Università degli studi di Ferrara
Dipartimento di Matematica
A.A. 2019/2020 – I semestre

STATISTICA MULTIVARIATA

SSD MAT/06

LEZIONE 14 – Analisi fattoriale confermativa

Docente: Valentina MINI

valentina.mini@unife.it

RICEVIMENTO: LUNEDI POMERIGGIO, su appuntamento previa mail

Indice dei contenuti

1. Comprendere l'AFC a partire dall'AFE
2. Procedimento analitico: dall'AFE alla AFC (note di riepilogo)
3. Esercizi suggeriti in R

1. Conprendere l'AFC a partire dall'AFE

AFE e AFC: introduzione

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

AFE e AFC: introduzione

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

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

AFE e AFC: due procedure "inverse"

AFE e AFC

La differenza si basa sulla diversa procedura

Analisi fattoriale Esplorativa:

Ridurre un insieme di variabili osservate ad un insieme inferiore di variabili non osservate o latenti (fattori, componenti, dimensioni)

Trasformare le variabili osservate in una **struttura più semplice** che contenga però le stesse informazioni dell'originale

Procedura induttiva

Analisi Fattoriale Confermativa (detta anche Restricted Factor Analysis (Hattie&Frases, 1988) o Structural Factor Analysis (McArdle,1996) o –più semplicemente- Measurement Model (Hoyle, 1991)

Basata su una **procedura DEDUTTIVA** di verifica delle ipotesi, circa l'esistenza di fonti non osservate di variabilità in grado di spiegare la varianza comune fra un set di variabili osservate

Dall'Analisi Fattoriale Esplorativa a quella Confermativa

Matrice fattoriale ruotata
Assi principali

	Fattore 1	2
x10	-.970	
x8	.946	
x3	.926	
x4	-.922	
x5	.843	
x6		.929
x7		.895
x9		-.891
x1		.890
x2		.872

- Questa è una soluzione fattoriale ottenuta nelle analisi fattoriali (esplorative) precedenti, da cui abbiamo eliminato i valori inferiori a .30
- L'analisi fattoriale confermativa si chiede se, eliminando le influenze "molto basse" del fattore sugli item, riusciamo a spiegare **abbastanza** varianza da "confermare" il modello teorico

Dall'Analisi Fattoriale Esplorativa a quella Confermativa

Es. (questionario di 18 domande, estrapolazione)

Item di un test, OCQ, per la valutazione degli aspetti del disturbo ossessivo compulsivo

1.	Ho paura di usare i bagni pubblici, anche se ben puliti perché sono troppo preoccupato dei germi	1	2	3	4	5
2.	Mi preoccupo eccessivamente dei germi e delle malattie	1	2	3	4	5
3.	Uno dei miei maggiori problemi è quello di essere eccessivamente preoccupato per la pulizia	1	2	3	4	5
4.	Ogni giorno impiego troppo tempo a prepararmi per uscire di casa perché devo fare ogni cosa in modo esatto	1	2	3	4	5
5.	Conto quasi sempre quando compio le azioni abituali	1	2	3	4	5
6.	Spesso mi sento costretto a memorizzare cose inutili (es. numeri di targa, scritte sulle etichette, ecc.)	1	2	3	4	5
7.	Spesso devo controllare più volte alcune cose come interruttori, rubinetti, elettrodomestici o porte	1	2	3	4	5
8.	Controllo ripetutamente che i fornelli siano spenti, anche se tento di resistere all'impulso di farlo	1	2	3	4	5
9.	Controllo ripetutamente che le porte e le finestre siano chiuse, anche se tento di resistere all'impulso di farlo	1	2	3	4	5
10.	Mi è molto difficile prendere anche le decisioni banali	1	2	3	4	5
11.	Mi rende molto ansioso dover prendere una decisione anche di minor importanza	1	2	3	4	5
12.	Solitamente, dopo aver deciso qualcosa, rimugino sulla mia decisione per molto tempo	1	2	3	4	5

Dall'Analisi Fattoriale Esplorativa a quella Confermativa

I dati raccolti sui 150 pazienti servono per eseguire lo studio esplorativo, quello in cui valutiamo la dimensionalità della scala. In questo caso eseguiamo un'AFE

Supponiamo che, dopo aver valutato i risultati delle analisi di dimensionalità e aver confrontato fra loro strutture fattoriali alternative, siamo giunti alla conclusione che la soluzione migliore è quella riportata

Si noti i nomi dei fattori

Dall'Analisi Fattoriale Esplorativa a quella Confermativa

Risultati di un'analisi fattoriale esplorativa sugli item del test OCQ. I valori in grassetto e sottolineato rappresentano le saturazioni principali

Item	*Checking*	*Contamination*	*Hoarding*	*Indecisiveness*	*Obsessions*	*Compulsions*
ocq01	-,08	**,64**	,04	,00	-,03	,06
ocq02	,00	**,73**	,04	-,02	,06	-,03
ocq03	,06	**,71**	-,04	-,05	-,04	,06
ocq04	-,01	,09	-,11	,04	-,13	**,67**
ocq05	-,03	,06	,00	-,05	,15	**,54**
ocq06	-,01	-,07	,03	-,03	,24	**,55**
ocq07	**,86**	-,02	-,01	-,01	,04	-,08
ocq08	**,56**	,19	-,10	,20	-,06	-,01
ocq09	**,84**	-,09	,08	-,09	,02	,05
ocq10	,02	-,12	,06	**,72**	-,11	,17
ocq11	,07	,00	,09	**,51**	-,02	,18
ocq12	-,05	,04	-,08	**,72**	,21	-,24
ocq13	-,03	-,05	**,93**	-,01	-,06	-,06
ocq14	-,02	,13	**,48**	,24	,03	,00
ocq15	,06	,05	**,52**	-,05	,05	-,04
ocq16	-,06	-,04	-,10	,16	**,63**	-,03
ocq17	,05	,12	,08	-,06	**,52**	,01
ocq18	,06	-,06	,03	-,03	**,61**	,12

