ANALISI IN COMPONENTI PRINCIPALI (ACP)

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 = $\begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1p} \\ \cdot & & & & \cdot \\ x_{i1} & \cdot & \cdot & \cdot & x_{ip} \\ \cdot & & & & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{np} \\ \cdot & & & & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{np} \end{bmatrix}$

ANALISI IN COMPONENTI PRINCIPALI (ACP)

- Analisi di regressione in caso di collinearità;
- problemi di classificazione per gruppi ben separati;
- **riduzione delle dimensioni;**
- identificazione degli outliers.

Zani S. e Cerioli A. (2007) Analisi dei dati e data mining per le decisioni aziendali, Milano: Giuffrè, cap. 6

- Matrice dei dati *Originale* $n * p$

- *Nuova* matrice dei dati $n * q$, con $q < p$

- Ogni nuova variabile è combinazione lineare delle variabili originali
- Le nuove variabili sono scelte in modo da catturare al massimo la variabilità presente nei dati originali
- le nuove variabili sono incorrelate (ortogonali)!
- hanno varianza pari al proprio autovalore
- Le nuove variabili sono dette “scores” o “componenti principali”

- Algebricamente le componenti principali (scores) sono combinazioni lineari delle variabili $X_1,\ldots,X_p$
- Geometricamente le CP rappresentano un nuovo sistema di coordinate ottenuto ruotando gli assi originali che hanno $X_1,\ldots,X_p$ come coordinate
- I nuovi assi rappresentano le direzioni principali (autovettori o eigenvectors), cioè le direzioni di variabilità minima

- In generale, da un insieme di p>2 variabili correlate, le prime CP tengono conto della maggior parte della variabilità nelle variabili originali
- Viceversa, le ultime CP identificano direzioni lungo le quali c'è poca variabilità

- La prima componente principale si determina imponendo il vincolo che la varianza sia massima sotto la condizione che il vettore sia normalizzato (per evitare l'esplosione della varianza)

- La soluzione si ottiene via Moltiplicatori di Lagrange.

La prima componente principale della matrice X dei dati è per definizione

$$Y_1 = \sum_{i=1}^{p} \alpha_{i1} X_i = X \underline{\alpha}_1$$

LA PRIMA COMPONENTE PRINCIPALE SI DETERMINA IMPONENDO IL VINCOLO CHE LA VARIANZA SIA MASSIMA SOTTO L'ULTERIORE CONDIZIONE CHE IL VETTORE $\underline{\alpha}_1$ SIA NORMALIZZATO (VALE A DIRE CHE LA SOMMA DEL QUADRATO DELLE COMPONENTI SIA PARI AD

$$\Rightarrow \begin{cases} \sigma^2_{Y_1} = \max \\ \underline{\alpha}_1' \underline{\alpha}_1 = 1 \end{cases}$$

$$\Rightarrow \begin{cases} \sigma^2_{y_1} = \max \\ \underline{\alpha}_1' \underline{\alpha}_1 = 1 \end{cases}$$

$$\sigma^2_{y_1} = \frac{1}{n} \sum_{i=1}^{n} (y_{i1} - \bar{y}_1)^2 = \frac{1}{n} (y_{\cdot 1} - \bar{y}_1 \underline{1})' (y_{\cdot 1} - \bar{y}_1 \underline{1}) =$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_{i1} \alpha_{j1} \sigma_{ij} =$$

media aritmetica delle prime componenti

$$= \underline{\alpha}_1' \Sigma \underline{\alpha}_1$$

Σ → matrice di varianze e covarianze

Si tratta di determinare $\underline{a}_1$ in modo che sia massima l'espressione:

$$\underline{a}_1' \Sigma_1 \underline{a}_1$$

subordinatamente alla

$$\underline{a}_1' \underline{a}_1 = 1$$

Si tratta di un problema di massimo condizionato risolubile con il metodo dei moltiplicatori di Lagrange.

$$\Rightarrow \phi_1 = \underline{\alpha}_1' \Sigma \underline{\alpha}_1 + \lambda_1 (\underline{\alpha}_1' \underline{\alpha}_1 - 1)$$

