L'analisi fattoriale

Scopo dell'analisi fattoriale e' quello di identificare alcune variabili latenti (**fattori**) in grado di spiegare i legami, le inter-relazioni e le dipendenze tra le variabili statistiche osservate x_1 ,x_2, …. x_k. Per esempio immaginiamo di avere i punteggi (o voti) in diverse materie scolastiche(x_j, j=1,…,9) : italiano, inglese, latino, filosofia, storia, matematica, fisica, chimica, geografia astronomica. Potremmo supporre che il rendimento scolastico sia legato a due fattori: attitudine scientifica e attitudine umanistica; ovvero che vi siano due fattori non osservabili (f_h , h=1,2) che determinano i legami e le inter relazioni tra le x_j. Si assume anche che vi siano dei fattori specifici (u_j): attitudini in una certa materia.

In termini piu' generali supponiamo sia:
$X_j = \mu_j + \lambda_{j1} f_1 + \ldots + \lambda_{jm} f_m + u_j$ $\qquad$ $j = 1,2,\ldots,k$
f_1 , …, f_m sono i "fattori comuni" (m<k), u_j sono i fattori specifici, λ_{jh} ,h = 1,…,m sono dei parametri incogniti (*pesi fattoriali*).
In sostanza il modello di analisi fattoriale postula che le variabili osservabili x_j (j=1,…,k) siano il risultato dell'azione di m (m<k) fattori comuni $f_1,\ldots,f_m$ inosservabili e di fattori specifici u_j (j=1,…,k) anch'essi non osservabili e di natura residuale.
Usualmente le ipotesi che si fanno sono:

- I fattori comuni oltre ad avere valore atteso nullo, sono tra loro incorrelati ed hanno varianza unitaria.
- I fattori specifici oltre ad avere valore atteso nullo sono incorrelati ed hanno varianza pari a ψ_j .
- I fattori comuni sono incorrelati con i fattori specifici $\forall h$ = 1,…,m $\forall j$=1,…,k.

Da tutto cio' discende che:

$$\mathrm{var}(x_j) = \Sigma_h \lambda^2_{jh} + \psi_j = c_j + \psi_j$$

(dove si e' posto $\Sigma_h \lambda^2_{jh} = c_j$, in genere denominata *comunalità*),

$$\text{cov}(x_j,x_t) = \Sigma_h \lambda_{jh}\lambda_{th}$$

La struttura di dipendenza lineare tra le variabili osservate e' interamente dovuta ai fattori comuni.

Stima dei λ_{jh} (*factor loadings* o *pesi fattoriali* o, nel linguaggio dell'output di spss: components)
Si standardizzano le x_j e si lavora sulla matrice di correlazione R (che sarebbe la matrice delle varianze delle x_j standardizzate).

- Se vogliamo tanti fattori quante variabili (m=k) si estraggono gli autovalori ed i corrispondenti autovettori di R. Gli m vettori dei pesi fattoriali saranno dati (a meno di un coefficiente di proporzionalità) dagli m autovettori della matrice R. Ciascun *component* $\boldsymbol{\lambda}_h$ contiene le correlazioni tra il fattore comune f_h e le k variabili x. In altri termini il generico elemento λ_{jh} di $\boldsymbol{\lambda}_h$ e' il coefficiente di correlazione tra x_j e f_h. Questo equivale esattamente a fare un'analisi delle componenti principali delle variabili x_j e a prendere come output non gli autovettori bensi' le correlazioni tra la h-esima componente principale e le diverse variabili x_j , j=1,…,k.
- Se vogliamo un numero di fattori m <k , dobbiamo ipotizzare che nel modello esistano anche i fattori specifici. Pertanto la matrice dei dati da cui estraiamo gli autovalori ed i corrispondenti autovettori sarà la matrice di correlazione "ridotta", ovvero: $R_0 = R-\Psi_0$. Ψ_0 e' la matrice (diagonale) che contiene le stime iniziale delle varianze dei fattori specifici. Essendo – dopo la standardizzazione - $\text{var}(x_j) = c_j + \psi_j = 1$, la matrice di correlazione ridotta e' una matrice di correlazione che contiene sulla diagonale principale non piu' tutti 1 bensi' le stime preliminari (c_{j0})delle comunalità. A differenza di R, R_0 non e' una matrice definita semi-positiva, ed alcuni suoi autovalori potranno essere negativi. Sia p il numero di autovalori positivi. Possiamo porre m=p. Possiamo ricavarci λ_{jh}, e tramite $\Sigma_h \lambda^2_{jh} = c_j$ una nuova stima della comunalità e

quindi degli elementi diagonali della matrice ridotta (R_1 adesso). Estrarre gli autovalori positivi (o anche solo quelli >1) e trovare m fattori comuni.

