Germano Rossi

Appunti di analisi fattoriale

Work in progress

Versione: 0.4.1.42

2012-2018

Quest'opera è stata rilasciata con licenza Creative Commons Attribuzione 4.0 Internazionale. Per leggere una copia della licenza visita il sito web http://creativecommons.org/licenses/by/4.0/it o spedisci una lettera a Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Versione: 0.4.1.42 (2 ottobre 2018)

Germano Rossi (germano.rossi@unimib.it)
Università degli Studi di Milano-Bicocca, Dipartimento di Psicologia

Questo simbolo indica che c'è la possibilità di scaricare un file dati o qualche altro tipo di file. Generalmente i file dati sono disponibili in formato CSV (Comma Separated Value) che è compatibile sia con la maggior parte dei tabelloni elettronici (Excel, OpenCalc, LibreCalc...) sia con i programmi statistici (IBM SPSS, R, Jasp, Jamovi...) oppure in formato SPSS (.sav). Facendo click sul simbolo, potrai vedere il nome del file che verrà scaricato.

Le parti in grigio fra parentesi quadre indicano argomenti da aggiungere o da approfondire.

Indice

1 Approccio intuitivo **1**
1.1 Introduzione . . . 2
1.2 Approccio intuitivo . . . 2
1.3 Aspetti teorici . . . 6
1.4 Sintesi . . . 9

2 Eseguire un AFE: introduzione **11**
2.1 Introduzione . . . 12
2.2 Cosa c'è prima di un AFE . . . 13
2.3 Scelta del campione . . . 14
2.3.1 Dai dati alla matrice di correlazione . . . 15
2.3.2 Analisi della matrice di correlazione . . . 16
2.4 Determinazione dei fattori . . . 18
2.4.1 Numero dei fattori da estrarre . . . 18
2.4.2 Metodi teorici . . . 18
2.4.3 Metodi dipendenti da un'analisi preliminare . . . 19
2.4.4 Metodi che si basano su analisi più complesse . . . 20
2.5 Metodi di estrazione . . . 21
2.6 Rotazione dei fattori . . . 24
2.6.1 Metodi di rotazione ortogonali . . . 26
2.6.2 Metodi di rotazione obliqui . . . 26
2.7 L'interpretazione dei fattori . . . 27
2.7.1 Passaggi per l'interpretazione . . . 27
2.7.2 Esempio di interpretazione . . . 28
2.7.3 Punteggi fattoriali . . . 29

3 Appendice **31**
3.1 Test di Sfericità di Bartlett . . . 31
3.2 Test di Haitovsky . . . 31
3.3 Test di adeguatezza campionaria di Kaiser-Meyer-Olkin . . . 31
Riferimenti bibliografici . . . 31

Capitolo 1

L'analisi fattoriale da un punto di vista intuitivo

Germano Rossi e Paola Venuti[1]

[1] A partire dai primi due capitoli di un precedente libro: Lis, Rossi, e Venuti (1986)

1.1 Introduzione

L'analisi fattoriale può genericamente essere pensata come "la ricerca di variabili latenti a partire da alcune variabili osservate". Una "variabile osservata" è una variabile che è stata effettivamente misurata, mentre una "variabile latente" (o nascosta o sottostante) è un tipo di variabile che non è stata misurata, che forse non è neppure misurabile direttamente e che perciò viene ipotizzata e "analizzata" attraverso i suoi effetti. I legami, le relazioni, le influenze che una variabile latente ha su altre variabili misurabili diventano un modo per risalire a questa variabile nascosta (la consideriamo "sottostante" sia perché non è misurabile sia perché potremmo non averla misurata).

Storicamente, nell'ambito dell'analisi fattoriale, le variabili latenti vengono chiamate "fattori". Nei suoi primi modelli matematici, l'analisi fattoriale era una tecnica di analisi *esplorativa* (o AFE, Analisi Fattoriale Esplorativa) appunto perché permetteva di "esplorare" le relazioni nascoste fra un gran numero di variabili. In tempi più recenti, la tecnica delle equazioni strutturali ha permesso di sviluppare una Analisi Fattoriale *Confermativa* (o AFC), che cioè permette di verificare se effettivamente i fattori ipotizzati servono a spiegare le variabili misurate.

Spesso, all'Analisi fattoriale esplorativa viene associata un'altra analisi simile (l'Analisi in Componenti Principali o ACP) perchè AFE e ACP condividono una tecnica matematica che estrae, da una matrice di correlazione, delle "componenti". Tuttavia, come vedremo, l'ACP non è propriamente un'analisi fattoriale.

Sigla	Analisi (in italiano)	corrispettivo inglese
ACP	Analisi in Componenti Principali	PCA - Principal Component Analysis
AFE	Analisi Fattoriale Esplorativa	EFA - Explorative Factor Analysis
AFC	Analisi Fattoriale Confermativa	CFA - Confirmatory Factor Analysis

Tabella 1.1 – Corrispondenza fra terminologia italiana e inglese

Contrariamente alla maggior parte delle analisi dei dati (quelle comunemente chiamate "inferenziali") l'AFE (analisi fattoriale *esplorativa*) non usa la verifica d'ipotesi (cioé non ha un'ipotesi nulla e una alternativa) anche se (al suo interno) può utilizzare dei test inferenziali per "assicurarsi" che certe decisioni siano meno aleatorie. Infatti, nel condurre un'AFE, si devono prendere svariate decisioni arbitrarie e tali decisioni sono a volte guidate dalla teoria, dalla prassi o da qualche test inferenziale.

Nel caso dell'AFC (fattoriale confermativa) invece si può avere un'ipotesi nulla e una alternativa. In termini semplicistici, l'ipotesi nulla ritiene che il modello che il ricercatore ha in testa sia vero e questo porterà ad una stima di "discrepanza" fra la realtà e i risultati dell'analisi. Tanto più il modello stimato (la teoria del ricercatore) si avvicina alla realtà (i dati misurati) tanto più il modello sarà esplicativo della realtà. Ma questa ipotesi la vedremo nella parte dedicata all'AFC.

1.2 Approccio intuitivo

L'analisi fattoriale (soprattutto quella esplorativa) è un metodo di elaborazione dei dati che *appare* molto complesso, in quanto si basa su una serie di tecniche matematiche piuttosto difficili da studiare (per chi non ha affinità con i numeri e le formule e le matrici) ma spiegabili in termini semplici e quindi facilmente comprensibili.

Il procedimento logico che sta alla base di questa tecnica e gli scopi per cui viene generalmente usata sono meno complessi da capire ed è quello che cercheremo di illustrare qui, partendo da un ipotetico esempio di ricerca basato su dati immaginari.

Supponiamo di voler fare una ricerca per vedere, al termine di una terza media inferiore, che tipo di abilità hanno acquisito gli allievi individualmente, per poterli così indirizzare verso la scelta scolastica del loro futuro.

Si decide di applicare ad ognuno dei 20 ragazzi di una classi terza (un piccolo numero di casi a puro titolo esemplificativo), una serie di prove (variabili) relative a varie abilità ed acquisizioni insegnate ed apprese nel corso degli anni scolastici.

Le prove sono le seguenti:

I1 - Riconoscimento di figure
I2 - Risoluzione di problemi aritmetici
I3 - Contare e ordinare sequenze numeriche
I4 - Comprensione del brano letto
I5 - Lettura di un testo
I6 - Descrizione orale di avvenimenti
I7 - Risoluzione di problemi geometrici
I8 - Invenzione e racconto di una storia
I9 - Risoluzione di rebus

Queste prove vengono somministrate in ordine casuale e ogni soggetto ottiene per ciascuna di esse un punteggio compreso fra 1 e 10 (in realtà nel nostro esempio oscillano fra 3 e 8). Questi risultati sono mostrati nella Tabella 1.2 (i dati sono disponibili qui) .

Già ad una prima osservazione, questi risultati mostrano che, generalmente:

- chi ha ottenuto un punteggio alto (osservazioni 6 e 16) nella prova 2 ha ottenuto risultati simili anche nelle prove 3 e 7;
- chi ha ottenuto un punteggio medio-alto (osservazioni 3 e 8) nella prova 4 ha ottenuto punteggi medio-alti anche nelle prove 5 e 6;
- chi ha ottenuto bassi punteggi (osservazioni 10 e 14) nella prova 9 li ha avuti anche nella 4.

Non si riesce però a notare un andamento costante in tutte le prove.

Si può a questo punto tentare un'ipotesi: tra alcune delle variabili esiste un particolare rapporto per cui i rispondenti che ottengono un determinato punteggio in una di queste prove ottengono punteggi analoghi nelle altre prove dello stesso gruppo. Allo stesso tempo fra altre variabili esiste un rapporto analogo al precedente anche se non sembra che vi sia una relazione precisa fra i due gruppi di variabili.

Tabella 1.2 – Risultati ipotetici ottenuti da 20 ragazzi di terza media in 9 prove

Cod	I1	I2	I3	I4	I5	I6	I7	I8	I9
1	7	4	3	7	4	6	4	7	6
2	5	3	3	4	6	4	3	4	7
3	4	4	4	5	5	5	3	4	7
4	7	6	6	6	3	7	7	6	3
5	3	3	4	7	3	7	4	7	6
6	4	7	7	3	7	3	6	4	4
7	6	7	7	7	3	6	6	7	3
8	5	3	3	5	6	5	3	5	7
9	3	5	5	6	4	7	5	7	5
10	7	7	6	3	6	4	7	3	4
11	3	6	7	6	3	7	6	7	3
12	3	4	4	7	3	6	6	7	5
13	6	5	6	7	4	7	5	6	4
14	7	4	4	4	7	4	3	4	6
15	7	6	5	6	4	6	6	7	3
16	5	7	6	3	6	3	7	3	7
17	4	7	7	5	4	4	6	5	5
18	3	6	7	4	7	3	7	3	6
19	5	5	5	4	7	4	5	4	7
20	6	7	6	5	8	6	6	3	8

Questa nostra ipotesi sull'andamento dei dati rilevati, però, non è facilmente dimostrabile. Si può perciò pensare di ricorrere ad una metodica statistica che permetta di rilevare in che misura le variabili siano tra loro in relazione. In statistica ne esiste una che effettivamente risponde alle nostre esigenze: il coefficiente di correlazione prodotto-momento di Bravais-Pearson (cfr. Cristante, Lis, & Sambin, 1982), generalmente indicato semplicemente come "correlazione". Ricordiamo che questo indice statistico misura il grado di relazione esistente fra due variabili: se le variabili mutano concordemente (nel senso che soggetti che hanno ottenuto punteggi alti nella prima prova li ottengono pure alti nella seconda, e soggetti che invece ottengono punteggi bassi nella prima lo ottengono anche nella seconda) si

avrà una correlazione elevata, se mutano indipendentemente l'una dall'altra non si avrà alcuna correlazione. Abbiamo utilizzato questo coefficiente di "accordo" considerando tutte le coppie che è stato possibile ottenere tra le variabili. La Tabella 1.3 riporta i valori di questi coefficienti in una tabella che è chiamata "matrice delle correlazioni". All'incrocio fra una riga e una colonna si trova l'indice di correlazione delle due variabili indicate rispettivamente all'inizio della riga ed in cima alla colonna.

Analizziamo la Tabella 1.3 considerando come elevate, a solo titolo esemplificativo, le correlazioni superiori o uguali a .80 (evidenziate in grassetto). La tabella presenta solo metà dell'intera matrice delle correlazioni perché la correlazione della variabile 2 con la variabile 3 (.89) è uguale alla correlazione della variabile 3 con la variabile 2. L'intera matrice, quindi, è simmetrica rispetto alla diagonale (definita principale), cioè le due metà risultano speculari l'una all'altra. Una seconda caratteristica di questa tabella è la presenza di un trattino ad indicare la correlazione di una variabile con se stessa che generalmente non viene calcolata oppure viene indicata con il valore 1 (ovvero una correlazione perfettamente positiva).

Tabella 1.3 – Matrice delle correlazioni fra prove finali

Matrice delle correlazioni fra prove finali									
Variable	I1	I2	I3	I4	I5	I6	I7	I8	I9
I1	-								
I2	0.16	-							
I3	-0.12	**0.89**	-						
I4	-0.05	-0.30	-0.18	-					
I5	0.14	0.13	0.00	**-0.80**	-				
I6	0.02	-0.22	-0.12	**0.88**	-0.72	-			
I7	0.00	**0.84**	**0.80**	-0.13	-0.09	-0.07	-		
I8	-0.10	-0.26	-0.15	**0.89**	**-0.88**	**0.80**	-0.08	-	
I9	-0.15	-0.41	-0.50	-0.34	0.61	-0.38	-0.48	-0.55	-

Osservando questa tabella ci rendiamo conto che la nostra ipotesi iniziale era esatta. Esistono infatti delle relazioni molto alte tra alcune variabili (ad esempio la 2 con la 3 e con la 7; la 3 con la 7, tutte con valori superiori a .80) e altri valori decisamente bassi (ad esempio la 1 con tutte le variabili).

