Costruzione degli indicatori di comportamento mediante l'analisi fattoriale(*)

Agostino Tarsitano
Università degli studi della Calabria
Dipartimento di Economia e Statistica
87030 Arcavacata di Rende (Cs)
agotar@unical.it

Riassunto.
L'analisi fattoriale è un modello formulato nei primi anni '30 del XX secolo rivelandosi fin da subito strumento flessibile e potente per le indagini su molti fenomeni delle scienze "soft". Il lavoro fornisce la guida per la sua utilizzazione nell'ambito di una indagine sulla "ricerca del posto, solidiarietà, adattamento nell'esperienza dei giovani meridionali". Di rilievo sono le considerazioni sull'uso dell'analisi fattoriale per dati su scala quantitativa ordinale.

keywords: analisi multivariata, riduzione della dimensionalità, indagini sociologiche

(*) Lavoro inserito come appendice metodologica nel libro: "Non lontano dai padri" di P. Botta. Edizioni lavoro, Roma, 1981

L'autore si è limitato a riportare in forma digitale il testo e, laddove possibile, le immagini e le formule dei lavori originali Quando il risultato della scansione si è rivelato inadeguato, formule e grafici sono stati riscritti. Solo qualche evidente errore ortografico od imperfezione nelle espressioni analitiche è stato modificato. Il riassunto è stato aggiunto in tempi recenti.

1. Perché questa scelta

I dati di un'indagine sociologica si presentano di solito nella forma di una matrice di *N* righe ed *m* colonne *(N> m)* di misurazioni simultanee delle *N* unità di osservazione rispetto a tutte le *m* variabili. L'analisi fattoriale di una matrice di dati sociologici contribuisce in maniera determinante, sia a definire con precisione i concetti complessi sia ad arricchire la spiegazione dei concetti indefiniti o mal definiti, sia a provocare la formulazione di concetti nuovi. Il suo uso nell'analisi di dati reali resta influenzato non solo dall'aspetto matematico che i problemi possono presentare, ma soprattutto dagli obiettivi che la ricerca si propone, dalla mole dei dati coinvolti e dall'adattabilità dei dati stessi ad un modello teorico di interpretazione.

L'analisi fattoriale poggia sull'idea che le variabili siano correlate in modo tale da renderne possibile la ricostruzione a partire da un ristretto numero di parametri che rappresenterebbe così la struttura portante dei dati in una forma sintetica e di più agevole lettura. In questo senso l'analisi fattoriale dà luogo, attraverso una procedura complessa ed articolata in più passi iterativi, a delle combinazioni lineari, dette fattori, delle variabili di partenza; i fattori sono da interpretare come delle misurazioni di aspetti del campione, non direttamente osservabili, ma che stanno dietro alle variabili che compongono i fattori stessi.

Nell'ambito del presente lavoro l'analisi fattoriale ha motivato la scelta, fra le 256 variabili del questionario (un numero eccessivo rispetto alle 1.225 persone intervistate) di un sottoinsieme che ne contiene 68 e che risultano più prossime alle ipotesi che di solito ricorrono nell'analisi multivariata e che meglio si prestano a rappresentare il campione nei suoi aspetti caratterizzanti. E' stata inoltre definita una separazione logica delle variabili in gruppi afferenti ognuno ad un'angolazione particolare da cui il campione poteva essere colto. Nell'ultima parte di questa appendice parleremo in dettaglio di questa classificazione.

La motivazione principale all'analisi era quella di ottenere, all'interno di ogni gruppo, degli indicatori che, da un lato riuscissero a mettere in risalto le caratteristiche comuni delle variabili e d'altra parte potessero servire come "indicatori sintetici" rappresentativi del gruppo in una successiva analisi di più gruppi, fase del resto poco praticabile qualora si utilizzassero tutte le variabili di partenza. Le variabili comprese in ogni gruppo per le loro caratteristiche di assimilabilità per certi versi e di specularità per altri inducono ad aspettarsi che solo poche avranno effettiva rilevanza; le altre o saranno da trascurare oppure serviranno a rafforzare le variabili "interpretative" in una direzione a discapito di altre, pure interpretative, ma in una direzione diversa.

In generale, quello che bisogna notare è che la composizione dei gruppi è studiata in maniera tale che più piccolo è il numero di fattori e più interpretabili ed interessanti saranno i risultati dell'analisi fattoriale. A questo fine si è applicato a ogni gruppo il modello esposto nel paragrafo 2 con le chiavi di interpretazione delineate nel paragrafo 3. Si è però preceduta l'analisi vera e propria con una serie di controlli e verifiche (paragrafo 4) che ne collocassero l'adeguatezza in un ragionevole quadro di generalizzazione.

