

Analisi fattoriale

Bozze in revisione

A. Tarsitano

Dicembre 2011

Indice

1	Il modello di analisi fattoriale	2
1.1	Introduzione	2
1.2	Analisi fattoriale: metodo degli assi principali	5
1.3	Analisi fattoriale: metodo di massima verosimiglianza (Joreskog-Lawley)	6
1.4	Confronto tra AF e ACP	8

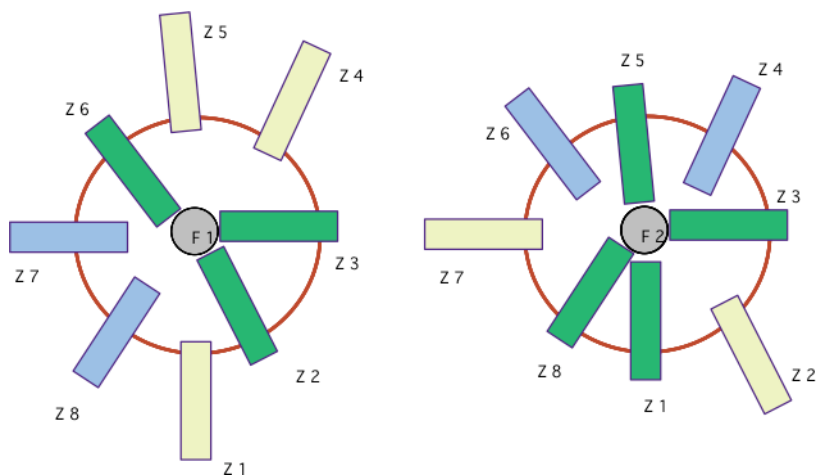
Capitolo 1

Il modello di analisi fattoriale

1.1 Introduzione

L'analisi fattoriale (AF) di una matrice di dati contribuisce a chiarire dei concetti difficili e ad arricchire la spiegazione di concetti indefiniti o mal definiti nonché a provocare la formulazione di concetti nuovi. Il suo uso nell'analisi dei dati -avviato da più di un secolo- è influenzato non solo dall'aspetto matematico che i problemi possono presentare, ma soprattutto dagli obiettivi che la ricerca si propone, dalla mole dei dati coinvolti e dall'adattabilità dei dati stessi ad uno specifico modello teorico di interpretazione.

L'analisi fattoriale poggia sull'idea che le variabili del data set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ siano correlate in modo tale da renderne possibile la ricostruzione a partire da un ristretto numero di nuove variabili o variabili latenti (diciamo $h < m$) che rappresenterebbero così la struttura portante dei dati in una forma sintetica e di più agevole lettura. In questo senso l'analisi fattoriale dà luogo, attraverso una procedura complessa ed articolata in più passi iterativi, a combinazioni lineari delle variabili di partenza dette fattori. Sotto l'aspetto del calcolo numerico, l'analisi fattoriale differisce poco dall'analisi delle componenti principali. Per molti versi il loro risultato scientifico è simile: i fattori, come le componenti principali, sono da interpretare come delle misurazioni di aspetti del campione, non direttamente osservabili, ma che stanno dietro alle variabili latenti che compaiono nei fattori stessi.



I due cerchi rappresentano i fattori comuni (due nell'esempio) e fra di essi esiste una gerarchia di importanza come per le componenti principali. Le variabili sono rappresentate con dei rettangoli (otto nell'esempio) che incidono in vario grado sui due fattori. La loro importanza è espressa dal grado di inclusione nel cerchio: piccola, media, grande. La configurazione dei pesi caratterizza il fattore (ipotizzando che il fattore possa in effetti essere caratterizzato). A differenza della ACP, le variabili dell'analisi fattoriale non si disperdono, in via di principio, totalmente nei fattori: se per m variabili l'estrazione di m componenti è solo una rotazione geometrica più o meno utile allo studio del data set, per l'analisi fattoriale è una contraddizione di termini perché qui le variabili mantengono sempre la cosiddetta specificità che non si può sovrapporre -logicamente- a nessun fattore comune o componente principale. Intendiamoci, anche i fattori della ACP lasciano fuori qualcosa dai fattori comuni, solo che non è interpretata come specificità, ma solo come quell'errore che si commette usando la scomposizione in valori singolari in cui solo una parte, quella più rilevante, forma l'approssimazione di rango prefissato della originale matrice dei dati.

