Università degli studi di Ferrara
Dipartimento di Matematica
A.A. 2018/2019 – I semestre

STATISTICA MULTIVARIATA

SSD MAT/06

analisi dell'interdipendenza

LEZION 9 – Analisi fattoriale esplorativa

Docente: Valentina MINI

valentina.mini@unife.it

RICEVIMENTO: su appuntamento previa mail

Indice della lezione

1. Introduzione
2. Intuizione
3. Trattazione teorica e formalizzazione
4. Procedimento analitico da seguire nella pratica
5. Esemplificazione e esercitazione in R

1 -INTRODUZIONE

Introduzione

- Abbiamo visto la ACP
- Oggi vediamo la AF metodo simile, **basata su un diverso modello** di base chiamato "*modello a fattori comuni*"
- Spesso usata per obiettivi simili a ACP
- Sottolineiamo le differenze
  - Esplicite assunzioni su come ogni variabile del dataset è misurata
  - Modello: la varianza osservata è attribuibile ad un piccolo numero di fattori comuni
- Obiettivo: identificare i fattori comuni e spiegare la loro relazione con i dati osservati

Introduzione

- obiettivo: ridurre il numero di variabili esplicative attraverso la creazione di nuove variabili chiamate fattori

- Metodo: trasformazione della struttura dei dati osservati in una nuova struttura tale che la variabilità dei dati è spiegata dai fattori

Introduzione

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

Esempio

In un'indagine sui consumi delle famiglie, viene registrato il livello dei consumi mensili **X** di *p* beni durevoli.

La variabilità e la covarianza delle *p* componenti di **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 **X**, perchè forniscono una migliore comprensionedel comportamento delle famiglie.

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

2 - INTUIZIONE

Proprietà dei fattori:

- Non correlati tra loro

- Variabili latenti non osservate (sconosciute a priori) che riproducono le correlazioni esistenti tra variabili originarie

- Modello sottostante

assunzioni:

- FA può essere applicata ad un set di variabili numeriche standardizzabili

- Numero di unità statistiche dovrebbe essere almeno 5 volte il numero delle variabili originarie
  (per ogni x → almeno 5 n)

Factor Model

- $X_1,\ldots,X_k$ variabile risposta, tale che

  - $E(X_j)= \mu_j$, $Var(X_j)= \sigma_{jj} = \sigma_j^2$, $Cov(X_jX_r)= \sigma_{jr}$; $j,r = 1,..,k$

- $X_j = \lambda_{j1} F_1 + \lambda_{j2} F_2 + \ldots + \lambda_{js} F_s + \ldots + \lambda_{jq} F_q + U_j + \mu_j$,
  $= \Sigma_s \lambda_{js} F_s + U_j + \mu_j$, $j=1,\ldots,k$

  - λ_{j1} , λ_{j2} , $\ldots$, λ_{jq} $(j=1,\ldots,k)$: parametero (constanti) definito peso fattoriale

  - F_1 , F_2 ,…, F_q fattori comuni (random variables)
  - U_j , fattore unico o specifico $(j=1,\ldots,k)$

Modello fattoriale

- assunzioni:

✓ $E(F_s)=0$,	$s=1,\ldots,q$
✓ $Var(F_s)=1$,	$s=1,\ldots,q$
✓ $Cov(F_s, F_t)=0$,	$s,t=1,\ldots,q;\ s \neq t$
✓ $E(U_j)=0$,	$j=1,\ldots,k$
✓ $Var(U_j)= {}_u\sigma_{jj} = {}_u\sigma_j^2$	$j=1,\ldots,k$
✓ $Cov(U_j, U_r)=0$	$j,r=1,\ldots,k;\ j \neq r$
✓ $Cov(F_s, U_j)=0$	$s=1,\ldots,q;$ $j=1,\ldots,k$

Modello fattoriale

Rappresentazione matriciale:

- $\boldsymbol{X} = [X_1, \ldots, X_k]'$ Vettore di variabili risposta
- $\boldsymbol{F} = [F_1, \ldots, F_q]'$ Vettore di fattori comuni
- $\boldsymbol{U} = [U_1, \ldots, U_k]'$ Vettore di fattori unici
- $\Lambda = [\lambda_{js}]$ *k× q matrice di costanti (parameters)*
- $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]'$ Vettore delle medie

- $\boldsymbol{X} = \Lambda \boldsymbol{F} + \boldsymbol{U} + \boldsymbol{\mu}$

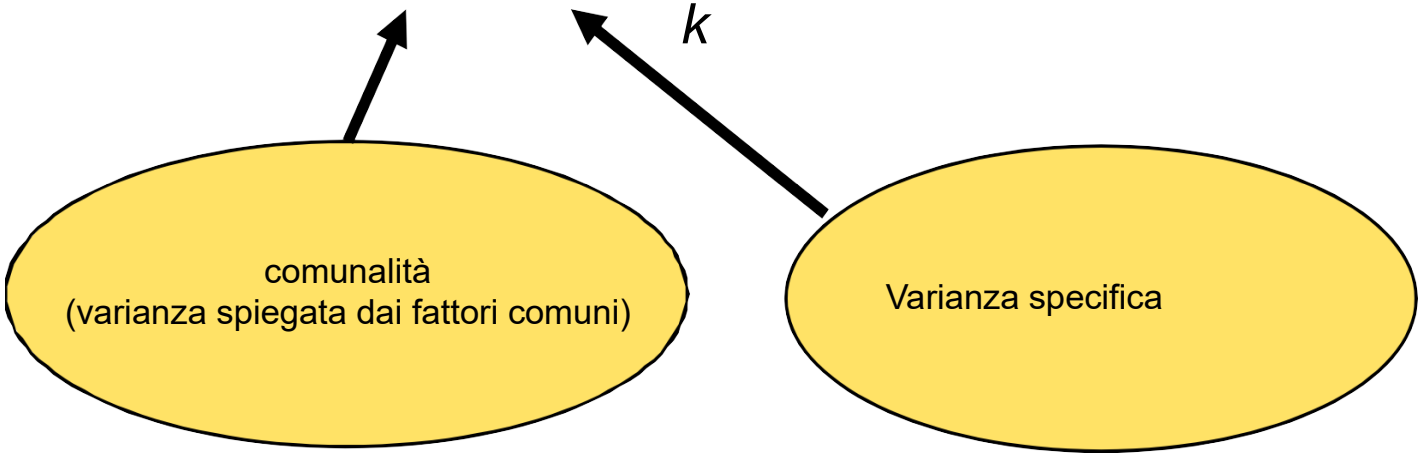
  - $\boldsymbol{E(X)= \mu}$, $\boldsymbol{Var(X)= \Sigma = [\sigma_{jr}]}$
  - $\boldsymbol{E(F)= 0}$, $\boldsymbol{Var(F)= I} = diag(1,1,\ldots,1)$
  - $\boldsymbol{E(U)= 0}$, $\boldsymbol{Var(U)=} diag({}_u\sigma_{11}, \ldots, {}_u\sigma_{kk}) = {}_u\Sigma$
  - $\boldsymbol{Cov(F,U)= 0}$

Modello fattoriale

Scomposizione della varianza (VD):

$Var(X_j) = \quad = \Sigma_s \lambda_{js}^2 + {}_u\sigma_{jj}$

$\sigma_{jj} \quad = h_j^2 + {}_u\sigma_{jj}, \quad j=1,\ldots,k$



$\lambda_{js} = E(X_j,F_s) = Cov(X_j,F_s) \rightarrow$ misura della dipendenza lineare tra X_j e F_s

Con notazione matriciale: $\Sigma = \Lambda\Lambda' + {}_u\Sigma$

Modello fattoriale

- FA applicabile a variabili numeriche standardizzabili
- Il numero di unità deve essere almeno 5volte il numero delle variabili originarie: $n \geq 5 \times k$
- I fattori comuni devono spiegare almeno il 70% della variabilità totale delle variabili originarie
- Il problema di identificazione di $\boldsymbol{F}$ **e** Λ non ha una unica soluzione

Stima dei parametri e rotazione fattoriale FA

Se le assunzioni del modello sono rispettate per ***F***, allora una rotazione di **F** definisce nuovi fattori **F*** per i quali le assunzioni sono ancora vere e per i quali si hanno diversi pesi fattoriali (factor loadings) Λ^* .

