8.1 Sintesi, descrizione, interpretazione

Molte diffuse tecniche di analisi statistica multivariata consentono di studiare simultaneamente un numero *elevato* di variabili sintetizzandone l'azione sinergica attraverso un numero *ridotto* di combinazioni lineari indipendenti. Vengono cioè individuate *nuove* variabili sottostanti all'insieme iniziale dei dati, ottenute come combinazioni lineari di quelle di partenza.

Se le nuove variabili spiegano buona parte della variabilità complessiva del sistema originario, diventa estremamente vantaggioso passare dalla difficile e spesso impossibile analisi simultanea di molte variabili al più accessibile studio di poche variabili capaci di sintetizzare statisticamente (in termini di variabilità) il fenomeno studiato.

Questo genere di analisi multivariata si caratterizza quindi per due fondamentali aspetti:

a) la *descrizione* di fenomeni complessi attraverso una necessaria riduzione dei dati, che corrisponde a una sensibile riduzione della dimensione dello spazio geometrico di riferimento;
b) l'*interpretazione* del nuovo sistema statistico ottenuto.

Il problema dell'interpretazione costituisce la parte operativamente più critica dei metodi di analisi multivariata: il *significato* che viene attribuito alle nuove variabili individuate rientra nella sfera delle *conoscenze* e delle *scelte* soggettive e come tale è suscettibile di variazioni in base al tipo di ricerca e ai ricercatori stessi.

A proposito del ruolo fondamentale e irrinunciabile, in qualsiasi contesto statistico, dell'interpretazione, Italo Scardovi e Paola Monari hanno scritto [Scardovi e Monari 1993]:

"Il calcolo statistico dà soltanto una sintesi quantitativa del variare di due o più grandezze, derivi esso da un legame causale diretto o rifletta invece una più complessa interazione. Ora, se l'analisi del *come* due o più variabili appaiono connesse invita alla ricerca del *perché*, non per questo i metodi […] possono tener luogo dei contenuti e sostituirsi alla conoscenza profonda della realtà indagata. Altrimenti la correlazione è sterile tautologia, l'analisi statistica buia elaborazione". Ogni risultato statistico "ha valore in quanto aggiusti un modello, integri una teoria, controlli un'ipotesi. Ed è sempre un'idea a dare significato al dato, ragione al calcolo, contenuto al paradigma teorico, ai suoi presupposti logico-formali".

"Correlazione statistica significa soltanto concomitanza di variazioni. Tale concomitanza può attestare influenza univoca" di una variabile sull'altra, "oppure reciproca interazione, o invece dipendenza comune da un terzo fattore: sono quindi implicite e possibili la dipendenza diretta, l'interdipendenza, la co-dipendenza; ma vi è pure la fortuita coincidenza, priva di tutti questi significati: ed essa è tanto più insidiosa quanto più l'osservazione è circoscritta". Ancora, quindi, "la rilevanza dei contenuti. Ancora il distinguo, e il confronto, tra *logos* ed *empeira*: ancora il dialogo della ragion critica con la ragion sperimentale".

"Il ricorso all'analisi statistica non significa mettere a tacere l'intelligenza critica, applicare meccanicamente tecniche più o meno raffinate senza la consapevolezza dei principi da cui discendono, nascondere l'assenza di idee dietro cortine di simboli: la trappola delle fallacie logiche è sempre pronta a scattare".

In tali fallacie "non è difficile cadere quando si assumano le tecniche di calcolo come regole sostitutive del ragionamento, così da fare del metodo statistico una sorta di arte divinatoria più vicina al rituale oscuro dell'aruspice che al ragionare galileiano dello scienziato".

8.2 Le origini della statistica multivariata

I primi metodi di analisi statistica multivariata nacquero tra la fine dell'Ottocento e l'inizio del Novecento.

Il modello di *regressione lineare multipla* fu ripreso da Galton nel 1877 dopo aver mosso i primi passi a inizio secolo grazie agli studi di Laplace e Gauss, che aveva sviluppato il metodo dei minimi quadrati.

