathbf{F}$ di ordine $I \times A$ della (4.8.15') e (5.14.6), che d'ora innanzi verrà indicata con $\underline{\mathbf{f}}_i$. E' un vettore colonna le cui A componenti sono le coordinate del profilo sugli assi fattoriali

$$\underline{\mathbf{f}}_i = (f_{i1} \;\; f_{i2} \;\; \dots f_{ia} \;\; \dots f_{iA})^T . \tag{6.4.1}$$

Geometricamente, gli I profili, interpretati come vettori, individuano una nuvola di I punti dotati di massa in uno spazio euclideo A-dimensionale.

Nella matrice $\mathbf{F}$ la riga $\underline{\mathbf{f}}_i$ individua le coordinate del profilo sugli A assi fattoriali e la colonna $\mathbf{f}_a$ quelle di tutti gli I profili sul medesimo asse a. In altri termini, le righe di $\mathbf{F}$ individuano i profili nello spazio fattoriale, le colonne i fattori. perciò la matrice $\mathbf{F}$ può esprimersi sia in termini delle sue I righe [1] che delle sue A colonne

$$\mathbf{F} = \begin{pmatrix} \underline{\mathbf{f}}_1^{\,T} \\ \dots \\ \underline{\mathbf{f}}_i^{\,T} \\ \dots \\ \underline{\mathbf{f}}_I^{\,T} \end{pmatrix} = (\mathbf{f}_1 \;\; \dots \;\; \mathbf{f}_a \;\; \dots \;\; \mathbf{f}_A) . \tag{6.4.2}$$

La massa di ogni profilo non muta cambiando sistema di riferimento ed è quindi $\overline{c}_i = n_{i+}/n_{++}$ se ottenuta con le Corrispondenze semplici e $\overline{c}_i = 1/I$ se con quelle multiple.

---

[1] Una *riga* di $\mathbf{R}$ è indicata con $\mathbf{r}_i^T$ e quella di $\mathbf{F}$ con $\underline{\mathbf{f}}_i^{\,T}$ perché, per le convenzioni fatte nella Sez. 2.2, un vettore è sempre definito come vettore colonna.

Nella partizione in H gruppi degli I profili $\underline{\mathbf{f}}_i$ con massa $\overline{c}_i$, il *numero di profili* del gruppo generico $h = 1, 2, \ldots, H$, verrà indicato con I_h, il profilo medio del gruppo con $\overline{\underline{\mathbf{f}}}(h)$ e la sua *massa* con $\overline{c}(h)$. Questa è la somma delle masse dei profili $\underline{\mathbf{f}}_{i'}$ del gruppo

$$\overline{c}(h) = \sum_{i'=1}^{I_h} \overline{c}_{i'} \tag{6.4.3}$$

e nel caso delle Corrispondenze Multiple la massa di un gruppo, per la (5.9.3), è la frazione di individui nel gruppo. Geometricamente, il profilo medio rappresenta il *baricentro* del gruppo, la cui coordinata su un asse fattoriale di rango $a = 1, 2, \ldots, A$ è la media ponderata delle coordinate sull'asse dei soli profili che costituiscono il gruppo

$$\overline{\underline{\mathbf{f}}}(h) = \frac{1}{\overline{c}(h)} \sum_{i'=1}^{I_h} \overline{c}_{i'}\, \underline{\mathbf{f}}_{i'} \qquad \text{e} \qquad \overline{\underline{f}}_a(h) = \frac{1}{\overline{c}(h)} \sum_{i'=1}^{I_h} \overline{c}_{i'}\, f_{i'a}. \tag{6.4.4}$$

Per esempio, nella Sez. 6.7 si vedrà che aggregando i profili delle regioni della matrice *Spettacoli*, le cui coordinate fattoriali e masse sono tratte dalla TAV. 6.3 e riportate nello specchietto qui sotto, il secondo gruppo stabile $h = \mathbf{2}$ risulta costituito da due sole regioni: Lazio e Molise

i	*Regione*	*Profilo*	*Coordinate*	*Massa*
12	Lazio	$\underline{\mathbf{f}}_{12}$	$+0.060, -0.188$	$+0.122$
14	Molise	$\underline{\mathbf{f}}_{14}$	$+0.112, -0.147$	$+0.002$

La massa di tale gruppo per la 6.4.3 è dunque

$$\overline{c}(\mathbf{2}) = \overline{c}_{12} + \overline{c}_{14} = 0.122 + 0.002 = 0.124,$$

mentre per la (6.4.4) il profilo medio del gruppo risulta

$$\overline{\underline{\mathbf{f}}}(\mathbf{2}) = \frac{0.122}{0.124} \times \begin{pmatrix} +0.060 \\ -0.112 \end{pmatrix} + \frac{0.002}{0.124} \times \begin{pmatrix} +0.112 \\ -0.147 \end{pmatrix} = \begin{pmatrix} -0.059 \\ -0.176 \end{pmatrix}.$$

Le sue coordinate sui due assi fattoriali sono dunque $\overline{\underline{f}}_1(\mathbf{2}) = -0.059$ e $\overline{\underline{f}}_2(\mathbf{2}) = -0.176$.

Il baricentro dei baricentri $\overline{\underline{\mathbf{f}}}(h)$ degli H gruppi di una partizione non è altro che il baricentro $\mathbf{0}_{\mathrm{A}}$ della nuvola di profili e la somma delle masse di

tutti i gruppi vale 1 , ossia

$$\sum_{h=1}^{H} \overline{c}(h)\, \underline{\overline{\mathbf{f}}}(h) = \mathbf{0}_{\mathrm{A}} \qquad \text{e} \qquad \sum_{h=1}^{H} \overline{c}(h) = 1.$$

Allo stesso modo, il numero di profili nei gruppi è pari al numero di profili rilevati

$$\sum_{h=1}^{H} I_h = I.$$

Il *numero di dimensioni* dello spazio fattoriale, e quindi di coordinate dei profili, per la (4.8.14') è $A = min\,(I, J) - 1$ se ottenuto da un'analisi delle Corrispondenze semplici di una matrice di contingenza di ordine $I \times J$ mentre, per la (5.14.3), è $A = J - Q$ nel caso delle Corrispondenze Multiple di una matrice del tipo $individui \times modalità$ con J modalità di Q variabili, codificata in forma disgiuntiva completa.

Abitualmente, ci si limita a considerare le proiezioni dei profili in un sottospazio individuato dai *primi* $A^* < A$ assi fattoriali. L'individuazione del numero A^* di assi da conservare è un punto delicato. Di solito è il numero di assi che si riesce ad interpretare. Nel caso della matrice Spettacoli, con $A = 7$, una buona approssimazione della configurazione si ottiene limitandosi ai primi $A^* = 2$. Invece, per l'ascolto delle trasmissioni radio, la nuvola di $I = 400$ individui è nello spazio fattoriale ad $A = 17$ dimensioni, ma di queste soltanto le prime $A^* = 8$ sono state considerate nella Sez. 5.20 in base al criterio che la loro inerzia che superava l'inerzia media. Il prossimo Capitolo 7 fornirà ulteriori strumenti per individuare A^* .

Questa operazione, da un punto di vista geometrico corrisponde ad un filtraggio, a uno 'smoothing' della configurazione geometrica della nuvola. Trascurare una parte dell'informazione geometrica è vantaggioso perché fa accelerare i calcoli ed elimina buona parte delle fluttuazioni statistiche che potrebbero confondere fenomeni importanti.

Coerentemente, la proiezione di un profilo $\underline{\mathbf{f}}_i$ di $\Re^A$ nel sottospazio $\Re^{A^*}$ di dimensione $A^* < A$ dovrebbe scriversi $\underline{\mathbf{f}}_i{}^*$, ma per non appesantire eccessivamente la simbologia, verrà indicata ancora con $\underline{\mathbf{f}}_i$. La dimensione del sottospazio in cui si trova verrà desunta dal contesto.

6.5 - Inerzie di una partizione

Nella Sez. 3.4 è stato dimostrato il teorema di Huygens relativo all'inerzia di una nuvola di profili e stabilita la relazione (3.4.3) tra le inerzie

riferite all'origine e al baricentro. Applicato ora alla nuvola degli I profili $\mathbf{r}_i$ delle righe nello spazio $\Re^J$, il teorema stabilisce che

$$In_{\mathbf{x}} = In_{\overline{\mathbf{r}}} + d^2_D(\mathbf{x}, \overline{\mathbf{r}}). \tag{6.5.1}$$

Dunque, l'inerzia complessiva della nuvola dei profili riferita a un punto qualsiasi $\mathbf{x}$, diverso dal baricentro $\overline{\mathbf{r}}$, può essere scomposta nella somma di due termini: l'inerzia complessiva della nuvola riferita al suo baricentro $\overline{\mathbf{r}}$ e l'inerzia del singolo punto $\mathbf{x}$ rispetto al baricentro della nuvola, considerando che nel punto $\mathbf{x}$ sia concentrata tutta la massa della nuvola che nel caso dei profili vale 1.

Se la nuvola dei profili è immersa nello spazio fattoriale $\Re^A$ ove il medesimo i^{mo} profilo ha le coordinate (6.4.1), la (6.5.1) diventa

$$In_{\mathbf{x}} = In_{\mathbf{0}_\mathrm{A}} + d^2(\mathbf{x}, \mathbf{0}_\mathrm{A}) \tag{6.5.2}$$

dove ora le coordinate x_a del punto $\mathbf{x}$ sono quelle fattoriali e la distanza è quella euclidea canonica e non più quella distribuzionale (4.8.1).

Espressa in termini di cordinate sugli assi fattoriali, la (6.5.2) si scrive

$$\sum_{a=1}^{A} \sum_{i=1}^{I} \overline{c}_i \, (f_{ia} - x_a)^2 = \sum_{a=1}^{A} \sum_{i=1}^{I} \overline{c}_i \, f_{ia}^2 + \sum_{a=1}^{A} x_a^2. \tag{6.5.3}$$

perché la massa complessiva della nuvola $\sum_i \overline{c}_i = 1$.

L'espressione (6.5.2) si riferisce a una nuvola considerata come gruppo unico, ma può essere estesa facilmente a una sua partizione in H gruppi, ciascuno con I_h profili e massa $\overline{c}(h)$. Per essi è $\sum_h I_h = I$ e $\sum_h \overline{c}(h) = 1$. L'*inerzia totale* dei profili ripetto al baricentro della nuvola, origine degli assi, è

$$In_{\mathbf{0}_\mathrm{A}} = \sum_{i=1}^{I} \overline{c}_i \, d^2(\underline{\mathbf{f}}_i, \mathbf{0}_\mathrm{A}) \tag{6.5.4}$$

ed è eguale alla somma delle inerzie, o autovalori, ottenuti dall'Analisi delle Corrispondenze.

Nella Sez. 3.4 si è visto che nella nuvola l'inerzia complessiva è minima quando è calcolata rispetto al baricentro. Ciò vale anche per ciascun gruppo della nuvola. Infatti, se si cerca di individuare le A coordinate di un punto $x_a(h)$, rispetto alle quali risulti minima l'inerzia interna di un gruppo

$$\sum_{a=1}^{A} \sum_{i'=1}^{I_h} \overline{c}_{i'} \, \left(f_{i'a} - x_a(h)\right)^2 = minimo$$

dove la somma su i' interessa soltanto gli I_h profili del gruppo h in considerazione. Individuare le coordinate significa esprimere le $x_a(h)$ in funzione delle $f_{i'a}$ degli I_h profili del gruppo. Derivando rispetto alla coordinata incognita $x_a(h)$ ed annullando la derivata si ottiene che su ogni asse $a = 1, 2, \ldots, A$, deve valere la condizione

$$2 \sum_{i'=1}^{I_h} \overline{c}_{i'} \, (f_{i'a} - x_a(h)) = 0 \quad \text{ossia} \quad \sum_{i'=1}^{I_h} \overline{c}_{i'} \, x_a(h) = \sum_{i'=1}^{I_h} \overline{c}_{i'} \, f_{i'a}$$

e quindi il minimo dell'inerzia nel gruppo h si ottiene quando la coordinata del punto su ogni asse è

$$x_a(h) = \frac{\sum_{i'=1}^{I_h} \overline{c}_{i'} \, f_{i'a}}{\sum_{i'=1}^{I_h} \overline{c}_{i'}} = \frac{\sum_{i'=1}^{I_h} \overline{c}_{i'} \, f_{i'a}}{\overline{c}(h)} = \underline{\overline{f}}_a(h)$$

ossia quando il punto è il baricentro del gruppo.

La dispersione dei profili del gruppo va dunque valutata rispetto al profilo medio ponderato del gruppo, il baricentro. Di conseguenza, l'*inerzia interna* al gruppo è così definita

$$In_w(h) \stackrel{\text{def}}{=} \sum_{i'=1}^{I_h} \overline{c}_{i'} \, d^2\big(\underline{\mathbf{f}}_{i'}, \underline{\overline{\mathbf{f}}}(h)\big) \tag{6.5.5}$$

dove la somma interessa i soli profili $\underline{\mathbf{f}}_{i'}$ che fanno parte del gruppo. L'inerzia interna misura la compattezza del gruppo: tanto più è piccola tanto più il gruppo è compatto. Geometricamente, ciò indica che i profili del gruppo sono molto vicini tra loro, matematicamente, che hanno coordinate sugli assi quasi eguali.

L'inerzia *nei gruppi*, o *within clusters*, è definita come somma delle inerzie interne (6.5.5) degli H gruppi che costituiscono la partizione

$$In_w \stackrel{\text{def}}{=} \sum_{h=1}^{H} In_w(h). \tag{6.5.6}$$

E' quindi un indicatore globale della compattezza dei gruppi della partizione: tanto più In_w è piccola tanto più gli H gruppi sono complessivamente compatti.

Per misurare la dispersione degli H gruppi rispetto al loro baricentro comune coincidente con quello $\mathbf{0}_{\mathrm{A}}$ della nuvola, si sostituisce alla nuvola degli I profili la nuvola degli H baricentri $\underline{\overline{\mathbf{f}}}(h)$ dei gruppi, assegnando a ciascuno la massa $\overline{c}(h)$ del gruppo.

L'inerzia *tra* un gruppo e il baricentro della nuvola è definita come il prodotto della massa complessiva del gruppo per il quadrato della distanza euclidea canonica tra i due baricentri: quello del gruppo e quello della nuvola, origine degli assi fattoriali.

$$In_b(h) \stackrel{\text{def}}{=} \overline{c}(h)\, d^2\big(\underline{\overline{\mathbf{f}}}(h), \mathbf{0}_{\mathrm{A}}\big)\,. \tag{6.5.7}$$

L'inerzia *tra gruppi*, o *between clusters*, e baricentro della nuvola è la somma delle inerzie (6.5.7) tra i baricentri di ogni gruppo della partizione e il loro baricentro comune,

$$In_b \stackrel{\text{def}}{=} \sum_{h=1}^{H} In_b(h) \tag{6.5.8}$$

ed è un indice complessivo della dispersione dei gruppi attorno al baricentro della nuvola: più è grande, più i gruppi sono ben separati. Il che non garantisce però che i gruppi siano ben distinguibili. L'inerzia In_b gioca un ruolo importante nel metodo di aggregazione gerarchico descritto nella Sez. 6.11.

Il *teorema di Huygens* applicato alla partizione, afferma che l'inerzia complessiva della nuvola (6.5.4) è pari alla somma dell'inerzia *nei* gruppi (6.5.6) e dell'inerzia *tra* gruppi (6.5.8) e baricentro della nuvola [1]

$$In_{\mathbf{0}_{\mathrm{A}}} = In_w + In_b. \tag{6.5.9}$$

Questa importante proprietà dell'inerzia ha come conseguenza che minimizzare l'inerzia In_w nei gruppi implica automaticamente massimizzare l'inerzia In_b tra gruppi perché l'inerzia complessiva $In_{\mathbf{0}_{\mathrm{A}}}$ è fissa per ogni nuvola di profili dal momento che è fissata la loro configurazione. In altri termini, compattezza e separazione dei gruppi sono legate: più i gruppi sono omogenei e compatti, più sono separati, anche se non sempre ben distinguibili, e viceversa.

La TAV. 6.1 illustra in forma grafica il teorema di Huygens nel caso di una configurazione piana, $A = 2$, di una nuvola costituita da $I = 5$ profili. In alto, l'inerzia complessiva della nuvola rispetto al suo baricentro, per la (6.5.4) è

$$In_{\mathbf{0}_2} = \sum_{1}^{5} \overline{c}_i\, d_i^2 = \overline{c}_1\, d_1^2 + \overline{c}_2\, d_2^2 + \ldots + \overline{c}_5\, d_5^2$$

---

[1] La scomposizione dell'inerzia (6.5.9) è analoga a quella dell'Analisi della Varianza nel caso di una sola variabile.

ove $\overline{c}_i$ è la massa del singolo profilo e $d_i^2 = d^2(\underline{\mathbf{f}}_i, \mathbf{0}_2)$ la sua distanza euclidea dal baricentro.

Se i 5 profili vengono ripartiti in $H = 2$ gruppi, in basso, il primo costituito da $I_1 = 2$ e il secondo da $I_2 = 3$ profili, le inerzie *interne* ai due gruppi, riferite ai loro baricentri, per la (6.5.5) sono

$$In_w(1) = \overline{c}_1\,\delta_1^2 + \overline{c}_2\,\delta_2^2 + \overline{c}_3\,\delta_3^2 \qquad \text{e} \qquad In_w(2) = \overline{c}_4\,\delta_4^2 + \overline{c}_5\,\delta_5^2$$

dove $\delta_i^2 = d^2(\underline{\mathbf{f}}_i, \overline{\underline{\mathbf{f}}}(h))$ con $h = 1, 2$.

L'inerzia complessiva interna ai gruppi, detta inerzia *nei* (*within*) gruppi, per la (6.5.6), è

$$In_w = In_w(1) + In_w(2) = \sum_1^3 \overline{c}_i\,\delta_i^2 + \sum_1^2 \overline{c}_i\,\delta_i^2.$$

L'inerzia tra (*between*) i 2 gruppi e il baricentro, per la (6.5.7) è

$$In_b = (\overline{c}_1 + \overline{c}_2 + \overline{c}_3)\,D_1^2 + (\overline{c}_4 + \overline{c}_5)\,D_2^2 = \overline{c}(1)\,D_1^2 + \overline{c}(2)\,D_2^2$$

dove le masse dei due gruppi sono rispettivamente

$$\overline{c}_1 + \overline{c}_2 + \overline{c}_3 = \overline{c}(1) \qquad \text{e} \qquad \overline{c}_4 + \overline{c}_5 = \overline{c}(2)$$

mentre le distanze dal baricentro della nuvola sono $D_h^2 = d^2(\overline{\underline{\mathbf{f}}}(h), \mathbf{0}_2)$.

In base al teorema di Huygens (6.5.9) nella partizione in esame di 5 profili in 2 gruppi, le inerzie sono così legate

$$\sum_1^5 \overline{c}_i\,d_i^2 = \Big(\sum_1^3 \overline{c}_i\,\delta_i^2 + \sum_1^2 \overline{c}_i\,\delta_i^2\Big) + \Big(\overline{c}(1)\,D_1^2 + \overline{c}(2)\,D_2^2\Big)$$
$$In_{\mathbf{0}_\mathrm{A}} = In_w + In_b.$$

Per dimostrare questa importante relazione, occorre prima ricordare la (2.5.3) che esprime la distanza euclidea canonica tra due profili $\mathbf{x}$ e $\mathbf{y}$ tramite il prodotto scalare: $d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T\,(\mathbf{x} - \mathbf{y})$. Questo può essere esteso al caso di più profili, per esempio a 3

$$(\mathbf{x}_1 - \mathbf{x}_2 + \mathbf{x}_2 - \mathbf{x}_3)^T\,(\mathbf{x}_1 - \mathbf{x}_2 + \mathbf{x}_2 - \mathbf{x}_3) =$$
$$(\mathbf{x}_1 - \mathbf{x}_2)^T(\mathbf{x}_1 - \mathbf{x}_2) + (\mathbf{x}_2 - \mathbf{x}_3)^T(\mathbf{x}_2 - \mathbf{x}_3) + 2\,(\mathbf{x}_1 - \mathbf{x}_2)^T\,(\mathbf{x}_2 - \mathbf{x}_3)$$

perché $(\mathbf{x}_1 - \mathbf{x}_2)^T(\mathbf{x}_2 - \mathbf{x}_3) = (\mathbf{x}_2 - \mathbf{x}_3)^T(\mathbf{x}_1 - \mathbf{x}_2)$ sono scalari, ossia numeri.

In una partizione con H gruppi l'inerzia complessiva (6.5.4) della nuvola di profili può essere scomposta negli H contributi che all'inerzia danno

i profili $\underline{\mathbf{f}}_{i'}$ dei singoli gruppi, ossia

$$In_{\mathbf{0}_\mathrm{A}} = \sum_{i=1}^{I} \overline{c}_i \, d^2(\underline{\mathbf{f}}_i, \mathbf{0}_\mathrm{A}) = \sum_{h=1}^{H} \sum_{i'=1}^{I_h} \overline{c}_{i'} \, d^2(\underline{\mathbf{f}}_{i'}, \mathbf{0}_\mathrm{A}) \tag{6.5.10}$$

La distanza può essere ora espressa tramite il prodotto scalare, tenendo anche conto di quanto appena visto poco sopra

$$\begin{aligned} d^2\left(\underline{\mathbf{f}}_{i'}, \mathbf{0}_\mathrm{A}\right) &= \left(\underline{\mathbf{f}}_{i'} - \overline{\underline{\mathbf{f}}}(h) + \overline{\underline{\mathbf{f}}}(h) - \mathbf{0}_\mathrm{A}\right)^T \left(\underline{\mathbf{f}}_{i'} - \overline{\underline{\mathbf{f}}}(h) + \overline{\underline{\mathbf{f}}}(h) - \mathbf{0}_\mathrm{A}\right) \\ &= \left(\underline{\mathbf{f}}_{i'} - \overline{\underline{\mathbf{f}}}(h)\right)^T \left(\underline{\mathbf{f}}_{i'} - \overline{\underline{\mathbf{f}}}(h)\right) + \left(\overline{\underline{\mathbf{f}}}(h) - \mathbf{0}_\mathrm{A}\right)^T \left(\overline{\underline{\mathbf{f}}}(h) - \mathbf{0}_\mathrm{A}\right) \\ &+ 2 \left(\underline{\mathbf{f}}_{i'} - \overline{\underline{\mathbf{f}}}(h)\right)^T \left(\overline{\underline{\mathbf{f}}}(h) - \mathbf{0}_\mathrm{A}\right) \\ &= d^2\left(\underline{\mathbf{f}}_{i'}, \overline{\underline{\mathbf{f}}}(h)\right) + d^2\left(\overline{\underline{\mathbf{f}}}(h), \mathbf{0}_\mathrm{A}\right) + 2 \left(\underline{\mathbf{f}}_{i'} - \overline{\underline{\mathbf{f}}}(h)\right)^T \left(\overline{\underline{\mathbf{f}}}(h) - \mathbf{0}_\mathrm{A}\right). \end{aligned}$$

Sostituendo questa espressione delle distanze in quella (6.5.10) dell'inerzia complessiva, si ha

$$\begin{aligned} In_{\mathbf{0}_\mathrm{A}} = &\sum_{h=1}^{H} \sum_{i'=1}^{I_h} \overline{c}_{i'} \, d^2\left(\underline{\mathbf{f}}_{i'}, \overline{\underline{\mathbf{f}}}(h)\right) + \sum_{h=1}^{H} \sum_{i'=1}^{I_h} \overline{c}_{i'} \, d^2\left(\overline{\underline{\mathbf{f}}}(h), \mathbf{0}_\mathrm{A}\right) \\ &+ 2 \sum_{h=1}^{H} \sum_{i'=1}^{I_h} \overline{c}_{i'} \left(\underline{\mathbf{f}}_{i'} - \overline{\underline{\mathbf{f}}}(h)\right)^T \left(\overline{\underline{\mathbf{f}}}(h) - \mathbf{0}_\mathrm{A}\right) \end{aligned}$$

Tenendo conto della (6.5.5) e (6.5.6) si vede che il primo termine della somma è l'inerzia nei gruppi In_w, somma delle inerzie interne ai singoli gruppi, e poi che per la (6.5.7) e (6.5.8) il secondo è l'inerzia In_b tra gruppi e origine e infine che il terzo è nullo. Infatti, per la proprietà (2.5.1') del prodotto scalare e per la definizione (6.4.4) di baricentro di un gruppo, è

$$\begin{aligned} &\sum_{h=1}^{H} \sum_{i'=1}^{I_h} \overline{c}_{i'} \left(\underline{\mathbf{f}}_{i'} - \overline{\underline{\mathbf{f}}}(h)\right)^T \left(\overline{\underline{\mathbf{f}}}(h) - \mathbf{0}_\mathrm{A}\right) = \\ &\qquad \sum_{h=1}^{H} \left(\overline{\underline{\mathbf{f}}}(h) - \mathbf{0}_\mathrm{A}\right) \left(\sum_{i'=1}^{I_h} \overline{c}_{i'} \underline{\mathbf{f}}_{i'} - \overline{c}(h) \overline{\underline{\mathbf{f}}}(h)\right)^T = 0 \end{aligned}$$

Il teorema di Huygens (6.5.9) è così dimostrato.

6.6 - Aggregazione a centri mobili

Primo dei due metodi di aggregazione presentati in questo Capitolo, è un metodo iterativo di tipo non gerarchico per costruire direttamente una partizione di I profili in H gruppi. Richiede che venga preventivamente assegnata una partizione provvisoria iniziale, o almeno i centri di aggregazione degli

H gruppi. La partizione viene 'migliorata' passo dopo passo riattribuendo i profili al loro centro più prossimo. Poiché ad ogni passo l'inerzia In_w nei gruppi diminuisce, come si vedrà al termine della Sezione, il metodo converge rapidamente verso una partizione finale, senza garanzia alcuna che sia quella ottimale. Il metodo ha tendenza a costruire dei gruppi 'sferici' e separati, nel senso che geometricamente non si compenetrano, ma che tuttavia possono essere contigui.

La velocità di convergenza rende questo metodo particolarmente utile quando i profili da aggregare sono molto numerosi, come ad esempio nel caso di vaste indagini, ma per rendere maggiormente comprensibile il metodo è opportuno prendere ad esempio un caso di ridotte dimensioni. Verranno perciò considerati i profili delle 20 regioni della matrice *Spettacoli* descritti dalle $A^* = 2$ prime coordinate fattoriali f_{i1} e f_{i2}. La matrice $\mathbf{F}^*$ è dunque di ordine 20×2 ed è riportata nella TAV. 6.2.

Con 20 profili, e in mancanza di altre informazioni, sembra ragionevole ricercare una partizione in $H = 4$ gruppi. Il procedimento si sviluppa allora nei passi seguenti:

passo 1: Invece di imporre la partizione iniziale, ossia l'appartenenza di ogni profilo a un gruppo, ci si limita ad estrarre a caso, senza reimmissione, 4 dei 20 profili. Risultano estratti i profili di queste regioni

i	*Regione*	*Profilo*	f_{i1}	f_{i2}	*Gruppo*
4	Trentino AA	$\underline{\mathbf{f}}_4$	+0.222	+0.080	1
8	Emilia Rom.	$\underline{\mathbf{f}}_8$	+0.103	+0.045	2
13	Abruzzi	$\underline{\mathbf{f}}_{13}$	+0.219	−0.062	3
18	Calabria	$\underline{\mathbf{f}}_{18}$	+0.061	+0.102	4

Questi profili vengono considerati provvisoriamente come *centri* di aggregazione e restano fissi per tutta la durata di questo primo passo. Ad essi viene assegnato un numero progressivo da 1 a H in base all'ordine di estrazione.

Per costruire la partizione iniziale, la matrice $\mathbf{F}^*$ di ordine 20×2 delle coordinate fattoriali dei profili viene letta sequenzialmente: un record, ossia una profilo, alla volta. Di ogni profilo $\underline{\mathbf{f}}_i$ si calcola la distanza [1] dai 4 centri provvisori. Per esempio, il profilo $\underline{\mathbf{f}}_1$ del

[1] Per distanza si deve sempre intendere il quadrato della distanza euclidea canonica.

Piemonte che dalla TAV. 6.2 risulta avere coordinate $f_{11} = 0.008$ e $f_{12} = 0.194$, ha queste distanze dai 4 centri

Gruppo	*Distanze di* $\underline{\mathbf{f}}_1$ *dai* 4 *centri provvisori*
1	$d^2(\underline{\mathbf{f}}_1, \underline{\mathbf{f}}_4) = (0.008 - 0.222)^2 + (0.194 - 0.080)^2 = 0.059$
2	$d^2(\underline{\mathbf{f}}_1, \underline{\mathbf{f}}_8) = (0.008 - 0.103)^2 + (0.194 - 0.045)^2 = 0.031$
3	$d^2(\underline{\mathbf{f}}_1, \underline{\mathbf{f}}_{13}) = (0.008 - 0.219)^2 + (0.194 + 0.062)^2 = 0.110$
4	$d^2(\underline{\mathbf{f}}_1, \underline{\mathbf{f}}_{18}) = (0.008 - 0.061)^2 + (0.194 - 0.102)^2 = 0.011$

Il profilo del Piemonte è perciò più vicino al centro 4 e viene dunque assegnato al quarto gruppo. Il procedimento di assegnazione è ripetuto per tutti i profili. Le distanze dei profili dai centri e il gruppo a cui vengono assegnati sono riportati nella TAV. 6.3. Riassumendo, i gruppi che risultano hanno questa consistenza

Gruppo	*Centro*		*Num.*	*Massa*
1	+0.222	+0.080	2	0.028
2	+0.103	+0.045	6	0.417
3	+0.219	−0.062	5	0.216
4	+0.061	+0.102	7	0.339

Lo specchietto va letto in questo modo. Il centro provvisorio per aggregare i profili nel gruppo 1 ha per coordinate $(0.222, 0.080)$ (sono quelle del Trentino AA. estratto a caso) ed è costituito da 2 profili: sono Trentino AA. e Valle d'Aosta, come risulta dall'ultima colonna della TAV. 6.3. La massa del gruppo è 0.028, somma delle masse dei due profili. E così via per gli altri gruppi. Si noti come il centro del gruppo *non* sia il baricentro del gruppo perché nell'attribuzione dei profili si è tenuto conto delle distanze, ma *non* delle masse.

passo 2: In questo passo vengono anzitutto calcolati i profili medi ponderati, o baricentri, dei 4 gruppi ottenuti nel passo precedente, tenendo conto ora delle coordinate *e anche* delle masse. Per esempio, dall'ultima colonna della TAV. 6.3, che riporta i dettagli della partizione ottenuta al passo precedente, si vede che il gruppo 1, costituito dai profili della Valle d'Aosta ($i = 2$) che ha per coordinate $f_{21} = 0.305$ e $f_{22} = 0.867$ e massa $\overline{c}_2 = 0.002$, e del Trentino ($i = 4$) con $f_{41} = 0.222$ e $f_{42} = 0.080$ e massa $\overline{c}_4 = 0.026$, le coordinate del

baricentro, in base alla (6.4.4) sono

$$\overline{f}_1(1) = (0.002 \times 0.305 + 0.026 \times 0.222)/(0.002 + 0.026) = 0.227$$
$$\overline{f}_2(1) = (0.002 \times 0.867 + 0.026 \times 0.080)/(0.002 + 0.026) = 0.126.$$

I risultati compaiono nella TAV. 6.5, nel riquadro in alto a sinistra, che riporta, ad ogni passo, le coordinate del baricentro del gruppo, il numero di profili che vi appartengono e la massa complessiva. I 4 baricentri vengono ora considerati come *nuovi centri* di aggregazione. La partizione precedente viene 'azzerata' e la matrice $\mathbf{F}^*$ letta di nuovo sequenzialmente ricalcolando le distanze di ogni profilo dai nuovi centri e riassegnando i 20 profili al gruppo del centro più vicino.

Dall'ultima colonna di TAV. 6.4 si ricava che la partizione ottenuta al termine di questo secondo passo ha la consistenza mostrata nello specchietto qui sotto che di ogni gruppo riporta le coordinate del (bari)centro calcolate all'inizio di questo passo, il numero di profili in ogni gruppo e la massa complessiva

Gruppo	*Centro*		*Num.*	*Massa*
1	+0.227	+0.126	2	0.028
2	+0.172	+0.045	6	0.045
3	+0.061	−0.223	5	0.216
4	−0.145	+0.130	7	0.339

passo 3: Di ogni gruppo appena individuato si calcola il baricentro che viene considerato come nuovo centro di aggregazione. La matrice $\mathbf{F}^*$ delle coordinate viene riletta per stabilire la distanza di ogni profilo dai 4 nuovi centri e l'appartenenza a uno dei gruppi. Le coordinate dei baricentri, la numerosità dei gruppi e la massa complessiva sono ora

Gruppo	*Centro*		*Num.*	*Massa*
1	+0.277	+0.126	4	0.058
2	+0.070	+0.004	9	0.483
3	+0.042	−0.223	3	0.194
4	−0.201	+0.109	1	0.265

passo 4: Come nei precedenti, anche in questo passo e nei successivi vengono calcolati i baricentri che vengono poi presi per nuovi centri di ag-

gregazione. Vengono calcolate le distanze dei profili dai nuovi centri e, in base a esse, stabilita l'appartenenza di ogni profilo a uno dei 4 gruppi, ecc. La procedura si arresta se in due passi successivi nessun profilo passa da un gruppo ad un altro, ossia se la partizione ottenuta non cambia, e / o se viene soddisfatto un prestabilito criterio d'arresto.

Anche se il metodo converge rapidamente in un numero di passi che di rado supera la decina, la sua automatizzazione richiede che venga precisato il criterio d'arresto dell'algoritmo. Il criterio abitualmente adottato è quello basato sull'arresto della diminuzione dell'inerzia In_w nei gruppi da un passo al successivo. L'inerzia $In_w(h)$ all'interno di un gruppo $h = 1, 2, \ldots, H$, è definita nella (6.5.5), mentre, per la (6.5.6), l'inerzia nei gruppi è in questo caso

$$In_w = In_w(1) + In_w(2) + In_w(3) + In_w(4).$$

Nel caso dell'esempio la diminuzione dell'inerzia In_w nei gruppi si arresta dopo il 5^o passo, come si vede dal seguente specchietto

	Inerzie interne ai gruppi				
Passo	$In_w(1)$	$In_w(2)$	$In_w(3)$	$In_w(4)$	In_w
1^o	0.001	0.002	0.014	0.044	0.061
2^o	0.005	0.002	0.001	0.022	0.030
3^o	0.002	0.007	0.001	0.014	0.024
4^o	0.001	0.007	0.001	0.000	0.009
5^o	0.000	0.007	0.001	0.000	0.008
6^o	0.000	0.007	0.001	0.000	0.008

Si noti come da un passo all'altro l'inerzia interna di un singolo gruppo possa anche aumentare perché aumenta il numero di profili nel gruppo, ma necessariamente l'inerzia In_w nei gruppi diminuisce ad ogni passo in cui ci sia uno spostamento di profili da un gruppo all'altro perché lo spostamento avviene soltanto se comporta una diminuzione della distanza da un centro. Questo si comprende se si tiene presente che a ogni passo la distanza dei profili è calcolata rispetto ai baricentri dei gruppi ottenuti nel passo precedente e si dimostra facilmente.

Al termine del passo p, con $p > 1$, il gruppo h sia costituito da $I_h(p)$ profili, ciascuno con massa $\overline{c}_i$. La massa del gruppo sia $\overline{c}(h, p)$. Il centro di aggregazione del gruppo non è altro che il baricentro dei profili che appartenevano allo stesso nel passo precedente $p - 1$, ossia $\underline{\overline{\mathbf{f}}}(h, p-1)$. Con

questi dati è possibile calcolare l'inerzia dei profili aggregati nel gruppo in questo passo, rispetto al loro centro di aggregazione

$$Q(p,h) = \sum_{i'=1}^{I_h(p)} \overline{c}_{i'}\, d^2\left(\underline{\mathbf{f}}_{i'}, \overline{\underline{\mathbf{f}}}(h,p-1)\right)$$

Sommando queste inerzie per ogni gruppo $h = 1, 2, \ldots, H$ si ottiene l'inerzia complessiva dei profili della nuvola rispetto ai centri di aggregazione durante il passo p

$$Q(p) = \sum_{h=1}^{H} Q(p,h). \tag{6.6.1}$$

Al termine del passo successivo $p+1$, i profili appartenenti al gruppo h siano diventati $I_h(p+1)$. Il centro di aggregazione $\overline{\underline{\mathbf{f}}}(h,p)$ è ora il baricentro del gruppo degli $I_h(p)$ profili che lo componevano nel passo p precedente. Analogamente a quanto fatto sopra, si può calcolare l'inerzia del gruppo rispetto al nuovo centro di aggregazione del passo $p+1$

$$Q(p+1,h) = \sum_{i'=1}^{I_h(p+1)} \overline{c}_{i'}\, d^2\left(\underline{\mathbf{f}}_{i'}, \overline{\underline{\mathbf{f}}}(h,p)\right)$$

e quella complessiva di tutti i gruppi

$$Q(p+1) = \sum_{h=1}^{H} Q(p+1,h). \tag{6.6.2}$$

Rispetto al medesimo centro, si può anche calcolare l'inerzia (6.5.5) *interna* al gruppo degli $I_h(p)$ profili, che lo costituivano nel passo p precedente, e dei quali il centro $\overline{\underline{\mathbf{f}}}(h,p)$ è il baricentro

$$In_w(p) = \sum_{i'=1}^{I_h(p)} \overline{c}_{i'}\, d^2\left(\underline{\mathbf{f}}_{i'}, \overline{\underline{\mathbf{f}}}(h,p)\right)$$

e l'inerzia (6.5.6) *nei* gruppi esistenti al passo p

$$In_w(p) = \sum_{h=1}^{H} In_w(h,p). \tag{6.6.3}$$

Si può ora mostrare che ad ogni passo p tra le tre inerzie vale sempre la diseguaglianza

$$Q(p) \geq In_w(p) \geq Q(p+1)$$

per cui l'inerzia nei gruppi $In_w(p)$ è funzione non crescente del passo p. Detto in altri termini, il metodo di aggregazione ai Cenntri Mobili crea partizioni successive che hanno gruppi sempre più omogenei. Che sia $Q(p) \geq In_w(p)$ deriva dal teorema di Huygens (6.5.9) applicato alle inerzie calcolate nel passo p

$$Q(p) = In_w(p) + \sum_{h=1}^{H} \overline{c}(h,p)\, d^2\left(\underline{\overline{\mathbf{f}}}(h,p+1), \underline{\overline{\mathbf{f}}}(h,p)\right).$$

Poiché l'inerzia tra baricentri non è mai negativa, la prima diseguaglianza è dimostrata. Invece, $In_w(p) \geq Q(p+1)$ deriva dal fatto che gli $I_h(p+1)$ profili sono più vicini al centro $\underline{\overline{\mathbf{f}}}(h,p)$ di quanto non lo fossero gli $I_h(p)$ profili nel passo p precedente. Nella riattribuzione di un profilo da un centro a un altro, la distanza dal centro che lo perde è maggiore di quella che ha dal centro a cui viene aggregato: le distanze non possono che diminuire durante ogni riallocazione. perciò, *complessivamente*, la riattribuzione dei profili fa diminuire l'inerzia In_w nei gruppi. Ciò significa che dal punto di vista della compattezza dei gruppi la nuova partizione è 'migliore' della precedente, anche se non 'la' migliore.

Il processo di aggregazione converge quindi sicuramente verso una partizione finale e viene arrestato quando l'inerzia *nei* gruppi In_w cessa di diminuire in modo sensibile da un passo al successivo. D'altra parte, questo criterio d'arresto puù rivelarsi inefficace con particolari configurazioni della nuvola che, per ragioni numeriche, provochino oscillazioni continue tra due partizioni. perciò di solito, a questo criterio si associa anche quello sul numero massimo di passi consentiti.

Se la rapidità di convergenza è il primo grande vantaggio del metodo di aggregazione ai Centri Mobili, esiste anche un secondo vantaggio non meno importante. Poiché la matrice $\mathbf{F}^*$ delle coordinate fattoriali dei profili viene letta sequenzialmente, un record alla volta ad ogni passo, essa puù risiedere in un file su disco ed essere quindi anche molto voluminosa. Il metodo è così applicabile senza difficoltà a casi anche con decine di migliaia di profili.

Il punto debole del metodo di raggruppamento attorno ai Centri Mobili, anzi un serio inconveniente, è che la composizione dei gruppi nella partizione finale dipende spesso dalla *scelta iniziale* dei centri provvisori [1] al passo 1, come evidenzia l'esempio schematizzato nella TAV. 6.5.

[1] La partizione finale dipende anche parzialmente dall'ordine dei profili nella matrice delle coordinate fattoriali.

6.7 - Strategia dei gruppi stabili

I due punti di debolezza del metodo dei Centri Mobili: conoscenza 'a priori' del numero H di gruppi e convergenza verso una partizione finale che frequentemente dipende dalla scelta iniziale dei centri di aggregazione, vengono in gran parte attenuati con la strategia dei Gruppi Stabili che sfrutta invece i due punti di forza del metodo: rapida convergenza e semplicità di calcolo. Sostanzialmente essa consiste nel ripetere due o più volte il metodo dei Centri Mobili, variando ogni volta i centri inizialmente estratti ed eventualmente anche il loro numero. I *gruppi stabili* sono gruppi non vuoti costituiti da profili che in tutte le partizione finali risultano aggregati sempre insieme perché fisicamente molto vicini. Geometricamente corrispondono a zone di maggiore densità di profili nella nuvola. Così da un lato si individuano i gruppi più compatti e dall'altro si mettono in evidenza quelli composti da profili isolati e di incerta attribuzione. Questo fatto attenua il primo punto di debolezza: la conoscenza 'a priori' del numero H di gruppi della partizione. Il numero di gruppi stabili dipende quindi dalla configurazione della nuvola di profili e non dal numero di gruppi imposto. In pratica bastano due sole partizioni per individuare i gruppi stabili. Ecco come ciò avviene.

Nel caso dell'esempio, si può eseguire una seconda volta il processo a Centri Mobili estraendo ancora $H = 4$ centri iniziali provvisori. Questa volta sono i profili 2-Valle d'Aosta, 14-Molise, 16-Puglia e 19-Sicilia che vengono numerati nell'ordine da 1 a 4. Come si vede nella TAV. 6.2, il metodo stavolta converge in 2 soli passi e la partizione finale, indicata con **II** e riportata nell'ultima colonna, è sensibilmente diversa da quella ottenuta precedentemente, indicata con **I**, che convergeva in 5 passi. Per esempio, il profilo 1-Piemonte è ora nel gruppo 3, mentre nella prima risultava nel gruppo 2. Invece il profilo 2-Valle d'Aosta compare nel gruppo 1 in entrambe le partizioni.

I gruppi stabili si ottengono incrociando le due partizioni ottenute, come è fatto nella TAV. 6.6. Nelle righe della tabella sono i 4 gruppi della prima partizione convergente in 5 passi ed indicata col simbolo **I**, e nelle colonne quelli della seconda convergente in 2 passi ed indicata con **II**. All'incrocio di una riga e di una colonna sono riportati gli indici i dei profili che nella prima partizione sono finiti nel gruppo della riga e nella seconda in quello della colonna. Per esempio, nella TAV. 6.2 si vede che il profilo di 1-Piemonte compare nel gruppo 2 nella prima partizione e nel gruppo 3 nella seconda. Nella TAV. 6.6 va quindi nella casella $(\mathbf{I}, 2)$ e $(\mathbf{II}, 3)$ e così via per

gli altri profili. Dei $4 \times 4 = 16$ gruppi stabili possibili nel caso dell'esempio, soltanto 6 non sono vuoti e sono numerati da (**1°**) a (**6°**) per valore decrescente della massa. Il numero d'ordine è indicato tra parentesi sotto ciascun gruppo.

I profili delle regioni 5-Veneto e 19-Sicilia che sono eccentrici ed isolati nella mappa della TAV. 37 formano gruppo a sé. Invece, il gruppo **1°** riunisce 11 profili e il 65% della massa complessiva. Questi profili sono collocati centralmente sulla mappa, intorno all'origine degli assi.

6.8 - Metodi aggregativi gerarchici

Il metodo di aggregazione a Centri Mobili della Sez. 6.6 appartiene ai cosidetti metodi *non gerarchici* perché conduce a un'unica partizione in cui il numero di gruppi è specificato 'a priori'.

Il metodo di Ward che verrà presentato nella Sez. 6.11 appartiene invece ai metodi aggregativi [1] *gerarchici ascendenti* i quali generano un'intera successione di partizioni organizzate gerarchicamente e con sempre meno gruppi. La successione inizia con la partizione della nuvola in I gruppi, considerando ogni profilo da aggregare come un singolo gruppo, e si conclude con la partizione in unico gruppo che riunisce tutti i profili. Ogni nuovo gruppo è riassunto e rappresentato dal baricentro dei due gruppi aggregati, facendo l'ipotesi quindi che i gruppi siano costituiti da profili simili, le cui caratteristiche sono quelle del loro profilo medio. La massa di ogni gruppo è la somma delle masse dei profili del gruppo.

