ANALISI FATTORIALE

L'analisi statistica *multivariata* studia le proprietà di un insieme di *p variabili* rilevate su un insieme di elementi $I = \{I_1, I_2, \ldots\ldots, I_n\}$ *(prodotti, marchi, aziende, individui, ……….)*

Matrice di dati multivariati

I dati consistono in una matrice in cui *p* variabili vengono rilevate su *n* di soggetti, oggetti o altre entità di interesse. Tali dati possono essere rappresentati da una matrice *X* ovvero la *matrice dei dati multivariati*

- X =

$$\begin{bmatrix} x_{11} & . & . & . & x_{1p} \\ . & & & & . \\ x_{i1} & . & . & . & x_{ip} \\ . & & & & . \\ x_{h1} & . & . & . & x_{hp} \\ . & & & & . \\ x_{n1} & . & . & . & x_{np} \end{bmatrix}$$

ANALISI FATTORIALE

- L’analisi fattoriale è un **insieme di tecniche statistiche utilizzate per ricercare l’esistenza di variabili latenti** a partire una serie di variabili osservate.
- **Una variabile è definita latente quando non è direttamente osservabili**.
- Ad esempio, concetti come il benessere, la soddisfazione del cliente, la salute, l’intelligenza non sono direttamente misurabili. Tuttavia, molte ricerche si basano proprio sull’indagare questi concetti utilizzando una serie di variabili misurabili.

ANALISI FATTORIALE

- L'obiettivo di questa analisi è **capire se queste variabili misurabili sono utili a spiegare un determinato fenomeno** che non può essere direttamente misurato.
- Agli inizi del 1900, Charles Edward Spearman (lo stesso dell'**indice di correlazione di Spearman**) formulò infatti questa tecnica con l'obiettivo di misurare in modo analitico l'intelligenza umana.

ANALISI FATTORIALE

- Questa tecnica è molto applicata nelle indagini di marketing.
  - Un esempio di applicazione di questa tecnica riguarda l’analisi di questionari, in questo caso l’analisi fattoriale aiuta a **capire se un questionario misura effettivamente quello per cui è stato progettato**

AFE ed AFC

- AFE (analisi fattoriale esplorativa)
- AFC (analisi fattoriale confermativa)
- **a scopi esplorativi**: l’analisi fattoriale esplorativa (AFE) permette di “esplorare” le relazioni nascoste fra una serie di variabili osservate.
- **a scopi confermativi**: l’analisi fattoriale confermativa (AFC) si utilizza quando si conosce già a priori la struttura delle variabili e si vuole “confermare” se tale struttura sia presente anche per i dati raccolti.

Analisi fattoriale e analisi delle componenti principali

- L’ACP è principalmente una tecnica di riduzione delle dimensioni che permette di semplificare l’interpretazione dei dati riducendo il numero di variabili senza perdere troppa informazione.
- l’AFE si basa sulla ricerca di variabili latenti mentre l’ACP valuta la correlazione esistente tra i dati ed è inserita tra i metodi di estrazione dei fattori.

ANALISI FATTORIALE

$$Z = \begin{pmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & & \\ z_{n1} & \cdots & z_{np} \end{pmatrix}$$

matrice contenente p var. standardizzate rilevate su n unità.

L'analisi fattoriale si propone di individuare le dimensioni fondamentali del campo descritto dalle p var. Vale a dire si tratta di verificare in che misura ciascuna delle p var. costituisce una ripetizione della descrizione effettuata dalle rimanenti p-1 e quindi se ∃ la possibilità di raggiungere la stessa efficacia

②

descrittive con un numero $q \le p$

di variabili non osservate dette

FATTORI.

Il modello FATTORIALE CLASSICO

prevede q fattori comuni a

tutte le p variabili

più un fattore specifico per ogni

variabile

$$Z = F S' + W C$$

Z matrice delle $(n \times p)$ var. osservate standardizzate

F " " $(n \times q)$ le cui colonne contengono i q fattori comuni NON OSSERVABILI

(ogni colonna di F ha media 0 e var. 1

S: matrice $p \times q$ dei coefficienti di correlazione tra le p variabili e i q fattori comuni.

W: matrice $n \times p$ contiene i p fattori specifici uno per ogni variabile.

C: matrice diagonale $p \times p$ che contiene i coeff. dei fattori specifici.

NB: Si determinano i fattori comuni

quelli specifici si ottengono per

residuo

$$W C = \underset{\text{note}}{Z} - \underset{\text{da determinare}}{F S'}$$

Date le p var. standardizzate misurate su n unità si tratta di:

a) ESTRARRE I FATTORI

Si devono determinare q vettori $\underline{V}_k$ $(p \times 1)$

$\underline{V}_k$ $(k = 1, 2, \ldots, q \quad q < p)$:

che trasformino la matrice Z:

$Y = Z V$

V matrice $p \times q$ composta dei vettori $\underline{V}_k$

Y matrice in cui ogni colonna Y_k contiene una variabile combinazione lineare delle p var. Z_j (di media 0 e var. λ_k)

Se le variabili Y_k sono tra loro 5
incorrelate $\Rightarrow$ la matrice di varianza e
covarianza

$$L = \frac{1}{n} Y'Y$$

sarà diagonale : $L = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_q \end{pmatrix}$

Standardizzando le variabili Y_k si ottengono q fattori

Y matrice in cui ogni colonna Y_k contiene una variabile combinazione lineare delle p var. z_j (di media 0 e var. λ_k)

$$f_k = \frac{1}{\sqrt{\lambda_k}} Y_k$$

che presi congiuntamente costituiscono la SOLUZIONE FATTORIALE

$$F = Y L^{-1/2}$$

La matrice F (nxp) di solito non viene determinata concretamente infatti la soluzione fattoriale è data dalla matrice S delle correlazioni tra le q var. z_j e i p fattori f_k

6

LA MATRICE DELLE CORRELAZIONI

$$S = \frac{1}{n} Z' F = \frac{1}{n} Z' Y L^{-1/2} = \frac{1}{n} Z' Z V L^{-1/2}$$

$$Y = Z V$$

$$= R V L^{-1/2}$$

Il generico elemento s_{jk} indica

IL COEF. DI CORRELAZIONE TRA LA VAR. z_j e IL FATTORE f_k

b) ROTARE I FATTORI

Lo scopo dell'analisi fattoriale è quello di individuare le dimensioni fondamentali di un certo fenomeno.

