Lezione 11

L'analisi fattoriale: il processo di estrazione dei fattori

Argomenti della lezione:

- **Metodi di estrazione dei fattori**
- **Metodi per stabilire il numero di fattori**

Metodi di Estrazione dei Fattori

Scopo fondamentale dell'AF: estrarre un numero sufficiente di fattori per "rendere conto" della matrice delle correlazioni tra le variabili osservate

Differenze tra metodi di estrazione

- **Metodi che spiegano il massimo di varianza delle variabili**
- **Metodi che massimizzano la "riproduzione" di R**
- **Metodi che richiedono una stima iniziale delle comunalità**
- **Metodi che richiedono una stima del numero di fattori da estrarre**

Analisi delle Componenti Principali (ACP)

Identifica una serie di combinazioni lineari ortogonali delle variabili originali X_i ($c_i=X_iV$, V = autovettori di R) che:

- **spiegano più varianza possibile**
- **riducono la complessità dei dati**

- **L'ACP analizza la varianza totale delle variabili**
- **Analizza la matrice R con valori 1 sulla diagonale principale: la varianza unica viene assorbita dai fattori comuni**
- **Le saturazioni fattoriali risultano gonfiate**

- **L'ACP estrae il massimo della varianza per ogni componente**
- **La prima componente spiega più varianza, la seconda spiega più varianza dopo la prima, ecc...**
- **R è perfettamente replicata se vengono estratte tante componenti quante sono le variabili**

Analisi dei Fattori Principali (AFP o PAF)

- **Massimizza lo stesso criterio della ACP; analizza R, ma con stime della comunalità inserite nella diagonale principale al posto di 1**
- **Analizza solo la varianza attribuibile ai fattori "comuni", per ottenere una soluzione non contaminata dalla varianza unica**

- **L' AFP estrae il massimo di varianza per ogni fattore, ma spiega meno varianza della ACP perché considera solo la varianza comune**
- **Passo preliminare: rimuovere dalla diagonale principale della matrice R la varianza unica (ovvero la componente $u^2=1-h^2$)**

Stima iniziale della comunalità

- **Coefficiente di correlazione multipla al quadrato (SMC)**
- **Coefficiente di correlazione più elevato**
- **Media delle correlazioni**

Minimi Quadrati (ULS e GLS)

- **Minimizza le differenze al quadrato tra le correlazioni osservate (R), e quelle riprodotte (R^) tramite i fattori estratti**
- **Minimizza le correlazioni residue (R- R^) cioè la parte di correlazione tra le variabili che non è spiegata dai fattori comuni**

Funzione dei minimi quadrati ordinari (ULS) minimizzata nel processo di estrazione dei fattori:

$$\Sigma_j\Sigma_k (r_{jk} - r_{jk}\hat{})^2$$

Viene ottimizzata la riproduzione dei coefficienti fuori della diagonale principale di R

Metodo dei minimi quadrati generalizzati o ponderati (GLS)

Massimizza la stessa funzione del metodo dei minimi quadrati ordinari (ULS)

Ponderazione delle variabili che penalizza quelle con varianza unica più elevata

- **Si inizia il processo stabilendo il numero di fattori**
- **Si stimano le saturazioni iniziali con l'ACP**

Le saturazioni vengono modificate iterativamente finché lo scarto tra R e R^ non è molto piccolo

Massima verosimiglianza (Maximum Likelihood, ML)

- **Stima i valori delle saturazioni nella popolazione**
- **Calcola le saturazioni che rendono massima la probabilità di osservare la matrice R**

- **Identifica la soluzione fattoriale che meglio riproduce R**
- **Stima le saturazioni della popolazione che hanno la massima verosimiglianza nel produrre la matrice delle correlazioni R**

- **Si considerano gli elementi fuori della diagonale principale**
- **Bisogna fornire il numero di fattori da estrarre**
- **Si stimano le saturazioni iniziali con l'ACP**

La stima delle saturazioni viene modificata iterativamente con una funzione complessa

Per utilizzare il criterio di massima verosimiglianza è necessario che:

$(n-k)^2 > (n + k)$

n = numero di variabili
k = numero di fattori

Es.: con n=5 possono essere estratti al massimo 2 fattori
Se k=3, $(n-k)^2 = 4$, e $(n + k) = 8$

Test di bontà dell'adattamento (goodness of fit)

→ Si ottiene dalle funzioni ML e GLS che vengono minimizzate, se le variabili seguono la distribuzione normale multivariata

Ipotesi nulla: H_0: R=R^
Segue la distribuzione del χ^2

Gradi di libertà:

$gdl = [(n-k)^2 - (n+k)]/2$

n = numero di variabili
k = numero di fattori

Se $(n-k)^2 < (n+k)$
i gradi di libertà sono negativi !

