https://datascience-project.eu/

Training Fiche Template

Title	Analisi delle Componenti principali (ACP)
Keywords (meta tag)	ACP, Correlazione, Variabili quantitative, Varianza spiegata, Autovalori.
Language	Italiano
Objectives / Goals / Learnig outcomes	**l'obiettivo di questo modulo è introdurre e spiegare la tecnica dell'Analisi delle Componenti Principali.** **Alla fine di questo modulo sarai in grado di:** **- Conoscere la logica dell'ACP;** **- Conoscere i requisiti** **- condurre un ACP** **-condurre un ACP in R con il comando FactorMineR**

Training course:	
Data Science Literacy	
Data Visualisation and Visual Analytics Module	X
Introduction to Data science for Human & Social Sciences	
Data Science for good	
Data Journalism and Storytelling	

Description	In questo modulo formativa verrà presentata la tecnica di analisi multidimensionale denominata Analisi in Componenti Principali (ACP) il cui obiettivo è quello di ridurre la dimensionalità di un fenomeno oggetto di indagine preservando l'informazione in quest'ultimo contenuta. La tecnica `e applicabile a fenomeni misurati con variabili quantitative, distinguendosi così da altre tecniche di riduzione della dimensionalità, come l'analisi delle corrispondeze semplici (AC) o l'analisi delle corrispondenze multiple (ACM), sviluppate per l'analisi di variabili qualitative. L'ultima parte del modulo sarà dedicata all'applicazione dell'ACP con il software R.

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Contents arranged in 3 levels	**1. INTRODUZIONE** L'analisi delle componenti principali (ACP) è una tecnica statistica di analisi multivariata per la riduzione delle dimensioni. In pratica, si utilizza quando all'interno di un dataset ci sono molte variabili correlate tra di loro e si vorrebbe ridurne il numero perdendo la minore quantità di informazione possibile. L'ACP ha proprio l'obiettivo di massimizzare la varianza, calcolando il peso da attribuire ad ogni variabile di partenza per poterle concentrare in una o più nuove variabili (dette componenti principali) che saranno combinazione lineare delle variabili di partenza. **2. I REQUISITI DELL'ACP** Affinché sia sensato condurre l'analisi delle componenti principali, è importante analizzare le variabili da utilizzare per avere chiare alcune loro caratteristiche. Nello specifico le variabili devono avere i seguenti requisiti: *- Le variabili devono essere di tipo Quantitativo* L'ACP è valida solo quando le variabili su cui si opera sono di tipo numerico. Se i caratteri hanno diverse unità di misura, bisogna standardizzare le variabili prima di procedere. Tuttavia, in alcuni casi è utilizzata anche per variabili su "scala Likert" e per "variabili binarie". Seppur a livello numerico i risultati siano molto simili tra loro, in questi casi sarebbe preferibile utilizzare metodi alternativi. *- Deve esserci una Correlazione lineare tra le variabili* La prima operazione da fare quando si effettua un'ACP è calcolare la matrice di varianza/covarianza o la matrice di correlazione di Pearson. l'ACP, infatti, è una tecnica utilizzabile quando sono rispettate le ipotesi dell'indice di correlazione lineare di Pearson. I coefficienti di correlazione di Pearson ti informano sul verso e sull'intensità della relazione lineare che intercorre tra i fenomeni. Per interpretarlo, ricordati che più l'indice è vicino a zero, più la relazione sarà debole, più si avvicina a -1 oppure a + 1 più la relazione sarà forte. Nell'ACP valori accettabili per questo indicatore si hanno per R>0,3 o R<-0,3. Se una variabile avesse indici di correlazione molto vicini a 0 con tutte le altre variabili allora quella variabile non dovrebbe essere inclusa nell'ACP. Questo perché il forzare tale variabile a fondersi con altre comporterà una perdita di

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

informazione molto elevata e questa è una situazione che in genere si preferisce evitare.

- *Assenza di outliers*

Come per tutte le analisi basate sulla varianza, singoli valori anomali possono influenzare i risultati soprattutto se molto estremi e se la numerosità campionaria è bassa.

A tal fine è utile realizzare dei *boxplot* oppure grafici a dispersione, detti *scatterplot*, dai quali è possibile dedurre relazioni lineari tra coppie di variabili.

- *Numerosità del campione abbastanza elevata*

Non vi è un valore soglia univoco, ma in generale è consigliabile avere almeno 5-10 unità statistiche per ogni variabile che vuoi includere nell'ACP. Se ad esempio vuoi provare a riassumere con delle nuove componenti 10 variabili, sarebbe consigliabile avere un campione composto da almeno 150 osservazioni.

