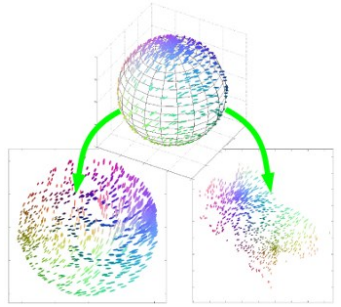
Natural and Artificial Cognition Lab Orazio Miglino
University of Naples Federico II

I metodi per lo studio della dimensionalità dei test psicologici: l'Analisi delle Componenti Principali e l'Analisi Fattoriale

Michela Ponticorvo

Natural and Artificial Cognition Lab Orazio Miglino
University of Naples Federico II

Studiare la dimensionalità



- Il concetto di dimensionalità è importante sia nel machine learning che in psicologia. In molti casi l'obiettivo è ridurre la dimensionalità
- Ridurre la dimensionalità significa ridurre il numero di caratteristiche o dimensioni in un dataset mantenendo quanta più informazione possibile
- Perchè? Per ridurre la complessità di un modello, per migliorare la performance di un algoritmo di apprendimento, per facilitare la visualizzazione dei dati…per comprendere la struttura dei test.
- Esistono molte tecniche: principal component analysis (PCA), singular value decomposition (SVD), linear discriminant analysis (LDA), analisi fattoriale. Ogni tecnica usa metodi diversi per proiettare i dati in uno spazio che ha meno dimesioni preservando le informazioni importanti.
- É un processo di trasformazione di dati con un grande numero di dimensioni in uno spazio con meno dimensioni che preserva l'essenza dei dati originali.

MMPI-2: 567 item → MMPI-2: 10 dimensioni cliniche

NAC
Natural and Artificial Cognition Lab Orazio Miglino
University of Naples Federico II

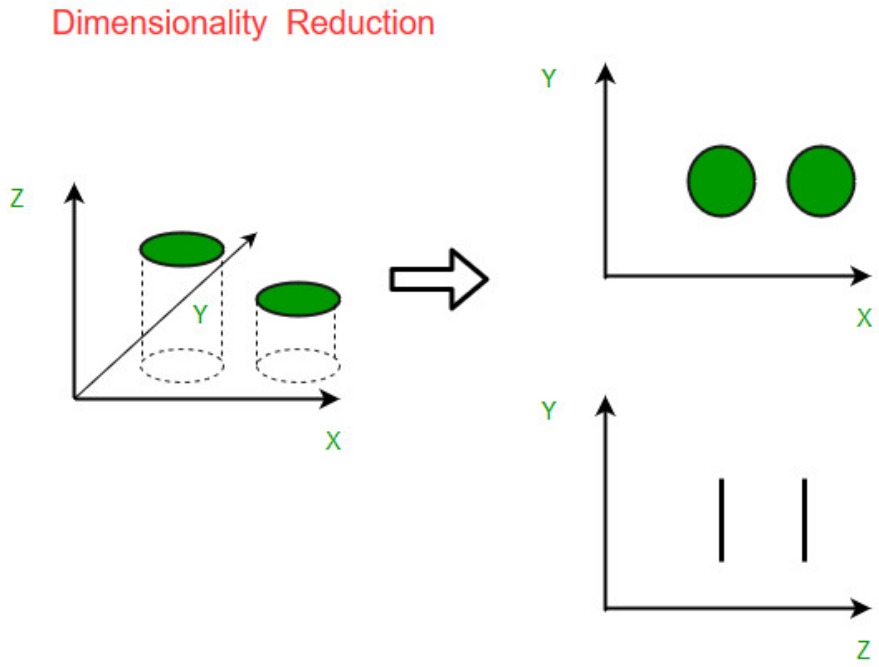
Studiare la dimensionalità

- Nell'ambito del machine learning, ci possono essere un gran numero di feature. La maledizione della dimensionalità è un problema, perchè la performance di un modello peggiora all'aumentare del numero di feature.
- Per la riduzione della dimensionalità ci sono 2 approcci principali: feature selection and feature extraction.

Feature Selection: si selezionano alcune delle feature originali, le più rilevanti

Feature Extraction: si creano delle nuove feature combinando o trasformando le feature originali. Queste nuove feature catturano l'essenza dei dati originali in uno spazio che ha meno dimensioni.

- Di questo secondo approccio fanno parte Principal component analysis (PCA), linear discriminant analysis (LDA), and t-distributed stochastic neighbor embedding (t-SNE).
- La PCA è una tecnica molto usata che proietta le feature originali in uno spazio con meno dimensioni, preservando l'informazione della varianza il più possible



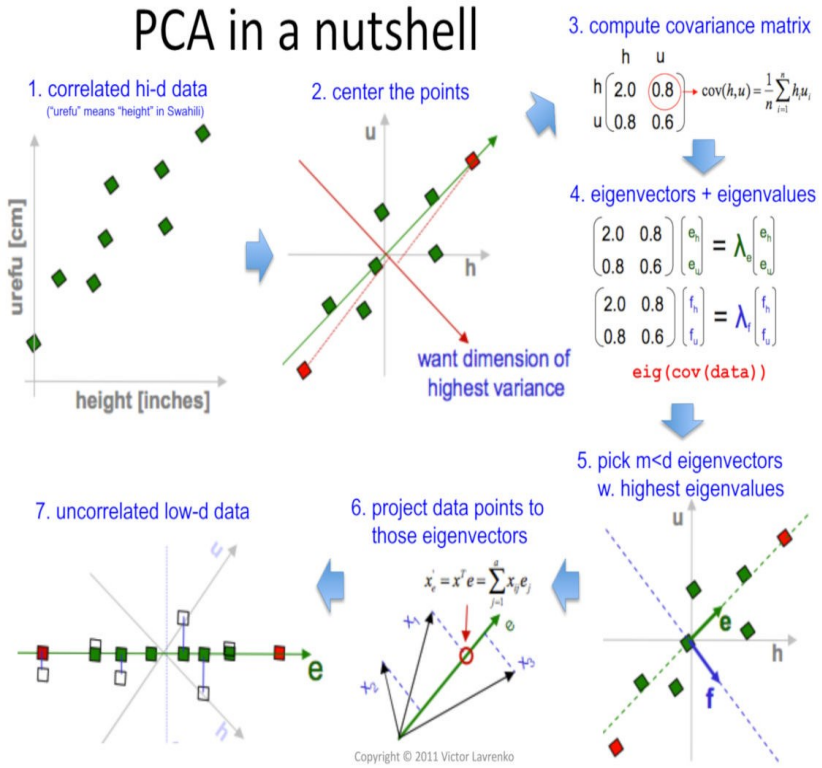
Analisi delle componenti principali

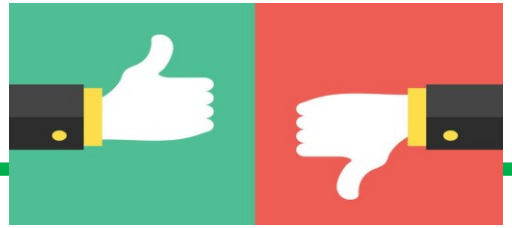
Metodo introdotto da Karl Pearson all'inizio del 1900.

È un caso particolare di Single Value Decomposition applicato ad una matrice quadrata.

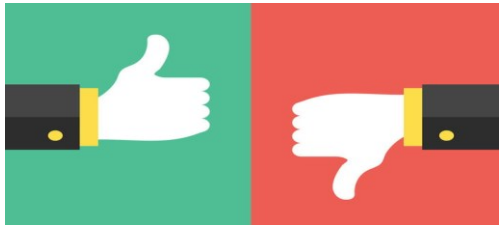
I dati in uno spazio con molte dimensioni vengono mappati in uno spazio con meno dimensioni in modo che la varianza dei dati nello spazio con poche dimensioni sia massima.

- Si costruisce la matrice di covarianza dei dati
- Si calcolano gli autovettori di questa matrice
- Gli autovettori che corrispondono agli autovalori più alti sono usati per ricostruire una porzione di varianza dei dati originali
- Avremo quindi un numero minore di autovettori e avremo perso delle informazioni in questo processo, ma la parte più consistente della varianza è mantenuta





- Aiuta nella compressione dei dati, riduce lo spazio di archiviazione ed il tempo di calcolo.
- Aiuta anche a rimuovere le informazioni ridondanti, se presenti.
- Visualizzazione migliorata: i dati ad alta dimensione sono difficili da visualizzare e le tecniche di riduzione della dimensionalità possono aiutare a visualizzare i dati in 2D o 3D, il che può aiutare ad avere una migliore comprensione e analisi.
- Prevenzione dell'overfitting: i dati ad alta dimensionalità possono portare all'overfitting nei modelli di machine learning e a scarse prestazioni di generalizzazione. La riduzione della dimensionalità può aiutare a ridurre la complessità dei dati e quindi prevenire l'overfitting.
- Estrazione di feature: la riduzione della dimensionalità può aiutare a estrarre feature importanti da dati ad alta dimensionalità, che possono essere utili nella selezione delle feature per i modelli di machine learning.
- Preelaborazione dei dati: la riduzione della dimensionalità può essere utilizzata come fase di preelaborazione prima di applicare algoritmi di apprendimento automatico per ridurre la dimensionalità dei dati e quindi migliorare le prestazioni del modello.
- Prestazioni migliorate: la riduzione della dimensionalità può aiutare a migliorare le prestazioni dei modelli di apprendimento automatico riducendo la complessità dei dati e quindi riducendo il rumore e le informazioni irrilevanti nei dati.



- Potrebbe portare ad una perdita di dati.
- La PCA tende a trovare correlazioni lineari tra le variabili.
- PCA fallisce nei casi in cui la media e la covarianza non sono sufficienti per definire i set di dati.
- Potremmo non sapere quanti componenti principali mantenere, in pratica vengono applicate alcune regole empiriche.
- Interpretabilità: le dimensioni ridotte potrebbero non essere facilmente interpretabili e potrebbe essere difficile comprendere la relazione tra le feature originali e le dimensioni ridotte.
- Sensibilità ai valori anomali: alcune tecniche di riduzione della dimensionalità sono sensibili ai valori anomali, che possono comportare una rappresentazione distorta dei dati.

Analisi delle componenti principali

Riassumendo….