Si può quindi dire che queste 3 variabili (2, 3, 7) sono strettamente collegate fra loro e quindi risulta che misurino abilità simili o addirittura aspetti diversi di una sola abilità. Se ciò è facile da individuare analizzando queste tre variabili oppure con 4, 5 e 6, perché sono le une vicine alle altre, per altre variabili che non presentano questa vicinanza sulla matrice (ad esempio la 4 e la 12) questo rapporto non è immediatamente visibile. Già evidenziando in grassetto i valori maggiori di .80 abbiamo ottenuto due possibili blocchi di item molto correlati al loro interno.

Proviamo a riorganizzare la matrice delle correlazioni in modo che la maggior parte delle variabili molto correlate fra loro si trovino le une vicino alle altre anche fisicamente nella matrice stessa.

Cerchiamo la correlazione più alta esistente nella matrice (scegliamo .89 tra la variabile 2 e la variabile 3) e riorganizziamo la variabile 2 in ordine decrescente di valori ignorando il segno (Tabella 1.4).

A questo punto riproduciamo tutta la matrice di correlazione secondo l'ordine delle variabili indicato dai valori decrescenti della variabile 2 (Tabella 1.5).

Possiamo vedere che le variabili 2, 3 e 7 sono tra loro altamente correlate, mentre le loro correlazioni con tutte le altre variabili sono minori. Analogamente per le variabili 4, 8, 6 e 5. Ci sono anche correlazioni un po' più piccole di .80 (indicate in corsivo) che probabilmente costituiscono un blocco comune con le variabili vicine. Al contrario le variabili 9 e 1 sono invece

Tabella 1.4 – Correlazioni della variabile 2 con le altre variabili in ordine decrescente e ignorando il segno

	I3	I7	I9	I4	I8	I6	I1	I5
I2	0.89	0.84	0.41	0.30	0.26	0.22	0.16	0.13

Tabella 1.5 – Matrice delle correlazioni ordinata secondo i valori decrescenti della variabile 2

Variable	I2	I3	I7	I9	I4	I8	I6	I1	I5
I2	1.00								
I3	**0.89**	1.00							
I7	**0.84**	**0.80**	1.00						
I9	-0.41	-0.50	-0.48	1.00					
I4	-0.30	-0.18	-0.13	-0.34	1.00				
I8	-0.26	-0.15	-0.08	-0.55	**0.89**	1.00			
I6	-0.22	-0.12	-0.07	-0.38	**0.88**	**0.80**	1.00		
I1	0.16	-0.12	0.00	-0.15	-0.05	-0.10	0.02	1.00	
I5	0.13	0.00	-0.09	*0.61*	**-0.80**	**-0.88**	*-0.72*	0.14	1.00

particolari. La variabile 9 sembra correlare più o meno con tutte le variabili e la variabile 1 con nessuna.

Variable	I2	I3	I7	I9	I4	I8	I5	I6	I1
I2	1.00								
I3	**0.89**	1.00							
I7	**0.84**	**0.80**	1.00						
I9	*-0.41*	*-0.50*	*-0.48*	1.00					
I4	-0.30	-0.18	-0.13	-0.34	1.00				
I8	-0.26	-0.15	-0.08	*-0.55*	**0.89**	1.00			
I5	0.13	0.00	-0.09	*0.61*	**-0.80**	**-0.88**	1.00		
I6	-0.22	-0.12	-0.07	-0.38	**0.88**	**0.80**	**-0.72**	1.00	
I1	0.16	-0.12	0.00	-0.15	-0.05	-0.10	0.14	0.02	1.00

Tabella 1.6 – Matrice delle correlazioni con i raggruppamenti delle variabili evidenziati

Ristrutturiamo allora la tabella delle correlazioni per evidenziare la vicinanza di queste variabili. Come risulta dalla Tabella 1.6 si possono individuare, quindi, 2 blocchi di variabili, strettamente unite al loro interno e separate dalle altre.

Per spiegarci il motivo di queste associazioni proviamo ad analizzare il significato delle variabili che li compongono.

Il primo raggruppamento è formato dalle variabili:

I2 - Risoluzione di problemi aritmetici
I3 - Contare e ordinare sequenze numeriche
I7 - Risoluzione di problemi geometrici

Chiaramente, queste variabili indicano tutte un'abilità legata alla conoscenza dei numeri, dei rapporti matematici e geometrici.

Il secondo raggruppamento è formato dalle variabili:

I4 - Comprensione del brano letto
I8 - Invenzione e racconto di una storia
I6 - Descrizione orale di avvenimenti
I5 - Lettura di un testo

Queste variabili indicano tutte delle abilità di tipo linguistico e verbale legate ai processi di lettura, scrittura e comprensione di materiale scritto o letto. Le ultime due variabili indicano delle capacità logiche legate a conoscenze linguistiche o ortografiche.

Le due variabili restanti sono:

I1 - Riconoscimento di figure
I9 - Risoluzione di rebus

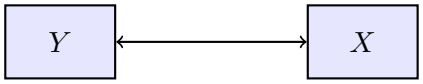
La variabile I1 non correla praticamente con nulla, per cui sembrebbe che il riconoscimento di figure non abbia a che fare né con le abilità numeriche né con quelle verbali. Al contrario, la variabile I9 ha correlazioni medie con tutte le altre variabili (ad eccezione di I1), per cui la soluzione di rebus sembra implicare sia aspetti numerici sia aspetti verbali.

L'aver raggruppato i dati in questo modo ci ha reso molto più evidente i rapporti esistenti tra le variabili e ci ha permesso con più facilità di interpretare i risultati. In questo caso, essendo l'esempio immaginario e i dati pochi, è stato possibile realizzare raggruppamenti in maniera empirica ottenendo dei risultati parecchio chiari.

1.3 Aspetti teorici

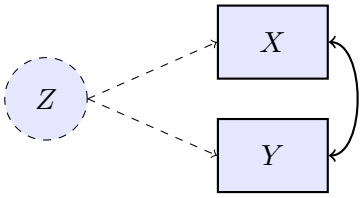
Nella realtà della ricerca psicologica i fatti non sono così semplici: il numero dei dati è sovente maggiore, sono presenti errori nel modo di rilevare i dati, il livello delle correlazioni non sempre risulta così nettamente differenziato... Da queste difficoltà è nata l'esigenza di elaborare una tecnica matematica e statistica che permettesse di ottenere dei raggruppamenti di variabili in base ad un tratto comune che li unifica e che consente di spiegare le correlazioni tra esse esistenti. Questa tecnica è l'**analisi fattoriale**, che (sebbene utilizzi sofisticati calcoli matematici) nei suoi processi più intuitivi e logici segue le fasi su esposte. Infatti il punto di partenza di questa tecnica è una matrice delle correlazioni tra le variabili.

Un indice di correlazione ci informa sull'andamento concomitante di due variabili, non è però possibile dire che, se A correla con B, allora A è causa di B oppure che B è causa di A. Di solito la relazione di correlazione è rappresentata graficamente in questo modo:



in cui i rettangoli indicano una variabile misurata e la linea (con frecce ad entrambi i lati) indica la co-relazione (in questo caso una correlazione, ma potrebbe indicare anche una covarianza).

Un alto indice di correlazione ci permette di capire che al variare di una, varia contemporaneamente anche l'altra. È impossibile dire se una delle due sia responsabile del variare della seconda oppure se entrambe siano legate ad una terza variabile (sconosciuta) che è responsabile del loro variare concomitante; situazione che possiamo rappresentare graficamente così:



in cui il cerchio tratteggiato rappresenta una variabile ipotizzata (cioè che non è stata misurata) e le frecce tratteggiate indicano la direzione dell'influenza.

Spearman, usando il concetto di "tetrade" (cioè tre variabili tra loro correlate), ipotizzò l'esistenza di una variabile latente (chiamata "fattore") che dovrebbe essere la responsabile dell'andamento concomitante delle tre.

Questo stesso ragionamento può essere esteso a quattro o più variabili tutte tra loro altamente correlate. Una di queste è responsabile del variare delle altre, oppure esiste un'ulteriore variabile a cui imputare il variare concorde di tutte?

L'Analisi Fattoriale parte appunto dall'ipotesi che esista questa ulteriore variabile che in qualche modo influenza ed agisce su un gruppo di variabili (che perciò risultano fra loro altamente correlate). Questa variabile sottostante (all'epoca definita *fattore*, oggi chiamata spesso "latente") agisce evidentemente su un particolare "tratto" comune a tutte le altre. L'analisi fattoriale, allora, si assume il compito d'individuare il o i fattori, cioè le variabili sottostanti ad un gruppo di altre variabili.

Per questo, nell'interpretare i fattori, si procede nel modo da noi seguito in precedenza. Si cerca cioè di trovare il significato comune alle variabili che confluiscono in un fattore, tenendo presente che tali variabili non hanno lo stesso peso nel determinare il significato di questo fattore. Infatti, in base ai calcoli matematici che poi spiegheremo, ad ogni variabile viene attribuito un valore, chiamato "saturazione" (in inglese *loading*, peso) che ci indica l'importanza che ha quella variabile nel determinare il significato del fattore. Allora, una variabile con una grande importanza in un fattore, ossia con un'alta saturazione in quel fattore, nella fase di interpretazione, influirà più di una variabile con una saturazione più bassa.

Possiamo quindi dire che, poiché l'analisi fattoriale si applica ad un'insieme di dati raccolti attraverso osservazioni, test, questionari... e tutta la gamma di misurazioni che sono comunemente utilizzate in psicologia, il suo scopo prioritario è quella di ridurre questa vasta quantità di misurazioni o variabili, attraverso delle variabili-fattore che siano in numero inferiore alle variabili di partenza e che spieghino tutte (o quasi) le correlazioni delle variabili di quel raggruppamento.

L'analisi fattoriale è stata inizialmente proposta da Charles Spearman nel 1904 e successivamente ampliata con il contributo di altri psicologi. Il contributo di alcuni matematici è stato importante non tanto nello sviluppo teorico, quanto nell'affinamento di alcune tecniche di calcolo.

Spearman sosteneva che tutte le misure di abilità mentale relative ad un test potessero essere spiegate come attribuibili ad un'*abilità generale* (G) e ad un'*abilità specifica* (S). Ogni abilità dipenderebbe (per Spearman) da due fattori: un fattore generale e un fattore specifico (G + S).

Applicando un test di Q.I. (Quoziente di Intelligenza) che comprenda una prova di abilità verbale, una di abilità numerica e una di abilità spaziale, il punteggio finale ottenuto (secondo Spearman) sarà determinato sia dai risultati ottenuti grazie all'abilità specifica del soggetto nei tre campi (fattori specifici) sia dal suo livello "generale" d'intelligenza (fattore generale).

Questa teoria, denominata teoria dei due fattori, ha orientato per vari anni le ricerche psicologiche finché Thurstone (1945) non propose una modifica, conosciuta come *teoria multifattoriale*.

Thurstone ritenne che non esistesse un unico fattore generale, ma piuttosto diversi fattori comuni (C). Questi fattori vengono chiamati "comuni" in quanto raccolgono solo alcune delle variabili in esame differenziandosi in tal modo dal fattore generale che risultava sottostante a tutte le variabili contemporaneamente. Ogni fattore comune ha caratteristiche diverse dagli altri fattori e dalle singole variabili che lo definiscono. Ciò che lo caratterizza in realtà è l'esprimere un tratto che è comune ad ognuna delle variabili in esso raccolte.

Ritorniamo al primo esempio del capitolo: consideriamo i due raggruppamenti che si sono formati tra le prove di abilità e di apprendimento somministrate a ragazzi di Scuola Media. In linea di massima potremmo dire che il primo raggruppamento corrisponde ad abilità di tipo verbale-linguistico, il secondo ad abilità legate alle conoscenze numeriche.

Secondo la teoria di Spearman il punteggio di un singolo partecipante sarà dovuto sia dal fattore generale sia dal fattore specifico di quello specifico item.

Secondo la teoria di Thurstone il punteggio di un certo soggetto in una prova (ad es. la risoluzione di problemi geometrici), sarà dovuta in gran parte ad abilità numerica, ma anche ad un piccolo contributo di tutte le altre abilità.