Le basi teoriche dell'analisi delle *componenti principali* trovano la loro origine nei tentativi di Galton nel 1889 e di Edgeworth nel 1891 di studiare alcune misure antropometriche attraverso strutture statisticamente indipendenti ottenute come combinazioni lineari delle variabili rilevate. Le stesse soluzioni proposte da Galton ed Edgeworth furono ottenute nel 1901 da Pearson partendo da un problema prettamente geometrico: l'obiettivo era infatti quello di individuare il sistema di rette e piani che potesse approssimare in modo ottimale un insieme di punti in un generico spazio *p*-dimensionale.

Le origini dell'analisi dei *fattori* vanno ricercate nell'ambito della ricerca psicologica di inizio Novecento: la prima formulazione del metodo fu quella di Spearman nel 1904.

Risalgono, infine, al 1936 sia l'analisi *discriminante*, introdotta da Fisher, che l'analisi della *correlazione canonica*, introdotta da Hotelling [Mignani e Montanari 1994].

8.3 Le variabili qualitative e l'ACM

Tutti i metodi sinora citati si adattano solamente allo studio di variabili *quantitative*, dotate di una propria metrica, cioè misurabili in modo univoco attraverso una coerente scala di misurazione.

Nello studio dei fenomeni sociali e socio-demografici si ha però spesso a che fare con variabili *qualitative*, che per la loro stessa natura sono difficilmente riconducibili a variabili metriche: è il caso di variabili demografiche di base come sesso e stato civile, delle variabili binarie "presenza/assenza" e di tutte quelle variabili che rilevano opinioni, atteggiamenti e comportamenti.

L'assenza di una precisa scala di misurazione rende impossibile, o perlomeno altamente sconsigliabile (perché richiederebbe strane o assurde trasformazioni delle variabili), il ricorso ai metodi quantitativi. La stessa analisi fattoriale, ampiamente utilizzata nella moderna ricerca psicologica e sociale, non si adatta alla fenomenologia studiata se questa è rappresentata operativamente anche o solo da variabili di natura qualitativa.

L'*analisi delle corrispondenze multiple* [→ Lebart, Morineau e Piron 1995; Mannetti; Monari 1997], brevemente indicata con la sigla *ACM*, risolve questi problemi di metodo in quanto si adatta sia allo studio di insiemi di variabili unicamente di natura qualitativa che allo studio di variabili tipologicamente non omogenee. Opera infatti su variabili qualitative

a) che erano già tali nell'ambito della ricerca a cui ci si riferisce (cioè già concettualizzate, operativizzate e rilevate sotto una veste qualitativa);
b) rese tali trasformando variabili quantitative con opportune aggregazioni in classi o con una ricodifica logica in variabili binarie presenza/assenza.

I fondamenti teorici di questa preziosa tecnica di analisi multivariata vanno principalmente ricercati negli studi di Guttman degli anni Quaranta, nei lavori di Burt e Hayashi degli anni Cinquanta e successivamente nelle estensioni proposte da Benzecri e da Masson.

In termini generali, *l'ACM si propone di descrivere la struttura delle relazioni sottese alla matrice dei dati oggetto di studio attraverso la collocazione e l'analisi dei punti-modalità delle variabili in uno spazio geometrico-statistico di dimensione ridotta*.

Per studiare le correlazioni tra variabili, i metodi quantitativi di analisi multivariata più utilizzati ricorrono al coefficiente di correlazione lineare di Pearson (*r*). Con l'ACM, invece, cambia la struttura matematico-statistica di riferimento. Si passa alla *metrica del* χ^2 (Chi quadro): le relazioni tra le variabili qualitative vengono misurate attraverso la statistica del χ^2, che valuta l'allontanamento delle variabili dalla situazione di *indipendenza*. Fondata su tale metrica, l'ACM prescinde sia dalle unità di misura delle variabili in esame che da qualsiasi assunto riguardante determinate relazioni funzionali tra esse, a differenza di altre tecniche di analisi fattoriale [1].

[1] Solitamente l'assunto teorico di base dei metodi quantitativi riguarda la presenza di relazioni lineari (o comunque riconducibili ad esse tramite opportune trasformazioni) tra le variabili.

8.4 La matrice logico-disgiuntiva e i profili

La tabella originale dei dati corrisponde a una matrice "unità x variabili", di dimensioni n x p, e rappresenta le n unità statistiche analizzate su cui sono state rilevate p variabili.

