6. ANALISI FATTORIALE

6.1 Introduzione

L'analisi fattoriale (AF) si è sviluppata nell'ambito della ricerca sulla misura di capacità attitudinali all'inizio del 1900. Spearman effettua molti test su campioni numerosi per cercare di misurare l'intelligenza degli individui sulla base delle correlazioni tra i punteggi ottenuti in diverse prove. Rilevando che il punteggio in ogni test era correlato positivamente con tutti gli altri, spiega questo risultato distinguendo due ordini di fattori:

1) un **fattore comune** che spiegava le correlazioni positive fra i test;

2) una **serie di fattori specifici**, espressione delle abilità specifiche individuali, che spiegavano perché tali correlazioni non fossero perfette.

L'ipotesi alla base era che la mente umana contiene un'abilità generale, l'intelligenza, che influisce sul rendimento in tutti i test, e un vasto insieme di abilità particolari. L'intelligenza non è direttamente osservabile, ma si può fare riferimento ad un insieme di indicatori (rilevabili e misurabili) che si ritengono correlati con questo fattore comune, (indicato anche come "**variabile latente**", perché non direttamente misurabile).

Questi indicatori dovranno quindi essere correlati con la variabile latente: nel caso dell'intelligenza gli individui più intelligenti devono poter ottenere valori più elevati degli indicatori considerati (tipicamente test attitudinali che misurano l'abilità linguistica, di calcolo e di logica).

Allo stesso modo, per valutare lo stato socio-economico delle famiglie ci si potrebbe basare, per esempio, sulle informazioni circa il reddito, l'occupazione e il livello di istruzione dei suoi componenti.

In generale l'AF viene utilizzata quando le variabili osservate su un insieme di individui possono essere considerate come funzioni lineari di un gruppo ristretto di una o più variabili latenti, non direttamente osservabili.

Questi fattori sono comuni a tutte le variabili osservate (o a gruppi di esse) ed il problema dell'AF consiste nell'identificare e denominare questi fattori. In generale, quindi, l'AF è un metodo che cerca di spiegare le correlazioni fra un insieme di *h* variabili osservate attraverso un insieme di *k* variabili non osservabili.

Un elevato coefficiente di correlazione lineare tra due variabili X_i e X_j può infatti dipendere dalla loro associazione con una terza variabile, *F*. In questo caso si ricorre al calcolo del coefficiente di correlazione parziale fra le due variabili, al netto di *F*, che consente di misurare l'associazione tra X_i e X_j dopo che si

è eliminato l'effetto lineare di F su ciascuna di esse. Se questo coefficiente è prossimo a zero, allora è proprio la variabile F quella responsabile dell'elevata relazione lineare esistente fra X_i e X_j.

In alcuni casi si può supporre che esista un'unica variabile latente, ma più spesso si ipotizza la presenza di più variabili.

C'è uno stretto legame tra l'AF e l'analisi in CP, che talvolta vengono confuse fra loro, ma fra i due tipi di analisi esistono delle differenze fondamentali:

- L'analisi in CP è un metodo di statistica descrittiva che ha l'obiettivo di ridurre la dimensione della matrice di dati, mentre l'AF è una tecnica basata su un modello che necessita di assunzioni sulla distribuzione congiunta delle variabili nella popolazione. Pertanto si utilizzano metodi di statistica inferenziale per misurare la bontà di adattamento, la significatività e la precisione delle stime.
- Una seconda differenza sta nel fatto che, mentre nell'analisi in CP si determinano delle opportune combinazioni lineari delle variabili osservate, nell'AF sono invece le variabili osservate a essere espresse come combinazioni lineari di fattori latenti.
- Infine l'analisi in CP cerca di spiegare con poche componenti una grande parte della varianza delle variabili osservate, mentre nell'AF si prendono in considerazione le covarianze, o le correlazioni, tra le variabili. L'ipotesi sottostante l'AF è che se le variabili osservate sono funzioni lineari di variabili comuni, dovranno risultare più o meno correlate fra di loro, per cui l'AF può essere considerata come la **ricerca delle origini delle correlazioni** fra le variabili. I fattori comuni determinano le **covarianze** fra le variabili, mentre i fattori specifici contribuiscono solo alla **varianza** della variabile a cui si riferiscono.

