dipl. ing. Zoran Radojičić

LINEARNI STATISTIČKI MODELI

Beograd 1999.

SADRŽAJ:

1. ELEMENTI MATRIČNE ALGEBRE

1.1 Vektori

Niz $\boldsymbol{x}$ realnih brojeva $x_1, x_2, \ldots, x_n$ se naziva ***vektor*** i piše

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

ili

$$x' = \begin{bmatrix} x_1 & x_2 & \vdots & x_n \end{bmatrix}$$

ili

$$x = \begin{bmatrix} x_1 & x_2 & \vdots & x_n \end{bmatrix}'$$

pri čemu se poslednja operacija naziva ***transponovanje kolone u vrstu.***

Nad vektorima $\boldsymbol{x}$ i $\boldsymbol{y}$ su definisane sledeće operacije:

$$cx = \begin{bmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_n \end{bmatrix}$$

$$x + y = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}$$

$$x'y = x_1y_1 + x_2y_2 + \ldots + x_ny_n \qquad \text{(unutrašnji proizvod)}$$

$$x'x = x_1^{\ 2} + x_2^{\ 2} + \ldots + x_n^{\ 2}.$$

Za dva vektora ***x*** i ***y*** koji su iste dimenzije se kaže da su ***linearno zavisni*** ako postoje konstante c_1 i c_2 različite od nule takve da je

$$c_1 x + c_2 y = 0.$$

U opštem slučaju, za skup vektora $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, ..., $\boldsymbol{x_k}$ se kaže da su linearno zavisni ako postoje konstante c_1, c_2, ..., c_k koje nisu sve jednake nuli tako da je

$$c_1 x_1 + c_2 x_2 + \ldots + c_k x_k = 0.$$

Linearna zavisnost implicira da se barem jedan vektor iz skupa može predstaviti kao linearna kombinacija ostalih.

Dužina vektora x, u oznaci L_x se definiše kao

$$L_x = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} = \sqrt{x'x}\,.$$

1.2 Matrice

Matrica je pravougaona šema sastavljena od p redova (vrsta) i n kolona sa $p \times n$ elemenata:

$$\underset{(p\times n)}{A} = \begin{bmatrix} a_{11} & a_{12} & \vdots & a_{1n} \\ a_{21} & a_{22} & \vdots & a_{2n} \\ \cdots & \cdots & + & \cdots \\ a_{p1} & a_{p2} & \vdots & a_{pn} \end{bmatrix} = \left[a_{ij}\right]_{p\times n}.$$

Transponovana matrica A' se dobija uzajamnom razmenom mesta redova i kolona i to onim redom kako se javljaju u matrici ***A***

$$\underset{(n\times p)}{A'} = \begin{bmatrix} a_{11} & a_{21} & \vdots & a_{p1} \\ a_{12} & a_{22} & \vdots & a_{p2} \\ \cdots & \cdots & + & \cdots \\ a_{1n} & a_{2n} & \vdots & a_{pn} \end{bmatrix} = A^T.$$

Množenje matrice konstantom c:

$$c\underset{(p\times n)}{A} = \begin{bmatrix} ca_{11} & ca_{12} & \vdots & ca_{1n} \\ ca_{21} & ca_{22} & \vdots & ca_{2n} \\ \cdots & \cdots & + & \cdots \\ ca_{p1} & ca_{p2} & \vdots & ca_{pn} \end{bmatrix}.$$

Dve matrice ***A*** i ***B*** istih dimenzija se mogu sabrati:

$$A+B=\left[a_{ij}\right]_{p\times n}+\left[b_{ij}\right]_{p\times n}=\left[a_{ij}+b_{ij}\right]_{p\times n}$$

Da bi se dve matrice ***A*** i ***B*** mogle međusobno pomnožiti, neophodno je da budu ***saglasne,*** tj. da matrica ***A*** ima onoliko kolona koliko matrica **B** ima redova:

$$\underset{(p\times k)}{A}\ \underset{(k\times n)}{B}=\underset{(p\times n)}{C},\qquad \text{tj.}\quad c_{ij}=\sum_{l=1}^{k}a_{il}b_{lj}.$$

Kvadratna matrica je simetrična ako je ***A=A'*** odnosno $a_{ij}=a_{ji}\ \forall\ i,j$.

Primer 1.1:

matrica $\begin{bmatrix}3 & 5\\ 5 & -2\end{bmatrix}$ je simetrična,

matrica $\begin{bmatrix}3 & 6\\ 4 & -2\end{bmatrix}$ nije simetrična.

Nula matrica predstavlja matricu čiji su svi elementi jednaki nuli i može biti bilo kog tipa:

$$0=\left[\begin{array}{cc:c}0 & 0 & 0\\ 0 & 0 & 0\\ \hdashline 0 & 0 & 0\end{array}\right].$$

Jedinična matrica I predstavlja matricu kod koje su svi članovi na glavnoj dijagonali jednaki jedinici a svi ostali nuli:

$$I=\left[\begin{array}{cc:c}1 & 0 & 0\\ 0 & 1 & 0\\ \hdashline 0 & 0 & 1\end{array}\right].$$

Determinanta kvadratne matrice ima formu

$$\det A=\left|A\right|=\left|\begin{array}{cc:c}a_{11} & a_{12} & a_{1n}\\ a_{21} & a_{22} & a_{2n}\\ \hdashline a_{n1} & a_{n2} & a_{nn}\end{array}\right|.$$

Minor elementa a_{ij} u determinanti ***A*** je determinanta reda $n-1$ koja se dobija iz *det A* kada se izbrišu njena i-ta vrsta i j-ta kolona.

Rang matrice je definisan kao veličina najveće determinante koja je različita od nule i koja se može formirati od matrice ***A***. Drugačije rečeno, rang matrice je najveći red minora te matrice koji je različit od nule.

Za kvadratnu matricu se kaže da je ***singularna*** ako je njena determinanta jednaka nuli. Ako je *det A≠0,* kvadratna matrica ***A*** je ***regularna*** (nesingularna). Matrica je sigurno singularna u sledećim slučajevima:

- ako su svi elementi jedne vrste ili kolone jednaki nuli,
- ako su vrste ili kolone matrice linearno zavisne.

Singularna matrica nema svoju inverznu matricu.

Kvadratna matrica $\boldsymbol{A^{-1}}$ se naziva ***inverzna matrica*** date kvadratne matrice ***A*** ako je:

$$A^{-1}A = AA^{-1} = I\,.$$

Za svaku ***dijagonalnu*** matricu

$$D = \left[\begin{array}{cc:c} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ \hdashline 0 & 0 & d_n \end{array}\right]$$

čiji su svi dijagonalni elementi $d_i \neq 0$, inverzna će biti:

$$D^{-1} = \left[\begin{array}{cc:c} \frac{1}{d_1} & 0 & 0 \\ 0 & \frac{1}{d_2} & 0 \\ \hdashline 0 & 0 & \frac{1}{d_n} \end{array}\right].$$

Pozitivno definitna matrica je takva kvadratna matrica ***A*** koja ima osobinu da je odgovarajuća kvadratna forma

$$X'AX > 0$$

za svaki vektor $\boldsymbol{X \neq 0}$.

Ako važi relacija

$$X'AX \geq 0$$

tada se matrica ***A,*** kao i odgovarajuća kvadratna forma, naziva ***pozitivno semidefinitnom***. Ako je simetrična matrica ***A*** pozitivno definitna onda je i matrica $\boldsymbol{A^{-1}}$ pozitivno definitna. Pozitivno definitna matrica ***A*** je regularna, tj. *det A≠0*.

Neka je $\boldsymbol{A}_{(k\times k)}$ kvadratna matrica i $\boldsymbol{I}_{(k\times k)}$ jedinična matrica. Skalari λ_1, λ_2, ..., λ_k koji zadovoljavaju jednačinu

$$|A-\lambda I|=0 \quad \text{odnosno} \quad \det(A-\lambda I)=0 \qquad (*)$$

nazivaju se ***karakteristični koreni*** matrice ***A.*** Jednačina (*) se naziva ***karakteristična jednačina***.

Primer 1.2:

Za sledeću matricu

$$A=\begin{bmatrix}1 & 0\\ 1 & 3\end{bmatrix}$$

se dobijaju dva karakteristična korena:

$$|A-\lambda I|=\left|\begin{bmatrix}1 & 0\\ 1 & 3\end{bmatrix}-\lambda\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right|=\begin{vmatrix}1-\lambda & 0\\ 1 & 3-\lambda\end{vmatrix}=(1-\lambda)(3-\lambda)=0$$

$$\lambda_1=1,\ \lambda_2=3.$$

Primer 1.3:

$$A=\begin{bmatrix}13 & -4 & 2\\ -4 & 13 & -2\\ 2 & -2 & 10\end{bmatrix}$$

$$|A-\lambda I|=\left|\begin{bmatrix}13 & -4 & 2\\ -4 & 13 & -2\\ 2 & -2 & 10\end{bmatrix}-\lambda\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\right|=\begin{vmatrix}13-\lambda & -4 & 2\\ -4 & 13-\lambda & -2\\ 2 & -2 & 10-\lambda\end{vmatrix}=$$

$$=-\lambda^3+36\lambda^2-405\lambda+1458=0$$

$$\lambda_1=9,\ \lambda_2=9,\ \lambda_3=18.$$

Neka je ***A*** matrica dimenzije $k\times k$ i λ njen karakteristični koren. Tada, ako važi

$$\underset{(k\times 1)}{x}\neq\underset{(k\times 1)}{0} \qquad \text{i} \qquad Ax=\lambda x$$

za ***x*** se kaže da predstavlja ***karakteristični vektor*** matrice ***A*** pridružen karakterističnom korenu λ.

Ekvivalentan uslov je:

$$(A-\lambda I)x=0.$$

Primer 1.4:

Odrediti karakteristične vektore matrice

$$A=\begin{bmatrix}1 & 0\\ 1 & 3\end{bmatrix}$$

Iz prethodnog primera smo dobili da je $\lambda_1=1$ i $\lambda_2=3$.

$$Ax=\lambda_1 x \Leftrightarrow \begin{bmatrix}1 & 0\\ 1 & 3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\lambda_1\begin{bmatrix}x_1\\ x_2\end{bmatrix}\Leftrightarrow \begin{bmatrix}1 & 0\\ 1 & 3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=1\begin{bmatrix}x_1\\ x_2\end{bmatrix} \qquad (*)$$

$$x_1=x_1$$
$$x_1+3x_2=x_2 \Rightarrow x_1=-2x_2$$

$$Ax=\lambda_2 x \Leftrightarrow \begin{bmatrix}1 & 0\\ 1 & 3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\lambda_2\begin{bmatrix}x_1\\ x_2\end{bmatrix}\Leftrightarrow \begin{bmatrix}1 & 0\\ 1 & 3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=3\begin{bmatrix}x_1\\ x_2\end{bmatrix} \qquad (**)$$

$$x_1=3x_1$$
$$x_1+3x_2=3x_2 \Rightarrow x_1=0$$

Iz (*) sledi da ima beskonačno mnogo rešenja za x_1 i x_2. Tako, ako uzmemo da je $x_2=1$ dobijamo da je $x_1=-2$, tj.

$$x=\begin{bmatrix}-2\\ 1\end{bmatrix}$$

karakteristični vektor koji odgovara karakterističnom korenu $\lambda_1=1$.