Il modello di analisi fattoriale deriva da un approccio molto simile al modello di regressione lineare multipla: le m variabili coinvolte nell'analisi, una volta standardizzate con media nulla e varianza unitaria: $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m\}$ possono essere espresse, al netto di un termine di errore $\mathbf{u}$, come combinazioni lineari di h supervariabili (fattori comuni o variabili latenti) $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_h\}$ tali che

$$\mathbf{Z}_i = \sum_{j=1}^h q_{ji} \mathbf{f}_j + \mathbf{u}_i \quad \text{per } j = 1, 2, \dots, m \quad (1.1)$$

dove i q_{ij} sono i pesi fattoriali e $\mathbf{u}_i$ è il termine di errore che racchiude al suo interno anche l'unicità della variabile i -esima (complemento della specificità). Il numero h di fattori deve essere almeno pari a due. Da notare che questa soglia non è contemplata nella ACP che anzi, in alcune applicazioni quali le graduatorie unidimensionali, utilizzano la prima e sola componente principale. In termini matriciali il modello di analisi fattoriale si può scrivere nel modo seguente:

$$\begin{matrix} \mathbf{Z} & = & \mathbf{F} & \mathbf{Q}^t & + & \mathbf{u} \\ (n \times m) & & (n \times h) & (h \times m) & & (n \times m) \end{matrix} \quad (1.2)$$

La matrice $\mathbf{Q}$ è la matrice dei pesi fattoriali. Sia questa procedura che quella delle componenti principali hanno lo stesso obiettivo: la riduzione della dimensionalità dal lato degli indicatori. La analisi fattoriale però tenta di raggiungerlo postulando un modello di relazione lineare tra variabili originarie e fattori comuni costruito su di una serie di premesse, alcune inverosimili ed altre di difficile asseveramento.

Ipotizziamo che i fattori siano delle variabili casuali standardizzate tali che

$$E(\mathbf{F}) = 0 \quad (1.3)$$

$$E(\mathbf{F}\mathbf{F}^t) = \mathbf{\Phi} \quad (1.4)$$

dove $\mathbf{\Phi}$ è una matrice $(h \times h)$ di correlazione tra fattori (quindi, con degli uno sulla diagonale). Ipotizziamo, inoltre, che gli errori abbiano media nulla, ma matrice di varianze-covarianze diagonale.

$$E(\mathbf{u}) = 0, E(\mathbf{u}\mathbf{u}^t) = \mathbf{\Psi} = \text{diag}(\psi_1^2, \psi_2^2, \dots, \psi_m^2) \quad (1.5)$$

Quindi gli errori sono eteroschedastici, ma incorrelati. La costante ψ_i^2 è nota come “specificità” teorica della variabile i -esima e la si può interpretare come quella parte di variabilità della $\mathbf{x}_i$ che confluisce nei fattori comuni. Supporremo infine, come avviene per il modello di regressione multipla, che gli errori ed i fattori siano incorrelati

$$E(\mathbf{F}^t \mathbf{u}) = E(\mathbf{u}^t \mathbf{F}) = 0 \quad (1.6)$$

La matrice di dispersione delle variabili $\mathbf{z}$ che, data la standardizzazione, coincide con la loro matrice di correlazione $\mathbf{R}$, può a questo punto essere scomposta nella somma di due componenti

$$\begin{aligned} R = E(\mathbf{Z}^t \mathbf{Z}) &= E \left[(\mathbf{F} \mathbf{Q}^t + \mathbf{u})^t (\mathbf{F} \mathbf{Q}^t + \mathbf{u}) \right] \\ &= E \left[\mathbf{Q} \mathbf{F}^t \mathbf{F} \mathbf{Q} + \mathbf{Q} \mathbf{F}^t \mathbf{u} + \mathbf{u}^t \mathbf{F} \mathbf{Q}^t + \mathbf{u} \mathbf{u}^t \right] \\ &= E \left[\mathbf{Q} \mathbf{F}^t \mathbf{F} \mathbf{Q} \right] + \mathbf{Q} E \left[\mathbf{F}^t \mathbf{u} \right] + E \left[\mathbf{u}^t \mathbf{F} \right] \mathbf{Q}^t + E \left[\mathbf{u} \mathbf{u}^t \right] \\ &= \mathbf{Q} E \left[\mathbf{F}^t \mathbf{F} \right] \mathbf{Q}^t + \mathbf{0} + \mathbf{0} + \mathbf{\Psi} = \mathbf{Q} \mathbf{\Phi} \mathbf{Q}^t + \mathbf{\Psi} \end{aligned} \quad (1.7)$$

