5.1 Introduzione

L'analisi geometrica dei dati (GDA) – un nome suggerito da Patrick Suppes nel 1996 – è l'approccio alla statistica multivariata avviato da J.-P.
Benzécri negli anni '60, noto nella letteratura francofona come Analyse des Données (Benzécri et al. 1973; Benzécri 1982b). Oltre al “caso principale” dell’analisi delle corrispondenze (CA), la GDA include l’analisi delle componenti principali (PCA), riformulata come metodo GDA, e l’analisi delle corrispondenze multiple (MCA), una conseguenza dell’AC. Le tre idee chiave della GDA sono la modellazione geometrica (costruzione di nuvole euclidee), l'approccio formale (algebra lineare astratta; GDA è propriamente l'approccio geometrico formale alla statistica multivariata) e la filosofia induttiva (l'analisi descrittiva viene prima della modellazione probabilistica). Nelle applicazioni vigono i due principi di omogeneità ed esaustività.

Per riassumere, GDA è la tradizione di statistica multivariata di Benzécri, con il teorema spettrale come strumento matematico di base: “Tutto sommato, fare un'analisi dei dati, in buona matematica, significa semplicemente cercare autovettori; tutta la scienza (o l’arte) sta solo nel trovare la giusta matrice da diagonalizzare”. (Benzécri et al. 1973: 289). Questa tradizione estende l’approccio geometrico ben oltre il ridimensionamento dei dati categorici, un fatto ben percepito da Greenacre (1981: 122): “L’approccio geometrico della scuola francese dà una visione molto più ampia dell’analisi delle corrispondenze, ampliando il suo campo di applicazione a altri tipi di matrici di dati oltre alle tabelle di contingenza.

Il presente capitolo, in linea con il libro di Le Roux e Rouanet (2004a), affonda le sue radici nella tradizione di Benzécri. È dedicato alle tabelle individui × variabili, un set di dati di base in molti studi di ricerca, tramite PCA (variabili numeriche) o MCA (variabili categorizzate). La distinzione tra numerico e categorizzato è importante dal punto di vista tecnico, ma non è essenziale dal punto di vista metodologico. Come afferma Benzécri (2003: 7): “Non si dovrebbe dire: 'grandezza numerica continua' 'dati quantitativi' vs. 'numero finito di categorie' 'dati qualitativi'. A livello di individuo statistico, infatti, un dato numerico non è da prendere di regola nella sua piena esattezza ma in base alla sua significatività; e da questo punto di vista non esiste alcuna differenza di natura tra età e (diciamo) professione”.

Il capitolo è organizzato come segue. Descrivo PCA e MCA come metodi GDA nella Sezione 5.2. Successivamente presenterò i fattori di strutturazione e l'analisi dei dati strutturati nella Sezione 5.3. Nelle Sezioni 5.4 e 5.5, descrivo due analisi di tabelle strutturate di individui × variabili, incorporando le tecniche ANOVA (analisi della varianza) nella struttura geometrica. Infine, i commenti conclusivi sono offerti nella Sezione 5.6.

5.2 PCA e MCA come metodi geometrici

5.2.1 PCA: dall'analisi multivariata alla GDA

Per evidenziare l'analisi dei dati geometrici, la PCA è un esempio calzante per due motivi: (a) la PCA preesisteva come procedura di analisi multivariata consolidata e (b) come sottolinea Benzécri (1992: 57): "A differenza dell'analisi delle corrispondenze, i vari metodi derivati dall'analisi delle componenti principali assegnano *ruoli chiaramente asimmetrici* agli individui e alle variabili".