Dall'Analisi Fattoriale Esplorativa a quella Confermativa

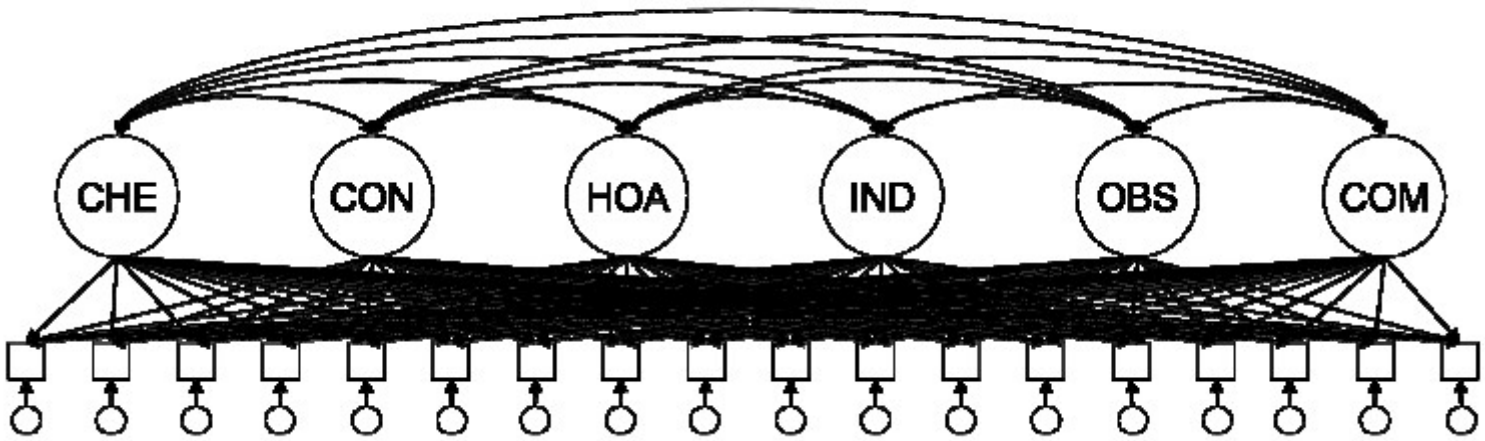
Il problema, a questo punto, è replicarlo su un

campione indipendente di soggetti.

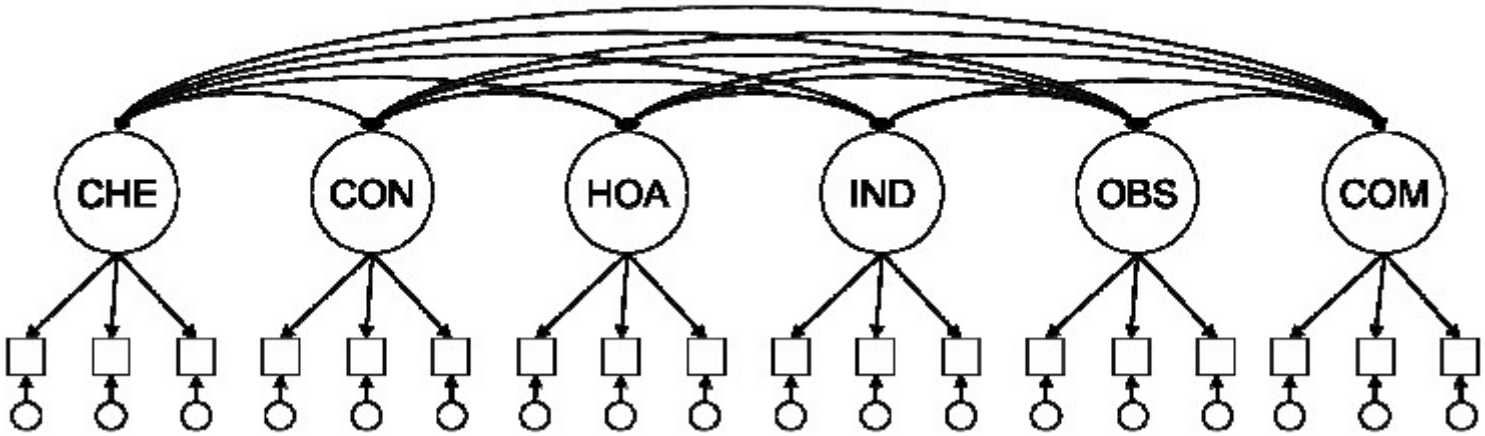
Che analisi facciamo? Certamente potremmo realizzare delle nuove analisi esplorative, ma in realtà noi sappiamo già quale *dovrebbe* essere la struttura fattoriale del test, perché quando abbiamo sviluppato il test sapevamo già che per definire il dominio di contenuto del disturbo ossessivo compulsivo avremmo dovuto utilizzare sei facet e risultati dello studio esplorativo hanno supportato questa ipotesi. Nel nuovo studio **confermativo**, quindi, abbiamo già una teoria e dei dati a supporto

