https://datascience-project.eu/

Fișă de învățare

Titlu	Analiza Componentelor Principale (ACP)
Cuvinte Cheie (meta tag)	ACP, Corelație, variabile cantitative, varianță explicată, valori proprii.
Limbă	Română
Obiective / Scop / Rezultatele învățării	**Scopul acestui modul este de a introduce și de a explica tehnica Analizei Componentelor Principale.** **La finalul acestui modul, vei fi capabil să:** **- Cunoști logica ACP** **- Cunoști cerințele** **- Realizezi o Analiză în Componente Principale** **- Realizezi ACP în R utilizând pachetul FactorMineR**

Curs:	
Data Science Literacy	
Vizualizarea Datelor și Modulul de Visual Analytics	X
Introducere în Data science pentru Științe sociale	
Data Science for good	
Data Journalism și Storytelling	

Descriere	În acest modul de învățare este prezentată tehnica multidimensională denumită Analiza Componentelor Principale (ACP), al cărei obiectiv este de a reduce dimensionalitatea unui fenomen investigat, păstrând în același timp informația conținută de acesta. Tehnica poate fi aplicată fenomenelor măsurate cu variabile cantitative, deosebindu-se astfel de alte tehnici de reducere a dimensionalității, cum ar fi analiza corespondențelor – cazul bidimensional sau multidimensional, dezvoltată pentru analiza variabilelor calitative. Ultima parte a acestui modul de învățare va fi dedicată aplicării tehnicii ACP în 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.

Conținutul este organizat pe 3 niveluri	**1. INTRODUCERE** Analiza componentelor principale (ACP) este o tehnică de analiză multivariată pentru reducerea dimensionalității. În practică, este utilizată atunci când există multe variabile corelate în setul de date, pentru a reduce numărul acestora, pierzând cât mai puțină informație. ACP are drept scop maximizarea varianței, calculând ponderea atribuită fiecărei variabile la start pentru a le putea concentra într-una sau mai multe variabile noi (denumite componente principale) care vor fi combinații liniare ale variabilelor de start. **2. Cerințele Analizei Componentelor Principale** Pentru a înțelege dacă are sens să se efectueze o analiză a componentelor principale, este important să fie analizate variabilele utilizate pentru a avea o imagine clară asupra caracteristicilor lor. Mai exact, variabilele trebuie să îndeplinească următoarele condiții: - *Variabilele trebuie să fie cantitative* ACP poate fi aplicată doar dacă variabilele sunt numerice. În cazul în care unitățile de măsură sunt diferite, variabilele trebuie standardizate înainte de a efectua procedura. Totuși, în unele cazuri tehnica este aplicată si pentru variabile măsurate pe scala ”Likert” și pentru variabile binare. Deși din punct de vedere numeric rezultatele sunt foarte asemănătoare, în aceste cazuri ar fi de preferat utilizarea unor metode alternative. - *Trebuie să existe o corelație liniară între variabile* Prima etapă care trebuie parcursă atunci când este realizată ACP este calculul matricii de varianță/covarianță sau a matricii de corelație Pearson. ACP este de fapt o tehnică care poate fi aplicată atunci când ipotezele coeficientului de corelație liniară Pearson sunt respectate. Coeficienții de corelație Pearson oferă informații despre direcția și intensitatea legăturii liniare dintre fenomene. Pentru a-l interpreta, să ne amintim că, cu cât are o valoare mai apropiată de zero, cu atât mai slabă va fi legătura dintre variabile, în timp ce cu cât e mai aproape de +1 sau -1, cu atât mai puternică va fi legătura. În ACP, valorile acceptabile pentru acest indicator sunt R>0.3 sau R<-0.3. Dacă o variabilă are coeficienți de corelație foarte apropiați de 0 cu toate

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.

celelalte variabile, atunci acea variabilă nu ar trebui să fie inclusă în ACP. Forțând variabila să fie agregată cu alte variabile va rezulta o pierdere foarte mare de informație, lucru care în general este bine să fie evitat.

- Lipsa valorilor extreme (outlier-ilor)

Așa cum este cazul pentru toate analizele bazate pe varianță, valorile extreme pot influența rezultatele analizei, în special dacă valorile extreme sunt foarte mari și eșantionul este de dimensiuni mici.

În acest sens, este utilă crearea de box-plot-uri și de scatter-plot-uri, din care este posibilă deducerea relațiilor liniare dintre perechi de variabile.