A partire da questa tabella, l'ACM richiede la costruzione della *matrice logico-disgiuntiva completa* di dimensioni n x q, dove q è il numero complessivo di tutte le modalità delle p variabili. Per ogni unità statistica (riga) si assegna una colonna a ciascuna modalità e si attribuisce

a) valore 0 se l'unità non presenta quella modalità;
b) valore 1 se l'unità presenta quella modalità.

Questo metodo di codifica è appunto

a) *disgiuntivo* perché le diverse modalità di ciascuna variabile si escludono a vicenda: ogni riga presenta un solo 1 per ogni variabile;
b) *completo* perché per tutte prevede l'assegnazione dei valori 0 e 1 per tutte le n unità e le q modalità delle variabili.

Le colonne della matrice logico-disgiuntiva completa individuano dunque *q variabili indicatrici*, che assumono valore 0 o 1 a seconda dell'assenza o della presenza delle rispettive modalità. In questo modo lo studio delle p variabili qualitative di partenza è riconducibile allo studio delle q variabili indicatrici riferite alle modalità, con $q > p$.

Indicato con k_{ij} il termine generico della matrice logico-disgiuntiva, , si ha che

$$\begin{cases} k_{ij}=1 \text{ se l'unità } i\text{-esima possiede la modalità } j\text{-esima;} \\ k_{ij}=0 \text{ altrimenti.} \end{cases} \qquad (8.1)$$

con

$$\begin{cases} i = 1, \ldots, n \\ j = 1, \ldots, q \end{cases} .$$

Inoltre è immediato notare che

$$\sum_{j=1}^{q} k_{ij} = p \ . \tag{8.2}$$

La matrice logico-disgiuntiva può essere facilmente trasformata in una *tavola di frequenze* (Figura 1).

Figura 1 ***Struttura della tavola di frequenze***

	1 , … **j** … , *q*	
1 , … ***i*** … , *n*	f_{ij}	f_i
	f_j	**1**

La frequenza della modalità *j*-esima relativa all'*i*-esima unità statistica viene definita come

$$f_{ij} = \frac{k_{ij}}{n \cdot p} \ . \tag{8.3}$$

In questo modo il totale complessivo della tavola risulta unitario:

$$\sum_{i=1}^{n}\sum_{j=1}^{q} f_{ij} = \frac{1}{n \cdot p}\sum_{i=1}^{n}\sum_{j=1}^{q} k_{ij} = \frac{1}{n \cdot p}\sum_{i=1}^{n} p = \frac{1}{n \cdot p} \cdot n \cdot p = 1 \ . \tag{8.4}$$

I totali marginali di riga f_i corrispondono alle somme per riga delle frequenze f_{ij}:

$$f_i = \sum_{i=1}^{n} f_{ij} \quad . \tag{8.5}$$

Analogamente, i totali marginali di colonna f_j corrispondono alle somme per colonna delle frequenze f_{ij}:

$$f_j = \sum_{j=1}^{q} f_{ij} \quad . \tag{8.6}$$

Ricordando la (8.4), vale la relazione

$$\sum_i \sum_j f_{ij} = \sum_i f_i = \sum_j f_j = 1 \quad . \tag{8.7}$$

I *profili riga* sono definiti dal rapporto tra le frequenze f_{ij} e i totali marginali di riga. Corrispondono quindi, al variare di i, ai vettori di frequenze relative

$$\left(\frac{f_{i1}}{f_i}, \ldots, \frac{f_{ij}}{f_i}, \ldots, \frac{f_{iq}}{f_i} \right) . \tag{8.8}$$

I *profili colonna* sono definiti dal rapporto tra le frequenze f_{ij} e i totali marginali di colonna. Corrispondono quindi, al variare di j, ai vettori di frequenze relative

$$\left(\frac{f_{1j}}{f_j}, \ldots, \frac{f_{ij}}{f_j}, \ldots, \frac{f_{nj}}{f_j} \right) . \tag{8.9}$$

Essendo il totale della tavola pari a 1, il *profilo marginale riga* corrisponde al vettore dei totali marginali di *colonna* [2]:

[2] Diversamente, invece, i totali marginali andrebbero rapportati al totale complessivo.

$$(f_1,...,f_j,...,f_q)\ . \tag{8.10}$$

Analogamente, il *profilo marginale colonna* corrisponde al vettore dei totali marginali di *riga*:

$$(f_1,...,f_i,...,f_n)\ . \tag{8.11}$$

Diventa allora possibile rappresentare geometricamente i *punti-modalità* delle variabili attraverso le coordinate dei rispettivi profili colonna. Parallelamente, i *punti-unità* sono collocabili nello spazio geometrico-statistico attraverso le coordinate dei profili riga.