I metodi gerarchici vengono impiegati quando non è chiaro 'a priori' quanti possano essere i gruppi. Sostanzialmente essi si differenziano tra loro per l'indice di dissimilarità tra gruppi che adottano, ossia per come è misurata la 'distanza' tra coppie di gruppi dotati di massa. Lo schema aggregativo tradizionale è di questo tipo. Dati I gruppi e scelto l'indice di dissimilarità, si procede a calcolare la matrice simmetrica di ordine $I \times I$ delle 'distanze' tra coppie di gruppi per individuare quella a 'distanza' *minima*. I due gruppi, considerati del tutto equivalenti, vengono aggregati in un nuovo gruppo. Sugli $I - 1$ gruppi rimasti si ripete il processo di calcolo delle

---

[1] Alternativi ai metodi aggregativi sono i metodi *divisivi*, o scissori, nei quali si procede per successive divisioni della nuvola di profili. Sono poco usati perché ad ogni passo è necessario esaminare tutte le divisioni possibili. Con H gruppi queste sono $2^{H-1} - 1$, mentre in un metodo aggregativo basta esaminare al massimo $(H-1)(H-2)/2$ distanze tra gruppi.

distanze e di aggregazione binaria ottenendo una successione di partizioni con sempre meno gruppi fino a quella finale con un unico gruppo. La successione delle partizioni è rappresentabile graficamente come un albero gerarchico con nodi e rami, descritto nella Sez. 6.12. La partizione 'ottimale' va individuata tra quelle della successione, in base a criteri di scelta che verranno esposti fra poco, ma prima di arrivare a questo occorre meglio definire che cosa si intende per partizione e per gruppo di una partizione, ma sopratutto occorre individuare il criterio di dissimilarità più adatto a misurare la 'distanza' tra due *gruppi di profili*. All'approfondimento di questi concetti sono dedicate le prossime due Sezioni.

6.9 - Gerarchia di partizioni e indice di dissimilarità

Formalmente, una *partizione* $\mathcal{P}$ di un insieme $\mathcal{X}$ di profili è la suddivisione dell'insieme in H sottoinsiemi $\mathcal{X}_1 \, , \mathcal{X}_2 \, , \ldots \, , \mathcal{X}_h \, , \ldots \, , \mathcal{X}_H$, detti *gruppi*, che soddisfano queste tre condizioni

1 - $\mathcal{X}_h \neq \emptyset \qquad h = 1, 2, \ldots, H$;
2 - $\mathcal{X}_h \cap \mathcal{X}_k = \emptyset \qquad k \neq h = 1, 2, \ldots, H$
3 - $\bigcup_{h=1}^{H} \mathcal{X}_h = \mathcal{X}$.

La prima condizione impone che nessun gruppo deve essere vuoto: deve contenere almeno un profilo; la seconda che i gruppi devono essere disgiunti: un profilo può appartenere a un solo gruppo [1] ; la terza, infine che ogni profilo deve trovar posto in un gruppo della partizione. Ne deriva che una partizione può anche essere costituita da tanti gruppi quanti sono i profili da aggregare: un profilo in ogni gruppo. Per esempio, la partizione in $H = 4$ gruppi degli $I = 4$ profili dell'insieme $\mathcal{X} = \{\underline{\mathbf{f}}_1, \, \underline{\mathbf{f}}_2, \, \underline{\mathbf{f}}_3, \, \underline{\mathbf{f}}_4\}$ è unica

$$\mathcal{P}_1 = \{\mathcal{X}_1 = \{\underline{\mathbf{f}}_1\}, \;\; \mathcal{X}_2 = \{\underline{\mathbf{f}}_2\}, \;\; \mathcal{X}_3 = \{\underline{\mathbf{f}}_3\}, \;\; \mathcal{X}_4 = \{\underline{\mathbf{f}}_4\}\} \, ,$$

mentre una partizione in $H = 2$ gruppi di $\mathcal{X}$ portebbe essere

$$\mathcal{P}_1 = \{\mathcal{X}_1 = \{\underline{\mathbf{f}}_1, \underline{\mathbf{f}}_4\}, \;\; \mathcal{X}_2 = \{\underline{\mathbf{f}}_2, \underline{\mathbf{f}}_3\}\} \, .$$

Invece, i sottoinsiemi

$$\mathcal{X}_1 = \{\underline{\mathbf{f}}_1, \underline{\mathbf{f}}_4\}, \;\; \mathcal{X}_2 = \{\underline{\mathbf{f}}_2\}$$

---

[1] Nei metodi di aggregazione sfumata (*fuzzy clustering* o *clumping*) si pongono nello stesso gruppo (clump) le coppie di profili il cui indice di dissimilarità non è inferiore a una soglia prefissata. I clump ottenuti possono non essere disgiunti perché un profilo può appartenere più di un clump. Gli algoritmi di clumping richiedono calcoli più complessi e spesso l'interpretazione dei clump risulta difficoltosa.

non costituiscono una partizione di $\mathcal{X}$ perché il profilo $\underline{\mathbf{f}}_3$ di $\mathcal{X}$ non trova posto in alcun gruppo.

Una sequenza di partizioni costituisce poi una *gerarchia di partizioni* quando i gruppi della partizione ottenuta a un certo livello della gerarchia comprendono i gruppi della partizione del livello precedente e ciò accade ad ogni livello della gerarchia.

Per esempio, se l'insieme $\mathcal{X}$ precedente si trova suddiviso al primo livello nella partizione $\mathcal{P}_1$ in $H = 4$ gruppi

$$\mathcal{P}_1 = \{\mathcal{X}_1 = \{\underline{\mathbf{f}}_1\},\ \mathcal{X}_2 = \{\underline{\mathbf{f}}_2\},\ \mathcal{X}_3 = \{\underline{\mathbf{f}}_3\},\ \mathcal{X}_4 = \{\underline{\mathbf{f}}_4\}\}$$

e le partizioni dei livelli successivi [1] fossero

$$\begin{aligned}
\mathcal{P}_2 &= \{\mathcal{X}_5 = \{\mathcal{X}_1, \mathcal{X}_2\},\ \mathcal{X}_3 = \{\underline{\mathbf{f}}_3\},\ \mathcal{X}_4 = \{\underline{\mathbf{f}}_4\}\} \\
\mathcal{P}_3 &= \{\mathcal{X}_5 = \{\mathcal{X}_1, \mathcal{X}_2\},\ \mathcal{X}_6 = \{\mathcal{X}_3, \mathcal{X}_4\}\} \\
\mathcal{P}_4 &= \{\mathcal{X}_7 = \{\mathcal{X}_5, \mathcal{X}_6\}\},
\end{aligned}$$

la successione a 4 livelli $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4$ costituisce una gerarchia perché i gruppi di un livello sono contenuti nei gruppi della partizione del livello successivo.

Una gerarchia di partizioni costituisce poi una *gerarchia indicizzata* di partizioni, quando a *ogni* livello della gerarchia corrisponde il valore di un *indice di dissimilarità* che indica la 'distanza' tra i due gruppi aggregati a quel livello. Esso è il *minimo* valore dell'indice di tutte le coppie di gruppi presenti nella partizione di quel livello. Questi valori minimi ad ogni livello della gerarchia permettono di tracciare l'albero gerarchico delle partizioni, detto anche dendrogramma, come verrà illustrato nella Sez. 6.12.

Nei metodi aggregativi gerarchici i gruppi sono riassunti e rappresentati dal profilo medio ponderato, ossia dal loro baricentro, per cui qualunque indice di dissimilaritità che si voglia adottare deve tenere conto, oltre che della distanza tra baricentri, anche delle loro masse perché possono essere molto diverse. Quello che viene correntemente utilizzato nell'aggregazione di profili è l'indice di dissimilarità di Ward generalizzato [2]. Se i baricentri di due gruppi sono $\overline{\underline{\mathbf{f}}}(h)$ e $\overline{\underline{\mathbf{f}}}(h')$ e hanno massa $\overline{c}(h)$ e $\overline{c}(h')$ rispettivamente,

---

[1] Abitualmente, in una gerarchia di partizioni i gruppi che via via si formano vengono numerati progressivamente.

[2] J. H. Ward, *Hierarchical grouping to optimize an objective function*, Journal of the American Statistical Association; (1963), vol. **58**, pag. 236-244.

l'indice di dissimilarità di Ward tra i due gruppi è così definito

$$d^2_W\left(\overline{\underline{\mathbf{f}}}(h),\overline{\underline{\mathbf{f}}}(h')\right) \stackrel{\text{def}}{=} \frac{\overline{c}(h)\,\overline{c}(h')}{\overline{c}(h)+\overline{c}(h')}\,d^2\left(\overline{\underline{\mathbf{f}}}(h),\overline{\underline{\mathbf{f}}}(h')\right). \tag{6.9.1}$$

In fisica, in un sistema costituito da due corpi di massa m_1 e m_2, il rapporto $m_1\,m_2/(m_1+m_2)$ è detto *massa ridotta* del sistema. perciò, l'indice (6.9.1) può interpretarsi come un'inerzia tra due gruppi ove la massa è la loro massa ridotta. In effetti, si vedrà nella prossima Sezione che l'indice (6.9.1) ha notevoli proprietà legate alle inerzie nella partizione, fatto questo che ne fa una scelta quasi obbligata in presenza di gruppi di profili, dato che l'Analisi delle Corrispondenze è proprio un metodo di scomposizione dell'inerzia, come si è visto nella Sez. 3.15. Esso inoltre si presta alla costruzione di un albero gerarchico delle partizioni e soddisfa al cosidetto 'assioma della mediana' il quale garantisce che l'aggregazione di due gruppi non rimette in discussione le aggregazioni precedenti. [1].

L'effetto della massa ridotta nell'indice (6.9.1) è quello di rendere più 'vicini' i gruppi con massa minore, privilegiandoli nell'aggregazione. Per esempio, se i baricentri di tre gruppi $\overline{\underline{\mathbf{f}}}(1)$, $\overline{\underline{\mathbf{f}}}(2)$ e $\overline{\underline{\mathbf{f}}}(3)$ sono ai vertici di un triangolo isoscele e le reciproche distanze sono assunte, per semplicità, unitarie, mentre le masse dei primi due gruppi sono 0.1 e quella del terzo, estremamente più piccola, è per esempio 10^{-6}, e quindi

$$d^2\left(\overline{\underline{\mathbf{f}}}(1),\overline{\underline{\mathbf{f}}}(2)\right) = d^2\left(\overline{\underline{\mathbf{f}}}(1),\overline{\underline{\mathbf{f}}}(3)\right) = d^2\left(\overline{\underline{\mathbf{f}}}(2),\overline{\underline{\mathbf{f}}}(3)\right) = 1$$

e

$$\overline{c}(1) = \overline{c}(2) = 0.1 \qquad \overline{c}(3) = 10^{-6}$$

i valori dell'indice (6.9.1) di dissimilarità fra i 3 gruppi risultano

$$d^2_W\left(\overline{\underline{\mathbf{f}}}(1),\overline{\underline{\mathbf{f}}}(2)\right) = \frac{0.1\times 0.1}{2\times 0.1}\times 1 = 0.05$$

$$d^2_W\left(\overline{\underline{\mathbf{f}}}(1),\overline{\underline{\mathbf{f}}}(3)\right) = d^2_W\left(\overline{\underline{\mathbf{f}}}(2),\overline{\underline{\mathbf{f}}}(3)\right) \simeq 10^{-7}/0.1 = 10^{-6}.$$

L'indice deforma quindi lo spazio dei profili in modo tale che pur essendo i primi due gruppi 'lontani' di 0.05, il terzo gruppo, estremamente leggero, risulta *simultaneamente* molto 'vicino' ad entrambi.

---

[1] Il lettore troverà la dimostrazione in Diday et al. *Eléments d'Analyse des données*, 1982, Dunod ed. alla pag. 172

6.10 - Proprietà dell'indice di dissimilarità di Ward

L'indice (6.9.1) gode di due notevoli proprietà, entrambe legate alle inerzie della partizione. La prima è che esso è una misura dell'inerzia di due gruppi rispetto al comune baricentro. La seconda, che misura anche la diminuzione dell'inerzia tra gruppi che avviene quando due gruppi vengono aggregati. Al solito, il baricentro riassume e rappresenta il gruppo, la cui massa è la somma delle masse dei profili del gruppo.

1 - verrà ora mostrata la prima proprietà dell'indice (6.9.1) di Ward. Se $\overline{\underline{\mathbf{f}}}(h)$ e $\overline{\underline{\mathbf{f}}}(h')$ sono i baricentri di due gruppi e $\overline{c}(h)$ e $\overline{c}(h')$ le loro masse, il loro baricentro comune è il profilo medio ponderato

$$\overline{\underline{\mathbf{f}}} = \frac{\overline{c}(h)}{\overline{c}(h)+\overline{c}(h')}\,\overline{\underline{\mathbf{f}}}(h) + \frac{\overline{c}(h')}{\overline{c}(h)+\overline{c}(h')}\,\overline{\underline{\mathbf{f}}}(h') \tag{6.10.1}$$

che ha una massa pari alla somma $\overline{c}(h)+\overline{c}(h')$ delle masse dei due gruppi. La coordinata $\overline{\underline{f}}_a$ del baricentro su ciascun asse $a = 1, 2, \ldots, A$ è la media ponderata delle coordinate dei due gruppi

$$\overline{\underline{f}}_a = \frac{\overline{c}(h)}{\overline{c}(h)+\overline{c}(h')}\,\overline{\underline{f}}_a(h) + \frac{\overline{c}(h')}{\overline{c}(h)+\overline{c}(h')}\,\overline{\underline{f}}_a(h'). \tag{6.10.2}$$

L'inerzia interna dei due gruppi rispetto al loro comune baricentro (6.10.1) si scrive

$$In_w(h) + In_w(h') = \overline{c}(h)\, d^2\left(\overline{\underline{\mathbf{f}}}(h), \overline{\underline{\mathbf{f}}}\right) + \overline{c}(h')\, d^2\left(\overline{\underline{\mathbf{f}}}(h'), \overline{\underline{\mathbf{f}}}\right). \tag{6.10.3}$$

Nella Sez. 2.5 si è visto che la distanza tra due profili può esprimersi tramite il prodotto scalare, per cui l'inerzia $In(h)$ del primo gruppo, tenendo conto della (6.10.1), risulta

$$\begin{aligned} In_w(h) &= \overline{c}(h)\, d^2\left(\overline{\underline{\mathbf{f}}}(h), \overline{\underline{\mathbf{f}}}\right) = \overline{c}(h)\,\left(\overline{\underline{\mathbf{f}}}(h) - \overline{\underline{\mathbf{f}}}\right)^T \left(\overline{\underline{\mathbf{f}}}(h) - \overline{\underline{\mathbf{f}}}\right) \\ &= \overline{c}(h)\,\frac{\overline{c}(h')^2}{\overline{c}(h)+\overline{c}(h')}\,\left(\overline{\underline{\mathbf{f}}}(h) - \overline{\underline{\mathbf{f}}}(h')\right)^T \left(\overline{\underline{\mathbf{f}}}(h) - \overline{\underline{\mathbf{f}}}(h')\right). \end{aligned}$$

Allo stesso modo si ottiene l'inerzia del secondo gruppo rispetto al baricentro comune

$$In_w(h') = \overline{c}(h')\,\frac{\overline{c}(h)^2}{\overline{c}(h)+\overline{c}(h')}\,\left(\overline{\underline{\mathbf{f}}}(h) - \overline{\underline{\mathbf{f}}}(h')\right)^T \left(\overline{\underline{\mathbf{f}}}(h) - \overline{\underline{\mathbf{f}}}(h')\right)$$

per cui la somma (6.10.3) delle inerzie dei due gruppi rispetto al loro baricentro comune, diventa

$$\begin{aligned} In_w(h) + In_w(h') &= \frac{\overline{c}(h)\,\overline{c}(h')}{\overline{c}(h)+\overline{c}(h')}\,\left(\underline{\overline{\mathbf{f}}}(h)-\underline{\overline{\mathbf{f}}}(h')\right)^T\left(\underline{\overline{\mathbf{f}}}(h)-\underline{\overline{\mathbf{f}}}(h')\right) \\ &= \frac{\overline{c}(h)\,\overline{c}(h')}{\overline{c}(h)+\overline{c}(h')}\,d^2\left(\underline{\overline{\mathbf{f}}}(h),\underline{\overline{\mathbf{f}}}(h')\right) = d_W^2\left(\underline{\overline{\mathbf{f}}}(h),\underline{\overline{\mathbf{f}}}(h')\right) \end{aligned} \tag{6.10.4}$$

che è appunto l'indice di dissimilarità (6.9.1).

2 - verrà mostrato ora che l'indice (6.9.1) di Ward corrisponde alla diminuzione dell'inerzia In_b tra gruppi e baricentro della nuvola che si verifica aggregando due gruppi con baricentri $\underline{\overline{\mathbf{f}}}(h)$ e $\underline{\overline{\mathbf{f}}}(h')$ e masse $\overline{c}(h)$ e $\overline{c}(h')$ rispettivamente.

Prima dell'aggregazione, le inerzie tra i due gruppi e il baricentro della nuvola per la (6.5.7) sono

$$\begin{aligned} In_b(h) &= \overline{c}(h)\,d^2(\underline{\overline{\mathbf{f}}}(h),\mathbf{0}_{\mathrm{A}}) = \overline{c}(h)\,\left(\underline{\overline{\mathbf{f}}}(h)-\mathbf{0}_{\mathrm{A}}\right)^T\left(\underline{\overline{\mathbf{f}}}(h)-\mathbf{0}_{\mathrm{A}}\right) \\ In_b(h') &= \overline{c}(h')\,d^2(\underline{\overline{\mathbf{f}}}(h'),\mathbf{0}_{\mathrm{A}}) = \overline{c}(h')\,\left(\underline{\overline{\mathbf{f}}}(h')-\mathbf{0}_{\mathrm{A}}\right)^T\left(\underline{\overline{\mathbf{f}}}(h')-\mathbf{0}_{\mathrm{A}}\right). \end{aligned}$$

La somma di queste due inerzie è il contributo dei due gruppi all'inerzia In_b tra gruppi della partizione e loro baricentro. Tenendone esplicitamente conto si può scrivere che prima dell'aggregazione dei due gruppi l'inerzia (6.5.8) tra gruppi della partizione e baricentro è

$$In_b(prima) = costante + In_b(h) + In_b(h') \tag{6.10.5}$$

dove *costante* è la somma dei contributi all'inerzia dei rimanenti $H-2$ gruppi. Esprimendo la distanza tramite i prodotti scalari, la (6.10.5) diventa

$$In_b(prima) = costante + \overline{c}(h)\,\underline{\overline{\mathbf{f}}}(h)^T\,\underline{\overline{\mathbf{f}}}(h) + \overline{c}(h')\,\underline{\overline{\mathbf{f}}}(h')^T\,\underline{\overline{\mathbf{f}}}(h') \tag{6.10.6}$$

Dopo l'aggregazione dei due gruppi, si crea un nuovo gruppo che ha per baricentro il profilo medio (6.10.1) e per massa la somma $\overline{c}(h)+\overline{c}(h')$ delle masse dei due gruppi. perciò, esplicitando il contributo del nuovo gruppo all'inerzia In_b tra gruppi della partizione e il loro baricentro si può scrivere

$$In_b(dopo) = costante + (\overline{c}(h)+\overline{c}(h'))\;d^2\left(\underline{\overline{\mathbf{f}}},\mathbf{0}_{\mathrm{A}}\right) \tag{6.10.7}$$

dove *costante* è la stessa che compare nella (6.10.5) e (6.10.6) perché è il contributo all'inerzia dei gruppi che non sono toccati dall'aggregazione. Esprimendo la distanza nella (6.10.7) tramite i prodotti scalari e tenendo conto

della (6.10.1) si ottiene

$$\begin{aligned} In_b(dopo) &= costante + \frac{1}{\overline{c}(h)+\overline{c}(h')} \\ &\qquad \left(\overline{c}(h)\,\underline{\overline{\mathbf{f}}}(h)+\overline{c}(h')\,\underline{\overline{\mathbf{f}}}(h')\right)^T \left(\overline{c}(h)\,\underline{\overline{\mathbf{f}}}(h)+\overline{c}(h')\,\underline{\overline{\mathbf{f}}}(h')\right) \\ &= costante + \frac{1}{\overline{c}(h)+\overline{c}(h')} \\ &\qquad \left(\overline{c}(h)^2\,\underline{\overline{\mathbf{f}}}(h)^T\underline{\overline{\mathbf{f}}}(h)+2\,\overline{c}(h)\,\overline{c}(h')\,\underline{\overline{\mathbf{f}}}(h)^T\underline{\overline{\mathbf{f}}}(h')+\overline{c}(h)^2\,\underline{\overline{\mathbf{f}}}(h')^T\underline{\overline{\mathbf{f}}}(h')\right). \end{aligned} \tag{6.10.8}$$

L'aggregazione provoca una diminuzione dell'inerzia In_b tra i gruppi della partizione. Essa continua a decrescere mano a mano che prodede il processo di aggregazione, fino ad annullarsi quando gli ultimi due gruppi si aggregano nel baricentro $\mathbf{0}_{\mathrm{A}}$ della nuvola. Indicando con

$$\Delta In_b = In_b(prima) - In_b(dopo) \tag{6.10.9}$$

la diminuzione dell'inerzia tra gruppi causata dall'aggregazione, questa per la (6.10.6) e (6.10.7) risulta

$$\begin{aligned} \Delta In_b &= \overline{c}(h)\,\underline{\overline{\mathbf{f}}}(h)^T\,\underline{\overline{\mathbf{f}}}(h)+\overline{c}(h')\,\underline{\overline{\mathbf{f}}}(h')^T\,\underline{\overline{\mathbf{f}}}(h') - \frac{1}{\overline{c}(h)+\overline{c}(h')} \\ &\qquad \left(\overline{c}(h)^2\,\underline{\overline{\mathbf{f}}}(h)^T\,\underline{\overline{\mathbf{f}}}(h)+2\,\overline{c}(h)\,\overline{c}(h')\,\underline{\overline{\mathbf{f}}}(h)^T\,\underline{\overline{\mathbf{f}}}(h')+\overline{c}(h)^2\,\underline{\overline{\mathbf{f}}}(h')^T\,\underline{\overline{\mathbf{f}}}(h')\right) \\ &= \left(\overline{c}(h)-\frac{1}{\overline{c}(h)+\overline{c}(h')}\right)\underline{\overline{\mathbf{f}}}(h)^T\,\underline{\overline{\mathbf{f}}}(h) - 2\,\frac{\overline{c}(h)\,\overline{c}(h')}{\overline{c}(h)+\overline{c}(h')}\,\underline{\overline{\mathbf{f}}}(h)^T\,\underline{\overline{\mathbf{f}}}(h') \\ &\qquad + \left(\overline{c}(h')-\frac{1}{\overline{c}(h)+\overline{c}(h')}\right)\underline{\overline{\mathbf{f}}}(h')^T\underline{\overline{\mathbf{f}}}(h') \\ &= \frac{\overline{c}(h)\,\overline{c}(h')}{\overline{c}(h)+\overline{c}(h')}\,\left(\underline{\overline{\mathbf{f}}}(h)^T\,\underline{\overline{\mathbf{f}}}(h)-2\,\underline{\overline{\mathbf{f}}}(h)^T\,\underline{\overline{\mathbf{f}}}(h')+\underline{\overline{\mathbf{f}}}(h')^T\,\underline{\overline{\mathbf{f}}}(h')\right). \end{aligned}$$

Quest'ultima espressione è il prodotto della massa ridotta dei due gruppi che si aggregano per la distanza, al quadrato, tra i loro baricentri, per cui la diminuzione dell'inerzia tra gruppi è

$$\begin{aligned} \Delta In_b &= In_b(prima) - In_b(dopo) \\ &= \frac{\overline{c}(h)\,\overline{c}(h')}{\overline{c}(h)+\overline{c}(h')}\,d^2\left(\underline{\overline{\mathbf{f}}}(h),\underline{\overline{\mathbf{f}}}(h')\right) = d_W^2\left(\underline{\overline{\mathbf{f}}}(h),\underline{\overline{\mathbf{f}}}(h')\right). \end{aligned} \tag{6.10.10}$$

Assumere come indice di dissimilarità quello di Ward (6.9.1) equivale dunque a misurare la 'distanza' tra gruppi in base alla diminuzione ΔIn_b dell'inerzia che si verificherebbe se venissero aggregati. Si può quindi considerare ΔIn_b come un nuovo 'indice di dissimilarità' che viene detto 'indice del livello gerarchico' o, più brevemente, 'indice di livello'. Ad ogni passo del processo ven-

gono aggregati i due gruppi che provocano la minima diminuzione dell'inerzia tra gruppi.

Una conseguenza importante diretta è che sommando le diminuzioni di inerzia ΔIn_b di tutti i livelli gerarchici $l = 1, 2, \ldots, I-1$, si ottiene l'inerzia totale della nuvola di profili

$$\sum_{l=1}^{I-1} \Delta In_b(l) = \sum_{l=1}^{I-1} (In_b(l-1) - In_b(l)) \tag{6.10.11}$$
$$= In_{\mathbf{0}_\mathrm{A}} - In_b(1) + \sum_{l=2}^{I-1} (In_b(l-1) - In_b(l)) = In_{\mathbf{0}_\mathrm{A}}$$

perché al livello iniziale l'inerzia complessiva degli I profili è $In_b(l=0) = In_{\mathbf{0}_\mathrm{A}}$, mentre all'ultimo con un unico gruppo è $In_b(l = I-1) = 0$ e le inerzie dei passi intermedi si cancellano a vicenda.

Poiché anche l'Analisi delle Corrispondenze scompone l'inerzia complessiva della nuvola in A inerzie sugli assi fattoriali, o autovalori, tra le due scomposizioni sussiste l'importante relazione

$$\sum_{l=1}^{I-1} \Delta In_b(l) = In_{\mathbf{0}_\mathrm{A}} = \sum_{a=1}^{A} \lambda_a. \tag{6.10.12}$$

In entrambi i casi la scomposizione è ordinata, nel senso che le $I-1$ diminuzioni di inerzia ΔIn_b e le A inerzie λ_a sugli assi fattoriali sono ordinate rispettivamente per valori crescenti e decrescenti

$$\Delta In_b(l=1) \leq \Delta In_b(l=2) \leq \ldots \qquad \text{e} \qquad \lambda_1 \geq \lambda_2 \geq \ldots$$

Le diminuzioni di inerzia sono fortemente dipendenti le une dalle altre a causa della struttura gerarchica, mentre gli assi fattoriali sui quali l'inerzia delle proiezioni dei profili è misurata, sono ortogonali. Inoltre, la più alta diminuzione d'inerzia, quella che si verifica all'ultimo passo di aggregazione, è sempre inferiore al primo e più alto autovalore $\Delta In_b(l = I-1) < \lambda_1$. La scomposizione fattoriale è più efficace nel rivelare la variabilità della struttura dei profili.

6.11 - Metodo gerarchico ascendente di Ward

L'indice (6.9.1) di dissimilarità tra gruppi caratterizza un particolare metodo di aggregazione gerarchica strettamente collegato all'Analisi delle Corrispondenze. Esso permette di costruire una gerarchia di partizioni via via sempre meno fine, privilegiando ogni volta l'aggregazione che comporta

la *minima* diminuzione dell'inerzia (6.10.9) tra gruppi. Questa diminuisce al crescere del livello della gerarchia: massima al livello iniziale quando ogni gruppo è costituito da un solo profilo e nulla al livello finale quando tutta la nuvola di profili si è ridotta per aggregazioni successive a un solo gruppo che contiene tutti i profili. In dettaglio, la partizione iniziale ha I gruppi e l'inerzia tra gruppi è $In_b = In_{\mathbf{0}_A}$. La successiva, al livello di aggregazione $l = 1$, ha $I - 1$ gruppi a causa dell'aggregazione dei due gruppi $\overline{\mathbf{f}}(h)$ e $\overline{\mathbf{f}}(h')$ e ad essa corrisponde la più alta inerzia tra gruppi possibile: $In_b = In_{\mathbf{0}_A} - \Delta In_b(l = 1)$ perché $\Delta In_b(l = 1)$ è la minima diminuzione che si può ottenere aggregando due gruppi della partizione iniziale. La partizione è dunque globalmente ottimale in base al criterio di aggregazione scelto, ma di nessun interesse pratico. Procedendo con le aggregazioni, al passo $l = 2$ si forma la partizione successiva con $I - 2$ gruppi e con la più alta inerzia $In_b(p = 2)$ *ottenibile* dalla partizione precedente a $I - 1$ gruppi. A partire dalla partizione con $I - 2$ gruppi in poi non vi è più certezza che le partizioni siano 'ottimali' tra tutte le possibili partizioni con quel numero di gruppi, ma soltanto che sono *buone* partizioni, in base all'indice adottato, tra quelle permesse dal metodo per cui ogni partizione deve essere ottenuta dalla partizione precedente. Una partizione è considerata 'buona' se $\Delta In_b(l)$ è grande, perché questo indica che i gruppi aggregati erano ben separati.

L'algoritmo *tradizionale* di aggregazione gerarchica ascendente è piuttosto semplice e si si presta bene ad essere programmato. Si sviluppa nei 4 passi seguenti, preceduti da un eventuale passo preliminare quando si tratta di aggregare direttamente dei profili [1]:

passo 0: Ciascun profilo, dotato di massa ed espresso in coordinate fattoriali, viene considerato come gruppo individuale. Si inizia cioè con tanti gruppi quanti sono i profili;

passo 1: Con l'indice di dissimilarità (6.9.1) di Ward si calcola la diminuzione dell'inerzia ΔIn_b che provocherebbe l'aggregazione di ogni coppia di gruppi, costruendo la matrice quadrata e simmetrica detta 'delle distanze';

passo 2: Si individua nella matrice la minima diminuzione dell'inerzia (se ci fossero due valori eguali si sceglie a caso uno dei due). I due gruppi corrispondenti vengono aggregti in un nuovo gruppo. Questi è lo-

[1] Nella Strategia Mista, esposta nella Sez. 6.14, si aggregano invece i gruppi stabili.

calizzato nel baricentro, il profilo medio ponderato, dei due gruppi e ha per massa la somma delle due masse;

passo 3: Si aggiorna la matrice delle 'distanze', ricalcolando le diminuzioni dell'inerzia tra il nuovo gruppo e tutti gli altri gruppi esistenti. I nuovi valori sostituiscono le due righe e le due colonne corrispondenti ai due profili aggregati. L'ordine della matrice si riduce così di 1 ;

passo 4: Si ripetono i passi 2, 3 e 4 finché non resta che un solo gruppo: tutti i profili sono riuniti nello stesso gruppo con baricentro in $\mathbf{0}_{\mathrm{A}}$, il baricentro dell'intera nuvola.

Il principale punto di forza del metodo è lo stesso dei Centri Mobili della Sez. 6.5: la tendenza a formare gruppi 'sferici' con un numero equilibrato di profili. Proprio questa caratteristica rende però difficile aggregare profili isolati o che si trovano in nuvole 'allungate' o 'sfilacciate'. Inoltre le partizioni che fornisce *non* sono ottimali in senso assoluto perché ogni partizione è vincolata dalla partizione formatasi al passo precedente che non viene *mai* rimessa in discussione. La partizione in H gruppi contiene quella del passo precedente con $H+1$ gruppi ed è contenuta in quella del passo successivo con $H-1$ gruppi.

Una difficoltà del metodo tradizionale appena esposto è il ricalcolo delle 'distanze' perché l'espressione (6.9.1) fa intervenire le coordinate dei baricentri dei gruppi. È possibile tuttavia aggiornare direttamente le 'distanze' sfruttando quelle calcolate nel passo precedente. Se $\overline{\underline{\mathbf{f}}}(h)$ e $\overline{\underline{\mathbf{f}}}(h')$ sono i baricentri dei due guppi che vengono aggregati e $\overline{\underline{\mathbf{f}}}$ della (6.11.1) quello del nuovo gruppo con massa $\overline{c}(h)+\overline{c}(h')$ somma delle masse dei due gruppi, la distanza ricalcolata è

$$d^2_W\left(\overline{\underline{\mathbf{f}}}\,,\,\overline{\underline{\mathbf{f}}}(k)\right)=\frac{1}{\overline{c}(h)+\overline{c}(h')+\overline{c}(k)}\Big(\left(\overline{c}(h)+\overline{c}(k)\right)\,d^2_W\left(\overline{\underline{\mathbf{f}}}(h),\overline{\underline{\mathbf{f}}}(k)\right)$$
$$+\left(\overline{c}(h')+\overline{c}(k)\right)\,d^2_W\left(\overline{\underline{\mathbf{f}}}(h'),\overline{\underline{\mathbf{f}}}(k)\right)-\overline{c}(k)\,d^2_W\left(\overline{\underline{\mathbf{f}}}(h),\overline{\underline{\mathbf{f}}}(h')\right)\Big).$$

L' algoritmo tradizionale richiede comunque che la matrice delle 'distanze' sia conservata nella memoria centrale del computer e, pur se simmetrica, in forma quadrata per maggiore rapidità di aggiornamento. L'occupazione di memoria è spesso notevole. Nel caso dell'esempio degli ascolti radiofonici la matrice conterrebbe $I\times I=400\times 400=160\,000$ distanze !

Si preferisce allora mantenere in memoria la matrice $\mathbf{F}^*$ delle coordinate fattoriali dei profili perché l'occupazione risulta inferiore [1] . Nel caso

[1] L'occupazione di memoria si riduce drasticamente con la Strategia Mista

dell'esempio degli ascolti radiofonici il numero di coordinate da conservare in memoria è $I \times A^* = 400 \times 8 = 32\,000$. C'è quindi un notevole risparmio di memoria, ma al prezzo di un aumento del tempo di calcolo perché ad ogni passo devono essere ricalcolate *tutte* le distanze tra coppie di gruppi. Tuttavia, il metodo dei *vicini reciproci* della Sez. 6.13 permette di sveltire enormemente i calcoli.

6.12 - Albero gerarchico e diagramma dei livelli

La successione delle aggregazioni di un processo Gerarchico Ascendente viene rappresentata graficamente sotto forma di albero formato da nodi e rami. La TAV. 6.9 nella parte superiore mostra quello ottenuto aggregando gerarchicamente i 6 gruppi stabili della TAV. 6.5, come prevede la Strategia Mista che verrà esposta nella Sez. 6.14. Per sfruttare la larghezza della pagina, l'albero è stampato orizzontalmente. Esso va letto come un albero genealogico in cui ogni nodo — la creazione di un nuovo gruppo — è considerato 'padre' dei *due* nodi 'figli' — i due gruppi che si aggregano — situati al livello immediatamente inferiore. Meno si deve risalire l'albero per aggregare due gruppi, più la 'parentela' tra i due gruppi è stretta. Il diagramma è noto anche col nome di *dendrogramma* ed è costruito in questo modo. Alla sua base, gli I profili da aggregare costituiscono altrettanti gruppi separati rappresentati da nodi numerati da 1 a I, e detti nodi di base o terminali. Da essi partono dei rami che si riuniscono a coppie in altri nodi — i nuovi gruppi che si formano — e che ricevono progressivamente un numero d'ordine da $I+1$ a $2I-1$. Quest'ultimo nodo alla sommità dell'albero, corrisponde alla partizione in cui tutti i profili sono riuniti in un unico gruppo rappresentato dal baricentro $\mathbf{0}_\mathrm{A}$ dell'intera nuvola. Ci sono dunque $I-1$ aggregazioni per completare la successione e quindi altrettanti nodi nell'albero gerarchico.

A fianco dell'albero una scala graduata a partire da 0, in corrispondenza della base, riporta le diminuzioni dell'inerzia $\Delta In_b(l)$ corrispondenti a ogni nodo. La scala riporta unicamente i *valori minimi*, ma per avere un unico campo di variazione, si preferisce il *tasso d'inerzia* $\nu(l)$ del nodo [1], ossia il rapporto percentuale tra la diminuzione d'inerzia del livello $l = 1, 2, \ldots, I-1$ e la somma di tutte le diminuzioni del processo di aggregazione che, per la

della Sez. 6.14 che prevede di conservare in memoria le coordinate dei soli gruppi stabili.

[1] Ovviamente $\nu(l)$ è calcolato per ogni coppia di gruppi della partizione, ma è solamente il valore minimo che viene riportato sulla scala.

(6.10.12) corrisponde all'inerzia totale della nuvola

$$\nu(l) = \frac{d_W^2\,(\overline{\underline{\mathbf{f}}}(h), \overline{\underline{\mathbf{f}}}(h'))}{\sum_l d_W^2\,(\overline{\underline{\mathbf{f}}}(h), \overline{\underline{\mathbf{f}}}(h'))} \times 100 = \frac{\Delta In_b(l)}{In_{\mathbf{0}_A}} \times 100. \qquad (6.12.1)$$

In tal modo nell'albero gerarchico ogni nodo ha una 'altezza' proporzionale al suo tasso d'inerzia $\nu(l)$.

Un ampio intervallo tra due indici di livello rivela che i due gruppi aggregati erano 'distanti' e suggerisce di effettuare il 'taglio' tra questi due nodi. Si ottengono così H gruppi.

La TAV. 6.9 riporta l'albero gerarchico delle aggregazioni dei 6 gruppi stabili ottenuti nella TAV. 6.8 in alto. La struttura dell'albero gerarchico evidenzia una parentela 'stretta' tra i gruppi $\mathbf{2^o} = \{\underline{\mathbf{f}}_{12}\,, \underline{\mathbf{f}}_{14}\}$ e $\mathbf{5^o} = \{\underline{\mathbf{f}}_{10}\,, \underline{\mathbf{f}}_{13}\,, \underline{\mathbf{f}}_{17}\}$ che, come mostra la TAV. 6.8 in basso, si aggregano formando il gruppo $\mathbf{7^o} = \{\,\mathbf{2^o}, \mathbf{5^o}\}$ al livello $\nu = 1.18\,\%$. Una 'parentela' più lontana sussiste tra i gruppi $\mathbf{5^o}$ e $\mathbf{6^o} = \{\underline{\mathbf{f}}_2\,, \underline{\mathbf{f}}_{20}\}$ che confluiscono in un unico gruppo soltanto a un livello più alto $\nu = 24.81\,\%$. Il gruppo $\mathbf{3^o} = \{\underline{\mathbf{f}}_5\}$ che contiene il profilo Veneto isolato dal primo asse fattoriale è l'ultimo ad essere aggregato.

La TAV. 6.8 riporta i valori dell'indice per tutte le coppie di gruppi stabili dell'esempio. I due gruppi più vicini sono il $\mathbf{2^o}$ e il $\mathbf{5^o}$ per i quali la diminuzione d'inerzia è $\Delta In_b(l=1) = d_W^2(\mathbf{2^o}, \mathbf{5^o}) = 0.000\,661$. Il nuovo gruppo diventa il $\mathbf{7^o}$. Il suo baricentro, ossia il profilo medio ponderato, è calcolato con la (6.4.3) e ha per coordinate $+$, e la sua massa è $\overline{c}(\mathbf{7^o}) = \overline{c}(\mathbf{2^o}) + \overline{c}(\mathbf{5^o}) = 0.124 + 0.036 = 0.286$.

Per quanto visto nella Sez. 6.10, il tasso d'inerzia ν di un nodo è una misura del grado di 'separazione' dei due gruppi che si aggregano. La differenza di livello ν sulla scala indica i gruppi meglio separati e quelli meno. L'indice (6.9.1) adottato, fa sì che il livello dei nodi della gerarchia sia proporzionale non alla distanza tra i baricentri dei due gruppi da aggregare, ma al suo quadrato. Ciò provoca un allungamento degli ultimi rami dell'albero e una compressione dei livelli dei primi. I gruppi aggregati inizialmente appaiono più vicini di quanto in realtà non siano.

La scelta della partizione, e quindi del numero e della composizione dei gruppi, è in parte soggettiva e molto legata ai fini dell'analisi. Si ottiene 'tagliando' il dendrogramma con un tratto orizzontale [1] a un livello di ν

[1] La linea di 'taglio' può anche essere sinousa. Deve soltanto incontrare ogni ramo in un solo punto.

tale per cui i gruppi risultino ben separati, ma non troppo 'poveri' di profili. La partizione ha tanti gruppi quanti sono i rami 'tagliati'. Quando non è ben chiaro a quale livello 'tagliare' il dendrogramma, conviene scegliere le 2 o 3 partizioni che appaiono più evidenti e conservare poi quella meglio interpretabile in termini di variabili attive ed illustrative.

Una rappresentazione che fornisce le stesse informazioni del dendrogramma, ma in forma più comoda per individuare la partizione, è il *diagramma dei livelli* di aggregazione dei nodi. Si tratta di un diagramma a barre orizzontali come quello nella TAV 6.9, in basso, che viene tracciato per gli ultimi nodi, quelli meglio separati e dunque di maggior interesse. Ogni barra corrisponde a un nodo e la sua lunghezza è proporzionale al tasso d'inerzia $\nu(p)$ di aggregazione del nodo: più è lunga più i due gruppi hanno 'resistito' all'aggregazione perché 'lontani'. Le barre sono ordinate dall'alto al basso per valore crescente del tasso d'inerzia. In tal modo le coppie di gruppi che si aggregano ad ogni nodo sono ordinate in base alla loro 'vicinanza' spaziale: in alto i gruppi più 'vicini' e via via scendendo quelli più 'lontani'. Per questo il diagramma va esaminato iniziando con le partizioni in pochi gruppi. Il diagramma va 'tagliato' con un tratto perpendicolare alle barre in corrispondenza delle barre più lunghe dovute a di differenze notevoli del tasso, dette in gergo 'pianerottoli' del diagramma, che rivelano i gruppi meglio separati. A ogni 'taglio' corrisponde una partizione. Va tenuto presente che ogni barra 'tagliata' conta per un gruppo, ma quella più in basso conta per due, dato che rappresenta l'aggregazione dei *due* gruppi finali. In pratica, raramente ci si accontenta di una sola partizione, ma se ne scelgono due o tre.

VOLLE 292 e segg.

6.13 - Algoritmo dei vicini reciproci

Nei metodi tradizionali di aggregazione gerarchica, e così anche in quello di Ward della Sez. 6.11, a ogni passo viene aggregata la sola coppia di gruppi a 'distanza' minima. Nel metodo di aggregazione gerarchica basato sul concetto di vicino reciproco vengono aggregate simultaneamente *tutte* le coppie di gruppi 'più vicini'. Così, le 'distanze' calcolate a ogni passo sono sfruttate al massimo e i passi di aggregazione e i tempi di calcolo notevolmente ridotti.

In una partizione in H gruppi, il gruppo con baricentro $\overline{\underline{\mathbf{f}}}(h)$ è il *prossimo vicino* del gruppo con baricentro $\overline{\underline{\mathbf{f}}}(h')$ se la sua distanza da $\overline{\underline{\mathbf{f}}}(h')$ è *inferiore* a ogni altra distanza di $\overline{\underline{\mathbf{f}}}(h')$ dai baricentri degli altri $H-1$ gruppi.

Questo si verifica quando nella matrice delle distanze la colonna del gruppo $\overline{\underline{\mathbf{f}}}(h')$ ha il valore minimo in corrispondenza della riga $\overline{\underline{\mathbf{f}}}(h)$. Per esempio, nella TAV. 6.7 che riporta le matrici delle distanze calcolate nei primi due passi di aggregazione dei 6 gruppi stabili, nel primo passo (matrice in alto) il gruppo **6°** è prossimo vicino del gruppo **5°** perché tra tutti è quello più vicino. Infatti $d^2_W(\mathbf{5°}, \mathbf{6°}) = 0.005\,253$ è la minima distanza che compare nella colonna del **6°** gruppo.

Due gruppi sono detti *vicini reciproci* se ciascuno è il più prossimo vicino dell'altro e viceversa. Anche se non si tratta della coppia a distanza minima in assoluto, due gruppi vicini reciproci sono a distanza minima rispetto a quelle che hanno dagli altri gruppi. perché questo accada è necessario che la distanza all'incrocio della riga corrispondente a $\overline{\underline{\mathbf{f}}}(h)$ e della colonna a $\overline{\underline{\mathbf{f}}}(h')$ sia la più piccola sia della riga che della colonna. Così si vede subito che i due gruppi **5°** e **6°** *non* sono vicini reciproci perché la loro distanza $(0.005\,253)$ è superiore a quella $(0.000\,661)$ tra il **2°** e il **5°** gruppo. L'unica coppia di vicini reciproci nel primo passo di aggregazione è quella a distanza minima, costituita appunto dai gruppi **2°** e **5°**.

L'algoritmo si basa sulla proprietà che se due gruppi vengono aggregati, come il **2°** e il **5°** nel **7°**, la distanza del nuovo gruppo **7°** dai gruppi restanti non è mai inferiore alla distanza che ciascuno dei due gruppi aggregati aveva dagli altri gruppi. Se $\overline{\underline{\mathbf{f}}}(k)$ indica uno qualunque dei gruppi restanti dopo l'aggregazione, escludendo il gruppo aggregato, allora

$$d^2_W\left(\mathbf{7°}, \overline{\underline{\mathbf{f}}}(k)\right) \geq min\left(d^2_W\left(\mathbf{2°}, \overline{\underline{\mathbf{f}}}(k)\right), d^2_W\left(\mathbf{5°}, \overline{\underline{\mathbf{f}}}(k)\right)\right) \qquad (6.13.1)$$

come si verifica facilmente nella matrice delle distanze della TAV. 6.7 calcolata nel secondo passo di aggregazione, ad esempio per $\overline{\underline{\mathbf{f}}}(k) = \mathbf{1°}$ perché $0.005\,926 \geq min\,(0.005\,937, 0.000\,923)$. In altri termini, ogni distanza ricalcolata è sempre più grande della distanza più piccola che rimpiazza. Questa proprietà è soddidfatta dall'indice di dissimilarità (6.9.1) di Ward.

Si può mostrare, Roux (1985), che se due gruppi sono vicini reciproci, essi costituiscono necessariamente un nodo della gerarchia di partizioni. Questo risultato permette di aggregare a ogni passo dell'algoritmo tutte le coppie di gruppi vicini reciproci invece della sola coppia a 'distanza' minima. Si riduce così il numero di passi del metodo di aggregazione gerarchica e, sopratutto, il numero di 'distanze' da ricalcolare.