- Di solito la stima preliminare della comunalità c_j (e quindi direttamente della matrice ridotta) viene fornita o dal quadrato del coefficiente di correlazione multipla della j-esima variabile con tutte le altre, o dal piu' grande dei coefficienti di correlazione tra la j-esima variabile e ciascuna delle altre.
- Cosi' come avviene nell'analisi delle componenti principali, per scegliere il numero m di fattori comuni da utilizzare per l'analisi ci si basa sugli autovalori della matrice di correlazione (o di quella ridotta). Il rapporto tra l'autovalore associato all'h-esimo fattore comune (= somma dei quadrati dei loadings, ovvero = $\Sigma_j \lambda^2_{jh}$) e la traccia della matrice di correlazione ci indica la quota della variabilità complessiva "spiegata" dall'h-esimo fattore.

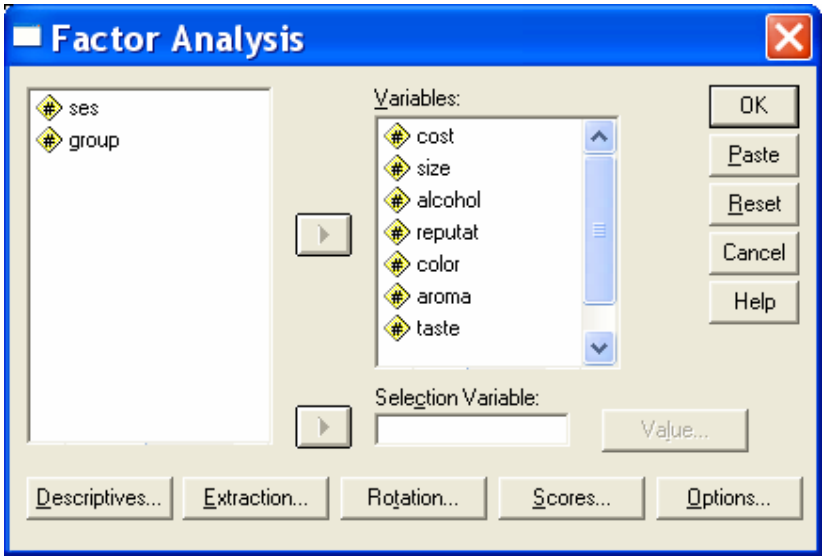
Interpretazione

Per l'interpretazione dei fattori comuni f_h ci si basa sui λ_{jh}, j=1,..,k. La situazione ideale e' quella per cui $\forall$j vi sono pochi λ_{jh} elevati (in valore assoluto) e tanti λ_{jh} prossimi allo zero. Tuttavia poiche' le soluzioni dell'analisi fattoriale sono univocamente determinate a meno di una trasformazione ortogonale e' possibile ruotare i fattori con lo scopo di cercare le soluzioni che si prestino meglio ad una interpretazione.

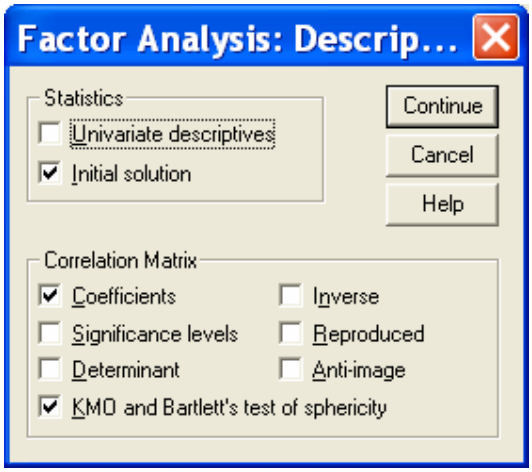
Esempio (dovuto a K. Wuensch)

Immaginiamo di essere interessati a cosa influenza il comportamento dei consumatori quando acquistano una birra. Chiediamo a 200 consumatori di dare un voto da 0 a 100 quanto considera importanti le seguenti qualità al fine di decidere di acquistare una confezione da 6: (basso) COST, (alta) SIZE (cioe'