Struttura grafica dei due modelli a confronto

(a) Modello di analisi fattoriale esplorativa

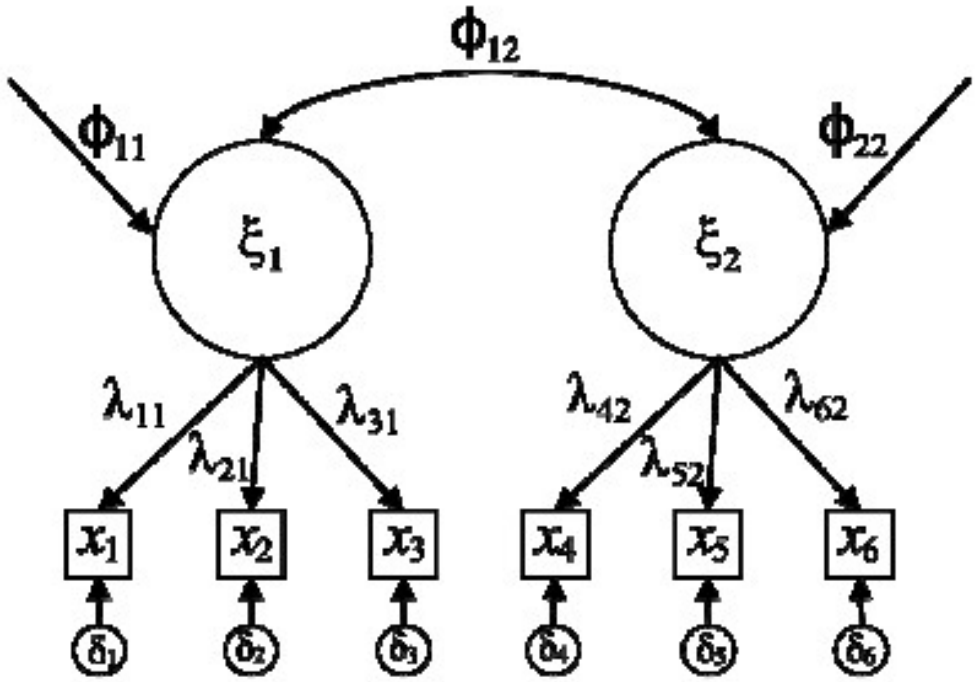


(b) Modello di analisi fattoriale confermativa



Il Modello Confermativo

$$x = \lambda\xi + \delta$$



Dall'Analisi Fattoriale Esplorativa a quella Confermativa

Procedimento di conferma

- L'analisi fattoriale esplorativa parte da una matrice di correlazione
- Il procedimento esplorativo stima il parametro di regressione con cui il fattore influenza l'item
- In questo modo possiamo :

1. "Ricostruire" la matrice di correlazione
2. Confrontare la matrice "osservata" ($\mathbf{R}$) con quella "ricostruita" ($\hat{\mathbf{R}}$)
3. Fare un'inferenza statistica usando $H_0 : \mathbf{R} = \hat{\mathbf{R}}$
4. Decidere se il modello da noi ipotizzato sia giustificato statisticamente

Contesto storico

Come sottolineano Nunnally e Bernstein (1994), i modelli delle strutture di covarianza rappresentano lo strumento statistico per la verifica di **teorie forti**. L'esempio più famoso di teoria forte è quello del fattore comune unico di intelligenza *g* di Spearman (1904). In questo caso, per provare la sua teoria Spearman doveva riuscire a dimostrare non solo che i punteggi osservati nelle varie prove potevano essere ricondotti ad un solo fattore, ma anche che questo fattore era effettivamente l'*unico* in grado di spiegare la loro covariazione. Nell'accezione di Nunnally e Bernstein, una **teoria debole** riguarda invece le modalità con cui le variabili si raggruppano: ad esempio, possiamo supporre che un gruppo di item di Estroversione item misurino l'Assertività, e un altro gruppo la Socievolezza. Non si assume che i fattori siano necessariamente indipendenti, e non ci si preoccupa del fatto che i due fattori ipotizzati riescano a spiegare completamente le relazioni fra gli item.

Contesto storico

1904, Charles Spearman: Teoria bifattoriale

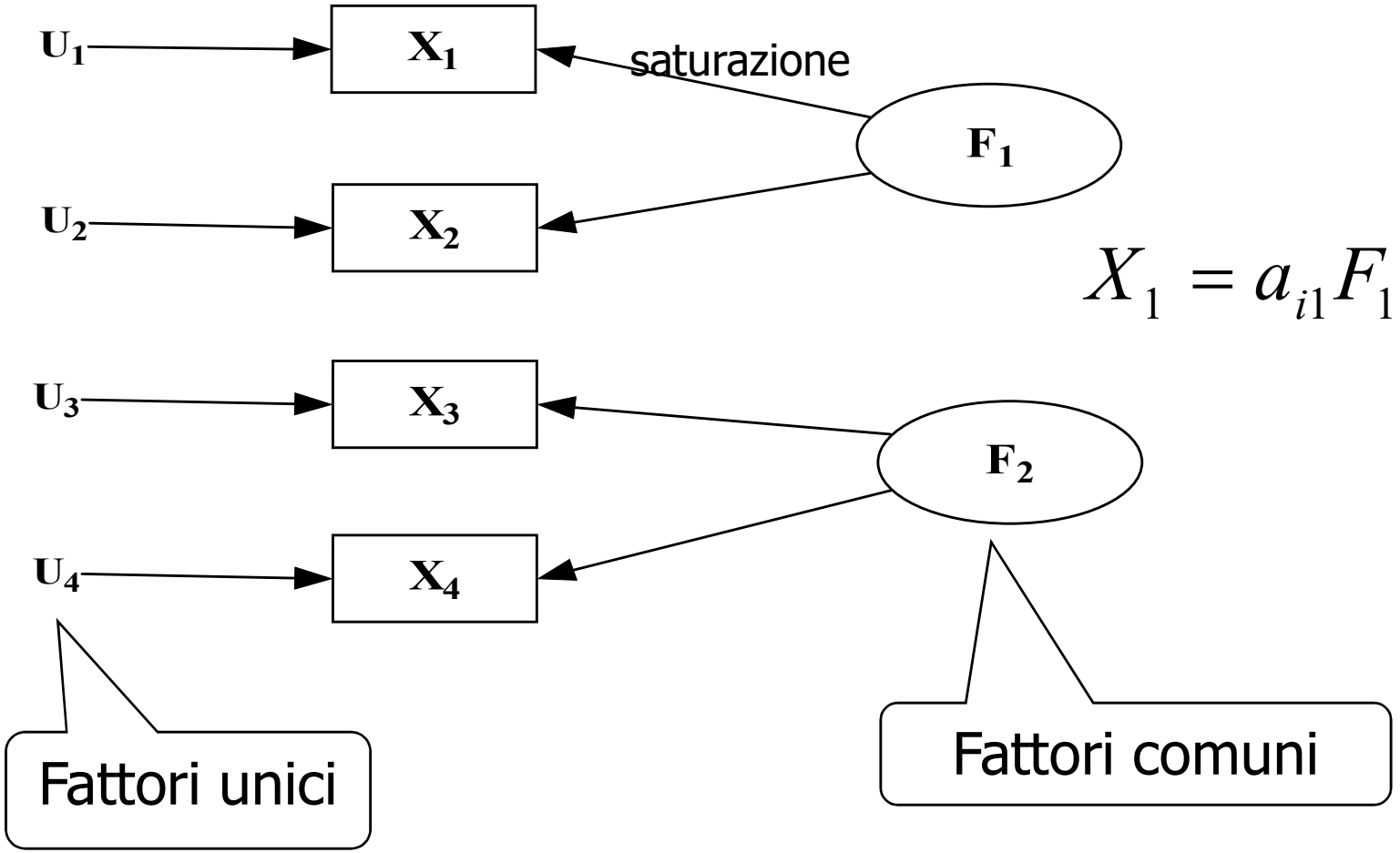
sosteneva che le **misure di abilità mentale** relative ad un test potevano essere spiegate come attribuibili ad **un'abilità generale** comune a tutte le abilità e ad un'abilità specifica e queste abilità dipendono ciascuna da un "fattore", chiamati da Spearman "Fattore generale" (G) e "fattore specifico o unico" (U).

