Analisi fattoriale

introduzione
vers. 1.0

Germano Rossi[1]
germano.rossi@unimib.it

[1] Dipartimento di Psicologia, Università di Milano-Bicocca

2009

Definizioni

Variabile indipendente una caratteristica che non dipende da altre (ad es. il sesso non dipende dall'intelligenza, dalla cultura. . .)

Variabile dipendente una caratteristica che può essere influenzata da altre caratteristiche (ad es. la ricchezza del vocabolario può dipendere dall'educazione e dalla famiglia di origine)

- La relazione esistente fra l'indipendente e la dipendente è spesso interpretata come influenza o spiegazione

Definizioni

Variabile osservata o misurata
: una caratteristica che si può misurare in modo diretto (es.: età, reddito, numero di errori)

Variabile latente
: una caratteristica che non può essere misurata direttamente ma che si ipotizza avere dei legami con altre variabili osservabili (es.: intelligenza, cultura, inconscio)

- Queste varie classificazioni vengono applicate in base al contesto
- Una variabile considerata indipendente in un certo contesto, può diventare dipendente in un'altro

Ripasso di statistica

- varianza e covarianza

$$s^2 = s_{xx} = \frac{(X-\bar{X})(X-\bar{X})}{N}$$

$$cov = s_{xy} = \frac{(X-\bar{X})(Y-\bar{Y})}{N}$$

- L'analisi fattoriale si calcola, in genere, sulle correlazioni
- In linea di massima la correlazione utilizzata è quella di Pearson

$$r = \frac{(X-\bar{X})(Y-\bar{Y})}{s_x s_y}$$

- La correlazione è una covarianza standardizzata

Ripasso di algebra matriciale

- ordine o dimensione: ogni matrice ha 2 dimensioni, sempre indicate come “numero di righe” e “numero di colonne” ($\mathbf{A}_{6x3}$). Se una matrice ha lo stesso numero di righe e di colonne (è una matrice quadrata) si può indicare una sola dimensione ($\mathbf{R}_6$)
- trasposta: è una matrice ottenuta scambiando fra loro le righe e le colonne. Si indica con un apice ($\mathbf{A}'$)
- vettori: Sono matrici con una sola riga ($\mathbf{v}_{1xn}$) o una sola colonna ($\mathbf{v}_{nx1}$). Per definizione i vettori riga sono considerati trasposti di un vettore colonna

Ripasso di algebra matriciale

- conformabilità: due matrici sono conformabili e si possono moltiplicare fra loro quando il numero di colonne della prima è uguale al numero di righe della seconda ($\mathbf{A}_{6x3}$ è conformabile con $\mathbf{B}_{3x7}$ ma non con $\mathbf{B'}_{7x3}$). La matrice risultante avrà il numero di righe della prima matrice e il numero di colonne della seconda ($\mathbf{A}_{6x3}\mathbf{B}_{3x7} = \mathbf{C}_{6x7}$)

$$\mathbf{A}_{2x3}\mathbf{B}_{3x2} = \begin{bmatrix} 1 & 3 & 2 \\ -1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} 7 & 15 \\ -2 & 10 \end{bmatrix} = \mathbf{C}_{2x2}$$

Ripasso di algebra matriciale

- matrice simmetrica: una matrice quadrata in cui gli elementi in posizione i, j sono uguali a quelli in posizione j, i. La matrice risulta speculare rispetto alla diagonale principale. Tutte le matrici di correlazione e di varianza/covarianza sono simmetriche
- matrice indentità ($\mathbf{I}$) o unitaria: una matrice quadrata tutta composta da 0 con l'eccezione della diagonale principale in cui vi è 1. Svolge la stessa funzione dello scalare 1 nell'algebra "normale": qualsiasi matrice moltiplicata per $\mathbf{I}$ rimane uguale a se stessa ($\mathbf{AI} = \mathbf{A}$)

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Ripasso di algebra matriciale

- determinante: è un numero caratteristico che rappresenta la matrice. Si indica ponendo fra $|$ il nome della matrice ($|\mathbf{R}|$). Si calcola prendendo in considerazione tutte le possibili combinazioni degli elementi della matrice in cui ogni riga e ogni colonna sia usata una sola volta.
- matrice inversa: è la matrice che svolge la funzione del reciproco dell'algebra scalare:

$$a \times \frac{1}{a} = 1 \qquad \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

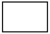
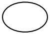
Si può calcolare solo per matrici quadrate e solo se il determinante è diverso da 0. L'inversa si usa nella "divisione" delle matrici.