Denotando *n* il numero di individui e *p* il numero di variabili, la tabella dei dati analizzata dalla PCA è una tabella *n* × *p* con voci numeriche. Il seguente estratto di Kendall e Stuart (1973: 276) descrive bene i due spazi coinvolti: "Possiamo impostare uno spazio euclideo di *p* dimensioni, uno per ciascuna variabile, e considerare ciascun insieme campione… come determinante un punto in esso, per cui il nostro campione è costituito da uno sciame di *n* punti; oppure possiamo impostare uno spazio di *n* dimensioni, una per ogni osservazione, e considerare ciascuna variabile in esso, in modo che la variazione sia descritta da *p* vettori (che si trovano in uno spazio p-dimensionale incorporato in uno spazio n-dimensionale)." Nella discussione che segue, questi spazi saranno chiamati rispettivamente "spazio degli individui" e "spazio delle variabili". Nell'analisi multivariata convenzionale – si veda, ad esempio, Kendall e Stuart (1973) e Anderson (1958) – lo spazio delle variabili è quello fondamentale; le variabili principali vengono ricercate come combinazioni lineari di variabili iniziali aventi la varianza maggiore sotto vincoli specificati. D'altro canto, nella PCA riformulata come metodo GDA, lo spazio di base è quello degli individui (vedi, ad esempio, Lebart e Fénelon 1971).

Nella PCA come metodo GDA, i passaggi della PCA sono i seguenti:

Passaggio 1: la distanza d(i,iÿ) tra gli individui *i* e iÿ è definita da una forma quadratica della differenza tra i loro profili descrittivi, eventualmente consentendo pesi diversi sulle variabili; si vedano, ad esempio, Rouanet e Le Roux (1993) e Le Roux e Rouanet (2004a: 131).

Passaggio 2: vengono determinati gli assi principali della nuvola (mediante i minimi quadrati ortogonali) e viene mantenuto un sottospazio principale.

Fase 3: La nuvola principale di individui viene studiata geometricamente, mostrando le distanze approssimative tra gli individui.

Passaggio 4: segue la rappresentazione geometrica delle variabili, che mostra correlazioni approssimative tra le variabili. Disegnare il *cerchio delle correlazioni* è diventata una tradizione nella PCA come metodo GDA.

5.2.2 MCA: un metodo GDA

Le variabili categorizzate sono variabili definite da (o codificate in) un finito insieme di categorie; il paradigma della tabella individui × variabili categorizzate è la tabella *I* × *Q* di un questionario in formato standard, dove per ogni domanda *q* esiste un insieme *Jq* di categorie di risposta — chiamate anche *modalità e* ogni individuo *che* scelgo per ciascuna domanda *q* una ed una sola categoria nell'insieme *Jq.* Per applicare l'algoritmo di CA a tali tabelle è necessaria una codifica preliminare. Quando ciascuna domanda *q* ha due categorie, una delle quali è distinta come “presenza”. della proprietà q", CA può essere applicata immediatamente dopo *la codifica logica:* "0" (assenza) vs. “1” (presenza); in questa procedura non c'è simmetria tra presenza e assenza. La preoccupazione per la simmetria, spesso a requisito metodologicamente desiderabile – ha portato naturalmente alla codifica dove ciascuna variabile categorizzata *q* è sostituita da variabili indicatore *Jq ,* cioè, variabili (0,1) (note anche come “variabili dummy”), producendo quindi (lasciando) una matrice di indicatori *I* × *J* a cui le variabili di base $J = \sum_{q=1}^{Q} J_q$ Viene applicato l'algoritmo CA. In questa procedura vengono forniti tutti gli individui pesi uguali. All'inizio degli anni '70 in Francia questa variante dell'AC divenne gradualmente uno standard per l'analisi dei questionari. La frase “analyse des corrispondences multiples” appare per la prima volta in l'articolo di Lebart (1975), dedicato all'MCA come metodo di proprio diritto. Uno speciale software MCA fu presto sviluppato e pubblicato (vedi Lebart et al. 1977).

I passaggi per MCA sono paralleli a quelli per PCA descritti sopra.

Passaggio 1: dati due individui *i* e iÿ e una domanda q, se entrambi gli individui scelgono la stessa categoria di risposta, la parte di la distanza dovuta alla domanda *q* è zero; se l'individuo *i* sceglie la categoria *j* e l'individuo iÿ la categoria jÿ, ÿ j, la parte di (al quadrato) distanza dovuta alla domanda *q* è $d_{iiq}^{2}(i,i') = \frac{1}{f_j} + \frac{1}{f_{j'}}$, dove sono *fj* e *fjÿ* le proporzioni degli individui che scelgono rispettivamente *j* e jÿ. La distanza complessiva d(i,iÿ) è quindi definita da $d^{2}(i,i') = \frac{1}{Q}\sum_{q=1}^{Q} d_{iiq}^{2}(i,i')$, (vedi Le Roux e Rouanet 2004a). Una volta definita la distanza tra gli individui, la nuvola degli individui è determinato.