Da un punto di vista grafico i fattori estratti si possono considerare come un sistema di assi

che individuano un iper-spazio
a q dimensioni nel quale ciascuna
delle p variabili z_j corrisponde
ad un punto

I coeff. di correlazione tra il
K-esimo fattore f_k e le p variabili z_j
sono le coordinate di queste
ultime sull'asse K.

Lo scopo è quello di individuare
un nuovo sistema di assi
in cui ciascun punto (raffigurante
una delle p variabili z_j) si disponga
il più vicino possibile ad uno degli
assi.

Si tratta di esprimere Z come combinazione lineare dei fattori comuni e dei fattori specifici: (8)

$$\underset{(n\times p)}{Z} = \underset{(n\times q)}{F}\ \underset{(q\times p)}{S'} + \underset{(n\times p)}{W}\ \underset{(p\times p)}{C}$$

es:

$$z_{ij} = s_{j1} f_{i1} + s_{j2} f_{i2} + \dots + s_{jp} f_{ip} + c_{jj} w_{ij}$$

i → OSS. j → VAR

Fissato il numero dei fattori si deve individuare la matrice S

IPOTESI:

1. La relazione che collega le var. latenti (fattori) e quelle osservabili sia lineare.

2. I fattori devono essere incorrelati (vale a dire ortogonali)

$$\frac{1}{n} F'F = I$$

$$\frac{1}{n} W'W = I$$

3. I fattori comuni e i fattori specifici devono essere incorrelati

$$\frac{1}{n} W'F = 0$$

ovvero

$$\frac{1}{n} F'W = 0$$

$$R = \frac{1}{n} Z'Z = \frac{1}{n}(FS' + WC)'(FS' + WC)$$

$\Downarrow$

matrice delle correlazioni = matr var. cov.

$$= \frac{1}{n}(SF' + C'W')(FS' + WC)$$

$$= \frac{1}{n}(SF'FS' + SF'WC + C'W'FS' + C'W'WC)$$

$$= S\underbrace{\left(\frac{1}{n}F'F\right)}_{I_p}S' + S\underbrace{\left(\frac{1}{n}F'W\right)}_{0}C +$$

$$+ C'\underbrace{\left(\frac{1}{m}W'F\right)}_{0}S' + C'\underbrace{\left(\frac{1}{n}W'W\right)}_{I_q}C$$

$$\Rightarrow R = SS' + C'C$$

$$R = \begin{bmatrix} s_{11} & \cdots & s_{1q} \\ \vdots & & \vdots \\ s_{p1} & \cdots & s_{pq} \end{bmatrix} \begin{bmatrix} s_{11} & \cdots & s_{p1} \\ \vdots & & \\ s_{1q} & \cdots & s_{pq} \end{bmatrix} + \begin{bmatrix} c_{11} & & 0 \\ & \ddots & \\ 0 & & c_{pp} \end{bmatrix} \begin{bmatrix} c_{11} & & 0 \\ & \ddots & \\ 0 & & c_{pp} \end{bmatrix}$$

$$\begin{bmatrix} s_{11}^2 + s_{12}^2 + \dots + s_{1q}^2 + c_{11}^2 & & \\ & \ddots & \\ & & s_{p1}^2 + s_{p2}^2 + \dots + s_{pq}^2 + c_{pp}^2 \end{bmatrix}$$

$$\hat{\sigma}_i^2 = \underbrace{s_{i1}^2 + \dots + s_{iq}^2}_{\text{somma per fattoriali al quadrato}} + \underbrace{c_{ii}^2}_{\text{coeff. fattore specifico}} = 1$$

COMUNALITA'   SPECIFICITA'

$$h_i^2 = \sum_j s_{ij}^2 \qquad\qquad c_{ii}^2$$

$$\Rightarrow \hat{\sigma}_i^2 = h_i^2 + c_{ii}^2 = 1 \qquad \Rightarrow h_i^2 = 1 - c_{ii}^2$$

L'estrazione dei fattori comuni utilizza R se il modello è

$$Z = FS'$$

Mentre se il modello è

$$Z = FS' + WC$$

Bisogna sostituire i valori 1 che si trovano sulla diagonale

principale con le stime delle comunalità che sono minori di 1

Nel modello generale si fa riferimento
alla matrice ridotta R_r

$$\begin{bmatrix} h_1^2 & r_{12} & \cdots & r_{1p} \\ \vdots & & & \\ r_{p1}, & r_{p2} & \cdots & h_p^2 \end{bmatrix} = R_r$$

Molte tecniche di analisi fattoriale si fondano su stime preliminari delle h_j^2

2 STRATEGIE PER DETERMINARE IL VALORE DELLE COMUNALITÀ:

a) fissare il numero di fattori da estrarre q e trovare i valori delle comunalità h_j^2 tale che R abbia rango q.

b) fissare i valori delle comunalità in modo che riducano il rango della matrice R_r senza alcuna considerazione preliminare sul n° di fattori da estrarre.

Si tratta di attribuire valori arbitrari alle comunalità.