λ_1 moltiplicatore di Lagrange

$$\frac{\partial \phi_1}{\partial \underline{\alpha}_1} = \left[\underline{\alpha}_1' \Sigma \underline{\alpha}_1 + \lambda_1 (1 - \underline{\alpha}_1' \underline{\alpha}_1) \right]'$$

$$= 2 \Sigma \underline{\alpha}_1 - 2 \lambda_1 \alpha_1 \Rightarrow$$

$$= 2 (\Sigma - \lambda_1 I) \alpha_1 = 0$$

$$\frac{\partial \phi_1}{\partial \lambda_1} = - \alpha_1' \alpha_1 + 1 = 0$$

$$\begin{cases} (\Sigma - \lambda_1 I)\,\underline{\alpha}_1 = 0 \\ 1 - \underline{\alpha}_1'\,\underline{\alpha}_1 = 0 \end{cases} \Rightarrow \underline{\alpha}_1'\,\underline{\alpha}_1 = 1 \quad (*)$$

$$(\Sigma - \lambda_1 I)\,\underline{\alpha}_1 = 0$$

sistema lineare omogeneo che ammette soluzione se

$$|\Sigma - \lambda_1 I| = 0$$

In questo caso $\exists$ p radici caratteristiche ognuna corrispondente ad un vettore.

Per determinare il vettore caratteristico:

$$(\Sigma - \lambda_1 I)\,\underline{\alpha}_1 = 0 \quad (\bullet) \quad (\text{premoltiplico per } \underline{\alpha}_1')$$

$$\underline{\alpha}_1'\,\Sigma\,\underline{\alpha}_1 - \lambda_1 \underbrace{\underline{\alpha}_1'\,\underline{\alpha}_1}_{=1} = 0$$

$$\underline{\alpha}_1'\,\Sigma\,\underline{\alpha}_1 = \lambda_1 = \sigma^2_{Y_1}$$

Poiché vogliamo che $\sigma^2_{Y_1}$ sia massima si sceglie per λ_1 il più elevato degli autovalori delle (*)

$$(\Sigma - \lambda_1 I)\,\underline{\alpha}_1 = 0 \qquad (*)$$

La prima componente principale Y_1 è una combinazione lineare delle p variabili $(X_1, X_2 \dots X_p)$ con coefficienti uguali alle componenti del vettore caratteristico associato al più grande autovalore della matrice di varianze e covarianze.

- Le altre componenti principali si ottengono ancora come combinazioni lineari delle variabili di partenza, sono tra loro incorrelate e spiegano la massima variabilità tolta quella spiegata delle precedenti.

La soluzione si ottiene via Moltiplicatori di Lagrange.

La seconde composante principale est :

$$Y_2 = \sum_{i=1}^{p} \alpha_{i2} X_i = X \underline{\alpha}_2$$

$\Rightarrow$ Var $(Y_2) = \sigma^2_{Y_2} = MAX$

sous les liens :

$$\begin{cases} \underline{\alpha}_2' \underline{\alpha}_2 = 1 \\ \underline{\alpha}_1' \underline{\alpha}_2 = 0 \end{cases}$$

$$\phi_2 = \underline{\alpha}_2' \Sigma \underline{\alpha}_2 + \lambda_2 (1 - \underline{\alpha}_2' \underline{\alpha}_2) + \lambda_3 \underline{\alpha}_1' \underline{\alpha}_2$$

$$\begin{cases} \dfrac{\partial \phi_2}{\partial \underline{\alpha}_2} = 2 \Sigma \underline{\alpha}_2 - 2 \lambda_2 \underline{\alpha}_2 + \lambda_3 \underline{\alpha}_1 \Rightarrow \\ \Rightarrow 2 (\Sigma - \lambda_2 I) \underline{\alpha}_2 + \lambda_3 \underline{\alpha}_1 = 0 \\ \dfrac{\partial \phi_2}{\partial \lambda_2} = - \underline{\alpha}_2' \underline{\alpha}_2 + 1 = 0 \\ \partial \phi_2 = \underline{\alpha}_1' \underline{\alpha}_2 = 0 \end{cases} \Rightarrow$$

$$\begin{cases} 2 (\Sigma - \lambda_2 I) \underline{\alpha}_2 + \lambda_3 \underline{\alpha}_1 = 0 \quad (*) \\ \underline{\alpha}_2' \underline{\alpha}_2 = 1 \\ \underline{\alpha}_1' \underline{\alpha}_2 = 0 \end{cases}$$