La definizione di profilo, infatti, pone in collegamento reciproco l'insieme delle unità statistiche e l'insieme delle modalità delle variabili. La *struttura dei profili* consente appunto di studiare la *struttura delle relazioni* sottese a questi due insiemi di interesse.

8.5 Nuvole di punti, centri di gravità e inerzia

La tavola delle frequenze permette di individuare due *nuvole di punti*, rispettivamente nello spazio geometrico-statistico *n*-dimensionale e *q*-dimensionale: N(I) e N(J).

La nuvola N(I) è costituita dai punti dello spazio $\Re^q$ che hanno per coordinate gli elementi dei profili riga:

$$x_{ij} = \frac{f_{ij}}{f_j} \quad . \tag{8.12}$$

A ogni punto x_{ij} è associata la *massa* f_i, che misura l'importanza relativa dell'unità *i*-esima rispetto all'insieme delle informazioni.

Le prossimità tra i punti sono direttamente interpretabili in termini di similarità tra profili: due osservazioni che hanno i profili identici individueranno due punti sovrapposti nella rappresentazione spaziale.

Il *centro di gravità* di N(I) è dato dal punto G avente come coordinate le medie aritmetiche ponderate, al variare di *j*, delle coordinate dei punti della nuvola, con pesi pari alle rispettive masse f_i:

$$g_j = \sum_{i=1}^{n} f_i \cdot x_{ij} \quad . \tag{8.13}$$

È immediato notare che il vettore delle coordinate di G così definito viene a corrispondere al profilo marginale riga. Infatti

$$g_j = \sum_{i=1}^{n} f_i \frac{f_{ij}}{f_i} = \sum_{i=1}^{n} f_{ij} = f_j \quad . \tag{8.14}$$

Quindi il profilo marginale riga è anche il *profilo medio di riga*.

La nuvola N(J) è costituita dai punti dello spazio $\Re^n$ che hanno per coordinate gli elementi dei profili riga:

$$y_{ij} = \frac{f_{ij}}{f_i} \ . \tag{8.15}$$

A ogni punto y_{ij} è associata la *massa* f_j della modalità j-esima, e il centro di gravità di N(J) è dato dal punto H avente come coordinate le medie aritmetiche ponderate, al variare di i, delle coordinate dei punti della nuvola, con pesi pari alle rispettive masse f_j:

$$h_i = \sum_{j=1}^{q} f_j \cdot y_{ij} \ . \tag{8.16}$$

Il vettore delle coordinate di H corrisponde al profilo marginale colonna:

$$h_i = \sum_{j=1}^{q} f_j \frac{f_{ij}}{f_j} = \sum_{j=1}^{q} f_{ij} = f_i = \frac{1}{n} \ . \tag{8.17}$$

Quindi il profilo marginale colonna è anche il *profilo medio di colonna*, pari a $1/n$ (essendo il totale della tavola di frequenze pari a 1 e l'insieme delle unità di numerosità n).

La distribuzione marginale $f_I = (f_i \mid i = 1, \ldots, n)$ è dunque uniforme su tutte le unità statistiche.

La variabilità complessiva di una nuvola di punti è data dalla dispersione dei punti stessi attorno al centro di gravità e prende il nome di *inerzia*, misurata attraverso la *distanza secondo la metrica del* χ^2 *dei profili riga o colonna dai corrispondenti profili marginali (profili medi)*.

