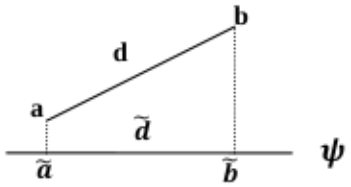
L'Analisi in Componenti Principali (a cura del Dott. Mario Candela):

L'analisi fattoriale è un insieme di tecniche aventi tutte la medesima logica, cambia l'approccio poiché cambia il tipo di dati da analizzare, nel nostro caso immaginiamo una matrice di dati quantitativi.
L'obiettivo di una tecnica dell'analisi fattoriale è di ridurre la dimensionalità del problema, quindi schiacciare lo spazio di analisi in uno spazio a dimensioni ridotte, più facilmente interpretabile.
Il motivo principale per cui nasce l'analisi fattoriale è di costruire indicatori complessi, l'idea è che abbiamo una serie di variabili statistiche, queste ultime esprimono parte della stessa informazione, ovvero sono tra loro correlate.
La domanda a cui risponde l'analisi fattoriale è: "Possiamo eliminare queste variabili iniziali, le quali risultano inefficienti, esprimendo parte della stessa informazione, con una sola variabile o con 2 o 3 variabili? Ovvero, possiamo ridurre il numero di variabili da $\boldsymbol{p}$ variabili molto numerose a un numero ridotto di variabili, esprimendo comunque la medesima informazione (conservando la struttura dei dati) e conservando la variabilità del problema?"
La variabilità è informazione e dobbiamo spiegarla.
Questo fa l'analisi fattoriale.
Quando le nostre variabili sono numeriche, parlando di metodi fattoriali, facciamo riferimento all'Analisi in Componenti Principali.
Il problema, dunque, consiste nel trovare i pesi migliori, e questi ultimi sono quelli che ci danno la migliore $\boldsymbol{\psi}$.
Alla fine in un metodo statistico si deve trovare la soluzione sulla base di un criterio di ottimo, il criterio di ottimo nel nostro caso è che vogliamo conservare quanta più informazione contenuta nei dati, ovverosia la variabilità; quindi, vogliamo che la nostra variabile di sintesi conservi quanto più possibile la variabilità contenuta nella nube originaria dei punti.
Significa che se 2 individui hanno valori di diversi di $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$, avranno anche punteggi diversi della variabile di sintesi $\boldsymbol{\psi}$; invece, se hanno valori simili di $\boldsymbol{x}$ devono avere punteggi simili della $\boldsymbol{\psi}$.
La 1ª variabile latente che tiriamo fuori coi metodi di analisi fattoriale è sempre quella nella direzione di massimo allungamento della nube, ossia quella che cattura la massima variabilità.
Quindi, andiamo a trovare quei pesi ($\boldsymbol{u}$) che fanno sì che la varianza della nostra $\boldsymbol{\psi}$ sia massima

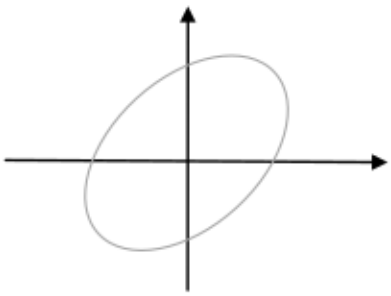
$$Var(\psi) = max$$

, questo è il criterio-guida nella scelta dei pesi; così fa l'ACP su variabili quantitative, mentre per l'ACM si tratta di variabili qualitative.
Nell'analisi fattoriale, nel momento in cui schiacciamo lo spazio, necessariamente un'azione di riduzione della dimensionalità (schiacciamento dello spazio del problema) comporta una distorsione nelle distanze tra gli individui, ossia tra le posizioni precedenti rispetto a quelle che otteniamo nello spazio proiettato.
Ad esempio, consideriamo 2 punti, $\boldsymbol{a}$ e $\boldsymbol{b}$, e $\boldsymbol{d}$ è la distanza originaria tra questi 2 punti; proiettando tali punti per ottenere la variabile di sintesi, $\boldsymbol{\psi}$, la distanza sarà $\tilde{\boldsymbol{d}}$.