Passaggi 2 e 3: questi passaggi sono gli stessi di PCA (sopra).

Passaggio 4: la *nuvola di categorie* è composta da *punti di categoria J.*

Osservazione 1

(i) Solo i disaccordi creano distanza tra gli individui. (ii) Il minore è la frequenza delle categorie di disaccordo, maggiore è la frequenza delle categorie di disaccordo

distanza tra gli individui. La proprietà (i) è essenziale; la proprietà (ii), che valorizza le categorie poco frequenti, è desiderabile fino a un certo punto. Categorie molto rare di domande attive devono essere raggruppate con altre; in alternativa, si può tentare di porli come elementi passivi riuscendo a preservare le regolarità strutturali degli MCA; si veda l'articolo di Benali e Escofier (1987), riprodotto in Escofier (2003), e il metodo dell'MCA *specifico* in Le Roux (1999), Le Roux e Chiche (2004), e Le Roux e Rouanet (2004a: cap. 5).

Osservazione 2

Esiste una *proprietà fondamentale* che lega le due nuvole. Consideriamo la sottonuvola degli individui che hanno scelto la categoria *j* e, quindi, il punto medio di questa sottonuvola *(punto medio della categoria);* indichiamo $\bar{F}_s$ la sua coordinata principale qc (nella nuvola degli individui) e *gs* la coordinata principale qc del punto *j* nella nuvola delle categorie; allora si ha: $f_j \bar{F}_s = \ddot{y}_s g_s$ dove $\ddot{y}_s$ denota il sesimo valore singolare del CA della tabella *I* × *J.* Questa proprietà fondamentale segue dalle formule di transizione; vedere Lebart et al. (1984: 94), Benzécri (1992: 410), e Le Roux e Rouanet (1998: 204). Di conseguenza, la nuvola di punti medi di categoria derivata è in corrispondenza biunivoca con la nuvola di punti di categoria, ottenuta mediante contrazioni per fattori di scala ÿs *lungo* gli assi principali *s* = 1, 2,…, S.

5.2.3 Una strategia per analizzare le tabelle individui × variabili

La stessa strategia può essere applicata ad ogni tabella da analizzare. La strategia è delineata nelle seguenti tre fasi (espresse in termini di MCA, da adattare per PCA).

Fase 1: Costruzione della tabella individui×variabili

- Condurre analisi statistiche elementari e codifica dei dati. • Scegliere individui attivi e supplementari, variabili attive e supplementari e fattori strutturanti (vedi Sezione 5.3).

Fase 2: Interpretazione degli assi

- Determinare gli autovalori, le coordinate principali e il contributo delle categorie agli assi, quindi decidere quanti assi interpretare.

- Interpretare ciascuno degli assi considerati esaminando le domande e le categorie importanti, utilizzando i contributi delle categorie.

- Disegnare diagrammi nella nuvola di categorie, mostrando per ciascun asse le categorie più importanti, e poi calcolare i contributi delle deviazioni (vedi Le Roux e Rouanet 1998).

Fase 3: Esplorare la nuvola degli individui • Esplorare la nuvola degli individui, in connessione con domande di interesse.

- Procedere ad una classificazione euclidea degli individui, e poi interpretare questa classificazione nel quadro dello spazio geometrico.

Ogni fase della strategia può essere più o meno elaborata, a seconda delle domande di interesse. Come esempi concreti, vedere l'"esempio culturale" in Le Roux e Rouanet (2004a) e gli insiemi di dati disponibili sul sito Web all'indirizzo http://www.math-info.univ-paris5.fr/lerb.

5.2.4 Utilizzo del GDA nella ricerca tramite sondaggio

Negli studi di ricerca, GDA (PCA o MCA) può essere utilizzato per costruire modelli geometrici di tabelle individui × variabili. Un esempio tipico è l'analisi dei questionari, quando l'insieme delle domande è sufficientemente ampio e allo stesso tempo sufficientemente diversificato da coprire diversi temi di interesse (tra i quali viene gestito un certo equilibrio), in modo da portare a rappresentazioni multidimensionali significative.