- La riduzione della dimensionalità è il processo di riduzione del numero di feature in un set di dati conservando quante più informazioni possibili.
- Questo può essere fatto per ridurre la complessità di un modello, migliorare le prestazioni di un algoritmo di apprendimento o semplificare la visualizzazione dei dati.
- Ogni tecnica proietta i dati su uno spazio dimensionale inferiore preservando le informazioni importanti.
- La riduzione della dimensionalità viene eseguita durante la fase di pre-elaborazione prima di costruire un modello per migliorare le prestazioni
- È importante notare che la riduzione della dimensionalità può anche scartare informazioni utili, quindi è necessario prestare attenzione.

PCA passo per passo

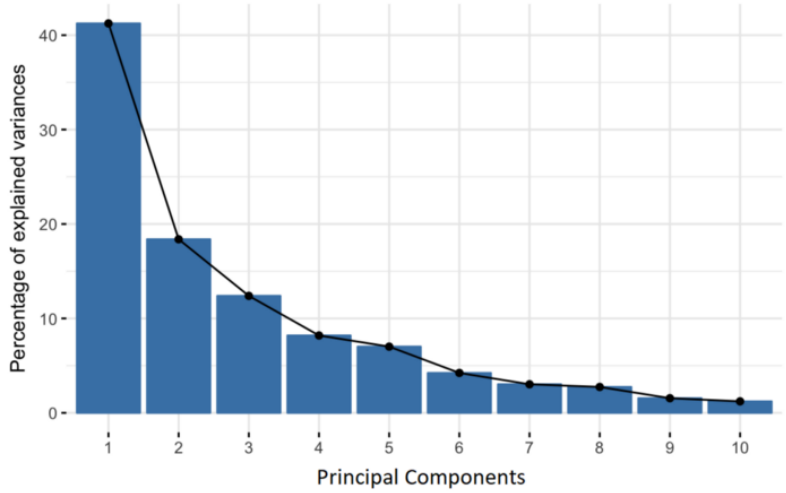
STEP 1 Standardizzazione

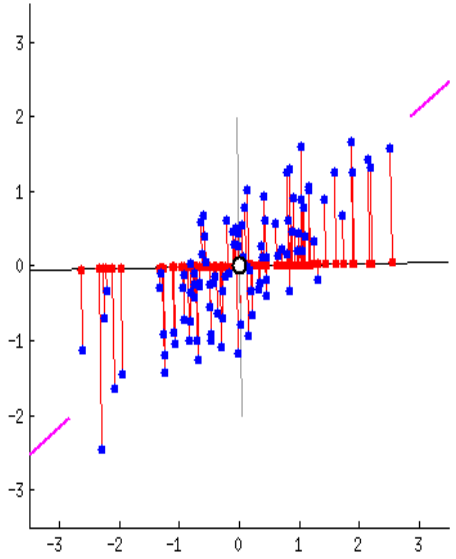
$$z = \frac{value - mean}{standard\ deviation}$$

STEP 2 Calcolo della matrice di covarianza

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

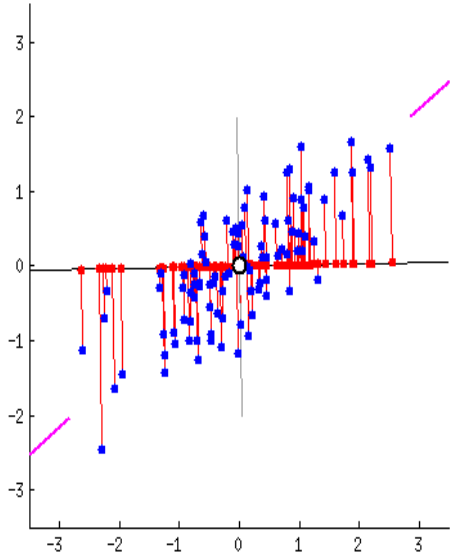
STEP 3 Calcolo degli autovettori ed autovalori della matrice di covarianza per identificare le componenti principali





Poiché ci sono tante componenti principali quante sono le variabili nei dati, le componenti principali sono costruite in modo tale che la prima componente principale rappresenti la massima varianza possibile nel set di dati. Ad esempio, supponiamo che il grafico a dispersione del nostro set di dati sia come mostrato nell'animazione, possiamo indovinare il primo componente principale?

Sì, è approssimativamente la linea che corrisponde ai segni viola perché passa per l'origine ed è la linea in cui la proiezione dei punti (punti rossi) è più estesa. O matematicamente parlando, è la linea che massimizza la varianza (la media delle distanze al quadrato dai punti proiettati (punti rossi) all'origine).



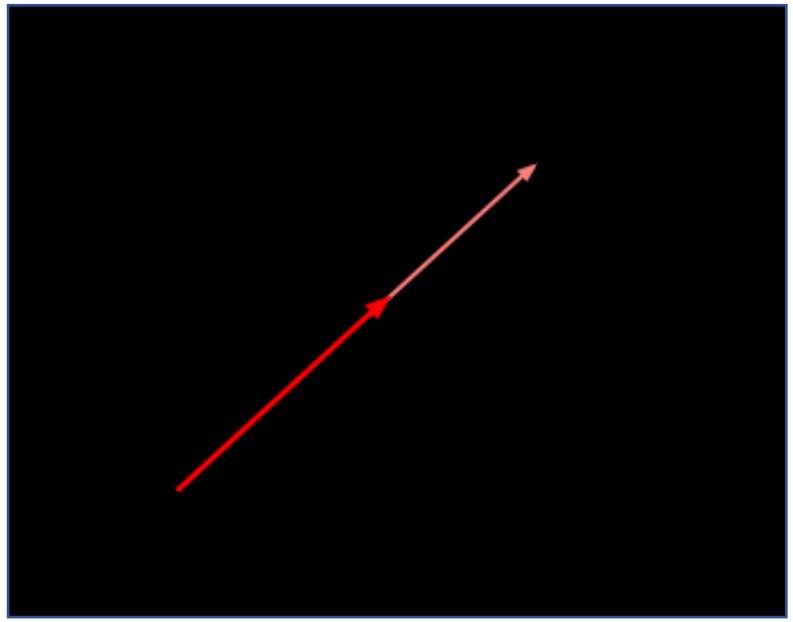
La seconda componente principale viene calcolata allo stesso modo, con la condizione che sia incorrelata con (cioè perpendicolare a) la prima componente principale e che tenga conto della successiva varianza più alta.

Questo continua fino a quando non è stato calcolato un totale di p componenti principali, pari al numero originale di variabili.

Gli autovettori della matrice di covarianza sono in realtà le direzioni degli assi dove c'è più varianza (più informazioni) e che chiamiamo componenti principali. Gli autovalori sono i coefficienti associati agli autovettori, che danno la quantità di varianza trasportata in ogni Componente Principale.

Classificando gli autovettori in ordine di autovalori, dal più alto al più basso, ottieni i componenti principali in ordine di importanza.

Eigenvalue



Un autovettore per la trasformazione lineare *L* è un vettore $x \neq 0$ che a seguito dell'applicazione di *L* non cambia la sua direzione, limitandosi ad essere moltiplicato per uno scalare, il rispettivo **autovalore**. Il vettore può quindi soltanto cambiare modulo (venendo amplificato o contratto) e verso (venendo ribaltato)

Nell'analisi fattoriale gli *eigenvalues* della matrice di correlazione sono in relazione (rappresentano) la varianza spiegata da un fattore. Solo valori alti sono da considerare

PCA passo per passo

STEP 4 Costruiamo il vettore delle componenti

Una matrice che ha in colonna gli autovettori delle componenti che decidiamo di tenere

STEP 5 Riportiamo i dati sugli assi dati dalle componenti principali

MATRICE DI VARIANZA E COVARIANZA

Se **X** è una matrice dei dati unità per variabili di dimensioni *n,k*, la matrice di varianze e covarianze **S** è:

$$\mathbf{X} = \begin{array}{c|cccccc} & x_1 & x_2 & x_3 & x_4 & x_j & x_k \\ 1 & x_{11} & x_{12} & x_{13} & x_{14} & x_{1j} & x_{1k} \\ 2 & x_{21} & x_{22} & x_{23} & x_{24} & x_{2j} & x_{2k} \\ 3 & x_{31} & x_{32} & x_{33} & x_{34} & x_{3j} & x_{3k} \\ 4 & x_{41} & x_{42} & x_{43} & x_{44} & x_{4j} & x_{4k} \\ 5 & x_{51} & x_{52} & x_{53} & x_{54} & x_{5j} & x_{5k} \\ 6 & x_{61} & x_{62} & x_{63} & x_{64} & x_{6j} & x_{6k} \\ i & x_{i1} & x_{i2} & x_{i3} & x_{i4} & x_{ij} & x_{ik} \\ n & x_{n1} & x_{n2} & x_{n3} & x_{n4} & x_{nj} & x_{nk} \end{array}$$

$$\mathbf{S} = \begin{array}{c|cccccc} & x_1 & x_2 & x_3 & x_4 & x_j & x_k \\ x_1 & \mathrm{Var}_{x1} & & & & & \\ x_2 & \mathrm{Cov}_{1,2} & \mathrm{Var}_{x2} & & & & \\ x_3 & \mathrm{Cov}_{1,3} & \mathrm{Cov}_{2,3} & \mathrm{Var}_{x3} & & & \\ x_4 & \mathrm{Cov}_{1,4} & \mathrm{Cov}_{2,4} & \mathrm{Cov}_{3,4} & \mathrm{Var}_{x4} & & \\ x_j & \mathrm{Cov}_{1,j} & \mathrm{Cov}_{2,j} & \mathrm{Cov}_{3,j} & \mathrm{Cov}_{4,j} & \mathrm{Var}_{xj} & \\ x_k & \mathrm{Cov}_{1,k} & \mathrm{Cov}_{2,k} & \mathrm{Cov}_{3,k} & \mathrm{Cov}_{4,k} & \mathrm{Cov}_{j,k} & \mathrm{Var}_{xk} \end{array}$$

Esempio di matrice di varianze e covarianze

Matrice dei dati **X** =

A	B	C	D
7,51	4,90	4,05	75,49
9,12	12,92	7,70	14,51
5,28	9,60	1,99	86,61
5,69	17,51	6,96	8,01
0,06	13,36	5,87	35,28
7,02	3,30	5,72	60,48
7,36	21,65	1,74	19,27
0,34	15,54	2,26	69,93
2,00	29,05	6,94	52,14
4,39	26,25	0,44	37,23
6,84	13,25	1,87	32,70
4,15	21,63	0,03	29,77
7,60	11,57	3,90	76,20

Matrice di varianze e covarianze **S** =

	A	B	C	D
A	7,65			
B	-7,99	54,35		
C	0,53	-3,77	6,32	
D	-8,28	-81,86	-11,23	617,84

Variabilità e autovalori

Matrice di varianze e covarianze **S =**

	A	B	C	D
A	7,65			
B	-7,99	54,35		
C	0,53	-3,77	6,32	
D	-8,28	-81,86	-11,23	617,84

La variabilità totale può essere sintetizzata dalla traccia della matrice **S**

traccia(**S**) **=** 7,64+54,35+6,32+617,84=686

A partire da una matrice, l'algebra lineare ci permette di calcolare dei valori, chiamati autovalori, i quali, nel caso di una matrice di var/cov, ricostruiscono la variabilità totale**.** Possiamo estrarre tanti autovalori quante sono le variabili in X. La somma di questi autovalori è uguale alla traccia della matrice S (la variabilità totale)

Autovalori e componenti principali

Ad ogni autovalore è associata una componente principale.