LE COMUNALITÀ

di correlazione

La matrice delle varianze e covarianze delle p var. standardizzate z_j è:

$$R = \frac{1}{n} Z'Z$$

La comunalità di una variabile z_j è la parte di varianza di z_j estratta dai fattori comuni e che quindi è uguale alla somma dei quadrati degli elementi della j-esima riga della matrice S:

$$h_j^2 = \sum_{1}^{q} s_{jj}^2 \qquad j = 1, 2, \ldots, p$$

Con l'analisi fattoriale si vuole riprodurre, mediante i fattori comuni, la matrice delle varianze e covarianze = alla matrice delle correlazioni R.

$$R = S S' + C'C$$

Sono stati trovati parecchi valori delle communalità

METODO DELLA MASSIMA CORRELAZIONE

a) ${}_1h_j^2 = \max_k (r_{jk})$

coeff. di correlazione più elevato in valore assoluto che si trova fuori della diagonale principale

METODO DELLA TRIADE

b) ${}_2h_j^2 = R_{jk} R_{jl} / R_{kl}$

con r_{jk} e R_{jl} coeff. di correlazione più elevati che si trovano sulla riga j di R

METODO DEL PRIMO CENTROIDE

c) ${}_3h_j^2 = (\sum_k R_{jk})^2 / \sum \sum R_{jk}$

il rapporto tra il quadrato della somma degli elementi della j-esima riga di R e la somma di tutti gli elementi di R

METODO DEL COEF. DI DETERMINAZIONE

d) $h_j^2 = R_j^2$ coeff. di determinazione tra la j-esima va e le rimanenti $p-1$

METODI DI ESTRAZIONE DEI FATTORI

Consiste nella trasformazione della matrice dei dati nella matrice dei valori fattoriali: 2 FASI

I° FASE

Si deve ricavare la matrice Y trasformata della matrice Z:

$$\underset{}{Y} = \underset{(n \times p)}{Z} \; \underset{(p \times q)}{V}$$

$$V = [\underline{v}_1, \ldots, \underline{v}_q]$$

$$Y = [\underline{y}_1, \ldots \underline{y}_q]$$

con $\underline{y}_k = Z \, \underline{v}_k$

$E(y_k) = 0$

$Var(y_k) = \lambda_k$

$cov(y_k, y_j) = 0 \quad \forall \, k \neq j$

II^a FASE

Ricercare le var. fattoriali che devono avere media nulla e varianza unitaria.

Data $L = \frac{1}{n} Y'Y = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_q \end{bmatrix}$

standardizzando le Y_k si ottengono q fattori

$$f_k = \frac{1}{\sqrt{\lambda_k}} Y_k$$

Si arriva alla matrice dei punteggi fattoriali

$$F = Y L^{-1/2}$$

e alla matrice dei pesi fattoriali

$$S = \frac{1}{n} Z'F \ldots = R V L^{-1/2}$$

Esistono diversi metodi tra cui:

- Il metodo delle componenti principali
- Il metodo dei fattori principali

DETERMINAZIONE DEL N° DEI FATTORI

Un criterio per valutare se il n° q di fattori è sufficiente a racchiudere l'informazione fornita dalle p var. di partenza è:

fissare una quota V_c di varianza delle var. Z_i

e fermarsi a q fattori se la quota di varianza cumulata estratta da questi è $> V_c$

Ora

$$h_i^2 = \sum_1^q s_{ij}^2 \qquad j = 1, \dots, p$$

quota di varianza della var. Z_i estratta dai fattori comuni

$R^* = SS'$ matrice degli h_i^2

L'ammontare complessivo di varianza delle p variabili originarie standardizzate estratta da q fattori comuni è:

$tr(R_{\bullet}^{*})$:

$\frac{tr(R_{\bullet}^{*})}{p}$: porzione di varianza delle p variabili originarie standardizzate estratte da q fattori comuni

Gli elementi della matrice dei residui $\bar{R} = R - R_{\bullet}^{*}$

possono fornire un'indicazione sul grado di accuratezza raggiunto dai q fattori nel riprodurre la matrice di var. e covar.

Un altro metodo è guardare allo scree plot e scegliere le componenti con autovalore>1

- Number of Factors to Retain

PROC FACTOR provides a variety of factor extraction methods that you can specify by using the METHOD= option. Typically, the maximum number of factors that can be extracted is the same as the number of variables that are being factor-analyzed. Because factor analysis can be viewed as a data-reduction technique, the actual number of factors that researchers want to retain after the extraction is usually much smaller. PROC FACTOR provides the following three options that you can control the number of factors to retain in factor solutions: MINEIGEN=, NFACTORS=, and PROPORTION=. The NFACTORS= option specifies explicitly the number of factors to retain. Each of the other two options specifies a criterion value to determine the number of factors. See the MINEIGEN= and PROPORTION= options for details about how these criteria work.

If you specify two or more of these options, the number of factors retained is the minimum number satisfying any of the criteria. (SAS help)

- MINEIGEN=p
- MIN=p
- specifies the smallest eigenvalue for which a factor is retained. If you specify two or more of the MINEIGEN=, NFACTORS=, and PROPORTION= options, the number of factors retained is the minimum number satisfying any of the criteria. The MINEIGEN= option cannot be used with either the METHOD=PATTERN or the METHOD=SCORE option. Negative values are not allowed. The default is 0 unless you omit both the NFACTORS= and the PROPORTION= options and one of the following conditions holds:

- If you specify the METHOD=ALPHA or METHOD=HARRIS option, then MINEIGEN=1.

- If you specify the METHOD=IMAGE option, then

- For any other METHOD= specification, if prior communality estimates of 1.0 are used, then

- When an unweighted correlation matrix is factored, this value is 1.

- PROPORTION=p
- PERCENT=p
- P=p
- specifies the proportion of common variance to be accounted for by the retained factors. The proportion of common variance is computed using the total prior communality estimates as the basis. If the value is greater than one, it is interpreted as a percentage and divided by 100.

Per la maggior parte dei metodi di estrazione dei fattori è possibile trovare un test statistico per verificare l'opportunità o meno di estrarre ulteriori fattori.