Nell'ACP le distanze si misurano con la distanza euclidea (trattandosi di uno spazio euclideo, considerando il segmento più breve che unisce i 2 punti); dunque, la distanza originaria rappresenta sempre l'ipotenusa, mentre la distanza proiettata è uno dei 2 cateti e, per tale motivo, è più piccola dell'ipotenusa e, quindi, della distanza originaria.
La distanza proiettata è sempre più piccola della distanza originaria.
Massimizzare la varianza di $\boldsymbol{\psi}$ significa la medesima cosa del volere che la distanza proiettata sia la più grande possibile, ossia quanto più vicina a quella originaria, conservando più o meno le posizioni degli individui originari.
Quindi, il problema possiamo vederlo da un punto di vista geometrico, ossia la misura di $\boldsymbol{\psi}$ che vogliamo è quella che schiaccia meno la distanza tra gli individui; il che corrisponde al volere la misura di $\boldsymbol{\psi}$ che conserva la variabilità, poiché schiacciando gli individui diminuisce la variabilità.

Ad esempio, se $\boldsymbol{a}$ fosse il punto medio, dunque la distanza tra $\boldsymbol{a}$ e $\boldsymbol{b}$ è la varianza, perché la distanza si calcola come $(valori\ di\ \boldsymbol{b} - valori\ di\ \boldsymbol{a})^2$ se ragioniamo in termini di media, mentre la distanza tra $\boldsymbol{a}$ e $\boldsymbol{b}$ è la devianza se ragioniamo in termini di somma; dunque, la distanza tra $\widetilde{\boldsymbol{a}}$ e $\widetilde{\boldsymbol{b}}$ è la varianza o la devianza dei punti proiettati.
Nel momento in cui vogliamo massimizzare la variabilità di $\boldsymbol{\psi}$, vogliamo massimizzare il segmento $\widetilde{\boldsymbol{d}}$, ossia le differenze di ogni punto proiettato rispetto alla media; se, invece, di confrontare un punto confrontiamo tutti i punti rispetto alla media è chiaro che significhino la stessa cosa.
Il teorema di Huygens, difatti, prevede che massimizzare la distanza proiettata, facendo sì che la variabile conservi le differenze tra gli individui, è la medesima cosa del dire che la sua variabilità è massima.
Nell'ACP, siccome vogliamo fare la somma pesata di variabili, è necessario che le variabili abbiano la medesima unità di misura, ossia tali variabili devono avere la stessa media e stessa varianza (se abbiamo una variabile con varianza molto grande e una variabile con varianza molto piccola, non possiamo confrontare le variazioni delle variabili).
Per poter confrontare variabili, per poter fare una somma delle variabili pesate le variabili devono essere standardizzate, eliminando cioè l'unità di misura di partenza trasformandola in un'unità di misura standard, in $\boldsymbol{\sigma - esimi}$; fatto ciò le variabili sono confrontabili perché hanno tutte media 0 e varianza 1.
Quindi, quando ragioniamo nell'ACP la 1ª cosa che facciamo è che le $\boldsymbol{x}$ di partenza sono tutte standardizzate, e avendo tali variabili tutte media 0, la nube è centrata nell'origine degli assi, poiché la media è l'origine



Se facciamo la distanza dall'origine stiamo calcolando la varianza.
Nell'ACP vogliamo massimizzare la varianza di $\boldsymbol{\psi}$ e, dunque, vogliamo trovare quegli $\boldsymbol{u}$ che massimizzano la varianza di $\boldsymbol{\psi}$; inoltre, quest'ultima è la somma pesata delle variabili originarie per i pesi, in forma matriciale

$$\Psi = x \cdot u$$

e la varianza di $\boldsymbol{\psi}$ è uguale a

$$Var(\psi) = E\big(\psi - media(\psi)\big)^2 =$$

$$= [(x \cdot u) - 0]^2 =$$

$$= (x \cdot u)^2$$

Alla fine il criterio che vogliamo massimizzare, in forma matriciale, è

$$\max Var(\psi) = u' \cdot x'x \cdot u$$

ossia vogliamo trovare quelle $\boldsymbol{u}$ che moltiplicate prima e dopo per $\boldsymbol{x'x}$, dove $\boldsymbol{x}$ è la matrice dei dati (ossia tutte le nostre variabili o i nostri item), siano massime.

