

Analisi fattoriale esplorativa

Carla Rampichini

rampichini@ds.unifi.it
Dipartimento di Statistica "G. Parenti"
Università di Firenze

Indice

- 1 Introduzione
- 2 Modello fattoriale ortogonale
 - Struttura di covarianza
 - Interpretazione del modello
 - Indeterminatezza soluzione AF
- 3 Metodi di stima

Analisi fattoriale

Paradigma comune a molte tecniche di analisi multivariata: modellare l'informazione rilevante (rappresentata da una variabile multivariata $\mathbf{X}$) con un numero limitato di *fattori* latenti

Esempio

- In un'indagine sui consumi delle famiglie, viene registrato il livello dei consumi mensili $\mathbf{X}$ di p beni durevoli.
- La variabilità e la covarianza delle p componenti di $\mathbf{X}$ possono essere spiegate da due o tre fattori di comportamento sociale della famiglia: il desiderio di comfort, il tentativo di raggiungere un certo livello sociale, o altri concetti sociali latenti possono spiegare la maggior parte dei comportamenti di consumo.
- I sociologi sono più interessati a questi *fattori non osservabili* che alle p variabili osservate $\mathbf{X}$, perchè forniscono una migliore comprensione del comportamento delle famiglie.

L'analisi dei fattori interessa molti settori: psicologia, marketing, economia, etc.

Obiettivi dell'analisi fattoriale (AF)

- Come nell'ACP, l'intento dell'AF è quello di *ridurre la dimensione dei dati* osservati
- La prospettiva però è diversa: si assume che esista un *modello*, detto **Modello Fattoriale**
- Il modello assume che le covarianze tra le p variabili di $\mathbf{X}$ possano essere spiegate tramite un numero limitato di fattori latenti.

Obiettivo dell'AF

Costruire un modello statistico che *spieghi* la correlazione tra le variabili osservate in termini di uno o più *fattori latenti*.

Argomenti trattati

- Esamineremo per primo il **Modello fattoriale ortogonale**, mostrando che non esiste un'unica soluzione.
- Mostreremo come trarre vantaggio da questa non unicità della soluzione per derivare delle tecniche che consentano di ottenere dei risultati più interpretabili.
- Vedremo che queste tecniche utilizzano delle **rotazioni** (geometriche) dei fattori.
- Mostreremo varie tecniche di stima del modello e definiremo una procedura di rotazione ottimale.
- Seguiremo un approccio empirico e vedremo diversi esempi.

Specificazione del modello fattoriale

Obiettivo dell'AF è descrivere la **matrice di covarianza** delle p variabili in $\mathbf{X}$ in termini di uno o più **fattori inosservabili**.

I fattori sono interpretati come **caratteristiche** latenti (non osservate) **comuni** delle $\mathbf{x} \in \mathbb{R}^p$.

- Sia $\mathbf{X}$ un vettore casuale a p componenti, di media μ e matrice di covarianza Σ
- Il modello fattoriale assume che le $\mathbf{X}$ siano linearmente dipendenti
 - da (pochi) fattori latenti $F_1, \dots, F_m$ detti **fattori comuni**,
 - p **fattori specifici**, o errori, $\varepsilon_1, \dots, \varepsilon_p$
- P.e. $\mathbf{X}$ può essere il vettore relativo a p punteggi di un test di intelligenza, che hanno come fattore latente comune il livello generale di intelligenza.
- Nel marketing, le $\mathbf{X}$ possono essere p item di un questionario sul livello di soddisfazione dei consumatori, che possono essere spiegate da fattori latenti comuni quali il livello di attrattività del prodotto, l'immagine della marca, ecc.