Anche ad un semplice livello intuitivo, si capisce che la capacità di una persona di risolvere problemi geometrici è legata alla capacità soggettiva di saper risolvere quella specifica prova, ma anche alla sua abilità più generale di tipo matematico, all'abilità di percepire gli elementi caratterizzanti la prova oltre che alla capacità di esprimere la soluzione in forma verbale o scritta. Si vede chiaramente che il peso di ciascuna di queste abilità è diverso e dipende dal tipo di problema; in questo caso l'abilità specifica e quella numerica saranno sicuramente più determinanti di quella linguistica o percettiva. Al contrario in una prova (ad esempio) di comprensione di un brano letto il peso dell'abilità linguistica sarà molto più determinante di quello dell'abilità matematica.

Questa teoria è tuttora alla base di tutti i modelli matematici di analisi fattoriale, modelli che sfruttano il principio della regressione multipla e sono, generalmente, di tipo additivo.

L'analisi fattoriale (sia esplorativa sia confermativa) è ormai una metodologia usata con larga diffusione sia in psicologia come anche, in generale, in tutte le scienze sociali. Essa si è rapidamente diffusa nei vari settori della psicologia perché il problema che si incontra solitamente nella ricerca psicologica è proprio quello di cercare quali componenti o fattori sottostanno ad uno stesso comportamento globale oppure, anche, quello di trovare quali relazioni esistono fra comportamenti, reazioni, manifestazioni psicologiche in generale emessi in determinate situazioni.

L'analisi fattoriale è stata ampiamente utilizzata in psicologia proprio perché permette:

a) di conoscere le relazioni che esistono tra le variabili relative ad un certo comportamento a manifestazione psicologica e su cui non è ancora possibile stabilire ipotesi di lavoro.

Esempio 1.1

Quando si parla di comportamento sociale di bambini in età prescolare s'intende riferirsi ad un gruppo di comportamenti, abbastanza facilmente definibili (quali il contatto, il gioco, la comunicazione, l'aggressività) che anche a livello intuitivo sono strettamente connessi tra loro e reciprocamente influenzabili e interdipendenti. Se è facile comprendere tutto questo, è più difficile capire fra quali comportamenti esiste realmente una relazione, quali sono invece indipendenti, qual è il "motivo" che fa mettere in rapporto alcuni comportamenti invece di altri. Per saperlo si ricorre all'analisi fattoriale. Raccolti i dati relativi al comportamento sociale attraverso un'osservazione diretta del comportamento di un campione di bambini, in cui vengono rilevati i comportamenti del tipo "giocare insieme ad altri bambini con un oggetto", "giocare insieme ad altri bambini senza oggetto", "giocare con l'adulto", "sorridere a qualcuno", "picchiare un altro bambino", "piangere", "giocare da soli", ecc. sarà possibile vedere come una serie di variabili indicanti i comportamenti osservati vengono a riunirsi insieme, dando origine ad una serie di raggruppamenti, che ci indicano il modo in cui vengono raggruppati i comportamenti di "gioco insieme con i coetanei", di "sorriso", "verbalizzazione", escludendo invece comportamenti quali "contatto aggressivo", "gioco da soli"...

b) la conferma di un'ipotesi specifica (che è anche obiettivo della cosiddetta "analisi fattoriale confermativa"), da cui si è partiti, sulle relazioni esistenti tra alcune variabili psicologiche.

Esempio 1.2

In un questionario elaborato nel 1975 da Dan Olweus e relativo all'autovalutazione dell'aggressività in preadolescenti, si era interessati a valutare le manifestazioni aggressive sotto certi particolari aspetti: le manifestazioni aggressive di tipo verbale, quelle di tipo fisico e le manifestazioni implicanti un'inibizione della risposta aggressiva. Si è quindi costruito un questionario che contenesse una serie di item relativi a questi tre aspetti particolari: per la manifestazione fisica dell'aggressività, item tipo "Faccio a botte con gli altri ragazzi della scuola", per quella verbale "Quando un adulto è ingiusto con me, mi arrabbio e protesto", per l'inibizione dell'aggressività "Quando un adulto è irritato con me, generalmente mi sento in colpa". L'ipotesi della ricerca era che ciascuno di questi item fosse relativo ad un diverso aspetto di un determinato tipo di aggressività. Si è quindi sottoposto il questionario, con gli item fra loro mischiati, a due gruppi di circa 80 soggetti l'uno. Ogni soggetto doveva giudicare su una scala da 1 a 5 se l'affermazione proposta dall'item era più o meno appropriata a se stesso. L'uso dell'analisi fattoriale ha permesso di verificare l'esattezza di quest'ipotesi, infatti gli item da Olweus selezionati come indicativi di manifestazione fisica dell'aggressività sono risultati appartenere ad un unico fattore. Analoghi risultati si sono ottenuti per gli altri due tipi di item.

c) verificare come i rispondenti interpretano le variabili (item) che compongono lo strumento sottoposto ad analisi.

Esempio 1.3

Con riferimento all'Esempio 1.2. Fra gli item di quel questionario c'erano 5 item cosiddetti *Riempitivi* il cui scopo era di verificare se i partecipanti stavano rispondendo coerentemente ai contenuti oppure stessero rispondendo in modo casuale oppure lasciandosi guidare dalla *desiderabilità sociale*. Item di *desiderabilità sociale* sono spesso usati per vedere se il desiderio di essere accettati dagli altri possa alterare le risposte. Uno di questi item era "Mi piace la musica rock" (risposta casuale) e un'altro "Sono sempre felice e contento" (desiderabilità sociale). Nella ricerca originale questi 5 item confluivano in un unico fattore. Nella nostra ricerca l'item "Mi piace la musica rock" è confluito in un fattore di *Aggressività verso l'adulto*, contrariamente a quanto ci aspettavamo: ascoltare musica rock non è un comportamento aggressivo e non sottintendeva aggressività (secondo gli autori dello strumento), ma i ragazzi e ragazze che hanno risposto al questionario l'hanno percepito come "aggressivo verso l'adulto".

Ribadiamo ulteriormente che l'analisi fattoriale è una metodologia statistico-matematica che permette di esplicitare e formalizzare due modi di operare tipici della psicologia: o raggruppando insieme eventi o variabili che si manifestano o mutano congiuntamente, per cogliere il legame esistente tra loro, oppure individuando le influenze, più o meno importanti, e i comportamenti che sottostanno ad ulteriori comportamenti o avvenimenti.

1.4 Sintesi

Abbiamo cercato di spiegare in questo capitolo in modo estremamente intuitivo cosa sia l'analisi fattoriale, mentre i capitoli seguenti forniranno delle spiegazioni più dettagliate e precise. Nel prossimo capitolo analizziamo quelle nozioni di analisi fattoriale necessarie a chi, pur volendola applicare, non desidera approfondirne, immediatamente, gli aspetti matematici.

Capitolo 2

Come si esegue un'analisi fattoriale esplorativa: indicazioni intuitive

2.1 Introduzione

In questo capitolo cercheremo di illustrare le varie fasi attraverso cui si passa nell'effettuare un'Analisi Fattoriale Esplorativa (AFE) mantenendoci ancora ad un livello intuitivo e rimandando i dettagli matematici. Come abbiamo già accennato lo scopo fondamentale di ogni teoria fattoriale è quello di cogliere le variabili latenti (i fattori) capaci di spiegare le relazioni comuni ad un ampio gruppo di variabili, di solito rilevate mediante l'uso di una misura di correlazione.

Ricordiamo che per correlazione si intende la misura del grado di interdipendenza reciproca fra due variabili: se all'aumentare di una aumenta anche l'altra e al diminuire di una corrisponde la diminuzione dell'altra si parlerà di correlazione positiva; se all'aumentare di una corrisponde il diminuire dell'altra, si parlerà di correlazione negativa.

Il punto di partenza di ogni analisi fattoriale è quindi l'insieme di tutte le correlazioni tra un gruppo N di variabili che, per facilitarne lo studio, viene disposto in una matrice delle correlazioni, cioè in una tabella a doppia entrata in cui all'incrocio fra una riga e una colonna si trova quel valore che rappresenta la correlazione tra le variabili indicate in quella stessa riga e colonna.

Ottenuta la matrice delle correlazioni, il passo successivo è quello di determinare quali e quanti siano i fattori che spiegano tale matrice. Si tratta quindi di rispondere alla domanda: "Quanti e quali fattori sono necessari (o sufficienti) per spiegare questa matrice delle correlazioni?". Questa domanda si suddivide in più parti, la prima è il numero di fattori da estrarre (cfr. 2.4.1), la seconda è come identificarli (cfr. 2.5), poi bisogna renderli interpretabili (cfr. 2.6) ed infine come interpretarli (cfr. 2.7).

Per il numero dei fattori si possono usare metodi teorici, pragmatici oppure metodi statistici. Alla seconda domanda è possibile dare una risposta attraverso dei metodi matematici chiamati "metodi di estrazione dei fattori", che consentono di individuarli.

Tuttavia, se da un punto di vista matematico, il numero dei fattori che si possono estrarre è ben determinato, per altri motivi, la loro determinazione non è per nulla semplice. Si deve ricorrere a dei metodi approssimati ed è spesso necessario utilizzare contemporaneamente più metodi.

Una volta identificato il numero dei fattori e dopo averli determinati secondo uno dei diversi metodi di estrazione, si è risolto solo il primo problema fondamentale dell'analisi fattoriale, quello appunto di trovare una possibile soluzione fattoriale, cioè un insieme di k fattori che spieghino la matrice delle correlazioni tra le n variabili.

Tuttavia, trovato che da una particolare matrice sia possibile estrarre k fattori, questo non ci dice nulla circa il modo di identificarli. Bisogna infatti precisare che, sempre a livello matematico, è possibile trovare infinite soluzioni di k fattori e ciascuna di queste soddisfa e spiega, matematicamente e in modo più o meno efficace, la stessa matrice delle correlazioni.

Si apre a questo punto una seconda questione (denominata problema dell'*indeterminazione fattoriale*) e si tratta allora di cercare una risposta alla successiva domanda: "Quale fra le infinite soluzioni scegliere?".

La risposta non può essere evidentemente di tipo matematico perché le soluzioni, da questo punto di vista, sono tutte ugualmente (o abbastanza) buone. Si dovranno invece, individuare dei criteri di ordine diverso, che si adeguino in ogni caso allo scopo dell'analisi fattoriale. Se teniamo presente che tale scopo è quello di individuare dei fattori che spieghino dei gruppi di variabili, si dovranno probabilmente cercare dei criteri del tipo "ogni fattore dovrebbe influenzare in maniera massima un gruppo di variabili ed in maniera minima tutte le altre".

I criteri fino ad oggi individuati si pongono proprio su questa linea e prendono il nome di criteri per la individuazione della *struttura semplice*. Da un punto di vista storico, si trattava inizialmente di criteri del tutto empirici, che poi sono stati via via ricondotti a formule matematiche.

Il compito di questo capitolo è appunto quello di indicare le possibili risposte a queste due questioni. In particolare, questi due problemi saranno meglio analizzati e spiegati, sempre in mo-

do intuitivo, nei paragrafi dedicati alla *Determinazione dei fattori* (cfr. 2.4) e in quello dedicato alla *Rotazione dei fattori* (cfr. 2.6). Ovviamente, in questo capitolo, si tratterà anche di tutti i passaggi necessari per poter giungere alla soluzione di tali problemi, partendo dai dati di cui si dispone.

Inoltre, poiché ormai l'analisi fattoriale viene effettuata con l'ausilio degli elaboratori elettronici, questo capitolo farà spesso riferimento al loro uso. Proprio per questo motivo, in alcuni casi, i temini italiani saranno accompagnati dalla relativa terminologia inglese. In questo modo si potranno più facilmente leggere e capire i tabulati contenenti i risultati di un'analisi fattoriale effettuata da un elaboratore.

2.2 Cosa c'è prima di un AFE

Un'analisi fattoriale non si può applicare a qualsiasi cosa e a qualsiasi variabile, qualunque essa sia. Ci sono dei presupposti di ricerca e dei tipi di ricerca che portano "naturalmente" all'uso di una tecnica di analisi fattoriale.

Un primo criterio è avere un ambito o un "dominio di ricerca" specifico (che potrebbe corrispondere ad un costrutto psicologico), ad es. la *religiosità*.