- Dimensiunea suficient de mare a eșantionului

Nu există o valoare prag univocă, dar în general este recomandat să existe cel puțin 5-10 unități statistice pentru fiecare variabilă care se dorește a fi inclusă în ACP. De exemplu, dacă se încearcă sistematizarea a 10 variabile în componente noi, ar fi de dorit ca eșantionul să aibă cel puțin 150 de observații.

3. Cum se realizează ACP

3.1 După verificarea condițiilor care trebuie îndeplinite de setul de date, verificarea faptului că variabilele au caracteristicile necesare pentru a realiza analiza componentelor principale, se vor parcurge următorii pași pentru a realiza ACP:

3.2 Verificarea gradului de adecvare a eșantionului, cu ajutorul:

- Testului Kaiser-Meyer-Olkin (KMO), care stabilește dacă variabilele considerate sunt consistente pentru a putea fi utilizate în analiza componentelor principale. Indicele ia valori înre 0 și 1, iar pentru a face sens într-o analiză a componentelor principale, trebuie să aibă o valoare mai mare de 0.5.

Acest indice poate fi calculat pe ansamblu, pentru toate variabilele incluse în ACP.

-Testului de sfericitate Bartlett: este un test de testare a ipotezei statistice, având drept ipoteză nulă că matricea de corelație coincide cu matricea identitate. Dacă acesta este cazul, nu are sens să fie realizată ACP, întrucât ar însemna că variabilele nu sunt deloc corelate liniar între

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.

ele. Ca orice testare de ipoteză statistică, decizia privind respingerea ipotezei nule se bazează pe *p-value*. În acest caz, pentru ca modelul să fie considerat valid, trebuie obținut un p-value mai mic de 0.05. Deci, ipoteza nulă poate fi respinsă la un prag de semnificație de 5%.

3.3 Extragerea componentelor principale:

Partea crucială a ACP este stabilirea numărului adecvat de factori care pot reprezenta cel mai bine variabilele de start.

Pentru a înțelege mai bine conceptul, să ne imaginăm că setul de date este ca un oraș necunoscut și fiecare componentă principală este o stradă în acest oraș. Dacă am vrea să cunoaștem orașul, câte străzi am vizita? Am începe probabil cu strada centrală (prima componentă principală) și apoi am explora celelalte străzi. Cât de multe străzi ar trebui să explorăm?

Pentru a putea spune că am ajuns că cunoaștem suficient de bine orașul, numărul de străzi variază în funcție de mărimea orașului și cât de similare sau de diferite sunt străzile. În mod similar, numărul de componente care vor fi extrase depinde de numărul de variabile care sunt incluse în analiza componentelor principale și cât de similare sunt acestea între ele. De fapt, cu cât sunt mai corelate, cu atât este mai scăzut numărul de componente principale necesare pentru a obține o imagine cât mai fidelă a variabilelor de start. Din contră, dacă gradul de corelare dintre ele este scăzut, va trebui extras un număr mai mare de componente principale pentru a avea informații cât mai precise pentru setul de date.

Criteriile utilizate pentru a alege numărul de componente sunt două: valori proprii mai mari de 1 și analiză paralelă.

Valori proprii mai mari de 1

Conform acestei reguli, se vor alege componentele care au asociată o valoare proprie mai mare de 1. Valoarea proprie este un număr care arată varianța explicată de componentă: întrucât inițial varianța explicată de fiecare variabilă este egală cu 1, nu ar face sens să fie aleasă o componentă (care este o combinație de variabile) cu varianța mai mică de 1. O valoare proprie mare corespunde unei variațe mai mari și un software ca SPSS sau R afișează aceste valori în tabel în ordine

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.

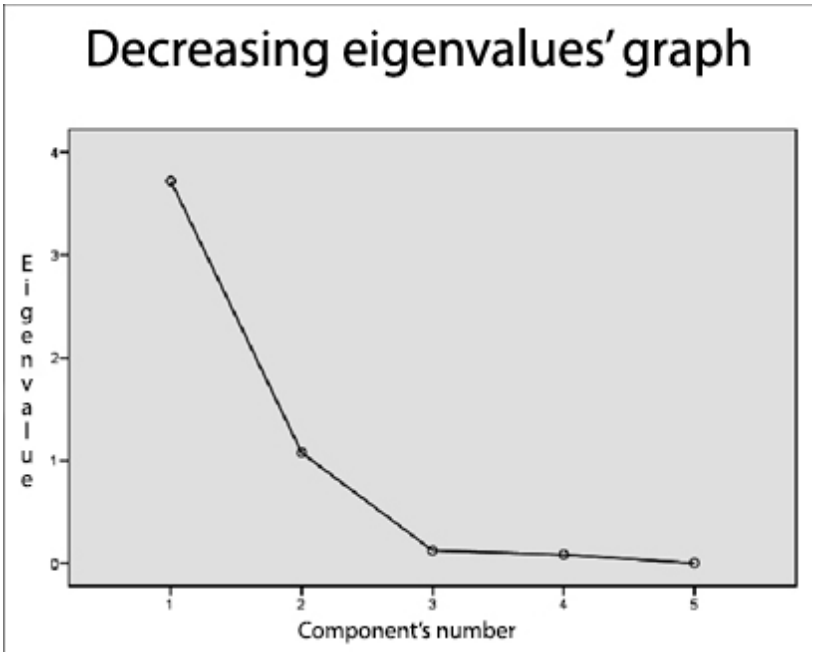
descrescătoare; astfel, prima componentă va fi asociată întotdeauna cu cel mai important factor.