Modello fattoriale

$$\begin{aligned}X_1 &= \mu_1 + \ell_{11}F_1 + \dots + \ell_{1m}F_m + \varepsilon_1 \\X_2 &= \mu_2 + \ell_{21}F_1 + \dots + \ell_{2m}F_m + \varepsilon_2 \\&\vdots \\X_p &= \mu_p + \ell_{p1}F_1 + \dots + \ell_{pm}F_m + \varepsilon_p\end{aligned}$$

- ℓ_{kj} sono i *factor loadings*
- F_j sono i *fattori comuni*
- ε_k sono i *fattori specifici*

Modello AF in forma matriciale

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\mu} + \mathbf{LF} + \boldsymbol{\varepsilon} \\ \mathbf{X} - \boldsymbol{\mu} &= \mathbf{LF} + \boldsymbol{\varepsilon}\end{aligned}$$

- $\mathbf{X}$ è il vettore matrice $p \times 1$ delle v.c., con vettore delle medie $\boldsymbol{\mu}$
- $\mathbf{L}$ è la matrice $p \times m$ dei coefficienti dei fattori, detti *factor loadings*
- $\mathbf{F}$ è il vettore $m \times 1$ dei *fattori comuni*
- $\boldsymbol{\varepsilon}$ è il vettore $p \times 1$ dei *fattori specifici*, ciascuno associato ad una sola variabile
- Ci sono $m + p$ fattori non osservabili $\Rightarrow$ il modello AF non è direttamente verificabile dai dati.
- $\Rightarrow$ fare ipotesi su $\mathbf{F}$ e $\boldsymbol{\varepsilon}$ che implicano una particolare struttura di $\boldsymbol{\Sigma}$ da verificare tramite i dati.

Assunzioni sui fattori latenti

Nel modello AF ortogonale con m fattori comuni, si fanno le seguenti assunzioni

- I fattori comuni
 - hanno media nulla $E(\mathbf{F}) = \mathbf{0}$ e varianza unitaria
 - sono incorrelati $E(\mathbf{F}\mathbf{F}') = \mathbf{I}$
 - sono incorrelati con i fattori specifici $Cov(\boldsymbol{\varepsilon}, \mathbf{F}) = E(\boldsymbol{\varepsilon}\mathbf{F}') = \mathbf{0}$
- i fattori specifici
 - hanno media nulla $E(\boldsymbol{\varepsilon}) = \mathbf{0}$
 - sono incorrelati con matrice di covarianza

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_p \end{pmatrix}$$

Modello fattoriale ortogonale

$$\underset{(p \times 1)}{\mathbf{X}} = \underset{(p \times 1)}{\boldsymbol{\mu}} + \underset{(p \times m)}{\mathbf{L}} \underset{(m \times 1)}{\mathbf{F}} + \underset{(p \times 1)}{\boldsymbol{\varepsilon}}$$

μ_k = media della k -ma variabile

ε = k -mo fattore specifico

F_j = j -mo fattore comune

ℓ_{kj} = loading della k -ma variabile sul j -mo fattore

con $j = 1, \dots, m, k = 1, \dots, p$.

- I vettori dei fattori latenti $\mathbf{F}$ e $\boldsymbol{\varepsilon}$ soddisfano le ipotesi:
 - $\mathbf{F}$ e $\boldsymbol{\varepsilon}$ sono indipendenti
 - $E(\mathbf{F}) = \mathbf{0}, \text{Cov}(\mathbf{F}) = \mathbf{I}$
 - $E(\boldsymbol{\varepsilon}) = \mathbf{0}, \text{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi} = \text{diag}\{\psi_j\}$
- Il modello e le ipotesi implicano una particolare struttura della matrice di covarianza $\boldsymbol{\Sigma}$ di $\mathbf{X}$

Struttura della matrice di covarianza

Il modello AF ortogonale implica una particolare struttura della matrice di covarianza Σ

- Ricordando il modello $\mathbf{X} - \mu = \mathbf{LF} + \varepsilon$ possiamo scrivere:

$$\begin{aligned}(\mathbf{X} - \mu)(\mathbf{X} - \mu)' &= (\mathbf{LF} + \varepsilon)(\mathbf{LF} + \varepsilon)' \\ &= \mathbf{LF}(\mathbf{LF})' + \varepsilon(\mathbf{LF})' + \mathbf{LF}\varepsilon' + \varepsilon\varepsilon'\end{aligned}$$