3. Come Condurre l'ACP

3.1 Dopo aver verificato i requisiti del dataset, controllato che le variabili abbiano le caratteristiche adatte per poter condurre l'analisi delle componenti principali, di seguito mostreremo i diversi passaggi per condurre un ACP:

3.2 Verificare l'Adeguatezza del campione attraverso:

- *Il test Kaiser-Meyer-Olkin*, (KMO), che stabilisce se effettivamente le variabili considerate sono coerenti per l'utilizzo di un'analisi delle componenti principali. Questo indice può assumere valori compresi tra 0 e 1 e, affinché abbia senso effettuare un'analisi delle componenti principali, deve avere un valore almeno superiore a 0,5.

Questo indice può essere calcolato complessivamente per tutte le variabili incluse nella ACP

- *Test di sfericità di Bartlett*: è un test d'ipotesi che ha come *ipotesi nulla* quella che la matrice di correlazione coincida con la matrice identità. Se così fosse, non avrebbe senso performare una ACP in quanto significherebbe che le variabili non sono per nulla correlate linearmente

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

tra loro. Come per tutti i test d'ipotesi, il valore su cui soffermarsi per decidere se rifiutare o meno l'ipotesi nulla è il *p-value*. In questo caso, perché il modello sia considerabile valido, bisogna ottenere un p-value inferiore a 0,05. In questo caso, infatti, si può rifiutare con un livello di significatività del 5% l'ipotesi nulla.

3.3 Estrazione delle componenti principali:

La parte fondamentale dell'ACP è stabilire il numero di fattori adeguato che possa meglio rappresentare le variabili di partenza.

Per comprendere meglio questo concetto, immagina che il tuo dataset sia una città a te sconosciuta, ed ogni componente principale sia una strada di questa città. Se tu volessi conoscere questa città, quante strade visiteresti? Probabilmente partiresti dalla via centrale (la prima componente principale) e poi ti addentreresti in altre vie. Ma in quante?

Per poter dire di conoscere a sufficienza una città, ovviamente il numero di vie da visitare cambia a seconda delle dimensioni della città e di quante le vie sono simili o diverse tra loro. Allo stesso modo, il numero di componenti da estrarre dipende da quante variabili hai scelto di includere all'interno dell'analisi delle componenti principali e da quanto queste sono simili tra loro. Più sono correlate infatti, minore sarà il numero di componenti principali necessario per ottenere una buona conoscenza delle variabili di partenza. Al contrario, meno sono correlate, maggiore sarà il numero di componenti principali da estrarre per poter avere un'informazione accurata del dataset.

I criteri utilizzati per la scelta del numero di componenti sono essenzialmente due: gli autovalori maggiori di 1 e l'analisi parallela.

Gli autovalori maggiori di 1

Secondo questa regola, si scelgono quelle componenti a cui è associato un autovalore superiore a 1. L'autovalore è un numero che fornisce la varianza spiegata dalla componente: siccome inizialmente la varianza spiegata da ogni singola variabile è pari a 1, non avrebbe senso prendere una componente (che è una combinazione di variabili) con varianza inferiore a 1. Ad un autovalore alto corrisponde una maggiore varianza e i software come SPSS o R restituiscono questa tabella con valori

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

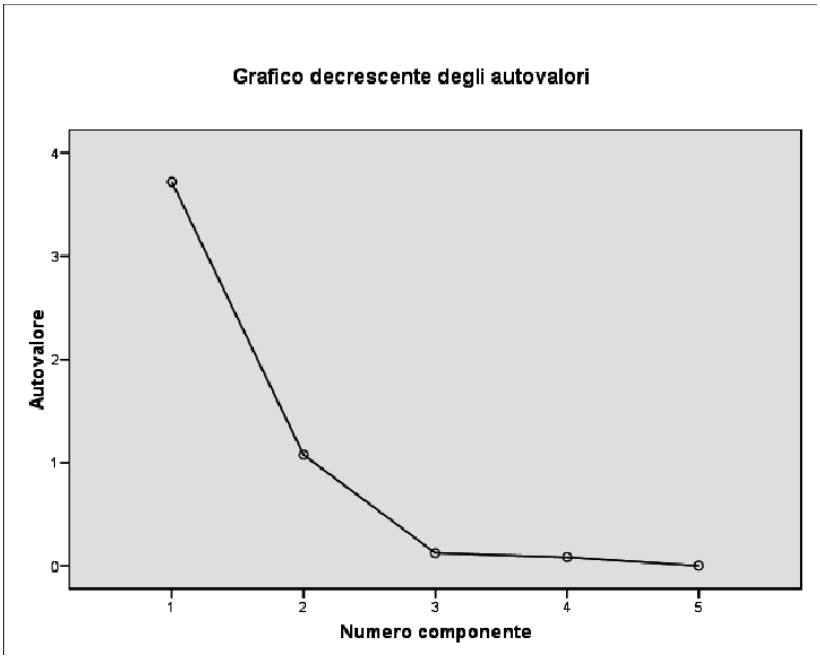
decrescenti; pertanto, il primo sarà sempre associato al fattore più importante.