Se Estrazione con il metodo delle componenti principali

⇒ TEST DI BARTLETT

Esistono diversi metodi tra cui:

- Il metodo delle componenti principali
- Il metodo dei fattori principali

1° Metodo delle componenti principali.

Si assume il modello

$$Z = S\,F$$

(si ipotizza la nullità o della matrice W o della matrice C)

Si tratta di determinare la matrice

$$V = [\underline{v}_1 \dots \underline{v}_q]$$

in modo da trasformare la Z in

$$Y = Z\,V$$

$\underline{v}_k$ sono tali che:

$$\underline{v}_k' \underline{v}_k = 1$$

le Y_k sono tali che

a) sono ortogonali

b) hanno media zero e varianza massima subordinatamente alla condizione di ortogonalità e di $\underline{v}_k' \underline{v}_k = 1$

c) $Var(Y_K) = \lambda_K$

k-esima radice latente della matrice R
e questa viene associato il vettore latente $\underline{v}_k$.

Per ricavare $\lambda_1 \dots \lambda_q$

si deve risolvere l'equazione

caratteristica:

$$|R - \lambda I| = 0$$

le $\lambda_1 \dots \lambda_q$ sostituite nel

sistema di eq.:

$$(R - \lambda I)\, \underline{v} = \underline{0}$$

permettono di passare dalle var. oss. alle latenti cui si assegna il significato di fattori.

Estrazione dei fattori:metodo delle componenti principali

La procedura FACTOR

Tipo di dati di input	Dati grezzi
Num. di record letti	10
Num. di record usati	10
N per test significatività	10

Medie e deviazioni standard da 10 osservazioni		
Variabile	Media	Dev std
maturita	46.700000	9.1414806
laurea	99.700000	7.1655038

Correlazioni		
	maturita	laurea
maturita	1.00000	0.79063
laurea	0.79063	1.00000

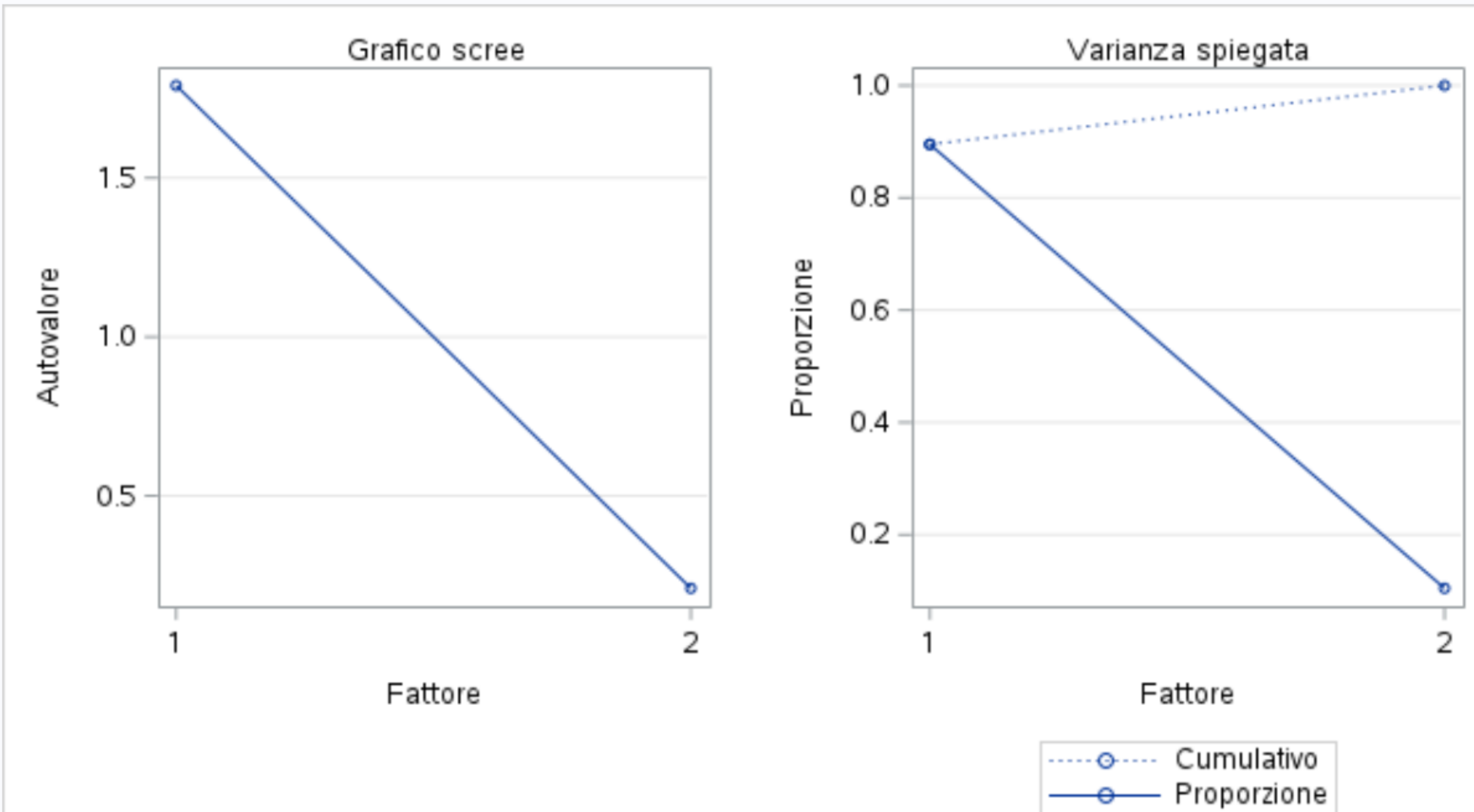
Proc factor (corr)

La procedura FACTOR
Metodo fattoriale iniziale: Componenti principali

Stime di comunanza a priori: ONE

Autovalori della matrice di correlazione: Totale = 2 Media = 1				
	Autovalore	Differenza	Proporzione	Cumulativa
1	1.79063006	1.58126013	0.8953	0.8953
2	0.20936994		0.1047	1.0000

1 fattore sarà mantenuto dal criterio MINEIGEN.