Questo metodo di analisi non è accettato da tutti gli studiosi in quanto può accadere di avere dei data set per i quali il modello fattoriale non fornisce un adeguato adattamento, nel senso che non si riesce ad individuare alcun fattore comune. Inoltre l'AF si basa spesso su scelte soggettive, a causa della difficoltà di individuare il numero corretto di variabili latenti e di darne una chiara interpretazione sulla base dei risultati ottenuti.

Cominciando a considerare il caso in cui le h variabili osservate dipendano da un unico fattore o variabile latente F, l'AF si basa sul modello lineare

$$X_i = \alpha_i F + U_i \qquad (i = 1, 2, \ldots, h) \qquad 6.1.1$$

che solo apparentemente somiglia al modello di regressione, dato che il fattore non è osservabile così che tutto quello che si trova a destra dell'uguaglianza 6.1.1 è incognito. Scopo dell'AF è spiegare l'interdipendenza esistente all'interno di un insieme numeroso di variabili X_i tramite un fattore F non osservabile sottostante.

Con il modello 6.1.1 si suppone che le intensità delle h variabili X_i rilevate su n unità dipendano da un unico fattore comune F e da un fattore specifico per ogni individuo, oltre che da una componente casuale.

Il parametro α_i indica il **peso del fattore** (o **peso fattoriale** o **loading factor**) per la variabile X_i, mentre U_i è una variabile le cui intensità sono influenzate sia dal fattore specifico sia da fattori casuali (che in genere non si ha interesse a misurare distintamente).

Nel modello a un fattore le ipotesi sottostanti sono:

- l'indipendenza dei fattori specifici U_i da F

- Le variabili U_i hanno distribuzione normale, di media nulla e varianza σ_i^2, e sono indipendenti tra di loro, per cui anche le X_i sono condizionatamente indipendenti, dato F.

- dato che F è una variabile non osservata, si assume (per semplicità di calcolo) che abbia media zero e varianza unitaria, due caratteristiche di comodo che non influiscono sulla forma dell'equazione di regressione.

Dato che la U_i è incorrelata con F la varianza della generica X_i corrisponde alla somma

$$V(X_i) = \alpha^2_i + \psi_i,$$

dove α^2_i è la **varianza spiegata** dal fattore, detta **comunalità** o **comunanza**, e ψ_i è la **varianza residua**, detta **varianza specifica**, **specificità** o **unicità**.

Si dimostra inoltre che $\text{Cov}(X_i,F) = \alpha_i$.

L'obiettivo dell'AF è la stima del valore dei coefficienti α_i, delle varianze delle U_i e dei valori di F.

Nel caso di più fattori comuni, date le h variabili X_i rilevate su n individui, si suppone che le loro intensità dipendano da q fattori comuni F_j (j = 1, 2, ..., q) e da un solo fattore unico.

Ipotizzando per semplicità che ogni fattore comune abbia media 0 e varianza 1, ogni X_i può essere scomposta nel modo seguente

$$X_i = \mu_i + \alpha_{i1} F_1 + \ldots + \alpha_{iq} F_q + U_i$$

dove α_{ij} è il peso del j-esimo fattore per la variabile X_i e corrisponde anche alla Cov(X_i, F_j), mentre U_i è una variabile le cui intensità sono influenzate dal fattore specifico e da fattori casuali.
La varianza di X_i assume la forma

$$V(X_i) = \sum_j \alpha_{ij}^2 + \psi_i$$

In questo caso le comunanze sono le componenti della varianza corrispondenti a

$$\sum_j \alpha_{ij}^2$$

e le varianze specifiche sono date sempre da ψ_i.
Come si è detto, per i fattori comuni si assumono le condizioni che abbiano media zero, varianza unitaria e siano incorrelati fra di loro, mentre si assume che i fattori specifici abbiano media zero, varianza ψ_i e siano incorrelati fra di loro e con i fattori specifici.
Se è valido il modello fattoriale, risulta che:
- la matrice di covarianza delle *h* variabili X_i è scomposta nella somma di due matrici sulle cui diagonali compaiono rispettivamente le comunanze e le specificità.
- la covarianza fra X_i e X_j è completamente spiegata dai fattori comuni,
- le covarianze fra fattori e variabili coincidono con i pesi fattoriali.