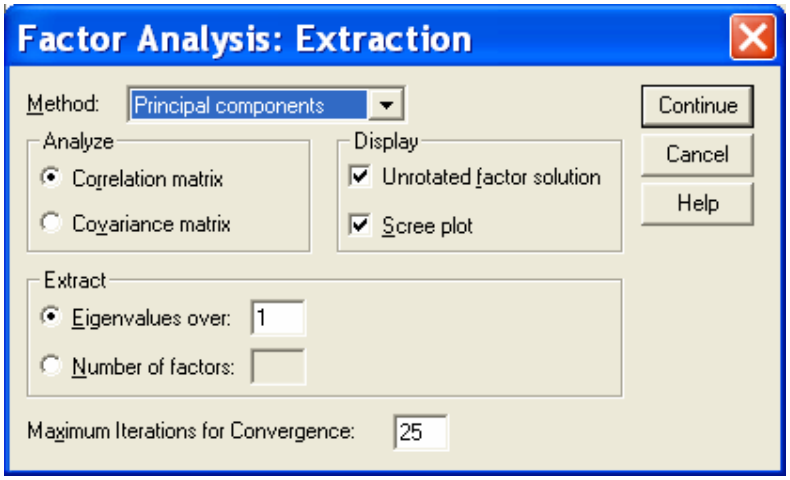
volume), (alta percentuale di) ALCOHOL, REPUTATION, COLOR, AROMA, (buon) TASTE.
Sulla riga dei comandi, click Analyze, Data Reduction, Factor. Inserisci le 7 variabili.



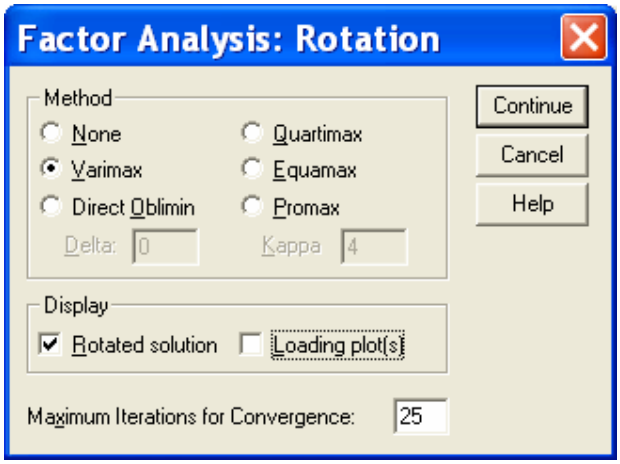
Click Descriptives and then check Initial Solution, Coefficients, Click Continue.



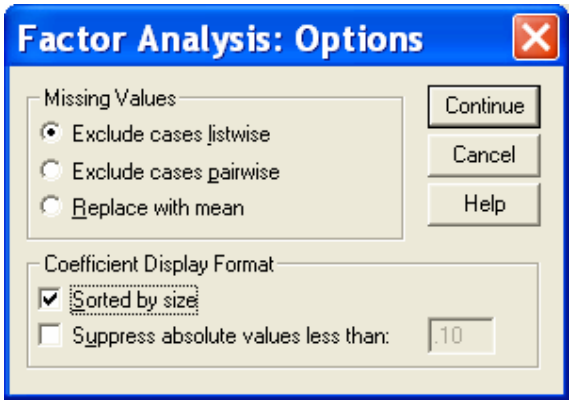
Click Extraction and then select Correlation Matrix, Unrotated Factor Solution, Scree Plot, and Eigenvalues Over 1. Click Continue.



Click Rotation. Select Varimax and Rotated Solution. Click Continue.



Click Options. Select Exclude Cases Listwise and Sorted By Size. Click Continue.



Click OK, and SPSS completes the Principle Components Analysis.

La matrice di correlazione e':

	COST	SIZE	ALCOHOL	REPUTAT	COLOR	AROMA	TASTE
COST	1.00	.83	.77	-.41	.02	-.05	-.06
SIZE	.83	1.00	.90	-.39	.18	.10	.03
ALCOHOL	.77	.90	1.00	-.46	.07	.04	.01
REPUTAT	-.41	-.39	-.46	1.00	-.37	-.44	-.44
COLOR	.02	.18	.07	-.37	1.00	.91	.90
AROMA	-.05	.10	.04	-.44	.91	1.00	.87
TASTE	-.06	.03	.01	-.44	.90	.87	1.00

Iniziamo a fare le componenti principali, ovvero estraiamo m=7 fattori.