Proporția din varianța totală explicată

Pe baza acestui criteriu, componentele extrase trebuie să asigure că cel puțin 70% din variabiltatea globală a variabilelor de start nu este pierdută. Mai mult, fiecare componentă extrasă ar trebui să aducă o creștere semnificativă în varianța de ansamblu (de exemplu, cel puțin 5% sau 10% mai mult la variabilitatea explicată).

Scree-plot

Această metodă se bazează pe un grafic în care valorile proprii sunt afișate pe axa verticală și toate componentele posibil a fi extrase sunt pe axa orizontală (care va fi deci egală ca număr cu cel al variabilelor de start). Prin unirea punctelor se va obține o linie frântă care în unele părți va avea o formă concavă, iar în alte părți o formă convexă.



După cum se poate observa din grafic, componentele sunt afișate pe axa Ox, în timp ce valorile proprii sunt pe axa Oy. Când curba de pe acest grafic formează un "elbow" (o cotitură), este momentul trasării unei linii si se vor lua în considerare numai factorii care se situează deasupra.

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.

Din graficul de mai sus, de exemplu, se poate observa că numărul de puncte deasupra ”elbow” (cotiturii) este 2.

Ultima parte a ACP constă în atribuirea unor nume componentelor principale găsite.

4. ACP în software-ul R

Utilizând un software statistic (cum ar fi SPSS, Jamovi sau R), ACP este o tehnică foarte simplă de realizat. Câteva click-uri sunt suficiente pentru a obține un tabel de rezultate care poate fi interpretat. Astfel, nu există un software care să fie preferat față de altul, ACP fiind o tehnică utilizată intensiv și deci toate programele statistice o au implementată, putând fi realizată ușor, fără a realiza calcule manuale. În acest modul vom arăta cum se poate efectua ACP cu ajutorul software-ului R.

Întregul proces de implementare a ACP în R va fi prezentat în documentul powerpoint atașat acestui modul, mai exact:

- ✓ Realizarea tuturor etapelor care se bazează pe demonstrații matematice (calcul matricial), geometrice sau statistice;
- ✓ Prin intermediul comenzii directe PCA (ACP) din pachetul FactoMineR.

În acest modul vom prezenta doar pachetul FactoMineR.

FactoMineR poate efectua o analiză a componentelor principale, reducând dimensionalitatea datelor multivariate la două sau trei variabile, care pot fi ulterior prezentate grafic cu o pierdere minimă de informație. Acest lucru poate fi realizat utilizând o singură comandă, **PCA,** inserând obiectul matrice de analiză între paranteze, în comandă.

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

library(FactoMineR)

res.pca = PCA(DATASET)
```

Cu ajutorul comenzii *summary* putem vedea importanța componentelor în termeni de abatere standard, proporție a variației explicate și varianță cumulată explicată, atât pentru elemente cât și pentru variabile.