All'interno di questo ambito dobbiamo selezionare un certo numero di variabili osservabili che ci permettano di riconoscere certi aspetti del costrutto e dobbiamo esprimere queste variabili in forma verbale. Ad es. per quanto riguarda la religiosità, possiamo distinguere le credenze religiose ("credo in ...") dai comportamenti religiosi ("faccio ..."). Poi si cercherà di rendere questi due aspetti in termini verbali: ad es. "Credo in Dio", "Credo nell'aldilà"... e "Prego", "Frequento le funzioni religiose". Sebbene sia possibile applicare un'analisi fattoriale su qualunque tipo di variabile (per ora consideriamo solo variabili misurate a livello "intervallo"), sarebbe preferibile fare in modo che le variabili sia più o meno simili fra loro nel modo di rispondere. Ad es. per le credenze si possono scrivere item che implicano la risposta ad una domanda principale del tipo "Quanto le affermazioni seguenti si applicano a te:" e poi usare come forma di risposta una dimensione del tipo "per nulla"-"perfettamente" graduata a sette, nove o undici gradini.

Successivamente si sottoporrà lo strumento che è stato costruito ad un gruppo numeroso di persone facendo in modo che sia il più possibile disomogeneo cioè che comprenda persone di diverse età, professioni, livelli di studio... per evitare che lo strumento rifletta il "costrutto" di un gruppo particolare di individui.

Sottoponendo questo questionario ad un campione vasto di persone ipotizziamo che tutte (o comunque una parte) le affermazioni che riguardano le credenze correlino fra loro, e che quelle che riguardano i comportamenti altrettanto, ma che le variabili di credenza non correlino con quelle di comportamento. Quest'ipotesi non è quasi mai realistica, ma noi *vorremmo* poter selezionare quelle affermazioni che meglio "rilevano", in modo separato e indipendente, le credenze e i comportamenti.

Passaggi per un'AFE

Nell'effettuare un'analisi fattoriale esplorativa bisogna essere sicuri di quali dati si possano usare e bisogna quindi verificare che si possa fare l'analisi controllando i livelli di misura delle variabili, la loro normalità e linearità, gli eventuali "outliers", il numero di variabili che si intendono usare e il numero di fattori che ci si attende di trovare rispetto al numero di soggetti.

In questo momento stiamo affrontando l'AFE classica che si applica a variabili di tipo intervallo. Bisogna quindi assicurarsi che tutte le variabili da usare siano effettivamente a livello intervallo o rapporto. Più avanti si vedrà che è possibile fare un'analisi fattoriale esplorativa sia con dati ordinali (Lisrel, MPlus, R) sia con dati nominali o misto (MPlus).

A livello intervallo (o rapporto), l'analisi verrà effettuata usando la correlazione di Pearson, quindi è bene verificare la normalità delle variabili e anche l'omoschedasticità. Se i dati sono stati

immessi a mano, è bene verificare che non ci siano valori anomali dovuti ad errori di immissione, verificando in primo luogo il minimo e il massimo di ogni variabile (se una variabile è una scala a 11 gradini e incontriamo un massimo di 33, c'è un errore di immissione). Eliminati gli errori di immissione, è bene verificare i valori di asimmetria e di curtosi (quindi una verifica della normalità) e successivamente i cosiddetti *outlier* cioè i valori che, pur essendo leciti, minano la normalità.

Un'altro punto da considerare è l'ampiezza del campione. Su questo punto ci sono indicazioni della letteratura diverse e a volte contrastanti che verranno affrontate successivamente (cfr. 2.3).

2.3 Scelta del campione

Il problema è già conosciuto all'interno della psicologia ma è importante sottolinearlo in relazione a questa metodologia. Si deve fare in modo di ridurre, al massimo, la possibilità di errore relativo al campione stesso e alle risposte date dal singolo soggetto, che costituiscono quella che in termini matematici viene chiamata varianza dell'errore, la quale interviene nel determinare la precisione delle saturazioni dei fattori estratti.

Gli aspetti da considerare nella scelta del campione sono la sua ampiezza rispetto ai tratti misurati e la sua rappresentatività rispetto alle caratteristiche che si ipotizzano correlate con i fattori. Se lo scopo della ricerca è semplicemente l'identificazione di fattori in una nuova area di ricerca, è preferibile un campione molto ampio. "Se comunque i risultati devono essere generalizzati, a una popolazione accuratamente definita" (Fruchter, 1954), il campione deve essere rappresentativo delle caratteristiche studiate.

A tal fine è estremamente importante decidere sulla eterogeneità o sulla omogeneità del campione rispetto alle variabili da studiare. Ad esempio, se si vogliono studiare delle variabili specifiche di una particolare zona geografica, il campione da esaminare dovrà essere rappresentativo di quella regione, vale a dire che nella selezione del campione bisognerà rispettare la proporzione esistente nella popolazione a cui ci si riferisce per quanto riguarda le variabili più importanti quali il genere, l'età, il livello socio-economico, il livello culturale... rispetto alla o alle variabili specifiche studiate.

Se invece si cerca di determinare l'eventuale esistenza di fattori comuni, cioè se si ipotizza che, rispetto a dei compiti proposti, esistono dei possibili raggruppamenti sottostanti la riuscita di essi, si dovrà perseguire al massimo l'eterogeneità del campione in modo da eliminare l'importanza di differenze individuali attribuibili a piccoli gruppi di soggetti. Così se si vuole standardizzare un test, un questionario, una scala o qualunque tipo di misura utilizzata in psicologia è necessario avere un campione estremamente eterogeneo comprendente cioè individui appartenenti ai possibili livelli economici, sociali, di genere, di età, di zone geografiche... in modo da evitare che una qualunque di queste variabili influenzi in modo determinante la standardizzazione del test. In questo caso la numerosità del campione e la sua eterogeneità servono per poter eliminare l'influenza sulla misurazione di alcune variabili individuali che non possono essere precedentemente controllate, misurate e neppure stimate quantitativamente.

Esempio 2.4

Nelle ricerche effettuate allo scopo di adattare il questionario di Olweus ad un campione rappresentativo dei pre-adolescenti e degli adolescenti sono stati scelti una serie di campioni omogenei rispetto alle variabili età e genere, zone geografiche di residenza, eterogenei rispetto alle variabili livello socio-economico e tipo di scuola frequentata. I campioni utilizzati sono stati i seguenti:

1. 91 soggetti (54 maschi e 37 femmine) di età compresa fra i 14 e i 19 anni, frequentanti un Istituto Professionale (58 ss.) e un Liceo Scientifico (33 ss.) di una città dell'Italia

Settentrionale di livello economico basso ed elevato;
2. 100 soggetti (50 maschi e 50 femmine) di età compresa fra i 17 e i 19 anni frequentanti un Istituto Magistrale (femmine) e un Istituto Tecnico (maschi) di una città dell'Italia Settentrionale e di livello socio-economico medio-basso;
3. 202 soggetti di età compresa fra i 10 e i 18 anni di entrambi i sessi, frequentanti le Medie Inferiori (82 ss.) e il Liceo Scientifico (130 ss.) di un paese dell'Italia Settentrionale.

I campioni scelti presentano aspetti di omogeneità e di eterogeneità sia se considerati individualmente sia rispetto all'insieme di tutti i soggetti. Infatti le variabili genere e zona geografica sono omogenee perché le abbiamo mantenute costanti nei vari campioni; ugualmente eterogenee sono le variabili età, livello socio-economico e tipo di scuola che risultano diverse nelle varie ricerche.

In conclusione il campione da utilizzare nell'analisi fattoriale deve essere abbastanza numeroso ed in genere si può ritenere tale un campione formato da un numero di soggetti maggiore delle variabili misurate. Ma quanto deve essere grande questo numero?

Nel corso del tempo sono stati proposti criteri diversi (alcuni logici e di buon senso, altri basati su principi matematici) che possiamo distinguere in due categorie e riassumere brevemente.

1. *prima dell'analisi dei dati*: questi criteri si possono applicare ancora prima di effettuare un'analisi, se non addirittura prima di raccogliere i dati.
   - più osservazioni abbiamo meglio è (Kline, 1994, p. 73);
   - più casi statistici che variabili e comunque non meno di 100;
   - 200 sono il minimo (Guilford, 1956);
   - 5 osservazioni per ogni variabile (Stevens, 1986);
   - almeno il doppio delle variabili che si intendono analizzare (Harris, 1967);
   - da 10 a 15 casi statistici per variabile (Nunnally, 1978);
   - da 5 a 10 partecipanti per variabile ma non meno di 300 (Kass & Tinsley, 1979)
   - minimo 5 casi per variabile ma non meno di 100 casi in totale (Gorsuch, 1983);
   - un campione di almeno 300 casi (Guadagnoli & Velicer, 1988; Tabachnick & Fidell, 2007) (p. 613);
   - 100 sono pochi, 1000 sono troppi, 300 sono sufficienti (Comrey & Lee, 1992)
2. *dopo l'analisi dei dati*: questi criteri invece, assumono che sia stata fatta almeno un'analisi iniziale dei dati, da cui poter ricavare saturazioni e comunalità:
   - ogni fattore deve avere almeno 4 saturazioni superiori a .60, non importa l'ampiezza del campione (Guadagnoli & Velicer, 1988);
   - ogni fattore deve avere almeno 10 saturazioni superiori a .40 e il campione dev'essere di almeno 150 casi (Guadagnoli & Velicer, 1988);
   - bastano anche 60 casi se tutte le comunalità sono maggiori di .60 (MacCallum, Widaman, Zhang, & Hong, 1999);
   - con comunalità attorno a .50, il campione dev'essere tra 100 e 200 casi (MacCallum et al., 1999).

Alcuni di questi criteri (quelli che utilizzano le saturazioni e le comunalità finali) implicano aver già effettuato un'analisi fattoriale esplorativa e di dover quindi ampliare eventualmente il campione.

2.3.1 Dai dati alla matrice di correlazione

Il primo passo di un'analisi fattoriale è il passaggio dai dati alla matrice delle correlazioni. Una misura di correlazione può essere utilizzata per misurare il grado di associazione delle variabili oppure dei soggetti, degli eventi...

In linea di massima si utilizza la correlazione prodotto-momento di Pearson, in quanto soddisfa al criterio di linearità che è una dei presupposti delle teorie fattoriali. È però possibile

utilizzare altri tipi di misure di correlazione ed in particolare il coefficiente *rho* di Spearman, senza che ciò influenzi in moda particolare i calcoli (Gorsuch, 1974). Le tecniche più recenti utilizzano anche matrici di associazione basati su correlazioni tetracoriche (due dicotomiche), poliseriali (una dicotomica e una intervallo) o policoriche (due ordinali tipo Likert).

Un'analisi fattoriale effettuata sulla matrice delle correlazioni fra variabili si chiama **R-factor analysis**, mentre quella effettuata sulla matrice delle correlazioni fra soggetti **Q-factor analysis**. Nel corso di questo libro ci si riferisce sempre ad una analisi di tipo R.

Se l'analisi viene effettuata tramite un elaboratore, è importante disporre i dati nel modo appropriato al tipo di analisi che si vuol ottenere. Nel caso di una R-analisi i dati devono essere disposti così come appaiono nella Tabella 1.2 del capitolo 1; vale a dire che, per ogni soggetto, bisognare elencare il punteggio da lui ottenuto in tutte le variabili.

Come abbiamo detto più volte, il punto di partenza di un'analisi fattoriale è la matrice delle correlazioni. Spesso, in questa matrice, lungo la diagonale principale devono essere inseriti dei particolari valori denominati "comunanze" o "comunalità" (*communality*) che vengono individuati con particolari procedimenti matematici di calcolo che saranno analizzati più avanti e che, come si vedrà, non possono essere calcolati in modo preciso (se non alla fine dell'analisi) ed è perciò necessario effettuarne una stima.

Nella pratica, poiché l'analisi fattoriale viene generalmente effettuata tramite l'ausilio di elaboratori, la scelta della stima della comunanze, da parte nostra, non è sempre necessaria. Infatti, la maggior parte dei programmi di analisi fattoriale sceglie autonomamente la stima della comunanza più appropriata al metodo di analisi prescelto.

2.3.2 Analisi della matrice di correlazione

Perché un'analisi fattoriale possa produrre dei fattori rilevanti bisogna che la matrice di correlazione contenga valori elevati accanto ad altri di bassa entità. Un primo modo per scoprirlo è quello di stampare e osservare la matrice di correlazione per vedere se ci sono correlazioni superiori a .30, ma la cosa non è facile se il numero di variabili è alto.