χ^2 non significativo ⇒
non vi sono più fattori da estrarre
Non si può rifiutare H_0:R=R^

χ^2 significativo ⇒
è necessario procedere all'estrazione di fattori ulteriori

Test fortemente dipendente dall'ampiezza del campione

Stabilire il numero di fattori

→ Decisione che ha conseguenze cruciali per la soluzione fattoriale

→ Salvaguardare la parsimonia della soluzione, e la sua adeguatezza (capacità di riprodurre R)

Metodi per stabilire il numero di fattori

→ Mineigen (Kaiser-Guttman rule)
→ Scree test degli autovalori
→ Test statistico (con GLS e ML)
→ Percentuale di varianza spiegata
→ Massima correlazione residua
→ Replicabilità della soluzione

Mineigen (Kaiser-Guttman rule)

→ **Estrae tutti quei fattori che hanno un autovalore maggiore di 1 quando viene analizzata la matrice R completa (diagonale principale)**

→ **I fattori devono spiegare almeno la stessa varianza spiegata dalle variabili osservate**

Il numero di autovalori maggiori di 1 è compreso tra 1/3 e 1/5 del numero delle variabili

Criterio inappropriato nel caso di soluzioni diverse dall'ACP

Scree test degli autovalori (Cattell e Vogelman)

→ **I primi fattori sono i più attendibili e i più validi, poiché spiegano una percentuale di varianza maggiore rispetto ai fattori rimanenti**

→ **Questi fattori hanno autovalori più grandi degli altri**

Gli autovalori rappresentano una progressione decrescente

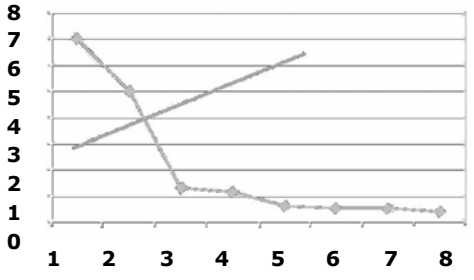
Grafico:

autovalore ⇒ in ordinata
numero del fattore ⇒ in ascissa

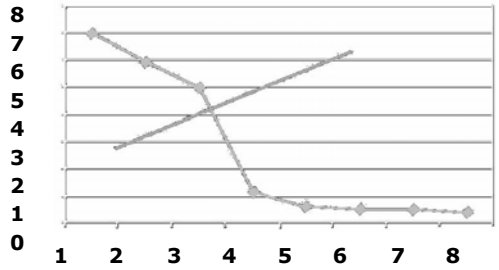
Si interrompe il processo di estrazione nel punto in cui la curva degli autovalori decrescenti cambia pendenza e diventa sostanzialmente piatta

Vanno presi quei fattori i cui autovalori sono al di sopra della linea piatta formata dagli autovalori dei fattori più piccoli

Soluzione con 2 fattori



Soluzione con 3 fattori



L'applicazione di questo test grafico è più attendibile quando:

- → il campione è grande
- → le comunalità sono elevati
- → i fattori saturano più variabili

Test statistico

→ Tende a sovrastimare il numero di fattori se il numero di soggetti è grande e le variabili sono molte

Percentuale di varianza spiegata

→ Specificare la proporzione di varianza totale che deve essere spiegata dall'ultimo fattore. Metodo troppo soggettivo

Massima correlazione residua

→ Per ogni elemento di R fuori della diagonale principale si può definire un residuo che è uguale a (r- r^), ovvero correlazione osservata meno correlazione riprodotta

Massima correlazione residua

→ Matrice dei residui: E = (R-R^). Se dopo aver effettuato l'estrazione di un certo numero di fattori tutti i residui sono minori di |.10|, non è necessario continuare il processo di estrazione: il nuovo fattore estratto avrebbe saturazioni molto basse

Replicabilità della soluzione

→ I fattori "validi" sono quelli che risultano più facilmente replicabili su campioni diversi da quelli nei quali sono stati individuati

→ I fattori "spuri" risultano poco generalizzabili e sono determinati sostanzialmente dall'errore campionario

CONCLUSIONE

- **Estrazione dei fattori**
- **Scelta del numero dei fattori**