- Considerando i valori attesi:

$$\begin{aligned}\Sigma &= E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)'] \\ &= \mathbf{LE}[\mathbf{FF}']\mathbf{L}' + E[\varepsilon(\mathbf{F})'](\mathbf{L}') + \mathbf{LE}[\mathbf{F}\varepsilon'] + E[\varepsilon\varepsilon'] \\ &= \mathbf{LL}' + \Psi\end{aligned}$$

(ricordare che per il modello AF ortogonale $\text{Cov}(\mathbf{F}\varepsilon') = \mathbf{0}$)

Matrice di covarianza dal modello AF

$$\Sigma = \mathbf{LL}' + \Psi = \begin{pmatrix} \sum_{j=1}^m \ell_{1j}^2 + \psi_1 & \sum_{j=1}^m \ell_{1j}\ell_{2j} & \cdots & \sum_{j=1}^m \ell_{1j}\ell_{pj} \\ \sum_{j=1}^m \ell_{2j}\ell_{1j} & \sum_{j=1}^m \ell_{2j}^2 + \psi_2 & \cdots & \sum_{j=1}^m \ell_{2j}\ell_{pj} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^m \ell_{pj}\ell_{1j} & \sum_{j=1}^m \ell_{pj}\ell_{2j} & \cdots & \sum_{j=1}^m \ell_{pj}^2 + \psi_p \end{pmatrix}$$

- Il modello fattoriale spiega la maggior parte della **varianza** di $\mathbf{X}$ attraverso un piccolo numero di fattori latenti $\mathbf{F}$ comuni alle sue p componenti: $\text{Var}(X_k) = \sum_{j=1}^m \ell_{kj}^2 + \psi_k$,
- e spiega completamente la **covarianza** tra le X_k :
 $\text{Cov}(X_k, X_s) = \sum_{j=1}^m \ell_{kj}\ell_{sj}$.
- I fattori specifici consentono di aggiustare il modello per catturare la **variabilità residua** ψ_j , non spiegata dai fattori comuni.

Osservazioni

- Il modello fattoriale si basa sulle assunzioni specificate
- Se le assunzioni non sono valide, l'analisi fornisce dei risultati spuri.
- Sebbene ACP e FA sembrino simili, la loro natura è molto diversa. Le CP:
 - sono *trasformazioni lineari* delle $\mathbf{X}$
 - costruite in modo tale da avere varianza massima
 - e con l'obiettivo di ridurre la dimensione dei dati.
- Nell'*analisi fattoriale*
 - si cerca di spiegare le variazioni di $\mathbf{X}$ utilizzando una trasformazione lineare di un numero fisso, limitato, di *fattori latenti*.
 - con l'obiettivo di trovare i *loadings* ℓ_{kj} e le varianze specifiche ψ_j
 - Le stime di $\mathbf{L}$ e $\mathbf{\Psi}$ sono dedotte dalla struttura di covarianza
$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi}$$

Interpretazione dei fattori

Se un modello fattoriale con m fattori risulta essere ragionevole, cioè se i fattori spiegano la maggior parte della (co)varianza delle p misure considerate, risulta naturale chiedersi cosa rappresentino questi fattori.

- Per interpretare i fattori si considerano le correlazioni tra le variabili osservate X_k e i fattori F_j .
- osserviamo che in base al modello si ha
 $(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}' = (\mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon})\mathbf{F}' = \mathbf{L}\mathbf{F}\mathbf{F}' + \boldsymbol{\varepsilon}\mathbf{F}'$
- quindi la covarianza tra fattori latenti e variabili osservate è pari a:

$$\text{Cov}(\mathbf{X}, \mathbf{F}) = \mathbf{L}E(\mathbf{F}\mathbf{F}') + E(\boldsymbol{\varepsilon}\mathbf{F}') = \mathbf{L}$$

- da cui si deriva la correlazione $\text{Corr}(\mathbf{X}, \mathbf{F}) = \mathbf{D}^{-1/2}\mathbf{L}$, con
 $\mathbf{D} = \text{diag}\{\sigma_{kk}\}$

Ipotesi di linearità

L'ipotesi di *linearità* è fondamentale nel modello fattoriale classico.