$$\begin{cases} 2(\Sigma - \lambda_2 I)\underline{\alpha}_2 + \lambda_3 \underline{\alpha}_1 = 0 \quad (*) \\ \underline{\alpha}_2' \underline{\alpha}_2 = 1 \\ \underline{\alpha}_1' \underline{\alpha}_2 = 0 \end{cases}$$

premoltiplico per $\underline{\alpha}_2'$ la (*)

$$2\underline{\alpha}_2'\Sigma\underline{\alpha}_2 - 2\lambda_2 \underbrace{\underline{\alpha}_2'\underline{\alpha}_2}_{1} + \lambda_3 \underbrace{\underline{\alpha}_2'\underline{\alpha}_1}_{0} = 0$$

$$\underline{\alpha}_2' \Sigma \underline{\alpha}_2 = \lambda_2$$

$$\begin{cases} 2(\Sigma - \lambda_2 I)\underline{\alpha}_2 + \lambda_3 \underline{\alpha}_1 = 0 \quad (*) \\ \underline{\alpha}_2' \underline{\alpha}_2 = 1 \\ \underline{\alpha}_1' \underline{\alpha}_2 = 0 \end{cases}$$

premoltiplico per $\underline{\alpha}_1'$ la $(*)$

$$2\underbrace{\underline{\alpha}_1' \Sigma \underline{\alpha}_2}_{0} - 2\lambda_2 \underbrace{\underline{\alpha}_1' \underline{\alpha}_2}_{0} + \lambda_3 \underbrace{\underline{\alpha}_1' \underline{\alpha}_1}_{1} = 0$$

$$\Rightarrow \lambda_3 = 0$$

ricordando che

$$(\Sigma - \lambda_1 I)\underline{\alpha}_1 = 0$$

premoltiplicando per $\underline{\alpha}_2'$

$$\underline{\alpha}_2' \Sigma \underline{\alpha}_1 - \lambda \underbrace{\underline{\alpha}_2' \underline{\alpha}_1}_{0} = 0$$

$$\Rightarrow \underline{\alpha}_2' \Sigma \underline{\alpha}_1 = 0$$

Oss.

Ogni autovalore è uguale alla varianza della corrispondente C P.

Esempio

- Es. da Zani Cerioli (2007)

laureati	maturita	laurea
1	60	110
2	54	100
3	36	99
4	40	95
5	36	88
6	58	105
7	44	100
8	42	102
9	42	90
10	55	108

Le componenti principali partendo dalla matrice delle var. covar.

La procedura PRINCOMP

Osservazioni	10
Variabili	2

Statistiche semplici		
	maturita	laurea
Media	46.70000000	99.70000000
StD	9.14148055	7.16550378

Matrice di covarianza		
	maturita	laurea
maturita	83.56666667	51.78888889
laurea	51.78888889	51.34444444

Le componenti principali partendo dalla matrice delle var. covar.

$$\begin{cases} \sigma^2_{Y_1} = \max \\ \underline{\alpha}_1' \underline{\alpha}_1 = 1 \end{cases}$$

$$\begin{cases} (\Sigma - \lambda_1 I)\,\underline{\alpha}_1 = 0 \\ 1 - \underline{\alpha}_1' \underline{\alpha}_1 = 0 \end{cases} \Rightarrow \underline{\alpha}_1' \underline{\alpha}_1 = 1 \quad (*)$$

$$(\Sigma - \lambda_1 I)\,\underline{\alpha}_1 = 0$$

sistema lineare omogeneo che ammette soluzione se

$$|\Sigma - \lambda_1 I| = 0$$

In questo caso $\exists$ p radici caratteristiche ognuna corrispondente ad un vettore.

Le componenti principali partendo dalla matrice delle var. covar.

- $\left|\begin{bmatrix} 83.567 & 51.789 \\ 51.789 & 51.344 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right| = 0$

- Con alcuni passaggi si arriva all'eq. caratteristica di secondo grado
- $\lambda^2 - 134.911\,\lambda + 1608.563 = 0$
- $\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{134.911 \pm \sqrt{134.911^2 - 4\ X\ 1608.563}}{2}$
- $\lambda_1 = 121.693$
- $\lambda_2 = 13.218$

Autovalori della matrice di covarianza				
	Autovalore	Differenza	Proporzione	Cumulativa
1	121.692599	108.474087	0.9020	0.9020
2	13.218512		0.0980	1.0000

$$\begin{cases} \sigma^2_{Y_1} = \max \\ \underline{\alpha}_1' \underline{\alpha}_1 = 1 \end{cases}$$