La distanza χ^2 tra due generici profili riga *i* e *i'* è definita dalla somma delle distanze tra i loro elementi al quadrato ponderate con l'inverso delle frequenze corrispondenti del profilo marginale di colonna:

$$d^2 = \sum_j \left(\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}} \right)^2 \cdot \frac{1}{f_j} \quad . \tag{8.18}$$

Analogamente, per i profili colonna si ha:

$$d^2 = \sum_i \left(\frac{f_{ij}}{f_j} - \frac{f_{ij'}}{f_{j'}} \right)^2 \cdot \frac{1}{f_i} \quad . \tag{8.19}$$

Ponderando in questo modo si ristabilisce un equilibrio tra le modalità, in quanto:

a) si rivaluta il contributo dato dalle modalità con frequenza più bassa;
b) si ridimensiona il contributo dato dalle modalità con frequenza più alta.

8.6 La matrice di Burt

Se si costruisce la tabella di contingenza che incrocia tra loro tutte le q modalità della matrice logico-disgiuntiva completa si ottiene una tabella di frequenze, di dimensioni $q \times q$, detta *matrice di Burt* o *matrice delle corrispondenze multiple*.

Per ognuno dei $q \times q$ punti di incrocio bivariati viene inserito il numero di unità statistiche che presentano entrambe le modalità, cioè che valgono 1 in entrambe le variabili indicatrici considerate.

La matrice di Burt

a) risulta *simmetrica* rispetto alla diagonale principale. È quindi definita e costruita in modo da poter essere letta sia nel senso delle righe che in quello delle colonne, dando luogo a un'analisi delle relazioni di tipo simmetrico.
b) nelle sottomatrici corrispondenti alle modalità di una stessa variabile presenta esternamente alla diagonale tutti valori nulli, per la già vista proprietà di disgiunzione (nessuna unità statistica può possedere contemporaneamente 2 modalità della stessa variabile).

La matrice di Burt non tiene conto delle informazioni sui singoli soggetti, ma solo delle frequenze delle associazioni (quindi già in forma aggregata); la matrice logico-disgiuntiva completa invece contiene tutte le informazioni della matrice originale dei dati (la "unità x variabili"). Si dimostra tuttavia che i risultati dell'ACM condotta sui due tipi di tabelle sono equivalenti, per cui è conveniente effettuare l'analisi in riferimento alla matrice di Burt.

8.7 La scelta del numero di fattori

Geometricamente le q modalità generano uno spazio a q dimensioni in cui vengono rappresentate le n unità statistiche. Ogni unità è rappresentata da un punto che ha come coordinate i q elementi del suo profilo riga.

Analogamente all'analisi fattoriale, lo scopo dell'ACM è quello di individuare un *sottospazio ottimale* di dimensione $k << q$ [3]

a) definito da k nuove variabili ottenute come combinazioni lineari di quelle di partenza;
b) che abbia come *origine* il centro di gravità del sistema originario.

Attraverso le combinazioni lineari ottimali dei *punti-modalità* delle variabili coinvolte nell'analisi, l'ACM provvede quindi a individuare variabili sintetiche denominate semplicemente *fattori* o *assi fattoriali*, come nell'analisi delle componenti principali e nell'analisi dei fattori.

I fattori

a) sono determinati in modo da essere ortogonali tra loro, cioè *indipendenti* l'uno dall'altro;
b) ordinati dal 1° al k-esimo spiegano in ordine decrescente la variabilità del sistema.

Individuati gli assi fattoriali, diventa possibile rivedere il sistema originario sotto la nuova veste delle *coordinate fattoriali*, che rappresentano le nuove posizioni dei punti-modalità nello spazio geometrico ridotto k-dimensionale.

In parallelo con l'analisi delle componenti principali e con l'analisi dei fattori, si dimostra che la quota di variabilità spiegata da ciascun fattore coincide con l'*autovalore* corrispondente λ

[3] k molto minore di q.

relativo alla *matrice delle distanze* dei punti-modalità dal baricentro del sistema [4].