- Il modello $\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon}$ è *lineare* rispetto ai fattori comuni.
- Se le X sono legate ai fattori latenti ma la relazione è NON lineare, p.e. $X_1 - \mu_1 = \ell_{11}F_1F_3 + \varepsilon_1$, allora la struttura di covarianza $\mathbf{L}\mathbf{L} + \boldsymbol{\Psi}$ non è adeguata

Comunalità e specificità

Il modello scompone la varianza di X_k in due quote

$$\underbrace{\sigma_{kk}}_{\text{Var}(X_k)} = \underbrace{\ell_{k1}^2 + \ell_{k2}^2 + \dots + \ell_{km}^2}_{\text{comunalità}} + \underbrace{\psi_k}_{\text{specificità}}$$

- la proporzione di varianza di X_k spiegata dagli m fattori comuni è detta k -ma **comunalità** e si indica con h_k^2
- la proporzione di varianza dovuta al fattore specifico ψ_k è detta **specificità**, o unicità
- $\sigma_{kk} = h_k^2 + \psi_k$
- vediamo un esempio numerico

Quanti parametri?

Σ ha p varianze e $\frac{p(p-1)}{2}$ covarianze



- in Σ ci sono in tutto $p + \frac{p(p-1)}{2} = \frac{2p+p^2-p}{2} = \frac{p(p+1)}{2}$ parametri
- il modello fattoriale ha
 - $p \times m$ parametri in $\mathbf{L}$
 - p specificità in Ψ
- totale parametri modello fattoriale: $m \times p + p = p(m+1)$

Tante variabili quanti fattori: $m = p$

Caso particolare: se $m = p$ la matrice Σ può essere riprodotta esattamente da $\mathbf{L}\mathbf{L}'$ con $\Psi = \mathbf{0}$

dimostrazione

- Poichè Σ è definita positiva, vale la scomposizione spettrale

$$\Sigma = \sum_{k=1}^p \lambda_k \mathbf{e}_k \mathbf{e}_k' = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$$

dove $\mathbf{\Lambda} = \text{diag}\{\lambda_k\}$, $\mathbf{P}' = [\mathbf{e}_1, \dots, \mathbf{e}_p]$

- ponendo $\mathbf{L} = \mathbf{P}\mathbf{\Lambda}^{1/2}$, con $\mathbf{\Lambda}^{1/2} = \text{diag}\{\sqrt{\lambda_k}\}$
- si ottiene

$$\Sigma = \mathbf{L}\mathbf{L}' = \mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}\mathbf{P}' = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$$

Modello parsimonioso

Se $m \ll p$ si ha una descrizione *parsimoniosa* di Σ .

Esempio:

- $p = 6$ variabili, $m = 2$ fattori comuni
- Σ ha 6 varianze e $(6 \times 5)/2 = 15$ covarianze $\Rightarrow 15 + 6 = 21$ parametri
- $\mathbf{LL}' + \Psi$ ha $m \times p = 2 \times 6 = 12$ loadings e $p = 6$ specificità $\Rightarrow 12 + 6 = 18$ parametri
- Non sempre è possibile trovare $m \ll p$ fattori comuni che rappresentino la struttura di covarianza $\Sigma = \mathbf{LL}' + \Psi$
- quindi non sempre esiste una soluzione (si veda es. pag. 486 JW)

Soluzione per $m > 1$

Quando $m > 1$ la soluzione del modello fattoriale è sempre *ambigua*.

- Sia $\mathbf{T}$ una matrice ortogonale, cioè: $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$
- Il modello fattoriale $\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon}$ può essere trasformato

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}^*\mathbf{F}^*\boldsymbol{\varepsilon}$$

- dove $\mathbf{L}^* = \mathbf{L}\mathbf{T}$ e $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$
- Questa trasformazione lascia immutate le proprietà del modello, infatti si ha
 - $E(\mathbf{F}^*) = \mathbf{T}'E(\mathbf{F}) = \mathbf{0}$
 - $Cov(\mathbf{F}^*) = \mathbf{T}'Cov(\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{I}$

indeterminatezza soluzione AF

È impossibile scegliere tra i *factor loadings* $\mathbf{L}$ e $\mathbf{L}^*$ solo sulla base delle osservazioni $\mathbf{X} \Leftrightarrow$ i fattori comuni $\mathbf{F}$ hanno le stesse proprietà statistiche dei fattori trasformati $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$

