Analisi Fattoriale e Analisi delle Componenti Principali

Giovanni Corradini

Free University of Bolzano

Contesto Generale

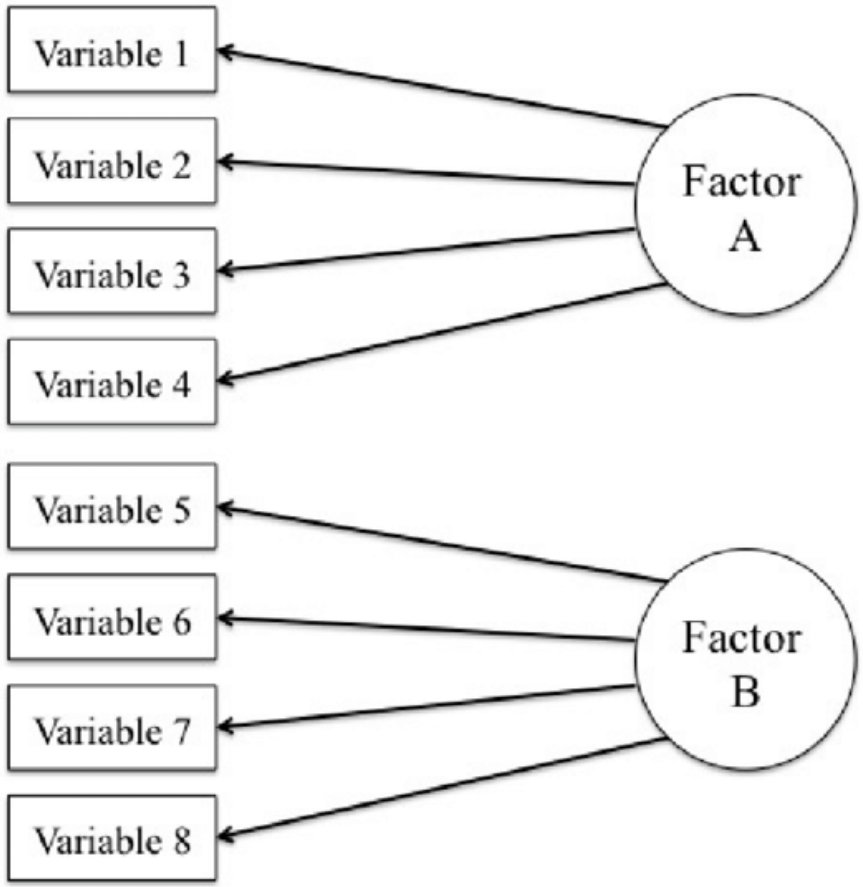
- Il **contesto** in cui ci muoviamo è quello in cui si dispone di un campione di un certo numero di individui, indicato con n, sui quali vengono misurate/osservate un certo numero di caratteristiche (le variabili), indicato con p, con $p > 2$.
- I dati vengono organizzati in una **matrice n x p**, con n righe e p colonne, chiamata anche matrice del disegno.
- I metodi seguenti hanno come assunto principale che le variabili siano sui **numeri reali**, ma vengono applicate anche su quelli **ordinali** (e.g. scala Likert).
- Il corrispettivo dell'analisi delle componenti principali e dell'analisi fattoriale per variabili **qualitative** (categoriali) è l'analisi delle corrispondenze.

Analisi Fattoriale

- L'analisi dei fattori (FA) è una tecnica che permette di descrivere la variabilità di un insieme di variabili osservate (quantitative e tendenzialmente sui reali) correlate fra loro, in termini di un **piccolo numero** di fattori sottostanti (chiamati anche fattori latenti).
- L'analisi dei fattori viene usata spesso anche quando si intende studiare una o più variabili **non misurabili** (ad es. l'intelligenza o il benessere), e si ricorre ad una serie di variabili misurabili che sono ritenute degli indicatori della variabile non osservabile (come test, questionari etc.).

Analisi Fattoriale

Schema generale di un'analisi fattoriale



Analisi Fattoriale

Ci sono due tipi di analisi fattoriale:

- **Esplorativa**: investigare le relazioni tra variabili manifeste e latenti, senza fare assunzioni su quali variabili osservate siano legate a quali fattori (a priori ignoto il numero di fattori).
- **Confermativa**: si cerca di valutare se uno specifico modello dei fattori, scegliendo a priori il numero di fattori, fornisce un buon adattamento ai dati.

Questi due tipi di analisi, sebbene differiscano per l'obiettivo, condividono la stessa interpretazione.

Analisi Fattoriale

- La determinazione del **numero di fattori** da utilizzare (FA esplorativa) o la conferma del numero di fattori scelto (FA confermativa) è molto importante perché i risultati del modello possono cambiare molto in base al numero di fattori inclusi.
- Dato che il modello viene stimato tramite la massima verosimiglianza, si dispone di un **test** per verificare che il numero di fattori introdotti nel modello sia sufficiente a descrivere la variabilità dei dati.