Proporzione di varianza spiegata

Seguendo questo criterio, le componenti da estrarre devono garantire che almeno il 70% della variabilità complessiva delle variabili di partenza non venga persa. Inoltre, ogni singola componente da estrarre dovrebbe apportare un incremento di rilievo alla varianza complessiva (ad esempio, almeno un 5% o un 10% in più di variabilità spiegata).

Scree-plot

Questo metodo, si basa su un grafico in cui sull'asse verticale sono riportati i valori degli autovalori e sull'asse orizzontale tutte le possibili componenti da estrarre (che saranno quindi in numero pari alle variabili di partenza). Unendo i punti si otterrà una linea spezzata che in alcune parti avrà una forma concava ed in altri convessa.



Come è possibile vedere dal grafico sull'asse x sono elencate le componenti, mentre sull'asse y ci sono gli autovalori. Quando la curva di

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

questo grafico fa un “gomito” è il momento per tracciare una linea, e prendere in considerazione solo i fattori che stanno sopra.

Dal grafico che puoi vedere qui sopra, ad esempio, si vede che il numero di punti che si trovano sopra al gomito è 2.

La parte conclusiva dell’ACP consiste nel dare un nome alle singole componenti principali trovate.

4. L’ACP con R

Con i software statistici (come ad esempio SPSS, Jamovi e R) l’ACP è un’operazione molto semplice. Bastano pochi click per riuscire ad ottenere un output da interpretare. Non c’è quindi un software preferibile agli altri in quanto è una tecnica molto utilizzata e tutti i programmi statistici ne permettono l’esecuzione in modo agevole e senza dover effettuare calcoli a mano. Ad ogni modo in questo modulo mostreremo come condurre l’ACP con il software R.

Nel power point correlato a questo modulo sarà rappresentato tutto il percorso per implementare l’ACP su R, ovvero:

- ✓ Svolgendo tutti i passaggi che si fondano sulle dimostrazioni matriciali, geometriche e statistiche;
- ✓ Attraverso il comando diretto PCA del pacchetto FactoMineR.

In questo modulo ci limiteremo a presentare il pacchetto FactoMineR.

FactoMineR è in grado di svolgere l'analisi in componenti principali riducendo la dimensionalità dei dati multivariati a due o tre che possono essere, così, visualizzati graficamente con una minima perdita di informazioni e ciò si può fare utilizzando un solo comando, ossia **PCA**, tra parentesi inseriremo la matrice oggetto di analisi

```
X <- as.matrix(DATASET)

library(FactoMineR)

res.pca = PCA(DATASET)
```

Con il comando *summary* possiamo vedere l'importanza delle componenti in termini di deviazione standard, proporzione della

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

varianza spiegata e la varianza spiegata cumulata, sia per quanto riguarda gli individui che le variabili.

summary(res.pca)

Con il comando *head*

head(ris.pca$eig)

invece si può calcolare l'importanza degli autovalori. Il comando, infatti, ci darà i valori degli autovalori, la percentuale della varianza spiegata e della varianza spiegata cumulata per ogni variabile.

Esempio dì ciò che vedremo su R

```{r}
head(ris.pca$eig)
```

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.8198226	70.495565	70.49557
comp 2	0.5141619	12.854049	83.34961
comp 3	0.3589118	8.972796	92.32241
comp 4	0.3071036	7.677590	100.00000

Infine per poter disegnare lo scree-plot degli autovalori tra parentesi inseriremo l'oggetto di analisi

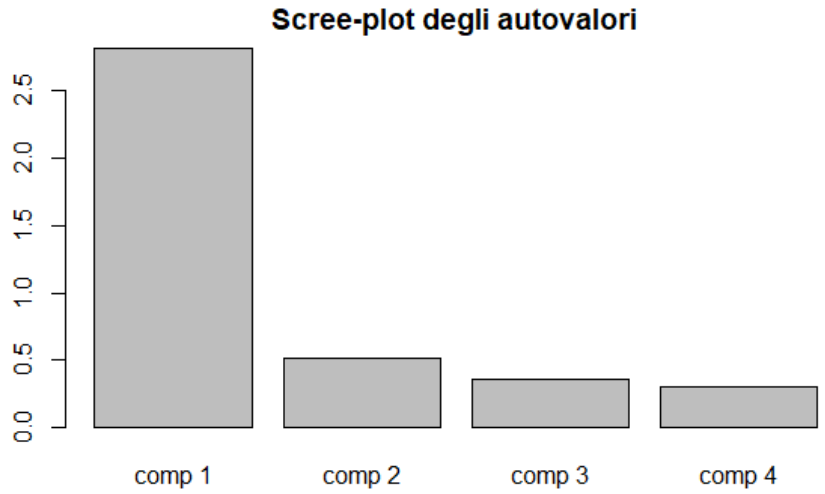
barplot(res.pca$eig[,1], main="Scree-plot degli autovalori")

Con il comando main indichiamo il titolo del grafico.