Se i fattori fossero tra di loro ortogonali si avrebbe

$$E(\mathbf{F} \mathbf{F}^t) = \mathbf{\Phi} = \mathbf{I} \quad (1.8)$$

che, a sua volta, comporterebbe una importante semplificazione:

$$\mathbf{R} = \mathbf{Q} \mathbf{Q}^t + \mathbf{\Psi} \quad (1.9)$$

Qui conosciamo solo il lato sinistro, ma sappiamo come determinare il primo addendo sul lato destro. E' la matrice $\mathbf{\Psi}$ che crea qualche fastidio che però si può eliminare. Esaminiamo le relazioni tra le variabili ed i fattori come si configura in base alle premesse effettuate:

$$\begin{aligned} E(\mathbf{F}^t \mathbf{Z}) &= E \left[\mathbf{F}^t (\mathbf{F} \mathbf{Q}^t + \mathbf{u}) \right] \\ &= E \left[\mathbf{F}^t \mathbf{F} \mathbf{Q}^t + \mathbf{Q} \mathbf{F}^t \mathbf{u} \right] \\ &= E \left[\mathbf{F}^t \mathbf{F} \right] \mathbf{Q}^t + \mathbf{Q} E \left[\mathbf{F}^t \mathbf{u} \right] \\ &= \mathbf{\Phi} \mathbf{Q}^t \end{aligned} \quad (1.10)$$

Se i fattori fossero incorrelati cioè $\mathbf{\Phi} = \mathbf{I}$, allora si avrebbe $E(\mathbf{F}^t \mathbf{Z}) = \mathbf{Q}^t$. In questo caso i coefficienti q_{ij} possono essere interpretati come correlazioni tra le variabili osservate e le variabili latenti. Queste relazioni sono il punto di partenza dell'analisi fattoriale.

La formulazione del modello [1.6] è simile a quella di un modello di regressione lineare multipla, ma con una sostanziale eccezione. Nel modello di analisi fattoriale la matrice $\mathbf{Q}$ è incognita come lo sono i fattori specifici $\mathbf{u}$ ed è facile accorgersi come questa procedura dovrebbe stimare un numero di parametri superiore al numero di osservazioni, con evidenti e insuperabili problemi di indeterminatezza. La possibilità di superare questo problema di identificazione è stato affrontato con metodi diversi che presuppongono la distribuzione gaussiana delle variabili $\mathbf{Z}$ ovvero si basano su ingegnose congetture sulle relazioni tra fattori e variabili.

Sia $\mathbf{G}$ una matrice non singolare di ordine $(h \times h)$. Se stabiliamo le seguenti definizioni

$$\begin{aligned}\mathbf{f}_0 &= \mathbf{G}\mathbf{f} \\ \mathbf{Q}_0 &= \mathbf{Q}\mathbf{G}^{-1} \\ \Phi_0 &= \mathbf{G}\Phi\mathbf{G}^{-1}\end{aligned}\tag{1.11}$$

Il modello [1.6] e la relazione finale di [1.7] possono essere riespressi come

$$\begin{aligned}\mathbf{Z} &= \mathbf{Q}_0\mathbf{f}_0 + \mathbf{u} \\ \mathbf{R} &= \mathbf{Q}_0\Phi_0\mathbf{Q}_0^t\end{aligned}\tag{1.12}$$

Poiché noi vediamo solo il lato sinistro di entrambe le relazioni, il lato destro potrebbe essere stato determinato tanto da $(\mathbf{f}_0, \mathbf{Q}_0)$ che da $(\mathbf{f}, \mathbf{Q})$. Ogni matrice $\mathbf{G}$ dotata di inversa determina una soluzione valida ai fini dell'analisi fattoriale. Anche se volessimo restringere il campo delle matrici $\mathbf{G}$ alle sole matrici ortonormali (matrici di rotazione) sarebbe sempre di fronte ad una soluzione indeterminata a causa delle specificità della matrice Φ . La procedura dell'AF opera tipicamente in due fasi: un primo stadio in cui si ottiene una soluzione ammissibile ed un altro in cui la soluzione ammissibile viene trasformata con una opportuna matrice di rotazione.