Nelle scienze sociali, il lavoro di Bourdieu e della sua scuola è esemplare delle "affinità elettive" tra la concezione spaziale dello spazio sociale e le rappresentazioni geometriche, descritte da Bourdieu e Saint-Martin (1978) e sottolineate nuovamente da Bourdieu (2001: 70): "Coloro che conoscono i principi dell'MCA coglieranno le affinità tra MCA e il pensiero in termini di campo".

Per Bourdieu, la MCA fornisce una rappresentazione dei due volti complementari dello spazio sociale, vale a dire lo spazio delle categorie – nelle parole di Bourdieu, lo spazio delle proprietà – e lo spazio degli individui. Rappresentare i due spazi è diventata una tradizione nella sociologia di Bourdieu.
A questo proposito, il punto è sottolineato da Rouanet et al. (2000) che fare analisi delle corrispondenze non è sufficiente per fare "analisi à la Bourdieu" e che dovrebbero essere tenuti presenti i seguenti principi:

1. Rappresentare gli individui: l'interesse a rappresentare la nuvola di individui è abbastanza evidente quando gli individui sono "persone conosciute"; è meno evidente quando gli individui sono anonimi, come nei sondaggi d'opinione. Quando, invece, esistono fattori strutturanti gli individui (istruzione, età, ecc.), l'interesse a rappresentare gli individui non come

si realizza presto una raccolta indifferenziata di punti, ma strutturata in sottonuvole, che porta naturalmente ad analizzare sottonuvole di individui. Ad esempio, si veda lo studio sugli elettorati nello spazio politico francese di Chiche et al. (2000).

2. *Unire teoria e metodologia:* una volta costruiti gli spazi sociali, il modello geometrico dei dati può portare a un *uso esplicativo* della GDA, fornendo risposte ai seguenti due tipi di domande: come possono essere spiegate le posizioni individuali nello spazio sociale fattori strutturanti? Come possono le posizioni individuali, a loro volta, spiegare le prese di posizione degli individui su questioni politiche o ambientali, tra le altre? Come esempi, si vedano gli studi sullo spazio editoriale francese di Bourdieu (1999), sul campo degli economisti francesi di Lebaron (2000, 2001), e sulla società norvegese di Rosenlund (2000) e Hjellbrekke et al. (in stampa).

5.3 Analisi dei dati strutturati

5.3.1 Fattori strutturanti

L'analisi geometrica di una tavola individui × variabili mette in evidenza le relazioni tra individui e variabili, ma non tiene conto delle strutture di cui possono essere dotati gli stessi insiemi di base. Per *fattori strutturanti* intendiamo descrittori dei due insiemi fondamentali che non servono a definire la distanza dello spazio geometrico; e per *dati strutturati* designiamo tabelle di dati i cui insiemi di base sono dotati di fattori di strutturazione. Chiaramente, i dati strutturati costituiscono la regola piuttosto che l'eccezione, portando a questioni di interesse che potrebbero essere centrali per lo studio del modello geometrico dei dati. Infatti, l'insieme degli individui statistici, detti anche unità, potrebbe dover essere costruito a partire da fattori strutturanti di base, un tipico esempio è il disegno soggetti x trattamenti, per il quale un'unità statistica è definita come una coppia (soggetto, trattamento). Allo stesso modo, l'insieme di variabili potrebbe dover essere costruito a partire da fattori strutturanti di base. Esemplificherò tali costruzioni per gli individui nello studio sul basket (vedi Sezione 5.4) e per le variabili nello studio Education Program for Gifted Youth (EPGY) (vedi Sezione 5.5).

Nella statistica convenzionale, esistono tecniche per gestire i fattori strutturanti, come l'analisi della varianza (ANOVA) - comprese le estensioni multivariate (MANOVA) - e la regressione; tuttavia, queste tecniche non sono generalmente utilizzate nell'ambito della GDA. Attraverso *l'analisi strutturata dei dati*