Poiché vogliamo che $\sigma^2_{Y_1}$ sia massima si sceglie per λ_1 il più elevato degli autovalori delle (*)

- Si sceglie come primo autovalore (radice caratteristica)
- $\lambda_1 = 121.693$
- Si sceglie come secondo autovalore (radice caratteristica)
- $\lambda_2 = 13.218$

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

Varianza totale	134.91111111

- $121.693 + 13.218 = 83.567 + 51.344$

- Inserendo i valori numerici nella

$$(\Sigma - \lambda_1 I)\,\underline{\alpha}_1 = 0$$

$$\left|\begin{bmatrix} 83.567 & 51.789 \\ 51.789 & 51.344 \end{bmatrix} - 121.693 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right| = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Con il vincolo $\underline{\alpha}_1' \underline{\alpha}_1 = 1$
- Si ottiene

Autovettori		
	Prin1	Prin2
maturita	0.805310	-.592853
laurea	0.592853	0.805310

- Analogamente

$$\left|\begin{bmatrix} 83.567 & 51.789 \\ 51.789 & 51.344 \end{bmatrix} - 13.218 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right| = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Si ottiene

Autovettori		
	Prin1	Prin2
maturita	0.805310	-.592853
laurea	0.592853	0.805310

- La prima componente principale si può interpretare come un indice sintetico della riuscita negli studi