Non unicità della soluzione per $m > 1$

I fattori comuni $\mathbf{F}$ hanno le stesse proprietà statistiche dei fattori trasformati $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$

- i *loadings* $\mathbf{L}$ sono in generale diversi dai *loadings* trasformati $\mathbf{L}^* = \mathbf{L}\mathbf{T}$
- ma generano lo *stesso modello* per la matrice di covarianza Σ !

$$\Sigma = \mathbf{L}\mathbf{L}' + \Psi = \mathbf{L} \underbrace{\mathbf{T}\mathbf{T}'}_{\mathbf{I}} \mathbf{L}' + \Psi = (\mathbf{L}^*)(\mathbf{L}^*)' + \Psi$$

- Questa molteplicità di soluzioni possibili giustifica la *rotazione* dei fattori, perchè la trasformazione attraverso una matrice ortogonale corrisponde a una rotazione delle coordinate per $\mathbf{X}$

Factor loadings e comunaltà

I *factor loadings* $\mathbf{L}$ sono determinati a meno della matrice ortogonale $\mathbf{T}$

$$\Rightarrow \mathbf{L} \text{ e } \mathbf{L}^* = \mathbf{L}\mathbf{T}$$

- $\mathbf{L}$ e $\mathbf{L}^*$ forniscono la stessa rappresentazione di Σ
- anche le comunaltà generate dalle possibili soluzioni trasformate sono le stesse, infatti le comunaltà sono date dalla diagonale di $\mathbf{L}^*(\mathbf{L}^*)'$, e per qualunque matrice ortogonale $\mathbf{T}$ vale:

$$\mathbf{L}^*(\mathbf{L}^*)' = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{L}' = \mathbf{L}\mathbf{L}'$$

Come si procede?

- 1 si impongono dei *vincoli* per ottenere una soluzione unica per $\mathbf{L}$ e Ψ
- 2 Si *ruota* la matrice dei *factor loadings* $\mathbf{L}$, moltiplicando $\hat{\mathbf{L}}$ per una matrice ortogonale $\mathbf{T}$: $\mathbf{L}^* = \mathbf{L}\mathbf{T}$
- 3 si sceglie $\mathbf{T}$ in modo da ottenere fattori più facili da interpretare
- 4 Ottenuti i *loadings* e le *specificità*, i fattori risultano identificati e si possono stimare i punteggi fattoriali (*scores*).

Stima del modello AF

- Date n osservazioni p -dimensionali $\mathbf{x}_1, \dots, \mathbf{x}_n$ l'analisi fattoriale cerca di rispondere alla domanda:
Il modello fattoriale ortogonale con un numero $q \ll p$ può rappresentare adeguatamente i dati?
- Da un punto di vista statistico, si tratterà di verificare se vale la relazione $\text{Cov}(\mathbf{X}) = \mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi}$
- La matrice campionaria $\mathbf{S}$ è uno stimatore di $\mathbf{\Sigma}$
- Il modello fattoriale non è appropriato se le covarianze in $\mathbf{S}$ sono piccole, o analogamente se le correlazioni in $\mathbf{R}$ sono vicine a zero: in tal caso le specificità assumono il peso maggiore e non è possibile determinare pochi fattori comuni rilevanti.
- Se la matrice $\mathbf{\Sigma}$ ha covarianze significativamente diverse da zero, il modello fattoriale è appropriato. Si tratterà prima di tutto di stimare i factor loadings e le specificità.

Metodi di stima

I metodi di stima più utilizzati sono:

- metodo delle componenti principali
- metodo della massima verosimiglianza
- la soluzione ottenuta può essere *ruotata* per *semplificare* l'*interpretazione* dei fattori
- è sempre meglio provare entrambi i metodi di stima: se il modello fattoriale è adeguato le due soluzioni risultano coerenti.
- La stima e la rotazione dei fattori richiedono metodi di calcolo iterativi che devono essere svolti con un computer.
- Un primo esempio: wisc.sas
- Manuale SAS: statugfactor.pdf