Iz izraza (**) dobijamo da je $x_1=0$ a za x_2 uzimamo proizvoljnu vrednost 1 i dobijamo

$$x=\begin{bmatrix}0\\ 1\end{bmatrix}$$

karakterističan vektor koji odgovara karakterističnom korenu $\lambda_2=3$.

1.3 Neke važnije osobine matrica

Neka su ***A*** i ***B*** *k×k* kvadratne matrice. Tada važi:

a) $|A| = |A'|$

b) Ako su svi elementi jedne vrste (kolone) matrice ***A*** jednaki nuli, tada je:

$$|A| = 0$$

c) Ako su bilo koje dve vrste (kolone) matrice ***A*** identične, tada je:

$$|A| = 0$$

d) Ako je ***A*** nesingularna matrica, tada je:

$$|A| = \frac{1}{|A^{-1}|}, \qquad \text{tj. } |A||A^{-1}| = 1$$

e) $|AB| = |A||B|$

f) $|cA| = c^k |A|$, gde je c skalar.

Neka je ***A*** *k×k* kvadratna matrica. ***Trag matrice A***, u oznaci *tr(A)* je jednak zbiru elemenata na glavnoj dijagonali:

$$tr(A) = \sum_{i=1}^{k} a_{ii} .$$

Neka su ***A*** i ***B*** *k×k* kvadratne matrice i c skalar. Tada važi:

a) $tr(cA) = c\,tr(A)$

b) $tr(A \pm B) = tr(A) \pm tr(B)$

c) $tr(AB) = tr(BA)$

d) $tr(B^{-1}AB) = tr(A)$

e) $tr(AA') = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}^2$

2. VRSTE PODATAKA I MERNE SKALE

Statistička obeležja mogu biti ***kvantitativna*** i ***kvalitativna*** (nominalna). Kvantitativna obeležja se dele na ***neprekidna*** (jedinica mere se može beskonačno deliti) i ***prekidna***.

$X = \left(X_1 \quad X_2 \quad \vdots \quad X_p \right)^T, \quad p \geq 1$ vektor obeležja (promenljivih),

$E = \{e_1, e_2, \ldots, e_N\}, \quad N \geq p$ skup jedinica posmatranja (entiteta).

Primer 2.1:

Iz skupa objekata (skup ljudi) izabrana su 3 objekta (3 muškarca) stara između 30 i 40 godina i nazvana e_j, e_k i e_q. Pretpostavimo da nas o objektima e_j, e_k i e_q zanimaju sledeće karakteristike:

X_1 – broj dece (dobijen brojanjem)
X_2 – broj članova uže porodice (dobijen brojanjem)
X_3 – telesna masa (dobijena merenjem u gramima)
X_4 – inteligencija (kao broj rešenih zadataka od ukupno 60)
X_5 – obrazovanje (stepen stručne spreme: 1 magistrat, 2 fakultet, 3 viša škola)
X_6 – sportska igra koju najviše vole.

Dobijeni su sledeći rezultati:

	X_1	X_2	X_3	X_4	X_5	X_6
e_j	2	4	64000	20	3	fudbal
e_k	0	1	86520	40	1	košarka
e_q	4	6	128000	0	5	fudbal

Tabela 2.1

Skup objekata predstavlja skup jedinica posmatranja (entiteta): {Pera, Mika, Žika}. Karakteristike predstavljaju obeležja. Tabelu 2.1 nazivamo ***matrica podataka*** i ona je oblika

$$X = \left[x_{ij} \right]_{N \times p}$$

odnosno, u našem slučaju

$$X = \left[x_{ij} \right]_{3 \times 6}$$

gde je x_{ij} ***realizacija*** x_i na e_j.

Kako za svaku od ispitanih karakteristika entiteti mogu imati različite rezultate to ćemo rezultate dobijene utvrđivanjem neke karakteristike nazivati ***promenljivim*** (varijablama).

Osobine definisanih varijabli:

X_1
- broj dece (prekidna kvantitativna varijabla)
- uzima vrednost iz skupa pozitivnih celobrojnih vrednosti uključujući i nulu
- rezultati se ne mogu na smislen način podvrgnuti nikakvoj transformaciji već moraju ostati takvi kakvi jesu
- kako su merne jedinice definisane onim što se meri reč je o ***apsolutnim skalama***

X_2
- broj članova uže porodice (prekidna kvantitativna)
- varijabla je tako definisana da i sam entitet ulazi u broj članova uže porodice (npr. e_k je očigledno samac bez dece)
- iz ovoga bi se moglo zaključiti da se u zaključivanju ništa nebi promenilo ako od svakog rezultata oduzmemo jedinicu pa prema tome X_2 dobija vrednosti:
  3
  0
  5
  prema tome, veličina varijabli ovog tipa je određena do neke proizvoljne konstante koju možemo dodati ili oduzeti pa se zato ove skale nazivaju ***aditivne skale***

X_3
- telesna masa (neprekidna kavntitativna)
- ovde je reč o ***skali odnosa*** koja ima sledeće osobine:
  a) količnik ma koje dve vrednosti ima smislenu interpretaciju
     npr. e_q je dva puta teži od e_j
  b) rastojanje između dva objekta mereno na ma kom delu ove skale je jednako
     npr. e_j ima 22520 grama manje od entiteta e_k, a e_q 64000 grama više od e_j
  c) vrednostima pozicioniranim na skali mogu se dodeliti rangovi od višeg ka nižem
     npr. e_q ima viši rang od e_k

X_4
- inteligencija (neprekidna kvantitativna)
- nema realnu nultu tačku (nema ljudskog stvora sa inteligencijom nula)
- merne jedinice su proizvoljne (bodovi na testu)
- poseduje osobine b i c (za drugi test druge vrednosti)
- ove skale se nazivaju ***intervalnim*** jer je jedino interval (razmak) između rezultata definisan
  npr. razlika u inteligenciji između entiteta e_j i e_q je jednaka kao i između entiteta e_j i e_k

X_5
- obrazovanje (kvalitativna varijabla koja u ovom slučaju uslovno može da prođe kao kvantitativna, kao stepen stručne spreme)
- vrednosti varijable ne označavaju količinu već rang, pa se zato nazivaju ***ordinarnim varijablama*** ili varijablama koje leže na nekoj ordinarnoj skali
- poseduju samo osobinu c

X_6
- sportska igra koju najviše vole (kvalitativna varijabla)
- rezultat ne govori ništa ni o količini ni o redosledu već samo o tome koje svojstvo ima neki entitet iz grupe međusobno isključivih svojstva
- kako oznaka pripadanja ne označava nikakvu veličinu, varijable ovog tipa se nazivaju ***kvalitativnim*** ili nominalnim varijablama

Klasifikacija metoda multivarijacione analize u zavisnosti od vrste promenljivih:

kvantitativne promenljive:

- metoda glavnih komponenti
- faktorska analiza
- klaster analiza
- višestruka korelacija
- višestruka regresija
- *man*-ova metoda (multidimenziona analiza varijanse)
- kanonična korelaciona analiza

kvalitativne promenljive:

- loglinearni modeli
- diskriminaciona analiza (samo zavisna promenljiva mora da bude kvalitativna)

3. VIŠEDIMENZIONALNE SLUČAJNE PROMENLJIVE

3.1 Sredina i kovarijaciona matrica

Označimo p jednodimenzionalnih slučajnih promenljivih sa $X_1, X_2, \ldots, X_p$. Skup ovih slučajnih promenljivih nazivamo ***slučajni vektor X*** (vektor obeležja)

$$X = \begin{pmatrix} X_1 & X_2 & \vdots & X_p \end{pmatrix}^T, \quad p \geq 1.$$

Realizaciju $\boldsymbol{X}$ na e_i označavamo sa

$$x_{(i)} = \begin{pmatrix} x_{i1} & x_{i2} & \vdots & x_{ip} \end{pmatrix}^T.$$

Za svaku jednodimenzionalnu slučajnu promenljivu možemo odrediti ***sredinu*** $\mu_i=E(X_i)$ i ***varijansu*** $\sigma_i^2=E(X_i-\mu_i)^2$, $i=1, 2, \ldots, p$.

Sredina slučajnog vektora $\boldsymbol{X}$ je

$$\mu = E(X) = \begin{pmatrix} E(X_1) & E(X_2) & \vdots & E(X_p) \end{pmatrix}^T = \begin{pmatrix} \mu_1 & \mu_2 & \vdots & \mu_p \end{pmatrix}^T.$$

Za ma koji par slučajnih promenljivih X_i i X_j definišemo ***kovarijansu***

$$\sigma_{ij} = E\left[(X_i - \mu_i)(X_j - \mu_j)\right] = Cov(X_i, X_j)$$
$$\sigma_{ii} = \sigma_i^2 = Var(X_i)$$
$$\sigma_{ij} = \sigma_{ji}.$$

Za slučajni vektor $\boldsymbol{X}$ se definiše $p \times p$ simetrična matrica koja se naziva ***kovarijaciona matrica*** od $\boldsymbol{X}$ sa znakom $\boldsymbol{\Sigma}$:

$$\Sigma = \left[\sigma_{ij}\right]_{p\times p} = \left[\begin{array}{cc:c} \sigma_{11} & \sigma_{12} & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{2p} \\ \hdashline \sigma_{p1} & \sigma_{p2} & \sigma_{pp} \end{array}\right].$$

Kovarijaciona matrica se može iskazati i kao očekivana vrednost ***slučajne matrice***

$$\Sigma = \left[\begin{array}{cc:c} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & (X_1-\mu_1)(X_p-\mu_p) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & (X_2-\mu_2)(X_p-\mu_p) \\ \hdashline (X_p-\mu_p)(X_1-\mu_1) & (X_p-\mu_p)(X_2-\mu_2) & (X_p-\mu_p)^2 \end{array}\right].$$

Na kraju, kovarijacionu matricu možemo iskazati kao očekivanu vrednost sledeće ***matrice odstupanja:***

$$W = \left[(X_i-\mu_i)(X_j-\mu_j)\right]_{p\times p}$$
$$\Sigma = E(W) = E\left[(X-\mu)(X-\mu)^T\right].$$

3.2 Podela vektora

Često se tokom istraživanja javlja potreba da p promenljivih posmatramo razdvojeno:

$$X = \left(\begin{array}{cc:c|cc:c} X_1 & X_2 & X_q & X_{q+1} & X_{q+2} & X_p \end{array}\right)^T.$$

U tom slučaju vršimo podelu sredine μ i kovarijacione matrice $\boldsymbol{\Sigma}$

$$\mu = (\mu_1 \quad \mu_2)^T$$
$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^{\ T} & \Sigma_{22} \end{bmatrix}$$

gde su

$$\Sigma_{11} = \left[\sigma_{ij}^{(1)}\right]_{q\times q}$$
$$\Sigma_{22} = \left[\sigma_{ij}^{(2)}\right]_{(p-q)\times(p-q)}$$
$$\Sigma_{12} = \left[\sigma_{ij}^{(3)}\right]_{(p-q)\times q}$$

3.3 Uzoračka sredina i kovarijaciona matrica

Primer 3.1:

Za matricu podataka

	X_1	X_2	X_3	X_4	X_5	X_6
e_j	2	4	64000	20	3	fudbal
e_k	0	1	86520	40	1	košarka
e_q	4	6	128000	0	5	fudbal

odrediti realizovanu vrednost sredine vektora $\boldsymbol{X}$ za prve dve varijable (X_1 – broj dece, X_2 – broj članova uže porodice).