L'algoritmo dei vicini reciproci si basa sul fatto che le aggregazioni di coppie di gruppi a distanza inferiore a quella di due vicini reciproci, non

modificano la proprietà di questi due gruppi di essere vicini reciproci l'uno dell'altro. Così erano e così restano fino al loro turno di aggregazione.

Siano $\overline{\underline{\mathbf{f}}}(h)$ e $\overline{\underline{\mathbf{f}}}(h')$ i baricentri di una coppia di gruppi vicini reciproci e $\overline{\underline{\mathbf{f}}}(k)$ e $\overline{\underline{\mathbf{f}}}(k')$ quelli della coppia a 'distanza' minima, e quindi la prima da aggregare. Non è possibile, per esempio, aggregare i gruppi $\overline{\underline{\mathbf{f}}}(h)$ e $\overline{\underline{\mathbf{f}}}(k)$ perché $d^2_W\left(\overline{\underline{\mathbf{f}}}(h)\,,\,\overline{\underline{\mathbf{f}}}(h')\right) < d^2_W\left(\overline{\underline{\mathbf{f}}}(h)\,,\,\overline{\underline{\mathbf{f}}}(k)\right)$ e quindi $d^2_W\left(\overline{\underline{\mathbf{f}}}(h),\,\overline{\underline{\mathbf{f}}}(k)\right)$ non è la distanza minima. Se vale in generale la condizione (6.13.1), dopo l'aggregazione dei due gruppi $\overline{\underline{\mathbf{f}}}(k)$ e $\overline{\underline{\mathbf{f}}}(k')$ in $\overline{\underline{\mathbf{f}}}$, si ha

$$d^2_W\left(\overline{\underline{\mathbf{f}}}\,,\,\overline{\underline{\mathbf{f}}}(h)\right) \geq min\left(d^2_W\left(\overline{\underline{\mathbf{f}}}(k),\overline{\underline{\mathbf{f}}}(h)\right)\,,\,d^2_W\left(\overline{\underline{\mathbf{f}}}(k'),\overline{\underline{\mathbf{f}}}(h)\right)\right)$$

ma, siccome i gruppi $\overline{\underline{\mathbf{f}}}(h)$ e $\overline{\underline{\mathbf{f}}}(h')$ sono vicini reciproci

$$d^2_W\left(\overline{\underline{\mathbf{f}}}(h)\,,\,\overline{\underline{\mathbf{f}}}(h')\right) < d^2_W\left(\overline{\underline{\mathbf{f}}}(h)\,,\,\overline{\underline{\mathbf{f}}}(k)\right) \quad \text{e} \quad d^2_W\left(\overline{\underline{\mathbf{f}}}(h)\,,\,\overline{\underline{\mathbf{f}}}(h')\right) < d^2_W\left(\overline{\underline{\mathbf{f}}}(h)\,,\,\overline{\underline{\mathbf{f}}}(k')\right)$$

e sostituendo nell'espressione precedente, si ottiene

$$d^2_W\left(\overline{\underline{\mathbf{f}}}\,,\,\overline{\underline{\mathbf{f}}}(h)\right) > d^2_W\left(\overline{\underline{\mathbf{f}}}\,,\,\overline{\underline{\mathbf{f}}}(h')\right).$$

Allo stesso modo si dimostra che

$$d^2_W\left(\overline{\underline{\mathbf{f}}}(h')\,,\,\overline{\underline{\mathbf{f}}}\right) > d^2_W\left(\overline{\underline{\mathbf{f}}}(h)\,,\,\overline{\underline{\mathbf{f}}}(h')\right).$$

Così, mano a mano che l'algoritmo tradizionale aggrega le coppie di gruppi, la distanza tra coppie aggregate aumenta via via, fino a che la distanza $d^2_W\left(\overline{\underline{\mathbf{f}}}(h)\,,\,\overline{\underline{\mathbf{f}}}(h')\right)$ di una coppia di vicini reciproci arriva a essere la minima delle distanze, rendendo la copia aggregabile.

L'algoritmo, quando è applicato alla matrice delle coordinate fattoriali, si sviluppa in 5 passi:

1) si calcolano le distanze di tutte le coppie presenti;
2) si memorizzano le coppie di gruppi vicini reciproci;
3) tra le coppie di gruppi vicini reciproci si aggrega quella a distanza minima e si calcola il baricenrtro del nuovo gruppo;
4) se restano altre coppie di gruppi vicini reciproci si ritorna al passo 3) a meno che non restino 2 gruppi;
5) si ritorna al passo 1).

La gerarchia di partizioni ottenute con il metodo dei vicini reciproci è esattamente eguale a quella che si otterrebbe con l'algoritmo tradizionale illustrato nella Sez. 6.11, ma con un risparmio di tempo che può superare il 30% .

6.14 - Strategia mista.

Quando il numero di profili da aggregare è molto numeroso, il metodo di aggregazione gerarchica non è direttamente applicabile perché la memoria del computer non riesce a contenere la matrice delle coordinate dei profili. La difficoltà può essere superata aggregando gerarchicamente non i profili, ma i gruppi stabili.

La Strategia Mista consiste appunto nel far seguire alla strategia dei Gruppi Stabili un'aggregazione gerarchica e poi un 'consolidamento' finale della partizione col metodo dei Centri Mobili. Negli anni passati questa era la strategia obbligata quando i profili da aggregare superavano il migliaio. Oggi, con computer dotati di ampie memorie centrali questo vincolo è in gran parte caduto, tuttavia la strategia è sempre raccomandabile qualunque sia il numero di profili da aggregare perché, sfruttando i punti di forza di entrambi i metodi di aggregazione, produce gruppi che si avvicinano molto alla composizione ottimale e che si rivelano più facilmente interpretabili.

La Strategia Mista prevede tre fasi. La prima è costituita dalla strategia dei Gruppi Stabili, già esaminata nella Sez. 6.7, e consistente nell'applicazione ripetuta del metodo dei Centri Mobili alla matrice $\mathbf{F}^*$ dei profili, variando ogni volta la scelta dei centri iniziali. Il numero di centri da imporre in questa prima fase dipende dal numero di profili da aggragare. Tanto per fissare le idee, con $2\,000$ profili è bene ottenere qualche decina di gruppi stabili. Con $10\,000$, si può arrivare anche a un centinaio. Ciò si ottiene per tentativi variando il numero H di centri. Per esempio, nel primo caso si può iniziare con $H = 10$ se le partizioni sono due. Se dei 100 gruppi stabili potenziali la percentuale di gruppi vuoti è alta si può provare con $H = 15$ e così via. Comunque, non esistono regole precise: si va per tentativi perché tutto dipende dalla struttura della nuvola dei profili. Al termine di questa prima fase si ottiene la matrice delle coordinate fattoriali dei gruppi stabili con lo stesso numero di colonne, ma con molte meno righe della matrice $\mathbf{F}^*$ dei profili.

Nella seconda fase, alla matrice delle coordinate dei gruppi stabili, conservata nella memoria centrale, si applica il metodo di aggregazione gerarchico della Sez. 6.11. La matrice è letta sequenzialmente più volte per calcolare la distanza $\nu(l)$ della (6.12.1) di ogni coppia di gruppi.

Soltanto le distanze dei vicini reciproci eventualmente presenti vengono memorizzate, insieme alle coppie di gruppi a cui si riferiscono. La seconda fase si conclude con l'esame del diagramma dei livelli di aggregazione e

la scelta della partizione finale. Questa, difficilmente sarà ottimale in senso assolut a causa del tipo di algoritmo.

Scelta la partizione finale, inizia la terza fase, detta di 'consolidamento', volta a 'migliorare' l'omogeneità dei profili nei gruppi. I baricentri dei gruppi sono considerati come centri di aggregazione e i profili sono riassegnati ai gruppi col metodo a Centri Mobili [1] . Ad ogni passo il tasso d'inerzia tra gruppi diminuisce e, quando il suo valore si stabilizza, si arrestano le iterazioni. Se la partizione è stata scelta bene, il miglioramento resta limitato, ma se un numero consistente di profili cambia di gruppo, è opportuno riconsiderare la scelta della partizione finale.

Per riassumere, una Strategia Mista di aggrevazione prevede i passi:

1 - Analisi delle Corrispondenze della matrice dei profili;
2 - strategia dei Gruppi Stabili;
3 - metodo Gerarchico Ascendente sui gruppi stabili;
4 - scelta di una partizione;
5 - consolidamento della partizione col metodo dei Centri Mobili.

Questa strategia richiede quattro interventi critici da parte dell'analista. Il primo, per fissare il numero $A^* < A$ di coordinate fattoriali da conservare dopo l'Analisi delle Corrispondenze; il secondo per fissare il numero H di centri di aggregazione nel metodo a Centri Mobili; il terzo per stabilire quante partizioni incrociare per ottenere i Gruppi Stabili, e infine, il quarto, per scegliere la partizione finale. Come si vede, la sua individuazione non è mai un processo automatico, ma scaturisce dall'interazione continua tra calcolo e riflessione.

6.15 - Valori-test di modalità e variabili

L'individuazione della partizione finale fissa automaticamente il numero e la composizione dei gruppi. Tuttavia, sapere quanti e quali sono i profili nei gruppi non basta, bisogna anche individuare le principali caratteristiche che differenziano e descrivono i gruppi. Mentre il processo di caratterizzazione è descritto in dettaglio nella prossima Sez. 6.16, in questa viene presentato il Valore-test che è lo strumento per individuare rapidamente le modalità più 'tipiche' di ogni gruppo.

---

[1] Vengono riassegnati soltanto i profili che alla conclusione della fase 1 costituivano da soli un gruppo stabile.

Si procede come in un test d'ipotesi classico, l'ipotesi di base essendo un'estrazione casuale senza reimmissione dei profili del gruppo da quelli complessivamente rilevati. In tale ipotesi e nel caso di una modalità illustrativa si calcola lo scarto tra la percentuale di profili con la modalità nel gruppo e la percentuale di profili, con la medesima modalità, nella nuvola. Nel caso invece di una variabile continua illustrativa, lo scarto è tra due valori medi: nel gruppo e globale. Il Valore-test 'relativizza' questi scarti per tener conto del differente numero di profili nei gruppi rendendo comparabile qualsiasi modalità in qualsiasi gruppo. Va ricordato che il Valore-test e la probabilità critica ad esso associata sono calcolati unicamente per stabilire una graduatoria al fine di caratterizzare i gruppi. Le modalità più caratteristiche di un gruppo sono quelle corrispondenti ai valori più grandi del Valore-test e a quelli più piccoli della probabilità critica.

Valori-test di modalità e frequenze

Nella partizione di I profili in H gruppi, è risultato che

I_j profili possiedono la modalità j di una variabile categorica illustrativa [1],

I_h sono finiti in uno stesso gruppo h,

I_{hj} sono nel gruppo h e possiedono anche la modalità j.

La situazione si può riassumere in una matrice di contingenza 2×2, ove $I_{hj} \leq min\,(I_h, I_j)$, mentre i valori mancanti, indicati con $-$, possono essere ricavati per differenza

	Con j	*Senza j*	*Rilevati*
Nel gruppo h	I_{hj}	$-$	I_h
Fuori gruppo	$-$	$-$	$-$
Rilevati	I_j	$-$	I

Chiedersi se la modalità j è caratteristica del gruppo h equivale a chiedersi se la frazione I_{hj}/I_h di profili nel gruppo è significativamente maggiore di quella globale I_j/I in tutti i rilevati.

Si può procedere come per un test d'ipotesi, in cui l'ipotesi di base H_0 è che le due frazioni siano eguali, mentre l'ipotesi alternativa unilaterale H_1 è che la frazione di elementi che possiedono la modalità j sia invece

[1] Nella Sez. 5.18 questa grandezza era indicata con $\tilde{z}_{+j}$. Qui si è mutato il simbolo per uniformare le notazioni.

molto più grande nel gruppo che tra i rilevati, ovvero che nel gruppo ci sia un significativo affollamento di profili con quella modalità

$$\begin{cases} H_0 : \dfrac{I_{hj}}{I_h} = \dfrac{I_j}{I} \\ H_1 : \dfrac{I_{hj}}{I_h} > \dfrac{I_j}{I}. \end{cases}$$

Affermare che l'ipotesi H_0 è vera è come dire che gli I_h profili del gruppo h sono stati estratti casualmente, e senza reimmissione, dagli I profili rilevati. Se con X_{hj} si indica la variabile casuale discreta 'numero di profili con modalità j nel gruppo h', essa si è realizzata nel gruppo in esame assumendo il valore $X_{hj} = x = I_{hj}$. Se è vera l'ipotesi di base H_0, i totali marginali della matrice sono fissi, e quindi X_{hj} è una variabile casuale *ipergeometrica* il cui valore atteso e varianza sono rispettivamente [1]

$$E\left(X_{hj} \mid H_0\right) = I_h \frac{I_j}{I} \qquad \text{e} \qquad VAR\left(X_{hj} \mid H_0\right) = I_h \frac{I_j}{I}\left(1 - \frac{I_j}{I}\right)\frac{I - I_h}{I - 1}.$$

La frazione I_j/I è una stima della probabilità di rilevare un profilo con la modalità e $1 - (I_j/I)$ quella complementare di non rilevarlo. Se il numero I dei profili complessivamente rilevati è nettamente superiore a quello I_h dei profili che costituiscono il gruppo in modo che questi si possano ritenere praticamente indipendenti e all tempo stesso I_h è abbastanza grande, la distribuzione ipergeometrica può essere sostituita con una normale. Se ciò è possibile, la variabile casuale

$$\frac{X_{hj} - E\left(X_{hj} \mid H_0\right)}{\sqrt{VAR(\, X_{hj} \mid H_0)}}$$

ha una densità di probabilità normale standard $\mathcal{N}(0, 1)$. Si definisce così il

$$\text{Valore} - \text{test} \stackrel{\text{def}}{=} \frac{I_{hj} - E(X_{hj} \mid H_0)}{\sqrt{VAR\left(X_{hj} \mid H_0\right)}} = \frac{\frac{I_{hj}}{I_h} - \frac{I_j}{I}}{\frac{I_j}{I}\left(1 - \frac{I_j}{I}\right)\frac{I - I_h}{I - 1}} \tag{6.15.1}$$

che valuta lo scarto tra la frazione di profili nel gruppo e quella globalmente rilevata in termini di deviazione standard per tener conto del differente numero di profili nei gruppi. Il suo valore assoluto misura in un certo senso quanto la modalità caratterizza il gruppo, il suo segno il senso della caratterizzazione. Se positivo, l'abbondanza relativa della modalità nel gruppo, se negativo la carenza.

---

[1] Si veda ad esempio D. Piccolo (1998), *Statistica*, Il Mulino ed. a pg. 439.

Accanto al Valore-test è sempre utile esaminare la probabilità critica del test. Si è visto che in base all'ipotesi H_0 il numero X_{hj} di profili con modalità j nel gruppo h è una variabile casuale ipergeometrica, per cui è possibile calcolare la probabilità che estraendo I_h elementi dagli I se ne trovino $x = 0, 1, \ldots, I_{hj}, \ldots, min(I_h, I_j)$ con la modatità j

$$Prob\,(X_{hj} = x \mid H_0) = \binom{I_h}{x}\binom{I - I_h}{I_j - x}\Big/\binom{I}{I_j}. \qquad (6.15.2)$$

La probabilità non cambia se nell'espressione (6.15.2) si sostituisce I_j a I_h e I_h a I_j. Quello che interessa, però, è la probabilità che $x \geq I_{hj}$, ossia la *probabilità critica* del test o *p-value* che vale

$$Prob\,(X_{hj} = x \geq I_{hj} \mid H_0) = \sum_{x=I_{hj}}^{min(I_h,I_j)} Prob\,(X_{hj} = x \mid H_0). \qquad (6.15.3)$$

Più I_{hj}, numero di profili osservati nel gruppo h ed aventi la modalità j, è grande, più questa probabilità è piccola e più appare infondata l'ipotesi di base H_0 che le due frazioni di profili nel rilevamento e nel gruppo siano eguali.

Nella pratica si riduce il tempo di calcolo approssimando la probabilità critica (6.15.3) con una equivalente [1] ottenuta da una distribuzione normale standard $\mathcal{N}(0,1)$ che è più facile da calcolare. In tal caso l'espressione della probabilità critica diviene

$$Prob\Big(\mid U \mid \geq \text{ Valore} - \text{test} \mid H_0\Big)$$

dove U è una variabile casuale normale standard e $\mid U \mid$ il suo valore assoluto.

Il procedimento può essere ripetuto per tutte le modalità delle variabili categoriche illustrative d'interesse, per cui alla fine le modalità possono venire *ordinate* in base ai valori delle loro probabilità critiche: le più caratteristiche del gruppo sono quelle con i valori più piccoli.

Valori-test di variabili continue illustrative

Nell'Analisi delle Corrispondenze Multiple accanto alle variabili attive, necessariamente di tipo categorico, si possono avere delle variabili illustrative di tipo continuo. Viene naturale chiedersi in quale misura il valore medio $\overline{x}(h)$ della variabile nel gruppo di I_h valori differisca dal valore medio

[1] W. Molemar (1973), *Simple Approximations to the Poisson, Binomial and Hypergeometric Distributions.*, Biometrics, vol. 29, pag 403-407.

$\overline{x}$ calcolato su tutti gli I valori rilevati. Sia poi s^2 la varianza calcolata sugli I valori rilevati.

Per rispondere alla domanda si confronta il valore medio $\overline{x}(h)$ osservato nel gruppo con il valore medio che ci si aspetterebbe nel caso in cui gli I_h valori fossero entrati casualmente nel gruppo. Il confronto viene fatto costruendo un Valore - test del quale è facilmente calcolabile la densità di probabilità. Si immagina dunque di estrarre un valore a caso tra gli I valori rilevati della variabile continua. Questa estrazione genera *una* variabile casuale $X_{i'}$ che ha valore atteso e varianza eguali a quelli della popolazione di I valori, ossia

$$E\left(X_{i'}\right) = \overline{x} \qquad \text{e} \qquad VAR\left(X_{i'}\right) = s^2.$$

Ripetendo l'estrazione I_h volte, e nell'ipotesi H_0 che essa avvenga *senza* sostituzione, si può costruire la variabile casuale media $\overline{X}(h) = \sum_{i'=1}^{I_h} X_{i'}/I_h$, il cui valore atteso e la varianza sono, secondo la teoria delle variabili casuali,

$$E\left(\overline{X}(h) \mid H_0\right) = \overline{x} \qquad \text{e} \qquad VAR\left(\overline{X}(h) \mid H_0\right) = s^2(h) = \frac{s^2}{I_h}\,\frac{I - I_h}{I - 1}.$$

dove la costante $(I - I_h)/(I - 1)$ è la cosidetta correzione per popolazione finita.

Si dispone ora di tutti gli elementi per costruire il

$$\text{Valore} - \text{test} \stackrel{\text{def}}{=} \frac{\overline{x}(h) - E\left(\overline{X}(h) \mid H_0\right)}{VAR\left(\overline{X}(h) \mid H_0\right)} = \frac{\left(\overline{x}(h) - \overline{x}\right)\sqrt{I_h}}{s}\sqrt{\frac{I-1}{I-I_h}} \quad (6.15.4)$$

in cui lo scarto da una situazione di estrazione casuale è misurato in unità della varianza che si avrebbe in tale situazione. Il valore assoluto del Valore-test indica quanto la variabile continua è caratteristica del gruppo. più il Valore-test è grande, più la media $\overline{x}(h)$ nel gruppo differisce dalla media globale $\overline{x}$. Il segno indica il senso della caratterizzazione: se il Valore-test è positivo (rispettivamente negativo), il gruppo è caratterizzato da valori della variabile alquanto superiori (risp. inferiori) alla media generale.

Anche al Valore-test (6.15.4) è possibile associare una probabilità critica, o *p-value*. Infatti, se il gruppo comprende un numero di valori I_h abbastanza grande da rendere il valor medio $\overline{x}(h)$ stabile e se I_h è anche molto più piccolo di quello I di valori rilevati, in modo che si possano ritenere le I_h estrazioni senza sostituzioni non molto difformi da estrazioni con sostituzione, nel qual caso le I_h variabili casuali $X_{i'}$ si possono ritenere indipendenti, per

il teorema centrale limite la variabile casuale

$$U = \frac{\overline{X}(h) - E\left(\overline{X}(h)\right)}{VAR\left(\overline{X}(h)\right)}$$

ha una densità di probabilità che può essere approssimata da una Normale standard $\mathcal{N}(0,1)$ con valore atteso nullo e varianza unitaria. Il valore critico è allora la probabilità

$$Prob\Big(\mid U \mid \geq \text{ Valore} - \text{test}\Big).$$

Il metodo ha il vantaggio di essere piuttosto veloce anche quando i gruppi da caratterizzare sono numerosi e il numero di variabili è grande.

Valore-test di variabili attive

Il Valore-test può essere calcolato anche per le variabili attive che hanno contribuito a costruire i gruppi della partizione. Si può così predisporre per ogni gruppo una graduatoria unica dei Valori-test delle modalità attive ed illustrative e delle eventuali variabili continue illustrative. Si riconoscono così le variabili più tipiche di ogni gruppo selezionando quelle con minore probabilità critica o, in modo del tutto equivalente con il Valore-test più grande. L'informazione fornita dal valore critico è complementare a quella fornita dal Valore - test. Così, un valore critico pari a $1.96 \simeq 2$ indica che lo scarto ha 5 probabilità su 100 di essere 'fortuito', ossia di essere ottenuto o superato nel caso che i valori nel gruppo fossero estratti a caso. Il procedimento si rivela molto utile per caratterizzare i gruppi, come si vedrà nella Sezione che segue.

6.16 - Interpretazione dei risultati

Nel caso, sempre raccomandabile, che si sia utilizzata la Strategia Mista, l'output prodotto consta di tre parti. La prima contiene i risultati delle singole partizioni ottenute con l'aggregazione a Centri Mobili e quindi le caratteristiche dei Gruppi Stabili che si sono formati. Questi risultati intermedi sono già stati illustrati nella Sez. 6.x e nelle TAV. 6.n. per cui non verranno riesaminati qui. Si raccomanda vivamente di esaminare sempre attentamente questi risultati, seppure siano intermedi. La seconda parte mostra l'albero gerarchico delle aggregazioni e il diagramma dei livelli dell'indice di dissimilarità e la terza ed ultima la composizione dei gruppi della partizione scelta, e l'elenco delle variabili attive ed illustrative che hanno maggiormente contribuito a produrli.

Una volta individuati i gruppi di profili, bisogna renderne intelligibile il significato caratterizzandoli in base alle modalità, alle frequenze o variabili continue illustrative. Come nella Sez. 4.x per caratterizzare gli assi fattoriali, si utilizza il Valore-test.

I gruppi sono stati individuati tenendo conto di tutte le coordinate e forniscono quindi più informazioni sulla configurazione geometrica della nuvola, informazioni che in parte potrebbero sfuggire esaminandone le proiezioni in sottospazi bi- o tri-dimensionali.

6.17 - Bibliografia essenziale

Per approfondire i concetti e le metodiche dell'Analisi dei gruppi, il lettore può consultare

Silvio Griguolo e Pier Carlo Palermo (1984). *Nuovi problemi e nuovi metodi di analisi territoriale*. Franco Angeli ed., Milano., 257 pg., che nell'ultima parte presenta un chiaro compendio dell'Analisi dei gruppi dal punto di vista matematico e statistico. Contiene una grande varietà di applicazione dell'Analisi delle Corrispondenze e dei Gruppi a problemi di condizione abitativa, dei mercati urbani, delle tendenze costruttive e dell'analisi territoriale in genere.

Una descrizione dei principali metodi dell'Analisi dei Gruppi, degli algoritmi, dei programmi di calcolo insieme a molti esempi, si trovano nel testo, facilmente leggibile anche dai non specialisti, di

Maurice Roux (1985). *Algorithmes de Classification*. Masson ed., Paris., 153 pg., ISBN 2-225-80683-7.

Il metodo di aggregazione dei vicini reciproci è descritto in dettaglio nei due articoli seguenti. Il primo è più di carattere metodologico, il secondo più orientato agli aspetti della programmazione

C. de Rham (1980). La classification hyérarchique ascendante selon la methode des voisins rèciproques. *Les Cahiers de l'Analyse des Données*, Vol. **V** $n^o 2$, pag. 135 - 144.

J. Juan (1982). Le programme HIVOR de classification ascendante hyérarchique selon les voisins réciproques et le critàre de la variance. *Les Cahiers de l'Analyse des Données*, Vol. **VII** $n^o 2$, pag. 173 - 184. Quest'ultimo articolo contiene le istruzione Fortran per programmare l'algoritmo.

PARTE PRIMA: IL METODO

CAPITOLO 7: Stabilità della configurazione

Sommario

La struttura che si è cercato di interpretare sussiste realmente o è il frutto di circostanze casuali? Come fare per capirlo?

CAPITOLO 7

7.1 -Introduzione

In questo capitolo si cercherà di dare una risposta a questi due unici quesiti: 'Quanti assi fattoriali conservare ?' 'Quanti gruppi ci sono ?' Gli assi da conservare sono quelli che generano un sottospazio in cui la proiezione della nuvola risulta stabile.

Le interviste sono state condotte col metodo della Quota, che consiste nell'assegnare agli intervistatori il numero di interviste da fare, lasciando però a loro la scelta del cliente da intervistare a condizione che rispettino le quote di intervistati che presentano le caratteristiche loro comunicate. Il problema della rappresentatività del campione intervistato non interessa l'Analisi delle Corrispondenze che non ha messun fine inferenziale, ma può diventare importante se si deve valutare la stabilità della configurazione dei profili sulle mappe, problema che verrà affrontato nel Capitolo 7.

vedere articolo in SUGITALIA 1997

Wasserman FAUST pag 56-59.

V. CHristensen SPRINGER, da pag. 2 a pag. 18

Ricampionare n volte le righe (o le colonne) significa dare peso 1/n alle matrici.

Estrarre da marginali fissi ....

Si è parlato di un metodo statistico, ma per quanto visto finora Si tratta di un metodo matematico e quindi puramente deterministico. Il lettore si chiederà: deve sta la Statistica? La Statistica sta nei dati d'ingresso e nelle considerazioni che si possono fare sui risultati, considerazioni in questo capitolo. Dati casuali si presentano molto di rado. Risalire al campione in base a considerazioni extra-statistiche.

La distribuzione dei conteggi nelle celle, dato il totale generale n_{++} fisso, è la multinomiale che però dipende dai totali marginali che sono sconosciuti. (Whit. pag. 287 in alto.). Si presentano due possibilità 1) tenere fissi i totali marginali, nel qual caso la distribuzione è nota: è la distribuzione

ipergeometrica. In questo caso si è interessati all'omogeneità; 2) si ipotizza che la distribuzione osservata è una stima consistente e non distorta della multinomiale sconosciuta: è il metodo di bootstrap.

The argument that the pattern of interactions in the observed table should be assessed against all other possible tables with the same margins, has some force. (Whitt. pag. 287)

Quanto verrà detto vale esclusivamente per le matrici di contingenza. Inoltre, secondo la prassi, le variabili casuali verranno indicate con una lettera latina maiuscola e le loro realizzazioni con una lettera latina minuscola.

La matrice dei dati è

$$\mathbf{N} = \begin{pmatrix} 0 & 2 & 2 \\ 3 & 4 & 7 \\ 3 & 6 & 9 \end{pmatrix} \qquad (7.x.1)$$

La struttura della matrice osservata va confrontata con la struttura di tutte le matrici di contingenza con caratteristiche simili, costituenti l'*insieme di riferimento* al quale appartiene anche la matrice osservata $\mathbf{N}$. Per poter disporre di queste matrici è necessario fare delle ipotesi sul meccanismo che ha generato i valori osservati nella matrice $\mathbf{N}$. Si tratta quindi di ipotizzare dei modelli particolari di generazione considerando i dati osservati, ossia gli elementi di $\mathbf{N}$ come realizzazioni di $I \times J$ variabili casuali (v. c.). Ipotizzando per esse delle specifiche distribuzioni di probabilità, è possibile calcolare le probabilità di osservare ogni matrice dell'insieme di riferimento.

7.10 - Sensibilità a perturbazioni dei dati

J. Benasseni(1991)

Quando i dati rilevati vengono organizzati in una matrice di contingenza, occorre preventivamente stabilire quante e quali saranno le modalità delle due variabili che si intende incrociare. Dal momento che l'Analisi delle Corrispondenze è in grado di analizzare agevolmente matrici di notevoli dimensioni, si sarebbe portati a conservare tutte le modalità rilevate, col risultato di avere nella matrice di contingenza molti elementi con valori esigui o addirittura nulli, e di conseguenza profili con scarsa sgnificatività statistica. D'altro canto un eccessivo accorpamento delle modalità potrebbe occultare aspetti significativi dei legami che intercorrono tra le modalità. C'è dunque una certa dose di arbitrarietà nella costruzione della matrice di contingenza ed è quindi importante esaminare come questa influisca sui risultati finali

dell'analisi, intendendo per questi gli autovalori e gli autovettori. Il problema risale a Escofier e Leroux (1976).

Nelle Sezioni che seguono viene affrontato analicamente il problema di stabilire in qual misura gli autovalori ed autovettori ottenuti dall'Analisi delle Corrispondenze sono sensibili a modifiche come l'aggregazione di righe o colonne e lo spostamento di conteggi da una cella ad un'altra della matrice $\mathbf{N}$. L'importanza pratica di questi metodi risiede nel fatto che permettono di individuare dei limiti di variazione, evitando così il rifacimento dell'analisi. Molti aspetti del problema restano comunque ancora da esplorare.

7.10 - Aggregazione di coppie di colonne o di righe

In questa Sezione verrà mostrato in qual modo l'aggregazione di due *colonne* nella matrice $\mathbf{N}$ di contingenza, si rifletta sugli *autovalori* e sugli *autovettori* e, tramite questi, sui *fattori* e le *mappe fattoriali*. Risultati analoghi, relativi all'aggregazione di due *righe*, sono riportati alla fine della Sezione. Nella Sez. 3.14 si è visto che nell'Analisi delle Corrispondenze, autovalori ed autovettori si ottengono risolvendo l'equazione (3.14.7)

$$\mathbf{C}\,\mathbf{R}^T\mathbf{u}^\star = \lambda\,\mathbf{u}^\star \qquad (7.x.1)$$

con la condizione che il vettore $\mathbf{u}^\star$, che ha origine nel baricentro della nuvola dei J profili colonna, sia di lunghezza $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitaria. Grazie alla (3.2.4) ed ai risultati della Sez. 3.9 e della Sez. 3.14, la matrice quadrata $\mathbf{C}\,\mathbf{R}^T$ di ordine $I \times I$, è esprimibile come

$$\mathbf{Q} = \mathbf{C}\,\mathbf{R}^T = \mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} = \left[\sum_{j=1}^{J} \overline{r}_j\,\mathbf{c}_j\,\mathbf{c}_j^T\right]\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} = \left[\sum_{a=1}^{A} \lambda_a\,\mathbf{u}_a^\star\,\mathbf{u}_a^{\star T}\right]\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \qquad (7.x.2)$$

perché è $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-simmetrica (APP. B) e perciò ha A autovalori reali λ_a, ai quali corrispondono A autovettori $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitari $\mathbf{u}_a^\star$ di ordine I, che risultano $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonali due a due quando corrispondono ad autovalori distinti. Il loro numero è $A = min\,(I, J) - 1$.

Quando due delle J colonne di $\mathbf{N}$, ad esempio la colonna k e la colonna l, vengono aggregate in una sola, a questa corrisponde il profilo colonna di ordine I,

$$\mathbf{c}_{k+l} = \left(\frac{n_{1k}+n_{1l}}{n_{+k}+n_{+l}}\;\frac{n_{2k}+n_{2l}}{n_{+k}+n_{+l}}\ldots\frac{n_{Ik}+n_{Il}}{n_{+k}+n_{+l}}\right)^T$$

$$= \left(\frac{1}{n_{+k}+n_{+l}}\right) \left((n_{1k}\ n_{2k}\ \ldots, n_{Ik})^T\ + (n_{1l}\ n_{2l}\ \ldots, n_{Il})^T\right)$$
$$= \frac{n_{+k}}{n_{+k}+n_{+l}}\,\mathbf{c}_k + \frac{n_{+l}}{n_{+k}+n_{+l}}\,\mathbf{c}_l = \frac{\overline{r}_k}{\overline{r}_k+\overline{r}_l}\,\mathbf{c}_k + \frac{\overline{r}_l}{\overline{r}_k+\overline{r}_l}\,\mathbf{c}_l. \quad (7.x.3)$$

Dopo l'aggregazione, la matrice da diagonalizzare resta di ordine $I \times I$,

$$\widetilde{\mathbf{Q}} = \left[\sum_{\substack{j=1\\ j\neq k,l}}^{J} \overline{r}_j\,\mathbf{c}_j\,\mathbf{c}_j^T + (\overline{r}_k+\overline{r}_l)\,\mathbf{c}_{k+l}\,\mathbf{c}_{k+l}^T\right] \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \quad (7.x.4)$$

dove $\overline{r}_k+\overline{r}_l$ è la *massa* del profilo $\mathbf{c}_{k+l}$ della colonna aggregata. La matrice al secondo termine entro parentesi, grazie alla (7.x.3), può scriversi

$$\begin{aligned}(\overline{r}_k+\overline{r}_l)\,\mathbf{c}_{k+l}\,\mathbf{c}_{k+l}^T &= \frac{1}{\overline{r}_k+\overline{r}_l}\left(\overline{r}_k^2\,\mathbf{c}_k\,\mathbf{c}_k^T + \overline{r}_k\,\overline{r}_l(\mathbf{c}_k\,\mathbf{c}_l^T + \mathbf{c}_l\,\mathbf{c}_k^T) + \overline{r}_l^2\,\mathbf{c}_l\,\mathbf{c}_l^T\right)\\ &= \overline{r}_k\,\mathbf{c}_k\,\mathbf{c}_k^T - \frac{\overline{r}_k\,\overline{r}_l}{\overline{r}_k+\overline{r}_l}\,\mathbf{c}_k\,\mathbf{c}_k^T + \frac{\overline{r}_k\,\overline{r}_l}{\overline{r}_k+\overline{r}_l}\,(\mathbf{c}_k\,\mathbf{c}_l^T + \mathbf{c}_l\,\mathbf{c}_k^T)\\ &+ \overline{r}_l\,\mathbf{c}_l\,\mathbf{c}_l^T - \frac{\overline{r}_k\,\overline{r}_l}{\overline{r}_k+\overline{r}_l}\,\mathbf{c}_l\,\mathbf{c}_l^T\\ &= \overline{r}_k\,\mathbf{c}_k\,\mathbf{c}_k^T + \overline{r}_l\,\mathbf{c}_l\,\mathbf{c}_l^T - \frac{\overline{r}_k\,\overline{r}_l}{\overline{r}_k+\overline{r}_l}\,(\mathbf{c}_k-\mathbf{c}_l)\,(\mathbf{c}_k-\mathbf{c}_l)^T.\end{aligned}$$

Perciò la matrice perturbata $\widetilde{\mathbf{Q}}$ della (7.x.4), grazie alla (7.x.2), può essere espressa come

$$\widetilde{\mathbf{Q}} = \mathbf{Q} - \frac{\overline{r}_k\,\overline{r}_l}{\overline{r}_k+\overline{r}_l}\,(\mathbf{c}_k-\mathbf{c}_l)\,(\mathbf{c}_k-\mathbf{c}_l)^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} = \mathbf{Q} - \epsilon_{kl}\,\mathbf{T} \quad (7.x.5)$$

dove

$$\epsilon_{kl} = \frac{\overline{r}_k\,\overline{r}_l}{\overline{r}_k+\overline{r}_l}$$

è detta *massa ridotta* del profilo $\mathbf{c}_{k+l}$ della colonna aggregata. Risulta quindi che la matrice perturbata può ottenersi dalla matrice originale sottraendole una perturbazione, che dipende da quali colonne si aggregano, espressa da una matrice $\mathbf{T} = (\mathbf{c}_k-\mathbf{c}_l)\,(\mathbf{c}_k-\mathbf{c}_l)^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$, anch'essa $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-simmetrica. La matrice perturbata $\widetilde{\mathbf{Q}}$ soddisfa la proprietà equidistributiva della Sez. 2.9, in quanto $\widetilde{\mathbf{Q}} = \mathbf{Q}$ quando i due due profili che si aggregano sono eguali.

La condizione (7.x.1) per la matrice perturbata $\widetilde{\mathbf{Q}}$ diventa

$$\widetilde{\mathbf{Q}}\,\tilde{\mathbf{u}}^\star = \tilde{\lambda}\,\tilde{\mathbf{u}}^\star \quad \text{con} \quad \tilde{\mathbf{u}}^{\star T}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\tilde{\mathbf{u}}^\star = 1 \quad (7.x.6)$$

che ammette come soluzione gli $\tilde{A} = min\,(I, J-1) - 1$ autovettori $\tilde{\mathbf{u}}_a^\star$ di ordine I, $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitari e $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonali due a due, in corrispondenza degli $\tilde{A}$ autovalori $\tilde{\lambda}_a$ reali e non negativi, che è verosimile ritenere tutti distinti.

La determinazione degli autovalori $\tilde{\lambda}_a$, e degli autovettori $\tilde{\mathbf{u}}_a^\star$, si può ottenere dall'Analisi delle Corrispondenze della matrice di contingenza di ordine $I \times (J-1)$ con le due colonne aggregate, ma spesso è possibile evitare il rifacimento dell'analisi perché R. Sibson (1979) ha mostrato come ottenere una *stima* approssimata di $\tilde{\lambda}_a$ e di $\tilde{\mathbf{u}}_a^\star$. Se ci si limita ad una approssimazione del *primo ordine*, di solito sufficiente nelle applicazioni pratiche, il Lemma 2.1 dimostrato da Sibson [1] afferma che se λ_a è uno degli A autovalori, supposti distinti, di $\mathbf{Q}$ e $\mathbf{u}_a^\star$ il suo corrispondente autovettore di lunghezza $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-unitaria, allora, se la matrice $\mathbf{Q}$ è perturbata sì da potersi esprimere così

$$\widetilde{\mathbf{Q}} = \mathbf{Q} + \epsilon_{kl}\,\mathbf{T} + \text{ altri termini dell'ordine di } \epsilon_{kl}^2$$

allora i suoi $\tilde{A}$ autovalori $\tilde{\lambda}_a$, supposti distinti, ed i corrispondenti autovettori $\tilde{\mathbf{u}}_a^\star$, a meno di termini dell'ordine di ϵ_{kl}^2, sono, in prima approssimazione,

$$\tilde{\lambda}_a \approx \lambda_a - \frac{\overline{r}_k\,\overline{r}_l}{\overline{r}_k + \overline{r}_l}\,\mathbf{u}_a^{\star T}\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,(\mathbf{c}_k - \mathbf{c}_l)\,(\mathbf{c}_k - \mathbf{c}_l)^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{u}_a^\star$$

$$\tilde{\mathbf{u}}_a^\star \approx \mathbf{u}_a^\star + \frac{\overline{r}_k\,\overline{r}_l}{\overline{r}_k + \overline{r}_l}\left[\sum_{\substack{b=1\\ b\neq a}}^{\tilde{A}} \frac{1}{(\lambda_b - \lambda_a)}\,\mathbf{u}_b^\star \mathbf{u}_b^{\star T}\right]\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,(\mathbf{c}_k - \mathbf{c}_l)\,(\mathbf{c}_k - \mathbf{c}_l)^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\mathbf{u}_a^\star.$$

Tutte le grandezze a secondo membro sono note, in quanto ottenute dall'Analisi delle Corrispondenze della matrice originale $\mathbf{N}$ di ordine $I \times J$. L'importanza di queste espressioni sta nel fatto che permettono di capire *come* l'aggregazione di due colonne influisca sui risultati dell'analisi, per esempio come comporti sempre una *diminuzione* degli autovalori, ossia una perdita d'inerzia. Le approssimazioni degli autovettori possono essere utilizzate per ottenere dalla (4.1.4) delle stime approssimate dei fattori delle colonne dopo l'aggregazione

$$\tilde{\mathbf{g}}_a = \widetilde{\mathbf{C}}^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\tilde{\mathbf{u}}_a^\star$$

dove $\widetilde{\mathbf{C}}$ indica la matrice $I \times (J-1)$ dei profili colonna dopo l'aggregazione e quindi delle mappe fattoriali. Le approssimazioni degli autovalori forniscono

---

[1] Il Lemma 2.1 è stato dimostrato da Sibson per matrici simmetriche, ma esso può essere esteso a matrici $\mathbf{D}$-simmetriche (APP. B) come la (7.x.5). La dimostrazione sfrutta la scomposizione (B.3.2) di tali matrici.

il valore, approssimato, dell'inerzia delle proiezioni dei profili su ciascun asse fattoriale.

Per concludere, *dopo* l'aggregazione di due colonne: 1) la matrice di contingenza diventa di ordine $I \times (J-1)$ e la nuvola di punti in $\Re^I$ si riduce a $J-1$ profili colonna; 2) la matrice $\widetilde{\mathbf{Q}}$ resta di ordine $I \times I$ e gli autovettori restano di ordine I; 3) il numero di autovalori distinti si riduce ad $\tilde{A} = A-1$ e tale diviene l'ordine dei fattori $\tilde{\mathbf{g}}_a$.

Quando, invece, nella matrice $\mathbf{N}$ di contingenza si aggregano due *righe*, per esempio le righe k^{ma} ed l^{ma}, dall'Analisi delle Corrispondenze della matrice di ordine $(I-1) \times J$, si ottengono gli $\tilde{A} = min\,(I-1, J) - 1$ autovalori $\tilde{\lambda}_a$, ed i corrispondenti $\tilde{A}$ autovettori $\tilde{\mathbf{v}}_a^\star$ di ordine J, $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$-unitari e $\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$-ortogonali due a due quando corrispondono ad autovalori distinti. Le approssimazioni del primo ordine, analogamente a quanto ottenuto per le colonne, risultano

$$\tilde{\lambda}_a \approx \lambda_a - \frac{\overline{c}_k\,\overline{c}_l}{\overline{c}_k + \overline{c}_l}\,\tilde{\mathbf{v}}_a^{\star T}\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\,(\mathbf{r}_k - \mathbf{r}_l)\,(\mathbf{r}_k - \mathbf{r}_l)^T\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\,\tilde{\mathbf{v}}_a^\star$$

$$\tilde{\mathbf{v}}_a^\star \approx \mathbf{v}_a^\star + \frac{\overline{c}_k\,\overline{c}_l}{\overline{c}_k + \overline{c}_l}\left[\sum_{\substack{b=1\\ b\neq a}}^{A} \frac{1}{(\lambda_b - \lambda_a)}\,\mathbf{v}_b^\star\,\mathbf{v}_b^{\star T}\right]\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\,(\mathbf{r}_k - \mathbf{r}_l)\,(\mathbf{r}_k - \mathbf{r}_l)^T\,\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\,\mathbf{v}_a^\star.$$

Queste approssimazioni del primo ordine, per l'accorpamento di due colonne o righe di $\mathbf{N}$, sono state ottenute da Gifi (1990) in un approccio più generale al problema della stabilità e da J. Bénasséni (1991). Questi mostra anche come determinare l'accuratezza delle approssimazioni, fornendo quindi un criterio obiettivo per valutare se il rifacimento dell'analisi può essere evitato. Il medesimo autore fornisce anche gli estremi dell'intervallo di variabilità delle stime $\tilde{\lambda}_a$.

7.11 - Aggregazioni multiple di colonne o righe

Quando l'aggregazione interessa più di due colonne della matrice di contingenza, le approssimazioni della Sez. precedente possono essere estese senza difficoltà, fornendo dalla (7.x.5)

$$\widetilde{\mathbf{Q}} = \mathbf{Q} - \sum_{(k,l)} \frac{\overline{r}_k\,\overline{r}_l}{\overline{r}_k + \overline{r}_l}\,(\mathbf{c}_k - \mathbf{c}_l)\,(\mathbf{c}_k - \mathbf{c}_l)^T\,\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$$

dove $\sum_{(k,l)}$ indica che la somma è fatta su tutte le coppie di colonne (k,l) che sono state aggregate. Limitandosi all'approssimazione del primo ordine,

si ha

$$\tilde{\lambda}_a \approx \lambda_a - \sum_{(k,l)} \epsilon_{kl}\, \mathbf{u}_a^{\star T}\, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\, (\mathbf{c}_k - \mathbf{c}_l)\, (\mathbf{c}_k - \mathbf{c}_l)^T\, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\, \mathbf{u}_a^{\star}$$

$$\tilde{\mathbf{u}}_a^{\star} \approx \mathbf{u}_a^{\star} + \sum_{(k,l)} \epsilon_{kl} \left[\sum_{\substack{b=1 \\ b \neq a}}^{\bar{A}} \frac{1}{(\lambda_b - \lambda_a)}\, \mathbf{u}_b^{\star} \mathbf{u}_b^{\star T} \right] \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\, (\mathbf{c}_k - \mathbf{c}_l)\, (\mathbf{c}_k - \mathbf{c}_l)^T\, \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}} \mathbf{u}_a^{\star}.$$

dove ϵ_{kl} è la massa ridotta delle coppie aggregate. Quando le modalità delle colonne sono ordinate, sia perché hanno un ordine intrinseco, sia perché ottenute suddividendo in classi contigue una variabile continua come nella Sez. 1.2, l'aggregazione deve necessariamente rispettare l'ordine. Così, ad esempio se la variabile è il *Titolo di Studio*, si accorperanno le colonne corrispondenti a Licenza Media e Diploma e non Licenza e Laurea.

Rsultati analoghi valgono per l'accorpamento multiplo di righe. Le espressioni dell'approssimazione lineare sono le stesse ottenute nella Sez. precedente con l'inclusione della somma $\sum_{(k,l)}$ dopo i segni $-$ e $+$. Ancora J. Bénasséni (1991) indica come stabilire l'accuratezza delle approssimazioni per valutare l'opportunità di ripetere l'analisi.