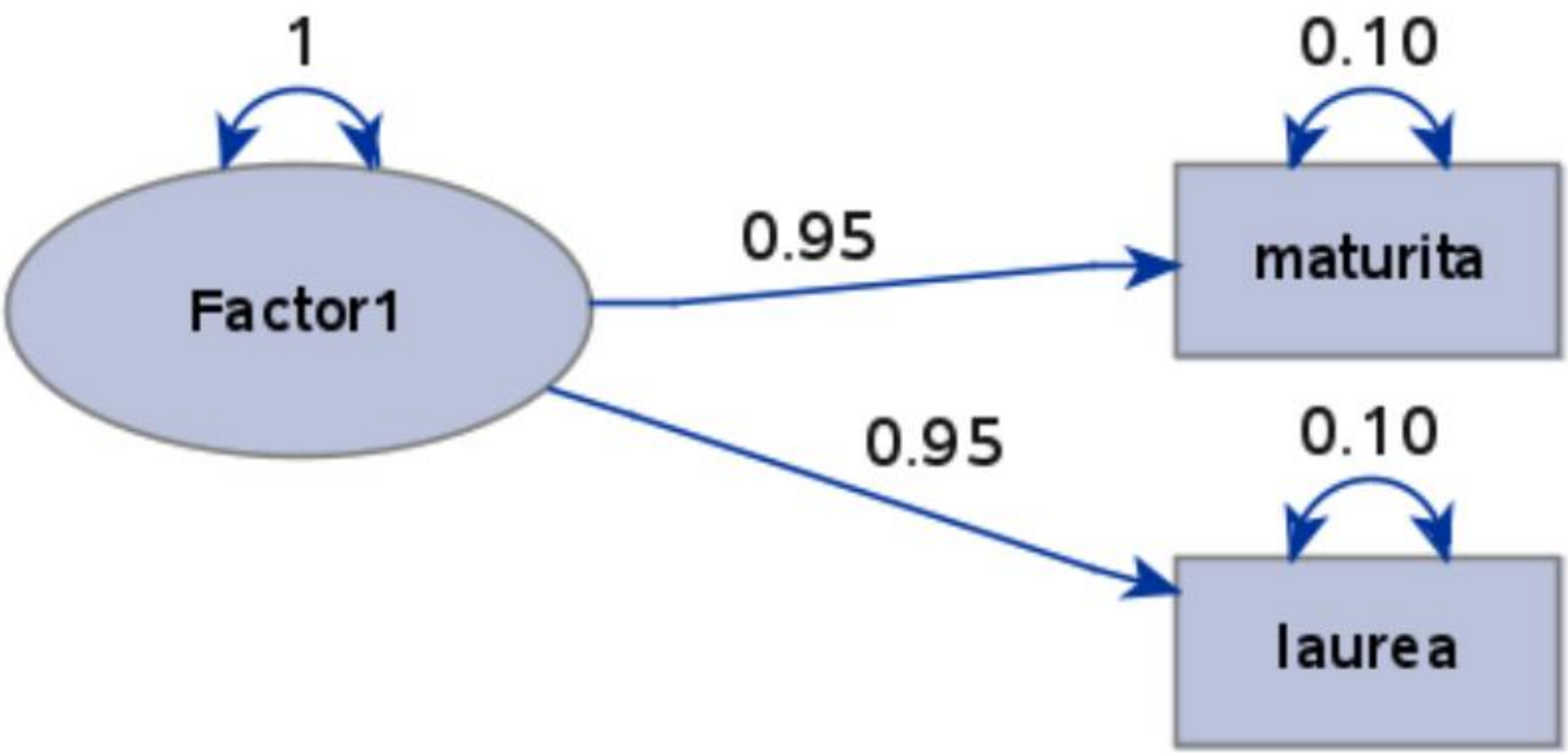


Pattern fattoriale	
	Factor1
maturita	0.94621
laurea	0.94621

Varianza spiegata da ogni fattore
Factor1
1.7906301

Stime di comunanza finali: Totale = 1.790630	
maturita	**laurea**
0.89531503	0.89531503

La procedura FACTOR
Metodo fattoriale iniziale: Componenti principali



1-0.9=0.1

1-0.9=0.1

Proc factor (cov)

La procedura FACTOR

Tipo di dati di input	Dati grezzi
Num. di record letti	10
Num. di record usati	10
N per test significatività	10

Medie e deviazioni standard da 10 osservazioni		
Variabile	Media	Dev std
maturita	46.700000	9.1414806
laurea	99.700000	7.1655038

La procedura FACTOR
Metodo fattoriale iniziale: Componenti principali

Stime di comunanza a priori: ONE

Autovalori della matrice di covarianza: Totale = 134.911111 Media = 67.4555556				
	Autovalore	**Differenza**	**Proporzione**	**Cumulativa**
1	121.692599	108.474087	0.9020	0.9020
2	13.218512		0.0980	1.0000

1 fattore sarà mantenuto dal criterio MINEIGEN.