Esempio dì ciò che vedremo su R

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Un altro pacchetto utile per l'ACP (ma che non tratteremo in questo modulo) è *factoextra* che fornisce alcune funzioni di facile utilizzo per estrarre e visualizzare i risultati che abbiamo dalle analisi multivariate, tra cui ACP (analisi in componenti principali), AC (analisi delle corrispondenze semplici), ACM (analisi delle corrispondenze multiple), AMF (analisi dei fattori multipli), HMFA (analisi gerarchica dei fattori multipla).

Self-assessment (multiple choice queries and answers)

1. L' Analisi in Componenti Principali ha come obiettivo:
   - A) L' aggregazione di unità statistiche in funzione della loro distanza
   - B) **La riduzione della dimensionalità di un fenomeno complesso**
   - C) La descrizione di un dataset

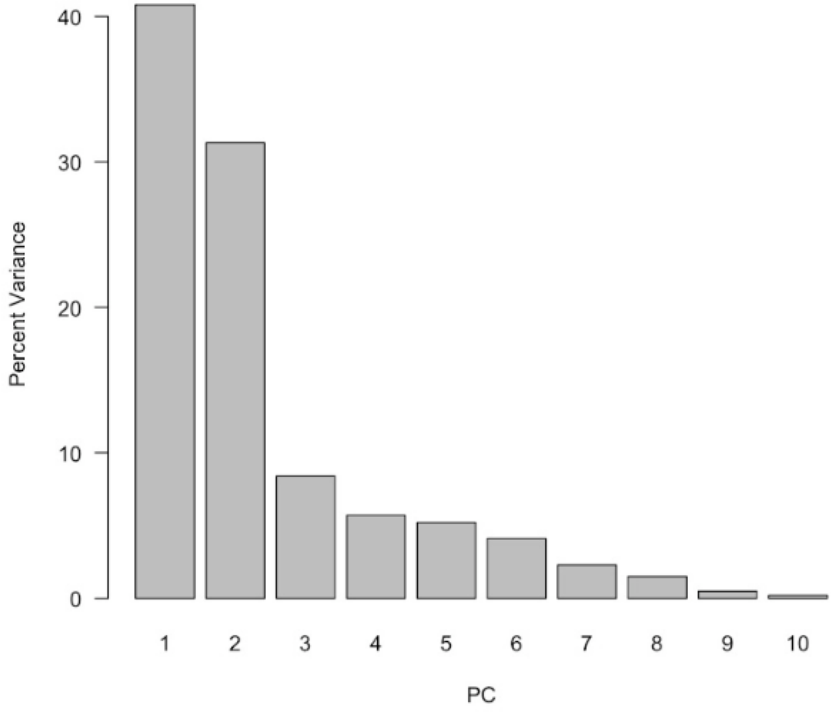
2. La matrice di dati di partenza di un ACP deve essere
   - A) Con dati qualitativi
   - B) Con dati standardizzati
   - **C) Con dati quantitativi**

3. Le componenti estratte nell'Analisi in Componenti Principali:
   - A) Sono combinazioni lineari delle variabili di partenza
   - B) Godono della proprietà equidistributiva
   - **C) Presentano tutte autovalori maggiori di 1**

4. Con quante dimensioni spieghereste il seguente fenomeno?

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



A. Una
B. Due
C. Tre

Resources (videos, reference link)

Pozzolo P., *Analisi delle componenti principali: da dove partire*, https://paolapozzolo.it/analisi-delle-componenti-principali-criteri/

Gilardone A., *Analisi delle componenti principali: 7 passaggi da eseguire* https://adrianogilardone.com/analisi-delle-componenti-principali/

Gilardone A., https://www.youtube.com/watch?v=OksC-g4K2gY

Vardanega A., L'Analisi in componenti principali

https://www.agnesevardanega.eu/wiki/r/analisi_esplorativa/analisi_in_componenti_principali

Zakaria Jaadi, *A Step-by-Step Explanation of Principal Component Analysis (PCA)*, https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Ian T. Jolliffe and Jorge Cadima, *Principal component analysis: a review and recent developments*, https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	Science Snippets Blog, *What Is Principal Component Analysis (PCA) and How It Is Used?*, 2020 https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186
Related material	
Related PPT	
Bibliography	
Provided by	[UNISALENTO/DEMOSTENE CENTRO STUDI]

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.