Slučajni vektor (vektor obeležja)

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \hdashline X_p \end{bmatrix} = \left(X_1 \quad X_2 \;\vdots\; X_p \right)^T$$

u našem slučaju je

$$X = \left(X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad X_6 \right)^T, \quad p = 6.$$

Skup jedinica posmatranja (entiteta)

$$E = \left\{ e_1, e_2, \ldots, e_N \right\}$$

u našem slučaju je

$$E = \left\{ e_j, e_k, e_q \right\}, \quad N = 3.$$

Matrica podataka

$$X = \left[x_{ij} \right]_{N \times p}$$

u našem slučaju je

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & x_{16} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & x_{26} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & x_{36} \end{bmatrix}$$

tj.

$$X = \begin{bmatrix} 2 & 4 & 64000 & 20 & 3 & fudbal \\ 0 & 1 & 86520 & 40 & 1 & kosarka \\ 4 & 6 & 128000 & 0 & 5 & fudbal \end{bmatrix}.$$

U daljem radu se nećemo sretati sa nominalnim varijablama u matrici podataka dok ćemo matricu podataka zadavati uglavnom na ovaj način.

U ovom primeru je reč o uzorku iz ***višedimenzionalnog rasporeda.*** Višedimenzionalni raspored se opisuje nekom ***merom centralne tendencije*** i nekom ***merom disperzije*** (rasipanja).

Za meru centralne tendencije najčešće se uzima ***sredina slučajnog vektora X*** koja se označava sa μ

$$\mu = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \hdashline E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \hdashline \mu_p \end{bmatrix} = \left(\mu_1 \quad \mu_2 \;\vdots\; \mu_p\right)^T.$$

Za meru varijabiliteta najčešće se uzima ***kovarijaciona matrica*** od ***X*** koja se označava sa *Cov(X)* ili Σ

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \vdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \vdots & \sigma_{2p} \\ \hdashline \sigma_{p1} & \sigma_{p2} & \vdots & \sigma_{pp} \end{bmatrix}$$

gde je $\sigma_{ii}=\sigma_i^2$, $i=1, 2, \ldots, p$.

Ako definišemo simetričnu matricu kvadrata

$$\begin{bmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \vdots & (X_1-\mu_1)(X_p-\mu_p) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \vdots & (X_2-\mu_2)(X_p-\mu_p) \\ \hdashline (X_p-\mu_p)(X_1-\mu_1) & (X_p-\mu_p)(X_2-\mu_2) & \vdots & (X_p-\mu_p)^2 \end{bmatrix}$$

koja je proizvod slučajnih vektora odstupanja od sredine, tj.

$$(X-\mu)(X-\mu)'$$

onda je njena očekivana vrednost jednaka kovarijacionoj matrici:

$$E\left[(X-\mu)(X-\mu)'\right] = \Sigma.$$

Iz višedimenzionalne raspodele se može izračunati i ***korelaciona matrica*** ρ i to preko kovarijacione matrice. U matričnoj notaciji to izgleda ovako:

$$\rho = \left(D^{\frac{1}{2}} \right)^{-1} \Sigma \left(D^{\frac{1}{2}} \right)^{-1}$$

gde je ***D*** dijagonalna matrica koja sadrži elemente na glavnoj dijagonali kovarijacione matrice ***(matrica standardnih devijacija)***, tj.

$$\rho = \left[\begin{array}{cc:c} \frac{1}{\sqrt{\sigma_{11}}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & 0 \\ \hdashline 0 & 0 & \frac{1}{\sqrt{\sigma_{pp}}} \end{array}\right] \left[\begin{array}{cc:c} \sigma_{11} & \sigma_{12} & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{2p} \\ \hdashline \sigma_{p1} & \sigma_{p2} & \sigma_{pp} \end{array}\right] \left[\begin{array}{cc:c} \frac{1}{\sqrt{\sigma_{11}}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & 0 \\ \hdashline 0 & 0 & \frac{1}{\sqrt{\sigma_{pp}}} \end{array}\right] =$$

$$= \left[\begin{array}{cc:c} 1 & \rho_{12} & \rho_{1p} \\ \rho_{21} & 1 & \rho_{2p} \\ \hdashline \rho_{p1} & \rho_{p2} & 1 \end{array}\right]$$

gde je ρ_{ij} koeficijent korelacije između X_i i X_j:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}.$$

Na osnovu uspostavljene relacije može se napisati da je

$$\Sigma = D^{\frac{1}{2}} \rho D^{\frac{1}{2}}.$$

Na osnovu matrice podataka smo u mogućnosti da izračunamo pokazatelje uzorka: sredinu uzorka, kovarijacionu matricu uzorka i korelacionu matricu uzorka.

Uzoračka sredina je data sa:

$$\overline{X} = \left(\overline{X}_1 \quad \overline{X}_2 \;\vdots\; \overline{X}_p \right)^T = \left[\begin{array}{c} \overline{X}_1 \\ \overline{X}_2 \\ \hdashline \overline{X}_p \end{array}\right]$$

gde je

$$\overline{X}_j = \frac{1}{N}\sum_{i=1}^{N} X_{ij}, \quad j = 1,2,\ldots,p\,.$$

Uzoračku kovarijacionu matricu definišemo preko matrice ***centriranih podataka X**** i matrice ***uzajamnih proizvoda centriranih podataka X*' X****

$$X^* = \left[\begin{array}{cc:c} X_{11}-\overline{X}_1 & X_{12}-\overline{X}_2 & X_{1p}-\overline{X}_p \\ X_{21}-\overline{X}_1 & X_{22}-\overline{X}_2 & X_{2p}-\overline{X}_p \\ \hdashline X_{N1}-\overline{X}_1 & X_{N2}-\overline{X}_2 & X_{Np}-\overline{X}_p \end{array}\right]$$

odnosno

$$X^{*\prime}X^* = \left[\begin{array}{ccc:c} \sum_{i=1}^{N}(X_{i1}-\overline{X}_1)^2 & \sum_{i=1}^{N}(X_{i1}-\overline{X}_1)(X_{i2}-\overline{X}_2) & & \sum_{i=1}^{N}(X_{i1}-\overline{X}_1)(X_{ip}-\overline{X}_p) \\ \sum_{i=1}^{N}(X_{i2}-\overline{X}_2)(X_{i1}-\overline{X}_1) & \sum_{i=1}^{N}(X_{i2}-\overline{X}_2)^2 & & \sum_{i=1}^{N}(X_{i2}-\overline{X}_2)(X_{ip}-\overline{X}_p) \\ \hdashline \sum_{i=1}^{N}(X_{ip}-\overline{X}_p)(X_{i1}-\overline{X}_1) & \sum_{i=1}^{N}(X_{ip}-\overline{X}_p)(X_{i2}-\overline{X}_2) & & \sum_{i=1}^{N}(X_{ip}-\overline{X}_p)^2 \end{array}\right]_{p\times p}$$

pa je uzoračka kovarijaciona matrica $\boldsymbol{S_{(p\times p)}}$

$$S = \frac{1}{N-1}X^{*\prime}X^* = \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\overline{X})(X_i-\overline{X})'.$$

Uzoračka korelaciona matrica je

$$R = \left[\begin{array}{ccc:c} 1 & r_{12} & r_{13} & r_{1p} \\ r_{12} & 1 & r_{23} & r_{2p} \\ r_{13} & r_{23} & 1 & r_{3p} \\ \hdashline r_{1p} & r_{2p} & r_{3p} & 1 \end{array}\right]$$

gde su r_{ij} ***uzorački koeficijenti korelacije***

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}, \quad i,j = 1,2,\ldots,p$$

a S_{ij} ***elementi uzoračke kovarijacione matrice***

$$S_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}(X_{ki}-\overline{X}_i)(X_{kj}-\overline{X}_j), \quad i,j = 1,2,\ldots,p\,.$$

Do uzoračke kovarijacione matrice smo mogli doći i koristeći relaciju koju smo uspostavili između kovarijacione i korelacione matrice populacije:

$$R = \left(D^{\frac{1}{2}} \right)^{-1} S \left(D^{\frac{1}{2}} \right)^{-1}.$$

Vratimo se primeru. Interesuje nas matrica podataka

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 0 & 1 \\ 4 & 6 \end{bmatrix} \qquad i = 1,2,3 \Rightarrow N = 3, \qquad j = 1,2 \Rightarrow p = 2.$$

Treba odrediti $\overline{X}$:

$$\overline{X} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_p \end{bmatrix} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{3}(2+0+4) \\ \frac{1}{3}(4+1+6) \end{bmatrix} = \begin{bmatrix} 2 \\ 3.67 \end{bmatrix}$$

što govori da je prosečan broj dece $\overline{X}_1 = 2$ a prosečan broj članova uže porodice $\overline{X}_2 = 3.67$.

Primer 3.2:

Realizovane vrednosti slučajnog uzorka su date matricom podataka

$$X = \begin{bmatrix} -2 & 3 \\ 4 & 2 \\ 1 & 1 \end{bmatrix}.$$

Odrediti realizovanu vrednost uzoračke kovarijacione matrice.

Prvo ćemo izračunati uzoračku sredinu $\overline{X}$:

$$\overline{X} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{3}(-2+4+1) \\ \frac{1}{3}(3+2+1) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad N = 3, p = 2.$$

Zatim računamo matricu centriranih podataka X^*:

$$X^* = \left[\begin{array}{cc|c} X_{11}-\overline{X}_1 & X_{12}-\overline{X}_2 & X_{1p}-\overline{X}_p \\ X_{21}-\overline{X}_1 & X_{22}-\overline{X}_2 & X_{2p}-\overline{X}_p \\ \hline X_{N1}-\overline{X}_1 & X_{N2}-\overline{X}_2 & X_{Np}-\overline{X}_p \end{array}\right].$$

Kako je $N=3$ i $p=2$

$$X^* = \begin{bmatrix} X_{11}-\overline{X}_1 & X_{12}-\overline{X}_2 \\ X_{21}-\overline{X}_1 & X_{22}-\overline{X}_2 \\ X_{31}-\overline{X}_1 & X_{31}-\overline{X}_2 \end{bmatrix} = \begin{bmatrix} -2-1 & 3-2 \\ 4-1 & 2-2 \\ 1-1 & 1-2 \end{bmatrix} = \begin{bmatrix} -3 & 1 \\ 3 & 0 \\ 0 & -1 \end{bmatrix}.$$

Nakon toga računamo matricu uzajamnih proizvoda centriranih podataka $X^{*\prime}X^*$

$$X^{*\prime}X^* = \begin{bmatrix} -3 & 3 & 0 \\ 1 & 0 & -1 \end{bmatrix}\begin{bmatrix} -3 & 1 \\ 3 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 18 & -3 \\ -3 & 2 \end{bmatrix}$$

pa je uzoračka kovarijaciona matrica

$$S = \frac{1}{N-1}X^{*\prime}X^* = \frac{1}{3-1}\begin{bmatrix} 18 & -3 \\ -3 & 2 \end{bmatrix} = \begin{bmatrix} 9 & -\frac{3}{2} \\ -\frac{3}{2} & 1 \end{bmatrix}.$$

Odavde je $S_1^2=9$, $S_2^2=1$ i $S_{12}=S_{21}=-\frac{3}{2}$ (negativna linearna veza).

Za domaću vežbu naći kovarijacionu matricu uzorka za podatke date u primeru 3.1.

Primer 3.3:

Za podatke iz prethodnog primera odrediti realizovanu vrednost uzoračke korelacione matrice koristeći relaciju uspostavljenu između nje i kovarijacione matrice.