La somma di tutti i *q* autovalori coincide con l'*inerzia totale* e corrisponde alla *traccia* della matrice, pari a

$$\frac{q-p}{p}$$

vale a dire

$$\frac{q}{p}-1 \quad . \tag{8.20}$$

Quindi il rapporto tra l'autovalore λ_j del fattore *j*-esimo e la somma di tutti gli autovalori individua la proporzione di inerzia spiegata dal fattore stesso:

$$\frac{\lambda_j}{\sum_{j=1}^{q}\lambda_j} \quad . \tag{8.21}$$

Il primo fattore dà la migliore sintesi della matrice dei dati di partenza. L'autovalore ad esso associato è quello più alto e spiega la quota più elevata dell'inerzia totale. Tale quota diminuisce poi progressivamente proseguendo con i fattori successivi al primo.

Aggiungere fattori porta dunque a un guadagno di precisione e aumenta la quota di *inerzia cumulativa* spiegata, ma nello stesso tempo aumenta il numero di dimensioni da interpretare.

Anche se non esistono regole rigide per l'individuazione del numero di fattori da considerare, solitamente il criterio seguito consiste nell'individuare il fattore oltre il quale l'inerzia cumulativa spiegata cumulativa inizia ad aumentare molto lentamente, per esempio al di sotto del 10% per ogni successivo fattore aggiunto.

[4] Questo escludendo il 1° autovalore, detto *banale*, pari a 1.

L'ACM porta a un rapido decremento dell'importanza dei fattori, e il software SPAD fornisce un istogramma che rappresenta graficamente gli autovalori tramite segmenti di lunghezza proporzionale a ciascuno di essi. Diventa così possibile dare una valutazione "a colpo d'occhio" dell'andamento decrescente degli autovalori, in analogia con lo scree-test dell'analisi delle componenti principali. Un autovalore graficamente ben separato da quello che lo precede e da quello che lo segue individua un asse fattoriale ben caratterizzato, che è produttivo interpretare.

Secondo Benzecri, gli autovalori forniti dall'ACM danno una valutazione riduttiva dell'inerzia spiegata dagli assi fattoriali estratti, perché solitamente il numero di modalità delle variabili su cui si opera è molto alto e porta a tenere conto di tanti fattori di importanza infinitesima.

L'autore propone allora una "formula ottimista" per attribuire più importanza ai primi fattori. Il metodo tiene conto solamente degli autovalori maggiori di $1/p$ e i nuovi autovalori λ_j' vengono calcolati in questo modo:

$$\lambda_j' = \left(\frac{p}{p-1}\right)^2 \cdot \left(\lambda_j - \frac{1}{p}\right)^2 \quad . \tag{8.22}$$

La quota di inerzia spiegata da ogni fattore viene così rivalutata sulla base dei λ_j', cioè attraverso il rapporto

$$\frac{\lambda_j'}{\sum_j \lambda_j'} \quad . \tag{8.23}$$

8.8 Le variabili attive e le variabili illustrative

Le variabili coinvolte nell'ACM possono assumere due differenti ruoli "strategici":

a) Le *p* variabili che entrano direttamente nell'analisi e concorrono alla formazione della matrice logico-disgiuntiva completa, della matrice di Burt e degli assi fattoriali sono le variabili *attive*.
b) Le variabili *illustrative*, o *supplementari*, sono invece escluse dalla fase di estrazione dei fattori e vengono utilizzate solo successivamente, considerando la loro posizione sugli assi fattoriali come ausilio interpretativo all'analisi e per studiare eventuali legami di interdipendenza con i fattori stessi.

Solitamente, quindi, le variabili attive corrispondono alle variabili d'interesse della ricerca e quelle illustrative alle variabili demografiche di base o socio-demografiche aggiunte.

Effettuata una prima ACM "esplorativa", il ricercatore può decidere di cambiare il ruolo di alcune variabili attive riducendole a illustrative (e diminuendo quindi *p*). Questo avviene nei casi in cui una variabile attiva dà scarso contributo alla formazione degli assi fattoriali, a causa

a) di particolari specificità che la rendono indipendente da tutte le altre variabili;
b) di una distribuzione di frequenza molto sbilanciata fra le sue modalità.

Cambiato il ruolo delle variabili considerate "critiche", si rieffettua l'ACM andando a vedere come i punti-modalità delle variabili stesse ridotte a illustrative si collocano rispetto agli assi fattoriali.

8.9 L'interpretazione degli assi fattoriali

Stabilito il numero *k* di fattori si passa alla fase "focale" e nello stesso tempo più critica dell'analisi, e cioè all'*interpretazione* delle singole dimensioni fattoriali.