X.2 - Modello di Poisson

L'insieme di riferimento $\Im_{\mathbf{W}}$, è l'insieme costituito da tutte le matrici di *contingenza* $\mathbf{W}$ dello stesso ordine, e che per la matrice $\mathbf{N}$ in (X.1.1) è del tipo

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & w_{1+} \\ w_{21} & w_{22} & w_{2+} \\ w_{+1} & w_{+2} & w_{++} \end{pmatrix}$$

Secondo questo modello i conteggi osservati w_{ij} sono realizzazioni di $I \times J$ variabili casuali W_{ij} *poissoniane indipendenti*. Ciascuna segue quindi una *distribuzione di Poisson* discreta che dipende da un unico parametro, il valore atteso $\overline{w}_{ij}$ della variabile casuale W_{ij}

$$\overline{w}_{ij} = E(W_{ij}) = Var(W_{ij}).$$

In base a questo modello, la probabilità di osservare *proprio* il conteggio w_{ij} è, per $i = 1, 2, \ldots, I$ e $j = 1, 2, \ldots, J$, data da

$$P(W_{ij} = w_{ij}) = \frac{e^{-\overline{w}_{ij}}\, \overline{w}_{ij}^{w_{ij}}}{w_{ij}!} \qquad (X.2.1)$$

Essendo le variabili casuali indipendenti per ipotesi, la probabilità di osservare proprio una *specifica* matrice $\mathbf{W}$, è la probabilità congiunta che la prima variabile casuale W_{11} assuma proprio il valore w_{11}, che la seconda assuma il valore w_{12}, e così via, e questa è il prodotto delle probabilità (X.2.1) per tutti gli $I \times J$ elementi

$$P(W_{11} = w_{11}, \ldots, W_{I,J} = w_{IJ}) = \prod_{i=1}^{I} \prod_{j=1}^{J} P(W_{ij} = w_{ij}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{e^{-\overline{w}_{ij}}\, \overline{w}_{ij}^{w_{ij}}}{w_{ij}!}.$$

Di conseguenza, la probabilità di osservare una *qualunque* matrice $\mathbf{W}$ dell'insieme $\Im_{\mathbf{W}}$ di riferimento risulta

$$\begin{aligned} P(\mathbf{W}) &= \frac{P(W_{11} = w_{11}, \ldots, W_{I,J} = w_{IJ})}{\sum_{\mathbf{W} \in \Im_{\mathbf{W}}} P(W_{11} = w_{11}, \ldots, W_{I,J} = w_{IJ})} \\ &= \frac{\prod_{i=1}^{I} \prod_{j=1}^{J} \frac{e^{-\overline{w}_{ij}}\, \overline{w}_{ij}^{w_{ij}}}{w_{ij}!}}{\sum_{\mathbf{W} \in \Im_{\mathbf{W}}} \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{e^{-\overline{w}_{ij}}\, \overline{w}_{ij}^{w_{ij}}}{w_{ij}!}} \end{aligned}$$

dove, al denominatore la somma è estesa a tutte le matrici dell'insieme di riferimento $\Im_{\mathbf{W}}$, costituito da un numero infinito di matrici $\mathbf{W}$.

In ciascuna matrice il numero si conteggi totali w_{++} varia ed è caratteristica peculiare del modello di Poisson che anche questo elemento sia una realizzazione di una v. c. $W_{++} = \sum_i \sum_j W_{ij}$ di Poisson con valore atteso $\overline{w}_{++} = E(W_{++})$. Perciò

$$P(W_{++} = w_{++}) = \frac{e^{-\overline{w}_{++}}\, \overline{w}_{++}^{w_{++}}}{w_{++}!} \qquad (X.2.2)$$

Essendo sconosciuti i valori attesi $\overline{w}_{++}$, le probabilità $P(\mathbf{W})$ non possono essere calcolate, a meno di fare ulteriori e più restrittive ipotesi sui valori che le v. c. possono assumere.

X.3 - Modello multinomiale

Dal momento che nella matrice dei dati $\mathbf{N}$, il campione osservato ha numerosità n_{++}, sembra ragionevole limitare l'insieme di riferimento a quelle matrici di contingenza con lo *stesso* totale generale della matrice osservata. Per la matrice (X.1.1), l'insieme di riferimento $\Im_{\mathbf{X}}$ conterrà tutte matrici $\mathbf{X}$ dello stesso ordine 2×2 e del tipo

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & x_{1+} \\ x_{21} & x_{22} & x_{2+} \\ x_{+1} & x_{+2} & \mathrm{n}_{++} \end{pmatrix}.$$

Imporre che l'ampiezza del campione sia fissa ed eguale ad n_{++}, ha due conseguenze. La prima è che le variabili casuali son sono più poissoniane, perché ogni conteggio non può superare n_{++}, e la seconda che non sono più indipendenti, dal momento che il valore assunto da una condiziona i possibili valori che possono assumere le altre. Partendo dal modello poissoniano con $\overline{x}_{ij} = E(X_{ij})$, si può ottenere la probabilità di osservare una *specifica* matrice $\mathbf{X}$ dell'insieme di riferimento: sarà la probabilità congiunta che la v. c. X_{11} assuma proprio il valore x_{11}, X_{12} il valore x_{12}, e così via, e *condizionata* dal vincolo che la v. c. $\sum_i \sum_j X_{ij} = X_{++} = n_{++}$,

$$P(X_{11} = x_{11}, \ldots, X_{I,J} = x_{IJ} \,|\, X_{++} = n_{++}) = \frac{P(X_{11} = x_{11}, \ldots, X_{I,J} = x_{IJ})}{P(\sum_i \sum_j X_{ij} = n_{++})}$$

$$= \frac{\prod_i \prod_j \frac{e^{-\overline{x}_{ij}} \, \overline{x}_{ij}^{x_{ij}}}{x_{ij}!}}{e^{-\sum_i \sum_j \overline{x}_{ij}} \, (\sum_i \sum_j \overline{x}_{ij})^{n_{++}} \, \frac{1}{n_{++}!}} = \frac{n_{++}!}{\prod_i \prod_j x_{ij}!} \frac{\prod_i \prod_j \overline{x}_{ij}^{x_{ij}}}{(\sum_i \sum_j \overline{x}_{ij})^{n_{++}}}$$

$$= \frac{n_{++}!}{\prod_i \prod_j x_{ij}!} \frac{\prod_i \prod_j \overline{x}_{ij}^{x_{ij}}}{\prod_i \prod_j (\sum_i \sum_j \overline{x}_{ij}^{x_{ij}})} = \frac{n_{++}!}{\prod_i \prod_j x_{ij}!} \prod_{i=1}^{I} \prod_{j=1}^{J} p_{ij}^{x_{ij}}. \qquad (X.3.1)$$

Questa è la *distribuzione multinomiale* [1], che dipende dall'ampiezza n_{++} del campione, che è nota, e dalle $I \times J$ probabilità, che sono incognite,

$$p_{ij} = \frac{\overline{x}_{ij}}{\sum_i \sum_j \overline{x}_{ij}} \qquad (X.3.2)$$

che una unità delle n_{++} disponibili vada ad incrementare proprio l'elemento di posizione (i, j).

La probabilità di osservare una *qualunque* matrice $\mathbf{X}$ dell'insieme $\Im_{\mathbf{X}}$ di riferimento è allora

$$P(\mathbf{X}) = \frac{\frac{n_{++}!}{\prod_i \prod_j x_{ij}!} \prod_i \prod_j p_{ij}^{x_{ij}}}{\sum_{\mathbf{X} \in \Im_{\mathbf{X}}} \frac{n_{++}!}{\prod_i \prod_j x_{ij}!} \prod_i \prod_j p_{ij}^{x_{ij}}}. \qquad (X.3.3)$$

Poiché le probabilità p_{ij} non sono note, $P(\mathbf{X})$ non può essere calcolata, ma nella Sez. 7.X verrà mostrato come sia possibile approssimarla con una distribuzione empirica.

X.4 - Modello multinomiale stratificato

[1] Quando le matrici $\mathbf{X}$ sono di ordine 2×2, la distribuzione si riduce ad una *binomiale*.

Un vincolo che si può ulteriormente imporre al criterio d'appartenenza di una matrice all'insieme di riferimento, è quello che i totali nella riga, oppure nella colonna marginale, siano eguali a quelli della matrice osservata. Questo vincolo comporta implicitamente che anche i totali generali siano eguali. Così, se per esempio, si sceglie di tener fissi i totali nella riga marginale, allora l'insieme di riferimento $\Im_{\mathbf{Y}}$ per la matrice (X.1.1) è costituito dalle matrici di contingenza $\mathbf{Y}$ di ordine 2×2 così strutturate

$$\mathbf{Y} = \begin{matrix} \begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix} & \begin{matrix} n_{1+} \\ n_{2+} \end{matrix} \\ \begin{matrix} y_{+1} & y_{+2} \end{matrix} & \mathrm{n}_{++} \end{matrix}$$

Secondo questo modello, le probabilità p_{ij} dipendono soltanto dalla riga che si considera e sono quindi le stesse per una stessa riga,

$$\begin{array}{ccccc} p_{11} & = p_{12} & \dots & = p_{1J} & = p_{j|1} \\ \dots & \dots & \dots & \dots & \dots \\ p_{i1} & = p_{i2} & = \dots & = p_{iJ} & = p_{j|i} \\ \dots & \dots & \dots & \dots & \dots \\ p_{I1} & = p_{I2} & = \dots & = p_{IJ} & = p_{j|I}. \end{array}$$

Qui $p_{j|i}$ è la probabilità che una unità delle n_{++} vada ad incrementare il j^{mo} elemento della prefissata i^{ma} riga e

$$p_{j|1} + p_{j|2} + \ldots + p_{j|i} + \ldots + p_{j|I} = 1.$$

I conteggi y_{ij} all'interno di ciascuna riga i della matrice sono quindi realizzazioni di J v. c. multinomiali Y_{ij} con parametri n_{i+} e $p_{j|i}$, per cui, in base all (X.3.1)

$$P(Y_{i1} = y_{i1}, \ldots, Y_{iJ} = y_{iJ}) = \frac{n_{i+}!}{\prod_i \prod_j y_{ij}!} \prod_{i=1}^{I} \prod_{j=1}^{J} p_{j|i}^{y_{ij}}.$$

Se le v. c. relative ad una riga sono indipendenti da quelle delle altre righe, la probabilità di osservare simultaneamente le I righe di una specifica matrice $\mathbf{Y}$ è il prodotto delle distribuzioni multinomiali di probabilità relative alle varie righe

$$P(Y_{i1} = y_{i1}, \ldots, Y_{IJ} = y_{IJ} \,|\, Y_{i+} = n_{i+}) = \prod_{i=1}^{I} \Big(\frac{n_{i+}!}{\prod_j y_{ij}!} \prod_{j=1}^{J} p_{j|i}^{y_{ij}} \Big). \qquad (X.4.1)$$

La probabilità di osservare una qualunque matrice dell'insieme di riferimento

è

$$P(\mathbf{Y}) = \frac{\prod_{i=1}^{I}\left(\frac{n_{i+}!}{\prod_j y_{ij}!}\prod_{j=1}^{J} p_{j|i}^{y_{ij}}\right)}{\sum_{\mathbf{Y}\in\Im_{\mathbf{Y}}}\prod_{i=1}^{I}\left(\frac{n_{i+}!}{\prod_j y_{ij}!}\prod_{j=1}^{J} p_{j|i}^{y_{ij}}\right)}.$$

Se vale il modello multinomiale stratificato, allora i conteggi $y_{+1}, y_{+2}, \ldots, y_{+I}$ nella riga marginale sono anch'essi realizzazioni di J v. c. multinomiali Y_{+j} con parametri n_{++} e $p(j|+)$

$$P(Y_{+1} = y_{+1}, \ldots, Y_{+J} = y_{+J} \,|\, Y_{++} = y_{++}) = \frac{n_{++}!}{\prod_j y_{+j}!}\prod_{j=1}^{J} p_{j|+}^{y_{+j}}. \qquad (X.4.2)$$

Se si fossero tenuti fissi i totali nella colonna marginale, si sarebbero ottenute espressioni analoghe, con lo scambio degli indici i e j.

X.5 - Modello ipergeometrico

Un'ulteriore e definitiva restrizione al criterio di appartenenza all'insieme di riferimento è quello di imporre che *tutti* i totali marginali siano eguali a quelli della matrice osservata. Nel caso della matrice $\mathbf{N}$ l'insieme di riferimento $\Im_{\mathbf{Z}}$ è costituito da tutte le matrici di contingenza $\mathbf{Z}$ di ordine 2×2 del tipo

$$\mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & n_{1+} \\ z_{21} & z_{22} & n_{2+} \\ n_{+1} & n_{+2} & \mathrm{n}_{++} \end{pmatrix}$$

Fissare i valori marginali implica che

$$p_{j|1} = p_{j|2} = \ldots = p_{j|i} = \ldots = p_{j|I} = p_{j|+}. \qquad (j = 1, 2, \ldots, J)$$

La distribuzione di probabilità congiunta, condizionata dai totali marginali fissi, è data da

$$\begin{aligned} &P(Z_{11} = z_{11}, .., Z_{IJ} = z_{IJ} \,|Z_{+1} = n_{+1}, .., Z_{+J} = n_{+J}, Z_{1+} = n_{1+}, .., Z_{I+} = n_{I+}) \\ &= \frac{P(Z_{11} = z_{11}, \ldots, Z_{IJ} = z_{IJ})}{P(Z_{+1} = n_{+1}, \ldots, Z_{+J} = n_{+J}, Z_{1+} = n_{1+}, \ldots, Z_{I+} = n_{I+})} \\ &= \frac{\prod_i\left(\frac{n_{i+}!}{\prod_j z_{ij}!}\prod_j p_{j|+}^{z_{ij}}\right)}{\frac{n_{++}!}{\prod_j n_{+j}!}\prod_j p_{j|+}^{n_{+j}}} = \frac{\prod_i n_{i+}!}{n_{++}!}\frac{\prod_j n_{+j}!}{\prod_i\prod_j z_{ij}!}. \end{aligned} \qquad (X.5.1)$$

L'ultima espressione deriva dal fatto che al numeratore

$$\prod_{i=1}^{I}\prod_{j=1}^{J} p_{j|+}^{z_{ij}} = \prod_{j=1}^{J}\prod_{i=1}^{I} p_{j|+}^{z_{ij}} = \prod_{j=1}^{J} p_{j|+}^{\sum_i z_{ij}} = \prod_{j=1}^{J} p_{j|+}^{n_{+j}}.$$

Questa distribuzione di probabilità è detta *ipergeometrica multipla* e ha il grande pregio di *non* dipendere da parametri incogniti.

Dal punto di vista geometrico l'insieme $\Im_{\mathbf{Z}}$ è costituito da matrici di contingenza i cui profili, sia righe che colonne, honno tutte lo stesso baricentro. Dal punto di vista matematico, invece, il fatto che ogni elemento di z_{ij} di $\mathbf{Z}$ sia condizionato dai totali marginali n_{i+} e n_{+j} implica che questo si realizza in una situazione di omogeneità

omogeneità

Esistono soltanto altre due matrici di contingenza di ordine 2×2 con gli stessi totali marginali della matrice $\mathbf{N}$ in (1.1.1), per cui l'insieme di riferimento $\Im_{\mathbf{Z}}$ è costituito da 3 matrici distinte. Le loro probabilità di realizzarsi sono

$$\mathbf{N} = \begin{pmatrix} 0 & 2 \\ 3 & 4 \end{pmatrix} \begin{matrix} 2 \\ 7 \end{matrix} \qquad P(\mathbf{N}) = \frac{2!\ 7!\ 3!\ 6!}{9!\ \ 0!\ 2!\ 3!\ 4!} = \frac{30}{72} = 0.417$$
$$3 \quad 6 \quad 9$$

$$\mathbf{Z_1} = \begin{pmatrix} 1 & 1 \\ 2 & 5 \end{pmatrix} \begin{matrix} 2 \\ 7 \end{matrix} \qquad P(\mathbf{Z}_1) = \frac{2!\ 7!\ 3!\ 6!}{9!\ \ 1!\ 1!\ 2!\ 5!} = \frac{36}{72} = 0.500$$
$$3 \quad 6 \quad 9$$

$$\mathbf{Z_2} = \begin{pmatrix} 2 & 0 \\ 1 & 6 \end{pmatrix} \begin{matrix} 2 \\ 7 \end{matrix} \qquad P(\mathbf{Z}_2) = \frac{2!\ 7!\ 3!\ 6!}{9!\ \ 2!\ 0!\ 1!\ 6!} = \frac{6}{72} = 0.083.$$
$$3 \quad 6 \quad 9$$

Quando la matrice di contingenza è di ordine 2×2, basta che sia noto uno solo degli elementi per ricavare per differenza gli altri 3. Se, ad esempio, è noto z_{11}, la matrice ha l'aspetto

$$\mathbf{Z} = \begin{pmatrix} z_{11} & [n_{1+} - z_{11}] & n_{1+} \\ [n_{+1} - z_{11}] & [n_{++} - n_{1+} - n_{+1} + z_{11}] & n_{2+} \\ n_{+1} & n_{+2} & n_{++} \end{pmatrix}$$

e la sua probabilità di essere osservata, in base alla (X.5.1) è

$$P(\mathbf{Z}) = \frac{n_{1+}!\ n_{2+}!\ n_{+1}!\ n_{+2}!}{n_{++}!\ \ z_{11}!\ [n_{1+} - z_{11}]!\ [n_{+1} - z_{11}]!\ [n_{++} - n_{1+} - n_{+1} + z_{11}]!}$$
$$= \frac{\binom{n_{+1}}{z_{11}} \binom{n_{++} - n_{+1}}{n_{1+} - z_{11}}}{\binom{n_{++}}{n_{1+}}} \qquad (X.5.2)$$

dove il coefficiente binomiale

$$\binom{a}{b} = \frac{a\,!}{b\,!\,(a-b)\,!}$$

X.2 - Bootstrap

Se si ipotizza che i vettori di ricampionamento $\mathbf{x}^*_{(b)}$, $(b = 1, 2 \ldots, B)$ sono estratti da una multinomiale, allora il

$$\text{numero di campioni } \mathbf{x}^*_{(b)} = \binom{2\,n_{++} - 1}{n_{++}}$$

dove n_{++} è l'ampiezza del campione e di $\mathbf{x}^*_{(b)}$. Questi sono i campioni che differiscono per almeno un elemento. Senza questa condizione, il numero di campioni di bootstrap sarebbe $n_{++}^{n_{++}}$.

i metodi dipendenti dal computer sostituiscono il tempo di calcolo all'analisi teorica Sebbene il metodo sia semplice nel suo principio, la sua applicabilità va sempre verificata in ogni applicazione attraverso lo studio delle sue proprietà asintotiche(Bickel e Friedman, 1981).

Ora, si può mostrare che l'inerzia rispetto al baricentro è legata all'indice Ψ χ^2 che misura complessivamente il legame tra due variabili nominali dalla relazione $\chi^2 = In_{\overline{\mathbf{r}}}/n_{++} = In_{\overline{\mathbf{c}}}/n_{++} = (1/n_{++}) \sum_a \lambda_a$. Questo indice segue una distribuzione χ^2 ad A gradi di libertà e permette di testare l'ipotesi di indipendenza tra le due variabili. Ora, un importante teorema [1] afferma che le percentuali d'inerzia estratte dai fattori sono indipendenti, nel senso che a questo termine si dà nella Teoria delle Probabilità, dall'inerzia rispetto al baricentro, e quindi anche dal χ^2 . Perciò anche se un il valore di $In_{\overline{\mathbf{r}}} = In_{\overline{\mathbf{c}}}$ non permette di respingere l'ipotesi di indipendenza del test, le prime percentuali di varianza possono essere significativamente alte e quindi l'Analisi delle Corrispondenze può essere impiegata anche per matrici di contingenza per le quali il valore dell'inerzia indica povertà d'informazione. Al contrario, può darsi il caso di matrici con valori d'inerzia che portano a respingere l'ipotesi di indipendenza, ma che tuttavia hanno percentuali di varianza non significative. In tali casi l'Analisi delle Corrispondenze non è il metodo più adatto a descrivere la dipendenza tra modalità. Comunque le percentuali d'inerzia vanno sempre comparate alle dimensioni della matrice (?)

---

[1] La dimostrazione può trovarsi in Lebart e al. (1984), citato nella bibliografia del Capitolo 5.

PARTE SECONDA: LE APPLICAZIONI

CAPITOLO 8: Matrici di contingenza e di incidenza

Sommario

Questo Capitolo approfondisce alcuni aspetti applicativi dell'analisi di una matrice di contingenza e costituisce perciò il naturale completamento dei Capitoli 2, 3 e 4. Le matrici prese ad esempio sono intenzionalmente di ridotte dimensioni per concentrare l'attenzione del lettore su un singolo aspetto dell'analisi e per contenere le distorsioni nella rappresentazione dei profili sulla mappa. Molte delle considerazioni qui svolte possono estendersi anche agli altri tipi di matrice che verranno passati in rassegna nei prossimi Capitoli.

L'attenta lettura di questo Capitolo, metterà il lettore in grado di comprendere

- la differenza che intercorre tra una matrice di contingenza e una di incidenza;
- le proprietà dell'Optimal Scaling e le sue connessioni con l'Analisi delle Corrispondenze;
- le trasformazioni preliminari all'analisi di matrici con valori negativi, con dati mancanti o con zeri strutturali;
- l'interpretazione di alcune forme tipiche delle nuvole dei profili sulle mappe fattoriali;
- il modo corretto di proiettare profili illustrativi sulle mappe fattoriali;
- le peculiarità dell'analisi di una matrice di profili;
- la capacità dell'Analisi delle Corrispontenze di rivelare associazioni tra modalità di più di due variabili categoriche.

CAPITOLO 8

8.1 - Matrici di contingenza e di incidenza

Le *matrici di contingenza* si incontrano di frequente perché, tra l'altro, permettono di riassumere in forma compatta i risultati parziali di sondaggi e di ricerche di mercato, che sono strumenti d'indagine piuttosto diffusi. L'Analisi delle Corrispondenze è stata concepita proprio per l'analisi di questo tipo di matrice, che si ottiene quando su ogni elemento di un insieme vengono rilevate contemporaneamente le modalità i e j di due variabili *qualitative*, aventi rispettivamente I e J modalità esclusive. Per evidenziare la numerosità delle $I \times J$ modalità congiunte, si ricorre ad una tabella a doppia entrata, o matrice, il cui elemento n_{ij} indica il numero di rilevamenti simultanei delle modalità i e j. La matrice ha I righe corrispondenti alle modalità della prima variabile e J colonne, corrispondenti alle modalità della seconda.

Per esempio, in *una* scuola ad un gruppo di scolari viene rilevato il colore degli occhi, classificato secondo I modalità, e quello dei capelli, classificato con J modalità. L'elemento n_{ij} indica il numero di scolari che in quella scuola hanno occhi del colore i e capelli del colore j.

Una matrice di contingenza si ottiene quindi ventilando gli elementi di *un* insieme secondo le modalità di due variabili qualitative e in questo senso è l'estensione bidimensionale della distribuzione unidimensionale di frequenza.

Le *matrici d'incidenza* sono invece ottenute affiancando distribuzioni di frequenza unidimensionali osservate su insiemi *diversi*. Ad esempio le J modalità del colore dei capelli degli alunni sono osservate in I scuole *diverse* ed n_{ij} indica il numero di scolari che nella scuola i hanno capelli di colore j.

Le matrici di contingenza sono sottoposte all'Analisi delle Corrispondenze senza trasformazioni preliminari [1], mentre su quelle d'incidenza è opportuno operare in certi casi una preventiva trasformazione. Per il fatto che

[1] Fanno eccezione le matrici con zeri strutturali, considerate nella Sez. 8.5.

i profili di queste matrici sono stati osservati su insiemi diversi, può capitare che le masse differiscano notevolmente, anche di qualche ordine di grandezza. Si supponga, ad esempio, che la matrice registri il numero di disoccupati in I settori industriali nei J paesi dell'Unione Europea. La massa del profilo Lussemburgo può risultare anche 1000 volte inferiore a quella della Germania, ma questo unicamente perché la Germania è più popolosa del Lussemburgo. Per ridurre le differenze e dare meno enfasi ad alcuni paesi, si può dividere ogni colonna per il numero di abitanti, o per il numero di occupati, ecc. prima di effettuare l'Analisi delle Corrispondenze. Altro esempio. In I parchi naturali europei si sono censite J specie animali. Non ha molto senso che la massa di un parco sia proporzionale al numero di animali presenti, perché ciò dipende dall'ampiezza del parco che è stata determinata da motivi storici, politici, ecc. Conviene quindi riesprimere i dati come frequenze per unità di area, dividendo ogni riga per la superficie del parco naturale.

A parte questi rari casi, il confine tra matrice di contingenza e d'incidenza è sfumato e dal punto di vista dell'Analisi delle Corrispondenze inessenziale. In questo Capitolo $\mathbf{N}$ indica indifferentemente una matrice di contingenza o di incidenza.

Un avvertimento: non si deve confondere la matrice d'incidenza appena definita con la matrice d'incidenza di un grafo, considerata nel Cap. 11.

8.2 - Strutture particolari della matrice

Nella Sezione 3.3 si è visto che l'indice complessivo di difformità tra profili è l'inerzia totale, ottenuta come media ponderata dei quadrati delle distanze distribuzionali dei profili dal loro profilo medio. Dal punto di vista geometrico l'inerzia misura la dispersione della nuvola dei profili intorno al loro baricentro.

Quando i profili hanno suppergiù la stessa forma per cui non c'è grande differenza tra un profilo e il profilo medio, la nuvola appare compatta e l'inerzia totale è prossima a zero. Tali sono anche gli autovalori, ossia l'inerzia delle proiezioni sugli assi fattoriali. Nel caso limite di profili tutti eguali tra loro e al profilo medio, ossia di una nuvola totalmente collassata nel baricentro, l'inerzia totale è nulla, nulli sono tutti gli autovalori e la matrice è quella omogenea $\mathbf{O}$ ottenuta dai totali marginali ed introdotta nella Sez. 1.9.

Il caso opposto, con profili completamente diversi tra loro, si ha

quando ogni profilo è concentrato in una sola modalità. In questo caso di perfetta dipendenza tra profili delle righe e delle colonne, l'inerzia totale raggiunge il massimo possibile che è $A = min\,(I-1, J-1)$, e le inerzie delle proiezioni dei profili sugli assi fattoriali sono tutte eguali a 1, ossia $\lambda_a = 1$ per ogni asse $a = 1, 2, \ldots, A$.

A parte queste forme degeneri che nei casi reali non si incontrano mai, esistono delle *forme tipiche* che la nuvola dei profili può assumere sulla mappa. Anche se la estrema varietà delle configurazzioni possibili non permette una rigorosa sistematizzazione dell'interpretazione di una mappa fattoriale, saper riconoscere queste forme tipiche è importante perché permette di individuare subito dei fenomeni importanti che potrebbero passare inosservati e anche di interpretare più speditamente, e correttamente, le mappe fattoriali. In tutti gli altri casi le mappe devono essere interpretate secondo le regole generali presentate nella Sez. 4.11. Le forme tipiche sono ovviamente configurazioni schematiche estremamente nitide che nella realtà si presentano in modo più sfumato.

Due nuvole separate

Sulla mappa principale i profili risultano divisi in due nuvole come nella FIG. 8.2.1, in alto a sinistra. In tal caso, se le righe (le colonne) della matrice $\mathbf{N}$ vengono riordinate per valori crescenti, o decrescenti, delle coordinate f_{i1} (g_{j1}) sul primo asse fattoriale, la matrice assume la forma a blocchi

$$\mathbf{N} \simeq \begin{pmatrix} \mathbf{N}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 \end{pmatrix}$$

con *due* blocchi quasi nulli fuori dai blocchi diagonali che contengono invece i valori più elevati. Ciò rivela che si è in presenza di due corrispondenze distinte. Occorre allora procedere ad analisi diverse. Se il numero di profili in una delle due nuvole, per esempio in $\mathbf{N}_2$ è piccolo, si procede all'Analisi delle Corrispondenze di $(\mathbf{N}_1\ \mathbf{0})$ considerando i profili delle righe di $(\mathbf{0}\ \mathbf{N}_2)$ come illustrativi. Se invece il numero di profili nelle due nuvole è suppergiù lo stesso, è bene fare due analisi separate: di $(\mathbf{N}_1\ \mathbf{0})$ con i profili delle righe di $(\mathbf{0}\ \mathbf{N}_2)$ come illustrativi e quindi di $(\mathbf{0}\ \mathbf{N}_2)$ con quelli di $(\mathbf{N}_1\ \mathbf{0})$ come illustrativi. L'analisi può essere anche fatta ripartendo la matrice $\mathbf{N}$ per colonne; ciò dipende dagli obiettivi dell'analisi.

Quando i sottoinsiemi dei profili appartenenti alle due nuvole sono perfettamente disgiunti, il primo autovalore non banale vale 1. Questa situazione non si presenta mai nei casi reali, ma tutte le volte che $\lambda_1 \simeq 1$ oc-

corre fare attenzione, specialmente quando si analizzano matrici di flusso, di import-export, di immigrazione-emigrazione, ecc. tutte trattate nel Cap. 11, perché il fatto rivela che alcuni percorsi sono privilegiati e altri praticamente inesistenti o proibiti.

Tre nuvole separate

La forma mostrata nella FIG. 8.2.1, in alto a destra, si presenta meno frequentemente, ma rivela che la matrice è ripartita in tre matrici di corrispondenza. Riordinando le righe o le colonne in base alle loro coordinate fattoriali sul primo asse, la matrice dei dati assume la forma

$$\mathbf{N} \simeq \begin{pmatrix} \mathbf{N}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{N}_3 \end{pmatrix}.$$

Anche in questo caso può essere utile analizzare separatamente le sottomatrici con più profili, considerando le altre come illustrative.

Nuvola a ferro di cavallo

La nuvola ha un aspetto parabolico che può pensarsi ottenuto intercalando alle tre nuvole separate del caso precedente degli altri profili fino ad ottenere un continuo. La forma a ferro di cavallo, o a 'croissant', si presenta frequentemente quando una delle variabili è ordinata, per esempio giudizi espressi in base a una scala di punteggi, o è diventata tale perché una variabile continua è stata suddivisa il classi ordinate: classi d'età, scaglioni di reddito, ecc. In questi casi, per meglio guidare l'occhio sulla mappa, si possono collegare ordinatamente con un tratto di matita i punti che rappresentano i profili delle classi. Una rappresentazione tipica è quella della FIG. 8.2.1 ove su un asse che di solito è il primo come in questo caso, appare una netta opposizione tra le classi di rango inferiore e quelle di rango più elevato, mentre l'altro asse oppone le classi intermedie a quelle estreme.

Il riordino di righe o colonne della matrice secondo i valori di f_{i1} o di g_{j1} rispettivamente, fa comparire una matrice con una banda diagonale di elementi a valore relativamente elevato ($n_{ij} \gg n_{i+}\, n_{+j}/n_{++}$), mentre quelli più lontani hanno valori inferiori ($n_{ij} \ll n_{i+}\, n_{+j}/n_{++}$), se non nulli o quasi. Il riferimento è la matrice omogenea $\mathbf{O}$ in cui tutti i profili sono eguali, introdotta nella Sez. 1.9. Le 'equazioni' dei fattori assumono la forma

$$\begin{aligned} f_{i2} &= a_1\, f_{i1}^2 + a_2\, f_{i1} + a_3 \\ g_{i2} &= b_1\, g_{i1}^2 + b_2\, g_{i1} + b_3. \end{aligned} \tag{8.2.1}$$

Anche se fin dalla Sez. 4.2 e dalla Sez. 4.8 è noto che i fattori non sono correlati linearmente due a due, ciò non esclude che tra essi possano intercorrere dei legami di tipo *non* lineare, ad esempio quadratico come nelle (8.2.1).

Una configurazione di questo tipo mette in luce il cosidetto *effetto Guttman*[1] che rivela un certo grado di ridondanza tra i profili delle due variabili quando si trovano 'sul' ferro di cavallo: dalla conoscenza del profilo riga $\mathbf{r}_i$ si desume quella del profilo colonna $\mathbf{c}_j$. Il terzo fattore è una funzione di terzo grado del secondo e così via. In altri termini, i fattori di rango successivo al primo non fanno che confermare l'informazione geometrica sulla configurazione della nuvola dei profili data dal primo. In pratica la situazione non è mai così netta, per cui spesso il secondo asse fattoriale affina ulteriormente l'interpretazione del primo.

In alcuni casi, all'interno dei due rami della parabola viene a trovarsi qualche profilo. In base alle relazioni di transizione, ciò significa che le modalità rappresentate da questi profili sono associate simultaneamente a modalità che si oppongono sul primo asse e che di solito sono esclusive.

Se possibile, è utile esaminare questa importante forma tipica in un sottospazio tridimensionale, generato dai primi assi fattoriali e, comunque, ogni allontanamento dalla regolarità deve mettere in allerta e spingere all'esame diretto delle matrici dei dati e dei profili: potrebbe trattarsi di un fenomeno interessante, ma anche di un errore.

Nuvola a triangolo

Anche la forma triangolare , mostrata nella FIG. 8.2.1, è abbastanza frequente: la nuvola dei profili ha l'aspetto di un tetraedro nel sottospazio generato dai primi tre assi fattoriali. Il primo asse oppone quindi dei profili che si differenziano poco sul secondo asse e molto invece sul terzo ad altri profili che si differenziano poco sul terzo asse e molto sul secondo.

Nuvola con profili aberranti

Questa forma si presenta, come nella FIG. 8.2, con uno o più profili lontani dal complesso della nuvola. Si tratta di profili la cui struttura si differenzia nettamente da quella media, rappresentata dal baricentro, e da quelle degli altri profili. È sempre bene controllare sulla matrice dei dati che non si tratti dell'errore di trascrizione di un qualche elemento. I profili aberranti vengono di solito retrocessi dal rango di attivi a quello di illustrativi,

---

[1] Louis Guttman: New York 1916, Gerusalemme, Minneapolis 1987.

pur con le cautele suggerite nella Sez. 4.11 *in fine* e nella Sez. 4.12, e l'analisi viene rifatta.

8.3 - Matrici con elementi negativi

Può presentarsi il caso in cui a una matrice di contingenza è necessario sottrarne un'altra: la matrice risultante, a causa delle fluttuazioni statistiche, può avere qualche elemento negativo. Analoga situazione si incontra con i residui lasciati dal fit di un modello log-lineare. L'Analisi delle Corrispondenze è in grado di analizzare matrici con alcuni elementi negativi purché siano *positivi* i totali marginali delle righe e delle colonne. I problemi possono sorgere col software d'analisi che di solito effettua una scansione preliminare della matrice, eliminando le righe, se $I \geq J$, o le colonne, se $J > I$, che contengono elementi negativi.

Quando anche i totali marginali non sono positivi, questi possono essere resi positivi *sommando* alla matrice una matrice costante. L'operazione, una traslazione dal punto di vista geometrico, non altera la dispersione dei profili rispetto al loro baricentro, e quindi l'inerzia totale che ne è la misura, ma altera i profili ed è quindi da evitare, a meno che la costante da sommare sia piccola per cui la perturbazione resti limitata.

Un'altra possibilità è quella di accorpare la riga o la colonna con elementi negativi con un'altra non negativa, purché l'accorpamento abbia senso e non stravolga l'obiettivo dell'analisi.

Il metodo per ottenere fattori e tassi d'inerzia di una matrice con alcuni elementi negativi, ma con totali marginali *positivi*, è dovuto a van der Heijden (1987). Supponendo per un momento che $\mathbf{N}$ indichi la matrice dei dati con qualche elemento negativo, essa può essere ricostruita tramite la (4.10.3), ove si è indicata con $\mathbf{O} = n_{++}\,\overline{\mathbf{c}}\,\overline{\mathbf{r}}^T$ la matrice omogenea in cui i profili non si distinguono dal loro profilo medio e che ha i medesimi totali marginali di $\mathbf{N}$,

$$\mathbf{N} = \mathbf{O} + n_{++}\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{F}\,\mathbf{D}_{\lambda}^{-\frac{1}{2}}\mathbf{G}^T\,\mathbf{D}_{\overline{\mathbf{r}}}.$$

Scegliendo un opportuno valore di una costante positiva c, si può senz'altro ottenere da $\mathbf{N}$ una matrice

$$\mathbf{N}^* = c\,\mathbf{O} + \mathbf{N}$$

senza elementi negativi. Dal momento che $\mathbf{O}$ ha i medesimi totali marginali di $\mathbf{N}$, il totale generale di $\mathbf{N}^*$ risulta

$$n^*_{++} = (1+c)\,n_{++} \qquad \text{mentre} \qquad \mathbf{D}^*_{\overline{\mathbf{c}}} = \mathbf{D}_{\overline{\mathbf{c}}} \quad \mathbf{D}^*_{\overline{\mathbf{r}}} = \mathbf{D}_{\overline{\mathbf{r}}}.$$

Sussiste allora la relazione

$$\begin{aligned}\mathbf{N}^* &= (1+c)\,\mathbf{O} + (1+c)\,n_{++}\,\mathbf{D}_{\overline{\mathbf{c}}}\,\frac{\mathbf{F}}{(1+c)}\,(1+c)\,\mathbf{D}_\lambda^{-\frac{1}{2}}\,\frac{\mathbf{G}^T}{(1+c)}\,\mathbf{D}_{\overline{\mathbf{r}}} \\ &= \mathbf{O}^* + n_{++}^*\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{F}^*\,\mathbf{D}_\lambda^{*-\frac{1}{2}}\,\mathbf{G}^{*T}\,\mathbf{D}_{\overline{\mathbf{r}}} \qquad (8.3.1)\end{aligned}$$

dove si è posto

$$\mathbf{O}^* = (1+c)\,\mathbf{O} \qquad n_{++}^* = (1+c)\,n_{++}$$

e

$$\mathbf{F}^* = \frac{\mathbf{F}}{(1+c)} \qquad \mathbf{G}^* = \frac{\mathbf{G}}{(1+c)} \qquad \mathbf{D}_\lambda^{*-\frac{1}{2}} = (1+c)\,\mathbf{D}_\lambda^{-\frac{1}{2}}.$$

Perciò dalla (8.3.1) si ricava che l'Analisi delle Corrispondenze della matrice $\mathbf{N}^*$ produce dei fattori delle righe e delle colonne che sono più piccoli di un fattore $(1+c)$ di quelli che si otterrebbero analizzando direttamente la matrice $\mathbf{N}$ con i valori negativi. Invece i fattori *standard* sono *eguali* perché dalla (4.8.16) si ottiene

$$\hat{\mathbf{F}}^* = \mathbf{F}^*\,\mathbf{D}_\lambda^{*-\frac{1}{2}} = \frac{\mathbf{F}}{(1+c)}\,(1+c)\,\mathbf{D}_\lambda^{-\frac{1}{2}} = \hat{\mathbf{F}} \quad \text{e} \quad \hat{\mathbf{G}}^* = \hat{\mathbf{G}}.$$

Anche i tassi d'inerzia sugli assi fattoriali sono gli stessi perché

$$\frac{1}{tr\,[\mathbf{D}_\lambda^*]}\,\mathbf{D}_\lambda^* = \frac{\frac{1}{tr\,[\mathbf{D}_\lambda]}}{(1+c)^2} \Big/ \frac{\mathbf{D}_\lambda}{(1+c)^2} = \frac{1}{tr\,[\mathbf{D}_\lambda]}\,\mathbf{D}_\lambda.$$

Pertanto, su ogni asse fattoriale $a = 1, 2, \ldots, A$,

$$\tau_a^* = \frac{\lambda_a^*}{\sum_a \lambda_a^*} = \frac{\lambda_a}{\sum_a \lambda_a} = \tau_a.$$

Prima dell'analisi, la matrice $\mathbf{N}^*$ può essere anche portata ad avere tutti i suoi elementi *interi* moltiplicandola per una opportuna costante positiva intera. Si è visto nella Sez. 3.2, che due matrici proporzionali hanno gli stessi fattori.

8.4 - Matrici con dati mancanti

Per dato mancante si intende un elemento di una matrice al quale per un qualche motivo non è stato possibile assegnare un valore. Questo può succedere ad esempio, quando i dati provengono da sensori dislocati sul territorio in una serie di stazioni di rilevamento. Se le righe riportano un certo numero I di date e le colonne si riferiscono a J stazioni di rilevamento, il valore n_{ij} dell'elemento (i,j) può indicare il numero di volte che alla data

i nella stazione j si è superata la soglia di attenzione. Se alla data i il rivelatore della stazione j era guasto, nell'elemento (i, j) viene scritto $n.p.$, dato non pervenuto.

L'Analisi delle Corrispondenze non è in grado di analizzare una tale matrice, a meno che ogni dato mancante non venga in qualche modo stimato o ricostruito e il modo tradizionale è quello di interpolare con una media ponderata i valori registrati alla data i nelle stazioni territorialmente più vicine alla stazione j.

Come ha mostrato Mutombo (1973), l'Analisi delle Corrispondenze, grazie alle formule di ricostruzione della Sez. 4.10, è in grado di fornire una stima del dato o dei dati mancanti con un procedimento iterativo che ricalca nella sostanza il metodo tradizionale. Il metodo è stato oggetto di ricerca da parte di molti altri autori che ne hanno stabilito la convergenza. In Bastin et al. (1980) si possono trovare casi più o meno semplici di ricostruzione di dati mancanti.

Quando nella matrice vi è *un solo* dato mancante ed è l'elemento (i, j), l'algoritmo si sviluppa nel modo seguente:

passo 1 - calcolare i totali marginali n_{i+} e n_{+j} e il totale generale n_{++} della matrice col dato mancante;

passo 2 - calcolare per l'elemento (i, j) il valore della corrispondente matrice omogenea: $o_{ij} = n_{i+}\, n_{+j}/n_{++}$ e porre $n_{ij} = o_{ij}$;

passo 3 - assegnare all'elemento (i, j) il valore n_{ij} appena ottenuto

$$(i, j) \longleftarrow n_{ij}.$$

Gli altri elementi conservano il valore che avevano inizialmente;

passo 4 - effettuare l'Analisi delle Corrispondenze della matrice aggiornata, ottenendo inerzie λ_a e fattori $\mathbf{f}_a$ e $\mathbf{g}_a$ per ogni asse a fino ad un rango A^*. L'esperienza mostra che per accelerare la convergenza è bene fissare A^* molto vicino al valore massimo $A = min\,(I - 1, J - 1)$;

passo 5 - calcolare i totali marginali n_{i+} e n_{+j} e il totale generale n_{++} della matrice aggiornata;

passo 6 - con i dati ottenuti nei passi *4* e *5* e con la formula di ricostruzione (4.10.1) calcolare

$$n_{ij} = \frac{n_{i+}\, n_{+j}}{n_{++}} \left(1 + \sum_{a=1}^{A^*} \frac{1}{\lambda_a}\, f_{ia}\, g_{ja}\right);$$

passo 7 - ritornare ad eseguire il *passo 3* a meno che questo nuovo valore n_{ij} non sia praticamente eguale a quello che già si trova in (i, j).

Nel caso piuttosto raro in cui i dati mancanti siano più di uno, è preferibile effettuare una stima *simultanea*, essendo il procedimento piuttosto gravoso in termini di tempo di calcolo. Il caso in cui manca o è indefinita l'intera *diagonale* principale di una matrice quadrata è trattato in dettaglio nella Sez. 11.x.

L'algoritmo di ricostruzione fornisce una stima tanto più soddisfacente quanto meglio i due profili $\mathbf{r}_i$ e $\mathbf{c}_j$ che si incrociano nell'elemento (i, j), sono approssimati da una combinazione lineare degli altri profili, ma non garantisce che essa sia unica, né che sia ottimale. Questo però importa relativamente perché l'algoritmo va visto come uno stratagemma per poter effettuare l'Analisi delle Corrispondenze della *intera* matrice. L'alternativa poco allettante sarebbe la drastica soppressione della riga i e della colonna j con tutte le informazioni che contengono.

8.5 - Matrici con zeri strutturali

Per monitorare nel tempo il grado di soddisfazione del servizio offerto, i supermercati effettuano delle interviste periodiche tra la clientela. Le indagini durano di solito una settimana e vengono ripetute in periodi diversi dell'anno. La matrice di TAV. 8.5.1 è tratta da un'indagine condotta in un ipermercato della Riviera adriatica e ripartisce il numero di clienti intervistati per fascia oraria e per giorno d'indagine. Le frequenze n_{ij} indicano il numero di intervistati nella fascia oraria i del giorno j e, per il modo con cui è stato selezionato il campione, possono ritenersi con buona approssimazione proporzionali all'affluenza della clientela.

Si è ripetuto più volte che l'Analisi delle Corrispondenze di una matrice di contingenza è sempre possibile. Ebbene, l'analisi della matrice nella TAV. 8.5.1 *non* lo è. Il motivo sono i 5 zeri della mattina del lunedì: il supermercato era *chiuso* e nessuna intervista era possibile. Questi zeri sono detti *strutturali* perché segnalano l'assenza di una qualsiasi relazione tra l'affluenza nelle ore mattutine e il lunedì. Non vanno confusi con gli zeri *statistici* dovuti a fluttuazioni accidentali perché, ad esempio, un violento nubifragio ha tenuto lontani i clienti. Ripetendo il rilevamento in un'altra settimana, gli zeri statistici possono sparire, ma quelli strutturali restano.