Na osnovu izračunate kovarijacione matrice

$$S = \begin{bmatrix} 9 & -\frac{3}{2} \\ -\frac{3}{2} & 1 \end{bmatrix}$$

formiramo dijagonalnu matricu

$$D = \begin{bmatrix} 9 & 0 \\ 0 & 1 \end{bmatrix}$$

a zatim izračunavamo korelacionu matricu

$$R = \left(D^{\frac{1}{2}} \right)^{-1} S \left(D^{\frac{1}{2}} \right)^{-1}$$

$$R = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 9 & -\frac{3}{2} \\ -\frac{3}{2} & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}.$$

Često se tokom istraživanja javlja potreba da p promenljivih posmatramo razdvojeno, odnosno da ih tretiramo kao da pripadaju dvema grupama promenljivih

$$X = \left(X_1 \quad X_2 \;\vdots\; X_q \;|\; X_{q+1} \quad X_{q+2} \;\vdots\; X_p \right)^T$$

tako da dobijamo

$$X_1 = \left(X_1 \quad X_2 \;\vdots\; X_q \right)^T \qquad X_1 = \left[x_{ij} \right]_{N \times q},$$
$$X_2 = \left(X_{q+1} \quad X_{q+2} \;\vdots\; X_p \right)^T \qquad X_2 = \left[x_{ij} \right]_{N \times s}.$$

Tako npr. ako ispitujemo uticaj demografskih faktora na potrošnju proizvoda široke potrošnje tada ćemo vektor $\boldsymbol{X}$ dimenzije $p \times 1$ podeliti na vektor $\boldsymbol{X_1}$ u koji će ući sledeći podskup promenljivih: pol, starost i stručna sprema, i na vektor $\boldsymbol{X_2}$ u koji će ući promenljive potrošnje hleba, mesa, mleka, povrća i voća.

X_1 – pol
X_2 – starost
X_3 – stručna sprema → podela vektora
X_4 – potrošnja hleba
X_5 – potrošnja mesa
X_6 – potrošnja mleka
X_7 – potrošnja povrća
X_8 – potrošnja voća

$$X = \left(X_1 \quad X_2 \quad X_3 \;|\; X_4 \quad X_5 \quad X_6 \quad X_7 \quad X_8 \right)^T$$
$$X_1 = \left(X_1 \quad X_2 \quad X_3 \right)^T, \quad q = 3$$

$$X_2 = \begin{pmatrix} X_4 & X_5 & X_6 & X_7 & X_8 \end{pmatrix}^T, \quad s = 5$$

U skladu sa podelom slučajnog vektora $\boldsymbol{X}$ vršimo podelu sredine μ i kovarijacione matrice $\boldsymbol{\Sigma}$:

$$\mu = (\mu_1 \quad \mu_2)^T$$
$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

gde su:

$\mu_{1(q\times 1)}$ *i* $\mu_{2(q\times 2)}$	sredine slučajnih vektora $\boldsymbol{X_1}$ i $\boldsymbol{X_2}$
$\Sigma_{11(q\times q)}$ *i* $\Sigma_{22(s\times s)}$	kovarijacione matrice slučajnih vektora $\boldsymbol{X_1}$ i $\boldsymbol{X_2}$

dok su elementi matrice $\Sigma_{12(q\times s)}$ kovarijanse između elemenata slučajnih vektora $\boldsymbol{X_1}$ i $\boldsymbol{X_2}$ sa osobinom $\Sigma_{21} = \Sigma_{12}{}^T$.

3.4 Nekoliko važnijih osobina uzoračke sredine i kovarijacione matrice

Neka su slučajne promenljive X_j i X_k linearno transformisane tj. neka su definisane nove slučajne promenljive

$cX_j + a$ *i* $dX_k + b$, *a, b, c* i *d* su realne konstante.

Tada, na osnovu definicije kovarijanse, sledi

$$Cov(cX_j + a,\, dX_k + b) = cd\, Cov(X_j,\, X_k).$$

Uopštimo ovaj slučaj na linearnu kombinaciju p slučajnih promenljivih iz slučajnog vektora $\boldsymbol{X}$ sa sredinom μ i kovarijacionom matricom $\boldsymbol{\Sigma}$. Linearnom kombinacijom

$$Y = a_1 X_1 + a_2 X_2 + \ldots + a_p X_p = a^T X$$

za dati vektor linearne kombinacije

$$a^T = \begin{pmatrix} a_1 & a_2 & \vdots & a_p \end{pmatrix}$$

definišemo novu slučajnu promenljivu $\boldsymbol{Y}$ čija je očekivana vrednost

$$\mu_Y = E(Y) = E(a^T X) = a^T \mu$$

i varijansa

$$\sigma_Y^2 = Var(Y) = Var(a^T X) = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j \sigma_{ij} = a^T \Sigma a .$$

Razmotrimo sada još opštiji slučaj *q* linearnih kombinacija *p* slučajnih promenljivih

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ Y_q &= a_{q1}X_1 + a_{q2}X_2 + \ldots + a_{qp}X_p \end{aligned}$$

ili ako ove linearne kombinacije izrazimo u matričnoj formi

$$Y = AX$$

gde je $\boldsymbol{Y_{(qx1)}}$ vektor, a $A = \left[a_{ij}\right]_{q\times p}$ matrica koeficijenata linearnih kombinacija, tada je

$$\begin{aligned} \mu_Y &= E(Y) = E(AX) = A\mu_X \\ \Sigma_Y &= Cov(Y) = Cov(AX) = A\Sigma_X A^T \end{aligned}$$

gde su μ_X i $\boldsymbol{\Sigma_X}$ sredina i kovarijaciona matrica slučajnog vektora $\boldsymbol{X}$, respektivno.

Na kraju, neka su date matrice

$$A = \left[a_{ij}\right]_{q\times p} \quad i \quad B = \left[b_{ij}\right]_{s\times r}$$

čiji su elementi realni brojevi. Sada možemo da formiramo *q* linearnih kombinacija *p* slučajnih promenljivih

$$Y = AX$$

i *s* linearnih kombinacija *r* slučajnih promenljivih $Z^T = \left(Z_1 \quad Z_2 \quad \vdots \quad Z_r\right)$

$$W = BZ .$$

Ako sa $\boldsymbol{\Sigma_{XZ}}$ dimenzije $p\times r$ označimo matricu kovarijansi između X_j, $j=1, 2, \ldots, p$ i Z_k, $k=1, 2, \ldots, r$, tada je

$$\Sigma_{YW} = Cov(Y,W) = Cov(AX, BZ) = A\Sigma_{XZ}B^T .$$

Primer 3.4:

Izračunati sredine, varijanse i kovarijanse sledeće dve linearne kombinacije na osnovu podataka iz primera 3.2:

a) direktno, na osnovu opservacija za Y_1 i Y_2
b) indirektno na osnovu izračunate uzoračke sredine $\overline{X}$ i uzoračke kovarijacione matrice ***S.***

$$A=\begin{bmatrix}-1 & 2\\ 3 & -2\end{bmatrix} \qquad X=\begin{bmatrix}-2 & 3\\ 4 & 2\\ 1 & 1\end{bmatrix}.$$

Treba da dobijemo

$$Y=\begin{bmatrix}y_{11} & y_{12}\\ y_{21} & y_{22}\\ y_{31} & y_{32}\end{bmatrix}.$$

Imamo dve linearne kombinacije:

$$Y_1 = a_{11}X_1 + a_{12}X_2 = a_1^T X = \begin{bmatrix}-1 & 2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} = -x_1 + 2x_2$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 = a_2^T X = \begin{bmatrix}3 & -2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} = 3x_1 - 2x_2$$

a) Za prvu linearnu kombinaciju

$$Y_1 = a_1^T X = \begin{bmatrix}-1 & 2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}$$

imamo realizovane vrednosti opservacija y_{11}, y_{21} i y_{31} tj.

$$y_{11} = (-1)(-2) + 2\cdot 3 = 8$$
$$y_{21} = (-1)\cdot 4 + 2\cdot 2 = 0$$
$$y_{31} = (-1)\cdot 1 + 2\cdot 1 = 1$$

Napomena:

$$X_1 = \begin{pmatrix}x_1 & x_2 & x_3\end{pmatrix}^T$$
$$\overline{y}_1 = \frac{1}{3}(8+0+1) = 3$$

$$S_{Y_1}^2 = \frac{1}{3-1}\left[(8-3)^2 + (0-3)^2 + (1-3)^2\right] = 19\,.$$

Za drugu linearnu kombinaciju je

$$Y_2 = a_2^T X = \begin{bmatrix} 3 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$y_{12} = 3\cdot(-2) + (-2)\cdot 3 = -12$$
$$y_{22} = 3\cdot 4 + (-2)\cdot 2 = 8$$
$$y_{32} = 3\cdot 1 + (-2)\cdot 1 = 1$$

pa je

$$\bar{y}_2 = \frac{1}{3}(-12+8+1) = -1$$

$$S_{Y_2}^2 = \frac{1}{3-1}\left[(-12+1)^2 + (8+1)^2 + (1+1)^2\right] = 103$$

$$Cov(Y_1, Y_2) = S_{12} \frac{\sum_{k=1}^{3}(y_{k1} - \bar{y}_1)(y_{k2} - \bar{y}_2)}{3-1} =$$

$$= \frac{(8-3)(-12+1) + (0-3)(8+1) + (1-3)(1+1)}{2} = -43$$

$$S_Y = \begin{bmatrix} 19 & -43 \\ -43 & 103 \end{bmatrix}.$$

b) Indirektno:

$$Y = AX \Leftrightarrow \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\bar{y} = A\bar{x} \Leftrightarrow \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 3 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}.$$

Za uzoračku kovarijacionu matricu imamo iste relacije kao i za populaciju, tj.

$$\Sigma_Y = A\Sigma_X A'$$
$$S_Y = AS_X A'$$

$$S_Y = \begin{bmatrix} -1 & 2 \\ 3 & -2 \end{bmatrix} \begin{bmatrix} 9 & -\frac{3}{2} \\ -\frac{3}{2} & 1 \end{bmatrix} \begin{bmatrix} -1 & 2 \\ 3 & -2 \end{bmatrix} = \begin{bmatrix} 19 & -43 \\ -43 & 103 \end{bmatrix}.$$

Primer 3.5:

Zavod za statistiku u svojoj redovnoj anketi o potrošnji domaćinstava prikuplja podatke o njihovim prihodima i rashodima po različitim kategorijama za poljoprivredna, mešovita i nepoljoprivredna domaćinstva. Izabran je slučajan uzorak od šest domaćinstava i izmereni su sledeći podaci:

X_1	X_2	X_3	X_4
6	9.4	6.2	0.20
7	12.1	7.2	0.10
4	10.7	3.1	0.22
5	9.3	5.4	0.12
2	11.2	2.1	0.26
3	13.3	1.8	0.30

X_1 – broj članova domaćinstva,
X_2 – godišnja raspoloživa sredstva,
X_3 – izdaci za ishranu,
X_4 – izdaci za obrazovanje i razonodu.

Izračunati kovarijacionu i korelacionu matricu uzorka.