2. Un modello di analisi fattoriale

L'espressione analitica del modello 1 di analisi fattoriale è la seguente

$$Z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jp}F_p + U_j \qquad j = 1, 2, \ldots, m$$

m - Numero delle variabili.
p - Intero positivo minore di m.
Z_j - La variabile standardizzata j-esima.
F_i - L'i-esimo fattore standardizzato.
a_{ji} - Il peso del fattore i-esimo sulla variabile j-esima.
U_j - Errore di specificazione per la j-esima variabile.

Con *N* osservazioni disponibili sulle *m* variabili il modello potrà essere espresso con la notazione matriciale:

$$\underset{(mxN)}{\boldsymbol{Z}} = \underset{(mxp)}{\boldsymbol{A}}\,\underset{(pxN)}{\boldsymbol{F}} + \underset{(mxN)}{\boldsymbol{U}}$$

Al fine di distinguere l'effetto su ***Z*** di ***F*** e di ***U*** è necessario che ogni colonna della matrice ***A*** abbia almeno due elementi non nulli. I pesi individuali a_{ij} dovrebbero essere inoltre o molto piccoli oppure piuttosto grandi in modo che ogni variabile possa essere collegata ad un numero limitato di fattori.
Le ipotesi di tipo più statistico sono soprattutto legate alle aspettative delle variabili coinvolte nel modello.
.
1. $E(\boldsymbol{U}\boldsymbol{U}^t) = \boldsymbol{D}$; dove ***D*** è una matrice diagonale. Questo significa che gli errori connessi ad una variabile non sono correlati con quelli di altre variabili.
2. $E(\boldsymbol{U}\boldsymbol{F}^t) = \boldsymbol{O}$; dove ***O*** è una matrice nulla. L'ipotesi implica che il modello assume l'assenza di legami Lineari più o meno forti tra i fattori e gli errori.

Le due ipotesi danno luogo a delle interessanti relazioni tra le matrici di dispersione.

$$\text{a) } E(\boldsymbol{ZZ}^t) = \boldsymbol{AS}_F\boldsymbol{A}^t + \boldsymbol{D}; \quad \text{b) } E(\boldsymbol{ZF}^t) = \boldsymbol{AS}_F$$

Dove $\boldsymbol{S}_{\mathrm{F}}$ è la matrice di dispersione dei fattori.

Per quanto attiene alla relazione a), se i fattori non sono correlati, allora $\boldsymbol{S}_{\mathrm{F}} = \boldsymbol{I}$, cioè la matrice di identità e di conseguenza:

$$E(\boldsymbol{ZZ}^t) = \boldsymbol{AA}^t + \boldsymbol{D}$$

Dal punto di vista formale, l'ipotesi di fattori incorrelati non è necessaria. Dato che la $\boldsymbol{S}_{\mathrm{F}}$ ammette, per come è definita, una decomposizione nella forma $\boldsymbol{S}_{\mathrm{F}} = \boldsymbol{GG}^{\mathrm{t}}$ con $\boldsymbol{G}$ di ordine $(p\ X\ p)$ nota come matrice di rotazione, se si pone $\boldsymbol{A}^+ = \boldsymbol{AG}$ allora:

$$E(\boldsymbol{ZZ}^t) = (\boldsymbol{AG})\,(\boldsymbol{AG})^t + \boldsymbol{D} = \boldsymbol{A}^+\boldsymbol{A}^{+t} + \boldsymbol{D}$$

Di fatto ci si trova di fronte solo alla seconda di queste relazioni e dato che non conosciamo $\boldsymbol{G}$ questa non è distinguibile dalla precedente $E(\boldsymbol{ZZ}^t)$. L'equivalenza algebrica dei due casi non implica che la rappresentazione delle variabili originarie sia la stessa tanto se tenuta con fattori correlati quanto con fattori che non lo siano. E' molto importante, ai fini dell'interpretazione del modello, tenere i due casi ben distinti.

La relazione b) definisce in sostanza il tipo di legame lineare esistente tra i fattori e le variabili del modello. Come si è detto prima è matematicamente irrilevante che i fattori siano correlati o meno, ma è chiaro che assumere l'una o l'altra ipotesi comporta una precisa valutazione sulla forza del legame che esiste tra fattori e variabili. Nell'ambito della ricerca condotta si è preferito assumere $\boldsymbol{S}_F \neq \boldsymbol{I}$ cioè fattori correlati, dato che le variabili coinvolte erano in effetti molto connesse fra di loro, almeno da un punto di vista logico-formale.