L'Analisi delle Corrispondenze di matrici con zeri strutturali deve essere fatta escludendo gli elementi con tali zeri o, almeno, tenendone conto il

meno possibile. Per capire come, occorre partire dalla formula di ricostruzione (4.10.1) di un elemento n_{ij} della matrice. Se la somma sui fattori viene arrestata ad un ordine $A^* \leq A = min\,(I-1, J-1)$, l'elemento ricostruito vale

$$n^*_{ij} = \frac{n_{i+}\, n_{+j}}{n_{++}} \left(1 + \sum_{a=1}^{A^*} \frac{1}{\sqrt{\lambda_a}}\, f_{ia}\, g_{ja} \right). \tag{8.5.1}$$

Il secondo membro di questa relazione può essere visto come un 'modello', il modello dell'Analisi delle Corrispondenze, che ha come parametri λ_a, f_{ia} e g_{ja} legati dalle relazioni di standardizzazione e di incorrelazione

$$\sum_{i=1}^{I} \overline{c}_i\, f_{ia} = \sum_{j=1}^{J} \overline{r}_j\, g_{ja} = 0 \qquad \text{e} \qquad \sum_{i=1}^{I} \overline{c}_i\, f^2_{ia} = \sum_{j=1}^{J} \overline{r}_j\, g^2_{ja} = \lambda_a$$

$$\sum_{i=1}^{I} \overline{c}_i\, f_{ia}\, f_{ib} = \sum_{j=1}^{J} \overline{r}_j\, g_{ja}\, g_{jb} = 0 \quad \text{quando} \quad a, b = 1, 2, \ldots, A^* \quad \text{e} \quad b \neq a$$

il modello è bilineare ed i suoi parametri si stimano...

I parametri del 'modello' si stimano minimizzando la somma ponderata dei quadrati degli scarti tra valori osservati e stimati col 'modello'

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{1}{n_{i+}\, n_{+j}} (n_{ij} - n^*_{ij})^2 = minimo \tag{8.5.2}$$

e il grado di bontà del fit si ottiene dalla somma degli autovalori di rango superiore ad A^*. Gli elementi con zeri strutturali devono essere esclusi dall'algoritmo di stima dei parametri e quindi dalla somma nella (8.5.2). Se in questi elementi si sostituisse agli 0 i valori n^*_{ij} ottenuti dalla (8.5.1), essi non darebbero più alcun contributo alla somma e neppure inciderebbero sulla bontà del fit.

I valori da sostituire agli zeri strutturali si possono ottenere con un algoritmo di tipo iterativo che ricalca quello della Sez. 8.4, ma snellito, perché di solito gli zeri strutturali sono più di uno. L'esperienza ha mostrato che l'algoritmo meno costoso in termini di tempo di calcolo è il seguente:

passo 1 - dai totali marginali n_{i+} e n_{+j} e dal totale generale n_{++} della matrice con gli zeri strutturali calcolare per ogni elemento (i, j) con uno zero strutturale il valore della corrispondente matrice omogenea: $o_{ij} = n_{i+}\, n_{+j}/n_{++}$ e porre $n_{ij} = o_{ij}$;

passo 2 - assegnare a ogni elemento (i, j) con zero strutturale il valore o_{ij} appena ottenuto. Gli altri consevano il valore che avevano inizialmente;

passo 3 - ripetere i passi *1* e *2* finché nessuno dei valori o_{ij} ottenuti è molto diverso da quello calcolato nell'iterazione precedente;

passo 4 - effettuare l'Analisi delle Corrispondenze della matrice completata e calcolare con la formula di ricostruzione (8.5.1) i valori n^*_{ij} per ciascun elemento che inizialmente conteneva uno zero strutturale. L'ordine A^* fino al quale spingere la somma nella (4.10.1) deve essere molto prossimo al valore massimo A ;

passo 5 - sostituire agli zeri strutturali i valori n^*_{ij} ottenuti al passo precedente.

Per la matrice di TAV. 8.5.1 sono occorse 12 iterazioni per arrivare a una convergenza soddisfacente. Nella stessa Tavola è riportata la matrice con i valori sostitutivi dei 5 zeri strutturali, arrotondati all'intero più prossimo.

In presenza di un solo zero strutturale, Greenacre (1984) suggerisce un algoritmo del tipo di quello presentato nella Sez. 8.4. In tutti i casi, comunque, gli elementi della matrice ai quali vengono assegnati dei valori ottenuti dall'algoritmo iterativo, danno un contributo praticamente nullo alla somma (8.5.2) e intervengono in modo trascurabile, anche se non sempre nullo, alla bontà del fit tra matrice e 'modello'.

8.6 - Impiego dei profili illlustrativi

Nella Sez. 4.12 si è visto che un profilo illustrativo viene posizionato su un asse fattoriale tramite le relazioni di transizione che ne determinano la coordinata come media delle coordinate dei profili dell'altra nuvola, ponderata con le componenti del profilo illustrativo. Si è visto anche che ogni profilo illustrativo viene proiettato *indipendentemente* da ogni altro eventuale profilo illustrativo che fosse presente. Le righe e le colonne illustrative di una matrice di contingenza attiva possono essere di 3 tipi: di frequenze, di presenza/assenza e di valori espressi in una unità di misura.

Righe o colonne di frequenze

Una riga o una colonna illustrativa costituita da una distribuzione di frequenze, sia assolute che relative, non richiede alcuna trasformazione preliminare. Talvolta però può tornare utile suddividere le frequenze in classi, per esempio in tre: frequenza alta, media e bassa, codificando poi le tre modalità in forma disgiuntiva completa, in modo che ogni riga o colonna

possieda una ed una sola modalità. Prima dell'analisi le modalità devono essere ricodificate nel modo spiegato qui appresso.

Righe o colonne di presenza / assenza

Si supponga per fissare le idee che illustrativa sia la colonna $\tilde{\mathbf{d}}$, in cui $\tilde{d}_i = 0$ indica l'assenza della caratteristica per la modalità i, e $\tilde{d}_i = 1$ indica invece la sua presenza. Il suo profilo $\tilde{\mathbf{c}}$ si ottiene dividendo ogni elemento per la loro somma, ossia per il numero di 1 che ci sono nella colonna $\tilde{\mathbf{d}}$ e che verrà indicato con $\#1$

$$\tilde{c}_i = \frac{\tilde{d}_i}{\sum_{i=1}^{I} \tilde{d}_i} \qquad \text{e quindi} \qquad \tilde{c}_i = 0 \quad \text{o} \quad \tilde{c}_i = \frac{1}{\#1}.$$

Se si applicano direttamente le relazioni di transizione (4.12.1) al profilo $\tilde{\mathbf{c}}$ si ottiene

$$\tilde{g}_a = \frac{1}{\sqrt{\lambda_a}} \sum_{i=1}^{I} \tilde{c}_i \, f_{ia} = \frac{1}{\sqrt{\lambda_a}} \sum \frac{1}{\#1} \, f_{ia} \tag{8.6.1}$$

dove $\sum$ indica che la somma va fatta sulle sole componenti i non nulle del profilo $\tilde{\mathbf{c}}$. L'espressione (8.6.1) mostra che la coordinata $\tilde{g}_a$ del profilo illustrativo è una media ponderata soltanto con le componenti non nulle di $\tilde{\mathbf{d}}$, per cui $\tilde{g}_a$ va a finire sull'asse dove i profili delle righe sono più *numerosi*. Nel caso limite di una colonna illustrativa costituita tutta di 1, colonna che non darebbe alcuna informazione sulla differenziazione dei profili delle righe, il profilo non verrebbe proiettato nell'origine come dovrebbe, dato che la media ponderata di un fattore deve essere nulla. Per avere la colonna illustrativa $\tilde{\mathbf{d}}$ posizionata correttamente sull'asse occorre ponderarne *preventivamente* ogni componente con le masse $\overline{c}_i$ dei profili attivi, creando la colonna $\tilde{\mathbf{d}}^*$ con componenti che valgono

$$\tilde{d}_i^* = \overline{c}_i \, \tilde{d}_i = \overline{c}_i \times 0 = 0 \qquad \text{oppure} \qquad \tilde{d}_i^* = \overline{c}_i \, \tilde{d}_i = \overline{c}_i \times 1 = \overline{c}_i. \tag{8.6.2}$$

Il suo profilo è

$$\tilde{c}_i^* = \frac{\tilde{d}_i^*}{\sum_{i=1}^{I} \tilde{d}_i^*} \qquad \text{e quindi} \qquad \tilde{c}_i^* = 0 \quad \text{o} \quad \tilde{c}_i^* = \frac{\overline{c}_i}{\sum \overline{c}_i}$$

che proiettato sull'asse si posiziona in

$$\tilde{g}_a^* = \frac{1}{\sqrt{\lambda_a}} \sum_{i=1}^{I} \tilde{c}_i^* \, f_{ia} = \frac{1}{\sqrt{\lambda_a}} \sum \frac{\overline{c}_i}{\sum \overline{c}_i} \, f_{ia} \tag{8.6.3}$$

dove $\sum$ indica che la somma va effettuata soltanto sulle componenti non nulle di $\tilde{\mathbf{d}}^*$.

A riprova della correttezza della ricodifica (8.6.2), si vede subito che se la colonna illustrativa $\tilde{\mathbf{d}}$ fosse costituita tutta da 1, la colonna trasformata $\tilde{\mathbf{d}}^*$ avrebbe tutte le sue I componenti eguali a $\overline{c}_i$ e nella (8.6.3) si avrebbe $\sum \overline{c}_i = 1$ e quindi $\tilde{g}^*_a = \sum \overline{c}_i \, f_{ia} = 0$, perché un fattore ha media ponderata nulla. Il profilo verrebbe proiettato correttamente nell'origine dell'asse.

Quando illustrativa è invece una riga $\tilde{\mathbf{b}}$, si procede in modo analogo a quanto esposto qui sopra, sostituendo a ogni 1 di $\tilde{\mathbf{b}}$ la massa $\overline{r}_j$ del corrispondente profilo colonna attivo e creando così la riga $\tilde{\mathbf{b}}^*$ con componenti

$$\tilde{b}^*_j = \overline{r}_j \, \tilde{b}_j = \overline{r}_j \times 0 = 0 \qquad \text{oppure} \qquad \tilde{b}^*_j = \overline{r}_j \, \tilde{b}_j = \overline{r}_j \times 1 = \overline{r}_j . \qquad (8.6.4)$$

Si procede poi all'Analisi delle Corrispondenze con $\tilde{\mathbf{b}}^*$ come riga illustrativa.

Soltanto nel caso di matrici con riga (colonna) marginale costante, per esempio matrici di profili $\mathbf{R}$ o $\mathbf{C}$ della Sez. 8.8, di matrici espresse in forma disgiuntiva completa del Cap. 5 o di matrici di punteggi complementate del Cap. 9, la riga (colonna) illustrativa codificata con 0 e 1 può essere proiettata *senza* ulteriore ricodifica perché quando le masse sono tutte eguali, nella (8.6.3) si ha $\sum \overline{c}_i = \#1 \times \overline{c}_i$ e quindi la (8.6.3) si riduce alla (8.6.1). La trasformazione non modificherebbe il risultato.

Righe o colonne di misure

Una variabile illustrativa di tipo quantitativo, i cui valori sono espressi in una unità di misura, deve essere *necessariamente* ricodificata prima dell'analisi. L'intervallo di variazione dei valori va suddiviso in classi, la variabile codificata in forma disgiuntiva completa e gli 1, attestanti l'appartenenza del valore a quella classe, sostituiti con le corrispondenti masse dei profili attivi, come si è visto sopra.

Un esempio servirà a chiarire il procedimento esposto in questa Sezione. La TAV. 8.6.1 riporta la situazione degli istituti carcerari del Veneto al 31 dicembre 1992. Nelle 7 province i reclusi sono suddivisi per Grado di Giudizio: 1 - in attesa di giudizio, 2 - appellante e 3 - giudicato definitivamente. È riportato anche l'Indice di Affollamento o rapporto tra il numero di reclusi e la capienza dell'istituto carcerario della provincia. Questa variabile quantitativa non può essere direttamente proiettata come illustrativa, per cui, prendendo come riferimento la mediana che vale 1.34, i suoi valori sono stati ripartiti in due classi, [0 - 1.34] e [1.35 - 1.60], creando così una nuova variabile a due modalità, indicate con D1 e D2, codificate in forma disgiuntiva completa come indicato nella Sez. 5.X. In D1 lo 0 indica che l'indice di affollamento è inferiore al valore mediano e 1 che lo supera. Nella colonna D2 è il contrario. T1 e T2

sono invece le colonne ottenute sostituendo in D1 e D2 ad ogni 1 la massa della provincia. Nella FIG. 8.6.2 è riprodotta la mappa ottenuta dall'Analisi delle Corrispondenze della matrice 7×3, con le colonne illustrative D1, D2 e T1 e T2. Se si collegano con un tratto di penna i due ultimi punti, si vede che il segmento passa per l'origine che è il loro baricentro. Questo non avviene congiungendo i punti D1 e D2 delle due modalità codificate con 0 e 1, perché come si è visto sopra, sono dislocati dove i profili delle righe sono più numerosi.

8.7 - Optimal scaling [1]

La matrice d'incidenza nella TAV. 8.7.1 raccoglie gli esiti di un intervento chirurgico innovativo al ginocchio registrati in 5 importanti strutture ospedaliere, selezionate per partecipare ad uno studio e qui indicate con delle sigle. Gli esiti dell'intervento sono classificati in base alla Ripresa Funzionale dell'Arto: Nulla o molto Limitata, Parziale e Completa.

Con matrici di questo tipo, in cui una delle variabili ha le modalità ordinate, ma non l'altra, si è talvolta interessati ad assegnare un punteggio, in inglese *score*, alle modalità di quest'ultima. Essendo lo score assegnato un numero reale, si possono redigere delle graduatorie, calcolare medie e varianze, o effettuare procedure statistiche più complesse, come una regressione.

Nel caso specifico, per essere considerato un indicatore della performance di un ospedale, lo score deve basarsi sui tre gradi dell'esito di ogni intervento, ai quali vanno assegnati pesi differenziati, per esempio, linearmente progressivi come 1, 2 e 3. In tal modo, un esito con Ripresa Parziale verrà valutato il doppio di un esito con Ripresa Nulla o Molto Limitata e uno con Ripresa Totale il triplo. Fissati i pesi delle modalità ordinate, si è in grado di quantificare lo *score* di ciascun ospedale, che può essere considerato un indicatore di performance. Per esempio, l'Ospedale H.A riceve lo score

$$(1 \times 13 + 2 \times 18 + 3 \times 16)/47 = 1 \times \frac{13}{47} + 2 \times \frac{18}{47} + 3 \times \frac{16}{47} = \frac{97}{47} = 2.06.$$

Basta quindi moltiplicare i pesi per le componenti del profilo $\mathbf{c}_1$ dell'ospedale H.A. La graduatoria dei 5 ospedali si ottiene dal loro *score*, come nello specchietto

---

[1] L'equivalente italiano del termine inglese è *Graduazione Ottimale*, ma viene raramente usato.

	Ospedali					
	H.S	*H.A*	*H.P*	*H.M*	*H.B*	*Media*
Score	1.60	2.06	2.23	2.34	2.35	*2.10*

Lo score medio si può ottenere o dalla media dei 5 score ponderata con le masse $\overline{\mathbf{r}}$, o direttamente dal profilo medio $\overline{\mathbf{c}}$. Come si vede dei tre ospedali con performance sopra la media, i primi due risultano a pari merito, con uno score praticamente eguale.

Parrebbe quindi che per l'intervento chirurgico in questione, fosse indifferente farsi ricoverare presso gli ospedali H.B o H.M. Ma le cose stanno proprio così? Evidentemente la classifica dipende strettamente da come si pesano i tre gradi dell'esito e, d'altra parte, la scelta di assegnare pesi equispaziati è del tutto arbitraria e difficile da giustificare, per cui sorge legittima la domanda: esiste una terna di pesi che sia 'maggiormente giustificabile' ? La risposta dipende dal significato che si intende dare dare a questo ultimo aggettivo. D'altra parte, se l'obiettivo finale è quello di differenziare le performance, allora è opportuno che gli score risultino separati al massimo per evitare ex-aequo e indecifrabili affastellamenti nella graduatoria. E poiché la misura tradizionale di dispersione è la varianza, si può cercare quella terna di pesi che produca degli score con varanza *massima*. Dato però che la varianza può essere resa grande a piacere semplicemente aumentando il valore dei pesi, occorre stabilire un limite all'ampiezza del loro intervallo di variazione, e quindi alla loro varianza, e scegliere poi tra le terne di pesi con la medesima varianza quella con valore medio nullo. Se i pesi sono *standardizzati* con valore medio nullo e varianza unitaria sull'intero campione di 367 esiti, il problema diventa risolubile e, sorprendentemente, si riconduce al calcolo dei fattori di un'Analisi delle Corrispondenze della matrice di TAV. 8.7.1.

Indicati con p_1 , p_2 e p_3 i tre pesi incogniti da assegnare ai tre gradi della Ripresa Funzionale, le due condizioni comportano che

$$(p_1 \times 90 + p_2 \times 149 + p_3 \times 128)/367 = 0$$
$$(p_1^2 \times 90 + p_2^2 \times 149 + p_3^2 \times 128)/367 = 1$$

e se $\mathbf{p} = (p_1\ p_2\ p_3)^T$ indica il vettore colonna dei tre pesi incogniti, il vincolo della standardizzazione espresso in forma matriciale si scrive

$$\mathbf{p}^T\,\overline{\mathbf{c}} = 0 \qquad\qquad \mathbf{p}^T\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{p} = 1. \tag{8.7.1}$$

Se con $\mathbf{s} = (s_1\ s_2\ \ldots\ s_5)^T$ si indica il vettore colonna incognito degli scores dei 5 ospedali, lo score di un ospedale j che ha profilo $\mathbf{c}_j$ si ottiene dai tre pesi come

$$s_j = \mathbf{c}_j^T\,\mathbf{p} \qquad \text{e quindi} \qquad \mathbf{s} = \mathbf{C}^T\,\mathbf{p} \tag{8.7.2}$$

dove $\mathbf{C}$ indica al solito la matrice di ordine 3×5 dei profili delle colonne, pure riportata nella TAV. 8.7.1.

Si tratta ora di cercare quel vettore $\mathbf{p}$ di pesi che produca score con la massima varianza (dispersione)

$$\mathbf{s}^T\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{s} = \textit{massimo} \qquad \text{col vincolo} \qquad \mathbf{p}^T\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{p} = 1. \tag{8.7.3}$$

La varianza degli score espressa in funzione dei pesi tramite la (8.7.2) diventa

$$\mathbf{s}^T\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{s} = (\mathbf{C}^T\mathbf{p})^T\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T\mathbf{p} = \mathbf{p}^T\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T\mathbf{p} \tag{8.7.4}$$

con il vincolo (8.7.3) che i pesi abbiano varianza unitaria. Per trovare il massimo vincolato della (8.7.4) si procede come indicato nella Sez. 3.8, costruendo la funzione di Lagrange

$$\mathcal{L}(\mathbf{p}, \lambda) = \mathbf{p}^T\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T\mathbf{p} + \lambda\,(1 - \mathbf{p}^T\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{p})$$

la si deriva (APP. A), ottenendo il vettore colonna delle derivate parziali che viene posto eguale a zero

$$\frac{\partial\mathcal{L}(\mathbf{p}, \lambda)}{\partial\mathbf{p}} = 2\,\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T\mathbf{p} - 2\,\lambda\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{p} = \mathbf{0}_3$$

$$\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T\mathbf{p} - \lambda\,\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{p} = \mathbf{0}_3$$

Premoltiplicando per $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$ ambo i membri e tenendo conto della (3.2.6) che fornisce $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\,\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}} = \mathbf{R}$, sostituendo si ottiene l'equazione agli autovalori

$$\mathbf{R}\,\mathbf{C}^T\mathbf{p} - \lambda\,\mathbf{p} = \mathbf{0}_3 \tag{8.7.5}$$

che è analoga all'equazione (4.9.10) in cui gli autovettori sono i fattori $\mathbf{f}$ delle righe. Il moltiplicatore di Lagrange, ossia l'autovalore λ, è eguale alla varianza dei 5 score, come si vede premoltiplicando la (8.7.5) per $\mathbf{p}^T\mathbf{D}_{\overline{\mathbf{c}}}$ e tenendo conto della (8.7.1) e ancora della (3.2.6)

$$\mathbf{p}^T\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{R}\,\mathbf{C}^T\mathbf{p} = \lambda\,\mathbf{p}^T\mathbf{D}_{\overline{\mathbf{c}}}\,\mathbf{p}$$

$$\mathbf{p}^T\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{C}^T\mathbf{p} = \lambda$$

$$\mathbf{s}^T\mathbf{D}_{\overline{\mathbf{r}}}\,\mathbf{s} = \lambda.$$

Questo indica che il massimo della varianza degli score che si sta cercando si ha in corrispondenza del *primo* autovalore non banale λ_1, e che perciò

gli score ottimali sono le componenti di $\mathbf{g}_1$, *primo* fattore (principale) delle colonne, ossia le ascisse dei profili delle colonne sul primo asse fattoriale. I pesi ottimali, per la (4.9.4), sono le componenti di $\mathbf{p} = \hat{\mathbf{f}}_1$, primo fattore (standard) delle righe che ha varianza unitaria.

La soluzione trovata è quindi la stessa dell'Analisi delle Corrispondenze limitatamente al primo autovalore e al corrispondente fattore. Nel caso dell'esempio i valori $\mathbf{p} = \hat{\mathbf{f}}_1$ dei pesi risultano essere

	Ripresa funzionale		
	limitata	*parziale*	*completa*
peso	−1.66	+0.20	+0.94

Come si vede, nell'ottica dell'*Optimal Scaling* i pesi da assegnare ai tre gradi della Ripresa Funzionale per differenziare maggiormente i 5 ospedali non sono affatto equispaziati e sostanzialmente discriminano tra una Ripresa Funzionale molto limitata o nulla e gli altri due gradi di Ripresa, in quanto la distanza tra i primi due pesi è di 1.86 mentre quella tra il secondo e il terzo è di 0.74, meno della metà. Le differenze con i tre pesi 1, 2 e 3 assegnati inizialmente, si possono evidenziare più chiaramente normalizzando i pesi ottimali in una scala da 1 a 3. In altri termini, posto -1.66 pari a 1 e +0.94 pari a 3, il peso intermedio $p_{2ott} = +0.20$, in base alla relazione di proporzionalità

$$p_{2norm} = 1 + (p_{2ott} - (-1.66)) \frac{(3-1)}{(0.94 - (-1.66))}$$

corrisponde al valore $p_{2norm} = +2.43$, ben lontano dal peso 2 assegnato inizialmente e molto spostato verso il valore estremo 3, come risulta dallo specchietto

	Pesi		
Ripresa	*ottimali*	*normal.*	*equisp.*
limitata	−1.66	1	1
parziale	+0.20	2.43	2
completa	+0.94	3	3

I corrispondenti score ottimali $\mathbf{s} = \mathbf{g}_1$ dei 5 ospedali risultano

	Ospedali				
	H.S	*H.A*	*H.P*	*H.M*	*H.B*
score	−0.69	−0.06	+0.19	+0.28	+0.34

Adesso non ci sono più ospedali ex-aequo: l'ospedale H.B surclassa nettamente tutti gli altri.

Mentre la varianza dei pesi è unitaria per costruzione, quella degli score ottimali è pari all'autovalore $\lambda_1 = 0.149$, quasi il doppio di quella che si poteva ottenere con i pesi iniziali equispaziati, che risultava essere di

$$\left((1.60 - 2.10)^2 + (2.06 - 2.10)^2 + \ldots + (2.35 - 2.10)^2\right)/5 = 0.078.$$

L'Otimal Scaling differenzia meglio le performance dei 5 Ospedali, perché 'sparpaglia' al massimo i loro score sull'asse, come si vede chiaramente sulla TAV. 8.7.1, che evidenzia una netta differenza di performance tra l'ospedale H.S da un lato, l'ospedale H.A e gli altri tre.

Nella Sez. 4.9 si è visto che i profili sono il baricentro dei vertici. Ad esempio, la coordinata dell'ospedale H.A per la (4.9.4) si ottiene ponderando con le 3 componenti del profilo $\mathbf{c}_1$ le coordinate $\mathbf{p}$ dei tre vertici

$$\mathbf{g}_1 = \mathbf{c}_1^T \hat{\mathbf{f}}_1 = 0.277 \times (-1.66) + .382 \times 0.20 + 0.340 \times 0.94 = -0.06.$$

Questo è proprio lo score di H.A riportato nello specchietto qui sopra che si può ottenere ponderando con i pesi ottimali p_1, p_2 e p_3 le 3 componenti del profilo $\mathbf{c}_1$ dell'ospedale H.A.

Nella matrice presa ad esempio, le modalità ordinate erano quelle delle righe. Quando invece sono quelle delle colonne, un procedimento del tutto analogo a quello sviluppato qui sopra conduce alla equazione agli autovalori

$$\mathbf{C}^T\mathbf{R}\,\mathbf{p} - \lambda\,\mathbf{p} = \mathbf{0}_J$$

ove ora il vettore $\mathbf{p}$ dei pesi è di ordine J e quello $\mathbf{s}$ degli score di ordine I. La soluzione è ancora fornita dall'Analisi delle Corrispondenze, limitatamente al primo autovalore λ_1 e ai due corrispondenti primi fattori:

$$\mathbf{p} = \hat{\mathbf{g}}_1 \qquad \text{e} \qquad \mathbf{s} = \mathbf{f}_1.$$

A causa dell'indeterminatezza del segno degli assi fattoriali, può capitare talvolta che i pesi risultino avere valori decrescenti, mentre le modalità hanno un ordine crescente intrinseco, o viceversa. In tal caso basta cambiare il segno ai pesi, ricordandosi però di fare lo stesso agli score ottenuti.

Quando le modalità ordinate sono ottenute suddividendo in classi una variabile continua, è prassi della Statistica considerare come pesi i valori centrali delle classi, sempre che la prima e/o l'ultima classe non siano aperte, cioè del tipo 'inferiore a ...' e 'oltre ...'. L'Optimal Scaling invece è in grado di assegnare comunque un peso a ogni classe, anche se aperta.

L'Optimal Scaling è una proprietà intrinseca dell'Analisi delle Corrispondenze, sia semplici che Multiple, indipendente dal fatto che le modalità siano esplicitamente ordinate o meno. Un esempio è presentato nella prossima Sez. 8.8.

Se gli score, o i pesi, vengono impiegati con altre variabili in una regressione, queste devono essere standardizzate, ad esempio con valore medio nullo e varianza unitaria. Anche gli score vanno allora rinormalizzati dividendoli per λ_1 .

Infine va tenuto sempre presente che l'Optimal Scaling è *uno* dei modi per ottenere pesi e score e questi sono ottimali solo se si cercano dei pesi standardizzati che differenzino gli score al massimo. Con un altro criterio la graduatoria potrebbe risultare diversa.

8.8 - Matrici di profili

L'Analisi delle Corrispondenze di una matrice di profili, sia delle righe, $\mathbf{R}$, che delle colonne, $\mathbf{C}$, non richiede alcuna trasformazione preliminare della matrice.

Se la matrice attiva è $\mathbf{R}$, le colonne illustrative codificate con 0 e 1 non richiedono trasformazioni prima della proiezione sui piani fattoriali, per quanto si è visto nella Sez. 8.6, mentre le righe illustrative devono essere ricodificate con 0 e $\overline{r}_j$. Inversamente, se la matrice attiva è $\mathbf{C}$, soltanto le colonne illustrative vanno ricodificate con 0 e $\overline{c}_i$.

Anche l'applicazione dell'Optimal Scaling a una matrice di profili richiede qualche attenzione. Un esempio chiarirà meglio come procedere. La matrice di TAV. 8.8.1 è tratta da una vasta indagine telefonica condotta intervistando oltre 200 responsabili finanziari delle maggiori imprese per stabilire quali banche, secondo loro, avessero la migliore immagine. I dati originali dell'indagine non sono disponibili, ma dai dati parziali resi pubblici è stato possibile costruire la matrice di profili di TAV. 8.8.1. Nelle colonne sono 5 banche, indicate con B1 - B5, e nelle righe 5 dei quesiti rivolti ai direttori finanziari, ed indicati con Q1 - Q5. L'elemento r_{ij} indica la quota di intervistati che alla domanda i hanno indicato la banca j come migliore. Poiché le frequenze assolute sono ignote, due banche possono avere profili eguali, anche se il numero n_{+j} di intervistati che le avevano indicate era notevolmente diverso. Non sarebbe corretto considerare il fattore $\mathbf{g}_1$, ottenuto dall'Analisi delle Corrispondenze della matrice 5×5 di TAV. 8.8.1 come il vettore degli score delle banche perché significherebbe dare la stessa importanza a quote

eguali r_{ij} che potrebbero derivare da numerosità diverse: una banca poco citata dagli intervistati può avere lo stesso profilo di una molto più apprezzata. Non disponendo di elementi per dare ad ogni banca la sua giusta importanza, la cosa che si può fare è di porre tutte le banche sullo stesso piano, dando ad esse la stessa massa, creando per ogni risposta una nuova *riga* che contenga $1 - r_{ij}$, ossia i complementi a 1 delle quote originali r_{ij} di intervistati che avevano indicato la banca j come migliore per la caratteristica i. Questa operazione è detta *complementazione* e le sue implicazioni saranno illustrate con maggiori dettagli nella Sez. 9.X. Il risultato è che ora ogni banca ha la medesima massa $\overline{r}_j$. La matrice complementata di ordine 10×5 è riportata nella TAV. 8.8.1 e l'Analisi delle Corrispondenze della matrice così creata produce la mappa standard mostrata nella TAV. 8.8.2, ove i profili sono quelli delle banche e i vertici sono i quesiti. Gli score sono le componenti del fattore $\mathbf{g}_1$ sul primo asse, ove le caratteristiche positive, i 5 profili originali, si oppongono nettamente a quelle negative, i 5 profili complementati, con la banca B1 fortemente associata ai primi e le banche B5, B4 e B3 a quelle negative

	Banche				
	$B5$	$B4$	$B3$	$B2$	$B1$
score	-0.33	-0.34	-0.11	$+0.24$	$+0.55$

La graduatoria ottimale vede la banca B1 al primo posto con la migliore immagine presso le imprese interpellate, almeno per quanto riguarda le 5 caratteristiche oggetto delle interviste. La varianza degli score risulta essere $\lambda_1 = 0.120$.

I pesi ottimali assegnati ai profili attivi e complementati sono le componenti del primo fattore standard $\hat{f}_{1i}$, ossia

	Quesiti									
	$Q1$	$C1$	$Q2$	$C2$	$Q3$	$C3$	$Q4$	$C4$	$Q5$	$C5$
pesi	2.04	-0.51	2.12	-0.54	1.69	-0.48	1.91	-0.48	2.09	-0.52

8.9 - Matrici concatenate

Per diffondere la cultura scientifica tra i giovani, un centro di ricerca ha organizzato una 'Giornata della Scienza' durante la quale alcune scolaresche potevano partecipare ad un Seminario di presentazione delle attività del centro, visitare una Mostra ed effettuare una Visita guidata ai laboratori di

ricerca. Al termine si doveva compilare un questionario esprimendo il grado di apprezzamento dei tre momenti della visita, in base a una scala a tre valori: basso, medio e alto. I questionari raccolti, complessivamente 165 , sono stati suddivisi per tipo di indirizzo di studio: scuola Media, Istituto professionale e Liceo.

In corrispondenza del tipo di scuola, le variabili rilevate simultaneamente sono *tre*: Seminario, Mostra e Visita e ciascuna ha 3 modalità. Dai questionari raccolti si possono costruire tre matrici di contingenza, anzi di incidenza, incrociando il Tipo di scuola (nelle colonne) con ciascuna delle tre variabili (nelle righe),ottenendo matrici di ordine 3×3 le cui numerosità differiscono leggermente perché alcuni questionari non erano completi.

Un primo tipo di analisi che si può fare è l'analisi *separata* delle tre matrici. Questa scelta sottintende l'ipotesi che le tre variabili siano in un certo senso indipendenti e che si possano ignorare le informazioni ricavabili dalle altre due matrici.

L'Analisi delle Corrispondenze è ancora in grado di rivelare le eventuali associazioni, o *interazioni*, tra le tre variabili, analizzando *globalmente* la matrice di ordine 9×3 costruita concatenando in senso verticale, ossia 'impilando' le tre matrici una sull'altra. La matrice è riportata nella FIG. 8.9.1 e la mappa risultante nella FIG. 8.9.2. Si noti innanzitutto come l'origine degli assi sia sempre contenuta nel triangolo che ha come vertici i tre gradi di apprezzamento di una variabile, per esempio: Mostra-scarso, Mostra-medio, Mostra-alto perché ne è il baricentro. Il primo asse oppone i giudizi meno favorevoli a quelli più favorevoli. Gli alunni delle medie hanno dato i giudizi più bassi, segno di scarsa comprensione e forse anche di noia, specialmente al Seminario. Gli studenti dell'Istituto professionale hanno invece apprezzato la Visita ai laboratori e la Mostra, mentrei liceali sono rimasti complessivamente soddisfatti della Giornata della Scienza e questo è un tipico esempio di associazione, o di interazione, tra le modalità di tre variabili che le analisi separate delle tre matrici non avrebbero mai messo in luce.

Invece che verticalmente, le matrici possono essere concatenate orizzontalmente. La scelta dipende dalla natura delle variabili in gioco.

Nell'analisi di matrici concatenate si ottengono di solito valori elevati dell'inerzia totale, valori che è bene accettare con cautela. Nel caso dell'esempio risulta infatti: $In_{\overline{\mathbf{r}}} = In_{\overline{\mathbf{c}}} = 0.392$.

8.10 - Bibliografia essenziale

Per l'analisi di matrici con elementi negativi il lettore può fare riferimento all'Appendice A del testo di

Peter G. M. van der Heijden (1987). *Correspondence Analysis of Longitudinal Categorical Data*. **II**. DSWO Press, Leiden. 270 pg. ISBN 90-6695-019-6.

Sulla ricostruzione dei dati mancanti in una matrice esiste un'ampia letteratura. Il primo lavoro è del compianto

F. K. Mutombo (1973). Traitement des données manquantes et rationalisation d'un réseau de stations de mesures. *Tesi di dottorato*, Università Pierre et Marie Curie, Paris.

mentre altri casi più complessi sono trattati in

Ch. Bastin e al. (1980). *Pratique de l'Analyse des Données. Abrégé teorique et études de cas modéle*. Bordas - Dunod ed., Paris. 470 pg. ISBN 2-04-011181-6.

Un algoritmo per trattare matrici con uno zero strutturale e relativo esempio è alle pag. 236 e 251 di Greenacre (1984), mentre in Greenacre (1993) l'intero Capitolo 15 è dedicato all'analisi di matrici concatenate. Entrambi i testi sono citati nella bibliografia del Capitolo 4 alla Sez. 4.17.

La proiezione sulle mappe fattoriali di profili illustrativi, detti talvolta supplementari, è trattata estesamente nell'articolo di

Pierre Cazes (1982). Note sur les éléments supplémentaires en Analyse des Correspondances. **II**. Tableau multiples. *Cahiers de l'Analyse des Données*, Vol. **VII**, n^o 2, pag. 133 - 154. Questo è il secondo articolo. Il primo è citato nella bibliogafia del Capitolo 4, nella Sez. 4.17.

TAVOLA 1 - Matrice di contingenza: formalismo.

La tavola riassume la simbologia impiegata nel testo per indicare una matrice di contingenza, i suoi elementi e altre grandezze relate:

$$\begin{aligned}
\mathbf{N} &= \text{indica una matrice di contingenza,} \\
i &= \text{indice della } i^{ma} \text{ riga} \\
I &= \text{numero di righe} \\
j &= \text{indice della } j^{ma} \text{ colonna} \\
J &= \text{numero di colonne} \\
n_{ij} &= \text{elemento generico di } \mathbf{N} \\
n_{+j} &= \text{elemento } j^{mo} \text{ della riga marginale} \\
n_{i+} &= \text{elemento } i^{mo} \text{ della colonna marginale} \\
n_{++} &= \text{totale generale di } \mathbf{N}
\end{aligned}$$

Gli elementi di $\mathbf{N}$, sono numeri interi *non negativi*. Gli elementi dei totali marginali ed il totale generale sono numeri interi *positivi*.

TAVOLA 2 - La matrice *Spettacoli*.

Fin dal 1936 la *Societa Italiana degli Autori ed Editori (SIAE)* pubblica annualmente un annuario contenente accurate statistiche relative a tutte le manifestazioni dello spettacolo e del trattenimento che hanno luogo in Italia. La matrice di contingenza riportata in questa tavola è stata costruita in base ai dati relativi alle attività teatrali e musicali, pubblicati nell'annuario del 1991 .

La matrice, evidenziata dalla cornice, indica come gli oltre 26 milioni di biglietti venduti in quell'anno sono distribuiti tra le 20 regioni del territorio nazionale e tra gli 8 principali tipi di spettacolo teatrale. Così all'incrocio della prima riga con la prima colonna si legge che nel 1991 sono stati venduti in Piemonte 639 074 biglietti per rappresentazioni di Prosa. Se, come è ragionevole supporre, ad ogni biglietto venduto corrisponde uno spettatore e quelli non paganti sono in numero trascurabile, si può anche affermare che in Piemonte 639 074 spettatori hanno assistito a rappresentazioni di Prosa.

Con l'intestazione TOTALE sono indicate la *riga marginale*, ossia la distribuzione degli spettatori negli 8 tipi di spettacolo prescindendo dalla regione, e la *colonna marginale*, la distribuzione degli spettatori nelle 20 regioni prescindendo dal tipo di spettacolo. Il *totale generale* è di 26 196 957 spettatori.

TAVOLA 2 - La matrice *spettacoli* 3

	Tipi di Spettacolo								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	
Regioni	*Prosa*	*Lirica e Balletti*	*Concerti di Musica Classica*	*Operetta*	*Rivista e Commedia Musicale*	*Musica Leggera e Folkloristica*	*Burattini e Marionette*	*Saggi Coreografici e Folkloristici*	*TOTALI*
Piemonte	639074	206759	307635	34806	92839	580477	21203	77698	1960491
e d'Aosta	921	224	3989	854	188	21190	1029	5588	42281
ombardia	1996069	408764	708246	70247	212141	1152501	51039	72965	4671972
lto Adige	298752	30273	115274	2291	3876	213294	12588	4733	681081
Veneto	788117	686236	240273	17649	75895	512118	6521	38573	2365382
zia Giulia	384141	105484	124314	24595	14042	217738	5797	23637	899748
Liguria	444970	102238	118088	11338	61791	239161	7409	13156	998151
Romagna	1195885	211587	366816	46856	89843	765922	37509	44893	2759311
Toscana	818062	238634	251210	23015	71817	608069	27398	27994	2066199
Umbria	167897	32864	58427	3699	6441	71815	5800	7497	354440
Marche	193816	54684	76818	10925	13082	177828	3855	13984	544992
Lazio	1508506	303187	494357	16865	203693	590141	68975	22425	3208149
Abruzzi	191543	23727	113664	4495	18981	98500	11406	12885	475201
Molise	25100	4252	7928	234	1943	10561	1567	511	52096
Campania	662165	117006	136160	4271	68216	388389	14264	26225	1416696
Puglia	293103	65237	129849	16894	16964	175109	6236	19240	722632
Basilica	60807	3226	9470	0	1899	29329	1288	1931	107950
Calabria	151822	30827	54387	2623	3884	103078	153	8492	355266
Sicilia	978485	200904	334022	37640	49324	187037	17956	17203	1822571
Sardegna	174260	41219	110849	5479	17362	314951	10316	17912	692348
TOTALI	10981793	2867332	3761776	334776	1024221	6457208	312309	457542	26196957

4 TAVOLA 3 - La matrice R dei profili delle righe

TAVOLA 3 - La matrice R dei profili delle righe.

Se nella matrice *Spettacoli* di TAV. 2 si dividono gli 8 elementi della prima riga (Piemonte) per il totale della prima riga, si ottiene

$\frac{639\,074}{1\,960\,491}$	$\frac{206\,759}{1\,960\,491}$	$\frac{307\,635}{1\,960\,491}$	$\frac{34\,068}{1\,960\,491}$	$\frac{92\,839}{1\,960\,491}$	$\frac{580\,477}{1\,960\,491}$	$\frac{21\,203}{1\,960\,491}$	$\frac{77\,698}{1\,960\,491}$
0.326	0.105	0.157	0.018	0.047	0.296	0.011	0.040
r_{11}	r_{12}	r_{13}	r_{14}	r_{15}	r_{16}	r_{17}	r_{18}

Questi 8 quozienti rappresentano la quota di spettatori agli 8 tipi di spettacolo nel Piemonte e costituiscono il *pro lo* della prima riga della matrice *Spettacoli*. I valori sono arrotondati alla terza cifra decimale e la loro somma è 1.000 . Ripetendo il procedimento per le altre 19 righe si ottiene la matrice riportata in questa Tavola.

La matrice dei profili delle righe, denotata col simbolo $\mathbf{R}$, è utile per confrontare la "produzione di spettatori" delle regioni. Se i profili di due regioni risultano eguali (o quasi), le due corrispondenti righe della matrice di contingenza *Spettacoli* di TAV. 2 sono proporzionali (o quasi).

TAVOLA 3 - La matrice R di profili delle righe 5

	Tipi di Spettacolo								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	
Regioni	*Prosa*	*Lirica e Balletti*	*Concerti di Musica Classica*	*Operetta*	*Rivista e Commedia Musicale*	*Musica Leggera e Folkloristica*	*Burattini e Marionette*	*Saggi Coreografici e Folkloristici*	*TOTALI*
Piemonte	0.326	0.105	0.157	0.018	0.047	0.296	0.011	0.040	1.000
e d'Aosta	0.218	0.005	0.094	0.020	0.004	0.501	0.024	0.132	1.000
ombardia	0.427	0.087	0.152	0.015	0.045	0.247	0.011	0.016	1.000
lto Adige	0.439	0.044	0.169	0.003	0.006	0.313	0.018	0.007	1.000
Veneto	0.333	0.290	0.102	0.007	0.032	0.217	0.003	0.016	1.000
ia Giulia	0.427	0.117	0.138	0.027	0.016	0.242	0.006	0.026	1.000
Liguria	0.446	0.102	0.118	0.011	0.062	0.240	0.007	0.013	1.000
Romagna	0.433	0.077	0.133	0.017	0.033	0.278	0.014	0.016	1.000
Toscana	0.396	0.115	0.122	0.011	0.035	0.294	0.013	0.014	1.000
Umbria	0.474	0.093	0.165	0.010	0.018	0.203	0.016	0.021	1.000
Marche	0.356	0.100	0.141	0.020	0.024	0.326	0.007	0.026	1.000
Lazio	0.470	0.095	0.154	0.005	0.063	0.184	0.022	0.007	1.000
Abruzzi	0.403	0.050	0.239	0.009	0.040	0.207	0.024	0.027	1.000
Molise	0.482	0.082	0.152	0.004	0.037	0.203	0.030	0.010	1.000
Campania	0.467	0.083	0.096	0.003	0.048	0.274	0.010	0.019	1.000
Puglia	0.406	0.090	0.180	0.023	0.023	0.242	0.009	0.027	1.000
Basilica	0.563	0.030	0.088	0.000	0.018	0.272	0.012	0.018	1.000
Calabria	0.427	0.087	0.153	0.007	0.011	0.290	0.000	0.024	1.000
Sicilia	0.537	0.110	0.183	0.021	0.027	0.103	0.010	0.009	1.000
Sardegna	0.252	0.060	0.160	0.008	0.025	0.455	0.015	0.026	1.000

6 TAVOLA 4 - Matrice R, profilo riga medio e masse

TAVOLA 4 - Matrice R, profilo riga medio e masse.

Alla matrice dei profili delle righe di TAV. 3, sono affiancati i profili della riga e della colonna marginale della matrice *Spettacoli*, ottenuti dividendo i totali per il totale generale. Così la riga marginale risulta

$\frac{10\,981\,793}{26\,196\,957}$	$\frac{2\,867\,332}{26\,196\,957}$	$\frac{3\,761\,776}{26\,196\,957}$	$\frac{334\,776}{26\,196\,957}$	$\frac{1\,024\,221}{26\,196\,957}$	$\frac{6\,457\,208}{26\,196\,957}$	$\frac{312\,309}{26\,196\,957}$	$\frac{457\,542}{26\,196\,957}$
0.419	0.109	0.144	0.013	0.039	0.246	0.012	0.017
$\frac{n_{+1}}{n_{++}}$	$\frac{n_{+2}}{n_{++}}$	$\frac{n_{+3}}{n_{++}}$	$\frac{n_{+4}}{n_{++}}$	$\frac{n_{+5}}{n_{++}}$	$\frac{n_{+6}}{n_{++}}$	$\frac{n_{+7}}{n_{++}}$	$\frac{n_{+8}}{n_{++}}$

Con procedimento analogo si ottiene il profilo della colonna marginale.

Il profilo della riga marginale si può ottenere anche come media ponderata dei profili riga. I pesi, detti *masse* dei profili delle righe, sono le componenti del profilo della colonna marginale. Così, per la prima componente,

$$\frac{n_{+1}}{n_{++}} = 0.326 \times 0.075 + 0.218 \times 0.002 + 0.427 \times 0.178 + \ldots + 0.252 \times 0.026$$
$$= 0.419.$$

I profili marginali si scambiano di ruolo (TAV. 5) per la matrice **C** dei profili colonna.