Rešenje:

$$S = \begin{bmatrix} 3.50 & -0.86 & 4.10 & -0.12 \\ -0.86 & 2.416 & -1.51 & 0.0524 \\ 4.10 & -1.51 & 5.152 & -0.1568 \\ -0.12 & 0.0524 & -0.1568 & 0.00608 \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & -0.2957 & 0.9655 & -0.8226 \\ & 1 & -0.4280 & 0.4323 \\ & & 1 & -0.8859 \\ & & & 1 \end{bmatrix}.$$

4. METODA GLAVNIH KOMPONENTI

Ideja metode glavnih komponenti je da se originalne promenljive transformišu u nove promenljive, odnosno, linearne kombinacije.

Polazimo od vektora obeležja

$$X = \left(X_1 \quad X_2 \quad \vdots \quad X_p \right)^T, \quad p \geq 1.$$

Ako je, na primer, *p=20* (20 obeležja) njihova međusobna povezanost iskazana je pomoću 190 koeficijenata korelacije što predstavlja problem kako za razumevanje, tako i za interpretaciju. Zato ćemo ovih *p* promenljivih zameniti manjim brojem linearnih kombinacija tipa

$$L_i' X, \quad i = 1,2,..$$

ali tako da su te linearne kombinacije nezavisne među sobom i da sadrže što je moguće više ***inicijalnog varijabiliteta.***

Prva glavna komponenta je konstruisana tako da obuhvata najveći deo varijanse originalnog skupa podataka. Druga glavna komponenta obuhvata najveći deo preostale varijanse originalnog skupa podataka i tako redom.

Analiza glavnih komponenti je bazirana na pretpostavci da će relativno mali broj glavnih komponenti dobro aproksimirati kovarijacionu strukturu skupa originalnih promenljivih. Ovim metoda glavnih komponenti postiže dva cilja:

1. vrši redukciju originalnog skupa podataka
2. olakšava interpretaciju kovarijacione strukture originalnih promenljivih na bazi manjeg broja međusobno nekorelisanih glavnih komponenti

Tako ćemo umesto *p=20* promenljivih u daljoj analizi koristiti 4 linearne kombinacije.

4.1 Definicija glavnih komponenti

Neka je

$$X = \begin{pmatrix} X_1 & X_2 & \vdots & X_p \end{pmatrix}^T, \quad p \geq 1$$

slučajni vektor sa kovarijacionom matricom $\boldsymbol{\Sigma}$. Prvu glavnu komponentu određujemo kao

$$Y_1 = l_{11}X_1 + l_{12}X_2 + \ldots + l_{1p}X_p = L_1'X$$

linearnu kombinaciju elemenata slučajnog vektora $\boldsymbol{X}$, gde su $l_{11}, l_{12}, \ldots, l_{1p}$ koeficijenti linearne kombinacije.

Od ranije znamo da je

$$Var(Y_1) = Var(L_1'X) = L_1'\Sigma L_1.$$

Naš zadatak je da odredimo vektor koeficijenata $\boldsymbol{L_1}$ tako da se maksimizira varijansa od Y_1. Ovde se uvodi uslov ortogonalnosti same transformacije jer se to pokazalo jako korisno u praksi i omogućuje se optimizacija, tj. uvodimo uslov

$$L_1'L_1 = 1$$

odnosno hoćemo da maksimiziramo $L_1'\Sigma L_1$ uz ograničenje $L_1'L_1 = 1$.

To se radi uz pomoć ***Lagranžovih množitelja.*** Diferenciranjem Lagranžove funkcije po koeficijentima $\boldsymbol{L_1}$ a zatim izjednačavanjem dobijenog izraza sa nulom dobijamo

$$\Sigma L_1 - \lambda L_1 = 0$$

ili

$$(\Sigma - \lambda I)L_1 = 0$$

gde je $\boldsymbol{I_{(p\times p)}}$ jedinična matrica, λ jedan od karakterističnih korena kovarijacione matrice $\boldsymbol{\Sigma}$ i to najveći $\lambda=\lambda_1$, a $\boldsymbol{L_1}$ je karakteristični vektor.

Da bi se dobilo netrivijalno rešenje za $\boldsymbol{L_1}$ treba da je ispunjeno

$$|\Sigma - \lambda I| = 0.$$

Svaku sledeću linearnu kombinaciju određujemo slično uz dodatni uslov da se radi o nekorelisanim veličinama (da je kovarijansa između prve i druge glavne komponente jednaka nuli)

$$\Sigma L_2 - \lambda L_2 = 0$$

ili

$$(\Sigma - \lambda I) L_2 = 0$$

gde za λ biramo što je moguće veću vrednost $\lambda=\lambda_2$. $\boldsymbol{L_2}$ je karakteristični vektor a linearna kombinacija

$$Y_2 = L_2' X$$

predstavlja drugu glavnu komponentu.

Rezime:

1. Ako su svi karakteristični koreni matrice $\boldsymbol{\Sigma}$ međusobno različiti (uredili smo ih u opadajući niz $\lambda_1>\lambda_2>\ldots>\lambda_p\geq 0$) tada postoji p glavnih komponenti

   $$Y_1, Y_2, \ldots, Y_p$$

   tj.

   $$Y_j = L_j' X, \quad j = 1,2,\ldots,p$$

2. Vektori koeficijenata $\boldsymbol{L_1}$, $\boldsymbol{L_2}$, ..., $\boldsymbol{L_p}$ predstavljaju karakteristične vektore matrice $\boldsymbol{\Sigma}$ koji su pridruženi karakterističnim korenima λ_j.

3.

$$E(Y_j) = 0$$
$$Var(Y_j) = L_j' \Sigma\ L_j = \lambda_j$$
$$Cov(Y_i, Y_j) = 0, \quad i \neq j$$
$$\Lambda = \left[\begin{array}{cc:c} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ \hdashline 0 & 0 & \lambda_p \end{array}\right]$$

Primer 4.1:

Za datu kovarijacionu matricu

$$\Sigma = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & -2 \\ 0 & -2 & 4 \end{bmatrix}$$

odrediti sve glavne komponente.

Glavne komponente su:

$$Y_1 = L_1' X$$
$$Y_2 = L_2' X$$
$$Y_3 = L_3' X$$

Prvu glavnu komponentu računamo iz uslova

$$\left(\Sigma - \lambda_1 I\right) L_1 = 0$$

odnosno češće se umesto L_1 piše P_1 ***(principal components)***

$$\left(\Sigma - \lambda_1 I\right) P_1 = 0$$

odnosno

$$\Sigma P_1 = \lambda_1 P_1$$

pa sledi da je potrebno izračunati karakteristične korene kovarijacione matrice Σ

$$\Sigma = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & -2 \\ 0 & -2 & 4 \end{bmatrix}$$

$$\left|\Sigma - \lambda_1 I\right| = \begin{vmatrix} 4-\lambda & 0 & 0 \\ 0 & 2-\lambda & -2 \\ 0 & -2 & 4-\lambda \end{vmatrix} = (4-\lambda)(\lambda^2 - 6\lambda + 4) = 0$$

$$\lambda_1 = 4,\ \lambda_2 = 0.769,\ \lambda_3 = 5.2361.$$

Šta ne valja?

Po definiciji glavnih komponenti, λ_1 je najveći karakteristični koren, λ_2 drugi po veličini, itd. pa je

$$\lambda_1 = 5.2361,\ \lambda_2 = 4,\ \lambda_3 = 0.769$$

pa će biti

$$\Sigma P_1 = \lambda_1 P_1$$

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & -2 \\ 0 & -2 & 4 \end{bmatrix} \begin{bmatrix} p_{11} \\ p_{21} \\ p_{31} \end{bmatrix} = 5.2361 \begin{bmatrix} p_{11} \\ p_{21} \\ p_{31} \end{bmatrix}$$

$$4p_{11} + 0p_{21} + 0p_{31} = 5.2361p_{11}$$
$$0p_{11} + 2p_{21} - 2p_{31} = 5.2361p_{21}$$
$$0p_{11} - 2p_{21} + 4p_{31} = 5.2361p_{31}$$

$$-1.2361p_{11} = 0 \Rightarrow p_{11} = 0$$
$$-3.2361p_{21} - 2p_{31} = 0$$
$$-2p_{21} - 1.2361p_{31} = 0$$

Rešavanjem ovog sistema dobija se

$$P_1' = \begin{pmatrix} 0 & -0.53 & 0.85 \end{pmatrix}.$$

Sada računamo drugu glavnu komponentu:

$$(\Sigma - \lambda_2 I) P_2 = 0$$

$$\Sigma P_2 = \lambda_2 P_2$$

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & -2 \\ 0 & -2 & 4 \end{bmatrix} \begin{bmatrix} p_{12} \\ p_{22} \\ p_{32} \end{bmatrix} = 4 \begin{bmatrix} p_{12} \\ p_{22} \\ p_{32} \end{bmatrix}$$

$$4p_{12} + 0p_{22} + 0p_{32} = 5.2361p_{12}$$
$$0p_{12} + 2p_{22} - 2p_{32} = 5.2361p_{22}$$
$$0p_{12} - 2p_{22} + 4p_{32} = 5.2361p_{32}$$

$$P_2' = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}.$$

Treća komponenta:

$$(\Sigma - \lambda_3 I) P_3 = 0$$

$$\Sigma P_3 = \lambda_3 P_3$$

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & -2 \\ 0 & -2 & 4 \end{bmatrix} \begin{bmatrix} p_{13} \\ p_{23} \\ p_{33} \end{bmatrix} = 0.769 \begin{bmatrix} p_{13} \\ p_{23} \\ p_{33} \end{bmatrix}$$

$$4p_{13} + 0p_{23} + 0p_{33} = 5.2361p_{13}$$
$$0p_{13} + 2p_{23} - 2p_{33} = 5.2361p_{23}$$
$$0p_{13} - 2p_{23} + 4p_{33} = 5.2361p_{33}$$

$$P_3' = \begin{pmatrix} 0 & -0.8507 & -0.5257 \end{pmatrix}.$$

Na kraju, glavne komponente su:

$$Y_1 = L_1' X = P_1' X = -0.53X_2 + 0.85X_3$$
$$Y_2 = L_2' X = P_2' X = X_1$$
$$Y_3 = L_3' X = P_3' X = -0.8507X_2 - 0.5257X_3$$

Zadovoljeni su uslovi normiranosti ***L'L=I***:

za $P_1' = \begin{pmatrix} 0 & -0.53 & 0.85 \end{pmatrix}$ imamo $0^2+(-0.5257)^2+0.8507^2=1$

za $P_2' = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$ imamo $1^2=1$

za $P_3' = \begin{pmatrix} 0 & -0.8507 & -0.5257 \end{pmatrix}$ važi isto kao i za $\boldsymbol{P_1'}$.

Primer 4.2:

Na prethodnom primeru pokazati da je varijansa dobijenih glavnih komponenti jednaka odgovarajućim karakterističnim korenima a da je kovarijansa između svakog para glavnih komponenti jednaka nuli.

Treba dokazati da je $Var(Y_j) = \lambda_j$ za *j=1, 2, ..., p,* tj.

$$Var(Y_1) = Var(-0.5257X_2 + 0.8507X_3) = \lambda_1 = 5.2361$$

itd.

Provera korelisanosti prve i druge glavne komponente:

$$Cov(Y_1, Y_2) = Cov(-0.5257X_2 + 0.8507X_3 \; ; \; X_1) = 0.$$

Napomena:

Generalizovana varijansa se određuje po dve definicije.