Per interpretare il significato degli assi fattoriali si utilizzano

a) le *coordinate fattoriali*, che indicano la proiezione dei *q* punti-modalità originali sugli assi dello spazio geometrico ridotto *k*-dimensionale ed esprimono le relazioni tra modalità e fattori;
b) alcuni importanti *indicatori* che permettono di valutare l'importanza che ogni variabile attiva, con le relative modalità, riveste nella formazione degli assi fattoriali.

Le coordinate fattoriali stabiliscono la posizione delle modalità sugli assi, sia in termini di distanza dal baricentro del sistema (cioè dal punto le cui coordinate corrispondono al vettore nullo *k*-dimensionale), sia in termini di "versante" positivo o negativo dell'asse considerato, in base al segno "+" (la modalità è proiettata sul semiasse positivo) o "-" (la modalità è proiettata sul semiasse negativo).

Le modalità che presentano i valori più alti sono di solito quelle che contribuiscono maggiormente alla formazione dell'asse fattoriale. Tuttavia, valore della coordinata e importanza della modalità non sono necessariamente proporzionali. Si può infatti verificare il cosiddetto "effetto delle modalità rare" per cui, a causa delle ponderazioni della metrica del χ^2, un punto può essere tanto più distante dal baricentro quanto più bassa è la frequenza marginale della relativa modalità.

È allora importante riferirsi agli altri indicatori, tutti forniti da SPAD:

a) La *massa*, o *peso relativo*, di ciascuna modalità è definita come rapporto tra la frequenza relativa della modalità e *p*.
b) La *distorsione*, o *distanza dall'origine*, di ciascuna modalità dà indicazioni sul suo carattere "periferico": solitamente a valori alti corrisponde una massa debole e quindi una scarsa rilevanza della modalità stessa; viceversa nel caso di valori bassi.
c) Il *contributo assoluto* di ciascuna variabile e delle singole modalità corrisponde alla percentuale di inerzia del fattore spiegata dalla modalità o dalla variabile a cui si riferisce rispetto all'insieme complessivo delle modalità o delle variabili.
d) Il *coseno quadrato* dell'angolo compreso tra l'asse fattoriale e il segmento corrispondente alla distanza del punto-modalità dal baricentro è detto anche *contributo relativo* o *qualità della rappresentazione* e consente di individuare quale asse fattoriale spiega meglio la posizione del punto rispetto al baricentro, cioè quale fattore descrive meglio la variabilità della modalità. Per ogni punto-modalità, l'asse fattoriale geometricamente più vicino corrisponde al valore più alto del coseno quadrato, che raggiunge il valore 1 se il punto è situato esattamente sull'asse.

Nell'applicazione dell'ACM ai dati in esame, illustrata nel prossimo Capitolo, si adottano i seguenti *criteri interpretativi*:

1ª fase: pre-selezione delle modalità

Vengono escluse a priori dall'analisi le modalità che all'interno della variabile presentano una frequenza assoluta inferiore o pari al 2% della frequenza assoluta massima, cioè del numero di unità statistiche. In SPAD il valore del 2% è consigliato per default.

2ª fase: scelta del numero di fattori

La scelta del numero di fattori da includere nell'analisi viene effettuata tramite

a) l'osservazione visiva dell'andamento degli autovalori nell'istogramma e l'individuazione del punto di gomito;
b) la reinterpretazione ottimistica di Benzecri;
c) lo studio dell'andamento degli incrementi percentuali dell'inerzia cumulativa calcolati sulla base dei nuovi autovalori ottenuti.

3ª fase: analisi dei contributi assoluti delle variabili

Una volta stabilito il numero k di assi fattoriali, il riferimento prioritario va ai contributi assoluti delle variabili, cioè alla percentuale di inerzia spiegata in corrispondenza di ciascuno dei k fattori scelti rispetto al sistema complessivo di variabili. In questo modo

a) viene valutato il contributo di ogni singola variabile all'inerzia complessiva del sistema;
b) per ogni variabile vengono individuati il fattore o i fattori corrispondenti ai contributi più rilevanti.