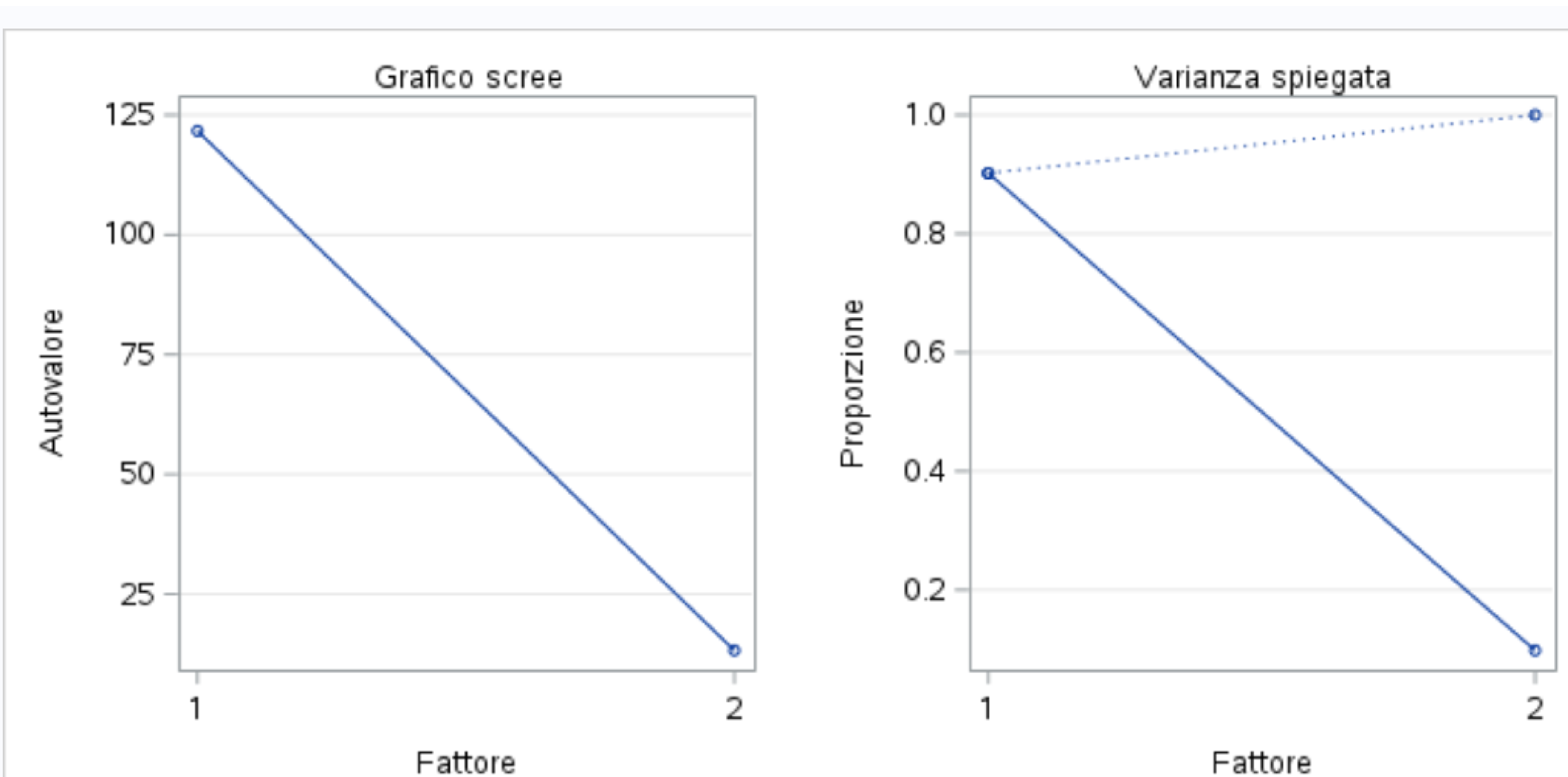


Grafico scree
Autovalore
125
100
75
50
25
1
2
Fattore
Varianza spiegata
Proporzione
1.0
0.8
0.6
0.4
0.2
1
2
Fattore

Pattern fattoriale	
	Factor1
maturita	0.97180
laurea	0.91271

Varianza spiegata da ogni fattore		
Fattore	Pesato	Non pesato
Factor1	121.692599	1.77744265

Stime di comunanza finali e pesi variabili		
Comunanza totale: Pesata = 121.69260 Non pesata = 1.777443		
Variabile	Comunanza	Peso
maturita	0.94440395	83.5666667
laurea	0.83303870	51.3444444