I problemi di caIcolo legati al modello di analisi fattoriale sono di:

a. Determinare una stima di A ed eventualmente di $\boldsymbol{S}_F$;
b. Determinare una stima della matrice $\boldsymbol{D}$;
c. Ottenere i p fattori con il minimo intero p.

Le procedure da seguire sono numerose. Quella prescelta dalla ricerca utilizza la tecnica delle componenti principali con iterazioni. Come si è detto prima, dal punto di vista matematico, tutte le rotazioni sono equivalenti. Cosic-

ché le ragioni per preferire una soluzione a tutte le altre devono scaturire dalla teoria che sta dietro al modello: possiamo scegliere quella rotazione, quando esista una matrice che si presti ad una più facile interpretazione, ovvero quella soluzione per cui l'esistenza di fattori $\boldsymbol{F}$ con coefficienti di correlazione $\boldsymbol{AS}_F$ rispetto alle variabili $\mathbf{Z}$ è più coerente alla teoria sottostante.

Il problema nella definizione della matrice $\boldsymbol{G}$ è che esiste un *continuum* di matrici di rotazioni possibili, cosicché è difficile dire se quella che si è determinata è la migliore oppure che non sia possibile ottenerne una adeguata. In ogni caso, la scelta deve sempre essere motivata dalla teoria del modello. Avviare l'analisi fattoriale significa presupporre che questo sia possibile. In caso contrario non esiste tecnica di analisi che possa fornire una teoria al modello. Consapevoli di questi limiti si è scelta una matrice di rotazione, da determinare analiticamente, ma che comportasse comunque fattori correlati (con coefficienti non troppo elevati, diciamo, in valore assoluto, compresi tra 0.30 e 0.50) in quanto si è fatta l'ipotesi che i fattori, per come si sono definiti i gruppi di variabili, riproducessero, in misura più blanda, la struttura della correlazione delle variabili di partenza.

3. Interpretazione del modello

I punti più importanti della fase di interpretazione di un modello di analisi fattoriale sono la scelta del numero di fattori e la loro denominazione. La scelta del numero di fattori è legata alla soluzione fattoriale diretta che è stata realizzata con la tecnica delle componenti principali. Quello che da questa procedura si ottiene sono gli autovalori ed autovettori della matrice di correlazione ridotta. Scegliere p vuol dice scegliere le p componenti di maggiore rilevanza ed in modo tale che quelle prescelte corrispondano alla gran parte della variabilità totale.

Il test di significatività che si è adottato per selezionare il numero di fattori è il criterio di Bartlett che si concretizza nella formula:

$$Q = -\left[(N-1) - \frac{2m+5}{6} - \frac{2}{3}p\right]\left\{(m-p)Ln\left(\frac{m-p}{p-\sum_{i=1}^{p}\lambda_{(i)}}\right)\right\}Tr(\boldsymbol{R})$$

dove i $\lambda_{(i)}$ sono gli autovalori di $\boldsymbol{R}$ ordinati in senso decrescente ed $\boldsymbol{R}$ è la matrice di correlazione ridotta cioè la matrice delle correlazioni tra le varia-

bili, ma con la gli “1” della diagonale sostituiti da valori opportuni ottenuti con una procedura iterativa.

La statistica *Q* ha una distribuzione, sotto l’ipotesi di normalità per le variabili del modello, del $\chi 2$ con $(m-p)\ (m-p+1)/2$ gradi di libertà. Se il suo valore è significativo, questo va interpretato come un indice di non trascurabilità dei fattori associati con gli autovalori che vanno da $p+1$ a m. Il test di Bartlett consente dunque di scegliere fra gli *m* fattori disponibili i *p* che a questo stadio sembrano promettere particolare idoneità alla interpretazione del modello di analisi fattoriale. In linea di massima si è però fissato a priori il numero di fattori da estrarre in base alla regola empirica $p < m/2$ Vale a dire che al massimo si dovranno estrarre dai dati un nu- mero di fattori pari alla metà delle variabili. Per il limite inferiore varranno soprattutto criteri di interpretabilità discussi più avanti.