Abbiamo $\boldsymbol{x}$, la matrice dei dati originari, $\boldsymbol{n}$ individui e $\boldsymbol{x_p}$ variabili originarie, queste ultime sono standardizzate, avendo

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_p}$
$\boldsymbol{s_1}$	$\frac{(x_1-\mu_1)}{\sigma_1}$	…	$\frac{(x_p-\mu_1)}{\sigma_1}$
$\boldsymbol{s_2}$	…	…	…
$\boldsymbol{s_n}$	$\frac{(x_1-\mu_n)}{\sigma_n}$	…	$\frac{(x_p-\mu_n)}{\sigma_n}$

Quando facciamo questa matrice per la sua trasposta, moltiplichiamo riga per colonna, ad es. 1ª riga per 1ª colonna; quindi, la matrice che ne deriva è una matrice in cui la dimensione sarà $\boldsymbol{p} \times \boldsymbol{p}$ dove le $\boldsymbol{p}$ sono le variabili.
Il prodotto matriciale si fa 1ª riga per 1ª colonna, ogni elemento l'uno per l'altro facendo, poi, la somma; quindi, nella cella avremo la sommatoria

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_p}$
$\boldsymbol{x_1}$	$\sum\left[\frac{(x_1-\mu_1)}{\sigma_1}\cdot\frac{(x_1-\mu_1)}{\sigma_1}\right]=\mathbf{1}$	$\sum\left[\frac{(x_1-\mu_1)}{\sigma_1}\cdot\frac{(x_2-\mu_2)}{\sigma_2}\right]$	…
$\boldsymbol{x_2}$	…	$\mathbf{1}$	…
$\boldsymbol{x_p}$	…	…	$\mathbf{1}$

Questa è la matrice di correlazione.
$\left[\frac{x_1\cdot(-\mu_1)}{\sigma_1}\cdot\frac{(x_1-\mu_1)}{\sigma_1}\right]$ è il coefficiente di correlazione di Pearson se avessimo 2 variabili diverse, $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$; difatti, la formula del coefficiente di correlazione di Pearson, al numeratore abbiamo il prodotto degli scarti e al denominatore il prodotto delle deviazioni standard

$$\rho = \frac{(x-\mu_x)\cdot(y-\mu_y)}{\sigma_x\cdot\sigma_y} \Rightarrow coeff.correlazione\ Pearson$$

$$(x-\mu_x)\cdot(y-\mu_y) = \boldsymbol{\sigma_{XY}}\ \{\text{covarianza tra X e Y}\}$$