Po prvoj definiciji generalizovana varijansa je jednaka determinanti kovarijacione matrice. ***Kovarijaciona matrica*** glavnih komponenti je dijagonalna matrica čiji su elementi karakteristični koreni matrice $\boldsymbol{\Sigma}$ i označava se sa $\boldsymbol{\Lambda}$

$$\Lambda = \left[\begin{array}{cc:c} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ \hdashline 0 & 0 & \lambda_p \end{array}\right].$$

Prema tome generalizovana varijansa vektora $\boldsymbol{Y}$ je jednaka determinanti

$$\left|\Lambda\right|.$$

Po drugoj definiciji generalizovana varijansa je jednaka tragu kovarijacione matrice. Trag kovarijacione matrice je jednak zbiru karakterističnih korena λ_j.

Korišćenjem ortogonalne matrice $\boldsymbol{P_{(pxp)}}$ čiji su redovi karakteristični vektori kovarijacione matrice $\boldsymbol{\Sigma}$, možemo izvršiti ortogonalnu dekompoziciju matrice $\boldsymbol{\Sigma}$ čiji su koreni različiti.

Imamo da je

$$\Sigma = P'\,\Lambda\,P$$

ili u razvijenom obliku

$$\Sigma = \left[L_1 \quad L_2 \;\vdots\; L_p\right]\left[\begin{array}{cc:c} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ \hdashline 0 & 0 & \lambda_p \end{array}\right]\left[\begin{array}{c} L_1' \\ L_2' \\ \hdashline L_p' \end{array}\right] = \sum_{j=1}^{p} \lambda_j L_j L_j'.$$

Takođe je

$$\Lambda = P\Sigma P'.$$

Uz korišćenje osobine traga matrice *tr(BC)=tr(CB)* imamo da je

$$tr(\Lambda) = tr(P\Sigma P') = tr(P'\,P\Sigma) = tr(\Sigma)$$

pošto je $\boldsymbol{P'P=I}$.

Primer 4.3:

Neka su za vektor $\boldsymbol{X}^T=(X_1, X_2, X_3)$ poznati korelacioni koeficijenti $r_{12}=0.5$, $r_{13}=0.3$ i $r_{23}=0.2$ i neka su date prve dve glavne komponente kao

$$Y_1 = 0.84X_1 + 0.79X_2 + 0.6X_3 \quad i \quad Y_2 = 0.17X_1 + 0.42X_2 - 0.79X_3.$$

Izračunati:

a) Koliko treba uzeti glavnih komponenti da bi se obuhvatilo barem 85% varijabiliteta i koliko procenata ukupnog varijabiliteta nosi svaka od glavnih komponenti Y_1, Y_2 i Y_3? Izračunati generalizovanu varijansu vektora $\boldsymbol{X}$ korišćenjem varijanse glavnih komponenata i pokazati zašto je to moguće.

b) Aproksimirati korelacionu matricu na osnovu projekcije trodimenzionalnog vektora $\boldsymbol{X}$ u prostor prve glavne komponente.

c) Korelaciju između prve glavne komponente Y_1 i varijable X_1 kao i između druge glavne komponente Y_2 i varijable X_2.

$$X^T = \begin{pmatrix} X_1 & X_2 & X_3 \end{pmatrix}$$
$$Y_1 = 0.84X_1 + 0.79X_2 + 0.6X_3$$
$$Y_2 = 0.17X_1 + 0.42X_2 - 0.79X_3$$
$$R = \begin{bmatrix} 1 & 0.5 & 0.3 \\ & 1 & 0.2 \\ & & 1 \end{bmatrix}$$

Da bi važio uslov $\alpha^T\alpha=1$, inicijalne glavne komponente treba podeliti pa se dobija:

$$P_1^T \begin{pmatrix} 0.646 & 0.608 & 0.461 \end{pmatrix}$$
$$P_2^T \begin{pmatrix} 0.187 & 0.462 & -0.868 \end{pmatrix}$$

a)

$$\begin{bmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 1 & 0.2 \\ 0.3 & 0.2 & 1 \end{bmatrix} \begin{bmatrix} 0.646 \\ 0.608 \\ 0.461 \end{bmatrix} = \lambda_1 \begin{bmatrix} 0.646 \\ 0.608 \\ 0.461 \end{bmatrix} \Rightarrow \lambda_1 = 1.685$$

$$\begin{bmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 1 & 0.2 \\ 0.3 & 0.2 & 1 \end{bmatrix} \begin{bmatrix} 0.187 \\ 0.462 \\ -0.868 \end{bmatrix} = \lambda_2 \begin{bmatrix} 0.187 \\ 0.462 \\ -0.868 \end{bmatrix} \Rightarrow \lambda_2 = 0.830$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 3 \Rightarrow \lambda_3 = 0.485$$

Prva glavna komponenta nosi 56.17% varijabiliteta, druga 27.67% a treća 16.17% varijabiliteta. Da bi se obuhvatilo 85% varijabiliteta potrebne su sve tri glavne komponente.

b)

$$\lambda_1 \alpha_1 \alpha_1^T = 1.685 \begin{bmatrix} 0.646 \\ 0.608 \\ 0.461 \end{bmatrix} \begin{bmatrix} 0.646 & 0.608 & 0.461 \end{bmatrix} = \begin{bmatrix} 0.703 & 0.662 & 0.502 \\ 0.662 & 0.623 & 0.472 \\ 0.502 & 0.472 & 0.358 \end{bmatrix}$$

c)

$$\rho_{Y_1 X_1} = \alpha_{11} \sqrt{\lambda_1} = 0.646 \sqrt{1.685} = 0.839$$
$$\rho_{Y_2 X_2} = \alpha_{22} \sqrt{\lambda_2} = 0.462 \sqrt{0.83} = 0.421$$

5. FAKTORSKA ANALIZA

Faktorska analiza je metod multivarijacione analize koji se koristi za opis međusobne zavisnosti velikog broja promenljivih korišćenjem manjeg broja osnovnih ali ***neopažljivih*** slučajnih promenljivih (faktora). Slično kao metoda glavnih komponenti i faktorska analiza se koristi za redukciju skupa podataka aproksimirajući kovarijacionu ili korelacionu matricu originalnih promenljivih.

Između metode glavnih komponenti i faktorske analize postoje sledeće razlike:

1. Analiza glavnih komponenti se bazira na varijansama – interesuju nas dijagonalni elementi kovarijacione matrice, dok nas kod faktorske analize interesuju vandijagonalni elementi.
2. Za razliku od analize glavnih komponenti, faktorska analiza pretpostavlja postojanje teorijskog modela kojim se uspostavlja relacija između opservacija p-dimenzionalne promenljive i manjeg broja zajedničkih faktora.
3. Često se analiza glavnih komponenti tretira kao specifičan slučaj ili prva faza u faktorskoj analizi ***(SPSS)***.
4. Analiza glavnih komponenti izučava ukupan varijabilitet skupa podataka. Nasuprot njoj, faktorska analiza polazi od razlaganja varijabiliteta na dva dela: zajednički deo i specifični deo. Zajednički deo je onaj deo varijacija promenljive koji ona deli sa ostalim promenljivama. Sa druge strane, specifični deo je onaj deo varijacija promenljive koji je poseban (specifičan) za tu promenljivu. Faktorska analiza izučava zajednički deo varijabiliteta za sve promenljive
5. U obe metode se javljaju dve vrste promenljivih: ***opažljive*** promenljive koje čini originalni skup podataka i ***neopažljive*** (latentne) promenljive. Međutim, u analizi glavnih komponenti na osnovu linearne kombinacije opažljivih promenljivih se formiraju glavne komponente kao neopažljive komponente, dok se u faktorskoj analizi na osnovu faktora kao neopažljivih komponenti izražavaju originalne promenljive

5.1 Istorijat

Krajem XIX veka *Galton* i *Spirman* postavljaju problem inteligencije:

Da li je inteligencija bazirana na jednom bazičnom tj. opštem faktoru ili se bazira na nekoliko zajedničkih faktora kao što su npr. verbalna sposobnost, matematička sposobnost, memorija?

Primer 5.1:

Čarls Spirman je 1904. godine izučavao pripremljenost dece za školu i dobio je korelacionu matricu rezultata testova iz klasike *(X_1)*, francuskog *(X_2),* engleskog *(X_3)* i matematike *(X_4)* u obliku

$$\begin{bmatrix} 1 & 0.83 & 0.78 & 0.70 \\ 0.83 & 1 & 0.67 & 0.67 \\ 0.78 & 0.67 & 1 & 0.64 \\ 0.70 & 0.67 & 0.64 & 1 \end{bmatrix}$$

Spirman je uočio proporcionalnost ma koja dva reda ili kolone u ovoj matrici (ako se zanemare elementi na glavnoj dijagonali). Tako za elemente drugog i trećeg reda imamo

$$\frac{0.83}{0.67} \approx \frac{0.78}{0.64} \approx 1.2 .$$

Zato *Spirman* predlaže redukciju problema sa *p=4* na *p=1* tako što rezultate svih testova *(X_i, i=1, 2, 3* i *4)* iskazuje preko modela

$$X_i = \beta_i F + \varepsilon_i, \quad i = 1,2,3,4$$

gde su

F – zajednički faktor,
β_i – koeficijenti (faktorska opterećenja),
ε_i – slučajne greške (specifični faktori).

Dakle, *Spirman* je formulisao dvofaktorsku teoriju testova inteligencije prema kojoj se rezultat svakog testa može dekomponovati na dva dela: prvi, koji je zajednički za sve testove *(F)* i koji se može interpretirati kao "opšta sposobnost" ili "inteligencija", i drugi, koji je specifičan za svaki test *(ε_i).*

Napomena:

Kasnijim istraživanjima prvobitni model je proširen tako što je uvedeno nekoliko zajedničkih faktora (pretpostavlja se da postoje posebne vrste inteligencija – sposobnosti), a i specifičan faktor je razložen na dva dela od kojih prvi pokazuje u kom stepenu se nečija individualna sposobnost u npr. matematici razlikuje od opšte sposobnosti i drugi koji

predstavlja činjenicu da su rezultati testiranja samo aproksimacija sposobnosti individue u konkretnoj oblasti (ipak samo test - papir).

5.2 Model faktorske analize

Neka je $\boldsymbol{X}$ p-dimenzioni vektor opažljivih promenljivih sa sredinom $\boldsymbol{\mu}$ i kovarijacionom matricom $\boldsymbol{\Sigma}$. Pretpostavimo da se vektor opažljivih promenljivih $\boldsymbol{X}$ može izraziti preko skupa od m neopažljivih promenljivih koje nazivamo zajednički faktori u oznaci $F_1, F_2, \ldots, F_m$ $(m<<p)$ i preko skupa $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ specifičnih ali neopažljivih faktora.

Tada model u razvijenom obliku izgleda

$$\begin{aligned}(X_1-\mu_1) &= \beta_{11}F_1+\beta_{12}F_2+\ldots+\beta_{1m}F_m+\varepsilon_1\\ (X_2-\mu_2) &= \beta_{21}F_1+\beta_{22}F_2+\ldots+\beta_{2m}F_m+\varepsilon_2\\ &\vdots\\ (X_p-\mu_p) &= \beta_{p1}F_1+\beta_{p2}F_2+\ldots+\beta_{pm}F_m+\varepsilon_p\end{aligned}$$

ili ekvivalentno u matričnoj notaciji

$$\underset{(p\times 1)}{X-\mu} = \underset{(p\times m)}{B}\ \underset{(m\times 1)}{F} + \underset{(p\times 1)}{\varepsilon}$$

gde je

$$X-\mu=\begin{bmatrix}X_1-\mu_1\\ X_2-\mu_2\\ \vdots\\ X_p-\mu_p\end{bmatrix} \qquad F=\begin{bmatrix}F_1\\ F_2\\ \vdots\\ F_m\end{bmatrix} \qquad \varepsilon=\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_p\end{bmatrix}$$

$$B=\left[\begin{array}{ccc}\beta_{11} & \beta_{12} & \beta_{1m}\\ \beta_{21} & \beta_{22} & \beta_{2m}\\ \hdashline \beta_{p1} & \beta_{p2} & \beta_{pm}\end{array}\right]$$

Poslednja matrica predstavlja matricu ***faktorskih opterećenja*** a β_{ij} faktorsko opterećenje i-te promenljive na j-ti faktor.