TAVOLA 4 - Matrice R, profilo riga medio e masse 7

	Tipi di Spettacolo								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	
Regioni	*Prosa*	*Lirica e Balletti*	*Concerti di Musica Classica*	*Operetta*	*Rivista e Commedia Musicale*	*Musica Leggera e Folkloristica*	*Burattini e Marionette*	*Saggi Coreografici e Folkloristici*	*MASSE*
Piemonte	0.326	0.105	0.157	0.018	0.047	0.296	0.011	0.040	0.075
a d'Aosta	0.218	0.005	0.094	0.020	0.004	0.501	0.024	0.132	0.002
ombardia	0.427	0.087	0.152	0.015	0.045	0.247	0.011	0.016	0.178
lto Adige	0.439	0.044	0.169	0.003	0.006	0.313	0.018	0.007	0.026
Veneto	0.333	0.290	0.102	0.007	0.032	0.217	0.003	0.016	0.090
ia Giulia	0.427	0.117	0.138	0.027	0.016	0.242	0.006	0.026	0.034
Liguria	0.446	0.102	0.118	0.011	0.062	0.240	0.007	0.013	0.038
Romagna	0.433	0.077	0.133	0.017	0.033	0.278	0.014	0.016	0.105
Toscana	0.396	0.115	0.122	0.011	0.035	0.294	0.013	0.014	0.079
Umbria	0.474	0.093	0.165	0.010	0.018	0.203	0.016	0.021	0.014
Marche	0.356	0.100	0.141	0.020	0.024	0.326	0.007	0.026	0.021
Lazio	0.470	0.095	0.154	0.005	0.063	0.184	0.022	0.007	0.122
Abruzzi	0.403	0.050	0.239	0.009	0.040	0.207	0.024	0.027	0.018
Molise	0.482	0.082	0.152	0.004	0.037	0.203	0.030	0.010	0.002
ampania	0.467	0.083	0.096	0.003	0.048	0.274	0.010	0.019	0.054
Puglia	0.406	0.090	0.180	0.023	0.023	0.242	0.009	0.027	0.028
Basilica	0.563	0.030	0.088	0.000	0.018	0.272	0.012	0.018	0.004
Calabria	0.427	0.087	0.153	0.007	0.011	0.290	0.000	0.024	0.014
Sicilia	0.537	0.110	0.183	0.021	0.027	0.103	0.010	0.009	0.070
Sardegna	0.252	0.060	0.160	0.008	0.025	0.455	0.015	0.026	0.026
MEDIA	0.419	0.109	0.144	0.013	0.039	0.246	0.012	0.017	

8 TAVOLA 5 - Matrice C, profilo colonna medio e masse

TAVOLA 5 - Matrice C, profilo colonna medio e masse.

Se nella matrice *Spettacoli* di TAV. 2 si dividono i 20 elementi della prima colonna (*Prosa*) per il totale della prima colonna, si ottiene il *pro lo* della prima colonna

$$\begin{array}{ccccc} \frac{n_{11}}{n_{+1}} & = & \frac{639\,074}{10\,981\,793} & = & 0.058 \\ \frac{n_{21}}{n_{+1}} & = & \frac{9\,219}{10\,981\,793} & = & 0.001 \\ \frac{n_{31}}{n_{+1}} & = & \frac{1\,996\,069}{10\,981\,793} & = & 0.182 \\ \ldots & = & \ldots & = & \ldots \\ \frac{n_{20\,1}}{n_{+1}} & = & \frac{174\,260}{10\,981\,793} & = & 0.016 \end{array}$$

I 20 quozienti, qui arrotondati alla terza cifra decimale, rappresentano la quota di spettatori di ciascuna regione alle rappresentazioni di Prosa. La loro somma vale 1.000 . Ripetendo il procedimento per le altre 7 colonne si ottiene la matrice 20×8 riportata in questa Tavola.

La matrice dei profili delle colonne, indicata con $\mathbf{C}$, è utile per confrontare il favore che i vari tipi di spettacolo teatrale incontrano nelle regioni.

Affiancano la matrice, il profilo colonna *medio*, media ponderata dei profili, e la riga dei pesi, o *masse* dei profili colonna. Rispetto alla TAV. 4 i ruoli di questi due profili sono scambiati.

TAVOLA 4 - Matrice R, profilo riga medio e masse 9

	Tipi di Spettacolo								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	
Regioni	*Prosa*	*Lirica e Balletti*	*Concerti di Musica Classica*	*Operetta*	*Rivista e Commedia Musicale*	*Musica Leggera e Folkloristica*	*Burattini e Marionette*	*Saggi Coreografici e Folkloristici*	*MASSE*
Piemonte	0.058	0.072	0.082	0.104	0.091	0.090	0.068	0.170	0.075
e d'Aosta	0.001	0.000	0.001	0.003	0.000	0.003	0.003	0.012	0.002
ombardia	0.182	0.143	0.188	0.210	0.207	0.178	0.163	0.159	0.178
lto Adige	0.027	0.011	0.031	0.007	0.004	0.033	0.040	0.010	0.026
Veneto	0.072	0.239	0.064	0.053	0.074	0.079	0.021	0.084	0.090
zia Giulia	0.035	0.037	0.033	0.073	0.014	0.034	0.019	0.052	0.034
Liguria	0.041	0.036	0.031	0.034	0.060	0.037	0.024	0.029	0.038
Romagna	0.109	0.074	0.098	0.140	0.088	0.119	0.120	0.098	0.105
Toscana	0.074	0.083	0.067	0.069	0.070	0.094	0.088	0.061	0.079
Umbria	0.015	0.011	0.016	0.011	0.006	0.011	0.019	0.016	0.014
Marche	0.018	0.019	0.020	0.033	0.013	0.028	0.012	0.031	0.021
Lazio	0.137	0.106	0.131	0.050	0.199	0.091	0.221	0.049	0.122
Abruzzi	0.017	0.008	0.030	0.013	0.019	0.015	0.037	0.028	0.018
Molise	0.002	0.001	0.002	0.001	0.002	0.002	0.005	0.001	0.002
Campania	0.060	0.041	0.036	0.013	0.067	0.060	0.046	0.057	0.054
Puglia	0.027	0.023	0.035	0.050	0.017	0.027	0.020	0.042	0.028
Basilica	0.006	0.001	0.003	0.000	0.002	0.005	0.004	0.004	0.004
Calabria	0.014	0.011	0.014	0.008	0.004	0.016	0.000	0.019	0.014
Sicilia	0.089	0.070	0.089	0.112	0.048	0.029	0.057	0.038	0.070
Sardegna	0.016	0.014	0.029	0.016	0.017	0.049	0.033	0.039	0.026
MASSE	0.419	0.109	0.144	0.013	0.039	0.246	0.012	0.017	

10 TAVOLA 6 - Diagrammi dei profili delle righe

TAVOLA 6 - Diagrammi dei profili delle righe.

Ciascuna riga della matrice $\mathbf{R}$ di TAV. 4 è rappresentata con un diagramma a 8 barre rettangolari: una barra per ogni tipo di spettacolo. L'altezza di ciascuna barra j per una regione i è proporzionale alla frequenza relativa $r_{ij} = n_{ij}/n_{i+}$. La larghezza delle barre, eguale per tutte nel grafico, non è rilevante. Per ragioni di spazio il profilo riga medio non è rappresentato.

TAVOLA 7 - Diagrammi dei profili delle colonne.

I profili degli 8 tipi di spettacolo sono qui rappresentati con 8 diagrammi a barre affiancati. Le altezze delle barre sono proporzionali alle frequenze relative $c_{ij} = n_{ij}/n_{+j}$ e le barre hanno tutte la stessa ampiezza. Anche in questa Tavola, per motivi di spazio, è stato omesso il diagramma del profilo colonna medio.

La rappresentazione grafica fornisce una visione d'insieme dei profili, ma il confronto tra diagrammi risulta piuttosto complesso quando si tratta di matrici di non trascurabile dimensione, come in questo caso. La realizzazione di grafici di questo tipo richiede un software specifico.

TAVOLA 8 - Matrice S degli scarti relativi dalla media.

Questa matrice è particolarmente utile perché consente di individuare rapidamente quegli elementi che, per eccesso o per difetto, si scostano dalla media. Può essere costruita in tre modi equivalenti

- dai profili delle righe (TAV. 3): rapportando ogni scarto tra la componente del profilo e quella del profilo riga medio, alla componente del profilo riga medio;
- dai profili delle colonne (TAV. 5): rapportando ogni scarto tra la componente del profilo e quella del profilo colonna medio, alla componente del profilo colonna medio;
- dalla matrice di contingenza (TAV. 2): rapportando lo scarto tra l'elemento della matrice ed il valore ottenuto in una situazione di omogeneità, a quest'ultimo valore.

Così lo scarto relativo tra Piemonte e Prosa, primo elemento della matrice, risulta

$$s_{11} = \frac{0.326 \quad 0.419}{0.419} = \quad 0.2 \quad \text{oppure} \quad s_{11} = \frac{0.058 \quad 0.075}{0.075} = \quad 0.2 \quad \text{oppure}$$

$$s_{11} = \frac{639\,074 \quad 821\,840}{821\,840} = \quad 0.2 \quad \text{perché} \quad \frac{10\,981\,793 \times 1\,960\,491}{26\,196\,957} = 821\,840.$$

TAVOLA 9 - Punti geometrici e vettori.

A e B sono due punti non coincidenti di un piano, dotato di un sistema di riferimento costituito da due assi ortogonali di coordinate. Da A si traccia la parallela al primo asse e da B la parallela al secondo. La loro intersezione avviene in Q. La *posizione* del punto B, relativamente a quella di A, può essere indicata dalla coppia ordinata di numeri a e b che indicano la lunghezza e la direzione dei due segmenti AQ e QB. I valori assoluti dei due numeri, ossia $|a|$ e $|b|$, sono le lunghezze dei segmenti. In questo esempio è $a = 3$ e $b = 2$ unità. I segni sono positivi o negativi a seconda che le direzioni da A a Q e da Q a B siano uguali o contrarie a quelle degli assi coordinati.

Si può così associare alla coppia ordinata di punti A e B un vettore colonna che si indica

$$A\overrightarrow{B} = \begin{pmatrix} a \\ b \end{pmatrix}.$$

Perciò, quando sono assegnati due punti A e B, esiste un *unico* vettore che indica la posizione del punto B rispetto a quella di A. Tutte queste considerazioni possono essere estese a spazi multidimensionali.

TAVOLA 10 - Punti geometrici e vettori. *(seguito)*

Da questa tavola appare che ad un assegnato vettore, ad esempio $\mathbf{v} = (2\ 3)^T$, possono essere associate molte coppie di punti diversi, perché si ha

$$A_1\,\overrightarrow{B_1} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \qquad A_2\,\overrightarrow{B_2} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \qquad A_3\,\overrightarrow{B_3} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \qquad O\,\overrightarrow{X} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

Questo è dovuto al fatto che i segmenti $A_1\,B_1$, $A_2\,B_2$, $A_3\,B_3$, $O\,X$ sono paralleli, hanno la stessa direzione (e non direzioni contrarie) e la stessa lunghezza. Al concetto matematico di vettore vengono così associati i concetti geometrici di direzione e di lunghezza. Di conseguenza,

1 - dato un vettore non nullo $\mathbf{v}$, esistono infinite coppie di punti A e B, tali che $A\overrightarrow{B} = \mathbf{v}$;

2 - dato un vettore non nullo $\mathbf{v}$ ed un punto A, esiste un solo punto B tale che sia $A\overrightarrow{B} = \mathbf{v}$.

In questo libro, si assume che l'origine dei vettori sia l'origine degli assi coordinati, salvo esplicita menzione di diversa origine, per cui il vettore $\mathbf{v} = (2\ 3)^T$ identifica *univocamente* il punto X.

TAVOLA 11 - Rappresentazione di operazioni con vettori.

La somma di vettori e la moltiplicazione di un vettore per un numero hanno una rappresentazione geometrica. La regola del *parallelogramma delle forze*, illustrata in figura, permette di individuare un punto che corrisponde alla somma di due vettori. Ad esempio, se $O\,A\,C\,B$ è un parallelogramma, allora se

$$O\overrightarrow{A} = \begin{pmatrix} 7 \\ 2 \end{pmatrix} \qquad O\overrightarrow{B} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} \qquad \text{è} \qquad O\overrightarrow{C} = \begin{pmatrix} 7+2 \\ 2+4 \end{pmatrix} = \begin{pmatrix} 9 \\ 6 \end{pmatrix}.$$

La medesima regola permette di dare una rappresentazione geometrica alla sottrazione di vettori, in quanto nel parallelogramma $O\,A\,B\,D$ risulta $O\overrightarrow{D} = O\overrightarrow{B} \quad O\overrightarrow{A} = (\ \ 5\ 2)^T$.

La moltiplicazione di un vettore non nullo, per esempio $O\overrightarrow{X} = (4\ \ 1)^T$, per un numero positivo $a = 2$ individua un punto Y tale che $O\overrightarrow{X}$ e $O\overrightarrow{Y}$ hanno la stessa direzione e la lunghezza di $O\overrightarrow{Y}$ è $a = 2$ volte la lunghezza di $O\overrightarrow{X}$, per cui

$$O\overrightarrow{Y} = 2\ O\overrightarrow{X} = 2 \begin{pmatrix} 4 \\ 1 \end{pmatrix} = \begin{pmatrix} 8 \\ 2 \end{pmatrix}.$$

Se il moltiplicatore è negativo, il vettore cambia di lunghezza ed inverte la direzione; se è nullo anche il vettore risultante è nullo.

TAVOLA 12 - Base canonica e vettori.

I tre vettori ortogonali ed unitari

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

costituiscono la *base canonica* dello spazio euclideo $\Re^3$. Riferito a questi vettori, un qualunque altro vettore, come $\mathbf{x} = (4\ 2\ 4)^T$, può esprimersi come somma di loro multipli

$$\mathbf{x} = 4\,\mathbf{e}_1 + 2\,\mathbf{e}_2 + 4\,\mathbf{e}_3.$$

Geometricamente, componendo con la regola del parallelogramma delle forze i tre vettori, multipli dei vettori di base, si individua il punto X.

Gli *elementi*, o *componenti*, del vettore coincidono con le *coordinate* del punto. Così, invece che con X, il punto dell'esempio può essere indicato direttamente col simbolo $\mathbf{x}$ del vettore.

TAVOLA 13 - Prodotto scalare di vettori.

Il *prodotto scalare* di due vettori dello stesso ordine, per es. $\mathbf{x} = (1.2\ 0.4)^T$ e $\mathbf{y} = (0.6\ 1.2)^T$ è un *numero* eguale alla somma dei prodotti delle loro componenti. Geometricamente, viene preso in considerazione l'angolo θ tra gli estremi dei due vettori, e quindi $0 \leq \theta \leq 2\,\pi$, per cui si può scrivere

$$\mathbf{x}^T\,\mathbf{y} = (1.2\ 0.4)\begin{pmatrix} 0.6 \\ 1.2 \end{pmatrix} = 1.2 \times 0.6 + 0.4 \times 1.2 = 1.2$$
$$= \overline{O\,X}\ \overline{O\,Y}\ \cos\theta = \sqrt{1.2^2 + 0.4^2} \times \sqrt{0.6^2 + 1.2^2}\ \cos\theta.$$

Il punto P, ottenuto proiettando ortogonalmente $\mathbf{y}$ su $\mathbf{x}$ è individuato dal vettore $\overrightarrow{O\,P}$, detto *vettore proiezione* di $\mathbf{y}$ su $\mathbf{x}$. La lunghezza di $\overline{O\,P}$ è eguale a $\overline{O\,Y}\ \cos\theta$. Perciò il prodotto scalare vale $\overline{O\,X}$ volte la lunghezza della proiezione di $\mathbf{y}$ su $\mathbf{x}$. O anche $\overline{O\,Y}$ volte la lunghezza $\overline{O\,Q} = \overline{O\,X}\ \cos\theta$ della proiezione di $\mathbf{x}$ su $\mathbf{y}$.

Il prodotto scalare è di notevole utilità, perché se $\mathbf{x}$ e $\mathbf{y}$ non sono nulli, permette di

1 - stabilire se i due vettori sono *ortogonali* o meno quando $\mathbf{x}^T\mathbf{y}$ è $= 0$ o $\neq 0$;
2 - calcolarne la *lunghezza* $\sqrt{\mathbf{x}^T\,\mathbf{x}} = \sqrt{1.6} = 1.26$ e $\sqrt{\mathbf{y}^T\,\mathbf{y}} = \sqrt{1.8} = 1.34$;
3 - determinare la *distanza* $d(\mathbf{x},\mathbf{y}) = \sqrt{(1.2 \quad 0.6)^2 + (0.4 \quad 1.2)^2} = 1$;
4 - ricavare l'*angolo* tra i vettori dall'espressione $\cos\theta = \mathbf{x}^T\,\mathbf{y}\,/(\mathbf{x}^T\mathbf{x}\,\mathbf{y}^T\mathbf{y})$ $= 1.2/(1.26 \times 1.34) = 0.711$, da cui $\theta \approx 45°$.

TAVOLA 14 - La matrice d'esempio *Spettacoli-3.*

La matrice di contingenza *Spettacoli-3* di ordine 3×8 (*in alto*) è ricavata dalla matrice *Spettacoli* di ordine 20×8 di TAV. 2, accorpando le righe corrispondenti alle regioni del Nord, del Centro e del Sud, dividendo quindi i totali per $10\,000$ ed arrotondando ogni elemento all'intero più prossimo. Il procedimento è illustrato nella Sez. 2.1.

Da questa sono desunte la matrice dei profili delle righe (*al centro*) e dei profili delle colonne (*in basso*). In questo libro, le tre matrici vengono indicate rispettivamente con $\mathbf{N}$, $\mathbf{R}$ e $\mathbf{C}$, ed impiegate per illustrare la metodologia dell'Analisi delle Corrispondenze. Si noti come il profilo medio e il profilo delle masse si scambino di ruolo e come a causa degli arrotondamenti la terza riga di $\mathbf{R}$ e la seconda e la settima colonna di $\mathbf{C}$ abbiano somma 1.001 invece che 1.000.

spettacoli 3

	Tipi di Spettacolo								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	
oni	*Prosa*	*Lirica e Balletti*	*Concerti di Musica Classica*	*Operetta*	*Rivista e Commedia Musicale*	*Musica Leggera e Folkloristica*	*Burattini e Marionette*	*Saggi Coreografici e Folkloristici*	*TOTALI*
ord	576	175	198	21	55	370	14	28	1437
tro	269	63	88	5	30	145	11	7	618
Sud	254	49	90	7	18	131	6	10	565
LE	10998	2878	3768	338	1038	6468	318	458	26208

spettacoli 3, profili righe

	Tipi di Spettacolo							
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	
Prosa	*Lirica e Balletti*	*Concerti di Musica Classica*	*Operetta*	*Rivista e Commedia Musicale*	*Musica Leggera e Folkloristica*	*Burattini e Marionette*	*Saggi Coreografici e Folkloristici*	*MASSE DELLE RIGHE*
0.401	0.122	0.138	0.015	0.038	0.257	0.010	0.019	0.548
0.435	0.102	0.142	0.008	0.049	0.235	0.018	0.011	0.236
0.450	0.087	0.159	0.012	0.032	0.232	0.011	0.018	0.216
0.419	0.109	0.143	0.012	0.039	0.246	0.012	0.017	

spettacoli 3, profili colonne

	Tipi di Spettacolo								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	
i	*Prosa*	*Lirica e Balletti*	*Concerti di Musica Classica*	*Operetta*	*Rivista e Commedia Musicale*	*Musica Leggera e Folkloristica*	*Burattini e Marionette*	*Saggi Coreografici e Folkloristici*	*COLONNA MEDIA*
l	0.524	0.610	0.527	0.636	0.534	0.573	0.452	0.622	0.548
o	0.245	0.220	0.234	0.152	0.291	0.224	0.355	0.156	0.236
l	0.231	0.171	0.239	0.212	0.175	0.203	0.194	0.222	0.216
)	0.419	0.109	0.143	0.012	0.039	0.246	0.012	0.017	

TAVOLA 15 - Simplesso dei profili.

I profili sono vettori particolari con componenti che non sono mai negative ed hanno per somma 1. Queste due peculiarità fanno sì che i punti da essi individuati siano contenuti in un sottospazio, detto *iperpiano*, con una dimensione in meno del loro spazio ambiente $\Re^I$, anzi in una regione limitata di questo, detta *simplesso* ad I vertici.

Così profili di ordine 2, come $\mathbf{p}_1 = (.4\ .6)^T$ e $\mathbf{p}_2 = (.8\ .2)^T$, giacciono su una retta (iperpiano di $\Re^2$) e precisamente su un segmento (simplesso a 2 vertici) di questa, che congiunge $\mathbf{e}_1 = (1\ 0)^T$ con $\mathbf{e}_2 = (0\ 1)^T$.

Profili di ordine 3, come $\mathbf{c}_1 = (0.524\ 0.245\ 0.231)^T$ di TAV. 14 che indica le quote di spettatori a rappresentazioni di Prosa al Nord, al Centro ed al Sud, giacciono su un piano (iperpiano di $\Re^3$) ed il simplesso a 3 vertici che li contiene è il triangolo equilatero che unisce i punti unitari dello spazio dei profili.

TAVOLA 16 - Diagramma ternario.

In un diagramma ternario gli assi coordinati sono i tre lati di un triangolo equilatero. Ogni lato è considerato di lunghezza 1 e graduato di conseguenza. Per posizionare un profilo di ordine 3, come ad esempio $\mathbf{c}_1 = (0.524\ 0.245\ 0.231)^T$ delle quote di spettatori a rappresentazioni di Prosa, si inizia riportando le tre componenti del profilo sui tre lati e tracciando quindi da questi tre punti la parallela all'asse che precede. Il senso di precedenza è quello orario. Il punto d'incontro delle tre parallele individua la posizione del profilo. In effetti di parallele basta tracciarne solo due: la terza, come si vede, risulta superflua.

I tre vertici indicano la massima polarizzazione delle affluenze di spettatori: $\mathbf{e}_1 = (1\ 0\ 0)^T$ ossia spettatori solo al Nord, $\mathbf{e}_2$ solo al Centro ed $\mathbf{e}_3$ solo al Sud. Il punto $\mathbf{c}_1$ risulta fortemente spostato verso il vertice che individua il Nord in quanto la quota di spettatori in quest'area è più che doppia delle altre.

TAVOLA 17 - Rappresentazione della distanza distribuzionale.

La nuvola degli 8 profili colonna dell'esempio è raffigurata (*in primo piano*) sul simplesso: un triangolo equilatero con vertici i profili $\mathbf{e}_1$, $\mathbf{e}_2$ ed $\mathbf{e}_3$. La distanza tra profili è la distanza distribuzionale di Sez. 2.8, per cui l'unità di misura è diversa sui tre assi di riferimento.

Dividendo ciascun elemento dei profili per $1/\sqrt{\overline{c}_i}$, ossia per il corrispondente elemento del profilo medio $\overline{\mathbf{c}}$, l'unità di misura sui tre assi diventa la stessa, come nello spazio che ci è familiare, ma il simplesso (*in secondo piano*) non è più equilatero. Le ascisse degli 8 profili sui tre assi sono ora le coordinate trasformate. La distanza distribuzionale tra gli elementi di due profili corrisponde così alla distanza euclidea canonica tra le corrispondenti coordinate trasformate. Sul simplesso trasformato, la distanza distribuzionale tra punti-profilo viene computata come una distanza euclidea canonica.

Gli elementi del profilo medio $\overline{\mathbf{c}}$ sono tutti positivi. Di conseguenza, le coordinate trasformate risultano sempre dilatate.

TAVOLA 18 - Visibilità e distorsione.

Tre ipotetiche fotografie di un oggetto tridimensionale - una casetta - ripresa secondo tre diverse inquadrature: di fronte (*in alto a sinistra*), di lato (*in alto a destra*) e di spigolo (*in basso*). Sono sostanzialmente tre diverse proiezioni di un oggetto tridimensionale su un piano. Le prime due possono fornire un numero limitato di dettagli, ma in compenso questi sono riportati senza alcuna distorsione. La terza, al contrario, presenta un maggior numero di dettagli, frontali e laterali, al prezzo però di una leggera distorsione. È chiaro comunque, che dovendo mostrare a un potenziale acquirente *una* immagine della casetta, verrà scelta l'ultima foto.

L'ultima foto è poi quella ove l'immagine proiettata occupa la maggior superficie, come appare dal confronto con l'area (*a tratteggio*) che si ha nella seconda foto. Si vede quindi che quanto più l'immagine proiettata di un oggetto risulta ampia e dispersa, tanto più numerosi sono i dettagli che si possono cogliere. Questo fatto è tenuto ben presente dai fotografi che devono riprodurre oggetti per cataloghi, brochures e depliants.

TAVOLA 19 - Teorema di Huyghens sull'inerzia.

Due profili $\mathbf{c}_1 = (0.2\ 0.8)^T$ e $\mathbf{c}_2 = (0.6\ 0.4)^T$ individuano due punti dotati, rispettivamente, delle masse $\overline{r}_1 = 0.6$ e $\overline{r}_2 = 0.4$. Il baricentro di questa nuvola di punti risulta avere le coordinate $\overline{c}_1 = 0.6 \times 0.2 + 0.4 \times 0.6 = 0.36$ e $\overline{c}_2 = 0.6 \times 0.8 + 0.4 \times 0.4 = 0.64$, per cui $\overline{\mathbf{c}} = (0.36\ 0.64)^T$. L'inerzia della nuvola di punti rispetto all'origine $\mathbf{0}_2$, definita nella Sez. 3.2, risulta

$$In_0 = 0.6 \times \left(\frac{0.2^2}{0.36} + \frac{0.8^2}{0.64} \right) + 0.4 \times \left(\frac{0.6^2}{0.36} + \frac{0.4^2}{0.64} \right) = 1.167.$$

L'inerzia rispetto al baricentro, per quanto visto nella Sez. 3.9, è invece

$$In_{\overline{\mathbf{c}}} = 0.6 \times \left[\frac{(0.2 \quad 0.36)^2}{0.36} + \frac{(0.8 \quad 0.64)^2}{0.64} \right] + 0.4 \times \left[\frac{(0.6 \quad 0.36)^2}{0.36} + \frac{(0.4 \quad 0.64)^2)}{0.64} \right] = 0.167.$$

Perciò vale 1 la differenza tra l'inerzia rispetto all'origine e l'inerzia rispetto al baricentro. Il teorema di Huyghens afferma che questa differenza non è altro che l'inerzia rispetto al punto dove si è calcolata l'inerzia, in questo caso l'origine, di tutta la massa della nuvola concentrata nel baricentro. Infatti, la massa complessiva della nuvola è $\overline{r}_1 + \overline{r}_2 = 0.36 + 0.64 = 1$ e la distanza distribuzionale del baricentro dall'origine vale $d_D^2(\overline{\mathbf{c}}, \mathbf{0}_2) = 0.36^2/0.36 + 0.64^2/0.64 = 1$.

TAVOLA 20 - Inerzia delle proiezioni su un vettore.

È assegnato il punto $\mathbf{0}_{\mathrm{I}}$, origine della base canonica di riferimento e un profilo-colonna $\mathbf{c}_j$ con massa $\overline{r}_j$, ed è anche *assegnato* un vettore $\mathbf{u}$ con origine in $\mathbf{0}_{\mathrm{I}}$ e lunghezza $\mathbf{D}_{\overline{\mathbf{c}}}{}^{\mathbf{1}}$-unitaria. Il vettore $h(\mathbf{c}_j)\,\mathbf{u} = (\mathbf{c}_j^T\,\mathbf{D}_{\overline{\mathbf{c}}}{}^{\mathbf{1}}\,\mathbf{u})\,\mathbf{u}$ è detto *proiezione* $\mathbf{D}_{\overline{\mathbf{c}}}{}^{\mathbf{1}}$-ortogonale di $\mathbf{c}_j$ su $\mathbf{u}$. Quindi $h(\mathbf{c}_j)$ è l'ascissa del punto proiezione su $\mathbf{u}$ e $\overline{r}_j\,h^2(\mathbf{c}_j)$ è l'inerzia di questa proiezione rispetto all'origine.

Quando, oltre all'origine $\mathbf{0}_{\mathrm{I}}$, i profili assegnati sono due o più, ad esempio $\mathbf{c}_j$ con massa $\overline{r}_j$ e $\mathbf{c}_k$ con massa $\overline{r}_k$, il vettore $\mathbf{u}$ può essere *individuato*, imponendo la condizione che l'inerzia delle due proiezioni sia massima

$$\overline{r}_j\,h^2(\mathbf{c}_j) + \overline{r}_k\,h^2(\mathbf{c}_k) = massima.$$

Si può immaginare il vettore $\mathbf{u}$ come un'asta con snodo in $\mathbf{0}_{\mathrm{I}}$, orientabile in tutte le direzioni: la sua estremità non vincolata descrive una circonferenza di raggio $\mathbf{D}_{\overline{\mathbf{c}}}{}^{\mathbf{1}}$-unitario. Variando l'orientamento, l'inerzia delle due proiezioni cambia di valore.

La condizione di inerzia massima su $\mathbf{u}$, può esprimersi in forma alternativa, ma del tutto equivalente. Siccome la distanza (distribuzionale) di un profilo dall'origine del sistema di riferimento è fissa, rendere massime le distanze delle proiezioni dall'origine equivale a rendere minime le distanze, pesate, dei profili dalle loro proiezioni

$$\overline{r}_j\,d_D^2(\mathbf{c}_j, h(\mathbf{c}_j)\,\mathbf{u}) + \overline{r}_k\,d_D^2(\mathbf{c}_k, h(\mathbf{c}_k)\,\mathbf{u}) = minima.$$

Quindi $\mathbf{u}$ può essere individuato cercando quell'orientamento che lo posiziona più 'vicino' ai profili. Questo criterio è detto dei minimi quadrati ponderati.

TAVOLA 21 - Base ortonormale e $\mathbf{D}_{\overline{\mathbf{c}}}{}^{\mathbf{1}}$-ortonormale.

La figura evidenzia tre basi di riferimento per i profili-colonna della matrice *Spettacoli-3* di TAV. 14. La prima è la base euclidea canonica costituita dai vettori $\mathbf{e}_1$, $\mathbf{e}_2$ ed $\mathbf{e}_3$ (Sez. 2.5) che individuano i vertici del simplesso, un triangolo equilatero, ove gli 8 profili-colonna ed il loro baricentro sono confinati. La base è *ortonormale*, ossia i suoi tre vettori sono due a due ortogonali ed unitari, ma non $\mathbf{D}_{\overline{\mathbf{c}}}{}^{\mathbf{1}}$-ortonormale, in quanto risulta soltanto $\mathbf{D}_{\overline{\mathbf{c}}}{}^{\mathbf{1}}$-ortogonale (Sez. 2.8).

La seconda è costituita dagli autovettori $\mathbf{u}_1$ e $\mathbf{u}_2$ (e da $\overline{\mathbf{c}}$, il vettore che individua il baricentro) con origine in $\mathbf{0}_3$. La base risulta $\mathbf{D}_{\overline{\mathbf{c}}}{}^{\mathbf{1}}$-*ortonormale* per costruzione (Sez. 3.13), ma non è ortonormale, nel senso che nello spazio euclideo canonico i suoi vettori non sono ne´ ortogonali, ne´ unitari. I vettori $\mathbf{u}_1$ ed $\mathbf{u}_2$ individuano un piano parallelo a quello del simplesso.

La terza base è costituita dagli autovettori $\mathbf{u}_1^\star$ e $\mathbf{u}_2^\star$ (e da $\overline{\mathbf{c}}$). Questa base si differenzia dalla precedente per avere l'origine nel baricentro (Sez. 3.14). Si tratta quindi di una traslazione rigida della precedente base, lungo la retta che collega origine e baricentro. Gli autovettori $\mathbf{u}_1^\star$ ed $\mathbf{u}_2^\star$ individuano gli *assi fattoriali d'inerzia* che risultano quindi contenuti nel piano del simplesso.

TAVOLA 22 - Asse fattoriale e retta di regressione.

Il primo asse fattoriale d'inerzia, nella figura a sinistra, è la retta passante per il baricentro dei profili e la più 'vicina' a questi, nel senso che rende minima la somma dei quadrati delle distanze distribuzionali ponderate, tra i profili e le loro proiezioni $\mathbf{D}_{\bar{\mathbf{c}}}^{\mathbf{1}}$-*ortogonali* sulla retta (Sez. 3.7 e TAV. 20). I pesi sono le masse dei profili.

Anche la retta di regressione semplice passa per il baricentro dei profili, ma rende minima la somma dei quadrati delle distanze - euclidee canoniche ora - calcolate *parallelamente* alla seconda dimensione, come nella figura a destra. Le seconde coordinate di ogni profilo sono considerate come osservazioni della variabile dipendente e si ritiene, per ipotesi, che questa sia 'misurata senza errore'. I pesi devono essere interi ed indicano la frequenza con cui un profilo è stato rilevato.

TAVOLA 23 - Proiezione di un profilo su un autovettore.

La figura mostra perché l'ascissa delle proiezioni del profilo colonna $\mathbf{c}_j$ sul primo autovettore $\mathbf{u}_1$ sia anche eguale alla proiezione di $\mathbf{c}_j - \overline{\mathbf{c}}$ sull'autovettore $\mathbf{u}_1^\star$. Si tratta dell'importante conseguenza di due fatti. Primo, il profilo $\overline{\mathbf{c}}$ è autovettore di entrambe le basi, corrispondente all'autovalore più grande nella base con origine in $\mathbf{0}_3$, e al più piccolo nella base con origine nel baricentro. Secondo, in entrambe le basi, gli autovettori sono $\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$-ortogonali due a due per costruzione. Questa proprietà vale per tutti i profili e per tutti gli assi fattoriali.

TAVOLA 24 - Mappa dei profili delle colonne. La mappa si riferisce ai profili delle 8 colonne della matrice *Spettacoli-3* di TAV. 14. La retta orizzontale rappresenta il primo asse fattoriale ove $\lambda_1 = 0.005$ e la percentuale d'inerzia è $\tau_1 = 71\,\%$; quella verticale il secondo, ove $\lambda_2 = 0.002$ e $\tau_2 = 29\,\%$ e il loro punto d'incrocio il baricentro $\overline{\mathbf{c}}$ della nuvola dei punti. La distanza unitaria sul primo asse è fissata dalla lunghezza dell'autovettore $\mathbf{u}_1^\star$, quella sul secondo dall'autovettore $\mathbf{u}_2^\star$. La scala è la medesima su entrambi gli assi. Ogni punto rappresenta un profilo colonna, ossia una distribuzione di frequenze condizionate. La posizione è indicata da un cerchietto la cui area è proporzionale alla massa del profilo. La distanza tra punti traduce la diversità di "forma" dei profili. Questa può essere grande, come tra gli spettacoli di Burattini e Marionette e le rappresentazioni di Operette, rivelando il fatto che nelle regioni in cui gli spettatori privilegiano un tipo di spettacolo, penalizzano l'altro.

Analisi della matrice *Spettacoli-3* di TAV. 14
Asse fattoriale 1

j	Profilo $\mathbf{c}_j$	Ascissa g_{j1}	Contributo $CTR_1(\mathbf{c}_j)$	Cos2 $COS_1^2(\mathbf{c}_j)$	Massa $\overline{r}_j$
1	PROSA	+.049	.208	.940	.419
2	LIRICA BALLET.	−.122	.342	.863	.109
3	CONC CLASSICA	+.043	.056	.536	.144
4	OPERETTA	−.179	.084	.720	.013
5	RIVISTA	+.032	.008	.047	.039
6	CONC LEGGERA	−.049	.122	.991	.246
7	BURAT. MARIO.	+.198	.097	.495	.012
8	SAGGI COREOG.	−.150	.081	.604	.017

Asse fattoriale 2

j	Profilo $\mathbf{c}_j$	Ascissa g_{j2}	Contributo $CTR_2(\mathbf{c}_j)$	Cos2 $COS_2^2(\mathbf{c}_j)$	Massa $\overline{r}_j$
1	PROSA	+.012	.028	.060	.419
2	LIRICA BALLET.	−.049	.116	.137	.109
3	CONC CLASSICA	+.040	.104	.464	.144
4	OPERETTA	+.111	.070	.280	.013
5	RIVISTA	−.142	.353	.953	.039
6	CONC LEGGERA	−.005	.002	.009	.246
7	BURAT. MARIO.	−.200	.211	.505	.012
8	SAGGI COREOG.	+.122	.114	.396	.017

TAVOLA 25 - Contributi relativi e Qualità della rappresentazione.

I due principali indicatori per l'interpretazione dei risultati dell'Analisi delle Corrispondenze sono positivi e possono variare tra 0 e 1. Il *contributo relativo* di un profilo all'inerzia di un asse, $CTR_a(\mathbf{c}_j)$, misura lo "sforzo" fatto da un profilo per orientare verso di sè l'asse fattoriale. La somma dei contributi relativi per ogni asse, ossia il totale della colonna Contributo, vale 1.000. Per esempio, per il primo asse: $0.208 + 0.342 + \ldots + 0.081 = 1.000$.

Il contributo relativo di un asse all'inerzia di un profilo, o *Qualità della rappresentazione* delle distanze, $COS^2_a(\mathbf{c}_j)$, misura la fedeltà della riproduzione su un asse della distanza tra profilo e baricentro della nuvola. Per ogni profilo, la somma dei Cos2 su tutti gli assi vale 1.000. Ad esempio, per il profilo Prosa: $0.940 + 0.060 = 1.000$.

TAVOLA 26 - Qualità della riproduzione delle distanze.

Geometricamente, il contributo relativo $COS_a^2(\mathbf{c}_j)$ di un asse fattoriale all'inerzia di un profilo è un coseno al quadrato.

Quando il profilo si trova vicino all'asse, il coseno quadrato è prossimo a 1, mentre l'angolo θ tra il vettore $\mathbf{c}_j - \overline{\mathbf{c}}$ e l'autovettore $\mathbf{u}_a^\star$ è prossimo a 0. In questo caso l'ascissa g_{ja} del profilo sull'asse rende conto abbastanza fedelmente dell'effettiva distanza $d_D(\mathbf{c}_j, \overline{\mathbf{c}})$ tra profilo e baricentro in $\Re^J$.

Quando invece $COS_a^2(\mathbf{c}_j)$ è quasi nullo, il profilo è lontano dall'asse e l'angolo è prossimo a 90° gradi. Essendo il profilo $\mathbf{D}_{\overline{\mathbf{c}}}^{-1}$-ortogonale all'asse, g_{ja} riproduce pessimamente la distanza $d_D(\mathbf{c}_j, \overline{\mathbf{c}})$ tra profilo e baricentro.

Assi fattoriali d′inerzia

		1	2	...	A	Totali
	1	$\overline{c}_1 f_{11}^2$	$\overline{c}_1 f_{12}^2$	...	$\overline{c}_1 f_{1A}^2$	$\overline{c}_1 \sum_a^A f_{1a}^2$
	2	$\overline{c}_2 f_{21}^2$	$\overline{c}_2 f_{22}^2$	...	$\overline{c}_2 f_{2A}^2$	$\overline{c}_2 \sum_a^A f_{2a}^2$
Righe	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	$\overline{c}_I f_{I1}^2$	$\overline{c}_I f_{I2}^2$	...	$\overline{c}_I f_{IA}^2$	$\overline{c}_I \sum_a^A f_{Ia}^2$
Inerzie		λ_1	λ_2	...	λ_A	$In_{\overline{\mathbf{r}}} = In_{\overline{\mathbf{c}}}$
	1	$\overline{r}_1 g_{11}^2$	$\overline{r}_1 g_{12}^2$	...	$\overline{r}_1 g_{1A}^2$	$\overline{r}_1 \sum_a^A g_{1a}^2$
	2	$\overline{r}_2 g_{21}^2$	$\overline{r}_2 g_{22}^2$	...	$\overline{r}_2 g_{2A}^2$	$\overline{r}_2 \sum_a^A g_{2a}^2$
Colonne	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	J	$\overline{r}_J g_{J1}^2$	$\overline{r}_J g_{J2}^2$	...	$\overline{r}_J g_{JA}^2$	$\overline{r}_J \sum_a^A g_{Ja}^2$

TAVOLA 27 - Scomposizione dell'inerzia per profilo e per asse.

Le inerzie totali $In_{\overline{\mathbf{r}}}$ dei profili delle righe rispetto al loro baricentro $\overline{\mathbf{r}}$ e $In_{\overline{\mathbf{c}}}$ dei profili delle colonne rispetto a $\overline{\mathbf{c}}$, sono eguali e possono scomporsi sugli assi fattoriali e tra i profili stessi in modo perfettamente simmetrico. La scomposizione ha molte analogie con quella della varianza. Le colonne nella Tavola mostrano i *contributi assoluti* dei profili (delle righe, sopra e delle colonne, sotto) all'inerzia su un asse fattoriale. Il contributo assoluto rapportato all'inerzia complessiva sull'asse dà il contributo relativo del profilo all'inerzia dell'asse, per cui ad es. $CTR_1(\mathbf{c}_1) = \overline{r}_1 g_{11}^2/\lambda_1$. Rapportato invece all'inerzia complessiva del profilo, fornisce il contributo relativo dell'asse all'inerzia del profilo, ossia Cos2, per cui ad esempio $COS_1^2(\mathbf{c}_1) = \overline{r}_1 g_{11}^2/\overline{r}_1 \sum_a g_{1a}^2 = g_{11}^2/\sum_a g_{1a}^2$.

Questi due contributi relativi costituiscono l'indispensabile supporto numerico per la corretta interpretazione dei risultati dell'Analisi delle Corrispondenze.

Analisi dei profili delle righe in $\Re^J$ quando $J < I$

Inerzia su $\mathbf{v}$ e $\mathbf{v}^\star$	$\mathbf{v}^T\mathbf{D}_{\overline{\mathbf{r}}}^{-1}\mathbf{R}^T\mathbf{D}_{\overline{\mathbf{c}}}\mathbf{R}\mathbf{D}_{\overline{\mathbf{r}}}^{-1}\mathbf{v}$	$\mathbf{v}^{\star T}\mathbf{D}_{\overline{\mathbf{r}}}^{-1}(\mathbf{R}-\overline{\mathbf{R}})^T\mathbf{D}_{\overline{\mathbf{c}}}(\mathbf{R}-\overline{\mathbf{R}})\mathbf{D}_{\overline{\mathbf{r}}}^{-1}\mathbf{v}^\star$
Matrice	$\mathbf{R}^T\mathbf{C}$	$(\mathbf{R}-\overline{\mathbf{R}})\,\mathbf{C}^T$
Autovalore banale	$\mu_0 = 1$	$\mu_{\mathrm{J}}^\star = 0$
Autovettore banale	$\mathbf{v}_0 = \overline{\mathbf{r}}$	$\mathbf{v}_{\mathrm{J}}^\star = \overline{\mathbf{r}}$
Numero autovalori [1]	$J-1$	$J-1$
Autovalori	μ_a	$\mu_a^\star = \mu_a$
Autovettori	$\mathbf{v}_a$	$\mathbf{v}_a^\star$
Origine	$\mathbf{0}_{\mathrm{J}}$	$\overline{\mathbf{r}}$
Coordinate su asse a	$f_{ia} = \sum_j r_{ij}\,\frac{1}{\overline{r}_j}\,v_{ja}$	$f_{ia} = \sum_j r_{ij}\,\frac{1}{\overline{r}_j}\,v_{ja}^\star$

Analisi dei profili delle colonne in $\Re^I$ quando $I < J$

Inerzia su $\mathbf{u}$ e $\mathbf{u}^\star$	$\mathbf{u}^T\mathbf{D}_{\overline{\mathbf{c}}}^{-1}\mathbf{C}\,\mathbf{D}_{\overline{\mathbf{r}}}\mathbf{C}^T\mathbf{D}_{\overline{\mathbf{c}}}^{-1}\mathbf{u}$	$\mathbf{u}^{\star T}\mathbf{D}_{\overline{\mathbf{c}}}^{-1}(\mathbf{C}-\overline{\mathbf{C}})\mathbf{D}_{\overline{\mathbf{r}}}(\mathbf{C}-\overline{\mathbf{C}})^T\mathbf{D}_{\overline{\mathbf{c}}}^{-1}\mathbf{u}^\star$
Matrice	$\mathbf{C}\,\mathbf{R}^T$	$(\mathbf{C}-\overline{\mathbf{C}})\,\mathbf{R}^T$
Autovalore banale	$\lambda_0 = 1$	$\lambda_{\mathrm{I}}^\star = 0$
Autovettore banale	$\mathbf{u}_0 = \overline{\mathbf{c}}$	$\mathbf{u}_{\mathrm{I}}^\star = \overline{\mathbf{c}}$
Numero autovalori [1]	$I-1$	$I-1$
Autovalori	λ_a	$\lambda_a^\star = \lambda_a$
Autovettori	$\mathbf{u}_a$	$\mathbf{u}_a^\star$
Origine	$\mathbf{0}_{\mathrm{I}}$	$\overline{\mathbf{c}}$
Coordinate su asse a	$g_{ja} = \sum_i c_{ij}\,\frac{1}{\overline{c}_i}\,u_{ia}$	$g_{ja} = \sum_i c_{ij}\,\frac{1}{\overline{c}_i}\,u_{ia}^\star$

Nota [1] : i primi $A = min\,(I-1, J-1)$ autovalori sono positivi, i rimanenti nulli.

TAVOLA 28 - Confronto tra le analisi dei profili.

Nello spazio dei profili delle righe (in alto) e delle colonne (in basso), l'Analisi delle Corrispondenze ricerca quegli assi che rendono massima l'inerzia, rispetto all'origine (penultima colonna) o al baricentro dei profili (ultima colonna), delle proiezioni dei profili sugli assi.