L'autovalore può essere interpretato come la varianza della componente principale associata.

La prima componente principale sarà quella con autovalore più alto, cioè quella in grado di spiegare più variabilità possibile.

La seconda componente principale sarà quella associata al secondo autovalore più grande, e così via.

Quando ci fermiamo?
Se prendiamo tutte le componenti (che saranno tante quante sono le variabili in **X**), stiamo spiegando tutta la variabilità totale ma non stiamo riducendo la dimensionalità del problema, quindi l'ACP risulterebbe inutile.
Quindi…

Criteri di scelta del numero di componenti

1. **Variabilità spiegata**

   Si fissa una soglia minima di variabilità spiegata (rapporto tra la somma dei primi k autovalori/variabilità totale)

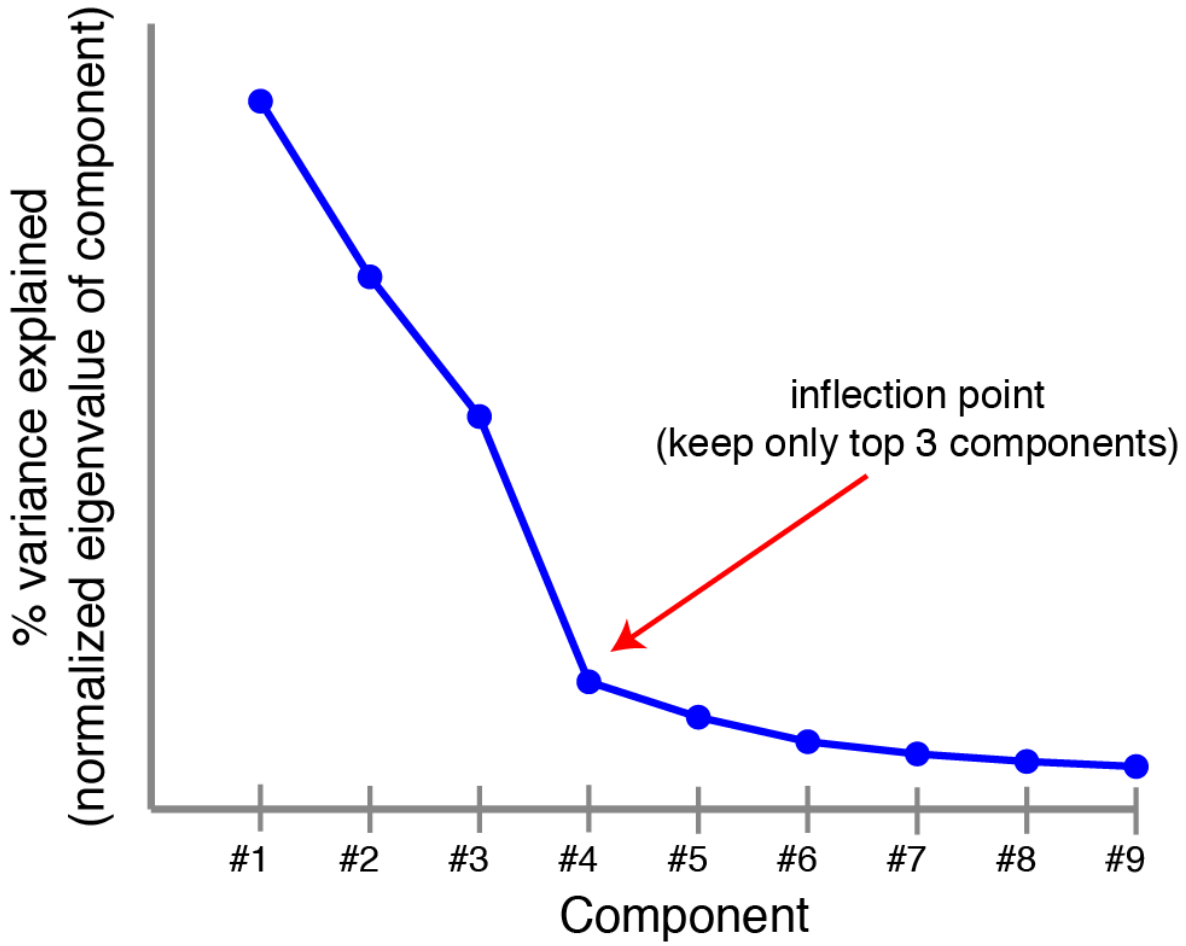
2. **Eigenvalue-one** (per variabili standardizzate)

   Poiché le variabili osservata hanno varianza unitaria si scelgono solo gli autovalori maggiori di uno (le componenti che spiegano più variabilità di una singola variabile osservata)

3. **Scree-Test**

   Si considerano le componenti i cui autovalori precedono il salto massimo di variabilità spiegata.

Esempio scree-plot



Sulle ascisse troviamo il numero di componenti.

Sulle ordinate la variabilità spiegata da ciascuna componente

Interpretazioni dei risultati

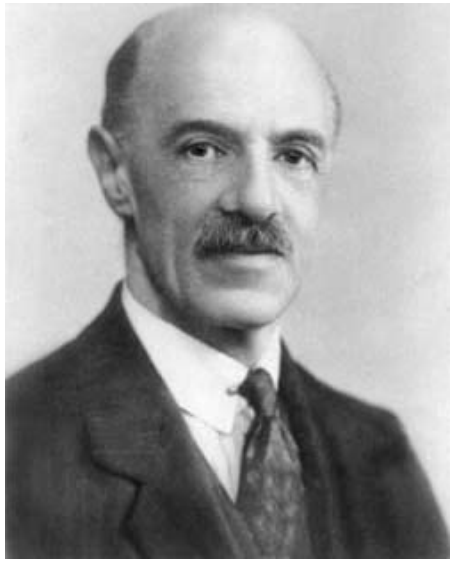
Non è prevista una validazione del modello nella PCA, perché non ha un modello di riferimento.

Di principale importanza nella PCA è l'interpretazione delle componenti principali in base ai loading, che misurano la loro relazione con le variabili osservate.

Per ogni variabile osservata si misura la percentuale di variabilità spiegata dalle componenti: comunalità

ANALISI FATTORIALE

Charles Edward Spearman (10 Settembre 1863 – 17 Settembre 1945)
Psicologo inglese



C. Spearman sviluppo le basi dell'Analisi Fattoriale (AF) all'inizio del 1900 per misurare l'intelligenza
L'idea di base è che le correlazioni tra le risposte fornite a un set di test di abilità siano descritte da un unico "fattore generale" di intelligenza

Uno dei maggiori contributi dell'Analisi Fattoriale è il concetto di "fattore", in altre parole il concetto di Variabile Latente (VL)

ANALISI FATTORIALE

L'analisi fattoriale è un metodo di statistica multivariata che permette di ottenere una riduzione della complessità del numero di variabili (osservate) che spiegano un fenomeno (metodo di riduzione della dimensionalità) eliminando la ridondanza di informazioni nei dati

Si propone quindi di determinare un numero di fattori minore rispetto al numero di variabili osservate in partenza (indicatori empirici, item, variabili manifeste), le quali possono essere altamente correlate

I fattori sono variabili latenti, non osservabili direttamente, che rappresentano la parte comune a più variabili osservate (ad esempio un tratto di personalità, un atteggiamento)

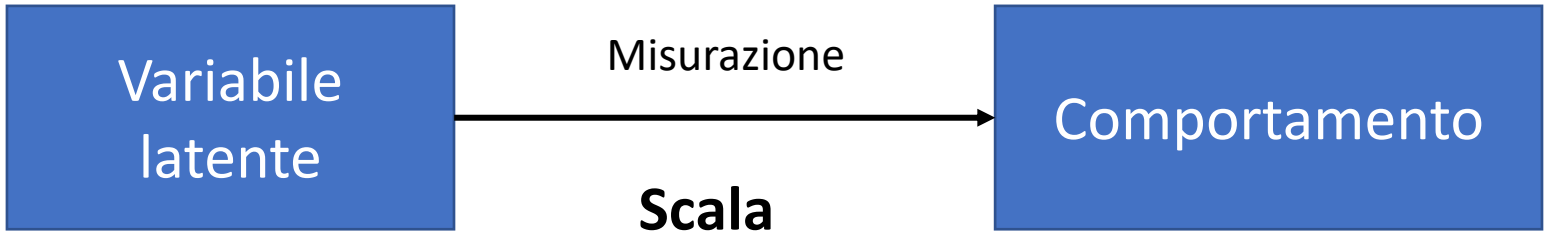
AF: VARIABILE LATENTE

I fattori sono variabili latenti, non osservabili direttamente, che rappresentano la parte comune a più variabili osservate (ad esempio un tratto di personalità, un atteggiamento)

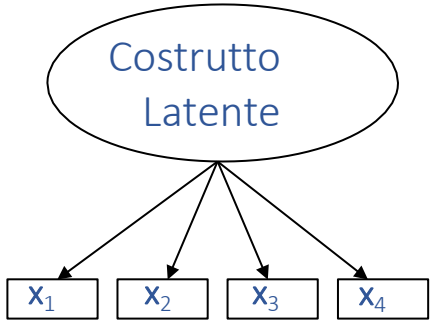
L'AF può utilizzata per costruire e validare di strumenti di misura (es. gli item sono coerenti con la definizione del costrutto?), oppure per verificare se la struttura teorica ipotizzata di un insieme di misure può essere confermata o meno (es. un atteggiamento consta di separate componenti cognitive, emotive e comportamentali?)