Per ottenere una soluzione ammissibile dobbiamo comunque aggiungere qualche vincolo alla scelta della matrice $\mathbf{Q}$ per annullare il problema della sua non identificabilità. La relazione [1.7] mostra che la varianza di ognuna delle variabili $\mathbf{Z}$ può essere espressa come somma di due componenti: l' i -esimo elemento della matrice $\mathbf{Q}\Phi\mathbf{Q}^t$ che rappresenta la comunalità della variabile $\mathbf{Z}_i$ e l' i -esimo elemento della matrice Ψ che rappresenta la specificità (od anche la varianza residua) della $\mathbf{Z}_i$. In formule si ha

$$\sigma_i^2 = \sum_{j=1}^h q_{ij}^2 + \sum_{j=1}^h \sum_{k \neq j} q_{ik} q_{jk} \phi_{jk} + \psi_i \quad \text{per } i = 1, 2, \dots, m \tag{1.13}$$

Qualora i fattori fossero ortogonali si otterrebbe la semplificazione

$$\sigma_i^2 = \sum_{j=1}^h q_{ij}^2 + \psi_i \quad \text{per } i = 1, 2, \dots, m \tag{1.14}$$

La trattazione completa delle tecniche di AF è reperibile in molti testi. Qui ci limiteremo a discuterne due fra le più diffuse: metodo degli assi principali e FA di massima verosimiglianza entrambe votate alla ricerca di fattori incorrelati.

1.2 Analisi fattoriale: metodo degli assi principali

Questo metodo comporta l'estrazione delle componenti principali dalla matrice $(\mathbf{R} - \Psi)$. I fattori così ottenuti sono ortonormali e l'equazione [1.7] può essere scritta come

$$(\mathbf{R} - \Psi) = \mathbf{Q}\mathbf{Q}^t \tag{1.15}$$

Nel presupposto che il modello sia corretto ovvero che vi siano in effetti h fattori incorrelati la matrice $(\mathbf{R} - \mathbf{\Psi})$ sarà simmetrica, di rango h , definita positiva ed avrà perciò h autovalori maggiori di zero che, noi -per convenienza, ma senza poter esserne certi- riteniamo distinti. Se gli autovalori $\lambda_1 > \lambda_2 > \dots > \lambda_h$ sono utilizzati per formare la matrice diagonale $\mathbf{\Theta}_h$ la matrice dei pesi $\mathbf{Q}$ verificherà la relazione

$$\mathbf{Q}^t \mathbf{Q} = \mathbf{\Theta}_h \quad (1.16)$$

Tale equazione specifica $\mathbf{Q}$ e, quindi, i fattori purché, come si è ipotizzato, gli elementi sulla diagonale di $\mathbf{\Theta}_h$ siano in effetti distinti.

L'ammontare totale di variabilità assorbito dal fattore j è dato da

$$\theta_j = \sum_{i=1}^m q_{ij}^2 \quad \text{per } j = 1, 2, \dots, h \quad \text{con } \theta_1 > \theta_2 > \dots > \theta_h \quad (1.17)$$

Per applicare il metodo degli assi principali devono essere risolte due questioni:

1. Come stimare $\mathbf{\Psi}$
2. Quanti fattori dovranno essere considerati.

Per quanto attiene al primo problema si potrebbe partire dalla considerazione che il quadrato del coefficiente di correlazione multipla tra la variabile $\mathbf{z}_i$ e le restanti variabili $\mathbf{z}$ costituisce un limite inferiore per la comunaltà della i -esima variabile cioè per gli elementi sulla diagonale di $(\mathbf{R} - \mathbf{\Psi})$. In genere si considera questo valore un tentativo iniziale

$$\psi_i(0) = 1 - r_i^2 \quad \text{per } i = 1, 2, \dots, m \quad (1.18)$$

che è necessario per la scomposizione in valori singolari di $\mathbf{R} - \mathbf{\Psi}(0)$. Da questa si estraggono le prime h componenti formando la prima approssimazione della matrice dei pesi $\mathbf{Q}_0$. La somma del quadrato delle sue righe fornirà una nuova stima delle specificità

$$\psi_i(1) = 1 - \sum_{j=1}^h q_{ij}^2(0) \quad \text{per } i = 1, 2, \dots, m \quad (1.19)$$

ed il ciclo ricomincia. Le iterazioni convergono comunque molto rapidamente.