Ripasso di algebra matriciale

- matrice mal condizionata (o *singolare* o *definita non positiva* o *positivamente non definita*: una matrice quadrata con determinante 0 che quindi non può essere invertita. Matematicamente significa che almeno una riga o una colonna è combinazione lineare di almeno una delle altre. Le possibili cause (non matematiche) sono campioni piccoli o l'uso delle opzioni *pairwise* (per le correlazioni)
- rango: se una matrice ha determinante nullo, si può eliminare una riga e una colonna e ricalcolare il determinante procedendo con tutte le possibili combinazioni. Se ancora i determinanti di tutte le sottomatrici così ottenute sono nulli, si elimina un'altra riga/colonna. Il rango è l'ordine della sottomatrice il cui determinante è diverso da 0. Indica il numero di righe/colonne che sono linearmente indipendenti dalle altre.

Modellazione grafica

È possibile rappresentare graficamente le relazioni fra le variabili (osservate/latenti, dipendenti/indipendenti)

Variabile osservata o manifesta	->	rettangolo	
Variabile latente	->	ellisse	
Relazione generica	->	segmento	$\overline{}$
Influenza, spiegazione	->	freccia	$\longrightarrow$
Relazione reciproca (correlazione, covarianza)	->	doppia freccia	$\longleftrightarrow$

Aspetti storici

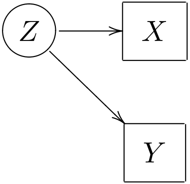
- Si pensa che la psicologia nasca attorno al 1879 con Wundt e generalmente si pensa anche che, alle sue origini, la psicologia fosse principalmente introspezionistica
- ma il francese Binet crea una prima versione della sua scala di misura dell'intelligenza già nei primi anni del 1900
- e subito si inizia a discutere:
  - alcuni ritengono che l'intelligenza sia una sola abilità e che chi è bravo a fare una cosa, sarà bravo a fare tutto
  - altri pensano che l'intelligenza dipenda dalle aree di misura e che l'abilità in qualcosa non permetta di predire l'abilità in un altro campo
- il primo a teorizzare quella che oggi chiamiamo *analisi fattoriale* è Spearman

Aspetti storici

- **1904, Charles Spearman: Teoria bifattoriale**
  sosteneva che le misure di abilità mentale relative ad un test potevano essere spiegate come attribuibili ad un'abilità generale comune a tutte le abilità e ad un'abilità specifica e queste abilità dipendono ciascuna da un "fattore", chiamati da Spearman "Fattore generale" (G) e "fattore specifico o unico" (U).
- **1945, Thurstone: Teoria multifattoriale**
  propose di sostituire il fattore generale con dei "fattori comuni" (F).
- La differenza è che i fattori comuni sono relativi solo ad alcuni item, quello generale li prendeva in considerazione tutti contemporaneamente.
- La teoria fattoriale esisteva, ma mancava la "capacità tecnica" (ovvero la parte matematica) per svilupparla
- solo con i computer divenne possibile utilizzarla

La correlazione

- Tutto si basa sul concetto e la logica della correlazione
- Quando due variabili correlano (molto) fra loro, noi sappiamo che hanno un andamento concomitante
- che hanno qualcosa in comune fra loro (pari a r^2)
- che è anche possibile che esista una terza variabile che le influenza entrambe



- questa terza variabile è il fattore o variabile latente
- tuttavia con due sole variabili non è possibile stimare una latente

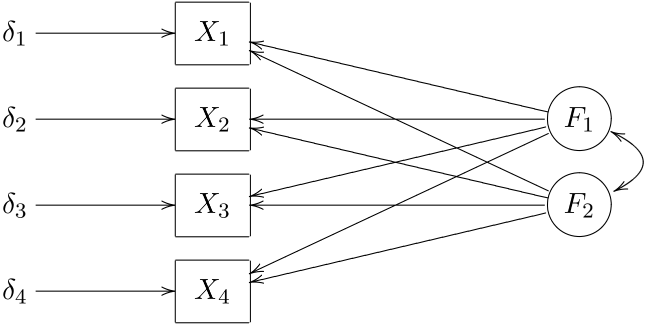
Matrice di correlazione

- Alla base della teoria di Spearman vi è la tetrade cioè tre variabili molto correlate fra loro

	X1	X2	X6	X3	X4	X5
X1	1					
X2	**0,891**	1				
X6	**0,842**	**0,799**	1			
X3	-0,300	-0,177	-0,161	1		
X4	0,126	0,000	-0,150	**-0,804**	1	
X5	-0,221	-0,124	-0,206	**0,876**	**-0,721**	1