Si dimostra che il modello fattoriale è invariante per cambiamenti di scala delle osservazioni.
Il problema dell'AF sta nel fatto che la matrice dei pesi fattoriali non è identificabile, ossia che non esiste una soluzione unica alla determinazione dei pesi fattoriali. Il problema viene risolto imponendo alcuni particolari vincoli che finiscono per imporre un limite al numero di fattori che possono essere stimati.
C'è inoltre da osservare che in alcuni casi la scomposizione della varianza può portare a stime negative che ovviamente non possono essere accettate, trattandosi di varianze.

Le comunanze delle variabili osservate, definite come la somma dei quadrati dei corrispondenti pesi fattoriali, danno una indicazione del grado in cui ciascuna variabile si sovrappone ai fattori, o più tecnicamente, rappresentano la proporzione di varianza delle variabili X_i che può essere spiegata dai punteggi nei fattori comuni.

Le singole comunanze consentono di valutare la bontà del modello nei confronti di ciascuna variabile osservata, mentre la comunanza media fornisce una valutazione di insieme.

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, però, non è possibile conoscere i valori esatti delle comunanze, ma soltanto stimarli e questo comporta come conseguenza che non tutti i fattori teoricamente estraibili siano dei "veri" fattori. Inoltre il numero esatto dei fattori potrebbe essere determinato solo se le correlazioni fossero calcolate sull'intera popolazione e non fossero stimate su base campionaria. Infine i fattori che l'analisi considera significativi da un punto di vista statistico, potrebbero non essere considerati tali dal ricercatore (per esempio perché si sta conducendo un'analisi in campo psicologico, ma il fattore non abbia alcun significato sotto questo punto di vista.

L'uso dei calcolatori consente di non indicare al programma quanti sono i fattori da estrarre, ma di fornire piuttosto un criterio da seguire per determinare il loro numero.

In tutti i metodi di stima dell'analisi fattoriale vengono determinati dei valori di partenza delle comunanze e questi valori di partenza variano a seconda del metodo scelto, anche se in genere i risultati finali ottenuti con l'AF non dipendono molto da questi valori iniziali. I metodi più utilizzati sono

- il metodo dei fattori principali, che prevede siano stabilite stime iniziali per le comunanze. La stima delle comunanze è ottenuta per iterazione e viene calcolata in base ai pesi fattoriali.
- il metodo di massima verosimiglianza, che si utilizza quando si può assumere che le variabili originarie abbiano una distribuzione normale multivariata.

Per interpretare più facilmente i pesi fattoriali di solito si effettuano delle **rotazioni degli assi fattoriali** che semplificano la struttura del sistema di pesi. Si cerca di suddividere le variabili in gruppi, in modo tale che i pesi all'interno di ciascun gruppo siano elevati su un singolo fattore e bassi o trascurabili sugli altri. Le soluzioni più utilizzate rispettano l'ortogonalità dei fattori. Le più usate sono

- <u>METODO VARIMAX:</u> minimizza il numero di variabili che hanno correlazioni alte con un fattore
- <u>METODO QUARTIMAX:</u> minimizza il numero di fattori che hanno correlazioni alte con una variabile
- <u>METODO EQUIMAX:</u> è una combinazione dei due metodi precedenti

In ogni caso la percentuale di varianza complessiva dei fattori ruotati rimane inalterata, mentre si modifica la percentuale di varianza spiegata da ciascun fattore

6.2 Proc factor

La procedura SAS per effettuare un'AF è la **`proc factor`**, che prevede opzioni diverse a seconda del metodo di stima prescelto.

Usualmente il punto di partenza per un'AF si basa sull'analisi della matrice di correlazione delle variabili rilevate, dato che se le correlazioni sono basse allora l'analisi rischia di non essere utile, in quanto basse correlazioni sono un sintomo della mancanza di fattori comuni.