Contesto storico

- 1945, Thurstone: Teoria multifattoriale

  propose di sostituire il **fattore generale con dei "fattori comuni" (F)**.
- La differenza è che i fattori comuni sono relativi solo ad alcuni item, quello generale li prendeva in considerazione tutti contemporaneamente.

ESEMPIO



$$X_1 = a_{i1} F_1 + U_1$$

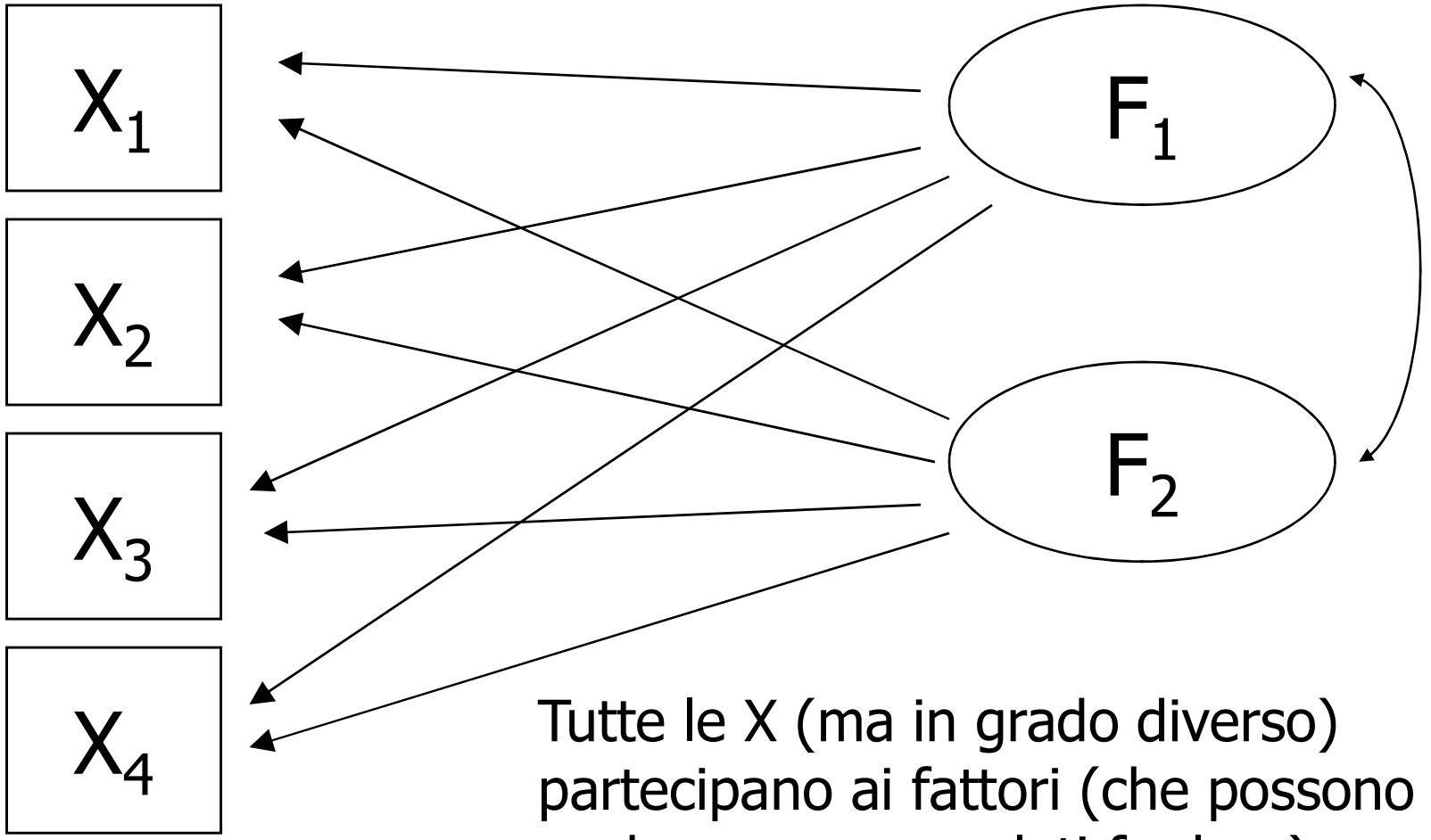
Esplorare/confermare

- L'analisi fattoriale esplorativa (AFE) serve per cercare le **variabili latenti all'interno delle osservate**: **non si hanno ipotesi a priori** su quali fattori influiscano sulle osservate.
- L'analisi fattoriale confermativa (AFC) serve quando si hanno idee abbastanza chiare (TEORIA) su **quali fattori influenzano quali variabili**. Quindi per verificare che certe relazioni ipotizzate fra le osservate e le latenti siano effettive.