- Ogni tetrade dovrebbe corrispondere ad una variabile latente che è possibile stimare a partire dalle correlazioni stesse

Esempio teorico



Considerando la variabile X_1 e pensandola come una regressione multipla:

$$X_1 = b_{11}F_1 + b_{21}F_2 + \delta_1$$

Analisi fattoriale come regressione

$$X_1 = b_{11}F_1 + b_{21}F_2 + \delta_1 \text{ (versione AF)}$$

$$Y_1 = b_{11}X_1 + b_{21}X_2 + \varepsilon_1 \text{ (versione regressione)}$$

- X_1 [Y_1] è la variabile che rappresenta l'"item" (la dipendente)
- F_1 e F_2 [X_1 e X_2] rappresentano i "fattori" (le indipendenti)
- b_{11} e b_{21} sono i parametri di regressione
- δ_1 [ε_1] è l'errore

- solo che F_1 e F_2 non li conosciamo (non sono stati misurati)

Scopo dell’analisi fattoriale

Sotto il nome generico di **analisi fattoriale** abbiamo diverse tecniche statistiche:

- L’*analisi in componenti principali* (ACP) è una tecnica matematica alla base dell’analisi fattoriale, dell’analisi delle corrispondenze e di altre tecniche
- L’*analisi fattoriale esplorativa* (AFE) che serve essenzialmente per esplorare la relazione fra variabili multivariate
- L’*analisi fattoriale confermativa* (AFC) che serve per verificare determinate relazioni fra variabili multivariate

Scopo dell’analisi fattoriale

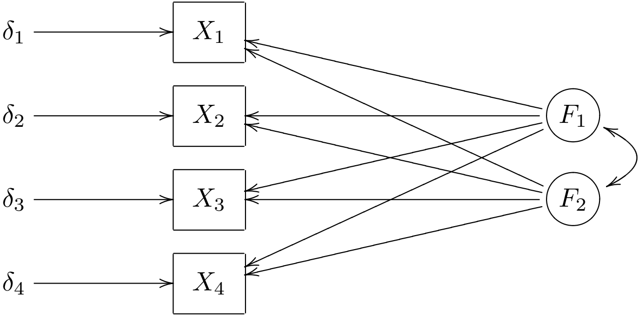
- Nel suo complesso, AFE e ACP servono per
  1. Trasformare le variabili osservate in una struttura più semplice che contenga però le stesse informazioni dell’originale
  2. Ridurre un insieme di variabili osservate ad un insieme inferiore di variabili non osservate o latenti (fattori, componenti, dimensioni)
- l’AFC è un’estensione recente dell’analisi fattoriale, sviluppata da Joreskog (1973). A partire da una possibile soluzione fattoriale si cerca di stimare la miglior soluzione possibile
- a partire dall’AFC, sono stati sviluppati nuovi metodi anche per l’AFE
- AFE e ACP sono matematicamente simili, mentre AFC richiede un approccio diverso

Esplorare/confermare

- L'analisi fattoriale **esplorativa** (AFE) serve per cercare le variabili latenti all'interno delle osservate: non si hanno ipotesi a priori su quali fattori influiscano sulle osservate.
- L'analisi fattoriale **confermativa** (AFC) serve quando si hanno idee abbastanza chiare su quali fattori influenzano quali variabili. Quindi per verificare che certe relazioni ipotizzate fra le osservate e le latenti siano effettive.
- Tuttavia non esiste un'analisi esclusivamente esplorativa e una esclusivamente confermativa
- È normale effettuare diverse analisi esplorative alla ricerca di una soluzione soddisfacente facendosi guidare da una teoria o cercando di avvicinarsi a delle ipotesi di partenza
- È altrettanto normale trovare che la prima analisi confermativa non funzioni e diventi necessario "aggiustare" il modello

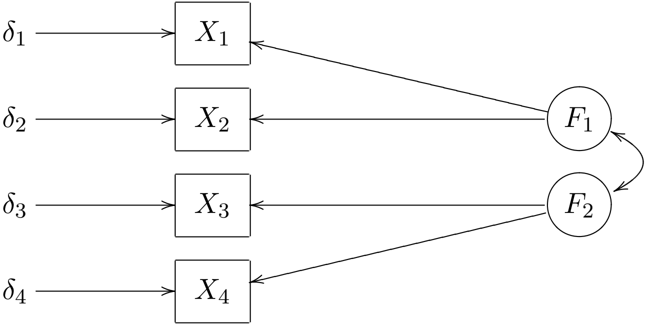
Analisi fattoriale esplorativa

- Serve per associare una o più variabili latenti (che non si conoscono) ad un gruppo di variabili osservate che si presuppone abbiano qualche cosa in comune,.
- Tutte le variabili osservate (ma in grado diverso) partecipano ai fattori (che possono anche essere correlati fra loro)