Nuvole di punti, assi fattoriali, fattori ed inerzie

	PUNTI-RIGA	PUNTI-COLONNA
Spazio	$\Re^J$	$\Re^I$
Matrice delle distanze	$\mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}$	$\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}$
Numero di punti	I	J
Punto profilo	${\mathbf{r}_i}^T$	$\mathbf{c}_j$
Nuvola di punti	$\mathbf{R}$	$\mathbf{C}$
Baricentro	$\overline{\mathbf{r}}$	$\overline{\mathbf{c}}$
Masse	$\overline{\mathbf{c}}$	$\overline{\mathbf{r}}$
Matrice delle masse	$\mathbf{D}_{\overline{\mathbf{c}}}$	$\mathbf{D}_{\overline{\mathbf{r}}}$
Numero di assi	$A = J - 1$	$A = I - 1$
Assi fattoriali	$\mathbf{v}_a^\star$	$\mathbf{u}_a^\star$
Equazione	$\mathbf{R}^T\mathbf{C}\ \mathbf{v}_a^\star = \lambda_a\ \mathbf{v}_a^\star$	$\mathbf{C}\ \mathbf{R}^T\mathbf{u}_a^\star = \lambda_a\ \mathbf{u}_a^\star$
Lunghezza	$\mathbf{v}_a^{\star T}\ \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\mathbf{v}_a^\star = 1$	$\mathbf{u}_a^{\star T}\ \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\mathbf{u}_a^\star = 1$
Ortogonalità	$\mathbf{v}_a^{\star T}\ \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\ \mathbf{v}_b^\star = 0$	$\mathbf{u}_a^{\star T}\ \mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\ \mathbf{u}_b^\star = 0$
Numero di fattori	$A = J - 1$	$A = I - 1$
Fattori	$\mathbf{f}_a = \mathbf{R}\ \mathbf{D}_{\overline{\mathbf{r}}}^{-\mathbf{1}}\mathbf{v}_a^\star$	$\mathbf{g}_a = \mathbf{C}^T\mathbf{D}_{\overline{\mathbf{c}}}^{-\mathbf{1}}\mathbf{u}_a^\star$
Media	$\overline{\mathbf{c}}^T\ \mathbf{f}_a = 0$	$\overline{\mathbf{r}}^T\ \mathbf{g}_a = 0$
Inerzia	$\mathbf{f}_a^T\mathbf{D}_{\overline{\mathbf{c}}}\ \mathbf{f}_a = \lambda_a$	$\mathbf{g}_a^T\mathbf{D}_{\overline{\mathbf{r}}}\ \mathbf{g}_a = \lambda_a$
Ortogonalità	$\mathbf{f}_a^T\ \mathbf{D}_{\overline{\mathbf{c}}}\ \mathbf{f}_b = 0$	$\mathbf{g}_a^T\ \mathbf{D}_{\overline{\mathbf{r}}}\ \mathbf{g}_b = 0$
Inerzia sull'asse a	λ_a	λ_a
Inerzia totale	$In_{\overline{\mathbf{r}}} = \sum_a^A \lambda_a$	$In_{\overline{\mathbf{c}}} = \sum_a^A \lambda_a$

TAVOLA 29 - Relazioni tra gli spazi dei profili.

Il prospetto riassume i risultati delle analisi dei profili delle righe, quando $I > J$, e delle colonne, quando $I < J$. I risultati finali delle due analisi sono legati dalle *relazioni di transizione*, diretta conseguenza delle trasformazioni effettuate simmetricamente sulle righe e sulle colonne della matrice di contingenza.

Analisi della matrice *Spettacoli* di TAV. 2

Scomposizione dell'inerzia $In_{\bar{\mathbf{r}}}$ e $In_{\bar{\mathbf{c}}}$ lungo gli assi fattoriali

Rango a	Inerzia λ_a	Percento τ_a	Cumulo $\sum_a \tau_a$	Diagramma delle percentuali ——10——20——30——40——
1	0.036 97	46.35 %	46.35 %	
2	0.025 22	31.62 %	77.97 %	
3	0.007 60	9.53 %	87.51 %	
4	0.004 96	6.22 %	93.72 %	
5	0.003 06	3.84 %	97.56 %	
6	0.001 35	1.69 %	99.25 %	
7	0.000 60	0.75 %	100.00 %	
$In_{\bar{\mathbf{r}}} =$ $In_{\bar{\mathbf{c}}} =$	0.079 76	100.00 %		

TAVOLA 30 - Tavola delle Inerzie.

Il primo risultato stampato da un programma di Analisi delle Corrispondenze di una matrice, in questo caso la matrice *Spettacoli* di ordine 20×8 della TAV. 2, è la Tavola delle inerzie che indica in qual modo l'inerzia dei profili rispetto al baricentro $In_{\bar{\mathbf{r}}} = In_{\bar{\mathbf{c}}} = 0.079\,75$ è stata ripartita sugli $A = min\ (I, J) - 1 = min\ (20, 8) - 1 = 7$ assi fattoriali, ordinati per inerzie sull'asse decrescenti. Le inerzie λ_a misurano la dispersione geometrica rispetto all'origine delle proiezioni dei profili sugli assi fattoriali. La percentuale d'inerzia sull'asse, τ_a, è ottenuta come percentuale dal rapporto $\lambda_a / In_{\bar{\mathbf{r}}}$, e sul primo asse vale $\tau_1 = 0.036\,97/0.079\,76 \times 100 = 46.35\ \%$. La terza colonna riporta il totale progressivo delle percentuali d'inerzia. Si vede che sul piano principale individuato dagli assi fattoriali di rango 1 e 2, l'inerzia delle proiezioni è il $77.97\ \%$ dell'inerzia che hanno i profili nel loro spazio ambiente a 7 dimensioni. Nel sottospazio principale tridimensionale, tale percentuale sale a $87.51\ \%$. Infine il diagramma permette di apprezzare con un colpo d'occhio le differenze relative tra percentuali d'inerzia, o tra autovalori, dato che sono proporzionali. Si nota chiaramente che i primi due assi sono predominanti, perché c'è un brusco calo tra gli assi di rango 2 e 3, a cui fa seguito una diminuzione regolare. L'esame scrupoloso di tutte le informazioni nella Tavola delle inerzie permette di farsi un'idea della dispersione della nuvola di profili sui diversi assi fattoriali e perciò di quali assi, piani, e spazi tridimensionali è prevedibile una buona caratterizzazione.

Analisi della matrice *Spettacoli* di TAV. 2

Asse fattoriale 1. Profili delle colonne

j	Profilo $\mathbf{c}_j$	Ascissa g_{j1}	Contributo $CTR_1(\mathbf{c}_j)$	Cos2 $COS_1^2(\mathbf{c}_j)$	Massa $\overline{r}_j$	Inerzia $INR_7(\mathbf{c}_j)$
1	PROSA	−.060	.040	.161	.419	.116
3	CONC CLASSICA	−.100	.039	.275	.144	.066
7	BURAT. MARIO.	−.282	.026	.383	.012	.031
6	CONC LEGGERA	−.057	.022	.048	.246	.208
4	OPERETTA	−.071	.002	.021	.013	.039
5	RIVISTA	−.035	.001	.009	.039	.069
8	SAGGI COREOG.	−.016	.000	.001	.017	.066
2	LIRICA BALLET.	+.542	.870	.998	.109	.404

TAVOLA 31 - Asse fattoriale 1. Profili delle colonne

In questa Tavola sono riunite tutte le informazioni che consentono di caratterizzare un asse fattoriale, in questo caso il primo, individuato dall'autovettore $\mathbf{u}_1^\star$ e sul quale la percentuale d'inerzia è $\tau_1 = 46.35\%$. Nell'ordine è presentato:

- j : numero d'ordine del profilo colonna $\mathbf{c}_j$.
- $\mathbf{c}_j$: modalità del profilo $\mathbf{c}_j$.
- g_{j1} : ascissa della proiezione del profilo $\mathbf{c}_j$ sul primo asse. I profili sono elencati separatamente in due gruppi: prima quelli con ascissa negativa, poi quelli con ascissa positiva.
- $CTR_1(\mathbf{c}_j) = \overline{r}_j\, g_{j1}^2 / In_{\overline{\mathbf{c}}}$: contributo relativo del profilo all'inerzia λ_1 del primo asse. Indica lo 'sforzo' fatto dal profilo per attrarre verso di sé l'asse fattoriale. I profili con ascisse del medesimo segno sono ordinati per valori decrescenti del contributo. La somma dei contributi degli 8 profili vale 1.
- $COS_1^2(\mathbf{c}_j) = g_{j1}^2 / \sum_a g_{ja}^2$: qualità della rappresentazione della distanza tra il profilo $\mathbf{c}_j$ e il baricentro $\overline{\mathbf{c}}$ sul primo asse. La somma su tutti gli $A = min\,(20, 8) - 1 = 7$ assi fattoriali vale 1.
- $\overline{r}_j = n_{ij}/n_{+j}$: massa del profilo. La somma per gli 8 profili vale 1.
- $INR_7(\mathbf{c}_j)$: quota d'inerzia del profilo in un sottospazio, che in questo caso essendo di dimensione 7 coincide con l'iperpiano del simplesso. Confrontando massa e quota d'inerzia si individuano nella configurazione i profili eccentrici e quelli più centrali.

Analisi della matrice *Spettacoli* di TAV. 2

Asse fattoriale 1. Profili delle righe

Percentuale d'inerzia $\tau_1 = 46.35\%$

i	Profilo $\mathbf{r}_i$	Ascissa f_{i1}	Contributo $CTR_1(\mathbf{r}_i)$	Cos2 $COS_1^2(\mathbf{r}_i)$	Massa $\overline{c}_i$	Inerzia $INR_7(\mathbf{r}_i)$
4	Trentino AA	−.222	.035	.457	.026	.035
8	Emilia R	−.103	.030	.598	.105	.024
13	Abruzzi	−.219	.023	.394	.018	.028
3	Lombardia	−.069	.023	.711	.178	.015
20	Sardegna	−.160	.018	.091	.026	.093
12	Lazio	−.060	.012	.061	.122	.090
15	Campania	−.070	.007	.118	.054	.028
17	Basilicata	−.239	.006	.365	.004	.008
2	Valle d'Aosta	−.305	.004	.073	.002	.026
16	Puglia	−.064	.003	.123	.028	.012
10	Umbria	−.064	.002	.119	.014	.006
18	Calabria	−.061	.001	.075	.014	.008
14	Molise	−.112	.001	.205	.002	.002
11	Marche	−.022	.000	.009	.021	.014
19	Sicilia	−.009	.000	.001	.070	.123
1	Piemonte	−.007	.000	.001	.075	.060
7	Liguria	−.009	.000	.004	.038	.011
5	Veneto	+.584	.832	.995	.090	.388
9	Toscana	+.021	.001	.028	.079	.016
6	Friuli VG	+.030	.001	.023	.034	.017

TAVOLA 32 - Asse fattoriale 1. Profili delle righe

Per il significato delle grandezze contenute in questa Tavola fare riferimento alla precedente TAV. 31. I profili sono divisi in base al segno della loro ascissa sul primo asse fattoriale e quindi elencati per valore decrescente del contributo relativo $CTR_1(\mathbf{r}_i)$ del profilo all'inerzia del primo asse, individuato dall'autovettore $\mathbf{v}_1^\star$.

Questo modo di presentare i risultati per ogni asse fattoriale facilita la caratterizzazione degli assi evitando al principiante errori marchiani.

Analisi della matrice *Spettacoli* di TAV. 2

Asse fattoriale 2

Percentuale d'inerzia $\tau_2 = 31.62\%$

j	Profilo $\mathbf{c}_j$	Ascissa g_{j2}	Contributo $CTR_2(\mathbf{c}_j)$	Cos2 $COS_2^2(\mathbf{c}_j)$	Massa $\overline{r}_j$	Inerzia $INR_7(\mathbf{c}_j)$
1	PROSA	−.126	.264	.720	.419	.116
5	RIVISTA	−.112	.019	.089	.039	.069
3	CONC CLASSICA	−.054	.017	.081	.144	.066
7	BURAT. MARIO.	−.098	.005	.046	.012	.031
2	LIRICA BALLET.	−.014	.001	.001	.109	.404
6	CONC LEGGERA	+.248	.600	.911	.246	.208
8	SAGGI COREOG.	+.369	.094	.449	.017	.066
4	OPERETTA	+.013	.000	.001	.013	.039

i	Profilo $\mathbf{r}_i$	Ascissa f_{i2}	Contributo $CTR_2(\mathbf{r}_i)$	Cos2 $COS_2^2(\mathbf{r}_i)$	Massa $\overline{c}_i$	Inerzia $INR_7(\mathbf{r}_i)$
19	Sicilia	−.340	.318	.821	.070	.123
12	Lazio	−.188	.172	.607	.122	.090
10	Umbria	−.097	.005	.272	.014	.006
7	Liguria	−.046	.003	.092	.038	.011
13	Abruzzi	−.062	.003	.032	.018	.028
14	Molise	−.147	.002	.353	.002	.002
3	Lombardia	−.015	.002	.033	.178	.015
17	Basilicata	−.034	.000	.007	.004	.008
20	Sardegna	+.484	.245	.838	.026	.093
1	Piemonte	+.194	.112	.588	.075	.060
2	Valle d'Aosta	+.867	.048	.588	.002	.026
11	Marche	+.210	.036	.847	.021	.014
9	Toscana	+.093	.027	.542	.079	.016
8	Emilia Romagna	+.045	.008	.113	.105	.024
4	Trentino AAdige	+.080	.007	.059	.026	.035
18	Calabria	+.102	.006	.210	.014	.008
5	Veneto	+.028	.003	.002	.090	.388
6	Friuli V. Giulia	+.030	.001	.023	.034	.017
16	Puglia	+.029	.001	.025	.028	.012
15	Campania	+.020	.001	.010	.054	.028

TAVOLA 33 - Asse fattoriale 2.

Analisi della matrice *Spettacoli* di TAV. 2

Asse fattoriale 3

Percentuale d'inerzia $\tau_3 = 9.53\%$

j	Profilo $\mathbf{c}_j$	Ascissa g_{j3}	Contributo $CTR_3(\mathbf{c}_j)$	Cos2 $COS_3^2(\mathbf{c}_j)$	Massa $\overline{r}_j$	Inerzia $INR_7(\mathbf{c}_j)$
5	RIVISTA	−.250	.323	.446	.039	.069
6	MUSICA LEGGE.	−.040	.052	.024	.246	.208
7	BURAT. MARIO.	−.153	.037	.113	.012	.031
1	PROSA	−.007	.003	.002	.419	.116
4	OPERETTA	+.392	.258	.629	.013	.039
3	MUSICA CLAS,	+.102	.196	.283	.144	.066
8	SAGGI COREOG.	+.240	.132	.190	.017	.066
2	LIRICA BALLET.	+.005	.000	.000	.109	.404

i	Profilo $\mathbf{r}_i$	Ascissa f_{i3}	Contributo $CTR_3(\mathbf{r}_i)$	Cos2 $COS_3^2(\mathbf{r}_i)$	Massa $\overline{c}_i$	Inerzia $INR_7(\mathbf{r}_i)$
12	Lazio	−.113	.207	.220	.122	.090
15	Campania	−.137	.134	.461	.054	.028
7	Liguria	−.105	.055	.476	.038	.011
9	Toscana	−.054	.030	.179	.079	.016
17	Basilica	−.087	.004	.008	.004	.008
20	Sardegna	−.030	.003	.093	.026	.093
14	Molise	−.061	.001	.002	.002	.002
5	Veneto	−.009	.001	.388	.090	.388
3	Lombardi	−.004	.000	.015	.178	.015
19	Sicilia	+.155	.219	.170	.070	.123
6	Friuli V. Giulia	+.162	.119	.681	.034	.017
16	Puglia	+.167	.102	.833	.028	.012
1	Piemonte	+.061	.037	.059	.075	.060
13	Abruzzi	+.115	.032	.110	.018	.028
10	Umbria	+.092	.015	.242	.014	.006
2	Valle d'Aosta	+.262	.015	.054	.002	.026
11	Marche	+.072	.014	.099	.021	.014
18	Calabria	+.084	.012	.142	.014	.008
4	Trentino AAdige	+.007	.000	.000	.026	.035
8	Emilia Romagna	+.002	.000	.000	.105	.024

TAVOLA 34 - Asse fattoriale 3.

TAVOLA 35 - Rappresentazione congiunta dei fattori.

Il fatto che le nuvole dei profili delle righe e delle colonne si trovino in sottospazi della medesima dimensionalità, $A = min(I, J) - 1$, pur di spazi diversi, $\Re^J$ e $\Re^I$, che l'inerzia delle loro proiezioni su assi fattoriali del medesimo rango sia la stessa, λ_a, e che per le relazioni di transizione ogni profilo riga sia il 'quasi-baricentro' dei profili delle colonne e viceversa, induce a rappresentare congiuntamente, ossia a sovrapporre, assi del medesimo rango per procedere alla loro caratterizzazione.

Analisi della matrice *Spettacoli* di TAV. 2

Piano fattoriale 1, 2

Percentuale d'inerzia $\tau_{(1,2)} = 77.97\%$

	Profilo	Distanza	Contributo	Cos2	Massa
j	$\mathbf{c}_j$	$\sqrt{g_{j1}^2 + g_{j2}^2}$	$CTR_{(1,2)}(\mathbf{c}_j)$	$COS^2_{(1,2)}(\mathbf{c}_j)$	$\overline{r}_j$
2	LIRICA BALLE.	0.542	0.518	0.999	0.109
6	MUSICA LEGG.	0.254	0.256	0.959	0.246
1	PROSA	0.139	0.131	0.881	0.419
i	$\mathbf{r}_i$	$\sqrt{f_{i1}^2 + f_{i2}^2}$	$CTR_{(1,2)}(\mathbf{r}_i)$	$COS^2_{(1,2)}(\mathbf{r}_i)$	$\overline{c}_i$
5	Veneto	0.584	0.496	0.997	0.090
19	Sicilia	0.340	0.129	0.822	0.070
20	Sardegna	0.510	0.110	0.930	0.026
12	Lazio	0.198	0.077	0.668	0.122
1	Piemonte	0.194	0.045	0.588	0.075
2	Valle d'Aosta	0.919	0.022	0.661	0.002
8	Emilia Romagna	0.113	0.021	0.711	0.105
11	Marche	0.211	0.015	0.856	0.021
3	Lombardia	0.071	0.014	0.743	0.178

TAVOLA 36 - Piano fattoriale 1, 2.

La Tavola elenca soltanto i profili con una soddisfacente qualità di rappresentazione sul piano e, per ciascuno di essi, riporta:

- Distanza : indica la distanza distribuzionale della proiezione di un profilo sul piano dall'origine degli assi.
- Contributo : è il contributo relativo del profilo all'inerzia del piano, ossia lo 'sforzo' fatto dal profilo per orientare verso di sè il piano fattoriale.
- Cos2 : indica la qualità della rappresentazione del profilo sul piano. Tanto più è vicino ad 1.000, tanto meglio la distanza della proiezione del profilo dall'origine traduce la distanza effettiva del profilo dal baricentro nello spazio A-dimensionale.
- Massa : è la massa del profilo che ne indica l'importanza relativa.

La mappa è mostrata nella TAV. 37 a fianco.

TAVOLA 37 - Mappa 1, 2.

Mappa principale ottenuta dall'analisi della matrice *Spettacoli* di TAV. 2. L'asse orizzontale raffigura il primo asse fattoriale, quello orizzontale il secondo. L'unità di scala è la stessa su entrambi gli assi. Sulla mappa sono mostrati soltanto i profili con una soddisfacente qualità di rappresentazione.

Analisi della matrice *Spettacoli* di TAV. 2

Piano fattoriale 2, 3

Percentuale d'inerzia $\tau_{(2,3)} = 41.15\%$

	Profilo	Distanza	Contributo	Cos2	Massa
j	$\mathbf{c}_j$	$\sqrt{g_{j2}^2 + g_{j3}^2}$	$CTR_{(2,3)}(\mathbf{c}_j)$	$COS^2_{(2,3)}(\mathbf{c}_j)$	$\overline{r}_j$
6	MUSICA LEGG.	0.251	0.473	0.934	0.246
1	PROSA	0.126	0.203	0.722	0.419
8	SAGGI COREO.	0.440	0.103	0.639	0.017
5	RIVISTA	0.274	0.090	0.535	0.039
4	OPERETTA	0.392	0.060	0.629	0.013
i	$\mathbf{r}_i$	$\sqrt{f_{i2}^2 + f_{i3}^2}$	$CTR_{(2,3)}(\mathbf{r}_i)$	$COS^2_{(2,3)}(\mathbf{r}_i)$	$\overline{c}_i$
19	Sicilia	0.373	0.295	0.991	0.070
20	Sardegna	0.485	0.189	0.841	0.026
12	Lazio	0.220	0.180	0.826	0.122
1	Piemonte	0.203	0.094	0.647	0.075
2	Valle d'Aosta	0.905	0.040	0.642	0.002
15	Campania	0.139	0.032	0.471	0.054
11	Marche	0.222	0.031	0.946	0.021
6	Friuli VG	0.165	0.028	0.704	0.034
9	Toscana	0.107	0.028	0.721	0.079
16	Puglia	0.170	0.024	0.857	0.028
7	Liguria	0.114	0.015	0.568	0.038
10	Umbria	0.134	0.007	0.515	0.014

TAVOLA 38 - Piano fattoriale 2, 3.

Le grandezze riportate in questa Tavola sono le stesse di TAV. 36. Sono elencati soltanto i profili con una Qualità di rappresentazione $COS^2 > 0.470$, così da garantire che le distanze tra proiezioni traducano in modo soddisfacente le reali distanze tra profili nel loro spazio ambiente.

La mappa relativa a questi profili è riprodotta nella TAV. 39 a fianco.

TAVOLA 39 - Mappa 2, 3.

In questa mappa l'asse orizzontale raffigura il terzo asse fattoriale e quello verticale il secondo. Questa rappresentazione è contraria alla prassi, ma è stata adottata per esigenze di spazio. L'unità di scala è la stessa per entrambi gli assi ed eguale a quella della mappa 1, 2 di TAV. 37.

Analisi della matrice *Spettacoli* di TAV. 2

Piano fattoriale 3, 4

Percentuale d'inerzia $\tau_{(3,4)} = 15.75\%$

	Profilo	Distanza	Contributo	Cos2	Massa
j	$\mathbf{c}_j$	$\sqrt{g_{j3}^2 + g_{j4}^2}$	$CTR_{(3,4)}(\mathbf{c}_j)$	$COS^2_{(3,4)}(\mathbf{c}_j)$	$\overline{r}_j$
5	RIVISTA	0.340	0.359	0.819	0.039
3	MUSICA CLAS.	0.128	0.187	0.447	0.144
4	OPERETTA	0.392	0.156	0.629	0.013
i	$\mathbf{r}_i$	$\sqrt{f_{i3}^2 + f_{i4}^2}$	$CTR_{(3,4)}(\mathbf{r}_i)$	$COS^2_{(3,4)}(\mathbf{r}_i)$	$\overline{c}_i$
15	Campania	0.154	0.102	0.583	0.054
6	Friuli VG	0.176	0.084	0.797	0.034
16	Puglia	0.169	0.062	0.844	0.028
7	Liguria	0.105	0.034	0.481	0.038

TAVOLA 40 - Piano fattoriale 3, 4.

Le grandezze riportate in questa Tavola sono le stesse di TAV. 36. Sono elencati soltanto i profili con una Qualità di rappresentazione COS^2 tale da garantire che le distanze tra proiezioni rappresentino con accettabile fedeltà le reali distanze tra profili nel loro spazio.

La mappa è riprodotta nella Tavola a fianco.

TAVOLA 41 - Mappa 3, 4.

Qui l'asse orizzontale raffigura il terzo asse fattoriale, quello verticale il quarto. Si noti come il profilo Concerti di Musica Classica *sembri* lontano dall'origine perché l'unità di scala è *doppia* di quella delle TAV. 37 e 39. Le mappe che interessano assi di rango elevato raffigurano di norma profili sempre più prossimi all'origine, ossia al profilo medio, dato che il tasso d'inerzia decresce, da $\tau_{(1,2)} = 77.97\%$ a $\tau_{(3,4)} = 15.75\%$. Il software che traccia le mappe non conserva sempre la stessa unità di scala, ma la varia in modo che ogni grafico occupi tutto lo spazio disponibile della pagina.

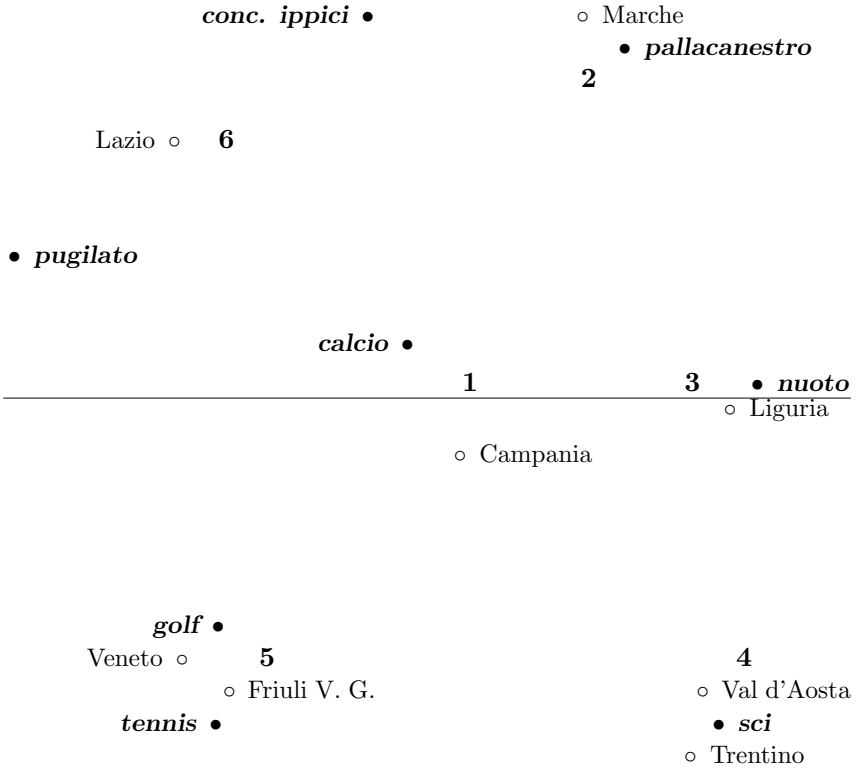

TAVOLA 42 - Interpretazione delle prossimità tra profili.

Questa tavola riassume 6 casi di prossimità tra profili delle righe e delle colonne che si presentano frequentemente nelle mappe fattoriali. I profili derivano da una ipotetica matrice che ripartisce per regione (righe) i biglietti venduti per assistere a manifestazioni sportive (colonne). Tutti i profili hanno una soddisfacente qualità di rappresentazione sul piano e, tranne quelli del caso 1, sono piuttosto eccentrici. I 6 casi sono commentati nell'ultima parte della Sez. 4.11.

$$\mathbf{N} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \qquad \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 \\ 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 \\ 0.0 & 0.0 & 0.5 \end{pmatrix} \quad \overline{\mathbf{c}} = \begin{pmatrix} 0.2 \\ 0.3 \\ 0.3 \\ 0.1 \\ 0.1 \end{pmatrix} \quad \tilde{\mathbf{c}}_1 = \begin{pmatrix} 0.333 \\ 0.333 \\ 0.333 \\ 0.000 \\ 0.000 \end{pmatrix} \quad \tilde{\mathbf{c}}_2 = \begin{pmatrix} 0.0 \\ 0.0 \\ 0.5 \\ 0.5 \\ 0.0 \end{pmatrix}$$

profilo	g_{j1}	g_{j2}	$\overline{r}_j$	COS_1^2	COS_2^2	QLT_2
$\mathbf{c}_1$	-0.500	$+1.936$	0.2	0.062	0.938	1.000
$\mathbf{c}_2$	-0.500	-0.646	0.6	0.375	0.625	1.000
$\mathbf{c}_3$	$+2.000$	$+0.000$	0.2	1.000	0.000	1.000
$\tilde{\mathbf{c}}_1$	-0.500	$+0.215$		0.844	0.156	1.000
$\tilde{\mathbf{c}}_2$	$+0.750$	-0.323		0.241	0.045	0.286

TAVOLA 43 - Profili illustrativi.

La Tavola mostra, in alto, una matrice di contingenza $\mathbf{N}$ di ordine 3×5 e due sue colonne illustrative e, al centro, la matrice dei profili delle colonne, il profilo colonna medio e i profili delle due colonne illustrative. In basso sono riportati i risultati dell'Analisi delle Corrispondenze per i profili delle colonne, attivi ed illustrativi: coordinate sui 2 assi fattoriali, massa dei profili attivi, qualità della rappresentazione sugli assi e sul piano che contiene tutti i profili, eccetto $\tilde{\mathbf{c}}2$ che ha $QLT_2 = 0.286 < 1$. La configurazione è mostrata nella figura.

Profili illustrativi della matrice *Spettacoli*

Piani fattoriali 1, 2 e 2, 3

profilo	distanza	Cos2	distanza	Cos2	qualità
$\tilde{\mathbf{c}}_j$	$\sqrt{g_{j1}^2 + g_{j2}^2}$	$COS^2_{(1,2)}$	$\sqrt{g_{j2}^2 + g_{j3}^2}$	$COS^2_{(2.3)}$	$QLT_7(\tilde{\mathbf{c}}_j)$
Trapani	0.112	[0.145]	0.176	[0.358]	1.000
Palermo	0.364	0.726	0.411	0.926	1.000
Messina	0.345	[0.353]	0.404	0.485	1.000
Agrigento	0.550	0.505	0.547	0.500	1.000
Caltaniss	0.323	[0.336]	0.172	[0.095]	1.000
Enna	0.517	[0.272]	0.605	[0.373]	1.000
Catania	0.521	0.700	0.529	0.722	1.000
Ragusa	0.326	0.791	0.065	[0.031]	1.000
Siracusa	0.260	0.700	0.150	[0.234]	1.000

TAVOLA 44 - Le province siciliane come illustrative.

Profili illustrativi della matrice *Spettacoli*

Piani fattoriali 1, 2 e 2, 3

profilo	distanza	Cos2	distanza	Cos2	qualità
$\tilde{\mathbf{c}}_j$	$\sqrt{g_{j1}^2 + g_{j2}^2}$	$COS^2_{(1,2)}$	$\sqrt{g_{j2}^2 + g_{j3}^2}$	$COS^2_{(2.3)}$	$QLT_7(\tilde{\mathbf{c}}_j)$
Verona	1.532	0.981	0.064	[0.002]	1.000
Vicenza	0.152	[0.351]	0.165	0.411	1.000
Belluno	0.586	0.803	0.530	0.656	1.000
Treviso	0.181	0.459	0.172	0.413	1.000
Venezia	0.107	[0.237]	0.173	0.626	1.000
Padova	0.371	0.536	0.298	[0.346]	1.000
Rovigo	0.214	[0.250]	0.268	[0.390]	1.000

TAVOLA 44 - Le province venete come illustrative.

I valori di COS^2 che indicano una pessima qualità di rappresentazione sul piano sono racchiusi tra [ ]. Tutti i 7 profili illustrativi sono contenuti nello sottospazio $\Re^7$, simplesso dei profili attivi.

TAVOLA 45 - Mappa 1, 2 con profili illustrativi.

La mappa 1, 2 di TAV. 37 è arricchita con i profili delle 9 province siciliane (o venete).

TAVOLA 46 - Mappa asimmetrica della matrice *Biglietti-3*.

In questa rappresemtazione le aree territoriali (le righe) sono rappresentate come vertici in coordinate standard ed i tipi di spettacolo (le colonne) come profili in coordinate principali. È essenziale che anche in questo tipo di mappa entrambi gli assi fattoriali rappresentati sulla mappa, abbiano le medesime unità di scala.

TAVOLA 47 - Distanze tra vertici e profili.

In ordinata è riportata la distanza distribuzionale, al quadrato, tra ii vertice i e il profilo j. In ascissa gli scarti relativi s_{ij} della TAV. XX. Così ad esempio, $s_{27} = 0..$ Le coppie di numeri indicano sinteticamente. le coppie (i, j).

TAVOLA 48 - Mappa asimmetrica della matrice *Biglietti.*

I tipi di spettacolo sono rappresentati come vertici (in coordinate standard) e le regioni come profili (in coordinate principali). Questa mappa asimmetrica va confrontata con quella simmetrica nella TAV. 37. Rispetto ad essa, per motivi di spazio, questa è stata ruotata, per cui qui l'asse 2 è orizzontale e l'asse 1 è verticale. Qui i vertici sono le modalità delle colonne, posizionate in coordinate standard $\hat{\mathbf{g}}_j$, ed i profili le righe, in coordinate principali $\mathbf{f}_i$. Rispetto alla mappa di TAV. 37 si vede una contrazione dei profili rappresentati o una dilatazione dei vertici.

Fasi di calcolo quando $I > J$

FORMULA	RISULTATO
$\mathbf{R}^T\,\mathbf{C}\,\mathbf{v}_a = \lambda_a\,\mathbf{v}_a$	equazione con matrice quadrata
$\mathbf{z}_a = \mathbf{D}_{\mathbf{r}}^{-\frac{1}{2}}\,\mathbf{v}_a$	trasform. per simmetrizzare
$\mathbf{D}_{\mathbf{r}}^{-\frac{1}{2}}\,\mathbf{R}^T\,\mathbf{D}_{\mathbf{c}}\,\mathbf{R}\,\mathbf{D}_{\mathbf{r}}^{-\frac{1}{2}}\,\mathbf{z}_a = \lambda_a\,\mathbf{z}_a$	autovalori λ_a e $\mathbf{z}_a$
$\mathbf{v}_a = \mathbf{D}_{\mathbf{r}}^{\frac{1}{2}}\,\mathbf{z}_a$	autovettori
$\mathbf{g}_a = \sqrt{\lambda_a}\,\mathbf{D}_{\mathbf{r}}^{-\mathbf{1}}\,\mathbf{v}_a$	fattori delle colonne attive
$\mathbf{f}_a = \sqrt{\lambda_a}\,\mathbf{R}\,\mathbf{g}_a$	fattori delle righe attive
$\tilde{f}_a = \frac{1}{\sqrt{\lambda_a}}\,\tilde{\mathbf{r}}^T\,\mathbf{g}_a$	coordinata della riga illustr. $\tilde{\mathbf{r}}$
$\tilde{g}_a = \frac{1}{\sqrt{\lambda_a}}\,\tilde{\mathbf{c}}^T\,\mathbf{f}_a$	coordinata della col. illustr. $\tilde{\mathbf{c}}$

Fasi di calcolo quando $I < J$

FORMULA	RISULTATO
$\mathbf{C}\,\mathbf{R}^T\,\mathbf{u}_a = \lambda_a\,\mathbf{u}_a$	equazione con matrice quadrata
$\mathbf{z}_a = \mathbf{D}_{\mathbf{c}}^{-\frac{1}{2}}\,\mathbf{u}_a$	trasform. per simmetrizzare
$\mathbf{D}_{\mathbf{c}}^{-\frac{1}{2}}\,\mathbf{C}\,\mathbf{D}_{\mathbf{r}}\,\mathbf{C}^T\,\mathbf{D}_{\mathbf{c}}^{-\frac{1}{2}}\,\mathbf{z}_a = \lambda_a\,\mathbf{z}_a$	autovalori λ_a e $\mathbf{z}_a$
$\mathbf{u}_a = \mathbf{D}_{\mathbf{c}}^{\frac{1}{2}}\,\mathbf{z}_a$	autovettori
$\mathbf{f}_a = \sqrt{\lambda_a}\,\mathbf{D}_{\mathbf{c}}^{-\mathbf{1}}\,\mathbf{u}_a$	fattori delle righe attive
$\mathbf{g}_a = \frac{1}{\sqrt{\lambda_a}}\,\mathbf{C}^T\,\mathbf{f}_a$	fattori delle colonne attive
$\tilde{f}_a = \frac{1}{\sqrt{\lambda_a}}\,\tilde{\mathbf{r}}^T\,\mathbf{g}_a$	coordinata della riga illustr. $\tilde{\mathbf{r}}$
$\tilde{g}_a = \frac{1}{\sqrt{\lambda_a}}\,\tilde{\mathbf{c}}^T\,\mathbf{f}_a$	coordinata della col. illustr. $\tilde{\mathbf{c}}$

TAVOLA 49 - Fasi di calcolo dei fattori

Se la matrice di contingenza $\mathbf{N}$ ha più righe che colonne, il tempo di calcolo si riduce diagonalizzando la matrice $\mathbf{R}^T\,\mathbf{C}$ di ordine $J \times J$. In caso contrario si diagonalizza $\mathbf{C}\,\mathbf{R}^T$ che è di ordine $I \times I$. In entrambi i casi si tratta di una matrice quadrata e non simmetrica che viene simmetrizzata per poter utilizzare le routines di diagonalizzazione per matrici simmetriche che sono più semplici, più precise e più veloci. Come risultato si ottengono gli autovalori ed i corrispondenti autovettori. Messa da parte la soluzione banale, le relazioni di transizione (4.9.2) e (4.9.3) forniscono i fattori delle righe e delle colonne attive e permettono di calcolare con la (4.13.1) le coordinate di eventuali righe e/o colonne illustrative.

Indagine sull'ascolto delle trasmissioni radiofoniche
Temi dell'indagine, quesiti e modalità di risposta

A - I programmi che vengono ascoltati

A2 Quali sono i suoi programmi musicali preferiti?
- 1 - musica italiana contemporanea
- 2 - musica italiana revival
- 3 - musica pop o folk contemporanea
- 4 - musica pop o folk revival
- 5 - musica rock contemporanea
- 6 - musica rock revival
- 7 - musica jazz contemporanea
- 8 - musica jazz revival
- 9 - musica lirica e sinfonica

A4 Quali sono i suoi programmi d'informazione preferiti?
- 1 - giornali radio
- 2 - trasmissioni con notizie a carattere locale
- 3 - servizi sportivi
- 4 - programmi culturali
- 5 - grandi dirette (fatti di cronaca, attualità, processi)
- 6 - programmi di economia e finanza
- 7 - nessuno

A6 Quali sono i suoi programmi d'intrattenimento preferiti?
- 1 - giochi e quiz
- 2 - telefonate e interviste in diretta
- 3 - show comici
- 4 - nessuno

TAVOLA 5.1 - Ascolto delle trasmissioni radiofoniche.

In questa Tavola e nelle due che seguono sono elencati 14 quesiti posti a un campione di 400 ascoltatori di trasmissioni radiofoniche in Emilia Romagna. Il sondaggio era volto ad indagare da un lato come e da chi venivano ascoltate le trasmissioni e dall'altro l'atteggiamento degli ascoltatori nei riguardi degli spot pubblicitari. Le domande si riferiscono a 4 aspetti, o temi, dell'indagine, indicati con le lettere A, C, D ed E. L'intervistato poteva scegliere una e una sola delle modalità di risposta elencate.

C - Aspettative e modalità d'ascolto

C1 Cosa fa abitualmente mentre ascolta la radio?
1 - mi lavo, mi vesto (curo la mia persona)
2 - consumo un pasto
3 - studio o lavoro
4 - sono in macchina o su un mezzo pubblico
5 - sono in un locale pubblico
6 - leggo (quotidiani, periodici, libri)
7 - faccio altro

C2 Complessivamente per quanto tempo ascolta la radio nella giornata?
1 - meno di 15 minuti
2 - da 15 a 30 minuti
3 - da mezz'ora a un'ora
4 - da 1 a 3 ore
5 - più di 3 ore

D - Comunicazione pubblicitaria

D1 Ritiene che i suoi coetanei siano influenzati dalla pubblicità radiofonica?
1 - quasi sempre
2 - molto spesso
3 - a volte
4 - raramente
5 - mai

D2 Come giudica la pubblicità radiofonica?
1 - mi infastidisce molto. La eliminerei
2 - interrompe troppo i programmi, ma so che è la fonte di sostentamento delle emittenti
3 - è una utile presentazione di prodotti e servizi

D3 Come è la qualità tecnica degli spot radiofonici?
1 - spesso sono realizzati male e quindi poco efficaci
2 - a volte sono originali e quindi facili da ricordare

D4 La musica facilita il ricordo degli spot?
1 - sì
2 - no

TAVOLA 5.2 - Ascolto delle trasmissioni radiofoniche (*seguito*).

D5 Quali prodotti o servizi pubblicizzati ricorda?
1 - prodotti alimentari
2 - prodotti per la casa
3 - abbigliamento
4 - beni durevoli (auto, ciclo, moto, hi-fi, videocamere, ...)
5 - servizi finanziari e assicurativi
6 - pub, ristoranti, discoteche, ...

E - Descrittori demo - sociali

E1 Sesso dell'intervistato:
1 - maschio
2 - femmina

E2 Età dell'intervistato (desunta dall'Anno di nascita):
1 - meno di 18 anni
2 - tra 18 e 25 anni
3 - tra 25 e 30 anni
4 - tra 30 e 40 anni
5 - tra 40 e 50 anni
6 - tra 50 e 60 anni
7 - oltre 60 anni

E3 Titolo di studio:
1 - licenza elementare
2 - licenza media inferiore
3 - licenza media superiore
4 - laurea

E4 Qualifica professionale:
1 - studente
2 - in cerca di occupazione
3 - operaio
4 - impiegato
5 - funzionario, dirigente
6 - imprenditore, libero professionista, agente di commercio
7 - casalinga
8 - insegnante

TAVOLA 5.3 - Ascolto delle trasmissioni radiofoniche (*seguito*).

				$j=$									
	$q=$						$j_q=$						
i	1	2	3		1	2	1	2	3	1	2	3	z_{i+}
1	2	3	1		0	1	0	0	1	1	0	0	3
2	1	3	2		1	0	0	0	1	0	1	0	3
3	1	1	3		1	0	1	0	0	0	0	1	3
4	2	2	2		0	1	0	1	0	0	1	0	3
5	2	3	2		0	1	0	0	1	0	1	0	3
6	1	2	3		1	0	0	1	0	0	0	1	3
7	1	3	2		1	0	0	0	1	0	1	0	3
8	1	2	3		1	0	0	1	0	0	0	1	3
9	2	1	1		0	1	1	0	0	1	0	0	3
10	1	1	1		1	0	1	0	0	1	0	0	3
11	1	3	1		1	0	0	0	1	1	0	0	3
12	2	1	3		0	1	1	0	0	0	0	1	3
13	2	1	1		0	1	1	0	0	1	0	0	3
14	1	3	2		1	0	0	0	1	0	1	0	3
15	2	2	1		0	1	0	1	0	1	0	0	3
				$z_{+j}=$	8	7	5	4	6	6	5	4	45

$$\mathbf{B}=\begin{bmatrix}
8 & 0 & 2 & 2 & 4 & 2 & 3 & 3 \\
0 & 7 & 3 & 2 & 2 & 4 & 2 & 1 \\
2 & 3 & 5 & 0 & 0 & 3 & 0 & 2 \\
2 & 2 & 0 & 4 & 0 & 1 & 1 & 2 \\
4 & 2 & 0 & 0 & 6 & 2 & 4 & 0 \\
2 & 4 & 3 & 1 & 2 & 6 & 0 & 0 \\
3 & 2 & 0 & 1 & 4 & 0 & 5 & 0 \\
3 & 1 & 2 & 2 & 0 & 0 & 0 & 4
\end{bmatrix}
\begin{matrix}
3\times 8 = b_{j+} = b_{+j} \\
3\times 7 = 21 \\
\\
3\times 5 = 15 \\
3\times 4 = 12 \\
3\times 6 = 18 \\
\\
3\times 6 = 18 \\
3\times 5 = 15 \\
3\times 4 = 12
\end{matrix}$$

TAVOLA 5.4 - Codifiche diverse dei risultati di un sondaggio.