Ovviamente, se la variabile è la medesima la correlazione è pari a 1, se le variabili sono diverse abbiamo, ad esempio, $\boldsymbol{\rho(x_1, x_2)}$.
Tutto il problema dell'analisi fattoriale è di trovare quei pesi che, moltiplicati per la matrice di correlazione, ci danno la varianza di $\boldsymbol{\psi}$; difatti, i pesi ricostruiscono la matrice di correlazione.
In una matrice di correlazione, se il fenomeno è multi-dimensionale, ci potrebbero essere più variabili latenti e non solo una, ossia avremmo più $\boldsymbol{\psi}$; poiché, una sola $\boldsymbol{\psi}$ può non bastare a spiegare tutta l'informazione.
Una volta trovata la 1ª $\boldsymbol{\psi}$, potremmo voler trovare la 2ª $\boldsymbol{\psi}$, ossia quella che massimizza la variabilità al netto della variabilità della 1ª $\boldsymbol{\psi}$.
Aggiungiamo, dunque, un altro vincolo nel nostro massimo, che la 2ª $\boldsymbol{\psi}$ resti incorrelata con la 1ª $\boldsymbol{\psi}$, ovvero che il prodotto del 1° vettore di pesi col 2° vettore di pesi sia pari a 0; perché, rispetto alle variabili originarie tali variabili devono essere efficienti (efficienti in questo caso vuol dire incorrelate), ovvero ogni $\boldsymbol{\psi}$ deve spiegare parti diverse dell'informazione.
Da questo sistema viene fuori una serie di vettori di pesi fino a $\boldsymbol{p}$, perché potremmo trovare tante variabili $\boldsymbol{\psi}$ quante sono le variabili originarie; l'obiettivo è trovarne meno però, se continuassimo a massimizzare, potremmo trovarne tante quante le variabili originarie, tuttavia le $\boldsymbol{\psi}$ sono efficienti rispetto alle variabili originarie perché sono incorrelate tra loro.
Ogni $\boldsymbol{\psi}$ è un vettore di pesi, il miglior vettore di pesi è quello che massimizza la variabilità; quindi, a questo vettore di pesi si aggiunge un'informazione, il coefficiente $\boldsymbol{\lambda}$, che è la varianza di $\boldsymbol{\psi}$

$$\lambda = Var(\psi)$$

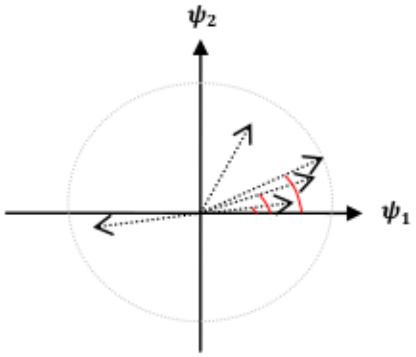
Immaginiamo ad esempio,

Vettori pesi	
$\boldsymbol{u_1}$	$\boldsymbol{\lambda_1 > \lambda_2}$
$\boldsymbol{u_2}$	$\boldsymbol{\lambda_2 > \lambda_3}$
$\boldsymbol{u_3}$	$\boldsymbol{\lambda_3 > \lambda_4}$
…	…
$\boldsymbol{u_p}$	$\boldsymbol{\lambda_p}$

Avendo tutti i vettori di pesi, il 1° vettore di pesi, quello migliore, che ci dà la 1ª $\boldsymbol{\psi}$ (la più importante), è quello che ha $\boldsymbol{\lambda}$ più grande, ossia la varianza più grande. Questa è la 1ª soluzione.
Se la 1ª $\boldsymbol{\psi}$ spiega tutta l'informazione interessante nella nube ci fermiamo, trovando una variabile latente; invece, se 1ª $\boldsymbol{\psi}$ non spiega tutta l'informazione, prendiamo il 2° vettore e otteniamo un'altra $\boldsymbol{\lambda}$, ottenendo la 2ª componente principale ($\boldsymbol{\psi_1}$), poi possiamo ottenere la 3ª e così via.
Per scegliere quante componenti principali ci servono per spiegare una parte sufficiente della informazione contenuta nella nube, consideriamo che la somma dei $\boldsymbol{\lambda}$ è nota e pari a $\boldsymbol{p}$; poiché, se ogni variabile originaria ha varianza 1 e abbiamo 10 variabili, la varianza totale della nube originaria è 10.
Se costruiamo tante variabili latenti $\boldsymbol{\psi}$ quante sono le variabili originarie, ricostruiamo tutta l'informazione della nube, nel caso dell'esempio con $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$ possiamo trovare 2 $\boldsymbol{\psi}$, la 1ª $\boldsymbol{\psi}$ è quella della massima direzione di allungamento della nube e la 2ª $\boldsymbol{\psi}$ spiega il resto dell'informazione e, essendo incorrelata con la 1ª, forma un angolo retto con la 1ª componente principale; quindi, la 2ª possibile $\boldsymbol{\psi}$ è per forza perpendicolare, perché abbiamo sostituito le 2 variabili originarie, tra loro correlate e, dunque, inefficienti, con 2 variabili efficienti.
Probabilmente, per spiegare l'informazione interessante contenuta nel fenomeno basta una sola $\boldsymbol{\psi}$.