Altro punto fondamentale è l’assegnazione di un “nome” ai fattori. Come si è detto nell”introduzione i fattori sono da considerarsi delle variabili *sui generis* che misurano le caratteristiche intrinseche dei dati. Queste ultime devono però risultare dalla disposizione dei pesi sui fattori. Una volta fissata una soglia al di sotto della quale i pesi sono da considerarsi praticamente nulli (0.25 ad esempio) e quindi con le rispettive variabili ininfluenti sul fattore resta da vedere come si distribuiscono i pesi più significativi. Se i pesi più elevati sono quelli associati a variabili che presentano una connessione ben delineata fra di loro, la connessione stessa darà il nome del fattore e l’interpretazione sarà completata. Se invece i pesi risultano Iegati a variabili fra le quali è difficile ipotizzare delle interrelazioni, oppure se tutte le variabili danno un contributo equivalente *(fattore overall),* verrà a cadere la possibilità di interpretazione ed occorre rivedere l’intero modello ed il campione di dati a cui è stato adattato.

In questo senso occorre passare dalla espressione del modello secondo la sua “ Pattern fattoriale ” dove le variabili sono in relazione lineare con i fattori, alla “ struttura fattoriale ” in cui i fattori compaiono a sinistra delle equazioni

$$r(Z_i F_j) = a_{j1} r(F_i F_1) + \ldots + a_{jp} r(F_i F_p)$$

dove $r(Z_i F_j)$ è il coefficiente di correlazione fra la i-ma variabile ed il j-mo fattore. Quello che si conosce dei fattori è solo la loro matrice di correlazione con le variabili del problema. Vale a dire la matrice $\boldsymbol{A}$. In fase di interpretazione occorre dunque chiedersi una volta che si sia identificato un fattore con un concetto coerente, se qualche altro insieme di variabili qualsiasi non possa aver dato luogo allo stesso schema di correlazione con i fattori.

La domanda non è oziosa. L'indeterminatezza dei fattori è una realtà matematica, sebbene nella maggior parte dei casi il dubbio dei ricercatori non arriva ad esplorare questa possibilità (si veda, per i dettagli analitici Torrens-Ibern, 1972). Nel lavoro citato viene presentato come soglia critica di identificabilità dei fattori il cosiddetto " Guttman Criterion " che discrimina fra i fattori " interpretabili " e quelli che non Io sono. Il criterio afferma in termini molto semplici che un dato fattore è connesso con quello specifico insieme di variabili se

$$\xi_i^2 \geq \sqrt{0.5(\gamma + 1)}$$

dove ξ_i^2 è il coefficiente di correlazione multipla tra l'i-mo fattore e le variabili ***Z***, mentre γ è il valore minimo accettabile del coefficiente di correlazione fra due fattori diversi, ma che hanno Io stesso Iegame lineare con le variabili del problema. Se ad esempio ($\gamma = 0.6$ si avrà $\xi_i^2 = 0.2$ e questo implica che l'interpretazione del fattore in termini teorici non è affidabile; esiste infatti almeno un altro fattore, oltre a quello ottenuto dalla procedura, che pur poggiando su variabili diverse, è simile a quello determinato e fra i due la correlazione è solo dello 0.20 (in pratica ortogonali). Ragione per cui non si può dare alcun significato particolare, a meno di alchimie teoriche, aI fattore calcolato. In questo lavoro, la soglia di ξ_i^2 è stata fissata al livello di 0.8 ovvero Ie variabili presenti in ogni fattore ne debbono spiegare almeno l'80 %. Il Guttman Criterion che ne consegue è dello 0.6 che può forse sembrare eccessivamente ridotto, ma elevandolo ancora si rischia di imporre coefficienti di correlazione multipla che ben di rado si realizzano in pratica.

Se la fase di interpretazione si è chiusa con risultati convincenti, si passa alla stima dei fattori ***F*** per tutte le *N* unità di osservazione. Anche in questo passo delI'analisi fattoriale la statistica offre una serie di procedure alternative. Quella scelta neI nostro caso è il metodo dei minimi quadrati. Secondo questo metodo bisogna rendere minimo il quadrato delle differenze fra i dati originali ed i risultati ottenuti stimando i fattori con i pesi ricavati nelle fasi precedenti. Il risultato è la determinazione degli ***F*** secondo la formula

$$\boldsymbol{F=Z\ (R^{-1}AS_F)}$$

dove ***R*** è la matrice di correlazione fra le variabili ***Z***.. Gli ***F*** riflettono a livello di unità i " tratti impliciti " rilevati a livello di variabili e costituiscono il prodotto finale dell'analisi fattoriale da usare come input in altre analisi.