$$Y_1 = \sum_{i=1} a_{i1} X_i = X a_1$$

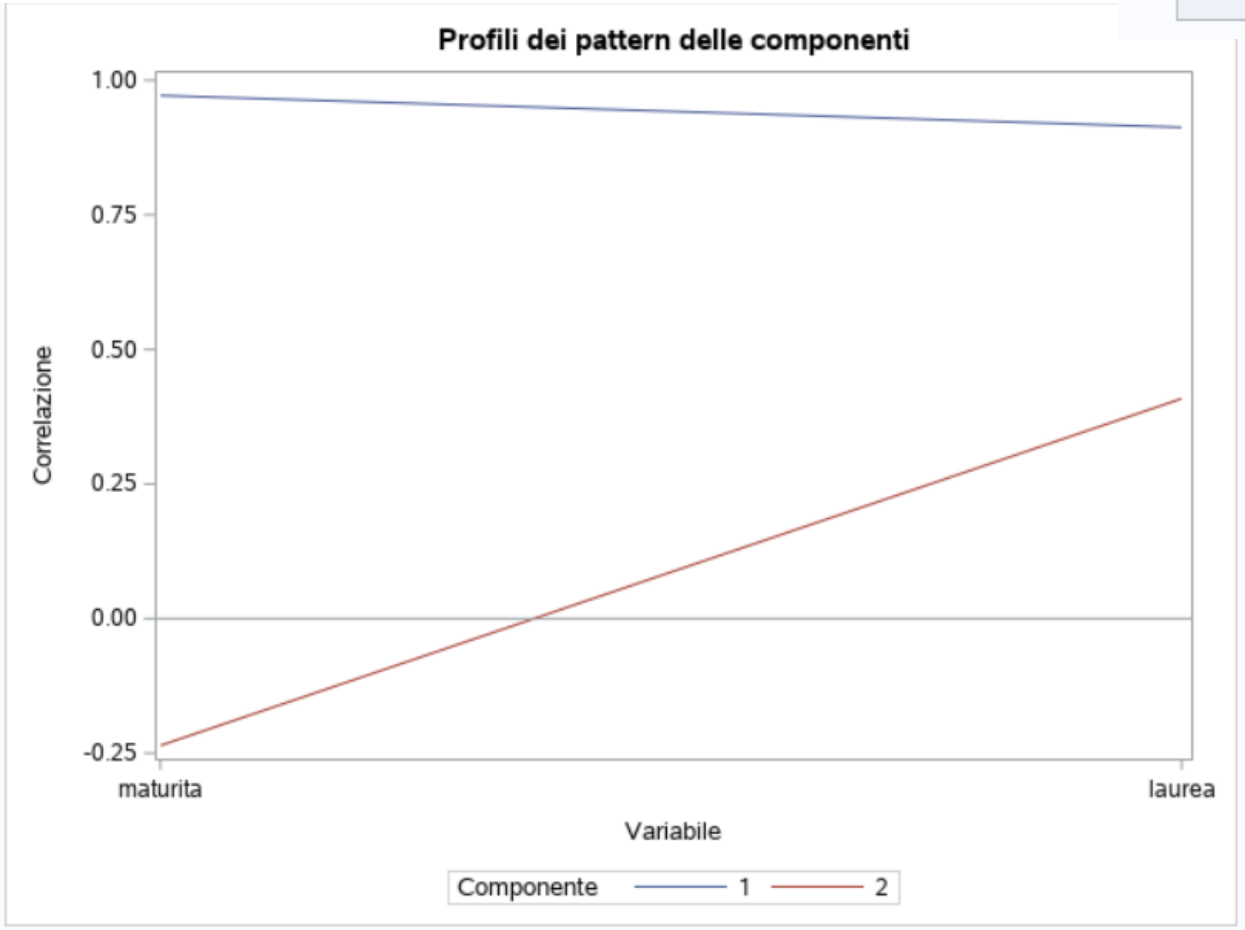
- y_{i1}=0.805 x_{i1} + 0.593 x_{i2}

  $x_{i1} x_{i2}$ *scarti dalla media*

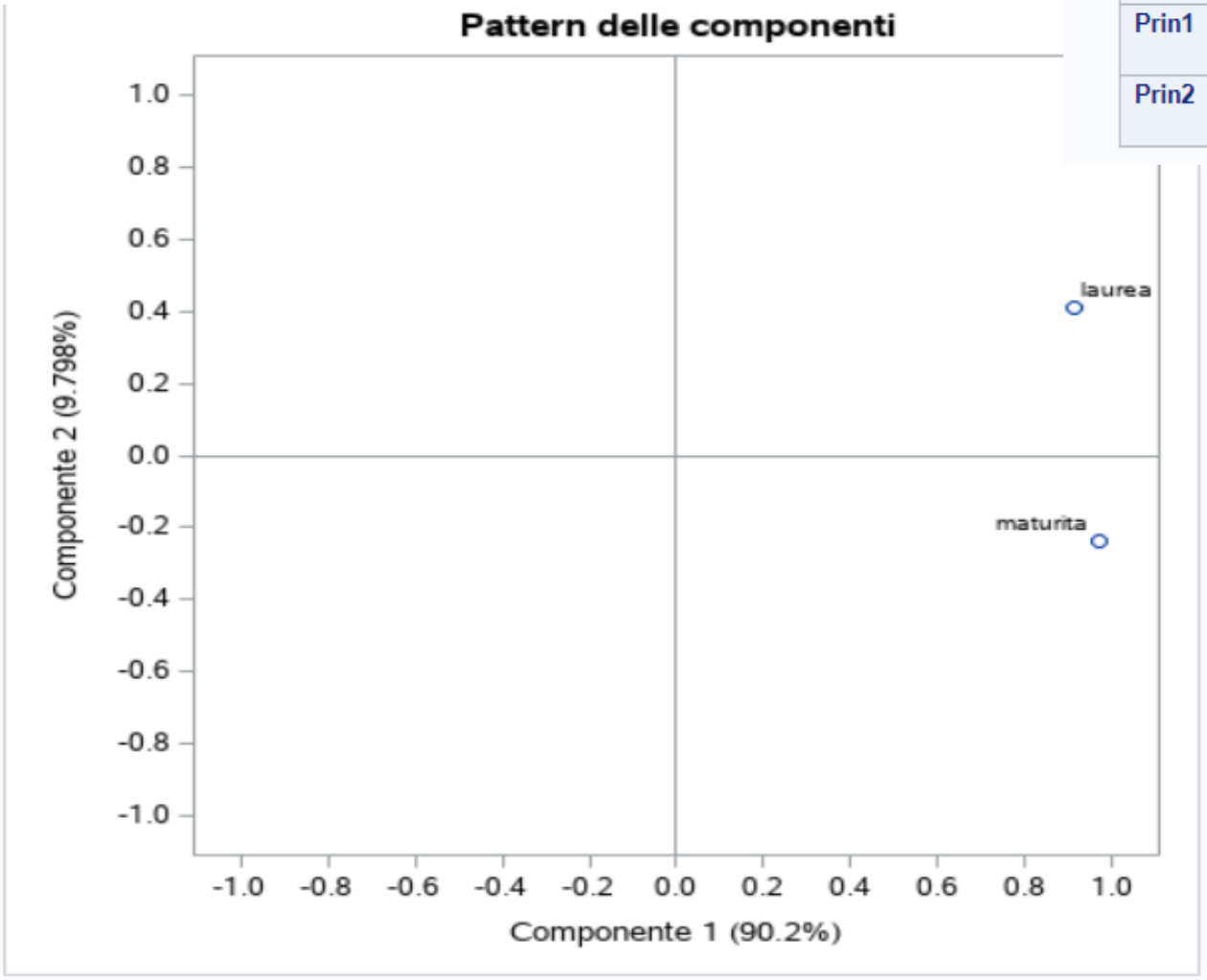
i=1,2,…,10

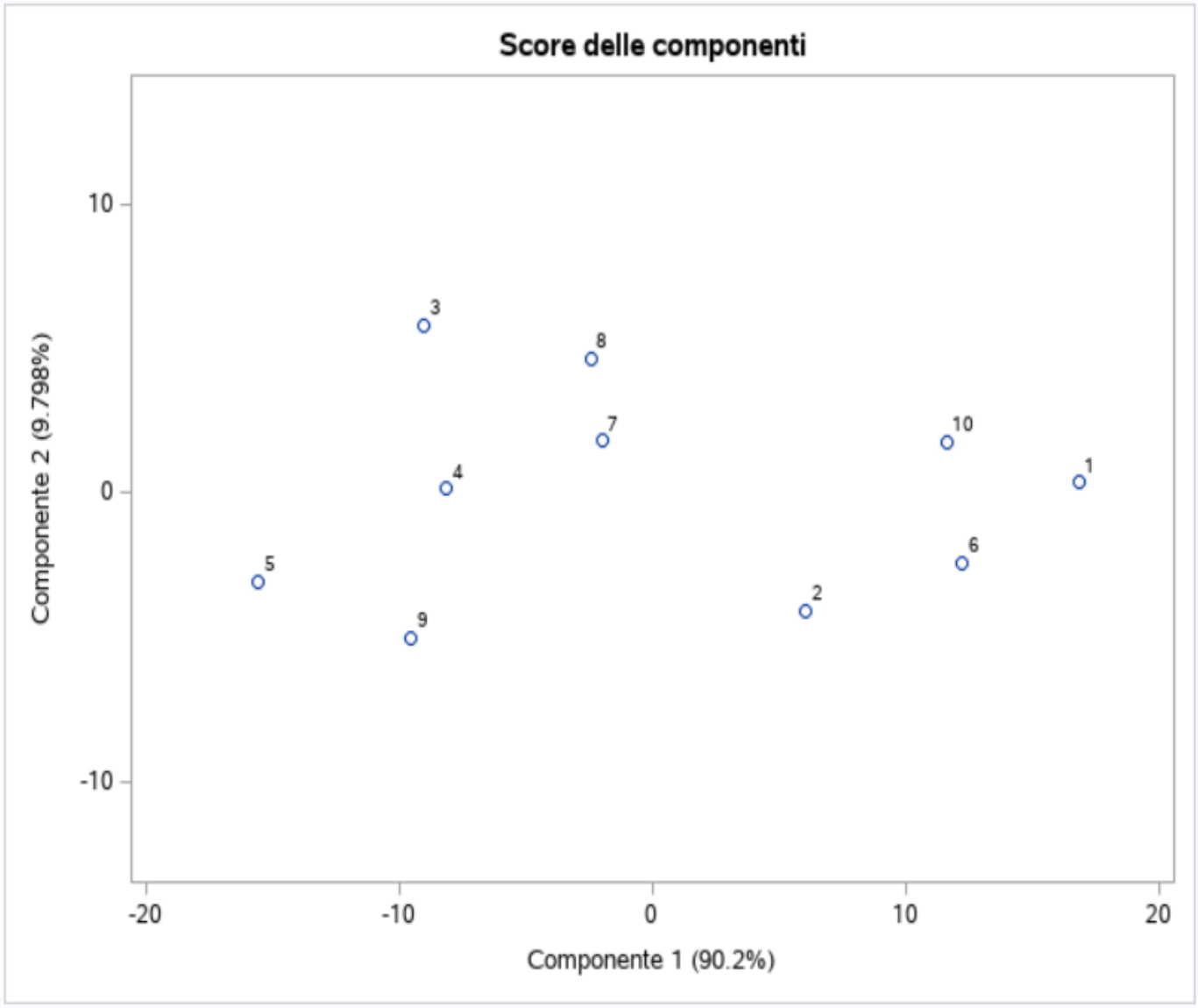
Prin1	Prin2
16.817018332	0.4097492037
6.0566225169	-4.086235883
-9.031819477	5.7798128801
-8.181990709	0.1871577803
-15.55320568	-3.078602438
12.242130908	-2.431096631
-1.996482318	1.842297032
-2.421396702	4.6386245818
-9.535636199	-5.02510122
11.604759332	1.7633946939