Oppure si può fare un'ispezione visiva della matrice di correlazione per vedere che abbia blocchi di correlazioni alte fra alcune variabili e basse con le alte (come abbiamo fatto nel primo capitolo) oppure tramite una rappresentazione grafica (Figura 2.1a, in questo caso fatta tramite il software R). Anche questa operazione non è facile con molte variabili ed inoltre bisogna provvedere a ri-ordinare le variabili per poter evidenziare i blocchi (Figura 2.1b).

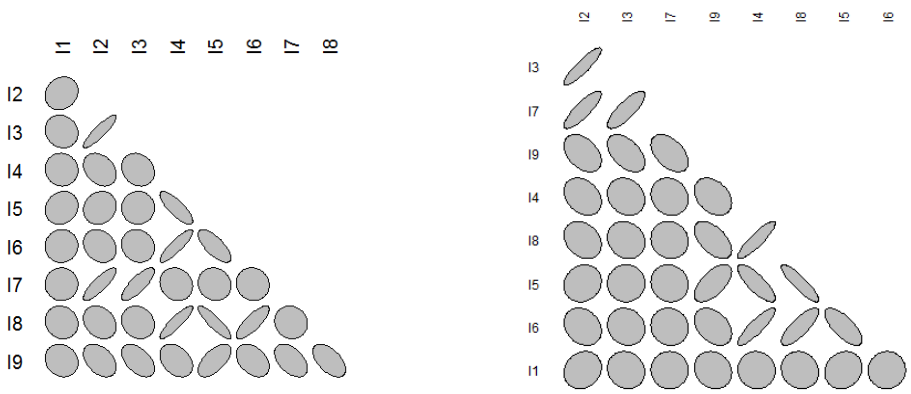


Figura 2.1 – Ispezione visiva della matrice di correlazione

Matrice di correlazione[a]

a. Determinante = ,000

Test di KMO e Bartlett

Misura di Kaiser-Meyer-Olkin di adeguatezza del campionamento.		,514
Test di sfericità di Bartlett	Appross. Chi-quadrato	159,374
	gl	36
	Sign.	,000

Figura 2.2 – Esempio di KMO

Il determinante della matrice di correlazione è un primo indice che si può utilizzare (ma con scarsi risultati). Se è nullo, non si può effettuare un'AFE (evento molto raro). Quando è elevato, possiamo dire che le correlazioni sono generalmente basse e viceversa (se tutte le correlazioni fossero uguali a zero, il determinante sarebbe 1). Quindi servirebbe un valore di determinante basso (che indicherebbe molte correlazioni alte) ma diverso da zero. Spesso un determinante inferiore a 0.00001 (Field, 2009, p. 648) viene considerato troppo basso (perché indicherebbe che ci sono troppe variabili multicollineari ovvero troppe variabili troppo correlate fra loro). Non è però molto facile stabilire "quanto buono" sia un determinato valore del determinante.

Una seconda possibilità è tramite il *Test della sfericità di Bartlett* che cerca di verificare l'ipotesi nulla che la matrice di correlazione $\mathbf{R}$ sia una matrice identità (H_0: $\mathbf{R} = \mathbf{I}$), ovvero se i valori fuori dalla diagonale principale siano zero e quelli lungo la diagonale principale siano 1. Si basa sul valore del determinante e si distribuisce come un chi-quadro (se interessati, la formula è a p. 31). Se è significativo, significa che $\mathbf{R}$ ha correlazioni sufficientemente elevate da non essere paragonabili a 0; se è non significativo le correlazioni sono basse e non si distinguono da 0 ma questo test dipende dal numero delle variabile e dalla numerosità del campione, quindi tende ad essere significativo all'aumentare del campione e del numero delle variabili anche se ci sono correlazioni basse.

Una terza indicazione è il *Test di adeguatezza campionaria di Kaiser-Meyer-Olkin* (KMO). Si basa sulle correlazioni di ogni variabile con ogni altra, parzializzate su tutte le restanti. Se ci sono variabili correlate fra loro (la base per ipotizzare un fattore comune), la loro correlazione "reale" senza le influenze delle altre dovrebbe essere molto bassa. Secondo Kaiser (19?7), se il valore di KMO è maggiore di 0.90 è eccellente; se compreso fra .80 e .90 buono; se fra .70 e .80 accettabile; fra .60 e .70 mediocre e se inferiore a .60, allora è meglio non fare l'analisi.Perciò un valore di KMO maggiore di 0.60 è generalmente considerato come l'indicazione che si può effettuare un'AFE con una certa probabilità di ottenere dei risultati. Nonostante questo test si chiami "adeguatezza campionaria" non ha nulla a che fare con il campione (come si può vedere dalla sua formula a p. 31).

Infine si potrebbe osservare la matrice anti-immagine, che contiene il complemento a 1 delle correlazioni parziali. Ovviamente funziona al contrario rispetto all'indice KMO e quindi cercheremo all'interno della matrice anti-immagine i valori elevati. Anche questa è una tecnica manuale lunga e poco utilizzata.

Nel Figura 2.2, vediamo un esempio dei risultati forniti da IBM SPSS relativamente a questi indicatori. Il determinante è visualizzato con soli 3 decimali e sembra essere un valore inferiore a .00049999 (altrimenti verrebbe stato arrotondato a .001) ma non sappiamo quanto sia inferiore. KMO equivale a .703, quindi superiore a .60 e il Test di Bartlett è significativo.

2.4 Determinazione dei fattori

2.4.1 Numero dei fattori da estrarre

Il numero dei fattori da estrarre dipende dal valore assunto dalle comunanze nel senso che una volta conosciute le comunanze, anche il numero dei fattori è esattamente determinato. Purtroppo, come abbiamo già osservato, non è possibile conoscere i valori esatti delle comunanze prima di aver effettuato un'analisi, ma soltanto stimarli e questo comporta, come conseguenza, che non tutti i fattori teoricamente estraibili siano dei fattori "veri" e non, invece, dei fattori di errore.

A questo primo problema si aggiunge il fatto che il numero esatto dei fattori può essere determinato solo quando le correlazioni sono calcolate sull'intera popolazione, mentre nella realtà esse sono misurate in un campione e questo potrebbe condurre ad altri fattori di errore.

Un terzo tipo di problema che influenza il numero dei fattori riguarda un aspetto che potremo definire di tipo psicologico: non sempre i fattori "significativi" da un punto di vista matematico, sono tali da un punto di vista "(psico)logico" e questo per varie e differenti ragioni. La più evidente è che un fattore deve saturare in maniera elevata un certo numero di variabili affinché gli si possa attribuire un significato psicologico; può accadere che un fattore "significativo" da un punto di vista matematico non soddisfi a questa condizione.

D'altra parte, proprio in relazione all'uso dei calcolatori è necessario, oggi, sapere in anticipo il numero dei fattori da estrarre oppure fornire al software un criterio che lo guidi nei calcoli e su cui possa basarsi per arrestare l'estrazione dei fattori.

Alcuni di questi metodi non trovano oggi una loro utilizzazione pratica in quanto non è sempre possibile disporre di alcuni dati intermedi dei calcoli (in quanto ogni software visualizza i risultati parziali che i loro programmatori ritengono utili). Cerchiamo di descriverli qui di seguito in maniera intuitiva e, soprattutto, operativa.

Possiamo dividerli in metodi che non richiedono un'analisi preventiva (quindi teorici e indipendenti dai dati), metodi che si basano su valori statistici ottenibili solo da un'analisi preliminare e metodi che si basano su analisi più complesse.

Si consideri tuttavia che già Gorsuch (1983) riteneva necessario effettuare più analisi, studiare le varie soluzioni e prendere in considerazione quei fattori che si mantengono abbastanza stabili nelle varie soluzioni.

2.4.2 Metodi teorici

Un primo criterio che si può seguire è quello di basarsi sull'ipotesi di ricerca: la teoria o lo scopo della ricerca possono suggerire un numero di fattori da estrarre; successivamente sarà possibile verificarne l'effettiva significatività ed eventualmente ri-effettuare l'analisi fattoriale stessa. Ad es., volendo confrontare i risultati ottenuti sui dati di un questionario italiano (tradotto da un'altra lingua) con quelli ottenuti sui dati della ricerca originale, si può estrarre lo stesso numero di fattori della ricerca di confronto.

Un secondo criterio può basarsi sulle caratteristiche dello strumento che si intende sottoporre ad analisi. Se lo strumento è stato costruito per avere un certo numero di sottoscale che misurino aspetti diversi dello stesso costrutto, allora quello potrà essere il numero di fattori da chiedere (almeno alla prima analisi).

Un terzo criterio teorico è legato al numero di variabili che saranno sottoposte ad analisi. Ci si aspetterebbe che ogni fattore spieghi almeno 5 variabili. Quindi un criterio potrebbe essere di estrarre tanti fattori in base a "numero di variabili diviso 5". Non è un criterio molto usato, perché implicherebbe che non si conosca assolutamente nulla sulle variabili che compongono lo strumento. Più spesso viene applicato *dopo* l'analisi per selezionare le variabili "migliori" di ogni fattore oppure per scartare quei fattori che raccolgono meno di 5 variabili.

2.4.3 Metodi dipendenti da un'analisi preliminare

Normalmente l'analisi fattoriale viene effettuata in due passaggi: si effettua una prima analisi fattoriale a carattere totalmente esplorativo. Se l'analisi è effettuata tramite l'ausilio di un elaboratore, ogni software estrarrà il numero di fattori che ritiene utile (in genere usando il criterio di Guttman). Una volta ottenuta questa prima matrice fattoriale si può chiederne una seconda in cui invece, il numero dei fattori sarà predeterminato utilizzando uno degli criteri seguenti.

Un primo criterio è quello proposto da Guttman (1954)(Kaiser, 1960), con il quale si decide di estrarre tanti fattori quanti sono gli autovalori (*eigenvalues*) maggiori o uguali a un numero prefissato che è generalmente 1 (colonna "Totale" di Figura 2.3). La logica di questo criterio dipende dal fatto che gli autovalori corrispondono alla varianza dei fattori e che, partendo da una matrice di correlazione in cui tutti i valori sono standardizzati (quindi con varianza pari a 1), i fattori estraibili debbano raccogliere più varianza di quanto farebbe una variabile da sola. Questa è la scelta adottata, in modo automatico nella maggior parte dei software. L'uso di questo criterio conduce però all'estrazione di un ampio numero di fattori che viene, generalmente, considerato *eccessivo* rispetto al numero di variabili analizzate.

Varianza totale spiegata

	Autovalori iniziali		
Fattore	Totale	% di varianza	% cumulata
1	4,961	49,611	49,611
2	3,743	37,429	87,041
3	,532	5,319	92,359
4	,272	2,724	95,084
5	,177	1,767	96,850
6	,102	1,020	97,870
7	,078	,785	98,655
8	,071	,710	99,365
9	,048	,480	99,845
10	,015	,155	100,000

Metodo di estrazione: Fattorizzazione dell'asse principale.

Figura 2.3 – Esempio di tabella della proporzione cumulativa della varianza totale

Sempre basandosi sugli autovalori, Cattell (1966) ha proposto un metodo grafico conosciuto con il nome di *scree test*. Il nome sembra far riferimento all'operazione di separazione della sabbia dai sassolini (*scree* equivale a ghiaia, pietrisco) che i cercatori d'oro facevano per separare i sassolini più grossi, lasciando scivolare via la sabbia e i sassolini più piccoli.

Si rappresentano graficamente su un piano cartesiano gli autovalori (in ordinata) relativi a ciascun fattore (sull'ascissa) con dei punti che verranno poi collegati da una linea. In corrispondenza del punto in cui la curva smettere di cadere e tende a divenire più simile ad una retta si pone il limite dei fattori da estrarre che risulteranno probabilmente significativi. La logica di questo metodo è quello di selezionare quei fattori i cui autovalori implichino una certa quantità di varianza e in cui un fattore successivo non spieghi troppo poca varianza rispetto al precedente (quindi la rappresentazione grafica non tende a diventare piatta). In termini più matematici, il numero di fattori dovrebbe coincidere con il punto di flesso della curva (Cattell, 1966), anche se Harmann (19??) propone di scegliere il fattore immediatamente precedente al punto di flesso.

Dal momento che è un criterio visivo, il rischio è quello di *estrarre meno fattori* di quanti potrebbero essere effettivamente significativi. Un secondo rischio è legato alla scala con cui viene rappresentato il grafico; infatti a seconda dell'unità che si sceglie per distanziare i fattori, la linea diverrà piatta più o meno velocemente (Runyon & Haber, 1976, cap. 1).