I risultati di un ipotetico sondaggio effettuato ponendo a 15 intervistati 3 domande, la prima con 2 e le altre con 3 modalità di risposta, sono presentati in forma compatta (*in alto a sinistra*), in forma disgiuntiva completa (*in alto a destra*) e organizzati in una matrice simmetrica di Burt (*in basso*). A fianco sono riportati i totali marginali.

$$\mathbf{D}_{\bar{\mathbf{c}}} = \begin{bmatrix}
3/45 & 0 & 0 & 0 & \ldots\ldots & 0 & 0 & 0 & 0 \\
0 & 3/45 & 0 & 0 & \ldots\ldots & 0 & 0 & 0 & 0 \\
0 & 0 & 3/45 & 0 & \ldots\ldots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 3/45 & \ldots\ldots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots\ldots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots\ldots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots\ldots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots\ldots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots\ldots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots\ldots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots\ldots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots\ldots & 3/45 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots\ldots & 0 & 3/45 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots\ldots & 0 & 0 & 3/45 & 0 \\
0 & 0 & 0 & 0 & \ldots\ldots & 0 & 0 & 0 & 3/45
\end{bmatrix}$$

$$\mathbf{D}_{\bar{\mathbf{r}}} = \begin{bmatrix}
8/45 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 7/45 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 5/45 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 4/45 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 6/45 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 6/45 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 5/45 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 4/45
\end{bmatrix}$$

TAVOLA 5.5 - Matrici diagonali delle masse

Le matrici diagonali delle masse dei profili delle righe (matrice $\mathbf{D}_{\bar{\mathbf{c}}}$ in alto) e delle colonne (matrice $\mathbf{D}_{\bar{\mathbf{r}}}$ in basso) si riferiscono ai dati fittizi dell'esempio della TAV. 5.4. Le due matrici sono di ordine 15×15 e 8×8 rispettivamente. Trattandosi di due matrici diagonali, le inverse $\mathbf{D}_{\bar{\mathbf{c}}}^{-1}$ e $\mathbf{D}_{\bar{\mathbf{r}}}^{-1}$ hanno per elementi diagonali i valori reciproci: $45/3$ per tutti gli elementi diagonali nella prima e $45/8, 45/7, \ldots$ nella seconda.

$$\mathbf{R} = \begin{bmatrix} 0 & 1/3 & 0 & 0 & 1/3 & 1/3 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 \\ 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 & 1/3 \\ 0 & 1/3 & 0 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 1/3 & 0 & 0 & 1/3 & 0 & 1/3 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 & 0 & 0 & 1/3 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 & 0 & 0 & 1/3 \\ 0 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 0 & 1/3 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 1/3 \\ 0 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \\ 3/45 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1/7 & 0 & 0 & 1/6 & 1/6 & 0 & 0 \\ 1/8 & 0 & 0 & 0 & 1/6 & 0 & 1/5 & 0 \\ 1/8 & 0 & 1/5 & 0 & 0 & 0 & 0 & 1/4 \\ 0 & 1/7 & 0 & 1/4 & 0 & 0 & 1/5 & 0 \\ 0 & 1/7 & 0 & 0 & 1/6 & 0 & 1/5 & 0 \\ 1/8 & 0 & 0 & 1/4 & 0 & 0 & 0 & 1/4 \\ 1/8 & 0 & 0 & 0 & 1/6 & 0 & 1/5 & 0 \\ 1/8 & 0 & 0 & 1/4 & 0 & 0 & 0 & 1/4 \\ 0 & 1/7 & 1/5 & 0 & 0 & 1/6 & 0 & 0 \\ 1/8 & 0 & 1/5 & 0 & 0 & 1/6 & 0 & 0 \\ 1/8 & 0 & 0 & 0 & 1/6 & 1/6 & 0 & 0 \\ 0 & 1/7 & 1/5 & 0 & 0 & 0 & 0 & 1/4 \\ 0 & 1/7 & 1/5 & 0 & 0 & 1/6 & 0 & 0 \\ 1/8 & 0 & 0 & 0 & 1/6 & 0 & 1/5 & 0 \\ 0 & 1/7 & 0 & 1/4 & 0 & 1/6 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 8/45 & 7/45 & 5/45 & 4/45 & 6/45 & 6/45 & 5/45 & 4/45 \end{bmatrix}$$

TAVOLA 5.6 - Matrici R e C dei profili

Profili delle righe e delle colonne ottenuti dalla tabella 15×8 dei dati codificati in forma disgiuntiva completa nella TAV. 5.4, in cui le righe (gli individui) sono 15 e le colonne (le modalità) sono 8. Sono riportate anche le masse dei profili, ricavate dai totali marginali della medesima tabella.

$$\mathbf{P}_r = \left[\begin{array}{cccccccc} 8/24 & 0 & 2/24 & 2/24 & 4/24 & 2/24 & 3/24 & 3/24 \\ 0 & 7/21 & 3/21 & 2/21 & 2/21 & 4/21 & 2/21 & 1/21 \\ \hline 2/15 & 3/15 & 5/15 & 0 & 0 & 3/15 & 0 & 2/15 \\ 2/12 & 2/12 & 0 & 4/12 & 0 & 1/12 & 1/12 & 2/12 \\ 4/18 & 2/18 & 0 & 0 & 6/18 & 2/18 & 4/18 & 0 \\ \hline 2/18 & 4/18 & 3/18 & 1/18 & 2/18 & 6/18 & 0 & 0 \\ 3/15 & 2/15 & 0 & 1/15 & 4/15 & 0 & 5/15 & 0 \\ 3/12 & 1/12 & 2/15 & 2/12 & 0 & 0 & 0 & 4/12 \end{array}\right]$$

$$\mathbf{P}_c = \left[\begin{array}{cccccccc} 8/24 & 0 & 2/15 & 2/12 & 4/18 & 2/18 & 3/15 & 3/12 \\ 0 & 7/21 & 3/15 & 2/12 & 2/18 & 4/18 & 2/15 & 1/12 \\ \hline 2/24 & 3/21 & 5/15 & 0 & 0 & 3/18 & 0 & 2/12 \\ 2/24 & 2/21 & 0 & 4/12 & 0 & 1/18 & 1/15 & 2/12 \\ 4/24 & 2/21 & 0 & 0 & 6/18 & 2/18 & 4/15 & 0 \\ \hline 2/24 & 4/21 & 3/15 & 1/12 & 2/18 & 6/18 & 0 & 0 \\ 3/24 & 2/21 & 0 & 1/12 & 4/18 & 0 & 5/15 & 0 \\ 3/24 & 1/21 & 2/15 & 2/12 & 0 & 0 & 0 & 4/12 \end{array}\right]$$

TAVOLA 5.7 - Matrici dei profili di B

A causa della simmetria della matrice di Burt della TAV. 5.4, le matrici dei profili delle sue righe, $\mathbf{P}_r$ (*in alto*), e delle sue colonne, $\mathbf{P}_c$ (*in basso*), sono l'una la trasposta dell'altra: le righe di $\mathbf{P}_r$ sono le colonne di $\mathbf{P}_c$. Si noti come il profilo di una modalità j sia la concatenazione dei Q profili, della stessa modalità, ottenuti dalle Q matrici di contingenza che incrociano due a due tutte le Q variabili. Questo deriva dal fatto che, a meno di un fattore Q, il totale marginale della colonna j è il medesimo per la tabella di indicatori in forma disgiuntiva completa e per la matrice di Burt.

Indagine sull'ascolto delle trasmissioni radio

i	Età	Classe	Codifica disgiuntiva completa						
1	28	3	0	0	1	0	0	0	0
2	18	2	0	1	0	0	0	0	0
3	37	4	0	0	0	1	0	0	0
4	51	5	0	0	0	0	0	1	0
5	55	7	0	0	0	0	0	1	0
6	21	2	0	1	0	0	0	0	0
7	16	1	1	0	0	0	0	0	0
8	23	2	0	1	0	0	0	0	0
9	60	7	0	0	0	0	0	0	1
10	42	5	0	0	0	0	1	0	0
11	66	7	0	0	0	0	0	0	1
12	78	7	0	0	0	0	0	0	1
13	17	1	1	0	0	0	0	0	0
14	74	7	0	0	0	0	0	0	1
15	38	4	0	0	0	1	0	0	0

Età	18	25	30	40	50	60
Classe	1	2	3	4	5	6	7

TAVOLA 5.8 - Codifica di una variabile numerica.

La variabile numerica attiva 'Età dell'intervistato', rilevata nell'indagine sull'ascolto delle trasmissioni radiofoniche descritta nella Sez. 5.3, è trasformata in una variabile categorica a 7 modalità. I valori qui riportati si riferiscono ai primi 15 dei 400 intervistati nell'indagine. L'intervallo di variazione delle 400 età rilevate è suddiviso in 7 fascie, o classi di valori, a ciascuna delle quali viene fatta corrispondere una modalità: 1 - 'Fino a 18 anni', 2 - 'Tra 18 e 24 anni', fino all'ultima 7 - 'Da 60 anni in poi'.

La trasformazione, o codifica, produce una variabile categorica ordinale, perché vi è un intrinseco ordine di successione tra le 7 modalità. L'ACM non tiene conto dell'ordine e considera la variabile come nominale pura.

Indagine sull'ascolto delle trasmissioni radio

Scomposizione dell'inerzia $In_{\bar{\mathbf{r}}}$ e $In_{\bar{\mathbf{c}}}$ lungo gli assi fattoriali

Rango	Inerzia	Percento	Cumulo	Diagramma delle percentuali
a	λ_a	τ_a	$\sum_a \tau_a$	3 6 9 12
1	0.5675	13.35 %	13.35 %	
2	0.4636	10.91 %	24.26 %	
3	0.3794	8.93 %	33.19 %	
4	0.3680	8.66 %	41.85 %	
5	0.3060	7.20 %	49.05 %	
6	0.2968	6.98 %	56.03 %	
7	0.2714	6.39 %	62.42 %	
8	0.2598	6.11 %	68.53 %	
9	0.2363	5.56 %	74.09 %	
10	0.2310	5.44 %	79.52 %	
11	0.2185	5.14 %	84.67 %	
12	0.1740	4.09 %	88.76 %	
13	0.1592	3.75 %	92.51 %	
14	0.1330	3.13 %	95.63 %	
15	0.0992	2.33 %	97.97 %	
16	0.0587	1.38 %	99.35 %	
17	0.0277	0.65 %	100.00 %	
$In_{\bar{\mathbf{r}}} =$				
$In_{\bar{\mathbf{c}}} =$	4.2500	100.00 %		

TAVOLA 5.9 - Tavola delle Inerzie.

La ripartizione delle inerzie sugli assi fattoriali è sempre il primo risultato stampato da un programma di Analisi delle Corrispondenze. In questo esempio, il numero di assi fattoriali, è $A = J - Q = 21 - 4 = 17$, mentre l'inerzia totale, somma delle inerzie λ_a sugli assi fattoriali è $In_{\bar{\mathbf{r}}} = In_{\bar{\mathbf{c}}} = J/Q - 1 = 21/4 - 1 = 4.25$.

Si noti la caratteristica attenuazione 'a mensola' delle inerzie all'aumentare del rango dell'asse. Questo andamento è tipico dell'ACM.

Le inerzie e le percentuali d'inerzia sugli assi sono indicatori pessimistici della dispersione delle proiezioni dei profili sugli assi. L'importanza dei primi assi fattoriali è superiore a quanto indicato.

Indagine sull'ascolto delle trasmissioni radio
Variabili attive

	modalità	massa	coordinate		contributi		coseni quad.	
			1^o	2^o	1^o	2^o	1^o	2^o
E1 - Sesso dell'intervistato								
1	*maschio*	13.38	+0.21	−0.31	1.0	2.8	0.05	0.11
2	*femmina*	11.62	−0.24	+0.36	1.2	3.2	0.05	0.11
	contibuto della variabile E1 =				*2.2*	*5.9*		
E2 - Età dell'intervistato								
1	*meno di* 18	2.81	−1.75	−0.51	15.1	1.6	0.39	0.03
2	*tra 18 e 25*	4.06	−0.59	−1.22	2.5	13.1	0.07	0.29
3	*tra 25 e 30*	4.19	+0.40	−0.52	1.2	2.4	0.03	0.05
4	*tra 30 e 40*	3.62	+0.89	−0.09	5.1	0.1	0.13	0.00
5	*tra 40 e 50*	3.50	+0.76	+0.30	3.5	0.7	0.09	0.01
6	*tra 50 e 60*	3.38	+0.66	+0.72	2.6	3.8	0.07	0.08
7	*oltre 60*	3.44	−0.71	+1.58	3.0	18.5	0.08	0.40
	contibuto della variabile E2 =				*33.0*	*40.2*		
E3 - Titolo di studio								
1	*licen. elementare*	2.19	−1.72	+0.70	11.4	2.3	0.28	0.05
2	*licen. media inf.*	7.25	−0.73	+0.30	6.8	1.4	0.22	0.04
3	*licen. media sup.*	11.56	+0.37	−0.47	2.7	5.6	0.12	0.19
4	*laurea*	4.00	+1.20	+0.44	10.2	1.7	0.28	0.04
	contibuto della variabile E3 =				*31.2*	*10.9*		
E4 - Qualifica professionale								
1	*studente*	7.25	−1.04	−0.93	13.8	13.6	0.44	0.35
2	*cerca occupazione*	1.12	+0.66	−0.80	0.9	1.6	0.02	0.03
3	*operaio*	2.12	−0.05	+0.23	0.0	0.2	0.00	0.00
4	*impiegato*	4.75	+0.57	−0.18	2.7	0.3	0.08	0.01
5	*funz., dirigente*	1.50	+1.38	+0.39	5.1	0.5	0.12	0.01
6	*imprend., agente*	4.38	+0.85	+0.20	5.5	0.4	0.15	0.01
7	*casalinga*	2.62	−0.97	+2.05	4.3	23.9	0.11	0.49
8	*insegnante*	1.25	+0.76	+0.96	1.3	2.5	0.03	0.05
	contibuto della variabile E4 =				*33.6*	*42.9*		

TAVOLA 5.10 - Masse, coordinate, contributi e coseni quadrati sui primi due assi fattoriali.

Per le 21 modalità delle 4 variabili attive, la tavola elenca nell'ordine: la massa del profilo, moltiplicata per 100; e, per i primi due assi: i fattori, i contributi relativi all'inerzia dell'asse moltiplicati per 100 - e, in corsivo, i contributi delle 4 variabili, somma dei contributi delle modalità, la cui somma vale quindi 100 - e infine i coseni quadrati o qualità della rappresentazione dei profili attivi.

Indagine sull'ascolto delle trasmissioni radio

Variabili illustrative del tema A: programmi ascoltati

	modalità	**num.**	**coordinate**		**valori test**	
			1^o	2^o	1^o	2^o
	A1 - Programmi musicali preferiti					
1	*musica italiana contemporanea*	129	−0.06	+0.31	−0.8	+4.3
2	*musica italiana revival*	37	−0.25	+1.10	−1.6	+7.0
3	*musica pop folk contemporanea*	61	+0.25	−0.19	+2.1	−1.6
4	*musica pop folk revival*	9	+0.63	−0.13	+1.9	−0.4
5	*musica rock contemporaneo*	121	−0.23	−0.56	−3.1	−7.3
6	*musica rock revival*	11	+0.49	−0.62	+1.6	−2.1
7	*musica jazz contemporaneo*	19	+0.65	−0.11	+2.9	−0.5
8	*musica jazz revival*	6	+1.07	+0.20	+2.6	+0.5
9	*musica lirica, sinfonica*	7	−0.03	+1.01	−0.1	+2.7
	A4 - Programmi d'informazione preferiti					
1	*giornali radio*	98	+0.12	+0.15	+1.4	+1.7
2	*notiziari locali*	50	+0.09	+0.41	+0.7	+3.1
3	*servizi sportivi*	40	+0.22	−0.09	+1.5	−0.6
4	*programmi culturali*	21	+0.96	+0.30	+4.5	+1.4
5	*grandi dirette, cronache, attualità*	33	+0.12	+0.66	+0.7	+3.9
6	*economia e finanza*	3	+0.70	+0.06	+1.2	+0.1
7	*nessuno di questi*	155	−0.33	−0.39	−5.3	−6.1
	A6 - Programmi di intrattenimento preferiti					
1	*giochi e quiz*	39	−0.46	−0.28	−3.0	−1.8
2	*telefonate e interviste*	49	−0.17	+0.05	−1.3	+0.3
3	*show comici*	69	−0.04	−0.44	−0.3	−4.0
4	*nessuno di questi*	243	+0.12	+0.16	+3.0	+4.0

TAVOLA 5.11 - Frequenze, coordinate e valori test sui primi due assi fattoriali.

Per le 20 modalità delle 3 variabili illustrative del tema A, la tavola elenca nell'ordine: le frequenze della modalità, la cui somma per ogni variabile raggiunge 400 che è il numero di individui intervistati; e, limitatamente ai primi due assi: le coordinate fattoriali dei profili illustrativi e il loro valore test. Quando questo supera 2 si può ritenere la modalità sufficientemente 'significativa' per 'illustrare' l'asse.

Indagine sull'ascolto delle trasmissioni radio
Variabili illustrative dei temi C e D

	modalità	num.	coordinate		valori test	
			1^o	2^o	1^o	2^o
	C1 - Cosa fa mentre ascolta la radio					
1	*si lava o si veste*	67	−0.12	−0.41	−1.1	−3.7
2	*consuma un pasto*	18	+0.19	+0.55	+0.8	+2.4
3	*studia o e lavora*	126	−0.33	+0.03	−4.5	+0.4
4	*è in auto o sul bus*	93	+0.37	−0.16	+4.1	−1.8
5	*in un locale pubblico*	0	+0.00	+0.00	+0.0	+0.0
6	*legge* (*quotidiani*, *periodici*, *libri*)	22	+0.15	+0.49	+0.7	+2.4
7	*fa altro*	74	+0.12	+0.24	+1.1	+2.3
	C2 - Per quanto tempo ascolta la radio					
1	*meno di* 15 *minuti*	2	+0.16	+0.26	+0.2	+0.4
2	*da 15 a 30 minuti*	44	+0.11	+0.06	+0.8	+0.4
3	*da 30 minuti a un'ora*	158	+0.17	+0.09	+2.8	+1.5
4	*da 1 a 3 ore*	117	−0.20	−0.01	−2.5	−0.2
5	*oltre 3 ore*	79	−0.10	−0.20	−1.0	−1.9
	D1 - Coetanei influenzati dalla pubblicità					
1	*quasi sempre*	16	+0.23	−0.20	+0.9	−0.8
2	*molto spesso*	60	−0.25	−0.15	−2.1	−1.3
3	*a volte*	150	−0.19	−0.20	−3.0	−3.1
4	*di rado*	112	+0.25	+0.02	+3.1	+0.3
5	*mai*	62	+0.19	+0.65	+1.6	+5.6
	D2 - Giudizio sulla pubblicità radiofonica					
1	*infastidisce e la eliminerei*	119	+0.08	−0.04	+1.1	−0.5
2	*interrompe*, *ma sostenta l'emittente*	231	−0.01	−0.09	−0.1	−2.0
3	*utile presentaz. di prodotti servizi*	50	−0.17	+0.50	−1.3	+3.8

TAVOLA 5.12 - Frequenze, coordinate, e valori test sui primi due assi fattoriali.

Per ogni modalità delle 2 variabili illustrative del tema C - Aspettative e modalità di ascolto, e per le prime due del tema D - Comunicazione pubblicitaria, la tavola elenca nell'ordine: il numero di individui che l'hanno indicata, e, per i primi due assi fattoriali: la coordinata del suo profilo sull'asse e il valore test. La modalità 'C15 - in un locale pubblico' viene abbandonata perché non è stata indicata da nessun intervistato.

Indagine sull'ascolto delle trasmissioni radio

Variabili illustrative del tema D: programmi ascoltati

	modalità	**num.**	**coordinate**		**valori test**	
			1^o	2^o	1^o	2^o
D3 - Qualità tecnica degli spot						
1	*realizzati male e poco efficaci*	175	+0.19	+0.10	+3.3	+1.7
2	*originali e facili da ricordare*	225	−0.15	−0.08	−3.3	−1.7
D4 - la musica facilita il ricordo						
1	*si'*	345	+0.00	+0.00	−0.1	−0.2
2	*no*	55	+0.01	+0.02	+0.1	+0.2
D5 - Prodotti e servizi che ricorda						
1	*prodotti alimentari*	32	−0.55	+0.90	−3.2	+5.3
2	*prodotti per la casa*	31	−0.03	+0.55	−0.1	+3.2
3	*abbigliamento*	53	−0.09	+0.16	−0.7	+1.2
4	*auto, bici, moto, hi − fi, video*	118	+0.22	−0.06	+2.9	−0.8
5	*servizi finanziari e assicurativi*	15	+1.14	+0.30	+4.5	+1.2
6	*pub, ristoranti, discoteche*	151	−0.13	−0.34	−2.1	−5.3

TAVOLA 5.13 - Frequenze, coordinate e valori test sui primi due assi fattoriali (*seguito*).

Per le 10 modalità delle ultime 3 variabili illustrative del tema illustrativo D - Comunicazione pubblicitaria, la tavola elenca nell'ordine: le frequenze della modalità, la cui somma per ogni variabile raggiunge 400 che è il numero di individui intervistati; e, limitatamente ai primi due assi: le coordinate fattoriali dei profili illustrativi e i loro valori test. Si noti come i valori test di una variabile a due modalità siano uno l'opposto dell'altro. Così, per la variabile D3 - Qualità tecnica degli spot, i valori test delle due modalità sono rispettivamente +3.3 e -3.3 sul primo asse. Questa è una conseguenza della proprietà baricentrica dei profili delle modalità di una stessa variabile, come è mostrato nella Sez. 5.18.

Indagine sull'ascolto delle trasmissioni radio
Variabili attive

TAVOLA 5.14 - Rappresentazione delle variabili.

Mappa dei rapporti di correlazione delle 4 variabili attive per i primi due assi fattoriali: il primo orizzontale, il secondo verticale. Il rapporto di correlazione è il rapporto tra l'inerzia sull'asse dovuta ai profili di una variabile e l'inerzia sull'asse di tutti i profili. Può variare tra 0, nessun legame tra variabile e asse, e 1 quando il legame è invece molto stretto.

La mappa rivela che la variabile '1 - Sesso' appare poco legata ai primi due assi e che le altre 3 sono legate al primo, ma solo la '2 - Età' e la '4 - Qualifica' sono legate anche al secondo. La prossimità fra questi due punti indica anche che i segmenti di individui stabiliti dalle due variabili sono molto simili.

Indagine sull'ascolto delle trasmissioni radio

TAVOLA 5.15 - Mappa principale delle modalità attive.

Sul piano fattoriale individuato dai primi due assi sono proiettate le 21 modalità del solo tema attivo 'E - Descrittori demo-sociali'. Appare, ad esempio, che il segmento dei 'Funzionari e Dirigenti' intervistati è costituito per la maggior parte da 'laureati' perché i loro profili sono vicini, ma appartengono a variabili diverse. Invece, i due segmenti vicini, 'Funzionari e Dirigenti' ed 'Imprenditori e Liberi professionisti', costituiti da individui necessariamente diversi perché appartenenti alla stessa variabile, hanno in gran parte in comune le stesse modalità delle altre tre variabili socio-demografiche. In questo caso: un'età da 30 a 60 anni, una laurea e una leggera prevalenza di maschi.

Indagine sull'ascolto delle trasmissioni radio

TAVOLA 5.16 - Mappa principale con modalità attive e illustrative.

Per evitare eccessivi affollamenti sulle mappe, conviene proiettare soltanto le modalità di un tema illustrativo per volta. Qui, sulla mappa principale della Tav. 5.15, sono rappresentati anche i profili illustrativi del tema 'A - Programmi che vengono ascoltati', ma di questi soltanto il 60% : quelli con il Valore-test più elevato sul primo asse fattoriale. Da 20 profili ci si riduce così a 12 . Si noti il mutare delle preferenze musicali degli intervistati con l'avanzare dell'età.

TAVOLA 5.17 - Forme tipiche delle spezzate su una mappa.

Per meglio guidare l'occhio e facilitare quindi l'interpretazione di una mappa fattoriale dell'ACM, i punti rappresentativi dei profili delle modalità ordinate di una variabile vengono collegati in successione con un tratto. La poligonale che si ottiene, detta *spezzata*, risulta sempre convessa perché il baricentro delle modalità di ogni singola variabile coincide col baricentro di tutte le modalità: l'origine degli assi fattoriali.

La Tavola presenta quattro tipi di ipotetiche spezzate che si presentano frequentemente nelle mappe fattoriali dell'ACM. La loro interpretazione si trova nella Sez. 5.20.

Indagine sull'ascolto delle trasmissioni radio

TAVOLA 5.18 - Mappa principale con individui e modalità.

Questo tipo di mappa è utile per farsi un'idea della configurazione geometrica della nuvola dei profili degli individui: le righe della matrice **R**. Insieme agli individui è spesso utile proiettare anche le modalità di una variabile, in questo caso del 'Titolo di studio'.

La lettura della mappa è immediata, dato che ogni modalità rappresenta l'individuo medio di un segmento. Scendendo dall'alto a sinistra in basso a destra, si arriva gradualmente agli individui più istruiti. I laureati stanno a parte e costituiscono il 16% degli intervistati.

Se è disponibile il software adatto, è sempre istruttivo esaminare la configurazione della nuvola in uno spazio tridimensionale.

Teorema di Huygens

TAVOLA 6.1 - Teorema di Huygens per una partizione.

Il teorema di scomposizione dell'inerzia può applicarsi alla partizione in gruppi di una nuvola, composta in questo caso da 5 profili dotati di massa, ed immersi in uno spazio bidimensionale. L'inerzia *complessiva* , in alto, è ottenuta sommando i 5 prodotti delle masse per i quadrati delle distanze dei profili dal baricentro della nuvola. Nelle figure in basso, la nuvola è stata ripartita in due gruppi. L'inerzia *nei* due gruppi, a sinistra, si ottiene sommando le due inerzie interne ai gruppi, ottenute come prodotto della massa per la distanza di ciascun profilo dal baricentro del gruppo a cui appartiene. L'inerzia *tra* gruppi, a destra, è data dal prodotto della massa di ciascun gruppo per la distanza tra baricentro del gruppo e baricentro della nuvola. Il teorema di Huygens stabilisce che l'inerzia *complessiva* è eguale alla somma dell'inerzia *nei* due gruppi e dell'inerzia *tra* i due gruppi e il baricentro della nuvola.

Metodo di aggregazione ai Centri Mobili

			F*		**I** - *Passi*					**II** - *Passi*	
i	*Regione*	*Massa*	f_{i1}	f_{i2}	1^o	2^o	3^o	4^o	5^o	1^o	2^o
1	Piemonte	0.075	+0.008	+0.194	4	4	2	2	2	3	3
2	Valle Ao.	0.002	+0.305	+0.867	1	1	1	1	1	*1*	1
3	Lombardia	0.178	+0.069	−0.015	2	2	2	2	2	3	3
4	Trent. AA	0.026	+0.222	+0.080	*1*	1	2	2	2	3	3
5	Veneto	0.090	−0.584	+0.028	4	4	4	4	4	3	3
6	Friuli VG	0.034	−0.030	+0.030	4	2	2	2	2	3	3
7	Liguria	0.038	+0.010	−0.046	2	2	2	2	2	3	3
8	Emilia R.	0.105	+0.103	+0.045	*2*	2	2	2	2	3	3
9	Toscana	0.079	−0.021	+0.093	4	4	2	2	2	3	3
10	Umbria	0.014	+0.064	−0.097	2	2	2	2	2	2	2
11	Marche	0.021	+0.022	+0.210	4	4	1	2	2	3	3
12	Lazio	0.122	+0.060	−0.188	3	3	3	3	3	2	2
13	Abruzzo	0.018	+0.219	−0.062	*3*	2	2	2	2	2	2
14	Molise	0.002	+0.112	−0.147	3	3	3	3	3	*2*	2
15	Campania	0.054	+0.070	+0.020	2	2	2	2	2	3	3
16	Puglia	0.028	+0.064	+0.029	2	2	2	2	2	*3*	3
17	Basilicata	0.004	+0.239	−0.034	3	1	2	2	2	2	2
18	Calabria	0.014	+0.061	+0.102	*4*	2	2	2	2	3	3
19	Sicilia	0.070	+0.009	−0.340	3	3	3	3	3	*4*	4
20	Sardegna	0.026	+0.160	+0.484	4	1	1	1	1	1	1

TAVOLA 6.2 - Masse, coordinate e partizioni.

I profili delle 20 regioni italiane, costituiti dalle coordinate f_{i1} e f_{i2} sui primi due assi fattoriali e dalle masse $\overline{c}_i$, sono presi come esempio per illustrare il metodo di aggregazione a centri mobili.

Nella parte destra sono presentate le partizioni ottenute con due processi indipendenti di aggregazione, imponendo in entrambe 4 centri. Il primo, indicato con **I** converge dopo 5 passi, il secondo, indicato con **II** dopo due soltanto, ma ad una partizione diversa dalla precedente. Ciò dipende dai profili scelti inizialmente come centri ed indicati in corsivo nel primo passo dei due processi.

Metodo di aggregazione ai Centri Mobili

Partizione ottenuta al 1^o passo

		Distanze dai centri				
i	*Regione*	*1*	*2*	*3*	*4*	*Gruppo*
1	Piemonte	0.059	0.031	0.110	0.011	4
2	Valle Aosta	0.626	0.716	0.870	0.645	1
3	Lombardia	0.032	0.005	0.025	0.014	2
4	*Trent. AA (1)*	*0.000*	0.015	0.020	0.026	*1*
5	Veneto	0.652	0.472	0.653	0.422	4
6	Friuli VG	0.066	0.018	0.070	0.013	4
7	Liguria	0.061	0.017	0.044	0.025	2
8	*Emilia R. (2)*	0.015	*0.000*	0.025	0.005	*2*
9	Toscana	0.059	0.018	0.082	0.007	4
10	Umbria	0.056	0.022	0.025	0.040	2
11	Marche	0.057	0.034	0.113	0.013	4
12	Lazio	0.098	0.056	0.041	0.084	3
13	*Abruzzo (3)*	0.020	0.025	*0.000*	0.052	*3*
14	Molise	0.064	0.037	0.019	0.065	3
15	Campania	0.027	0.002	0.029	0.007	2
16	Puglia	0.028	0.002	0.032	0.005	2
17	Basilicata	0.013	0.025	0.001	0.050	3
18	*Calabria (4)*	0.026	0.005	0.052	*0.000*	*4*
19	Sicilia	0.222	0.157	0.121	0.198	3
20	Sardegna	0.167	0.196	0.302	0.156	4

TAVOLA 6.3 - Risultati al 1^o passo.

La prima colonna riporta l'indice del profilo e la seconda il nome del profilo. I 4 profili estratti casualmente come centri provvisori di aggregazione sono evidenziati in corsivo. Tra parentesi è indicato il numero del centro, ossia del gruppo, che viene assegnato in base all'ordine di estrazione. Le quattro colonne successive riportano le distanze di ciascun profilo dai 4 centri e la quinta il gruppo al quale il profilo è assegnato. Per esempio, il profilo del Piemonte è assegnato al gruppo 4 perché la distanza dal centro di questo gruppo, pari a 0.011 , è la minima delle quattro. L'ultima colonna indica quindi la partizione, provvisoria, dei 20 profili in 4 gruppi che si ha alla conclusione del primo passo del processo iterativo.

Metodo di aggregazione ai Centri Mobili

Partizione ottenuta al 2^o passo

		Distanze dai baricentri				
i	*Regione*	*1*	*2*	*3*	*4*	*Gruppo*
1	Piemonte	0.053	0.041	0.177	0.028	4
2	Valle Ao.	0.555	0.802	1.248	0.747	1
3	Lombardia	0.045	0.000	0.043	0.067	2
4	Trent. AA	0.002	0.029	0.118	0.137	1
5	Veneto	0.667	0.431	0.479	0.203	4
6	Friuli VG	0.075	0.011	0.072	0.023	2
7	Liguria	0.077	0.006	0.034	0.055	2
8	Emilia R.	0.022	0.003	0.074	0.069	2
9	Toscana	0.063	0.017	0.107	0.017	4
10	Umbria	0.076	0.010	0.016	0.095	2
11	Marche	0.049	0.046	0.189	0.034	4
12	Lazio	0.126	0.036	0.001	0.143	3
13	Abruzzo	0.035	0.026	0.051	0.169	2
14	Molise	0.088	0.024	0.008	0.143	3
15	Campania	0.036	0.000	0.059	0.058	2
16	Puglia	0.036	0.001	0.064	0.054	2
17	Basilicata	0.026	0.029	0.067	0.175	1
18	Calabria	0.028	0.010	0.106	0.043	2
19	Sicilia	0.265	0.121	0.016	0.244	3
20	Sardegna	0.133	0.240	0.510	0.219	1

TAVOLA 6.4 - Risultati al 2^o passo.

La Tavola riporta le distanze dei 20 profili dai 4 baricentri dei gruppi individuati nel passo d'aggregazione precedente e che in questo secondo passo vengono considerati come nuovi centri. I profili vengono poi assegnati al gruppo del baricentro più vicino, come indicato nell'ultima colonna. Così il profilo del Piemonte resta nel quarto gruppo perché la distanza dal baricentro del quarto gruppo, 0.028 , è la più piccola delle quattro.

Metodo di aggregazione ai Centri Mobili

TAVOLA 6.5 - Convergenza verso partizioni diverse.

Il metodo di aggregazione a Centri Mobili può condurre ad una partizione finale che dipende dalla scelta iniziale dei centri di aggregazione. In questa rappresentazione schematica di una nuvola di 8 profili di eguale massa (*in alto a sinistra*), al primo passo la scelta a caso di due centri di aggregazione individua i due profili indicati con un una crocetta e col numero del gruppo. I profili della nuvola vengono poi assegnati al centro più vicino: i 4 profili della fila superiore al primo gruppo e i 4 di quella inferiore al secondo. Al secondo passo (*in alto a destra*) sono anzitutto calcolati i baricentri, indicati con una crocetta, dei due gruppi individuati al primo passo e i profili riassegnati al baricentro più vicino, per cui i 4 profili della fila superiore restano nel primo gruppo, quelli della fila inferiore nel secondo. Questa è la partizione finale perché l'assegnazione ai gruppi non è mutata e il processo si arresta.

Se, (*in basso a sinistra*), al primo passo la scelta casuale avesse individuato invece i due profili indicati con una crocetta, i 4 profili di sinistra della nuvola verrebbero assegnati al primo gruppo, i 4 di destra al secondo. Al secondo passo, (*in basso a destra*), la configurazione dei gruppi non cambierebbe e il processo di aggregazione si arresta fornendo una partizione diversa da quella trovata precedentemente.

Incrocio di 2 partizioni a Centri Mobili

	II, 1	**II, 2**	**II, 3**	**II, 4**
I, 1	2 20 .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .
	(**6°**)			
I, 2	. . .	10 13 17	1 3 4	. . .
	. . .	. . .	6 7 8	. . .
	. . .	. . .	9 11 15	. . .
	. . .	. . .	16 18 .	. . .
		(**5°**)	(**1°**)	
I, 3	. . .	12 14 .	. . .	19 . .
	. . .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .
		(**2°**)		**4°**
I, 4	. . .	. . .	5 . .	. . .
	. . .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .
			(**3°**)	

TAVOLA 6.6 - Individuazione dei Gruppi stabili.

La Tavola incrocia due partizioni, indicate con **I** nelle righe e **II** nelle colonne, ottenute con due processi a Centri Mobili a 4 gruppi convergenti in 6 e 2 passi rispettivamente. I profili delle regioni sono indicati con il loro numero d'ordine da 1 a 20. Si vede che i profili 2-Valle d'Aosta e 20-Sardegna si aggregano sempre insieme nel primo gruppo di entrambe le partizioni e formano il (**6°**) gruppo stabile. Invece i gruppi stabili (**3°**) e (**4°**) hanno un solo profilo, perché questi risultano sempre isolati in entrambe le partizioni. Su 16 gruppi stabili possibili, soltanto 6 non sono vuoti. Sono numerati per massa decrescente.

TAVOLA 8.2.1 - Forme tipiche della nuvola di profili

Le configurazioni sulla mappa principale rivelano una matrice di contingenza che può essere ripartita in 2 e 3 blocchi (in alto, a sinistra e a destra) e .....

TAVOLA 8.2.2 - Forme tipiche. *(seguito)*

.....

	Giorni di rilevamento						
Ore	*Lunedì*	*Martedì*	*Mercol.*	*Giovedì*	*Venerdì*	*Sabato*	*Totale*
9	0	14	12	13	16	24	79
10	0	17	19	20	24	31	111
11	0	38	24	24	24	36	146
12	0	27	28	30	22	41	148
13	0	24	20	21	28	19	112
17	13	24	14	15	23	13	102
18	34	19	19	19	15	35	141
19	35	35	23	24	28	31	176
20	32	16	33	36	32	24	173
Tot.	114	214	192	202	212	254	1188

	Giorni di rilevamento						
Ore	*Lunedì*	*Martedì*	*Mercol.*	*Giovedì*	*Venerdì*	*Sabato*	*Totale*
9	*19*	14	12	13	16	24	*98*
10	*26*	17	19	20	24	31	*137*
11	*35*	38	24	24	24	36	*181*
12	*35*	27	28	30	22	41	*183*
13	*27*	24	20	21	28	19	*139*
17	13	24	14	15	23	13	102
18	34	19	19	19	15	35	141
19	35	35	23	24	28	31	176
20	32	16	33	36	32	24	173
Tot.	*256*	214	192	202	212	254	*1330*

TAVOLA 8.5.1 - Zeri strutturali.

La matrice in alto riporta le frequenze di campionamento di una indagine svolta in un ipermercato. Le frequenze si possono ritenere con buona approssimazione proporzionali all'affluenza dei clienti e sono ripartite per fascia oraria (*nelle righe*) e per giorno della settimana (*nelle colonne*). Risulta che tra le 17 e le 18 del lunedì sono stati intervistati 13 clienti. I 5 zeri derivano dal fatto che l'ipermercato resta chiuso nella mattina del lunedì. Questi zeri, non imputabili a fluttuazioni statistiche, sono detti *strutturali*. Per poter effettuare l'analisi della matrice, occorre sostituire agli zeri delle stime dell'affluenza ottenute col metodo illustrato nella Sez. 8.5. Il risultato è la matrice in basso che può ora essere analizzata regolarmente.

	Grado di giudizio			*Affollamento*				
Provin.	*attesa*	*appel.*	*defin.*	*ind.*	$D1$	$D2$	$T1$	$T2$
Verona	152	113	85	1.59	0	1	0	0.172
Vicenza	80	62	91	1.11	1	0	0.115	0
Belluno	24	42	77	1.25	1	0	0.070	0
Treviso	50	36	126	1.58	0	1	0	0.105
Venezia	94	84	182	1.34	1	0	0.178	0
Padova	89	69	497	1.39	0	1	0	0.323
Rovigo	23	17	35	0.00	1	0	0.037	0
Totale	*512*	*423*	*1093*					

		Profili illustrativi			
a	λ_a	$\tilde{\mathbf{g}}_a(D1)$	$\tilde{\mathbf{g}}_a(D2)$	$\tilde{\mathbf{g}}_a(T1)$	$\tilde{\mathbf{g}}_a(T2)$
1	0.137	+0.334	+0.041	+0.344	−0.231
2	0.005	+0.771	−0.492	+0.669	−0.449

TAVOLA 8.6.1 - Profili illustrativi.

La matrice (7×3) presenta la situazione carceraria nel Veneto a fine 1992, ripartendo i detenuti per provincia (*nelle righe*) e per grado di giudizio (*nelle colonne*). Così 23 detenuti nel carcere di Rovigo sono in attesa di giudizio. Come colonna illustrativa si considera l'indice di affollamento (*ind.*), il cui valore mediano è 1.34. Viene creata una nuova variabile a due modalità esclusive, D1 e D2, codificate in forma disgiuntiva completa. Un'ulteriore variabile a due modalità esclusive, T1 e T2, è invece ricavata da D1 e D2 sostituendo ogni 1 con $\bar{c}_i$, massa dei 7 profili attivi delle righe, ossia delle province.

(*segue*)

TAVOLA 8.6.2 - Profili illustrativi. *(seguito)*

Sulla mappa ottenuta dall'analisi della matrice (7×3) della TAV. 8.6.1. sono stati proiettati come illustrativi i profili D1 e D2 codificati con 0 e 1 e T1 e T2 codificati con 0 e $\overline{c}_i$. Soltanto il segmento che collega questi due ultimi punti passa per l'origine che ne è il baricentro, confermando la correttezza della codifica adottata.

	Ospedali					
Ripresa	*H.A*	*H.B*	*H.M*	*H.P*	*H.S*	*Totale*
limitata	13	8	5	21	43	*90*
parziale	18	36	10	56	29	*149*
completa	16	35	16	51	10	*128*
Totale	*47*	*79*	*31*	*128*	*82*	*367*

	Ospedali					
Ripresa	*H.A*	*H.B*	*H.M*	*H.P*	*H.S*	$\bar{\mathbf{c}}$
limitata	0.277	0.101	0.161	0.164	0.524	*0.245*
parziale	0.383	0.456	0.322	0.437	0.354	*0.406*
completa	0.340	0.443	0.516	0.398	0.122	*0.349*
$\bar{\mathbf{r}}$	*0.128*	*0.215*	*0.084*	*0.349*	*0.223*	

Optimal scaling

Pesi 1, 2 e 3

TAVOLA 8.7.1 - Optimal Scaling.

La matrice in alto raccoglie i 367 esiti di un intervento chirurgico al ginocchio, effettuato con una tecnica innovativa in 5 strutture ospedaliere (*nelle colonne*), indicate con delle sigle. Gli esiti sono stati classificati (*nelle righe*) in base a 3 gradi di ripresa funzionale dell'arto operato: nulla o molto limitata, parziale e completa. L'elemento n_{ij} indica il numero di pazienti sottoposti all'intevento con esito i nell'ospedale j.

Più sotto è riportata la matrice $\mathbf{C}$ dei profili delle colonne, il profilo medio $\bar{\mathbf{c}}$ e il profilo $\bar{\mathbf{r}}$ delle masse dei profili. In basso sono confrontati graficamente gli score degli ospedali ottenuti: con l'Optimal Scaling e assegnando pesi equispaziati 1, 2 e 3 ai tre gradi dell'esito. In quest'ultimo caso agli score è stato sottratto il valore 2 per agevolare il confronto. L'unità di scala è la stessa in entrambe le rappresentazioni. Con i pesi equispaziati la varianza ottenuta è 0.078, mentre è $\lambda_1 = 0.149$ con l'Optimal Scaling: gli score sono più 'sparpagliati' sull'asse e la graduatoria dei 5 ospedali risulta più chiara.

	Banche					
Quesito	*B*1	*B*2	*B*3	*B*4	*B*5	*Totale*
Q1	0.43	0.28	0.18	0.07	0.04	*1.00*
Q2	0.45	0.29	0.14	0.06	0.07	*1.00*
Q3	0.42	0.31	0.19	0.08	0.10	*1.00*
Q4	0.38	0.36	0.10	0.09	0.07	*1.00*
Q5	0.44	0.27	0.19	0.04	0.06	*1.00*
$\bar{\mathbf{r}}$	*0.41*	*0.29*	*0.16*	*0.07*	*0.07*	*1.00*

	Banche					
Quesito	*B*1	*B*2	*B*3	*B*4	*B*5	*Totale*
Q1	0.43	0.28	0.18	0.07	0.04	*1.00*
C1	0.57	0.72	0.82	0.93	0.96	*4.00*
Q2	0.45	0.29	0.14	0.06	0.07	*1.00*
C2	0.55	0.71	0.86	0.94	0.93	*4.00*
Q3	0.42	0.31	0.19	0.08	0.10	*1.00*
C3	0.58	0.69	0.81	0.92	0.90	*4.00*
Q4	0.38	0.36	0.10	0.09	0.07	*1.00*
C4	0.62	0.64	0.90	0.91	0.93	*4.00*
Q5	0.44	0.27	0.19	0.04	0.06	*1.00*
C5	0.56	0.73	0.81	0.96	0.94	*4.00*
Totale	*5.00*	*5.00*	*5.00*	*5.00*	*5.00*	*25.00*
$\bar{\mathbf{r}}$	*0.20*	*0.20*	*0.20*	*0.20*	*0.20*	*1.00*

TAVOLA 8.8.1 - Matrici di profili.

La matrice è tratta dai risultati di un'indagine telefonica volta a stabilire quali banche (*nelle colonne*) avessero la migliore immagine presso le grandi imprese. Gli oltre 200 direttori finanziari interpellati dovevano indicare la banca ritenuta migliore per gli aspetti principali del servizio bancario (*nelle righe*). I quesiti posti erano: Quale è secondo Lei la banca che

Q1 - offre i più bassi tassi d'interesse sui prestiti ?
Q2 - è più rapida nel concedere finanziamenti ?
Q3 - è più presente nell'euromercato ?
Q4 - sostiene maggiormente le imprese nel commercio estero ?
Q5 - offre i migliori rendimenti nella gestione del portafoglio ?

segue

TAVOLA 8.8.2 - Matrici di profili *(seguito)*.

I dettagli dell'indagine non sono noti, ma è disponibile la matrice di profili (*nella TAV. 8.8.1, in alto*). Per stabilire con l'Optimal Scaling quale banca abbia la migliore immagine tra le imprese, conviene dare a tutte le banche la medesima importanza, ossia la stessa massa, creando 5 profili complementari di risposta (*in basso nella TAV. 8.8.1*) indicati con C1 - C5. La mappa principale (*qui sopra*) ottenuta da questa matrice mostra i profili complementari riuniti nella parte negativa del primo asse fattoriale, indicando che B1 è la banca meglio classificata con lo score più elevato: $g_1(B1) = +0.55$.

			Scuole			
i	*Apprezzamento*		*Medie*	*Ist.prof.*	*Licei*	*Totale*
1	Semin.	basso	20	15	2	
2		medio	20	16	25	
3		alto	0	28	37	*163*
4	Mostra	basso	16	0	0	
5		medio	20	35	11	
6		alto	4	23	53	*162*
7	Visita	basso	8	2	4	
8		medio	31	24	15	
9		alto	1	35	45	*165*
	Totale		*120*	*178*	*192*	*490*

Analisi dei profili delle righe della matrice 9×3 *:*

i	*Apprezzamento*	$\overline{c}_i$	f_{i1}	CTR_{i1}	COS^2_{i1}	f_{i2}	CTR_{i2}	COS^2_{i2}
1	Semin. basso	0.03	+1.68	0.27	0.91	+0.52	0.17	0.09
2	medio	0.13	+0.26	0.03	0.31	−0.39	0.40	0.69
3	alto	0.16	−0.55	0.14	0.84	+0.24	0.18	0.16
4	Mostra basso	0.07	+0.78	0.13	0.94	−0.20	0.06	0.06
5	medio	0.12	+0.14	0.01	0.36	+0.19	0.09	0.64
6	alto	0.13	−0.58	0.13	0.99	−0.06	0.01	0.01
7	Visita basso	0.03	+0.68	0.04	0.78	+0.37	0.07	0.22
8	medio	0.14	+0.49	0.10	0.99	−0.03	0.00	0.01
9	alto	0.16	−0.55	0.15	0.99	−0.06	0.01	0.01

Analisi dei profili delle colonne della matrice 9×3 *:*

j	*Apprezzamento*	$\overline{r}_j$	g_{j1}	CTR_{j1}	COS^2_{j1}	g_{j2}	CTR_{j2}	COS^2_{j2}
1	Medie	0.25	+0.98	0.69	0.98	+0.12	0.06	0.01
2	Ist. Prof.	0.36	−0.11	0.01	0.12	−0.30	0.62	0.88
3	Licei	0.39	−0.51	0.30	0.86	+0.20	0.31	0.14

TAVOLA 8.9.1 - Matrici concatenate.

.....

L'elevato valore di λ_1 indica che . . . . . .
. . . .

TAVOLA 8.9.2 - Matrici concatenate *(seguito)***.**

.....