4ª fase: analisi dei contributi assoluti delle modalità

Nel contesto di ciascuna variabile si passa poi all'analisi delle modalità *limitatamente* al fattore o ai fattori più importanti, andando a vedere come il contributo assoluto si distribuisce all'interno della variabile, dando luogo ai singoli contributi assoluti delle modalità.

5ª fase: analisi dei contributi relativi

Per l'analisi delle modalità un ulteriore riferimento va al coseno quadrato, che dà una valutazione rigorosamente geometrica delle distanze dei punti dagli assi fattoriali e offre quindi informazioni sulla qualità della loro rappresentazione, consentendo tra l'altro, se necessario, di individuare in modo univoco l'asse fattoriale più vicino ad ogni punto-modalità considerato.

6ª fase: scelta finale delle modalità più rappresentative

L'analisi congiunta dei contributi assoluti e dei coseni quadrati guida la scelta delle modalità da ritenere più significative ai fini della sintesi.

7ª fase: analisi delle coordinate delle modalità scelte

Individuate le modalità più importanti, si procede a una loro "rilettura" secondo le coordinate del nuovo sistema fattoriale di riferimento, mettendo in evidenza:

a) le *combinazioni* lineari, corrispondenti a comportamenti "dello stesso segno" in riferimento agli assi fattoriali;
b) i *contrasti* lineari, corrispondenti a comportamenti "di segno opposto".

8.10 I valori-test per le modalità illustrative

Una modalità illustrativa non risulta caratterizzare un dato fattore se la sua distanza sull'asse fattoriale dal baricentro del sistema non è statisticamente significativa.

In termini inferenziali, parlare di distanza dal baricentro statisticamente non significativa equivale a ipotizzare che le unità statistiche che presentano la modalità in esame siano estratte casualmente dall'insieme complessivo delle unità. Per ogni *m*-esima modalità illustrativa, quindi, l'ipotesi nulla H_0 da controllare è l'ipotesi di estrazione casuale senza reintroduzione di n_m unità tra le *n* della popolazione di riferimento.

La distanza dal baricentro del sistema dell'*m*-esima modalità sul *j*-esimo asse fattoriale è data direttamente dalla sua coordinata sull'asse stesso: $x_m(j)$. Definito l'indice $i_m=1,\ldots,n_m$, $x_m(j)$ è individuata dalla media aritmetica delle coordinate $x_{i_m}(j)$ delle unità che presentano la modalità moltiplicate per un coefficiente dipendente dall'autovalore λ_k:

$$x_m(j)=\frac{1}{\sqrt{\lambda_k}}\sum_{i_m}\frac{x_{i_m}(j)}{n_m} \quad . \tag{8.24}$$

La distanza $x_m(j)$ viene a corrispondere a una variabile aleatoria *X* che sotto l'ipotesi H_0 segue una legge di probabilità normale dai parametri noti e facilmente riconducibile a una distribuzione normale standardizzata:

$$\begin{cases} E(X \mid H_0) = 0 & (8.25) \\ V(X \mid H_0) = \dfrac{n-n_m}{n-1}\cdot\dfrac{1}{n_m} \quad . & (8.26) \end{cases}$$

I *valori-test* sono quindi tratti dalla normale standardizzata e individuano le *probabilità critiche*

$$p_m(j) = Prob\ \{X \geq x_m(j) \mid \mathrm{H}_0\}\ . \qquad (8.27)$$

Com'è noto dalla statistica inferenziale, la probabilità critica corrisponde alla probabilità di sbagliare nel rifiutare H_0, cioè in questo caso alla probabilità di considerare significativa una modalità in realtà non importante.

All'aumentare del valore-test diminuisce la probabilità critica e diventa sempre più inverosimile l'ipotesi nulla: più è alto il valore-test, più la modalità concorre significativamente alla caratterizzazione del fattore.

Nel prossimo Capitolo saranno considerate significative le modalità illustrative che portano a valori-test maggiori di +2 o minori di –2, corrispondenti a una probabilità critica inferiore a circa il 5% [5].

[5] È l'impostazione di default di SPAD. In corrispondenza del 5% esatto il valore in ordinata della distribuzione normale standard sarebbe pari a 1,96.