Napomena:

Na prvi pogled model faktorske analize liči na model višestruke regresije, međutim ovde p odstupanja $(X_1-\mu_1), \ldots, (X_p-\mu_p)$ izražavamo preko $m+p$ slučajnih promenljivih $F_1, F_2, \ldots, F_m$ i $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ koje su neopažljive za razliku od regresionog modela gde su nezavisne promenljive opažljive.

U model se uvode sledeća dodatna ograničenja:

$$E(F) = 0$$
$$Cov(F) = E(FF') = \Phi$$

i dalja izlaganja se baziraju na specifičnom slučaju faktorske analize kada je

$$\Phi = I$$

odnosno, radi se o ortogonalnom modelu kod koga su faktori F_1, F_2, ..., F_m međusobno nezavisni.

Što se tiče specifičnih faktora važi

$$E(\varepsilon) = 0$$
$$Cov(\varepsilon) = E(\varepsilon\varepsilon') = \Psi = \begin{bmatrix} \Psi_1 & 0 & 0 \\ 0 & \Psi_2 & 0 \\ 0 & 0 & \Psi_p \end{bmatrix}$$

Takođe pretpostavljamo da su zajednički faktori nezavisni od specifičnih tj. da je

$$Cov(\varepsilon, F) = E(\varepsilon F') = 0 .$$

Vezu između odstupanja opažljivih promenljivih od njihove sredine i neopažljivih faktora

$$X - \mu = B F + \varepsilon$$

odnosno

$$\underset{(p\times 1)}{X} = \underset{(p\times 1)}{\mu} + \underset{(p\times m)}{B}\ \underset{(m\times 1)}{F} + \underset{(p\times 1)}{\varepsilon}$$

zajedno sa navedenim pretpostavkama nazivamo ***ortogonalni model*** faktorske analize sa m zajedničkih faktora. Ovaj model omogućuje razlaganje kovarijacione matrice Σ

$$\Sigma = Cov(X) = BB' + \Psi .$$

Napomena:

U opštem slučaju, kad ne važi uslov ortogonalnosti tj. kada je $\Phi \neq I$, razlaganje kovarijacione matrice Σ je

$$\Sigma = B\Phi B' + \Psi .$$

Kako je

$$Cov(X,F) = E(BF+\varepsilon)F' = BE(FF') + E(\varepsilon F') = B$$

znači da su elementi matrice faktorskih opterećenja kovarijanse između originalnih promenljivih i faktora.

Korelacionu matricu promenljivih $\boldsymbol{X}$ i faktora $\boldsymbol{F}$ nazivamo matrica ***faktorske strukture***.

Napomena:

U slučaju ortogonalnog modela faktorske analize važi da je

$$Cov(X,F) = Cor(X,F).$$

U opštem slučaju se ove dve vrednosti međusobno razlikuju.

Na osnovu razlaganja kovarijacione matrice kod ortogonalnog modela faktorske analize

$$\Sigma = BB' + \Psi$$

imamo da je varijansa *i*-te promenljive

$$\sigma_{ii} = Var(X_i) = \beta_{i1}^2 + \beta_{i2}^2 + \ldots + \beta_{im}^2 + \Psi_i$$

tj. varijansa *i*-te originalne promenljive je podeljena na dva dela:

- prvi deo je

$$h_i^2 = \beta_{i1}^2 + \beta_{i2}^2 + \ldots + \beta_{im}^2$$

  i predstavlja varijansu objašnjenu zajedničkim faktorima i nazivamo ga ***zajednička varijansa*** ili ***komunalitet***
- drugi deo nazivamo specifična varijansa

Primer 5.2:

Data je kovarijaciona matrica

$$\Sigma = \begin{bmatrix} 15 & 8 & 3 & 16 \\ 8 & 10 & 1 & 10 \\ 3 & 1 & 16 & 2 \\ 16 & 10 & 2 & 24 \end{bmatrix}$$

i jednakost

$$\begin{bmatrix} 15 & 8 & 3 & 16 \\ 8 & 10 & 1 & 10 \\ 3 & 1 & 16 & 2 \\ 16 & 10 & 2 & 24 \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 2 & 1 \\ -1 & 3 \\ 4 & 2 \end{bmatrix} \begin{bmatrix} 3 & 2 & -1 & 4 \\ 2 & 1 & 3 & 2 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}.$$

Dokazati da važi

$$\Sigma = BB' + \Psi .$$

Iz jednakosti sledi postojanje dva zajednička faktora *(m=2)* pa je

$$B = \begin{bmatrix} 3 & 2 \\ 2 & 1 \\ -1 & 3 \\ 4 & 2 \end{bmatrix} \quad \Psi = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}.$$

Potrebno je da na osnovu ortogonalnog modela faktorske analize odredimo komunalitete sve četiri originalne promenljive i razložimo njihove varijanse.

Komunalitet prve promenljive:

$$h_1^2 = \beta_{11}^2 + \beta_{12}^2 = 3^2 + 2^2 = 13$$

jer je

$$B = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \\ \beta_{41} & \beta_{42} \end{bmatrix}$$

pa je razlaganje varijanse σ_{11} na komunalitet i specifičnu varijansu

$$\sigma_{11} = h_1^2 + \Psi_1$$
$$15 = 13 + 2$$

jer je

$$\Psi = \begin{bmatrix} \Psi_1 & 0 & 0 & 0 \\ 0 & \Psi_2 & 0 & 0 \\ 0 & 0 & \Psi_3 & 0 \\ 0 & 0 & 0 & \Psi_4 \end{bmatrix}.$$

Slično se dobija da je

$$\sigma_{22} = h_2^2 + \Psi_2 \Leftrightarrow 10 = 5 + 5$$
$$\sigma_{33} = h_3^2 + \Psi_3 \Leftrightarrow 16 = 10 + 6$$
$$\sigma_{44} = h_4^2 + \Psi_4 \Leftrightarrow 24 = 20 + 4$$

odnosno

$$h_2^2 = \beta_{21}^2 + \beta_{22}^2 = 2^2 + 1^2 = 5$$
$$h_3^2 = \beta_{31}^2 + \beta_{32}^2 = (-1)^2 + 3^2 = 10$$
$$h_4^2 = \beta_{41}^2 + \beta_{42}^2 = 4^2 + 2^2 = 20$$

Napomena:

Korišćenjem korelacione umesto ovarijacione matrice dobijamo da se jedinična varijansa standardizovane promenljive sastoji iz dva dela:

$$Var(X_i) = 1 = h_i^2 + \Psi_i .$$

Generalizovana varijansa (ukupna varijansa) originalnog skupa promenljivih je oblika

$$tr(\Sigma) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \sum_{j=1}^{m} \beta_{ij}^2 + \sum_{i=1}^{p} \Psi_i .$$

Ako sa h označimo ukupan komunalitet od $\boldsymbol{X}$

$$h = \sum_{i=1}^{p} h_i^2 = \sum_{i=1}^{p} \sum_{j=1}^{m} \beta_{ij}^2$$

tada se može pisati

$$tr(\Sigma) = h + tr(\Psi) .$$

Znači da je generalizovana varijansa od $\boldsymbol{X}$ jednaka zbiru dve komponente: ukupnog ***komunaliteta*** i ukupne ***varijanse*** specifičnih faktora.

Formiranjem količnika

$$\frac{\sum_{i=1}^{p} \beta_{ij}^2}{h}, \quad j = 1, 2, \ldots, m$$

dobija se proporcija ukupnog varijabiliteta koja se može pripisati j-tom zajedničkom faktoru.

Primer 5.3:

Na osnovu ortogonalnog modela faktorske analize iz prethodnog primera odrediti ukupnu varijansu, ukupan komunalitet, ukupnu specifičnu varijansu i doprinose svakog faktora ukupnom komunalitetu.

$$\Sigma = \begin{bmatrix} 15 & 8 & 3 & 16 \\ 8 & 10 & 1 & 10 \\ 3 & 1 & 16 & 2 \\ 16 & 10 & 2 & 24 \end{bmatrix} \quad B = \begin{bmatrix} 3 & 2 \\ 2 & 1 \\ -1 & 3 \\ 4 & 2 \end{bmatrix} \quad \Psi = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

Ukupna varijansa:

$$tr(\Sigma) = \sigma_{11} + \sigma_{22} + \sigma_{33} + \sigma_{44} = 15 + 10 + 16 + 24 = 65.$$

Ukupan komunalitet:

$$h = h_1^2 + h_2^2 + h_3^2 + h_4^2 = 13 + 5 + 10 + 20 = 48.$$

Ukupna specifična varijansa:

$$tr(\Psi) = \Psi_1 + \Psi_2 + \Psi_3 + \Psi_4 = 2 + 5 + 6 + 4 = 17.$$

Doprinos prvog zajedničkog faktora ukupnom komunalitetu, *j=1*:

$$\frac{\beta_{11}^2 + \beta_{21}^2 + \beta_{31}^2 + \beta_{41}^2}{h} = \frac{3^2 + 2^2 + (-1)^2 + 4^2}{48} = \frac{30}{48} = 62.5\%.$$

Doprinos drugog zajedničkog faktora ukupnom komunalitetu, *j=2*:

$$\frac{\beta_{12}^2 + \beta_{22}^2 + \beta_{32}^2 + \beta_{42}^2}{h} = \frac{2^2 + 1^2 + 3^2 + 2^2}{48} = \frac{18}{48} = 37.5\%.$$

Napomena:

Faktorska analiza se najčešće obavlja na ***standardizovanim*** promenljivama, odnosno vršimo razlaganje korelacione matrice.

5.3 Metode izdvajanja faktora (metode ocenjivanja)

Na osnovu uzorka uzetog iz *p*-dimenzione populacije ocenjuje se model faktorske analize, odnosno proveravamo da li važi razlaganje kovarijacione (ili korelacione) matrice koje je implicirano tim modelom.

Postoje dve osnovne metode ocenjivanja modela faktorske analize:

- metoda glavnih komponenti (glavnih faktora)
- metoda maksimalne verodostojnosti

Metoda glavnih komponenti se javlja u dve varijante:

1. direktno korišćenje metode glavnih komponenti na kovarijacionu ili korelacionu matricu radi istovremenog ocenjivanja komunaliteta i matrice faktorskih opterećenja (metoda glavnih komponenti)
2. korelaciona matrica se modifikuje u duhu modela faktorske analize a zatim se iterativno i odvojeno ocenjuju komunaliteti i faktorska opterećenja primenom metode glavnih komponenti na tu modifikovanu korelacionu matricu (metoda glavnih faktora)

5.3.1 Metoda glavnih komponenti

U uzoračkom slučaju postupak primene metode glavnih komponenti u ocenjivanju faktorskog modela zahteva određivanje karakterističnih korena $\tilde{\lambda}_J$ i karakterističnih vektora $\tilde{\alpha}_J$.