Nella tabella seguente è riportata, per esempio, la matrice di correlazione relativa ai risultati ottenuti da 34 atleti in alcune gare del decathlon. Le sei variabili rilevate sono:

X_1 - tempo ottenuto nella corsa dei 100 metri,

X_2 - tempo nei 400 metri,

X_3 - tempo nei 110 ad ostacoli

X_4 - risultato in metri nel salto con l'asta,

X_5 - risultato nel lancio del disco,

X_6 - risultato nel lancio del giavellotto.

Tabella 6.2.1
matrice di correlazione di alcuni risultati ottenuti nel decathlon

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1,000	0,698	0,751	-0,627	-0,353	-0,344
X_2	0,698	1,000	0,655	-0,521	-0,154	-0,150
X_3	0,751	0,655	1,000	-0,709	-0,403	-0,350
X_4	-0,627	-0,521	-0,709	1,000	0,620	0,557
X_5	-0,353	-0,154	-0,403	0,620	1,000	0,618
X_6	-0,344	-0,150	-0,350	0,557	0,618	1,000

Dalla tabella si nota che i risultati relativi alle corse sono abbastanza correlati fra di loro in modo positivo e che lo stesso vale anche per i risultati del secondo gruppo.

Le correlazioni fra i due gruppi sono negative, ma i risultati migliori nelle discipline del primo gruppo corrispondono a tempi più bassi, mentre per le altre variabili è vero il contrario, cosicché in realtà i risultati sono tutti correlati positivamente.

Si nota anche che le correlazioni più alte fra le variabili del primo gruppo e le variabili del secondo sono quelle tra le corse ed il salto con l'asta, dato che in quest'ultima disciplina probabilmente la fase della rincorsa è più importante che nei due lanci.

Per capire se un'analisi fattoriale potrà produrre dei fattori rilevanti si può calcolare l'indice di Kaiser-Meyer-Olkin (KMO) che confronta la grandezza delle correlazioni totali rispetto alle correlazioni parziali (dove la correlazione parziale fra due variabili si calcola eliminando l'influenza delle restanti variabili). Se le correlazioni parziali sono elevate rispetto alle correlazioni totali significa che i fattori specifici dominano sui fattori comuni e l'AF non potrà quindi fornire buoni risultati. Kaiser suggerisce di utilizzare l'indice in questo modo: un valore > 0.90 è eccellente, fra 0.80 e 0.90 è buono, fra 0.70 e 0.80 è accettabile, fra 0.60 e 0.70 è mediocre.

Il calcolo di questo indice può essere richiesto al SAS utilizzando l'opzione `msa` all'interno della **`proc factor`**.

Per far eseguire l'AF al SAS a partire dai dati contenuti nella tabella 6.2.1 è necessario dichiarare che il data set in questione è una matrice di correlazione, così come mostrato nel listato successivo

```
data a(/* N=34 */ TYPE=CORR);

   input _NAME_ $ X1-X6;
   if _N_=1 then _TYPE_='N    ';
        else _TYPE_='CORR';
   label
 X1='100 metri' X2='400 metri' X3='110 ostacoli' X4='salto con asta' X5='lancio
disco' X6='giavellotto'
      ;
      cards;
 N    34     34      34       34     34      34
 X1 1.000   0.698 0.751 -0.627      -0.353      -0.344
 X2 0.698   1.000 0.655 -0.521      -0.154      -0.150
 X3 0.751   0.655 1.000 -0.709      -0.403      -0.350
 X4 -0.627  -0.521      -0.709      1.000 0.620 0.557
 X5 -0.353  -0.154      -0.403      0.620 1.000 0.618
 X6 -0.344  -0.150      -0.350      0.557 0.618 1.000
;
proc print;
proc factor priors=smc msa;
run;
```

L'opzione `(TYPE=CORR)` nel data step, specifica che i dati in input corrispondono ai coefficienti di correlazione di una matrice, mentre il commento in verde serve per ricordare la numerosità delle unità statistiche utilizzate, `/* N=34 */`. Inoltre le istruzioni

```
if _N_=1 then _TYPE_='N    ';
         else _TYPE_='CORR';
```

servono per indicare al SAS che la prima riga dopo cards, indicata da `_N_=1`, contiene il numero delle osservazioni su cui è stato calcolato ciascun coefficiente di correlazione, mentre le righe successive si riferiscono alla matrice di correlazione vera e propria.
In questo caso l'indice KMO fornisce un risultato pari a 0.81, come risulta dalla tabella successiva, per cui l'AF dovrebbe fornire risultati soddisfacenti.