```
summary(res.pca)
```

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.

În schimb, cu ajutorul comenzii *head*

head(ris.pca$eig)

se poate calcula importanța valorilor proprii. De fapt, comanda oferă valorile proprii, procentul din varianța explicată și varianța cumulată explicată pentru fiecare variabilă.

Exemplu cu ceea ce se obține în R

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

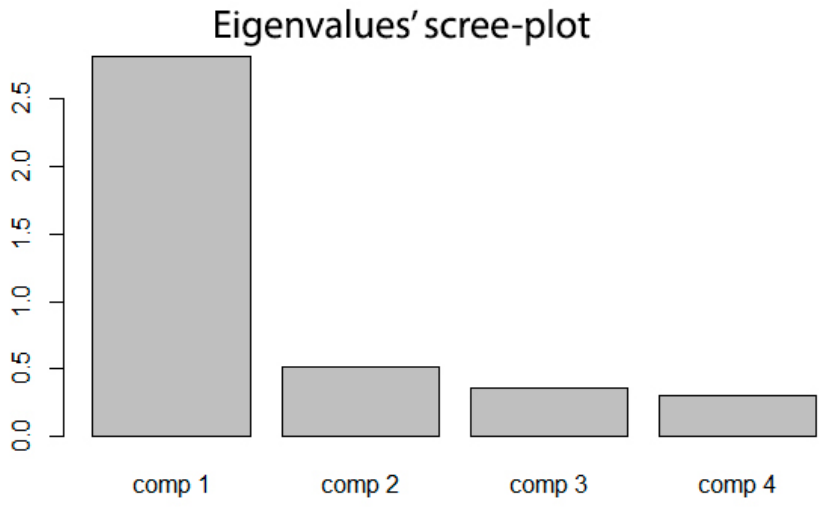
	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.8198226	70.495565	70.49557
comp 2	0.5141619	12.854049	83.34961
comp 3	0.3589118	8.972796	92.32241
comp 4	0.3071036	7.677590	100.00000

În final, pentru a vizualiza scree-plot-ul valorilor proprii, vom insera obiectul de analiză între paranteze.

barplot(res.pca$eig[,1], main="Eigenvalues' scree-plot")

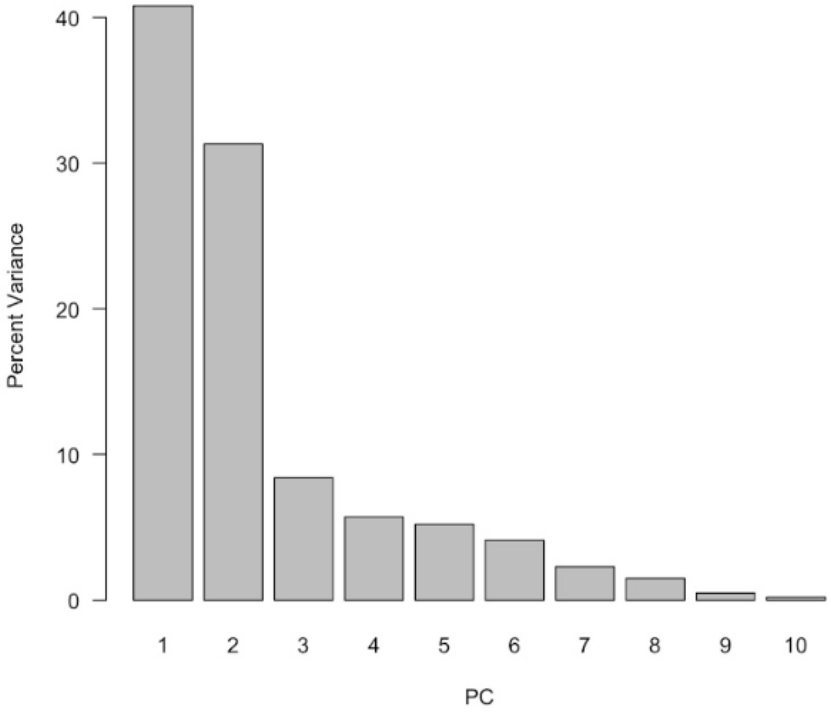
Cu ajutorul comenzii Main se indică titlul graficului.

Exemplu cu ceea ce se obține în R



Un alt pachet util pentru ACP (pe care nu îl vom aborda în acest modul) este *factoextra*, care oferă anumite funcții ușor de utilizat pentru a extrage și a vizualiza rezultatele obținute din analizele multivariate,

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.

	incluzând analiza componentelor principale, analiza corespondențelor – cazurile bi- și multidimensionale, analiză factorială multiplă, analiză factorială multiplă ierarhică.
Auto-evaluare (întrebări cu răspuns multiplu și răspunsuri)	1. Obiectivul Analizei Componentelor Principale este: A) Agregarea unităților statistice în funcție de distanță B) Reducerea dimensionalității unui fenomen complex C) Descrierea unui set de date 2. Matricea de date de start în ACP trebuie să fie: A) Cu date calitative B) Cu date standardizate C) Cu date cantitative 3. Componentele extrase în urma Analizei Componentelor Principale: A) Sunt combinații liniare ale variabilelor de start B) Au proprietatea de echidistribuție C) Toate au valori proprii mai mari de 1 4. Cu câte dimensiuni se poate explica următorul fenomen?  A. Una B. Două C. Trei

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.

Resurse (video, link-uri)	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 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
Materiale adiționale	
PPT	
Bibliografie	
Realizat de:	[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.