Variabili latenti

- Data la loro natura intangibile, i costrutti che sottendono comportamenti osservabili sono anche chiamati variabili latenti
- I costrutti si misurano attraverso una serie di correlazioni tra comportamenti. Gli strumenti che misurano le variabili latenti sono anche conosciute come scale



Costrutto latente o emergente



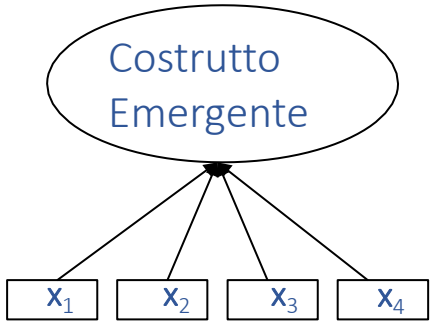
Indicatori Riflessivi, es. Intelligenza

Le VL **generano le corrispondenti VM** (la VL sottostante ogni blocco è unica: blocco unidimensionale)

Le VL sono **antecedenti rispetto alle VM**

Gli indicatori riflessivi di uno stesso blocco devono essere fortemente legati tra loro, in altre parole devono **covariare**

La consistenza Interna al blocco deve essere verificata (ad esempio usando l'alpha di Cronbach)



Indicatori Formativi

es. Status Sociale

- La VL è una **combinazione linerare delle corrispondenti VM** (ogni VL è quindi un constutto multidimesionale)
- Non è detto che gli indicatori fomativi di uno stesso blocco debbano covariare tra loro.
- La consistenza Interna non necessita di essere verificata.
- La variabile latente è determinata da più variabili osservate

AF: OBIETTIVO

In generale, l'obiettivo di una Analisi Fattoriale (AF) è quello di spiegare la variabilità e correlazione esistente tra una serie di variabili direttamente osservate, in termini di un numero ridotto di variabili latenti: i fattori

L'ipotesi di base è che la correlazione tra le variabili sia determinata da dimensioni non osservabili (i fattori) che in qualche modo sono causa o determinano i punteggi osservati nelle variabili osservate

Da un punto di vista statistico la variabilità è informazione: quanto più elevata la variabilità, tanto maggiore il contenuto informativo nei dati

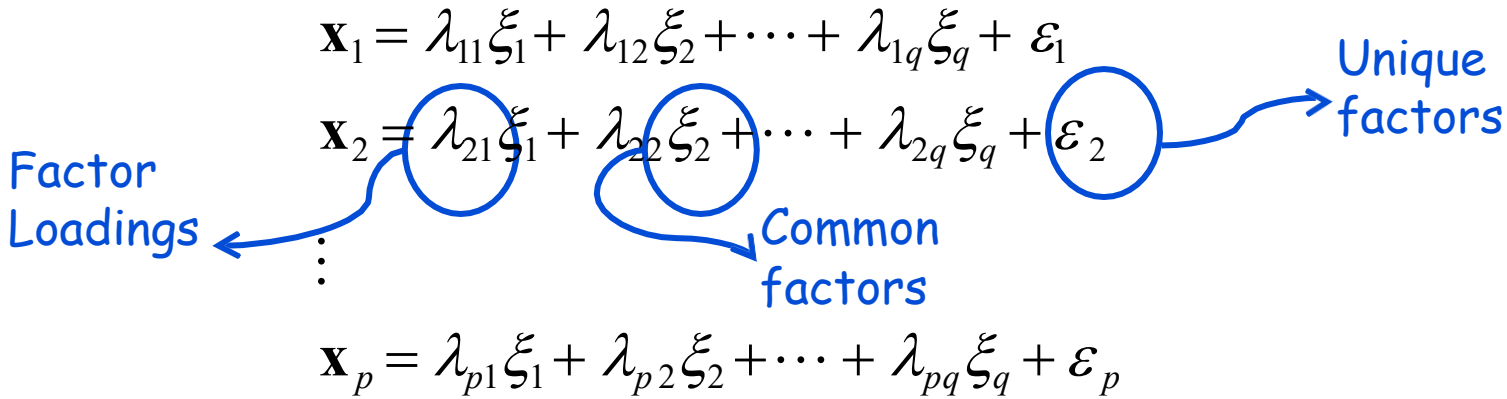
Nella statistica univariata, la variabilità è rappresentata dagli indici di variabilità: Varianza, deviazione standard, devianza, ecc.
Nella statistica multivariata, la variabilità è definita a partire dalla matrice di varianze e covarianze

NAC
Natural and Artificial Cognition Lab Orazio Miglino
University of Naples Federico II

Modello dell' analisi fattoriale

L'AF consiste nella stima di un **modello** che riproduca la struttura della covarianza tra le variabili osservate.

In termini più formali:
date ***p*** variabili manifeste osservate su n individui $x_1 \ldots x_p$ nell'AF ciascuna delle p variabili manifeste viene espressa come funzione lineare di ***q*** fattori "comuni" (con q <p), responsabili della correlazione della specifica variabile manifesta con le altre variabili manifeste, ed un unico errore di misura, responsabile della variabilità della variabile stessa.



Modello dell’ analisi fattoriale

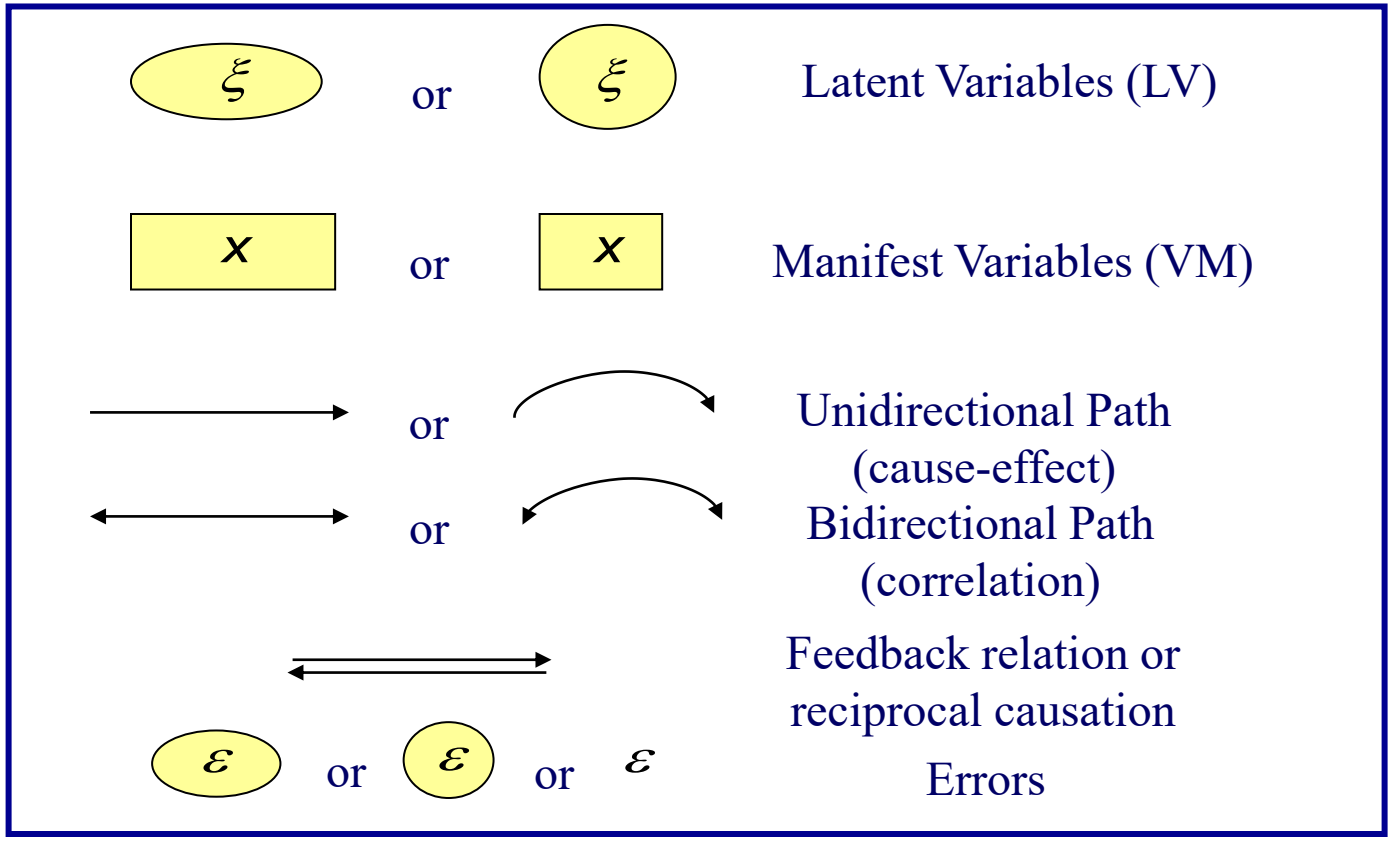
Fattore comune = variabile latente (non direttamente osservabile), inferito attraverso variabili osservate.