Gli autovalori e le proporzioni di varianza spiegata dalle 7 componenti:

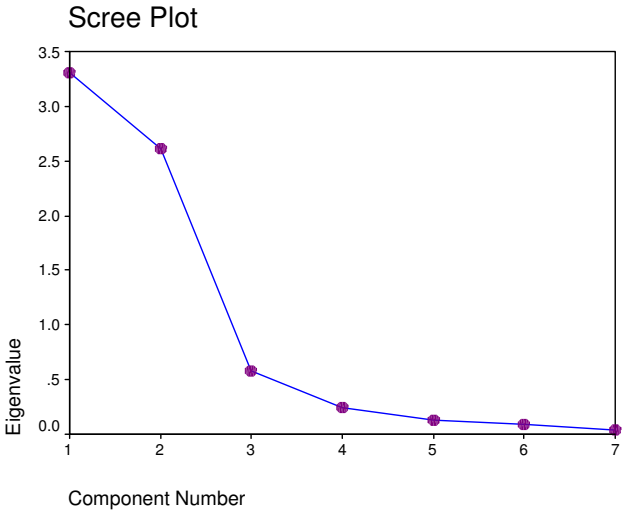
	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	3.313	47.327	47.327
2	2.616	37.369	84.696
3	.575	8.209	92.905
4	.240	3.427	96.332
5	.134	1.921	98.252
6	9.E-02	1.221	99.473
7	4.E-02	.527	100.000

Extraction Method: Principal Component Analysis.

Quante componenti tenere?

- solo quelle con autovalore maggiore di 1
- tante da avere una % di varianza spiagata almeno pari a 80%
- guardare scree plot per trovare il punto a destra del quale la pendenza decresce "poco".

Nel nostro caso



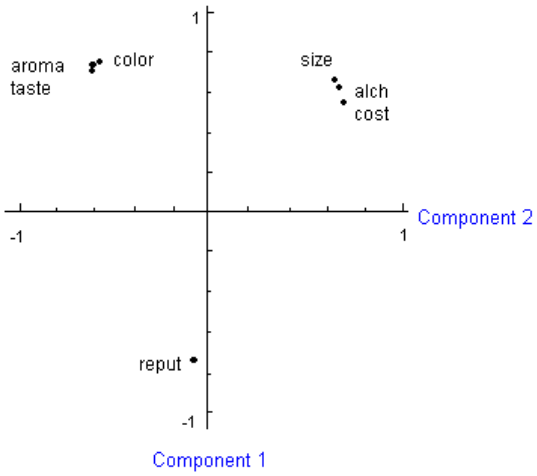
lo scree plot indicherebbe come migliore scelta quella di 2 o 3 componenti; la varianza spiegata tende ad individuarne 2, che sono anche i soli due autovalori superiori all'unità. Le prime due componenti sono riportate nella tabella, menrte nel grafico si ha il cerchio delle correlazioni. Entrambi questi output servono ad interpretare i fattori estratti.

Component Matrix[a]

	Component	
	1	2
COLOR	.760	-.576
AROMA	.736	-.614
REPUTAT	-.735	-.071
TASTE	.710	-.646
COST	.550	.734
ALCOHOL	.632	.699
SIZE	.667	.675

Extraction Method: Principal Component Analysis.

a. 2 components extracted.



Si puo' notare che quasi tutte le variabili hanno un coefficiente elevato (in valore assoluto) sulla prima componente. Tutti positivi tranne che per la reputazione.

La seconda componente e' piu' interessante: ha tre coefficienti alti e positivi e tre coefficienti alti e negativi. La prima componente sembra riflettere considerazioni riguardanti spesa e qualità contrapposte a reputazione. La seconda sembra contrapporre spesa/economicità a qualità.

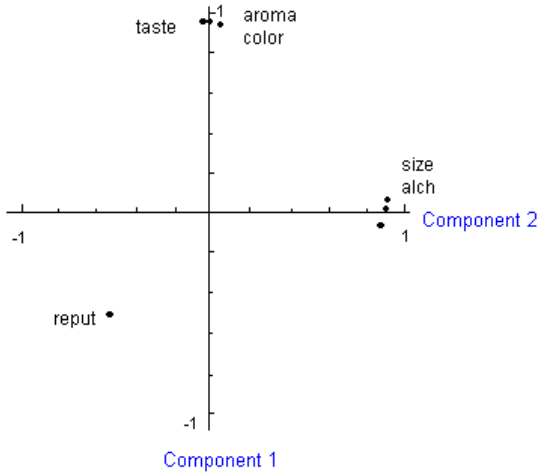
Effettuando una rotazione varimax:

Rotated Component Matrix[a]

	Component	
	1	2
TASTE	.960	-.028
AROMA	.958	1.E-02
COLOR	.952	6.E-02
SIZE	7.E-02	.947
ALCOHOL	2.E-02	.942
COST	-.061	.916
REPUTAT	-.512	-.533

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.