Analisi fattoriale esplorativa

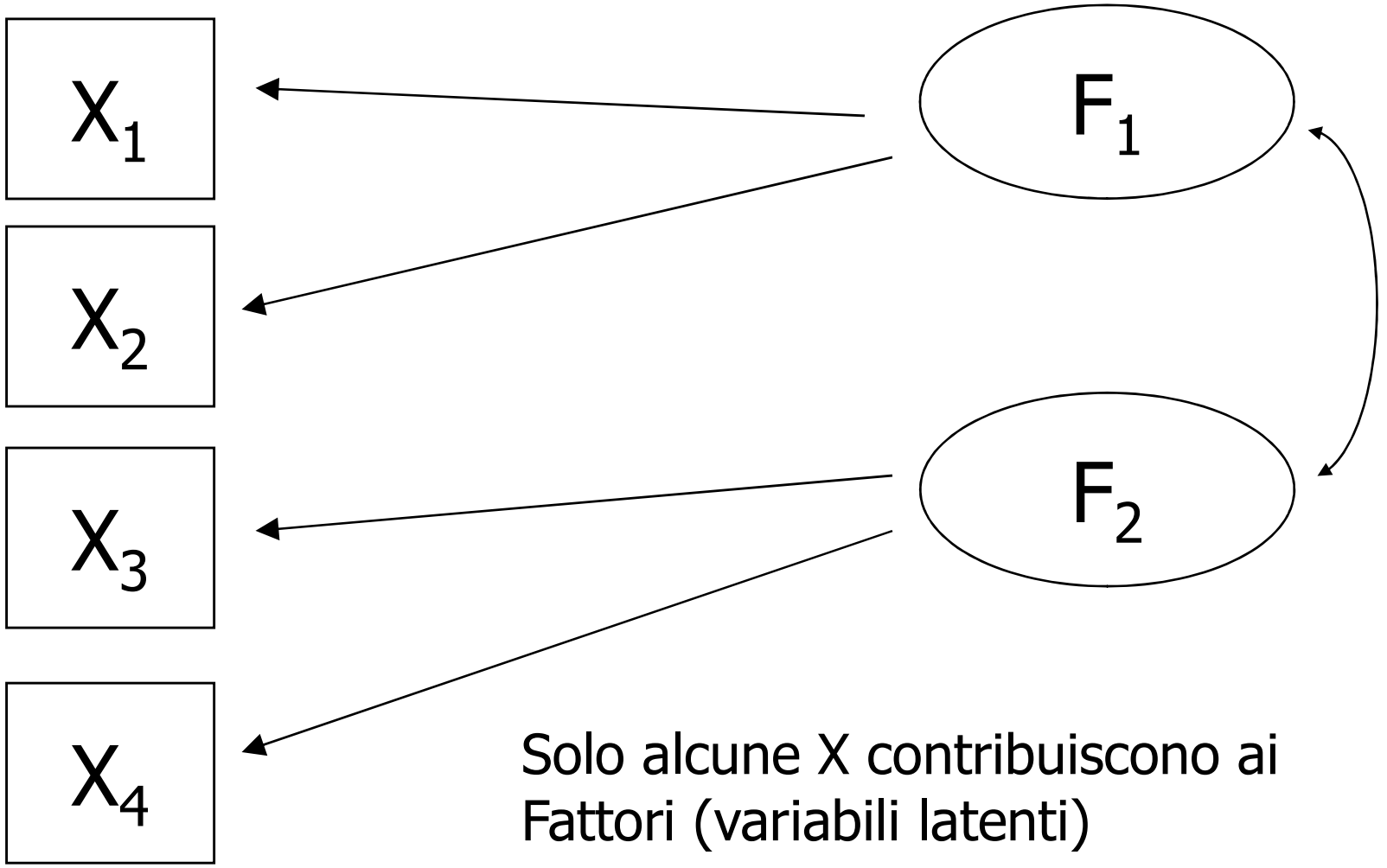
- Serve per **associare una o più variabili latenti** (che non si conoscono) **ad un gruppo di variabili osservate** che si presuppone abbiano qualche cosa in comune, ma non si sa esattamente 'cosa'.
- Questo "**qualcosa in comune**" viene chiamato ***Fattore***

Analisi fattoriale esplorativa



Tutte le X (ma in grado diverso) partecipano ai fattori (che possono anche essere correlati fra loro)

Analisi fattoriale confermativa



Implicazioni

- un **fattore può influire in una o più** variabili osservate
- **fattori diversi** possono influire su variabili osservate diverse
- la **differenza osservata** fra due individui in una stessa variabile osservata dipende, almeno parzialmente, **dalla loro differenza nel fattore**
- due variabili osservate **influenzate dal medesimo fattore devono correlare molto fra loro**

2. Procedimento analitico : dall'AFE alla AFC (note riepilogative)

Analisi fattoriale esplorativa

Teorema fondamentale $\mathbf{R=AA'+U^2}$

[storicamente, l'AFE si è svolta a partire da una matrice di correlazione, quindi con dati completamente standardizzati]

Assunzioni

- I fattori unici non correlano con i fattori comuni
- I fattori unici non correlano fra di loro
- I fattori comuni possono essere correlati fra di loro (soluzione obliqua) o non essere correlati (soluzione ortogonale)

Risultati

- Dall'analisi fattoriale di un **insieme di variabili osservate** (item di un questionario, misure psicometriche eseguite con vari test) si ottiene una **matrice fattoriale**, ossia una matrice di correlazioni fra le variabili latenti e le variabili osservate, che devono essere interpretate.
- Se la **soluzione trovata è ritenuta soddisfacente e adeguata**, si possono **stimare i punteggi fattoriali**, che sono le coordinate di ciascun partecipante su ciascuna dimensione latente (o detta "espresse in punti zeta").