Factor Loading = è un coefficiente di correlazione che mostra la forza della relazione tra ciascun fattore comune con la corrispondente variabile manifesta (l’importanza – peso- di ogni variabile nel definire un fattore

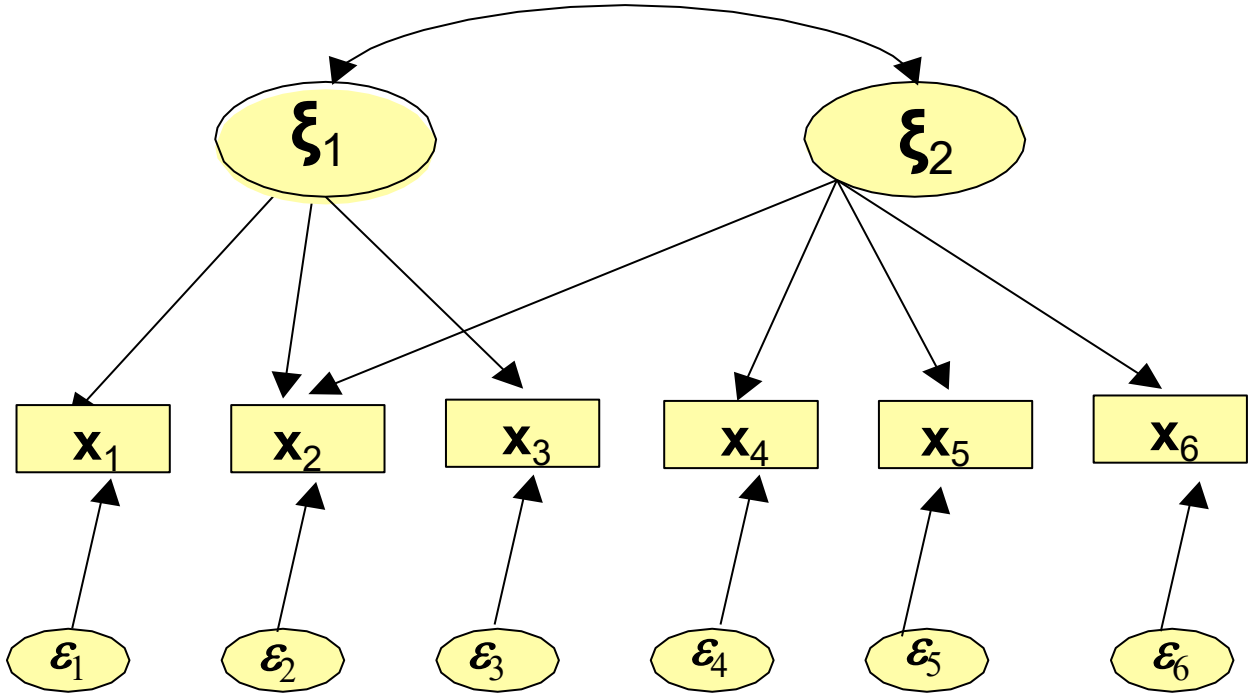
Il modello della diapositiva precedente somiglia solo apparentemente a quello di regressione multipla, infatti q fattori non sono osservabili (non abbiamo valori osservati su queste variabili): tutto ciò che giace a destra dell’equazione è dunque incognito.

Path Diagram

Un modello di analisi fattoriale può essere rappresentato graficamente attraverso il path diagram (mostra le relazione tra tutte le variabili, comprendendo fattori di errori).



Esempio di path diagram



L'Analisi Fattoriale Esplorativa e Confermativa

Analisi Fattoriale Esplorativa

Ha come obbiettivo quello di determinare come e in che modo le variabili manifeste osservate sono legate ad uno o più fattori latenti.

Le relazioni tra le variabili osservate e le variabili latenti sono quindi sconosciute o incerte.

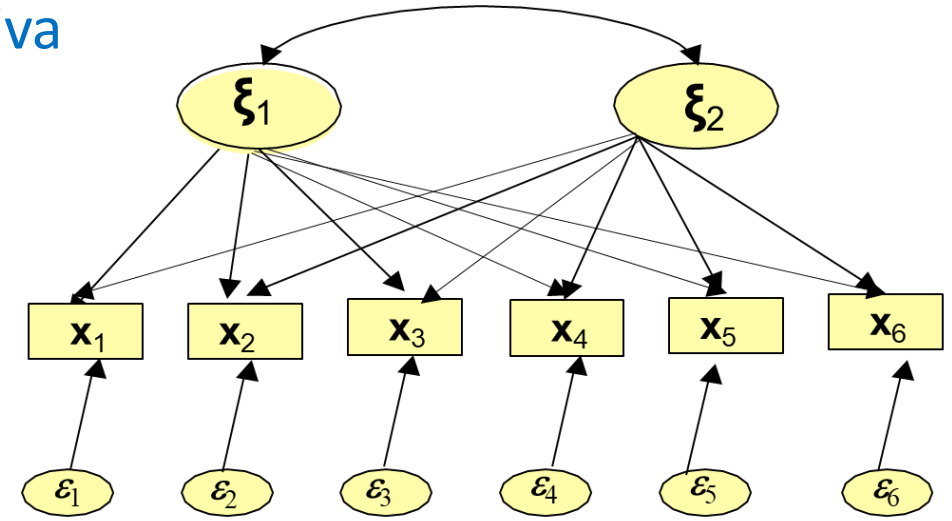
Analisi Fattoriale Confermativa

Ha come scopo quello di testare statisticamente le relazioni causali esistenti tra le variabili manifeste e uno o più fattori latenti.

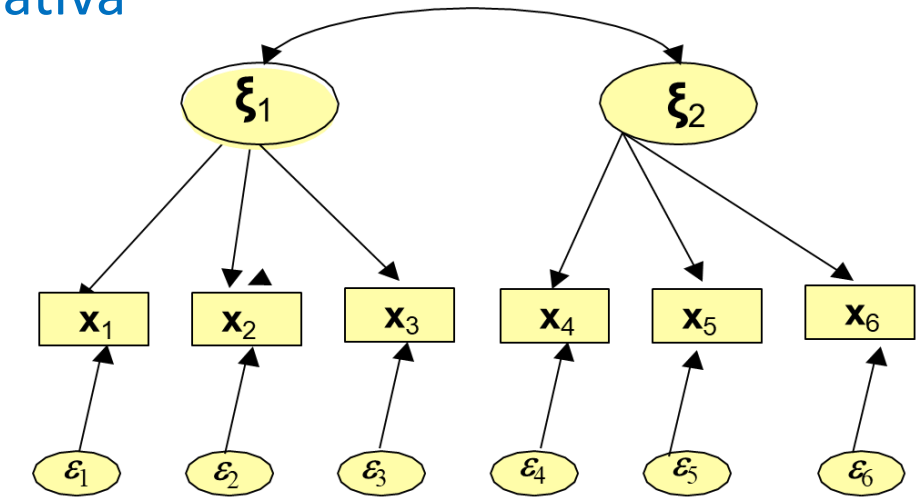
Le relazioni tra le variabili manifeste e le variabili latenti sono quindi note a priori sulla base di teorie o sulla base si esperimenti empirici

L'Analisi Fattoriale Esplorativa e Confermativa

Esempio Analisi Fattoriale Esplorativa



Esempio Analisi Fattoriale Confermativa



NAC
Natural and Artificial Cognition Lab Orazio Miglino
University of Naples Federico II

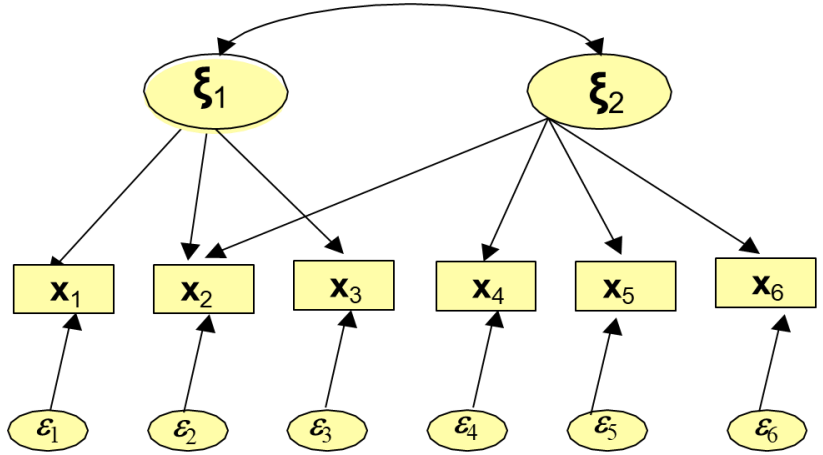
Matrice varianza e covarianza implicata dal modello

La matrice di var/cov S tra le variabili osservate può essere riscritta in termini dei parametri del modello di analisi fattoriale.

Si denota con Σ la matrice si var/cov riprodotta dai parametri del modello

Esempio per il calcolo di un elemento di Σ

$$
\begin{aligned}
x_1 &= \lambda_{11}\,\xi_1 + \varepsilon_1\,, \\
x_2 &= \lambda_{21}\,\xi_1 + \lambda_{22}\,\xi_2 + \varepsilon_2 \\
x_3 &= \lambda_{31}\,\xi_1 + \varepsilon_3 \\
x_4 &= \lambda_{42}\,\xi_2 + \varepsilon_4 \\
x_5 &= \lambda_{52}\,\xi_2 + \varepsilon_5 \\
x_6 &= \lambda_{62}\,\xi_2 + \varepsilon_6
\end{aligned}
$$



$$
\mathbf{S} = \begin{bmatrix}
\mathrm{var}(x_1) & & & & & \\
\mathrm{cov}(x_2,x_1) & \mathrm{var}(x_2) & & & & \\
\mathrm{cov}(x_3,x_1) & \mathrm{cov}(x_3,x_2) & \mathrm{var}(x_3) & & & \\
\mathrm{cov}(x_4,x_1) & \mathrm{cov}(x_4,x_2) & \mathrm{cov}(x_4,x_3) & \mathrm{var}(x_4) & & \\
\mathrm{cov}(x_5,x_1) & \mathrm{cov}(x_5,x_2) & \mathrm{cov}(x_5,x_4) & \mathrm{cov}(x_5,x_4) & \mathrm{var}(x_5) & \\
\mathrm{cov}(x_6,x_1) & \mathrm{cov}(x_6,x_2) & \mathrm{cov}(x_6,x_4) & \mathrm{cov}(x_6,x_4) & \mathrm{cov}(x_6,x_5) & \mathrm{var}(x_6)
\end{bmatrix}
$$

$$
\mathrm{var}(x_3) = \lambda_{31}^2\,\phi_{11} + \mathrm{var}(\varepsilon_3) \qquad \phi_{11} = var(\xi_1)
$$

Metodo di stima del modello

I parametri incogniti del modello fattoriale sono stimati minimizzando la «distanza» tra la matrice di var/cov **S** osservata e la matrice di var/cov Σ riprodotta dai parametri del modello.

I valori stimati dei parametri in Σ saranno tali che:

$$\mathbf{S - \Sigma = min}$$

(la differenza tra le due matrici deve essere più piccola possibile)

Di conseguenza, i metodi stima dei parametri del modello di analisi fattoriale cercano di ricostruire la matrice di var/cov osservata **S** definendo dei fattori comuni che spieghino nel miglio modo possibile la struttura di varianza e covarianza osservata.