4 Considerazioni sui dati

Per analizzare dei dati secondo il " modo " dell'analisi fattoriale non è necessario utilizzare come misura di similarità il coefficiente di correlazione e la corrispondente matrice ***R***. Qualsiasi altra matrice di coefficienti di associazione, purché presenti Ie stesse caratteristiche di simmetria e di positività, potrebbe essere adoperata come base di partenza dell'analisi fattoriale.

Nella presente ricerca si è in effetti adoperato il coefficiente di correlazione lineare. Le variabili erano però tutte misurate su delle scale ordinali e con dei campi di variazione abbastanza diversificati. E' stato dunque necessario prendere in considerazione i problemi connessi alla misura su scaIa ordinale. Era possibile utilizzare il coefficiente di correlazione per ranghi come indice di similarità, il che sarebbe stato più coerente alla scala di misura delle variabili. Comunque i risultati ottenuti sono stati perfettamente analoghi nei due casi e quindi si è mantenuta la correlazione lineare come misura di similarità.

In generale l'uso dei dati ordinali comporta, in positivo, una attenuazione dei problemi legati aIla diversità delle scale di misurazione i quali possono viziare i risultati. Consegue anche una attenuazione del problema della stima della matrice di correlazione in presenza di valori remoti nel campione. D'altra parte l'affidarsi ai dati ordinali preclude in sostanza l'uso dei test di significatività; dato che tali test poggiano sull'ipotesi di distribuzione normale delle variabili e su quelle ordinali si può porre solo a prezzo di ulteriori ipotesi ancora meno plausibili. Non si tratta di una gran perdita in quanto i test possono ben di rado applicarsi in maniera rigorosa anche con variabili di tipo metrico.

L'analisi fattoriale è una tecnica i cui risultati non possono essere valutati in maniera separata dalla particolare trasformazione (o codifica) adottata sulle variabili. Per esempio il coefficiente di correlazione lineare cambia di segno se si modifica il senso della codifica (assegnando per esempio il valore più alto allo *status* che si trova all'inizio della scala graduale) senza che si possa ritenere modificata l'intensità del legame lineare fra le variabili. Altro elemento di cui occorre tenere conto nel calcolo di r(x, y) è che se l'intervallo di codifica della X e/o della Y viene cambiato allora può cambiare anche il valore di r(X, Y) con conseguenze negative sulla corretta interpretazione e confrontabilità della misura. In realtà quando alcuni gradi di misurazione della X o della Y vengono aggregati o soppressi allora r(X, Y) calcolato sullo stesso numero di unità di osservazione tende ad essere più elevato dando così una idea distorta sulla forza della interrelazione esistente fra le due variabili. Questi inconvenienti si verificano in maniera più o meno accentuata a secondo dell'ampiezza del campione su cui si opera.

Si è dovuto dunque procedere ad una verifica sia della struttura di codifica che della misura di similarità al fine di saggiare l'eventualità di conclusioni diverse a partire da diverse numerosità campionarie. Il controllo è stato effettuato per ogni gruppo di variabili secondo i tre passi che seguono:

1. Estrazione di due campioni casuali con riposizione ciascuno di numerosità *N/2*;
2. Calcolo del vettore delle medie e della matrice di dispersione sia nei subcampioni che neI campione totale;
3. Esecuzione del test sulla differenza dei vettori di valori medi ottenuti al punto 2.

I tre passi suddetti sono stati eseguiti tre volte per ogni gruppo. L'ipotesi di uguaglianza dei valori medi è stata accettabile circa due volte su tre in ogni gruppo. Questo da un lato conforta sulla base di dati su cui si è costruita l'analisi fattoriale e tra l'altro indica che la numerosità del campione era probabilmente eccessiva, data la accentuata omogeneità della popolazione

Riferimenti bibliografici

Clark C.(1977). Correlation, principal component, and the problem of multistate data. *Geoforum*, 8, pp. 69-72, 1977.

Eiffers H., J. Bethlehem, R. Gill (1978). Indeterminacy problems and the interpretation of factor analysis problems, *Statitica Neerlandica*, 32,181-199 Frane J., M. Hill (1976). Factor analysis as a tool for data analysis *Communications in Statistics, Theor. Meth*. 6, 487-506.

Rao R. (1973). *Linear statistical inference and its applications*, John Wiley & Sons, New York

Torrens-lbern J. (1972). *Modèle et method de l'analyse factorielle*, Dunod, Paris 1972.