Formalmente:

Data la matrice ortogonale $q \times q$ = ***G*** (tale per cui ***GG′*** = ***I***)

$$
\begin{aligned}
\boldsymbol{X} = \quad & \Lambda \boldsymbol{F} + \boldsymbol{U} + \boldsymbol{\mu} = \\
& = \Lambda \boldsymbol{GG'} \boldsymbol{F} + \boldsymbol{U} + \boldsymbol{\mu} = \\
& = (\Lambda \boldsymbol{G})(\boldsymbol{G'} \boldsymbol{F}) + \boldsymbol{U} + \boldsymbol{\mu} = \\
& = \Lambda^* \boldsymbol{F^*} + \boldsymbol{U} + \boldsymbol{\mu}
\end{aligned}
$$

Stima dei parametri e rotazione fattoriale FA

- Metodi di rotazione fattoriale:
  - Varimax: rotazione ortogonale
  - Equimax
  - Quartimax
  - …

3 -TEORIA E FORMALIZZAIZONE

Obiettivo analisi fattoriale

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 **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*.

Obiettivo analisi fattoriale

Esamineremo per primo il Modello fattoriale ortogonale, mostrando che non esiste un'unica soluzione.

Mostreremo come trarre vantaggio da questa non unicità della soluzioneper 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.

Specificazione del modello fattoriale

Obiettivo dell'AF è descrivere lamatrice 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 $\mathbf{x} \in R^p$.

Sia $\mathbf{X}$ un vettore casuale con p componenti, di media $\boldsymbol{\mu}$ e matrice di covarianza $\boldsymbol{\Sigma}$

Il modello fattoriale assume che le $\mathbf{X}$ siano linearmente dipendenti

da (pochi) fattori latenti $F_1, \ldots, F_m$ detti *fattori comuni*, p*fattori specifici* , o errori, $\varepsilon_1, \ldots, \varepsilon_p$

P.e. $\mathbf{X}$ può essere il vettore realtivo 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, spiegabili da fattori latenti comuni quali il livello di attrattività del prodotto, l'immagine della marca, ecc.

Modello fattoriale

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

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

Modello AF in forma matriciale

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

- $\mathbf{X}$ è il vettore matrice $p \times 1$ delle v.c., con vettore delle medie μ
- $\mathbf{L}$ è la matrice $p \times m$ dei coefficienti dei fattori, detti *factor loadings*
- $\mathbf{F}$ è il vettore $m \times 1$ dei *fattori comuni*
- ε 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 ε che implicano una particolare struttura di $\mathbf{\Sigma}$ 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 $Cov(\varepsilon, \mathbf{F}) = E(\varepsilon\mathbf{F}') = \mathbf{0}$
- i fattori specifici
  - hanno media nulla $E(\varepsilon) = \mathbf{0}$
  - 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 $j = 1, \ldots, m$, $k = 1, \ldots, p$.

- I vettori dei fattori latenti F e ε soddisfano le ipotesi:
  - $\mathbf{F}$ e ε sono indipendenti
  - $E(\mathbf{F}) = \mathbf{0}$, $Cov(\mathbf{F}) = \mathbf{I}$
  - $E(\varepsilon) = \mathbf{0}$, $Cov(\varepsilon) = \boldsymbol{\Psi} = diag\{\Psi_j\}$
- 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 ortogononale implica una particolare struttura della matrice di covarianza $\mathbf{\Sigma}$

- 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:

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

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

Matrice di covarianza dal modello di analisi fattoriale descrittiva

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

- Il modello fattoriale spiega la maggior parte della varianza di **X** attraverso un piccolo numero di fattori latenti **F** comuni alle sue p componenti: $Var(X_k) = \sum_{j=1}^{m} \ell_{kj}^2 + \psi_k$,
- e spiega completamente la covarianza tra le X_k: $Cov(X_k, X_s) = \sum_{j=1}^{m} \ell_{kj}\ell_{sj}$.
- 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 **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 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

$$(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}^{j} = (\mathbf{LF} + \boldsymbol{\varepsilon})\mathbf{F}^{j} = \mathbf{LFF}^{j} + \boldsymbol{\varepsilon}\mathbf{F}^{j}$$

quindi la covarianza tra fattori latenti e variabili osservate è pari a:

$$Cov(\mathbf{X}, \mathbf{F}) = \mathbf{L}E(\mathbf{FF}^{j}) + E(\boldsymbol{\varepsilon}\mathbf{F}^{j}) = \mathbf{L}$$

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

Ipotesi di linearità

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

- Il modello $\mathbf{X} - \mu = \mathbf{LF} + \varepsilon$ è *lineare* rispetto ai fattori comuni.
- Se le X sono legate ai fattori latenti ma la relazione è NON lineare, p.e. $X_1 - \mu_1 = \ell_{11} F_1 F_3 + \varepsilon_1$, allora la struttura di covarianza $\mathbf{LL} + \mathbf{\Psi}$ non è adeguata

Il concetto e la funzione di comunalità

Il modello scompone la varianza di X_k in due quote

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

- 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

Scelta del numero dei parametri

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

$\Downarrow$

in $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Psi}$

totale parametri modello fattoriale: $m \times p + p = p(m + 1)$

Numero di variabili = numero di fattori

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

dimostrazione

- Poichè $\mathbf{\Sigma}$ è definita positiva, vale la scomposizione spettrale

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

  dove $\mathbf{\Lambda} = diag\{\lambda_k\}$, $\mathbf{P}' = [\mathbf{e}_1, \ldots \mathbf{e}_p]$
- ponendo $\mathbf{L} = \mathbf{P\Lambda}^{1/2}$, con $\mathbf{\Lambda}^{1/2} = diag\{\sqrt{\lambda_k}\}$
- si ottiene

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

Trade-off tra completezza e modello parsimonioso

Se $m < p$ si ha una descrizione *parsimoniosa* di $\boldsymbol{\Sigma}$.

Esempio:

p = 6 variabili, m = 2 fattori comuni

$\boldsymbol{\Sigma}$ ha 6 varianze e $(6 \times 5)/2 = 15$ covarianze $\Rightarrow 15 + 6 =$ 21 parametri

$\mathbf{L}\mathbf{L}^{j} + \boldsymbol{\Psi}$ ha $m \times p = 2 \times 6 = 12$ *loadindgs* e $p = 6$ specificità $\Rightarrow 12 + 6 =$ 18parametri

Non sempre è possibile trovare $m < p$ fattori comuni che rappresentino la struttura di covarianza $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^{j} + \boldsymbol{\Psi}$

quindi non sempre esiste una soluzione

Quando m>1

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

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

$$\mathbf{X} - \mu = \mathbf{LF} + \varepsilon = \mathbf{LTT}'\mathbf{F} + \varepsilon = \mathbf{L}^*\mathbf{F}^*\varepsilon$$

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

Soluzione indeterminata

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 soluzione unica 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 $\mathbf{L}^* = \mathbf{LT}$
- ma generano lo *stesso modello* per la matrice di covarianza $\mathbf{\Sigma}$!

$$\mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi} = \mathbf{L}\underbrace{\mathbf{TT}'}_{\mathbf{I}}\mathbf{L}' + \mathbf{\Psi} = (\mathbf{L}^*)(\mathbf{L}^*)' + \mathbf{\Psi}$$

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

Valori fattoriali e comunalità

I *factor loadings o valori fattoriali* **L** sono determinati attraverso la matrice ortogonale **T**

$\Rightarrow$ **L** e $\mathbf{L}^* = \mathbf{LT}$

L e $\mathbf{L}^*$ forniscono la stessa rappresentazione di $\boldsymbol{\Sigma}$