Validazione del modello

Alcuni step del processo di validazione:

- Verifica della bontà di adattamento (fit) globale del modello ai dati
- Percentuale di variabilità spiegata dai fattori
- Test statistici sui parametri dei modelli (loadings, varianze degli errori ecc.)

Test sulla bontà di adattamento

Un test di bontà di adattamento del modello ai dati si basa sulla funzione di DISCREPANZA $f\,\mathbf{S},\,\Sigma$, costruita a partire da $(\mathbf{S}-\Sigma)$, che si basa quindi sull'analisi dei residui del modello, cioè degli scarti tra matrice di var/cov osserva e quella implicata dal modello.

Si può dimostrare che:

$$f\ (\mathbf{S},\ \Sigma)\ \approx \chi^2$$

si distribuisce come

Test sulla bontà di adattamento

Il test del Chi-2

$H_0 : S = \Sigma$ ⟶ Buon adattamento del modello ai dati $H_1 : H_1 : S \neq \Sigma$

Regola di decisione:

se p-value > 0,05 allora accettiamo H_0 □ Buon adattamento

Questo test dipende molto dalla numerosità campionaria (N):

- Se N è grande, rischio di rifiutare il modello anche con un buon adattamento ai dati!
- Difficile procedere al confronto tra il fit di modelli di numerosità diversa

Test sulla bontà di adattamento

Il Goodness of Fit Index (GFI)

$$GFI = 1 - \frac{F}{F_{\text{Null}}}$$

Misura la percentuale di varianza e covarianza in **S** spiegata dalla matrice **Σ,** che equivale a testare quanto il modello considerato si adatta meglio di un «modello nullo» (tutti i parametri sono fissati a zero).

- Varia tra 0 e 1, ma in casi estremi può succedere che si osservano valori al di fuori di quest'intervallo

Regola di decisione:

Un modello è accettato se GFI > 0,9

Percentuale di varianza spiegata da fattori

Comunalità

La comunalità, h^2, associata a ciascuna variabile osservata, esprime la proporzione di varianza della variabile spiegata dai fattori comuni corrispondenti.

$$\text{var}(x_3) = \lambda_{31}^2 \, \phi_{11} + \text{var}(\varepsilon_3)$$

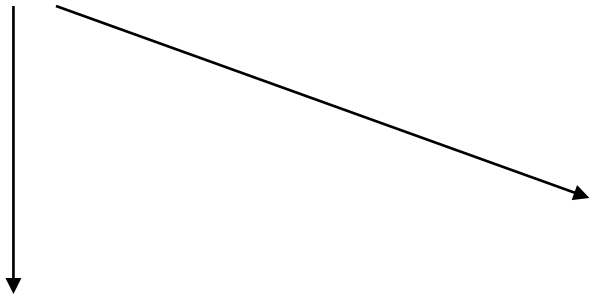
$$\phi_{11} = var(\xi_1)$$

h^2 è proporzionale a questa quantità

Unicità: parte di variabilità totale spiegata dal fattore unico

COMUNALITA’: parte di varianza totale che viene spiegata dai fattori comuni

UNICITA’: parte di varianza totale che spiegata dal fattore unico



Varianza dovuta all’errore di misurazione

Varianza attribuibile a processi che agiscono sistematicamente solo su una variabile (specificità)

Percentuale di varianza spiegata da fattori

Comunalità

Essendo una proporzione varia tra 0 e 1

Più h^2 si avvicina a 1, tanto più i fattori considerati saranno in grado di spiegare la varianza della variabile osservata.

Vanno tendenzialmente tenuti in considerazione variabili che abbiano un valore di comunalità di almeno 0.5.

Analisi fattoriale

- Considerando tutti gli indicatori ($m = p$) possiamo spiegare il 100% della varianza del fenomeno indagato
- La factor analysis consente di identificare $m (< p)$ fattori, capaci di sintetizzare in modo efficiente gli indicatori e di ridurne la complessità;
- i fattori, però, spiegano una quantità di varianza inferiore al 100%
- I fattori, cioè, descrivono la varianza comune tra gli indicatori, ma non la parte di varianza "unica" dei singoli indicatori
- Quindi, con la factor analysis, si decide di "sacrificare" una parte della varianza spiegata a favore di una maggiore semplicità

Analisi fattoriale

- Attraverso l'analisi della matrice delle correlazioni tra gli indicatori viene generato un set di fattori
- Per ogni indicatore, rispetto a ogni fattore, sarà possibile calcolare un coefficiente fattoriale (**factor loading**);
- Più è alto (in valore assoluto) un coefficiente fattoriale, più la variabile latente sarà descritta da quel fattore
- Uno degli output principali della factor analysis è la **matrice dei coefficienti fattoriali**, che presenta gli indicatori sulle righe e i fattori sulle colonne
- Gli indicatori più correlati tra loro mostreranno coefficienti alti sullo stesso fattore

I requisiti minimi

- Dati quantitativi misurati su scale a intervalli o rapporti regolari
- Variabili con distribuzione normale (o almeno non troppo diversa dalla normale) - ASSUNZIONE
- Esclusione dei valori anomali che alterano le correlazioni - outliers
- Più soggetti che variabili (almeno 100): rapporti 1:2 minimo, 1:3

- I fattori (o dimensioni latenti o componenti) non possono superare il numero di variabili osservate
- Il numero di soggetti non può essere inferiore al numero di variabili osservate
- Il numero di soggetti dovrebbe essere elevato (almeno 100-200). La stabilità completa (ripetibilità) si ottiene solo su 3-4000 casi.

Estrazione dei fattori

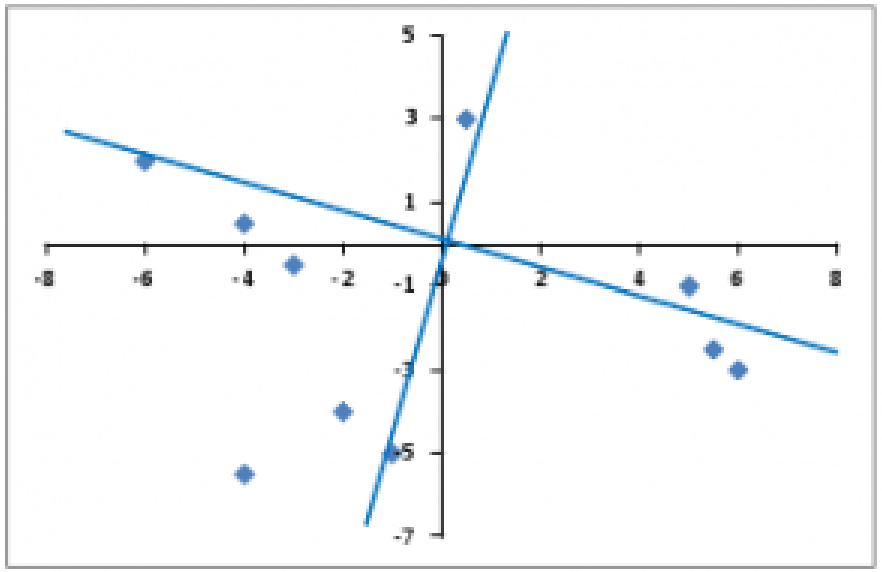
- Ci sono diversi metodi per estrarre i fattori ovvero per stimare la similarita’ tra la matrice di correlazione originale e quella determinata dai fattori:
  - Fattori principali (stima iniziale della comunalità sulla diag.)
  - Massima verosimiglianza
  - Minimi quadrati
- Per alcuni software statistici (SPSS)
  - Componenti principali

Struttura minima

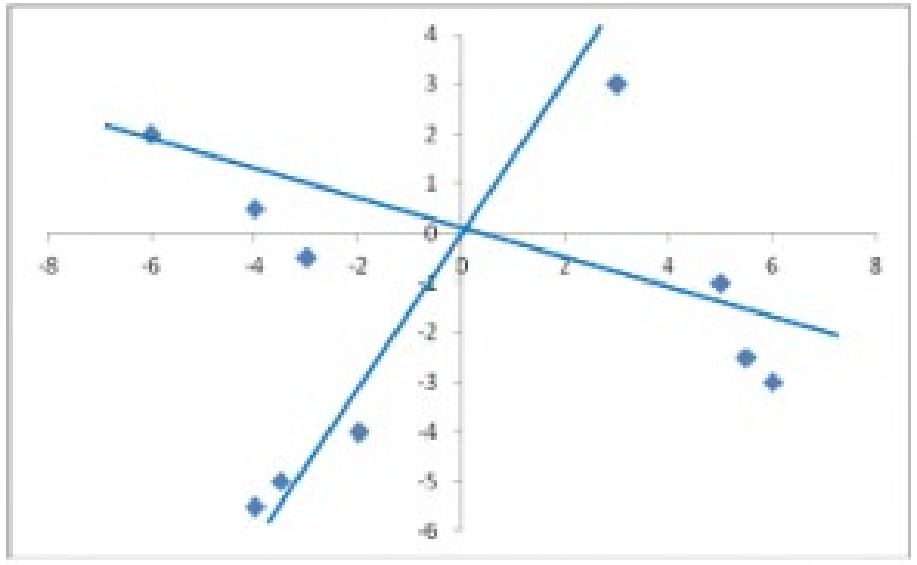
- Lo scopo dell'analisi fattoriale e' quello di trovare la struttura minima di fattori che possa spiegare la matrice di correlazione
- Le soluzioni offerte dall'analisi sono **infinite**
- La rotazione degli assi fattoriali (rotazione dei fattori) per mette di avvicinare o allontanare i fattori dai punteggi osservati
- Attraverso la rotazione si possono trovare soluzioni che abbiano dei *loading* piu' marcati

Rotazioni

Ortogonale



Obliqua



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 Varimax.
  - Oblimin (obliquità minima): permette di fissare l'inclinazione degli assi e quindi le loro intercorrelazioni

La rotazione ortogonale

- La rotazione 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

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:

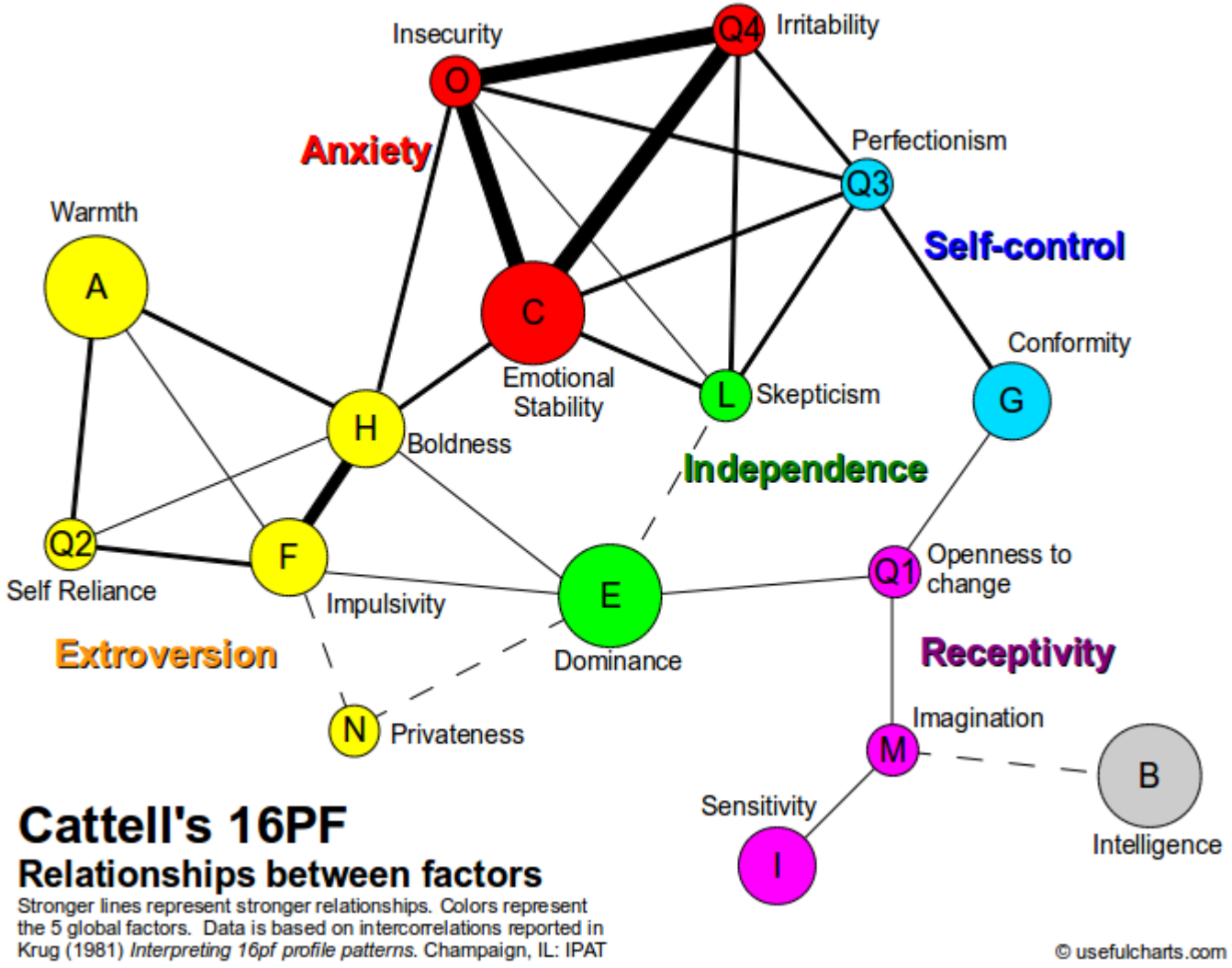
items x10, x8, x3, x4, x5

Fattore 2:

items x6, x7, x9, x1, x2

NAC
Natural and Artificial Cognition Lab Orazio Miglino
University of Naples Federico II

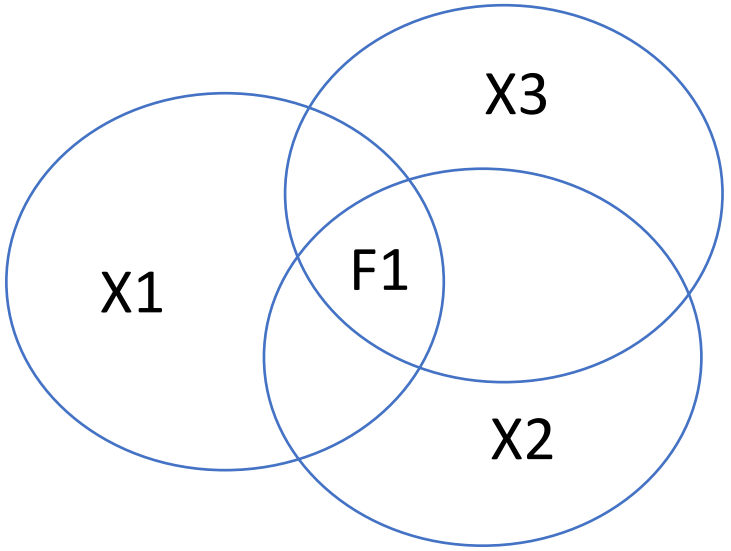
Costrutti e fattori



Analisi fattoriale

X1, X2, X3 = punteggi

F1 = fattore



- I fattori rappresentano la parte di covarianza condivisa da più indicatori, variabili
- A livello formale, i fattori vengono definiti analizzando la matrice delle correlazioni tra gli indicatori: gli indicatori che mostrano correlazioni alte tra di loro e basse rispetto ad altri indicatori genereranno un fattore

Riprendiamo la PCA

L’analisi fattoriale e l’analisi delle componenti principali hanno diversi aspetti in comune. Infatti, il software statistico SPSS, per esempio, tratta entrambi i metodi nella stessa procedura.

Tuttavia, i due metodi hanno differenze sostanziali e non sono due approcci diversi per la stima dei parametri di uno stesso modello

Riprendiamo la PCA

La PCA è un metodo di statistica multivariata che ha l'obiettivo di ridurre la complessità presente in una matrice di dati in maniera tale da esprimere la sua struttura in un numero ridotto di dimensioni (metodo di riduzione della dimensionalità), eliminando la ridondanza di informazioni nei dati.

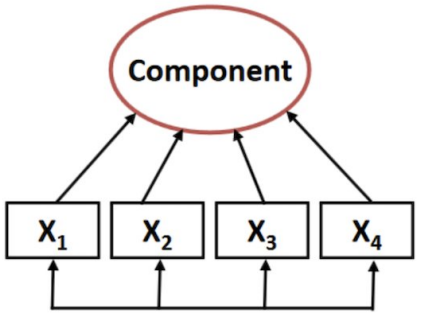
Tuttavia, a differenza dell'AF, la PCA non si basa su un **modello** che richiede una serie di assunzioni. La PCA non prevede affatto un modello sottostante.

La PCA individua delle particolari trasformazione delle variabili osservate, le componenti principali.
Le componenti principali costruite devono essere tra loro incorrelate e spiegano gran parte della variabilità totale.

L'ABBIAMO VISTO…

PCA vs. ANALISI FATTORIALE

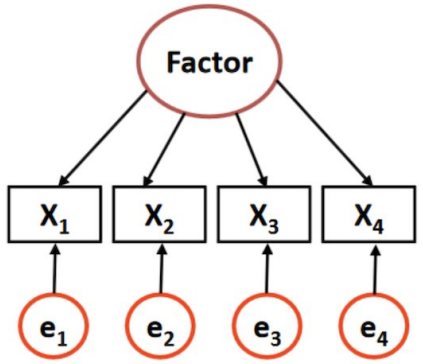
Principal Component Analysis



Linear combination that explains most of the total variance

The observable variables cause the component

Factor Analysis



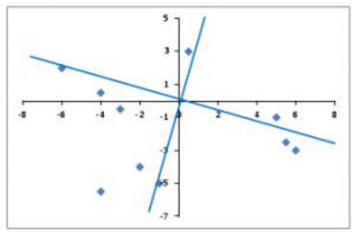
Unobservable (common) factors that explain variable variances and covariances

The factors cause the observable variables

If error term, e, is small, then both tools are roughly similar

PCA	FA
• Observed variables are relatively error-free. • Unobserved latent component is a perfect linear combination of its variables. • Ideal if data reduction and composite- construction are the goals.	• Error represents a portion of the total variance. • The observed variables are only indicators of the latent factors. • Ideal in well-specified theoretical applications.

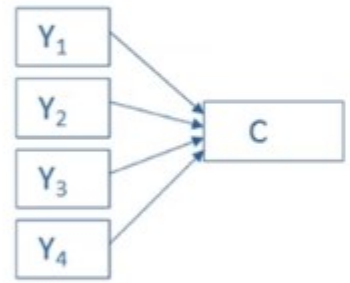
PCA - Principal Component Analysis	*PFA - Principal Factor Analysis*
• PCA is purely a descriptive method (a simple transformation of the data to simplify description • uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components** • the components are actual orthogonal linear combinations that maximize the total variance • varimax – orthogonal • transformation of the original variables into a new set that are mutually orthogonal. • The first new component maximizes the variance • answers the question, "What linear combination of my variables has largest variance? • 1. Run principal component analysis If you want to simply reduce your correlated observed variables to a smaller set of important independent composite variables	• FA is a formal statistical method (with statistical assumptions and tests that allow inference to a larger population). • factor solutions can be rotated if useful for interpretation and theoretically logical • the factors are linear combinations that maximize the shared portion of the variance--underlying "latent constructs" • oblique – oblimin • starts with a model about how the variables are related and where variation comes from in the data • 2. Run factor analysis if you assume or wish to test a theoretical model of latent factors causing observed variables



- ❖ Rotations that allow for correlation are called **oblique rotations**; rotations that assume the factors are not correlated are called **orthogonal rotations**. Our graph shows an orthogonal rotation.
- ❖ the angle between the two factors is now smaller than 90 degrees, meaning the factors are now correlated

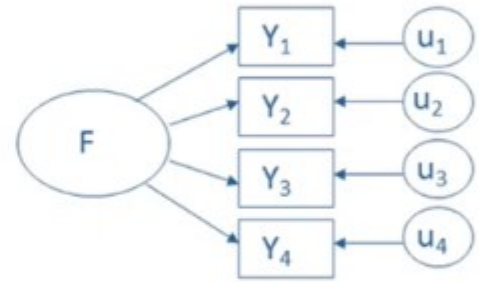
What is PCA?