Total Variance Explained

Component	Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	3.017	43.101	43.101
2	2.912	41.595	84.696

Extraction Method: Principal Component Analysis.

Communalities

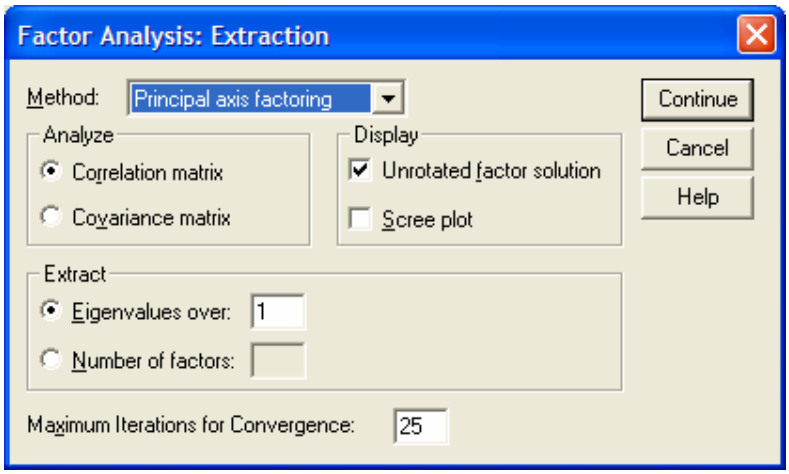
	Initial	Extraction
COST	1.000	.842
SIZE	1.000	.901
ALCOHOL	1.000	.889
REPUTAT	1.000	.546
COLOR	1.000	.910
AROMA	1.000	.918
TASTE	1.000	.922

Extraction Method: Principal Component Analysis.

Possiamo anche andare a calcolare la SSL per ciascuna variabile, ovvero: $\Sigma_h \lambda^2_{jh} = c_j$, la cosi' detta comunalità della j-esima variabile = R^2 della variabile "prevista" attraverso le componenti. Con i dati sulla birra risulta:

COST, .84; SIZE, .90; ALCOHOL, .89; REPUTAT, .55; COLOR, .91; AROMA, .92; and TASTE, .92.

Si potrebbe usare un altro metodo, quello dell'asse principale. Si parte da una matrice di correlazione ridotta, con le comunalità al posto degli 1 sulla diagonale principale.



Communalities

	Initial	Extraction
COST	.738	.745
SIZE	.912	.914
ALCOHOL	.866	.866
REPUTAT	.499	.385
COLOR	.922	.892
AROMA	.857	.896
TASTE	.881	.902

Extraction Method: Principal Axis Factoring.

Per fare una analisi fattoriale con il metodo principal axis iterato SPSS stima preliminarmente le comunalità tramite R^2 's, di ciascuna variabile e poi procede all'analisi. Si itera l'estrazione di componenti e la stima delle comunalità finche' non si raggiunge la convergenza .

Total Variance Explained

	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.123	44.620	44.620	2.879	41.131	41.131
2	2.478	35.396	80.016	2.722	38.885	80.016

Extraction Method: Principal Axis Factoring.

Rotated Factor Matrix[a]

	Factor	
	1	2
TASTE	.950	-2.17E-02
AROMA	.946	2.106E-02
COLOR	.942	6.771E-02
SIZE	7.337E-02	.953
ALCOHOL	2.974E-02	.930
COST	-4.64E-02	.862
REPUTAT	-.431	-.447

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Questo risultato e' assai simile a quello ottenuto con il metodo delle componenti principali.

Per quanto riguarda le **rotazioni** anche qui si puo' scegliere. Tra le rotazioni ortogonali oltre a VARIMAX c'e' anche QUARTIMAX e EQUAMAX. Ortogonalità significa che le componenti (= fattori comuni) restano incorrelate tra loro , e gli assi ortogonali.