Per questo motivo, un ulteriore criterio consiste nel controllare la tabella della proporzione cumulativa della varianza totale e chiedere un numero di fattori tale da raggiungere almeno

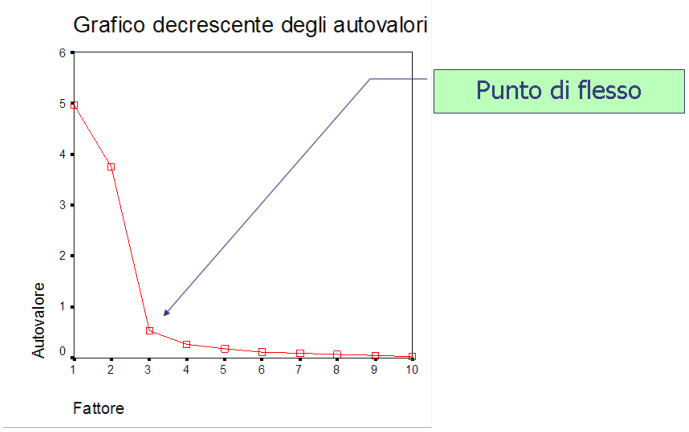


Figura 2.4 – Esempio di screetest

Test della bontà di adattamento

Chi-quadrato	gl	Sign.
116,134	53	,000

Figura 2.5 – Chi quadro per la bontà dell'adattamento

il 60/75%. Questo criterio risulta particolarmente valido quando la matrice delle correlazioni contiene molti valori particolarmente alti, mentre è del tutto inutile quando i valori della matrice di correlazione tendono ad essere generalmente bassi. In questo caso infatti sono necessari molti fattori per poter raggiungere il 75% della varianza totale e la maggior parte di questi fattori spiega una parte piccolissima di varianza. Il rischio di questo metodo è quello di estrarre come validi un *numero di fattori maggiore* di quanti sarebbero effettivamente significativi.

2.4.4 Metodi che si basano su analisi più complesse

Alcuni metodi di estrazione dei fattori (ad es. la massima verosimiglianza) possono applicare un test di chi-quadro (Figura 2.5) sulla soluzione trovata che verifichi se, con quel numero di fattori, riproduce abbastanza bene la matrice di correlazione di partenza (un approccio usato anche nell'Analisi fattoriale confermativa). Se il chi-quadro è statisticamente significativo, il numero di fattori estratto non è sufficiente e bisognerebbe aggiungerne un altro. Il problema di questa tecnica è che il chi-quadro tende ad essere significativo all'aumentare del numero di variabili e quindi sarebbe bene "verificare" questo risultato con qualche indice ulteriore. Considerando che IBM SPSS riporta solo alcune informazioni, l'unico ulteriore indice utilizzabile è l'RMSEA da calcolare a mano utilizzando la formula:

$$\text{RMSEA} = \sqrt{\frac{\chi^2 - gl}{N * gl}}$$

In IBM SPSS questa tecnica viene calcolata solo con i metodi della Massima verosimiglianza e dei Minimi quadrati generalizzati (cfr. 2.5), mentre in altri software (come R), è stata ulteriormente ampliata. Qualsiasi sia il metodo di estrazione, di rotazione e il numero dei fattori estratti, con i risultati ottenuti si "stima" la matrice di correlazione corrispondente che viene poi sottoposta al test di chi-quadro. Inoltre si possono applicare anche ulteriori indici statistici di adattamento come l'RMSEA e indici simili (indici di *fitting* che si vedranno più avanti quando si parlerà di Analisi fattoriale confermativa).

Un'altra tecnica è quella dell'*analisi parallela* proposta da Horn (1965). La procedura è descrivibile in questo modo: si genera un certo numero di campioni (stesso numero di variabili, stesso numero di casi) con dati casuali su cui si calcolano gli autovalori. I campioni possono essere generati in modo completamente casuale, in modo casuale ma con dati simili a quelli del nostro campione oppure con permutazioni dei nostri dati. In quest'ultimo caso, i "nostri" dati servono per generare campioni in cui i singoli valori possono comparire in posizione diversa (per osservazione e per variabile). In tutti i casi avremo una distribuzione campionaria degli autovalori dei campioni casuali. Di tutti gli autovalori così calcolati (in genere si generano dai 500 ai 1000 campioni) viene stimata la media suddivisa per posizione. Confrontando il primo autovalore del nostro campione con quello generato casualmente, si terrà il corrispondente fattore se il nostro autovalore è superiore di quello casuale.

Considerando i dati della seguente tabella, su 4 variabili ipotetiche:

	Attuali	Casuali	
1	2,776	1,756	si tiene
2	1,085	1,082	si tiene
3	0,141	0,797	si scarta (anche le seguenti)
4	0,098	0,364	
	4,00	4,00	

Il primo autovalore del nostro campione è maggiore di quello generato casualmente (2,776 vs 1,756, quindi "spiega" di più della media del primo autovalore dei campioni casuali generati) e lo teniamo. Analogamente per il secondo. Invece il terzo autovalore è inferiore a quello casuale (0,141 vs 0,797) e quindi il nostro fattore (oltre ad essere inferiore a 1) spiegherebbe meno di quanto spiegherebbe il caso.

Applicando questo metodo, dovremmo estrarre 2 fattori.

Un'altra tecnica è quella della *Struttura molto semplice* (*Very Simple Structure*, VSS) di Revelle e Rocklin (1979). Questa procedura parte dall'ipotesi che la maggior parte delle volte, l'AFE si utilizza per cercare di ottenere una struttura semplice, ovvero il fatto che un item abbia saturazione elevate in un fattore e il più basse possibili in tutti gli altri. VSS (disponibile in R) calcola una soluzione fattoriale con un numero di fattori predefinito, quindi ricostruisce la matrice di correlazione utilizzando il principio della soluzione semplice e la confronta con quella originale, calcolando un indice di complessità. Ripete l'operazione con un fattore in meno e così via fino a un solo fattore. A questo punto si sceglie il numero di fattori con indice di complessità minore.

2.5 Metodi di estrazione

Una volta stabilito il criterio per il numero dei fattori, il passo successivo consiste nell'ottenere una prima soluzione fattoriale, chiamata matrice fattoriale non ruotata (*unrotated factor matrix*).

Esistono diversi metodi per ottenere l'estrazione dei fattori, ognuno dei quali presenta delle caratteristiche particolari, dei vantaggi e degli svantaggi. La scelta del metodo è collegata ai motivi specifici della ricerca, alla possibilità di utilizzare dei metodi automatici di calcolo, al numero delle variabili che compongono la ricerca, alle caratteristiche della matrice di correlazione...

Elenchiamo e descriviamo brevemente i principali metodi disponibili su calcolatore, dedicando un breve spazio ad alcuni dei metodi di importanza storica, anche se non più utilizzati.

a) metodi di estrazione dei fattori non disponibili su calcolatore

Esistono molti metodi sviluppati prima dell'introduzione dell'uso dei calcolatori, che oggi non sono affatto utilizzati. La maggior parte di questi sono stati sviluppati da Thurstone (1931,

1945) o da Burt (1937) nella prima metà del XX secolo. La tendenza generale di questi metodi consiste nel tentativo di semplificare al massimo i calcoli a scapito della semplicità logica. Una tendenza del tutto opposta a quella ricercata dai metodi disponibili sui calcolatori.

Metodo diagonale (*diagonal method*)

Questo metodo ha il privilegio di essere molto semplice e veloce, ma ha il difetto di necessitare, per i calcoli, dei valori della comunanza. Si può utilizzare quando si conosce, a priori, il valore delle comunalità. Ciò però è un evento molto raro e, proprio per questo, il metodo diagonale è usato solo come momento intermedio nell'ambito di altri metodi di estrazione dei fattori. È un metodo che si può teoricamente utilizzare con ogni tipo di matrice di correlazione (singolare o non-singolare), ma è molto sensibile ai valori scelti per la comunanza che deve perciò essere stimata accuratamente (Fruchter, 1954; Metelli, s.d.; Lis, 1977).

Metodo centroide (*centroid method*)

Il metodo centroide era tra i metodi più usati un tempo; infatti è abbastanza semplice nei calcoli, anche se non permette di giungere ad un'unica soluzione fattoriale. È necessario utilizzare una stima della comunalità che andrà a sostituire il valore lungo la diagonale della matrice di correlazione. Solitamente si usa, per questo metodo, la correlazione più alta (in valore assoluto) di ogni colonna (Fruchter, 1954; Metelli, s.d.; Lis, 1977).

Metodo multigruppo (*multiple-group method*)

Questo metodo è una variazione di quello centroide. La differenza principale è che, anziché estrarre un fattore alla volta, vengono estratti più fattori assieme. È consigliabile inserire lungo la diagonale una buona stima della comunanza. Inoltre è necessario raggruppare le variabili in gruppi di almeno 3 o 4 in base ad un'ipotesi teorica a priori o dopo aver effettuato qualche analisi dei cluster fra le variabili. Si aggiunga che le variabili incluse in un gruppo devono essere tra loro linearmente dipendenti. Con questo metodo si ottengono fattori obliqui (cfr. il paragrafo seguente) che si possono, eventualmente, ortogonalizzare (Fruchter, 1954; Metelli, s.d.; Harman, 1967).

b) metodi di estrazione dei fattori disponibili al calcolatore

Ogni software usa diversi metodi di estrazione e li può chiamare in modo leggermente diverso.

Componenti principali (*principal component*) e Assi principali (*principal axis*)

La proprietà basilare di questi metodi di estrazione dei fattori è che ogni fattore estratto cerca di essere il più possibile esplicativo rispetto ai dati di partenza. Al di là dei calcoli matematici ciò significa che il primo fattore estratto avrà la massima importanza, in quanto "spiega" una percentuale maggiore di variabilità dei dati, rispetto agli altri fattori. Non significa però che avrà le saturazioni maggiori (anche se è spesso così). Il secondo fattore estratto spiega la massima varianza possibile di quanta ne è rimasta dopo il primo fattore... e cosi via fino all'ultimo fattore estratto.

Altre caratteristiche del metodo sono: vengono estratti pochi fattori che sono tra loro indipendenti (ortogonali, cioè non correlati); la soluzione fattoriale non ruotata è unica. Accanto a questi aspetti positivi vi è il fatto che richiede calcoli matematici lunghi, laboriosi e ripetitivi (Hotelling, 1935). Ma questo aspetto negativo è stato ampiamente superato dall'utilizzo dell'elaboratore elettronico.

Quando lungo la diagonale principale della matrice delle correlazioni si pone, come stima della comunalità, il valore 1, il metodo viene designato "componenti principali". In questo caso dovrebbero essere teoricamente estratti tanti fattori quanti sono le variabili implicate nell'analisi. Purtroppo per diversi problemi di calcolo, non tutti questi fattori sarebbero significativi da un punto di vista psicologico. I fattori estratti, perciò, sono in numero inferiore e il metodo prende il nome di "componenti troncate". Va chiarito però che l'analisi delle componenti principali non è propriamente un'analisi fattoriale.

Quando invece, lungo la diagonale della matrice delle correlazioni si pone una stima della comunanza diversa dall'unità (il valore 1), allora il metodo prende il nome di "assi principali" o "fattori principali". La maggior parte dei programmi per elaboratore, in questo caso, usa il metodo iterativo per la stima delle comunanze.

Il metodo degli assi principali, secondo Thurstone (1947, p. 509) permette di ottenere la soluzione fattoriale migliore tra tutte quelle ottenibili con altri metodi, a parità di fattori estratti, proprio perché ogni fattore rende conto della massima percentuale possibile di varianza.

Se all'epoca di Thurstone (e fino agli anni '60) questo metodo era poco usato (Reuchlin, 1964), grazie alle possibilità offerte dai computer (e soprattutto di alcuni software) è diventato molto sfruttato anche a scapito di altri metodi altrettanto validi.

Sono entrambi disponibili in SPSS.

Analisi fattoriale canonica (Rao's canonical factoring)

Questo metodo di estrazione dei fattori stato inizialmente è sviluppato da Rao (1955) a partire dai procedimenti di calcolo della correlazione canonica, quindi ampliato da Narris (1962). Lo scopo è quello di trovare una soluzione fattoriale in cui la correlazione fra i fattori e le variabili sia la massima possibile. È quanto fa anche il metodo della massima verosimiglianza, che ha infatti sostituito, nella pratica statistica, quello di Rao.