Un singolo punteggio

$$z_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1f}F_f + u_1$$

z_1 è il punteggio standardizzato di una persona nella variabile 1
F_f è il punteggio standardizzato di una persona nel fattore f
a_{11} è la saturazione fattoriale della variabile 1 nel fattore 1
u_1 è il punteggio standardizzato di una persona nel fattore unico della variabile 1

Passaggi per una AFE

Verificare che l'AFE **si possa fare** (livelli di misura, normalità, valori anomali, numero di:

- »variabili,
- »fattori latenti
- »Soggetti

Verificare la matrice di correlazione (adeguatezza)

Estrarre i fattori (metodo)

Numero di fattori da estrarre (parsimoniosità vs spiegazione)

Interpretazione

Prima di un AFE

- Identificare un dominio di ricerca
- selezionare un certo numero di variabili osservabili che verranno misurate su un buon numero di unità statistiche (= partecipanti)
- le osservate che correlano molto fra loro possono sottintende un fattore
- le variabili che non correlano con nessun altra, vengono scartate (non sono incluse)

Analisi fattoriale esplorativa

$Z=FA'+U$ **$R=AA'+U^2$ (ipotesi ortogonale)**

$R=P\Phi P'+ U^2$ (ipotesi obliqua)

Z=dati grezzi standardizzati	nxm	n=soggetti
F=Fattori comuni	nxf	m=osservate
A,P=saturazioni/pesi	mxf	f=latenti
U=fattori unici	nxm	
R=matrice correlazioni	mxm	
Φ=correlazioni fra fattori	fxf	

La matrice di correlazione è riproducibile tramite una matrice di saturazioni fattoriali (dipendenti dai fattori comuni) moltiplicata per la sua trasposta e aggiungendo un termine "d'errore" corrispondente ai fattori unici

I requisiti minimi (o desiderabili)

- Dati **quantitativi** (scale a intervallo o a rapporto)
- Variabili con **distribuzione normale** (o almeno non troppo diversa dalla normale)
- Esclusione dei **valori anomali** (che alterano le correlazioni)
- Più **casi che variabili** (almeno 100)
- **I fattori** (o dimensioni latenti o componenti) non possono superare il numero di variabili osservate
- **Il numero di casi** non può essere inferiore al numero di variabili osservate
- Il numero di casi dovrebbe essere elevato (almeno 100-200). La stabilità completa (ripetibilità) si ottiene solo su 3-4000 casi.

Verificare l’adeguatezza

- La matrice di correlazione deve avere alti coefficienti tra le coppie di variabili

  (Nb determinante: se è alto, le correlazioni sono basse; se è basso, ci sono correlazioni alte)

Esempio matrice correlazioni

Correlazione di Pearson

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1										
X2	,891**									
X3	-,300	-,177								
X4	,126	,000	-,804**							
X5	-,221	-,124	,876**	-,721**						
X6	,842**	,799**	-,127	-,086	-,074					
X7	,747**	,704**	-,161	-,150	-,206	,878**				
X8	-,258	-,151	,885**	-,880**	,798**	-,084	-,001			
X9	-,763**	-,787**	,280	,106	,273	-,787**	-,850**	,201		
X10	,264	,141	-,884**	,928**	-,820**	,076	,036	-,915**	-,109	

**.

F1?=x1, x2, x6, x7, x9

F2?=x3,x4,x5,x8,x10

Come estrarre i fattori

- Ci sono diversi metodi per estrarre i fattori
  - Massima verosimiglianza (test sui fattori)
  - Minimi quadrati (test sui fattori)
  - *Alfa factoring*
  - *Image factoring*
  - *…*

Con un numero di variabili elevato, si equivalgono tutti

Quanti fattori estrarre

- Autovalori maggiori di 1
- Almeno l'x% (60-75%) di varianza spiegata
- Scree-test di Cattell
- Teoria
- Analisi parallela
- Chi-test

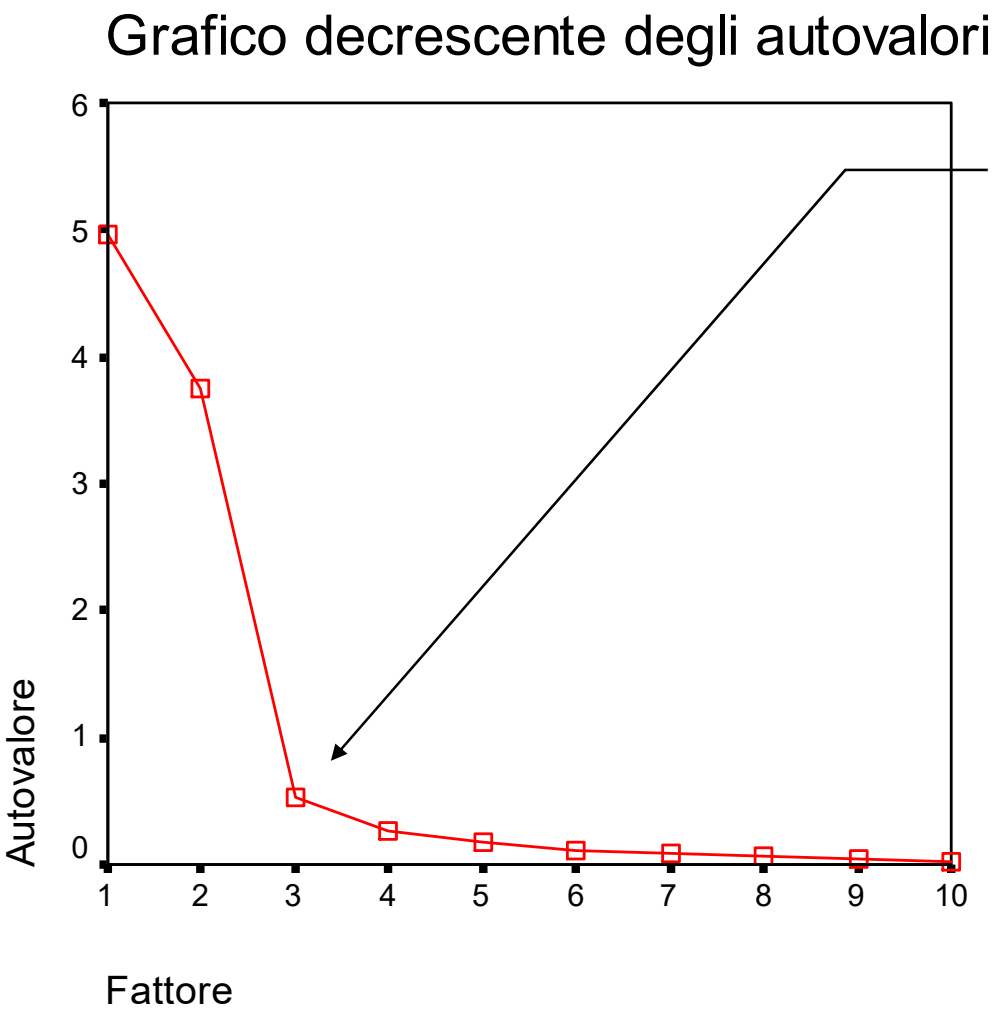
75% di varianza spiegata

Varianza totale spiegata

Fattore	Autovalori iniziali		
	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.