Coefficienti di correlazione di Pearson, N = 10 Prob > |r| sotto H0: Rho=0				
	maturita	laurea	Prin1	Prin2
maturita	1.00000	0.79063 0.0065	0.97180 <.0001	-0.23579 0.5119
laurea	0.79063 0.0065	1.00000	0.91271 0.0002	0.40861 0.2410
Prin1	0.97180 <.0001	0.91271 0.0002	1.00000	0.00000 1.0000
Prin2	-0.23579 0.5119	0.40861 0.2410	0.00000 1.0000	1.00000

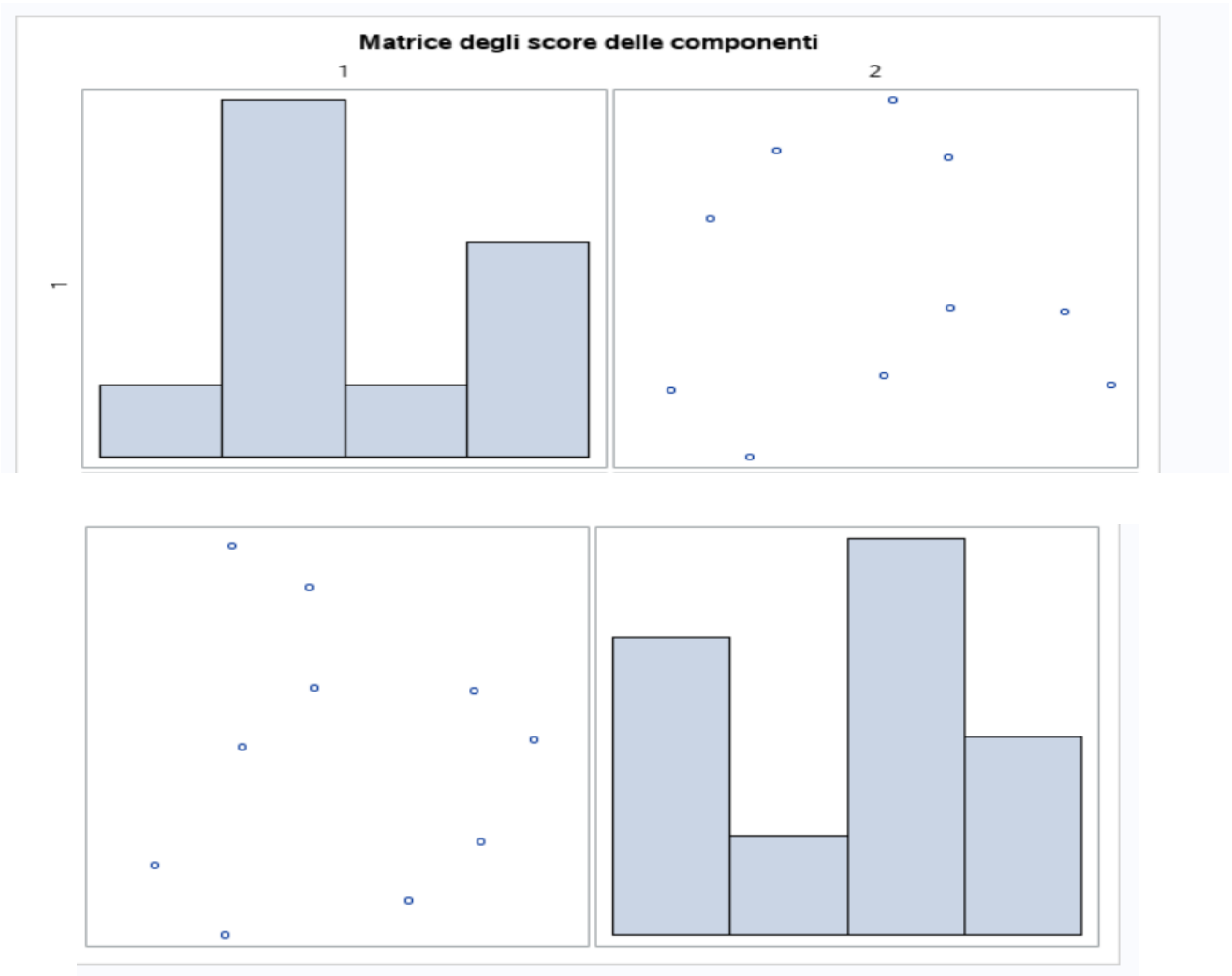


Coefficienti di correlazione di Pearson, N = 10 Prob > |r| sotto H0: Rho=0				
	maturita	laurea	Prin1	Prin2
maturita	1.00000	0.79063 0.0065	0.97180 <.0001	-0.23579 0.5119
laurea	0.79063 0.0065	1.00000	0.91271 0.0002	0.40861 0.2410
Prin1	0.97180 <.0001	0.91271 0.0002	1.00000	0.00000 1.0000
Prin2	-0.23579 0.5119	0.40861 0.2410	0.00000 1.0000	1.00000





Prin1	Prin2
16.817018332	0.4097492037
6.0566225169	-4.086235883
-9.031819477	5.7798128801
-8.181990709	0.1871577803
-15.55320568	-3.078602438
12.242130908	-2.431096631
-1.996482318	1.842297032
-2.421396702	4.6386245818
-9.535636199	-5.02510122
11.604759332	1.7633946939



shows a matrix plot of component scores for the two principal components. The histogram of each component is displayed in the diagonal element of the matrix. The histograms indicate that the first principal component is …. and the second principal component is … .

Le componenti principali partendo dalla matrice delle correlazioni

La procedura PRINCOMP

Osservazioni	10
Variabili	2

Statistiche semplici		
	maturita	laurea
Media	46.70000000	99.70000000
StD	9.14148055	7.16550378

Matrice di correlazione		
	maturita	laurea
maturita	1.0000	0.7906
laurea	0.7906	1.0000

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

Autovettori		
	Prin1	Prin2
maturita	0.707107	0.707107
laurea	0.707107	-.707107

OSS

·) In generale