Nella relazione [1.7] si è ipotizzato che la matrice a sinistra sia positiva definita, ma non è detto che tale rimanga in tutte le iterazioni. Non è quindi consigliabile considerare un numero di fattori prossimo al numero di variabili perché si rischia un eccesso di fattorizzazione inflazionando la varianza comune e, d'altra parte, un numero piccolo di fattori è proprio lo scopo della AF.

1.3 Analisi fattoriale: metodo di massima verosimiglianza (Joreskog-Lawley)

Ipotizziamo che i fattori $\mathbf{f}$ e gli errori $\mathbf{u}$ abbiano una distribuzione normale m -dimensionale con matrice di varianze-covarianze $\mathbf{\Omega}$ e che il data set $\mathbf{X}$ sia il frutto di un campione casuale

di ampiezza n di osservazioni da essa estratti. Se $n > m$ la matrice di varianze-covarianze è

$$\hat{\Omega} = \frac{1}{n-1} \sum_{k=1}^n (\mathbf{u} - \mu)(\mathbf{u} - \mu)^t \quad (1.20)$$

Il logaritmo della funzione di verosimiglianza per la stima di Ω corrisponde a

$$\log(L) = -\frac{(n-1)}{2} \left[\log(|\Omega|) + Tr(\hat{\Omega}\Omega^{-1}) \right] + \mathbf{M} \quad (1.21)$$

dove $\mathbf{M}$ è una quantità che non dipende dal parametro incognito Ω . Come si è accennato in precedenza, anche la soluzione di massima verosimiglianza mira all'ottenimento di una soluzione ortogonale $\mathbf{W}$, pertanto, il problema si riduce all'ottenimento di una stima $\hat{\mathbf{Q}}$ e $\hat{\Psi}$ che verifichino la relazione matriciale

$$\Omega = \mathbf{Q}\mathbf{Q}^t + \Psi \quad (1.22)$$

e che corrispondano ad un massimo per $\log(L)$. Per ottenere le due matrici si impiega un algoritmo a due stadi orientati alla minimizzazione di

$$L_1(\mathbf{Q}, \Psi) = \log(|\Omega|) + Tr(\hat{\Omega}\Omega^{-1}) - \log(|\hat{\Omega}|) - m \quad (1.23)$$

che equivale alla massimizzazione di per $\log(L)$. Nel primo stadio la L_1 è minimizzata per $\mathbf{Q}$ ipotizzando Ψ fissata. Sul risultato di questa fase si basa poi una nuova stima di Ψ per $\mathbf{Q}$ fissata al livello raggiunto nel precedente stadio. Lo schema si ripete fino alla convergenza. L'algoritmo è il seguente:

$$\begin{aligned} \mathbf{Q}_{i+1} &= \Psi_i^{0.5} \mathbf{W}_i [\mathbf{H}_i - \mathbf{I}] \\ \mathbf{W}_i &= \Psi_i^{-0.5} \mathbf{Q}_i [\mathbf{H}_i - \mathbf{I}]^{-0.5} + \mathbf{I} \\ \mathbf{H}_i &= \mathbf{I} + \mathbf{Q}_i \Psi_i^{-1} \mathbf{Q}_i^t \\ \Psi_{i+1} &= \text{diag} [\mathbf{W}_i - \mathbf{Q}_{i+1} \mathbf{Q}_{i+1}^t] \end{aligned} \quad (1.24)$$

soggette al vincolo che

$$\mathbf{Q}_i \Psi_i^{-1} \mathbf{Q}_i^t = \text{diagonale} \quad (1.25)$$

che assicura l'unicità della soluzione di massima verosimiglianza. A questo fine ci si deve anche limitare a considerare i soli autovalori maggiori dell'unità. Per avviare l'algoritmo si può usare la seguente stima iniziale per Ψ

$$\Psi_0 = \text{diag} [\psi_1(0), \psi_2(0), \dots, \psi_h(0)] \quad \text{con} \quad \psi_i = \left(1 - \frac{h}{2m}\right) \left(\frac{1}{\omega^{ii}}\right) \quad (1.26)$$

dove ω^{ii} è l'elemento sulla diagonale della matrice Ω^{-1} .