Scree-test



Punto di flesso

Per Harman si esclude (fattori 2)

per Cattell si include (fattori 3)

Test sui fattori

- **Massima verosimiglianza e minimi quadrati** permettono di calcolare una statistica di significatività (**un chi-quadro**) **sull'adattamento del modello fattoriale in base al numero dei fattori**.

- **Se il chi-quadro è *non significativo*, possiamo dire che la soluzione con *q* fattori si adatta bene** (accettabile bontà del modello)

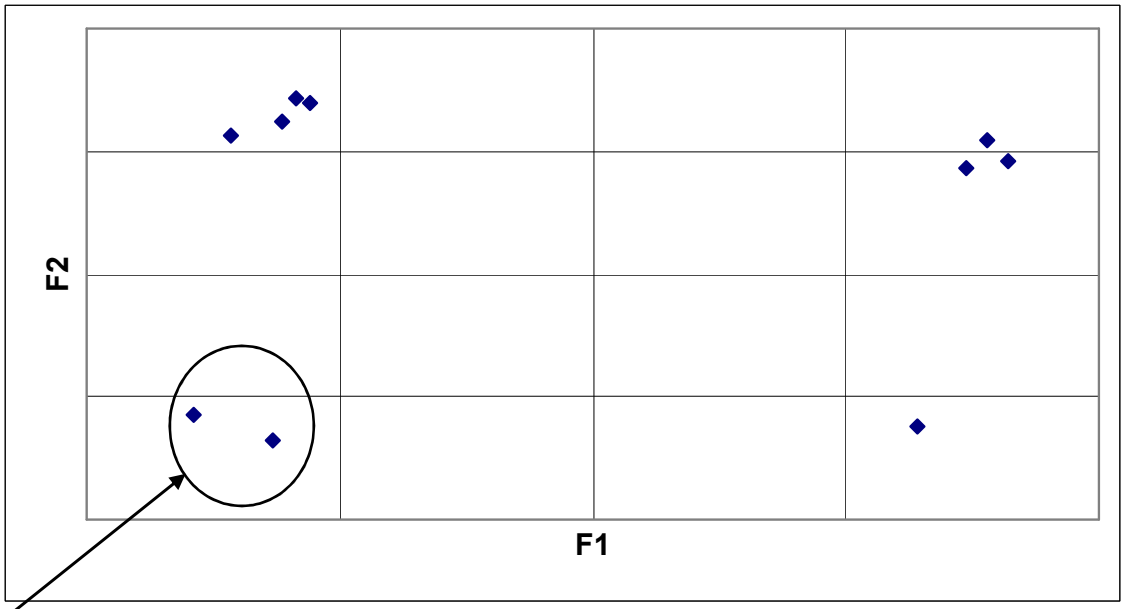
Metodi di rotazione

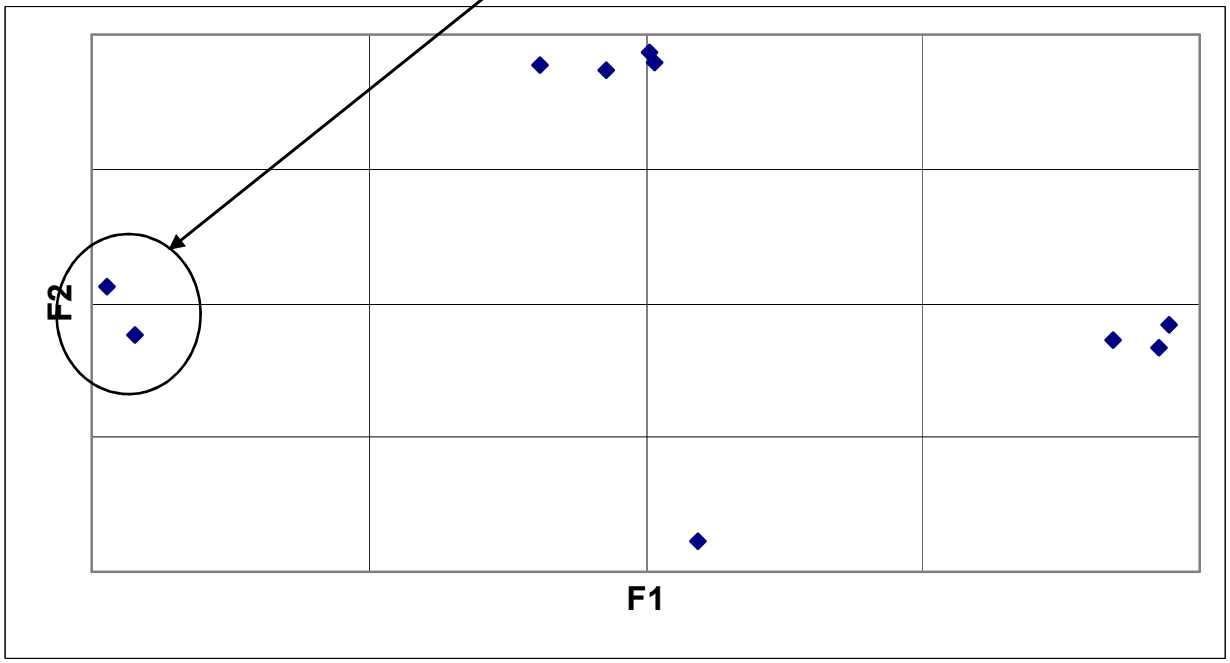
- Metodi ortogonali
  - **Varimax** (semplifica le righe: ogni variabile osservata è correlata massimamente con un fattore e nulla con gli altri). Metodo quasi sempre usato, per la sua efficacia semplificativa
  - Quartimax (semplifica le colonne: ogni colonna è massimamente correlata con tutte le variabili osservate e poco con le restanti)
  - Equamax (bilancia i due criteri)
- Metodi obliqui
  - **Promax**: rende gli assi obliqui in funzione di una soluzione iniziale (ortogonale – varimax).
  - Oblimin (obliquità minima): permette di fissare l'inclinazione degli assi e quindi le loro intercorrelazioni