Principal Component Analysis



$$C = w_1(Y_1) + w_2(Y_2) + w_3(Y_3) + w_4(X_4)$$

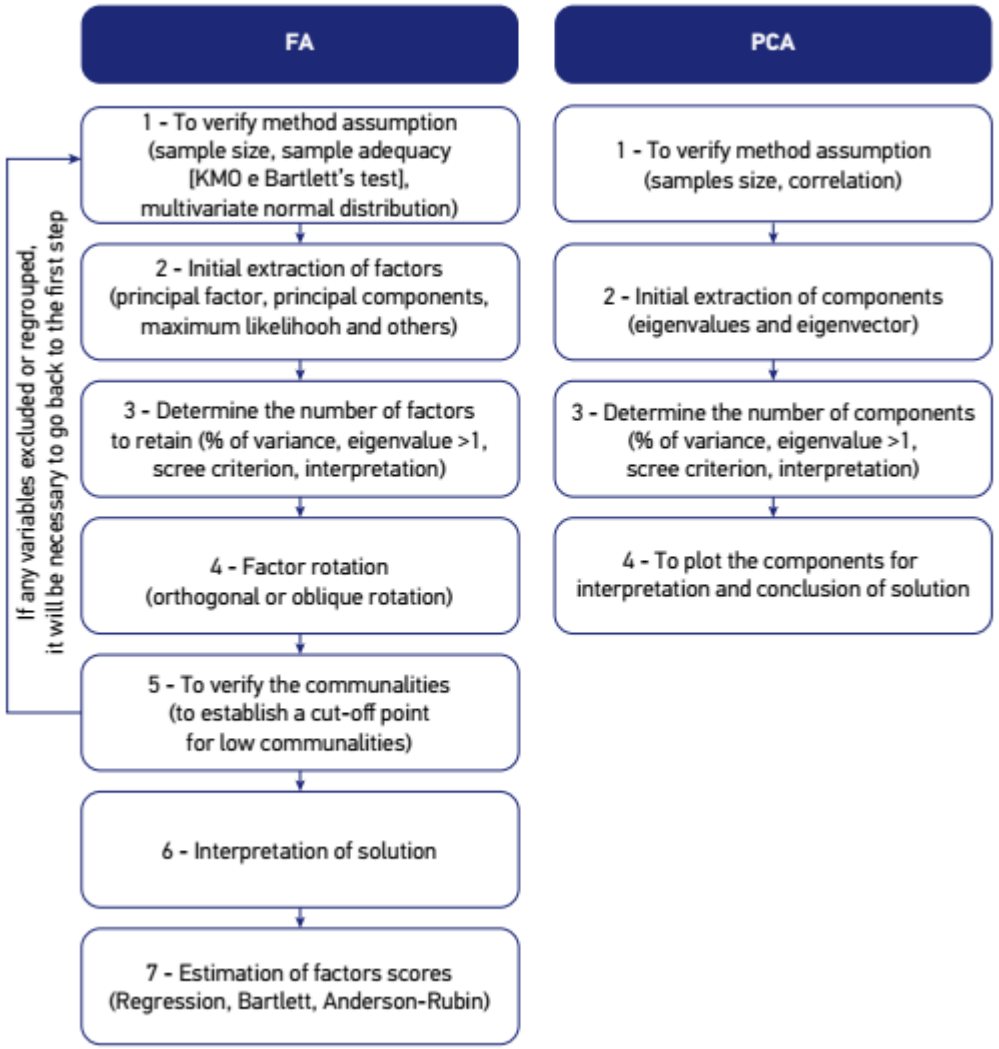
Factor Analysis



$$Y_1 = b_1*F + u_1$$
$$Y_2 = b_2*F + u_2$$
$$Y_3 = b_3*F + u_3$$
$$Y_4 = b_4*F + u_4$$

PCA vs. ANALISI FATTORIALE

Factor analysis	Principal component analysis
Number of factors predetermined	Number of components evaluated ex post
Many potential solutions	Unique mathematical solution
Factor matrix is estimated	Component matrix is computed
Factor scores are estimated	Component scores are computed
More appropriate when searching for an underlying structure	More appropriate for data reduction (no prior underlying structure assumed).
Factors are not necessarily sorted	Factors are sorted according to the amount of explained variability
Only common variability is taken into account	Total variability is taken into account
Estimated factor scores may be correlated	Component scores are always uncorrelated
A distinction is made between common and specific variance	No distinction between specific and common variability
Preferred when there is substantial measurement error in variables	Preferred as a preliminary method to cluster analysis or to avoid multicollinearity in regression
Rotation is often desirable as there are many equivalent solutions	Rotation is less desirable, unless components are difficult to interpret and explained variance is spread evenly across components

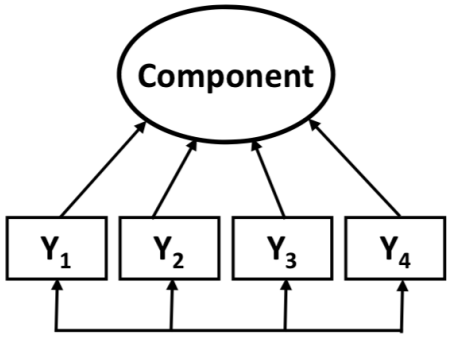


Exploratory Factor Analysis vs. Principal Components Analysis

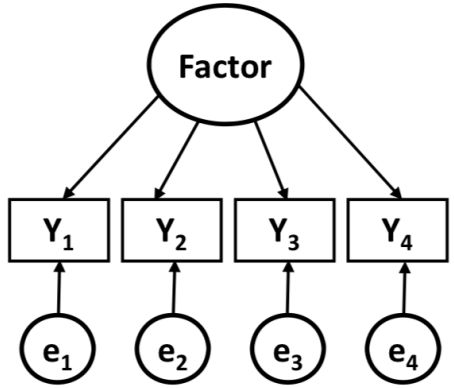
Exploratory Factor Analysis	Principal Components Analysis
Researcher assumes that there is a smaller set of unobserved constructs that underlie the measured variables	Researcher is trying to derive statistically (using variances) a relatively small number of variables to use to convey as much of the information in the measured variables as possible
Directed at understanding the relationships among variables by understanding underlying constructs	Used to enable researcher to use fewer variables to obtain the same information as would be gathered with more variables
Used when there is a theory about how the variables fit together	Used when researcher is looking to use fewer variables to provide the same information

PCA vs. ANALISI FATTORIALE

PCA vs. EFA/CFA

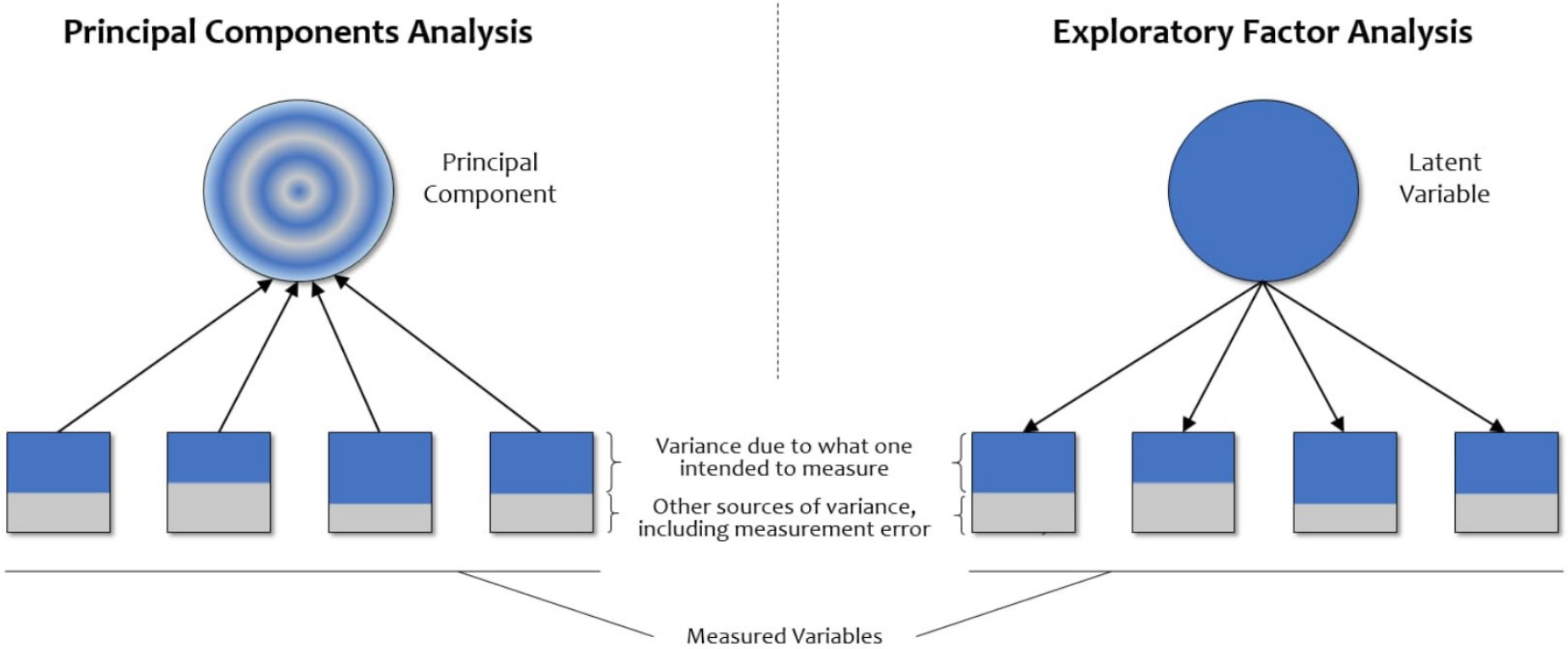


This is not a testable measurement model, because how do we know if we've combined stuff "correctly"?



This IS a testable measurement model, because we are trying to predict the observed covariances between the indicators by creating a factor – **the factor IS the reason for the covariance.**

PCA vs. ANALISI FATTORIALE



GRAZIE PER L'ATTENZIONE!

NAC
Natural and Artificial Cognition Lab Orazio Miglino
University of Naples Federico II