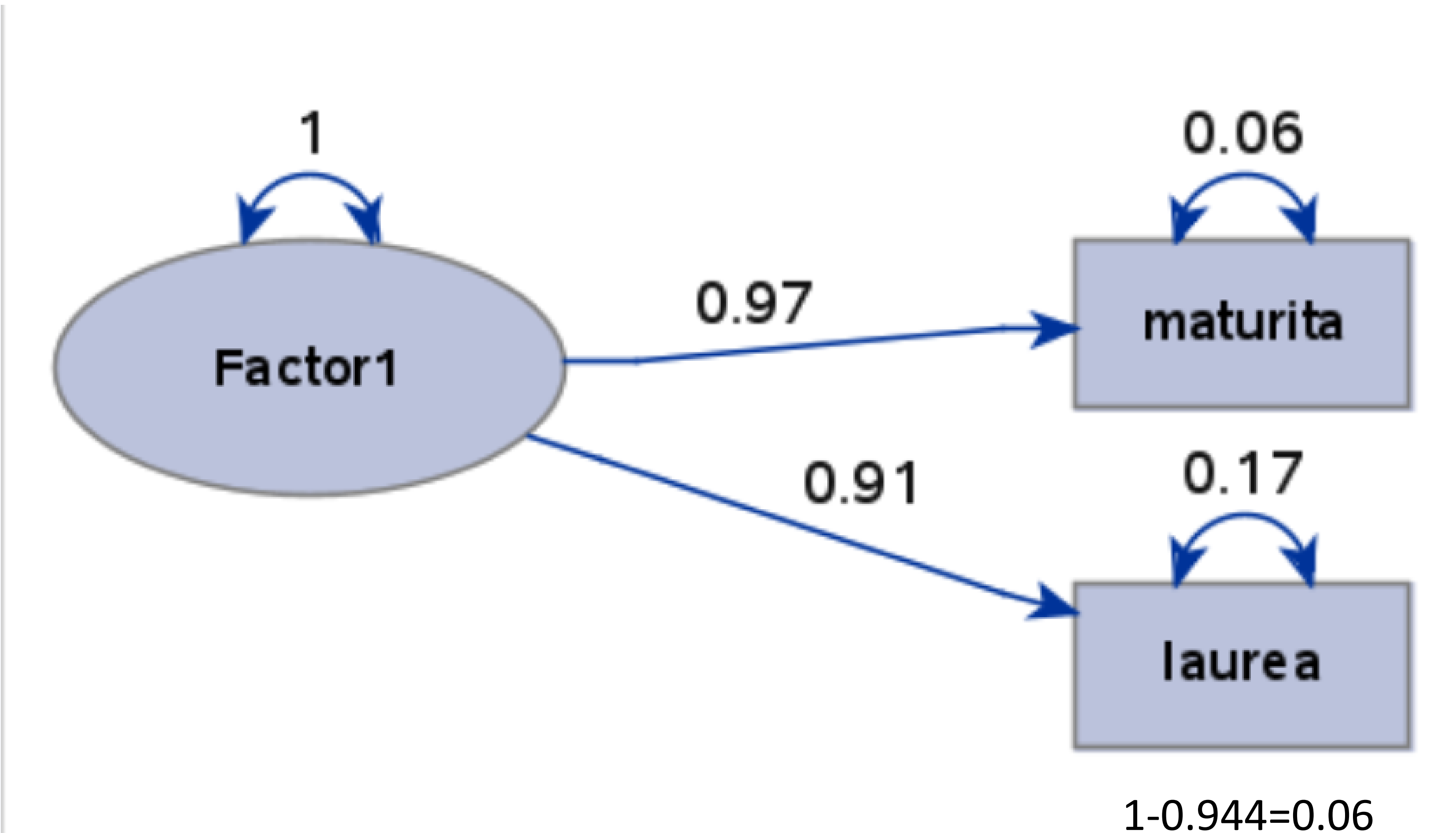
- Si verifica che la somma degli autovalori è uguale alla varianza totale

Varianza totale	134.9111111

- $121.693 + 13.218 = 83.567 + 51.344$

Peso= varianze

1.777=0.944+0.833



1-0.944=0.06
1-0.833=0.17

METODO DEI FATTORI PRINCIPALI

Si parte dal modello che tiene conto anche dei fattori specifici:

$$Z = FS' + WC$$

In questo caso si ricorre alla matrice di correlazione ridotta

$$R_r = \begin{bmatrix} h_1^2 & r_{12} & \cdots & r_{1p} \\ \vdots & & & \\ \vdots & & & \\ r_{p1} & r_{p2} & \cdots & h_p^2 \end{bmatrix} \qquad r(R_r) = g < q$$

Si tratta di risolvere l'equazione caratteristica:

$$|R_r - \lambda I| = 0$$

Si ottengono

$$\lambda_1 \; \cdots \; \lambda_g$$

Ritenendo che la corr. del punteggio fattoriale di una var. con se stessa non sia uno si arriva alla tecnica di estrazione detta dei fattori principali analogo a quello delle componenti principali eccetto che sulla diagonale principale ci sono le stime delle comunalità che sostituiti nel sistema

$$(R_r - \lambda I)\,\underline{v} = \underline{0}$$

danno la matrice di trasformazione

$$V = [\,\underline{v}_1 \ldots \underline{v}_q\,]$$

Estrazione dei fattori:metodo dei fattori principali

La procedura FACTOR
Metodo fattoriale iniziale: Fattori principali

Stime di comunanza a priori: MAX	
maturita	**laurea**
0.79063006	0.79063006

Autovalori della matrice di correlazione ridotta: Totale = 1.58126013 Media = 0.79063006				
	Autovalore	**Differenza**	**Proporzione**	**Cumulativa**
1	1.58126013	1.58126013	1.0000	1.0000
2	0.00000000		0.0000	1.0000

1 fattore sarà mantenuto dal criterio PROPORTION.