Massima verosimiglianza (maximum likelihood factor)

La teoria multifattoriale di Thurstone, presuppone che si lavori sulla matrice delle correlazioni fra le variabili calcolate sull'intera popolazione mentre invece si utilizza una matrice che è stata calcolata su un campione. Questo metodo, al contrario, è ben consapevole di utilizzare misure effettuate su un campione e cerca perciò di calcolare una stima delle correlazioni sulla popolazione. A tal fine, utilizza un procedimento tipicamente matematico, chiamato "massima verosimiglianza". Associato a questo specifico metodo, è stato sviluppato un test di significatività sul numero di fattori da estrarre, basato su un'approssimazione alla funzione di chi-quadro e denominato "indice di adeguamento al campione". Per il suo utilizzo pratico basterà dire che se il valore di questo indice è significativamente elevato, è necessario un numero maggiore di fattori di quanti ne sono stati estratti.

Disponibile in SPSS.

Fattorizzazione image (*Image factoring*)

L'analisi fattoriale per immagine è stata proposta da Kaiser (1963) ampliando alcuni suggerimenti di Guttman. Per l'identificazione della varianza comune, questo metodo fa uso della "teoria dell'immagine", che considera non la matrice delle correlazioni fra le variabili, ma una matrice che contiene le "proiezioni" di ogni variabile su tutte le altre. Questa matrice chiamata "matrice immagine" perché contiene l'immagine che una variabile proietta sulle altre. Secondo Guttman la matrice immagine è quella che si avvicina maggiormente alla matrice delle correlazioni fra variabili una volta che sia stato eliminato il peso dei fattori specifici.

Il numero dei fattori teoricamente estraibili è generalmente pari alla metà delle variabili, anche se equivale ad ottenere un numero di fattori maggiore di quelli che il ricercatore si aspetta. È possibile che fra questi fattori ne compaiano alcuni banali e non significativi ed è anche possibile che i fattori banali compaiano solo dopo una rotazione ortogonale (Kaiser, 1963).

Disponibile in SPSS.

Fattorizzazione alfa (*Alpha factoring*)

Questo metodo cerca di ottenere fattori che abbiano il valore massimo di "Generalizability". Tale valore viene misurato tramite il coefficiente di fedeltà di Kuder-Richardson o con il coefficiente alpha di Cronbach (Kaiser & Caffry, 1965).

Disponibile in SPSS.

Kaiser's second generation little jiffy

Nel tentativo di migliorare l'analisi fattoriale per immagine, Kaiser (1970) ha sviluppato quest'altro metodo d'analisi che sostanzialmente è costituito dal precedente, a cui viene applicata una rotazione particolare chiamata ortobliqua (Kaiser & Rice, 1974).

2.6 Rotazione dei fattori

Una prima, importante, precisazione è che bisogna ricordare che il procedimento di rotazione si può fare solo con almeno due fattori.

Una volta ottenuta la prima soluzione fattoriale (matrice fattoriale non ruotata), si deve ora passare alla fase di rotazione dei fattori. Rotazioni si chiamano le operazioni che i calcolatori effettuano alla ricerca di soluzioni fattoriali alternative a quella individuata, purché soddisfino, sia matematicamente sia logicamente, al criterio della struttura semplice.

Al tempo in cui i calcoli venivano effettuati a mano, era più comodo cercare di raggiungere la struttura semplice utilizzando un procedimento di rotazione di assi sulle rappresentazioni grafiche delle variabili e dei fattori. Proprio da questo procedimento deriva l'attuale nome di rotazione dei fattori.

Per meglio chiarire il meccanismo della rotazione della matrice fattoriale, anche noi useremo il metodo grafico su una matrice fattoriale a due fattori ottenuta da una matrice di correlazione di 5 variabili (Tabella 2.1a). Possiamo rappresentare graficamente le saturazioni delle 5 variabili considerando i due fattori come gli assi di un piano cartesiano. Per ogni variabile avremo una coppia di valori (le saturazioni nel primo e nel secondo fattore), che può essere perciò rappresentata con un punto (Figura 2.6a).

(a) Non ruotata

	F1	F2
1	.92	.08
2	-.68	.76
3	.64	.60
4	.72	.48
5	.40	.40

(b) Ruotata

	F1r	F2r
1	.72	-.57
2	.02	.96
3	.87	.00
4	.85	-.14
5	.01	.58

Tabella 2.1 – Matrici delle saturazioni fattoriali

Proprio perché le variabili possono essere rappresentate come punti su un piano cartesiano (e le saturazioni costituiscono le coordinate di questi punti) è possibile anche spostare (o meglio ruotare) gli assi di riferimento e calcolare le coordinate di questo nuovo piano (che corrisponderanno a nuove saturazioni negli stessi fattori).

È possibile effettuare un numero infinito di rotazioni, ma solo alcune hanno importanza e significato per i nostri scopi.

Nella Figura 2.6a vediamo la rappresentazione grafica delle saturazioni della Tabella 2.1a. Se spostiamo l'asse delle x (cioè il fattore F1) in modo che passi esattamente sul punto n.3 e spostiamo coerentemente l'asse delle y (cioè il fattore F2) e quindi calcoliamo le nuove coordinate, otterremo le saturazioni di Tabella 2.1b.

Anche se non staremo a spiegare il procedimento matematico necessario per effettuare una rotazione di assi cartesiani, si può comunque empiricamente osservare (Figura 2.6b) che se il punto n.3 (e quindi la variabile n.3) cade esattamente sull'ascissa, il valore che esprime la sua ordinata è 0. Tale sarà perciò la sua saturazione nel fattore F2r (rappresentato dall'asse delle y). Analogamente, la variabile n.2 la cui rappresentazione grafica cade molto vicino all'asse delle y, avrà saturazione quasi nulla con il fattore F1r.

(a) Non ruotata **(b)** Ruotata

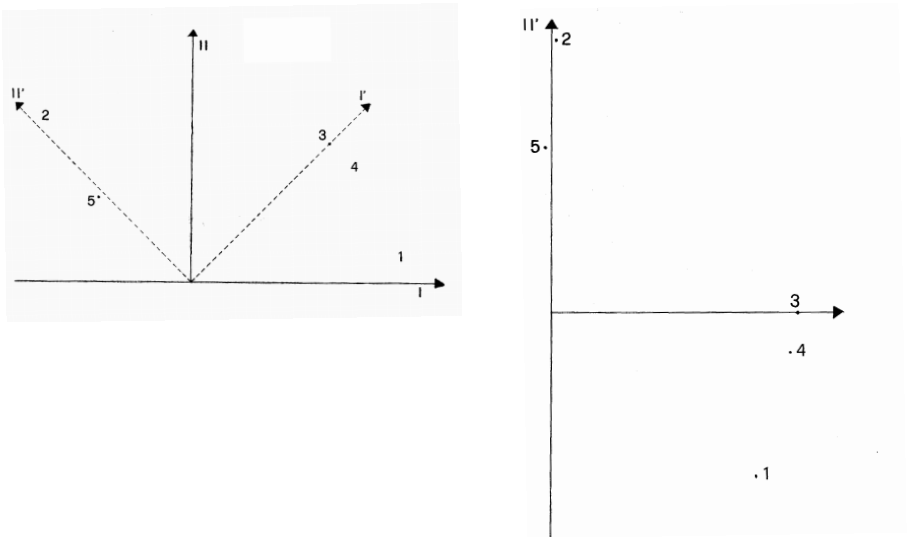


Figura 2.6 – Rappresentazione ruotata

La differenza sostanziale fra le due soluzioni fattoriali è che nel primo caso (matrice fattoriale non ruotata) tutte le variabili erano spiegate in parti molto simili da entrambi i fattori (ad eccezione della variabile 1), mentre nella seconda soluzione (matrice fattoriale ruotata), avendo reso massimo il contributo di un fattore e minimo quello dell'altro per ogni variabile, abbiamo che la terza e la quarta variabile sono spiegate principalmente dal primo fattore e la seconda e la quinta dal secondo fattore.

In questo modo ci si è avvicinati, quanto più possibile, alla cosiddetta *struttura semp1ice* proposta da Thurstone, per cui ogni variabile dovrebbe appartenere ad un unico fattore comune, cioè dovrebbe essere molto satura in un fattore e debolmente o per niente satura in tutti gli altri.

Questo è il motivo per cui si effettua la rotazione della matrice dei fattori. Bisogna sottolineare che, dopo la rotazione, i fattori non sono più necessariamente ordinati in modo decrescente di importanza.

È possibile effettuare la rotazione in molti e diversi modi, che possono però essere riassunti in due categorie: i metodi ortogonali e quelli obliqui.

I *metodi ortogonali* presuppongono che fra i fattori non esista alcun legame e cioè che siano fra loro indipendenti e che quindi la loro correlazione sia sempre zero. Graficamente ciò si esprime tramite un angolo retto e quindi la rotazione è stata chiamata "ortogonale".

Al contrario, i *metodi obliqui* presuppongono che fra i fattori esista un certo legame di dipendenza e che quindi la loro correlazione (in valore assoluto) sia compresa fra zero e uno. Graficamente ciò si esprime tramite una serie di angoli non retti che non sono facilmente disegnabili, soprattutto se le dimensioni del grafico (cioè il numero dei fattori) sono molte.

Il problema relativo alla scelta del metodo di rotazione consiste proprio nella possibilità o meno di ipotizzare la dipendenza reciproca o l'indipendenza dei fattori che si intende estrarre. Purtroppo non esiste un metodo matematico sicuro per decidere tra un metodo obliquo e uno ortogonale di rotazione. Anzi, Coleman (1964) afferma che l'ipotesi dell'ortogonalità o meno dei fattori riguarda esclusivamente l'interpretazione dei risultati.

I metodi ortogonali hanno il vantaggio di presentare fattori fra loro indipendenti, una struttura maggiormente stabile rispetto alla soluzione fattoriale che non è legata al campione, una maggiore semplicità nei calcoli e nell'interpretazione dei fattori, nonché la possibilità di rappresentazioni grafiche della matrice fattoriale che permette una migliore visione della disposizione delle variabili nello spazio fattoriale.

I metodi obliqui, proprio perché presuppongono una dipendenza di un fattore da un altro, sono più aderenti alla realtà dei dati e, quasi sempre, anche alla realtà delle variabili stesse; inoltre permettono di sottoporre la matrice di correlazione tra fattori ad un'analisi fattoriale per individuare i fattori di ordine superiore, cioè quanto hanno in comune fra loro i fattori comuni.

La decisione sulla scelta della rotazione dev'essere guidata soprattutto dalle ipotesi che è possibile formulare sulle variabili (Reuchlin, 1964, p. 69). Se si può presumere che fra i fattori che si estrarrà ci possa essere una certa dipendenza, allora converrà utilizzare una rotazione obliqua. Ad esempio, se dobbiamo analizzare i dati ottenuti da un questionario psicosomatico e ci aspettiamo che i fattori raggruppino fra loro item omogenei rispetto ai disturbi circolatori, a quelli respiratori, a quelli digestivi..., possiamo anche tranquillamente ipotizzare una dipendenza reciproca dei tre fattori in quanto le funzioni svolte da cuore, polmoni e stomaco sono sicuramente tra loro relate.

Anche gli scopi della ricerca possono influire sulla scelta. Se stiamo analizzando un test o un questionario per una sua futura standardizzazione, abbiamo bisogno di una rotazione ortogonale che ci isoli gli item molto saturi in un solo fattore. Nella ricerca sull'aggressività di Olweus, ad esempio, il ricercatore ha utilizzato un metodo ortogonale di rotazione perché desiderava verificare una sua ipotesi di lavoro e cioè che un determinato item misurasse un certo tipo di manifestazione aggressiva e non un altro tipo. Se invece avesse voluto semplicemente indagare sul più ampio fenomeno dell'aggressività, avrebbe potuto utilizzare un metodo obliquo, infatti non è possibile ipotizzare che fra diversi tipi di manifestazioni dell'aggressività non esista alcuna correlazione.

Altri autori (psicologi) ritengono ad es. che nessun fattore misurato e misurabile in ambito psicologico possa essere considerato come non correlato con gli altri, per cui suggeriscono di usare sempre una rotazione obliqua. Mentre Reuchlin (1964) consiglia di effettuare dapprima una rotazione ortogonale e di procedere eventualmente in seguito ad effettuare delle rotazioni oblique. In quest'ultimo caso, consiglia di effettuare anche un'analisi fattoriale di secondo ordine.

2.6.1 Metodi di rotazione ortogonali

Le rotazioni ortogonali sono quelle che lavorano mantenendo l'indipendenza fra i fattori, quindi fattori che non correlano fra loro e, quindi, la loro rappresentazione grafica dovrebbe essere ad angolo retto. Va ricordato che, anche se i fattori non correlano fra loro, i punteggi fattoriali possono correlare.