1.4 Confronto tra AF e ACP

Prima di passare a discutere alle applicazioni è necessario un confronto di massima con la procedura rivale: l'analisi fattoriale. Cominceremo la discussione su questo punto cruciale delle analisi territoriali richiamando una affermazione contenuta nel manuale del "Systat" (uno dei packages statistici più diffusi):

Nonostante qualcuno tra gli utilizzatori dell' analisi fattoriale sia molto suscettibile su questo punto e voglia riservare il termine "analisi fattorial" per il solo modello dei fattori comuni, i risultati ottenuti utilizzando su dati reali l'analisi fattoriale e l'analisi delle componenti principali raramente presentano differenze significative

La differenza tra analisi fattoriale e analisi delle componenti principali non è tanto nei risultati o nei metodi di calcolo, quanto nello schema teorico da cui muovono. L'idea di fondo del modello di analisi fattoriale è che gli m indicatori siano riconducibili a poche variabili latenti che ne determinano le interazioni. L'analisi delle componenti principali si basa su ipotesi molto meno specifiche e quindi non porta a verifiche di particolari ipotesi sulla struttura delle relazioni tra gli indicatori. Tale procedura costituisce solo un modo diverso, più semplice, di rappresentare con una trasformazione affine le stesse informazioni.

Il confronto tra analisi fattoriale e analisi delle componenti è esaurientemente condotto in Jolliffe (1986) dove sono messe in risalto analogie e difformità tra le due tecniche e che qui riprendiamo sommariamente. Innanzitutto, l'analisi fattoriale produce in genere meno fattori comuni di quanti non ne suggerisca l'analisi delle componenti e questo perché laddove in questa sono possibili delle componenti separate (connesse ad una sola variabile), nessun fattore comune della AF può invece essere legato a meno di due variabili (se così non fosse il corrispondente fattore comune confluirebbe, confondendosi, nel fattore specifico/errore $\mathbf{u}$). Inoltre, come si è visto per il metodo di verosimiglianza, nessun autovalore può essere inferiore all'unità per poterne considerare il fattore associato. Nella ACP questo vincolo non sussiste e si possono considerare tutte le componenti con autovalore maggiore di zero. Ai fini della riduzione della dimensionalità sembrerebbe quindi più efficiente l'analisi fattoriale, anche se "ingabbiare" in un fattore unico due o più indicatori altrimenti distinti potrebbe portare ad interpretazioni forzose.

Un altro punto rilevante è che entrambe le tecniche hanno come oggetto la matrice di varianze-covarianze $\mathbf{\Omega}$ (di indicatori variamente trasformati e di solito standardizzati), ma, mentre l'analisi delle componenti principali si concentra soprattutto sugli elementi della diagonale in quanto cerca di massimizzare la varianza delle componenti, l'AF cerca di massimizzare la rappresentazione degli elementi fuori della diagonale. Infatti, la $\mathbf{\Omega}$ è qui considerata come la somma di altre due matrici e poiché una di queste è diagonale il termine dei fattori comuni è più influenzato dalle covarianze che non dalle varianze. Da notare poi che i "punteggi fattoriali" usati in entrambe le tecniche sono ottenibili in maniera esatta dall'analisi delle componenti principali dato che queste sono delle funzioni lineari deterministiche delle variabili originarie. Lo stesso non è possibile per l'analisi fattoriale in quanto, nella relazione che lega le variabili ai fattori, compare il vettore incognito degli errori/specificità ed i punteggi possono solo essere stimati.

Infine, c'è la diversa reattività delle due tecniche alla alterazione del numero di componenti utilizzate. Se h passa da h_1 a h_2 con $h_2 > h_1$ si introducono $(h_2 - h_1)$ nuove

componenti, ma le prime h_1 non scompaiono e non si modificano. Nell'analisi fattoriale, l'aumento del numero di fattori, porta alla completa ridefinizione di tutti i fattori e fra i nuovi h_2 potrebbero non comparire uno o più dei fattori ottenuti per $h = h_1$. In conclusione si può dire che la scelta tra le due procedure dovrà essere guidata dalle finalità dell'analisi: se si è interessati ad una esplorazione dei dati che prescindano da particolari modelli allora la "semplicità" teorica dell'analisi delle componenti principali è preferibile. Se invece i dati si prestano alla formulazione in termini di variabili endogene e fattori esogeni diviene allora appropriata l'analisi fattoriale. Le due tecniche non sono in competizione, anzi possono essere utilmente impiegate sullo stesso insieme di dati.