Tabella 6.2.2

Misura di adeguatezza campionaria di Kaiser: MSA totale = 0.81458371					
X1	**X2**	**X3**	**X4**	**X5**	**X6**
0.82787206	0.79788019	0.83718750	0.83096032	0.76632091	0.79936697

Il SAS ottiene due soli autovalori positivi pari rispettivamente a 3.17 e 0.79 circa, come risulta dai valori riportati nella tabella successiva. Sono proprio questi i fattori che vengono considerati dal SAS in questo caso, sulla base del criterio "proportion" che è quello utilizzato per default.

Tabella 6.2.3
Autovalori della matrice 6.2.1 (metodo dei fattori principali)

	Autovalore	**Differenza**	**Proporzione**	**Cumulativa**
1	3.17125792	2.38437376	0.8967	0.8967
2	0.78688417	0.82988711	0.2225	1.1192
3	-.04300294	0.03603046	-0.0122	1.1070
4	-.07903340	0.06831670	-0.0223	1.0847
5	-.14735010	0.00475386	-0.0417	1.0430
6	-.15210396		-0.0430	1.0000

Con questo criterio il numero di fattori corrisponde a quelli necessari per ottenere per la prima volta una proporzione cumulata pari a 1, ma il valore del criterio può essere modificato a piacere. Per esempio, le istruzioni

```
proc factor priors=smc msa proportion=0.80;
```

indicano al SAS di considerare solo i fattori per i quali si ottiene una proporzione cumulata supera per la prima volta il valore 0.8, per cui in questo caso sarebbe considerato solo il primo fattore, la cui proporzione cumulata è pari a 0.8967.

I coefficienti dei due fattori considerati dal SAS sono riportati nella tabella successiva, dai quali risulta che il primo fattore è fortemente influenzato dai valori di X_1 e di X_3 in modo diretto e dai valori di X_4 in modo inverso, ma risulta abbastanza correlato anche con le restanti variabili. Il secondo fattore, invece, è più fortemente correlato con X_2, X_5 e X_6, ma i pesi non risultano molto diversi, per cui sembra opportuno procedere ad una rotazione degli assi.

Tabella 6.2.4
Pattern fattoriale

		Factor1	**Factor2**
X1	**100 metri**	0.80466	0.26812
X2	**400 metri**	0.66269	0.44194
X3	**110 ostacoli**	0.83301	0.20637
X4	**salto con asta**	-0.85128	0.15389
X5	**lancio disco**	-0.59897	0.49087
X6	**giavellotto**	-0.55432	0.46094

La varianza spiegata da ciascun fattore risulta rispettivamente pari a circa 3.17 e 0.79, per un totale di 3.96, come si nota dai risultati elencati nella tabella 6.2.5, mentre le stime finali delle comunanze sono riportate nella tabella 6.2.6

Tabella 6.2.5
Varianza spiegata dai fattori

Factor1	Factor2
3.1712579	0.7868842

Tabella 6.2.6
Stime di comunanza finali: Totale = 3.958142

X1	X2	X3	X4	X5	X6
0.71936268	0.63447256	0.73649311	0.74836876	0.59971458	0.51973040

La matrice dei pesi è stata quindi ruotata utilizzando l'opzione varimax, esplicitata con la seguente opzione prevista dal SAS

```
proc factor priors=smc rotate=varimax;
```

In questo modo si sono ottenuti i nuovi risultati riportati nelle tabelle successive

Tabella 6.2.7
Pattern fattoriale ruotato

		Factor1	Factor2
X1	**100 metri**	0.79805	-0.28720
X2	**400 metri**	0.79406	-0.06281
X3	**110 ostacoli**	0.78212	-0.35325
X4	**salto con asta**	-0.57358	0.64759
X5	**lancio disco**	-0.16687	0.75622
X6	**giavellotto**	-0.15030	0.70508

Tabella 6.2.8
Varianza spiegata dai fattori

Factor1	Factor2
2.2585446	1.6995975

Tabella 6.2.9

Stime di comunanza finali: Totale = 3.958142

X1	X2	X3	X4	X5	X6
0.71936268	0.63447256	0.73649311	0.74836876	0.59971458	0.51973040

I nuovi pesi rendono più chiara l'identificazione dei fattori. Il primo fattore infatti risulta correlato con le tre variabili relative alle corse e più debolmente con i risultati nel salto con l'asta. Il secondo fattore è correlato esclusivamente con il salto con l'asta e con i lanci, cosicché si può concludere che i due fattori in senso lato corrispondono alla potenza delle gambe e a quella delle braccia.