Con rotazioni ortogonali, le saturazioni **sono** a tutti gli effetti correlazioni fra la variabile e i fattori. In tal caso il loro quadrato corrisponde alla proporzione di varianza spiegata dal fattore per quella variabile.

Fra i vari tipi di rotazione, abbiamo i seguenti metodi:

- **Varimax**: lavora cercando di massimizzare le saturazioni alte, minimizzando quelle basse, all'interno dei singoli fattori. È consigliato se si vuole ottenere una netta separazione fra i fattori o se non si hanno criteri precisi da seguire. Come defetto, va indicato che privilegia il primo fattore.
- **Quartimax**: lavora cercando di sparpagliare la varianza entro le singole variabili, ma in realtà tende comunque a privilegiare il primo fattore. Questo dovrebbe facilitare la lettura delle variabili e non dovrebbe produrre fattori generali. Spesso però non ci riesce.
- **Equamax**: lavora equamente sulle variabili e sui fattori mantenendo costante la varianza spiegata dall'intera soluzione. Spesso non riesce ad ottenere la soluzione semplice.

2.6.2 Metodi di rotazione obliqui

Le rotazioni oblique sono quelle che lavorano senza mantenere l'indipendenza fra i fattori, quindi variabile latenti che possono correlano fra loro e, quindi, la loro rappresentazione grafica non sarà ad angolo retto.

Con le rotazioni oblique, le saturazioni finali **non sono** correlazioni ma parametri di regressione

Fra i vari tipi di rotazione, abbiamo i seguenti metodi:

- **Oblimin** diretto: il grado di associazione fra i fattori è determinato da un parametro δ generalmente posto a 0; valori positivi (fino a .8) aumentano la correlazione, negativi (fino a -.8) diminuiscono la correlazione. Tutti gli autori consigliano di non toccarlo.
- **Promax**: è sostanzialmente uguale a *oblimin*, ma pensato per essere più veloce e lavorare con grandi quantità di dati.

Entrambi i metodi si basano una prima rotazione Varimax, su cui poi si innesta un procedimento interattivo che cerca di aumentare le saturazioni alte e ridurre quelle basse avvicinando gli assi.

2.7 L'interpretazione dei fattori

Dopo la rotazione dei fattori, la matrice finale deve essere interpretata. Questa è una fase particolarmente delicata in quanto non esistono specifiche indicazioni che guidino il lavoro interpretativo. Non è stato infatti possibile elaborare un test statistico per la significatività delle saturazioni. Dipende, perciò, dalla capacità e dall'esperienza del ricercatore cogliere il significato comune delle variabili confluite in un fattore, attenendosi alla realtà delle singole variabili senza costruire interpretazioni fantastiche.

Questa fase è importante soprattutto in quanto, sia la scelta del metodo di estrazione dei fattori, sia il problema del numero dei fattori da estrarre e quello del metodo con cui effettuare la rotazione, rendono molto arbitraria l'interpretazione di un'unica soluzione fattoriale. Tutti questi aspetti e momenti in cui il ricercatore può e deve scegliere fra diversi metodi contribuiscono a creare una situazione di indeterminazione e di dubbio sui risultati che si sono ottenuti. Alcuni autori (a partire da Gorsuch, 1983) consigliano addirittura di effettuare analisi diverse per metodi di estrazione, di rotazione e numero dei fattori per poi scegliere come definitiva "quella che si adatta meglio ai dati osservati o che è più aderente agli scopi del ricercatore" (Sadocchi, 1980, p. 147).

Si ricordi che, se è stata effettuata una rotazione ortogonale si otterrà una sola matrice ruotata (*rotated factor matrix* corrispondente alla composizione fattoriale) mentre se è stata effettuata una rotazione obliqua si otterranno due matrici ruotate, la *pattern matrix* (composizione fattoriale) e la *structure matrix* (struttura fattoriale). In questo caso è la *pattern* che si usa per l'interpretazione.

2.7.1 Passaggi per l'interpretazione

Qui di seguito mostriamo i passaggi teorici necessari per interpretare una matrice fattoriale ruotata mentre nel paragrafo seguente presentiamo un esempio pratico.

1. È necessario definire un livello arbitrario per le saturazioni che ci indichi il limite oltre il quale non riteniamo le variabili sufficientemente importanti per caratterizzare quel determinato fattore. In linea di massima il valore generalmente prescelto è .40, ma è possibile usare livelli inferiori o superiori (ad esempio, da .30 a .50) a seconda che si abbia un numero ristretto o troppo ampio di variabili da interpretare.

   Ripetere i passi da 2 a 4 per ogni fattore.

2. Ordinare le saturazioni delle variabili del fattore in ordine decrescente (in valore assoluto), fermandosi al livello prescelto.
3. Scrivere accanto ad ogni saturazione la denominazione della variabile corrispondente o il testo dell'item se si tratta di un questionario.
4. Si deve ora cercare quale tratto, caratteristica, aspetto... abbiano in comune queste variabili, in modo da poter attribuire una denominazione al fattore che definisce questo tratto comune.

(a) Si tenga presente che più una variabile è satura più contribuisce alla definizione del carattere comune del fattore e, viceversa, ciò che è stato individuato come tratto comune delle variabili deve comparire in particolare e in maggior grado nelle variabili più sature.
(b) Nell'interpretazione dei fattori, si tenga conto che il segno negativo di una saturazione indica solamente un'opposizione rispetto alle saturazioni positive (di solito dipende dal processo di rotazione). Il tratto comune alle variabili dovrebbe essere pensato come un continuum che passa dalla sua massima presenza alla sua negazione o al suo opposto. Per procedere all'interpretazione conviene iniziare dalle variabili il cui segno è più frequente e considerarle come se fossero positive. Di conseguenza, le altre (siano esse di segno positivo o negativo) devono essere considerate di segno opposto.
(c) Nel caso in cui non si riesca a riscontrare nessun tratto comune alle variabili del fattore, si dovrà concludere che i1 fattore non è interpretabile e che le variabili sono state tra loro associate per un errore implicito nei calcoli e attribuibile o al campione o alla misurazione delle variabili stesse o perché, ancora, la varianza attribuibile al fattore è molto bassa. Normalmente i "primi" fattori estratti sono facilmente interpretabili mentre gli "ultimi", soprattutto se ne sono stati estratti molti o se la matrice delle correlazioni iniziale fra le variabili contiene molti valori bassi, sono spesso difficilmente interpretabili o saturi di una sola variabile e quindi fattori specifici di quella variabile. In linea di massima se i fattori non interpretabili sono molti converrà riesaminare tutti i punti nodali che contribuiscono all'indeterminazione della soluzione fattoriale per vedere se non ci sia stato qualche errore di impostazione. Quindi, se ci sono molti fattori ininterpretabili è meglio non considerare affatto i risultati dell'analisi fattoriale.

2.7.2 Esempio di interpretazione

Dimostriamo praticamente come è possibile procedere nell'interpretazione dei fattori; per questo scopo utilizzeremo la matrice di **??** ottenuta effettuando un tipo di rotazione, denominata Varimax, sulla matrice di Fig. 2.2. In questa tabella si possono vedere 6 colonne indicanti le saturazioni di ogni variabile nei 6 fattori estratti.

Tabella 2.2 – Add caption

1	Mi piace leggere argomenti che riguardano la mia religione
2	Frequento la chiesa perché mi aiuta ad incontrare altre persone
3	Non è molto importante quello in cui credo, l'importante è essere una buona persona
4	Per me è importante passare tempo in meditazione e in preghiera
5	Spesso ho avuto la forte sensazione che Dio fosse presente
6	Prego principalmente per avere aiuto e comprensione
7	Trovo difficile vivere tutta la mia vita secondo i miei principi religiosi
8	Quello che la religione in prevalenza mi offre è il confronto nei momenti di difficoltà e di dolore
9	La preghiera dà pace e felicità
10	Sebbene sia religioso, non permetto che questo influenzi la mia vita quotidiana
11	Frequento la chiesa soprattutto per passare il tempo con altre persone
12	Il mio approccio alla vita è interamente basato sulla mia religione
13	Frequento la chiesa sostanzialmente perché mi piace incontrare le persone che conosco lì
14	Sebbene io creda alla mia religione, ci sono molte altre cose più importanti nella mia vita.

2.7.3 Punteggi fattoriali

Capitolo 3

Appendice

3.1 Test di Sfericità di Bartlett

$$\chi^2 = -[N - 1 - \frac{1}{6}(2p + 5)]ln|R|$$

con $(p^2 - p)/2$ gradi di libertà; dove $|R|$ è il determinante della matrice di correlazione, N è la numerosità del campione, e p indica il numero delle variabili.

3.2 Test di Haitovsky

$$\chi^2_H = [1 + \frac{(2p + 5)}{6}]ln(1 - |R|)$$

con $p(p - 1)/2$ gradi di libertà e dove p indica il numero delle variabili, N è la numerosità del campione e $|R|$ è il determinante della matrice di correlazione

3.3 Test di adeguatezza campionaria di Kaiser-Meyer-Olkin

$$KMO = \frac{\sum_i \sum_j r^2_{ij}}{\sum_i \sum_j r^2_{ij} + \sum_i \sum_j p^2_{ij}}$$

dove r_{ij} sono le correlazioni e p_{ij} sono le correlazioni parzializzate su tutte le altre.

Riferimenti bibliografici

Burt, C. (1937). Methods of factor analysis with and without successive approximation. *British Journal of Educational Psychology*, *7*, 172-195.

Cattell, R. B. (1966). The scree-test for number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Coleman, J. S. (1964). *Introduction to mathematical sociology.* New York: The Free Press.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cristante, F., Lis, A., & Sambin, M. (1982). *Statistica per psicologi.* Firenze: Giunti Barbèra.

Field, A. (2009). *Discovering statistics using Spss* (3rd ed.). Los Angeles-London-New Delhi: Sage.

Fruchter, B. (1954). *Introduction to factor analysis.* Princeton: Van Nostrand.

Gorsuch, R. L. (1974). *Factor analysis.* Philadelphia: Saunders.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265-275. doi: 10.1037//0033-2909.103.2.265

Guilford, J. P. (1956). *Psychometric methods.* New York: McGraw-Hill.

Harman, H. H. (1967). *Modern factor analysis.* Chicago: University of Chicago Press.

Harris, C. W. (1967). On factors and factors scores. *Psichometrika*, *32*, 363-379.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-185.

Hotelling, H. (1935). Simplified calculation of principal components. *Psichometrika*, *1*, 27-35.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141-151.

Kaiser, H. F. (1963). Image analysis. In C. W. Harris (Ed.), *Problems in measuring change.* Madison: University of Wisconsin Press.

Kaiser, H. F. (1970). A second generation little jiffy. *Psichometrika*, *35*, 401-415.

Kaiser, H. F., & Caffry, J. (1965). Alpha factor analysis. *Psichometrika*, *30*, 1-14.

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark iv. *Educational and Psychological Measurement*, *34*, 111-117.

Kass, R. A., & Tinsley, H. E. A. (1979). Factor analysis. *Journal of Leisure Research*, *11*, 120-138.

Kline, P. (1994). *An easy guide to factor analysis.* New York: Routledge.

Lis, A. (1977). *Appunti di analisi fattoriale 2.* (Ciclostilato dell'Istituto di Psicologia, Università di Padova)

Lis, A., Rossi, G., & Venuti, P. (1986). *L'analisi fattoriale in psicologia con applicazioni di psicologia clinica.* Padova: Cleup.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84-99.

Metelli, F. (s.d.) *Analisi fattoriale.* Firenze: Giunti.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Olweus, D. (1975). *Development of a multi-faceted aggression inventory for boys* (Tech. Rep.). Reports from The Institute of Psychoilogy, University of Bergen, Norway.

Rao, C. A. (1955). Estimation and tests of significance in factor analysis. *Psichometrika*, *20*, 93-111.

Reuchlin, M. (1964). *Methodes d'analyse factorielle a l'usage des psychologues.* Paris: Presses Universitaire de France.

Revelle, W., & Rocklin, T. (1979). Very simple structure - alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*(4), 403-414.

Runyon, R. P., & Haber, A. (1976). *Fundamentals of behavioral statistics* (3rd ed.). Reading, MA: Addison-Wesley. (Trad. it. Fondamenti di statistica per le scienze del comportamento. Amsterdam: InterEuropean Editions.)

Sadocchi, S. (1980). *Manuale di analisi statistica multivariata per le scienze sociali.* Milano: Franco Angeli.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston-New York: Pearson.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, *38*, 406-427.

Thurstone, L. L. (1945). *Multiple factor analysis.* Chicago: University of Chicago Press.