Analisi Fattoriale

Assumendo k fattori latenti e p variabili osservate, con $k < p$, il **modello fattoriale** si può scrivere come:

$$
\begin{aligned}
x_1 &= \mu_1 + \lambda_{11} f_1 + \lambda_{12} f_2 + \ldots + \lambda_{1k} f_k + u_1 \\
x_2 &= \mu_2 + \lambda_{21} f_1 + \lambda_{22} f_2 + \ldots + \lambda_{2k} f_k + u_2 \\
&\vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \\
x_p &= \mu_p + \lambda_{p1} f_1 + \lambda_{p2} f_2 + \ldots + \lambda_{pk} f_k + u_p
\end{aligned}
$$

Dove $x = (x_1, \ldots, x_p)$ è il vettore delle variabili osservate, $\mu = (\mu_1, \ldots, \mu_p)$ è il vettore delle medie delle variabili osservate x, $f = (f_1, \ldots, f_k)$ il vettore dei fattori latenti che si relazione alle x tramite la matrice dei Λ che contiene i vari **pesi fattoriali** λ_{ij} (i *loadings*), per $i = 1, \ldots, p$ e per $j = 1, \ldots, k$, e $u = (u_1, \ldots, u_p)$ il vettore delle **unicità** delle x (le *uniquenesses*).

Analisi Fattoriale

- Il peso fattoriale λij rappresenta quanto l'$i-esima$ variabile osservata dipende dal $j-esimo$ fattore latente: è la quantità di maggior interesse nel modello in quanto, per avere una **buona interpretabilità**, vorremmo avere che ogni x_i avesse un λij elevato in relazione a solamente un fattore.
- l'unicità u_i indica la parte di variabilità di x_i che non viene colta dai fattori: se questa quantità è molto bassa per tutte le variabili osservate (situazione ideale), significa che il modello fattoriale ha colto quasi tutta la **variabilità dei dati**.

Analisi Fattoriale

- Dato che il modello viene stimato utilizzando la matrice di varianza e covarianza (e non correlazione) delle variabili osservate, queste devono essere sulla stessa scala: se non lo sono bisogna **standardizzarle**.
- Standardizzazione per rendere le variabili osservate confrontabili: $(x_i - \mu_i)/\sigma_i$. Questo affinché tutte le variabili osservate abbiano media pari a zero e varianza unitaria

Analisi Fattoriale

- Non esiste una **soluzione unica** per il modello dei fattori, in quanto si possono ruotare e trovare soluzioni altrettanto valide, ma con cui si possono "separare" meglio fattori e variabili osservate così da garantire una maggiore interpretabilità.
- Con **rotazioni** ortogonali come *varimax* i fattori rimangono incorrelati, mentre con rotazioni oblique come *oblimin* no.

Analisi Fattoriale

- Spesso è utile vedere come ogni fattore si manifesta tra i vari soggetti e quindi mappare l'insieme delle variabili osservate sul piano dei fattori.
- Per fare ciò si utilizzano i **punteggi fattoriali** e i più comunemente utilizzati sono quelli di Bartlett e quelli di Thompson.

Analisi delle Componenti Principali

- L'analisi delle componenti principali (PCA) serve a riassumere p variabili osservate attraverso k variabili (le componenti principali), con $k < p$: queste PC sono **incorrelate** tra di loro si ottengono attraverso **trasformazioni algebriche** applicate ai dati originari.
- La derivazione delle PC avviene sequenzialmente, ottenendo variabili che hanno **importanza via via decrescente**, con la prima PC che è quella che discrimina maggiormente le osservazioni.
- Anche qui la scala delle variabili originali è di fondamentale importanza e pertanto se non sono sulla stessa scala (e quindi se non hanno dimensioni confrontabili) bisogna standardizzarle.