La h-esima componente principale Y_h è una combinazione lineare delle p variabili $(X_1, \dots X_p)$ con coefficienti uguali alle componenti del vettore caratteristico associato all' h-esimo autovalore in ordine decrescente.

·) Y_1 incorpora la massima variabilità complessiva, Y_2 incorpora la massima variabilità una volta eliminata quella incorporata da Y_1

Fino a che punto spiegare il procedimento?

Dipende dalle esigenze del ricercatore

) Il contributo della h-esima componente principale alla spiegazione della variabilità complessiva si può misurare con il rapporto:

$$\frac{\lambda_h}{\sum_i^p \lambda_i} \times 100$$

Autovalori della matrice di covarianza				
	Autovalore	Differenza	Proporzione	Cumulativa
1	121.692599	108.474087	0.9020	0.9020
2	13.218512		0.0980	1.0000

•) La percentuale di varianza spiegata dalle prime h componenti si può misurare con il rapporto:

$$\frac{\sum_{1}^{h} \lambda_i}{\sum_{1}^{p} \lambda_i} \times 100$$

Autovalori della matrice di covarianza				
	Autovalore	Differenza	Proporzione	Cumulativa
1	121.692599	108.474087	0.9020	0.9020
2	13.218512		0.0980	1.0000

$$\sum_{i=1}^{p} \sigma^2_{X} = \text{tr}\,\Sigma = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \sigma^2_{Y_i}$$ (1)

Tale somma ha senso solo se si opera su variabili espresse nelle stesse unità di misura

⇒ Se si opera su variabili standardizzate ⇒

$$\text{tr}\,\Sigma = K = \sum_{1}^{p} \lambda_i$$

-) Se la matrice di correlazione è diagonale cioè se tutte le variabili sono incorrelate ⇒ le componenti principali sono le stesse p variabili.

) Esistono anche metodi grafici o scree plot

Problema:

E’ possibile assegnare un nome alle componenti?

- Nel es. (partendo dalla matrice di var. covar.) la prima componente principale si può interpretare come un indice sintetico della riuscita negli studi