La somma delle $\boldsymbol{\lambda}$ fa $\boldsymbol{p}$ perché se le $\boldsymbol{\lambda}$ sono standardizzate e hanno varianza pari a 1, la loro somma, nel nostro esempio, è pari a 10.
Immaginiamo che $\boldsymbol{p = 10}$, ossia abbiamo 10 variabili originarie, e che $\boldsymbol{\lambda_1 = 7}$, ciò vuol dire che il coefficiente $\boldsymbol{\lambda_1}$ sintetizza tanta informazione quanto prima era contenuta in 7 variabili originarie, dunque da sola $\boldsymbol{\psi_1}$ spiega il 70% dell'informazione originaria $\left(\frac{7}{10}\right)$; se non ci basta il 70%, prenderemo anche la variabile latente successiva $\boldsymbol{\psi_2}$, ad es. $\boldsymbol{\lambda_2 = 2}$, spiega 2 volte quanto spiegava una singola variabile, ottenendo così il 90% dell'informazione originaria $\left(\frac{2}{10} + \frac{7}{10}\right)$, però originariamente servivano 10 variabili (ossia 10 indicatori statistici).
Il 90% raggiunto con sole 2 variabili di sintesi ci soddisfa, quindi vuol dire che alle variabili $\boldsymbol{x}$ originarie sostituiremo queste 2.
Questo è un criterio.
Un altro criterio è di prendere le variabili che hanno $\boldsymbol{\lambda > 1}$, siccome $\boldsymbol{\lambda}$ è la varianza di $\boldsymbol{\psi}$ e la varianza di una variabile originaria è 1, tutti i $\boldsymbol{\lambda > 1}$ vuol dire che le $\boldsymbol{\psi}$ sono variabili che sintetizzano più di una variabile originaria, mentre le $\boldsymbol{\psi}$ che hanno $\boldsymbol{\lambda < 1}$ sintetizzano meno di una variabile originaria, dunque non sono efficienti.

Per l'ACP abbiamo un grafico di questo tipo,



osserviamo che le variabili originarie $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, e $\boldsymbol{x_3}$ sono tra loro fortemente correlate, avendo tutte la medesima direzione, e sono correlate con $\boldsymbol{\psi_1}$; quindi, valori alti di $\boldsymbol{\psi_1}$ stanno sintetizzando $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, e $\boldsymbol{x_3}$.

Anche un'altra variabile, nell'altro quadrante, è fortemente correlata con $\boldsymbol{\psi_1}$, però in maniera negativa; quindi, gli individui che hanno un valore elevato di $\boldsymbol{\psi_1}$ hanno bassi valori di $\boldsymbol{\psi_2}$ e viceversa.
Le variabili originarie che formano un angolo di quasi 90° con $\boldsymbol{\psi_1}$ sono incorrelate con quest'ultima e sono sintetizzate da $\boldsymbol{\psi_2}$.
Gli angoli che le variabili formano sono le correlazioni, sono i pesi ($\boldsymbol{u}$) che diamo alle variabili, ossia quanto queste variabili sono correlate rispetto alla variabile di sintesi $\boldsymbol{\psi}$; difatti, quella con la maggiore correlazione avrà maggior peso.
Dal grafico dell'ACP vogliamo 2 informazioni, l'angolo ci dice il grado di correlazione con $\boldsymbol{\psi}$ e la lunghezza del segmento di una variabile ci dice quanto di quella variabile è sintetizzata da $\boldsymbol{\psi}$.