La rotazione

- La rotazione **ortogonale** degli assi fattoriali rende interpretabili le dimensioni latenti (o fattori), mantenendo **l'indipendenza fra i fattori**.
- La rotazione **obliqua** permette un migliore adeguamento degli assi fattoriali alle variabili osservate **ma il criterio di indipendenza statistica fra i fattori non è più osservato**.

Non ruotata





Ruotata

La rotazione

- Nella soluzione **ortogonale**, le saturazioni possono essere interpretate come le correlazioni fra le variabili e i fattori.
- In tal caso il **loro quadrato corrisponde alla proporzione di varianza spiegata dal fattore per quella variabile**

Saturazione ^2=%var spiegata dal fattore

Varianza: chiave dell'interpretazione

La varianza dell'osservata X, può essere suddivisa in una parte dovuta ai fattori unici e una parte dovuta ai fattori comuni:

$var(X)=var(F)+ var(U)$

Il rapporto fra var(F) e var(X) si chiama "comunalità" (h^2), mentre var(U) si chiama "unicità" (u^2).

Essendo la $var(x) = 1 = h^2 + u^2$

L'unicità può essere ulteriormente suddivisa in varianza specifica dell'item ed varianza d'errore, ma l'AF non fa distinzione fra le due

Soluzione non ruotata

	Factor 1	Factor 2	Unique Var
	---	---	---
VAR1	0.333	**0.843**	0.178
VAR2	0.208	**0.846**	0.240
VAR3	**-0.916**	-0.036	0.160
VAR4	**0.919**	-0.256	0.089
VAR5	**-0.842**	-0.022	0.290
VAR6	0.141	**0.918**	0.138
VAR7	0.102	**0.901**	0.178
VAR8	**-0.941**	0.054	0.112
VAR9	-0.197	**-0.872**	0.200
VAR10	**0.976**	-0.076	0.041

h^2=
.333^2
+
.843^2)

u^2=.178

1=
h^2+u^2=
.822+178

Soluzione non ruotata

Matrice fattoriale[a]

	Fattore	
	1	2
X3	-,819	,464
X10	,788	-,571
X8	-,780	,541
X5	-,739	,431
X1	,716	,563
X9	-,642	-,624
X6	,588	,719
X7	,561	,698
X4	,633	-,681
X2	,612	,625

Metodo estrazione: fattorizzazione dell'asse principale.

a. 2 fattori estratti. 6 iterazioni richieste.

Varimax-Rotated F. Loadings

	Factor 1	Factor 2	Unique Var
	---	---	---
VAR1	0.209	**0.882**	0.178
VAR2	0.085	**0.867**	0.240
VAR3	**-0.901**	-0.166	0.160
VAR4	**0.946**	-0.122	0.089
VAR5	**-0.831**	-0.142	0.290
VAR6	0.009	**0.929**	0.138
VAR7	-0.027	**0.906**	0.178
VAR8	**-0.939**	-0.081	0.112
VAR9	-0.071	**-0.892**	0.200
VAR10	**0.977**	0.064	0.041

1-
(.209^2
+.882^2)
=.178

Soluzione ruotata semplificata

Matrice fattoriale ruotata[a]

	Fattore	
	1	2
X10	-,970	
X8	,946	
X3	,926	
X4	-,922	
X5	,843	
X6		,929
X7		,895
X9		-,891
X1		,890
X2		,872

Metodo estrazione: fattorizzazione dell'asse principale.
Metodo rotazione: Varimax con normalizzazione di Kaiser.

a. La rotazione ha raggiunto i criteri di convergenza in 3 iterazioni.

Fattore 1:

items x10, x8, x3, x4, x5

Fattore 2:

items x6, x7, x9, x1, x2

Promax-Rotated F. Loadings

	Factor 1	Factor 2	Unique Var
VAR1	**0.875**	0.141	0.178
VAR2	**0.869**	0.017	0.240
VAR3	-0.106	**-0.895**	0.160
VAR4	-0.188	**0.963**	0.089
VAR5	-0.086	**-0.826**	0.290
VAR6	**0.936**	-0.065	0.138
VAR7	**0.916**	-0.099	0.178
VAR8	-0.017	**-0.940**	0.112
VAR9	**-0.894**	-0.001	0.200
VAR10	-0.003	**0.980**	0.041

1-
(.875^2
+.141^2)
diverso da
.178

Factor Correlations

	Factor 1	Factor 2
Factor 1	1.000	
Factor 2	0.070	1.000

Punteggi (scores) fattoriali

Punteggio che ogni osservazioni assume in un certo fattore

Tutti i **programmi calcolano** i punteggi fattoriali e usano varie forme di regressione multipla

Metodo congenerico (punteggi fattoriali compositi): si sommano (o si fa la media) delle sole osservate che fanno parte del fattore

3. Esercizi in R

Problem 1 – Passito

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

- A partire dall’AFE effettuata, impostare una AFC

Problem 2 – centro commerciale

- Eseguire una ANALISI FATTORIALE ESPLORATIVA sulle 5 variabili risposta per individuare $q<5$ nuove variabili che "spiegano" i dati
- A partire dall'AFE effettuata, impostare una AFC

Problem 3 – abitudini alimentari

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

- A partire dall'AFE effettuata, impostare una AFC