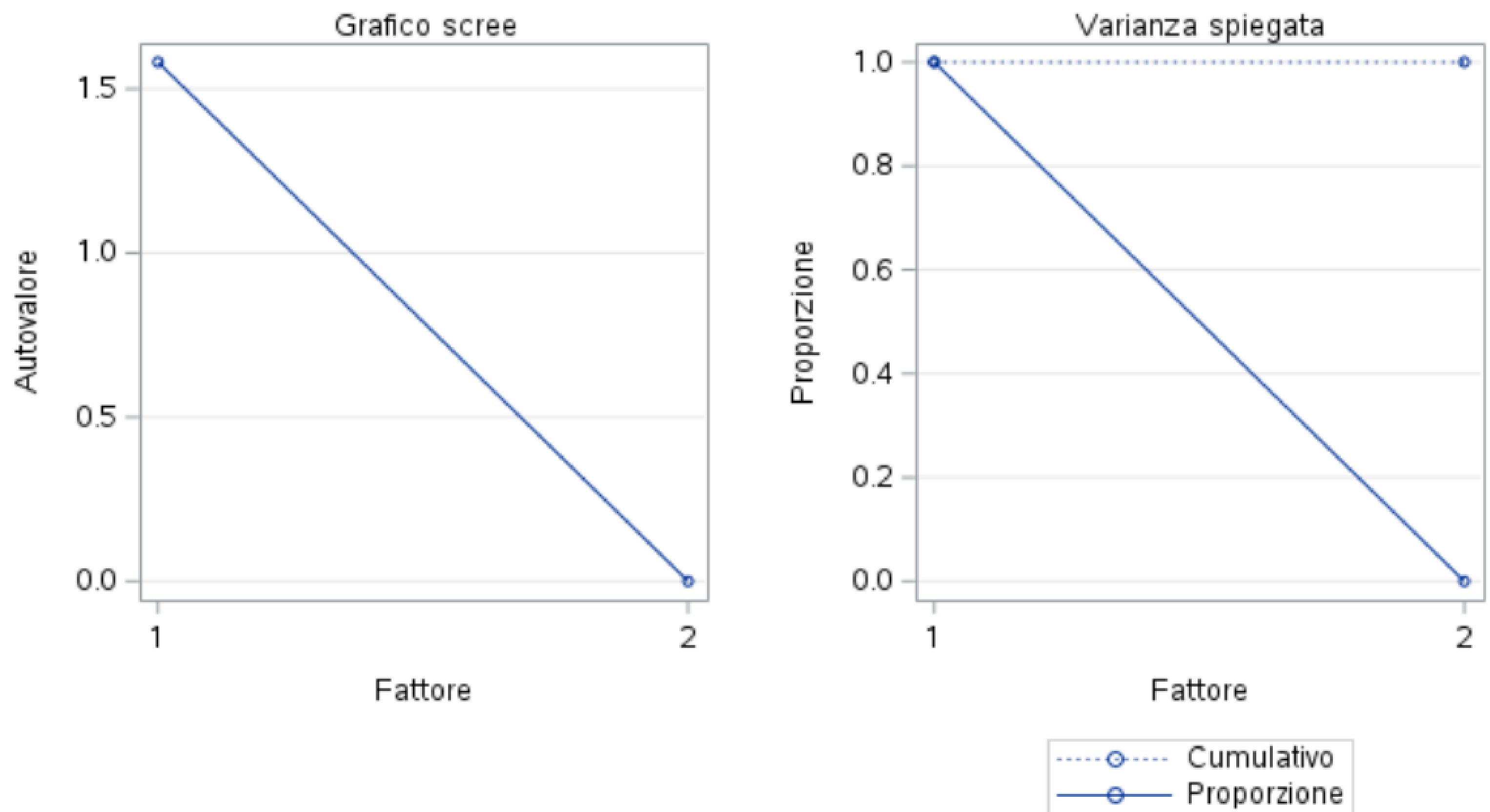
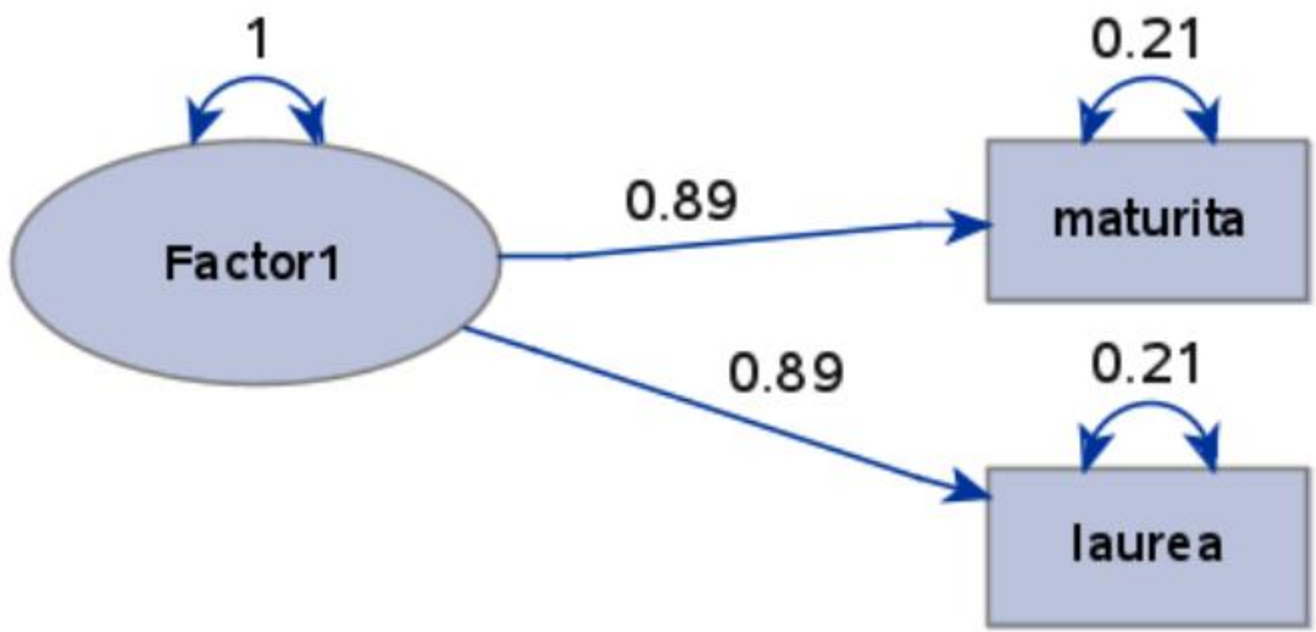


Grafico scree
Autovalore
1.5
1.0
0.5
0.0
1
2
Fattore
Varianza spiegata
Proporzione
1.0
0.8
0.6
0.4
0.2
0.0
1
2
Fattore
Cumulativo
Proporzione

Pattern fattoriale	
	Factor1
maturita	0.88917
laurea	0.88917

Varianza spiegata da ogni fattore
Factor1
1.5812601

Stime di comunanza finali: Totale = 1.581260	
maturita	laurea
0.79063006	0.79063006



1-0.79=0.21

METODI DI ROTAZIONE DEI FATTORI:

a) Metodi di rotazione ortogonale
b) Metodi di rotazione obliqua.

a) La rotazione ortogonale è un caso particolare della rotazione obliqua.

Negli anni '50 sono stati introdotti 4 criteri per la rotazione ortogonale sintetizzati dal metodo quartimax.
Successivamente è sorto il metodo varimax e poi l'equamax - normal - varimax