La varianza spiegata da questi nuovi fattori è rispettivamente pari a 2.26 e 1.70 circa, ottenendo sempre una somma pari a 3.96 circa.

Se sui dati della tabella 6.2.1 si vuole effettuare la stima con il metodo di massima verosimiglianza, il listato SAS assumerà la forma seguente

```
proc factor method=ml;
```

Anche in questo caso si ottengono due soli autovalori positivi, come risulta dalla tabella seguente

Tabella 6.2.10

Autovalori della matrice 6.2.1 (metodo di massima verosimiglianza)

	Autovalore	Differenza	Proporzione	Cumulativa
1	8.61477301	6.89579374	0.9271	0.9271
2	1.71897927	1.83417783	0.1850	1.1121
3	-.11519855	0.10177298	-0.0124	1.0997
4	-.21697153	0.08544682	-0.0234	1.0764
5	-.30241835	0.10495202	-0.0325	1.0438
6	-.40737037		-0.0438	1.0000

In questo caso il SAS effettua automaticamente anche un test di significatività sul numero di fattori, i cui risultati sono riportati nella tabella 6.2.11

Tabella 6.2.11
Test di significatività basati su 35 osservazioni

Test	DF	Chi-quadrato	Pr > ChiQuadr
H0: Nessun fattore comune	15	108.0530	<.0001
HA: Almeno un fattore comune			
H0: 2 fattori sono sufficienti	4	1.1134	0.8921
HA: altri fattori sono necessari			

Tralasciando alcune informazioni fornite dal SAS circa la matrice di correlazione ridotta, i pesi fattoriali risultano quelli riportati nella tabella seguente

Tabella 6.2.12
Pattern fattoriale

		Factor1	Factor2
X1	**100 metri**	0.82107	0.25465
X2	**400 metri**	0.69593	0.44914
X3	**110 ostacoli**	0.85151	0.18020
X4	**salto con asta**	-0.85247	0.18407
X5	**lancio disco**	-0.60189	0.58341
X6	**giavellotto**	-0.54301	0.49870

mentre la varianza spiegata è pari a circa 3.27 per il primo fattore e a 0.92 per il secondo (considerando i fattori non pesati).
Sempre relativamente al caso non pesato, le stime finali della comunanza sono

Tabella 6.2.13

Stime di comunanza finali: Totale = 4.189397

Variabile	Comunanza
X1	0.73900320
X2	0.68604671
X3	0.75755081
X4	0.76059021
X5	0.70264706
X6	0.54355854

Anche in questo caso si è effettuata la rotazione Varimax ottenendo i risultati seguenti

Tabella 6.2.14

Pattern fattoriale ruotato

		Factor1	Factor2
X1	**100 metri**	0.81592	-0.27070
X2	**400 metri**	0.82733	-0.03972
X3	**110 ostacoli**	0.79735	-0.34896
X4	**salto con asta**	-0.58615	0.64577
X5	**lancio disco**	-0.14999	0.82471
X6	**giavellotto**	-0.15140	0.72155

Tabella 6.2.15

Varianza spiegata

Fattore	Non pesato
Factor1	2.37496049
Factor2	1.81443605

Tabella 6.2.16

Stime di comunanza finali: Totale = 4.189397

Variabile	**Comunanza**
X1	0.73900320
X2	0.68604671
X3	0.75755081
X4	0.76059021
X5	0.70264706
X6	0.54355854

La varianza spiegata da questi nuovi fattori risulta rispettivamente pari a 2.37 ed a 1.81 circa e, come si vede, i risultati ottenuti con i due diversi metodi sono molto simili fra di loro.