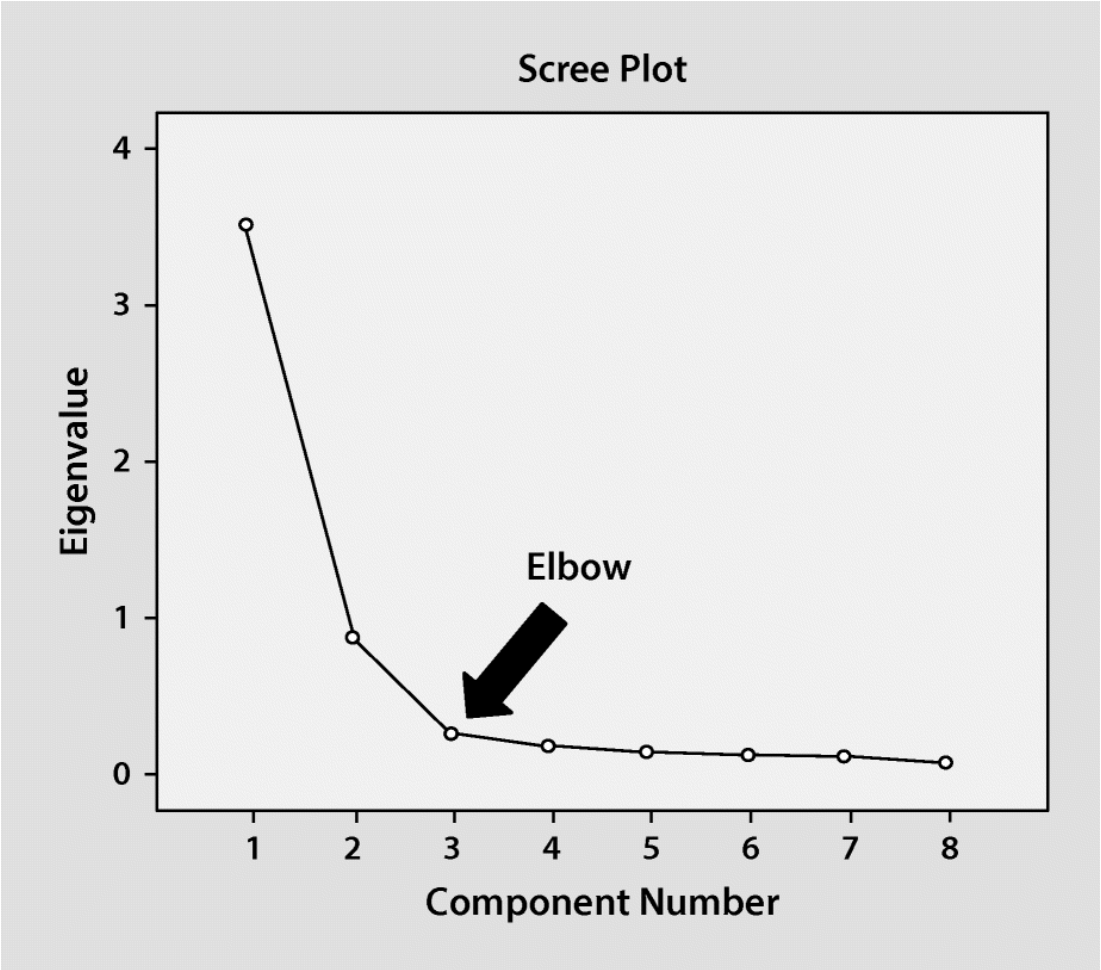
Analisi delle Componenti Principali

A differenza della FA, non si passa direttamente da p variabili osservate a k fattori latenti, ma si modificano i dati originali ottenendo p componenti principali e quindi bisogna scegliere il **numero di PC** da tenere. I 3 metodi più usati sono:

- Scegliere le prime k PC che contribuiscono a spiegare il 70 – 80 % della variabilità dei dati iniziali.
- Scegliere le k PC che hanno varianza superiore ad 1 (regola del Kaiser)
- Scegliere le prime k PC tali che la k – *esima* PC sia relativa ad un gomito nello screeplot, ovvero nel grafico che mette in relazione le PC ordinate con la varianza che spiegano

Analisi delle Componenti Principali

Schema generico di uno screeplot con relativo gomito



PCA vs FA

Sia l'analisi delle componenti principali sia l'analisi dei fattori tentano di spiegare un insieme di dati con un ridotto numero di dimensioni. Ma ci sono importanti differenze:

- La PCA è solo una **trasformazione dei dati**, non fa alcuna assunzione sulla forma della matrice di covarianza, mentre nell'analisi dei fattori si assume un **modello** ben definito.
- Nella PCA si trasformano p variabili osservate in p componenti principali e poi si scelgono le prime k componenti, mentre nella FA si trasformano k variabili latenti in p variabili osservate.
- Nella PCA non serve calcolare "a parte" gli **scores** (come nella FA), ma sono direttamente disponibili in quanto si trasforma la matrice $n \times p$ originaria in un'altra matrice $n \times p$ che ha come variabili le PC... Questi scores possono essere utilizzati come punto di partenza per altre analisi, come analisi di regressione, test, machine learning, clustering...

PCA vs FA

Quando usare la PCA e quando la FA?

- La PCA permette di trovare le componenti che spiegano la maggior quantità di varianza nel minor numero di variabili possibile, e questo la rende più utile per **riduzione della dimensionalità** (usata in modellistica, clustering, machine learning . . .)
- La FA invece, grazie alla sua possibilità di applicare rotazioni differenti ai fattori, la rende uno strumento più utile nei contesti dove si vuole dare un'**interpretazione** alle variabili latenti trovate (contesti come psicologia, marketing . . .)

P.S. Se il modello fattoriale è valido e le unicità delle variabili osservate sono piccole, tendenzialmente PCA e FA danno risultati simili.