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

$$\mathbf{L}^*(\mathbf{L}^*)' = \mathbf{LTT}'\mathbf{L}' = \mathbf{LL}'$$

Come si procede →

1. si impongono dei *vincoli* per ottenere una soluzione unica per **L** e $\boldsymbol{\Psi}$
2. Si *ruota* la matrice dei *factor loadings* **L**, moltiplicando $\hat{\mathbf{L}}$ per una matrice ortogonale **T**: $\mathbf{L}^* = \mathbf{LT}$
3. si sceglie **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 di analisi fattoriale

Date n osservazioni p-dimensionali $\mathbf{x}_1, \ldots, \mathbf{x}_n$ l'analisi fattoriale cerca di rispondere alla domanda:

Il modello fattoriale ortogonale con un numero q <p può rappresentare adeguatamente i dati?

Da un punto di vista statistico, si tratterà di verificare se vale la relazione

$$Cov(\mathbf{X}) = \mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi}$$

- La matrice campionaria **S** è uno stimatore di **Σ**
- Il modello fattoriale non è appropriato se le covarianze in **S** sono piccole, o analogamente se le correlazioni in **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 **Σ** ha covarianze significativamente diverse da zero, il modello fattoriale è appropriato. Si tratterà prima di tutto di stimare i factor loadings e le specificità.

4 – PROCEDIMENTO ANALITICO in R

ESEMPIO

Label	Description		Coding
ID	Personal ID of the interviewed		Increasing integer number
AgeClass	Age of the person		Age (years)
AGE_CLASS	Age class of the person		1-6
SEX	Sex of the person		M or F
PROV	Province where the interviewed lives		Province code
LIKE_WINE	How much do you like drinking wine?		Integrer number from 1 to 7
FREQ_HOME	How often do you drink wine at home with meals?		Integrer number from 1 to 5
FREQ_BAR	How often do you drink wine in bars/pubs?		Integrer number from 1 to 5
FREQ_REST	How often do you drink wine at restaurants with meals?		Integrer number from 1 to 5
KNOW_PAS	Do you know the wine Passito?		Integrer number from 1 to 7
FREQ_PAS	How often do you drink Passito?		Integrer number from 1 to 5
FREQ_P_HOL	How often do you drink Passito on holidays and celebrations?		Integrer number from 1 to 5
FREQ_P_ALO	How often do you drink Passito when you are alone?		Integrer number from 1 to 5
FREQ_P_MEA	How often do you drink Passito at the end of meals?		Integrer number from 1 to 5
FREQ_P_OFF	How often do you drink Passito offered by someone?		Integrer number from 1 to 5
HOW_MUCH	How much wine do you drink in one year?		Integrer number from 1 to 4
LIKE_PAS	How much do you like drinking Passito?		Integrer number from 1 to 7
LIKE_AROMA	How much do you like aroma and smell of Passito?		Integrer number from 1 to 7
LIKE_SWEET	How much do you like the sweetness of Passito?		Integrer number from 1 to 7
LIKE_ALCOHOL	How much do you like the alcohol content of Passito?		Integrer number from 1 to 7
LIKE_TASTE	How much do you like the intensity of taste of Passito?		Integrer number from 1 to 7
PRICE	How much could you pay for one bottle of Passito? (0.5 litre)		Integrer number from 1 to 5

5 - ESERCIZI IN R

Problem 1 – Passito

- Eseguire una ANALISI FATTORIALE ESPLORATIVA sulle17 variabili risposta che rappresentano il questionario sulle abitudini, il comportamento e le preferenze dei consumatori di vino (dalla variabile LIKE_WINE alla variabile PRICE) per identificare *q<17* nuove variabili che “spiegano” I dati

Problem 2 – centro commerciale

- Eseguire una ANALISI FATTORIALE ESPLORATIVA sulle 5 variabili risposta per individuare $q<5$ nuove variabili che "spiegano" i dati

Problem 3 – abitudini alimentari

- Eseguire una ANALISI FATTORIALE ESPLORATIVA sulle12 variabili risposta osservate (da *Alcoholic.Beverages a Milk*) per individuare $q<12$ nuove variabili che "spiegano" i dati