Analisi fattoriale confermativa

- Ciò che in esplorativa è chiamato *Fattore*, in confermativa è chiamato **variabile latente**
- Particolarità: non tutte le variabili osservate sono spiegate da tutte le latenti; al contrario, ogni latente spiega solo alcune osservate



Implicazioni fra osservate/latenti-fattori

- ogni fattore spiega tutte le osservate
- una latente deve spiegare almeno una osservata
- latenti diverse possono influire su osservate diverse
- la differenza misurata fra due casi statistici di una stessa osservata dipende, almeno in modo parziale, dalla loro differenza nel fattore/latente
- due osservate influenzate dal medesimo fattore/latente devono correlare molto fra loro

Analisi fattoriale esplorativa

Teorema fondamentale $\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{U}^2 \quad \mathbf{R} = \mathbf{P}\mathbf{\Phi}\mathbf{P}' + \mathbf{U}^2$
ovvero: la matrice di correlazioni fra le variabili osservate è riproducibile tramite a) una matrice di saturazioni fattoriali delle latenti (**A** o **P**) moltiplicata per la propria trasposta (ed eventualmente per la correlazione fra le latenti Φ) e b) una matrice (diagonale) di fattori unici (**U**)

Assunzioni

- I fattori unici (**U**) non correlano con i fattori comuni (**A** o **P**)
- I fattori unici (**U**) non correlano fra di loro
- I fattori comuni possono essere correlati fra di loro (soluzione obliqua) o non essere correlati (soluzione ortogonale)

Analisi fattoriale esplorativa

$\mathbf{Z} = \mathbf{F}\mathbf{A}' + \mathbf{U}$	$\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{U^2}$	(ipotesi ortogonale)
	$\mathbf{R} = \mathbf{P}\mathbf{\Phi}\mathbf{P}' + \mathbf{U^2}$	(ipotesi obliqua)

Z=dati grezzi standardizzati	nxm	n=soggetti
F=Fattori comuni	nxf	m=osservate
A,**P**=saturazioni/pesi	mxf	f=latenti
U=fattori unici	nxm	
R=matrice correlazioni	mxm	
Φ=correlazioni fra fattori	fxf	

La matrice di correlazione è riproducibile tramite una matrice di saturazioni fattoriali (dipendenti dai fattori comuni) moltiplicata per la sua trasposta e aggiungendo un termine ”d’errore” corrispondente ai fattori unici

Un singolo punteggio

$$z_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1f}F_f + u_1$$

- z_1 è il punteggio standardizzato di un caso nella variabile 1
- a_{11} è il parametro di regressione (saturazione fattoriale) della variabile 1 nel fattore 1
- F_1 è il punteggio standardizzato di un caso nel fattore 1
- u_1 è il punteggio standardizzato di un caso nel fattore unico (o specifico) della variabile 1

Definizioni

- L'analisi fattoriale "consiste in una serie di tecniche statistiche che hanno lo scopo di semplificare insiemi complessi di dati" (Kline, p. 8)
- "un fattore è essenzialmente una dimensione o un costrutto che descrive condensandole le relazioni tra un insieme di variabili" e "un costrutto definito operativamente dalle sue saturazioni fattoriali" (Kline, p.11)
- la "saturazione fattoriale è la correlazione tra una variabile e un fattore" (Kline, p. 11)

Traduzioni

Factor analysis (FA)	Analisi fattoriale (AF)
Principal components analyis (PCA)	Analisi delle componenti principali (ACP
Exploratory FA	AF esplorativa
Confirmatory FA	AF confermativa
Eigenvector	autovettore
Eigenvalue (characteristic root)	autovalore (radice caratteristica)
factor /component	fattore / componente
factor loadings	saturazioni
factor scores	punteggi fattoriali
pattern matrix	matrice dei modelli
structure matrix	matrice di struttura
common variance / communality	varianza comune /comunalità
unique varianza	varianza unica
random variance	varianza d'errore

Riferimenti bibliografici

- Kline, P. (1994). *An Easy Guide to Factor Analysis*. London-New York: Routledge. Trad. it. *Guida facile all'analisi fattoriale*. Roma, Astrolabio, 1997.
- Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology*, 15, 201-293.
- Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis*. Washington, DC: American Psychological Association.