Na osnovu prvih *m* faktora formiramo matricu ocenjenih faktorskih opterećenja prema sledećem izrazu:

$$\tilde{B} = \left[\sqrt{\tilde{\lambda}_1}\tilde{\alpha}_1 \quad \sqrt{\tilde{\lambda}_2}\tilde{\alpha}_2 \;\vdots\; \sqrt{\tilde{\lambda}_p}\tilde{\alpha}_p \right].$$

Ocenjene specifične varijanse dobijamo kao dijagonalne elemente matrice

$$S - \tilde{B}\tilde{B}'$$

pa je

$$\tilde{\Psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{\beta}_{ij}^2, \quad i = 1,2,\ldots,p$$

a ocenjeni komunaliteti su

$$\tilde{h}_i^2 = \tilde{\beta}_{i1}^2 + \tilde{\beta}_{i2}^2 + \ldots + \tilde{\beta}_{im}^2, \quad i = 1,2,\ldots,p.$$

Broj zadržanih faktora možemo odrediti *a priori* ili posmatranjem ***matrice reziduala***

$$S-(\tilde{B}\tilde{B}'+\Psi)$$

koja je dobijena kao rezultat aproksimacije uzoračke kovarijacione matrice sa prvih *m* faktora. Broj faktora koje smo uključili u aproksimaciju povećavamo sve dok ne procenimo da su elementi reziduala dovoljno mali.

Primer 5.4:

Na osnovu uzorka iz *p*-dimenzione populacije *(p=5)* dobijena je uzoračka korelaciona matrica

$$R=\begin{bmatrix} 1 & 0 & 0.8 & 0.3 & 0 \\ 0 & 1 & 0.1 & 0.6 & 0.8 \\ 0.8 & 0.1 & 1 & 0.4 & 0.1 \\ 0.3 & 0.6 & 0.4 & 1 & 0.7 \\ 0 & 0.8 & 0.1 & 0.7 & 1 \end{bmatrix}.$$

Karakteristični vektori i koreni matrice ***R*** su dati u tabeli:

Karakteristični vektori $\tilde{\alpha}_J$				
prvi	**drugi**	**treći**	**četvrti**	**peti**
0.2775	0.6365	-0.1965	0.6831	-0.1123
0.4893	-0.3508	-0.6102	-0.1294	-0.4985
0.3426	0.5871	-0.1201	-0.6802	0.2468
0.5510	-0.0451	0.7568	-0.0213	-0.3481
0.5124	0.3538	-0.0444	0.2314	0.7462
Karakteristični koreni $\tilde{\lambda}_J$				
2.5835	1.7166	0.3338	0.1941	0.1720

Korišćenjem metode glavnih komponenti oceniti model faktorske analize.

Potrebno je prvo da odredimo broj glavnih komponenti koje ćemo da zadržimo u analizi. Kada se koristi korelaciona matrica uobičajen kriterijum je da se zadrže one glavne komponente kod kojih je varijansa (karakteristični koren) veća od jedinice. Ovaj kriterijum se naziva "***kriterijum jediničnog korena*** - *Kaiser*-ov kriterijum".

Prema ovom kriterijumu ćemo zadržati prve dve glavne komponente jer je

$$\tilde{\lambda}_1=2.5835 \;\; i \;\; \tilde{\lambda}_2=1.7166.$$

Oni vuku

$$\frac{\tilde{\lambda}_1+\tilde{\lambda}_2}{p}=\frac{2.5835+1.7166}{5}=0.86$$

tj. 86% uzoračke varijanse.

Sada je potrebno da uporedimo korelacionu matricu na osnovu modela faktorske analize sa dva zajednička faktora

$$\tilde{B}\tilde{B}'+\Psi$$

sa uzoračkom korelacionom matricom ***R,*** i ako je dobijena razlika zanemariva, konstatujemo da ova dva zajednička faktora uspešno reprodukuju korelacionu strukturu pet originalnih promenljivih.

Napomena:

Za ortogonalni model važi da je $Cov(X,F)=Cor(X,F)$.

Daklem, treba da odredimo

$$\tilde{B}\tilde{B}'+\tilde{\Psi}$$

$$\tilde{B}=\left[\sqrt{\tilde{\lambda}_1}\tilde{\alpha}_1 \quad \sqrt{\tilde{\lambda}_2}\tilde{\alpha}_2\right]$$

$$\tilde{B}=\begin{bmatrix}\sqrt{2.5835}\cdot 0.2775 & \sqrt{1.7166}\cdot 0.6365\\ \sqrt{2.5835}\cdot 0.4893 & \sqrt{1.7166}\cdot(-0.3508)\\ \sqrt{2.5835}\cdot 0.3426 & \sqrt{1.7166}\cdot 0.5871\\ \sqrt{2.5835}\cdot 0.5510 & \sqrt{1.7166}\cdot(-0.0451)\\ \sqrt{2.5835}\cdot 0.5124 & \sqrt{1.7166}\cdot(-0.3538)\end{bmatrix}=\begin{bmatrix}0.4460 & 0.8339\\ 0.7865 & -0.4596\\ 0.5507 & 0.7692\\ 0.8857 & -0.0591\\ 0.8236 & -0.4635\end{bmatrix}$$

Da bi odredili $\tilde{\Psi}$ potrebno je da odredimo komunalitet

B_1	B_2	$h_i^2=B_1^2+B_2^2$	$\tilde{\Psi}_i=1-h_i^2$
0.4460	0.8339	$0.4460^2+0.8339^2=0.8944$	0.1056
0.7865	-0.4596	0.8297	0.1703
0.5507	0.7692	0.8949	0.1051
0.8857	-0.0591	0.7879	0.2121
0.8236	-0.4635	0.8932	0.1068

pa je

$$\tilde{B}\tilde{B}'+\Psi=\begin{bmatrix}0.4460 & 0.8339\\ 0.7865 & -0.4596\\ 0.5507 & 0.7692\\ 0.8857 & -0.0591\\ 0.8236 & -0.4635\end{bmatrix}\begin{bmatrix}0.4460 & 0.7865 & 0.5507 & 0.8837 & 0.8236\\ 0.8339 & -0.4596 & 0.7692 & -0.0591 & -0.4635\end{bmatrix}+$$

$$+\begin{bmatrix}0.1056 & 0 & 0 & 0 & 0\\ 0 & 0.1703 & 0 & 0 & 0\\ 0 & 0 & 0.1051 & 0 & 0\\ 0 & 0 & 0 & 0.2121 & 0\\ 0 & 0 & 0 & 0 & 0.1068\end{bmatrix}=$$

$$=\begin{bmatrix}1 & -0.0325 & 0.8871 & 0.3451 & -0.0192\\ & 1 & 0.0796 & 0.7237 & 0.8608\\ & & 1 & 0.4423 & 0.0971\\ & & & 1 & 0.7569\\ & & & & 1\end{bmatrix}.$$

Na osnovu poređenja ove matrice sa matricom ***R*** možemo zaključiti da zajednički faktori relativno uspešno reprodukuju korelacionu strukturu originalnih promenljivih. Kvalitet aproksimacije smo takođe mogli sagledati i na osnovu matrice reziduala:

$$R-(\tilde{B}\tilde{B}'+\Psi)=\begin{bmatrix}0 & 0.0325 & -0.0871 & -0.0457 & 0.0192\\ & 0 & 0.0204 & -0.1237 & -0.0608\\ & & 0 & -0.0423 & 0.0029\\ & & & 0 & -0.0569\\ & & & & 0\end{bmatrix}.$$

5.3.2 Metoda glavnih faktora

Ako analizu zasnivamo na korelacionoj matrici, nju razlažemo prema faktorskom modelu

$$\rho = BB' + \Psi$$

gde su dijagonalni elementi

$$\rho_{ii} = 1 = h_i^2 + \Psi_i, \quad i = 1, 2, \ldots, p.$$

Zajednički faktori objašnjavaju vandijagonalne elemente matrice $\boldsymbol{\rho}$ i deo dijagonalnih elemenata h_i^2. Zato formiramo novu matricu

$$\rho - \Psi = BB' = \left[\begin{array}{cc:c} h_1^2 & \rho_{12} & \rho_{1p} \\ \rho_{21} & h_2^2 & \rho_{2p} \\ \hdashline \rho_{p1} & \rho_{p2} & h_p^2 \end{array}\right]$$

koja predstavlja korelacionu matricu zajedničkih faktora i naziva se ***redukovana korelaciona matrica.***

Do ocene korelacione matrice zajedničkih faktora dolazimo na osnovu razlike ocenjene korelacione matrice i matrice specifičnih varijansi.

Kako je

$$\Psi_i = 1 - h_i^2$$

to je potrebno oceniti komunalitet pre nego što se primeni metod glavnih faktora za ocenu modela faktorske metode. Postoje dve osnovne grupe metoda za ocenjivanje komunaliteta:

1. Za ocenu *i*-tog komunaliteta uzima se maksimalna vrednost koeficijenta korelacije *i*-te promenljive i ma koje od preostalih *p-1* promenljivih. Apsolutna vrednost tog koeficijenta korelacije zamenjuje zatim *i*-tu jedinicu na glavnoj dijagonali korelacione matrice (dobra metoda za veliko *p*).
2. Koristi se celokupna korelaciona matrica i ocena za h_i^2 je kvadrat višestrukog koeficijenta korelacije promenljive X_i i svih preostalih *p-1* promenljivih. Ova metoda daje donju granicu ocene komunaliteta i bazira se na matrici $\boldsymbol{R^{-1}}$, tako da je ocena komunaliteta

$$\tilde{h}_i^2 = 1 - \tilde{\Psi}_i = 1 - \frac{1}{r^{ii}}, \quad i = 1, 2, \ldots, p$$

gde je r^{ii} *i*-ti dijagonalni element matrice $\boldsymbol{R^{-1}}$.

Nakon ocene komunaliteta i formiranja redukovane korelacione matrice vrši se izbor broja faktora istim postupcima kao kod metode glavnih komponenti. Potom se određuju karakteristični vektori redukovane korelacione matrice koji odgovaraju prvim *m* pozitivnim karakterističnim korenima te matrice.

Prema tome, ocena matrice faktorskih opterećenja je

$$B^* = \left[\sqrt{\tilde{\lambda}_1^*}\,\tilde{\alpha}_1^* \quad \sqrt{\tilde{\lambda}_2^*}\,\tilde{\alpha}_2^* \;\vdots\; \sqrt{\tilde{\lambda}_m^*}\,\tilde{\alpha}_m^* \right]$$

gde su

$$\left(\tilde{\lambda}_j^*, \tilde{\alpha}_j^*\right), \quad j = 1,2,\ldots,m$$

parovi karakterističnih korena i pridruženih karakterističnih vektora redukovane korelacione matrice.

Ocena specifične varijanse je

$$\tilde{\Psi}_i^* = 1 - \sum_{j=1}^{m} \tilde{\beta}_{ij}^{*2}, \quad i = 1,2,\ldots,p$$

gde su $\tilde{\beta}_{ij}^*$ ocenjena faktorska opterećenja.

Na osnovu ovako ocenjene specifične varijanse možemo ponovo oceniti komunalitete

$$\tilde{h}_i^{*2} = \sum_{j=1}^{m} \tilde{\beta}_{ij}^{*2}, \quad i = 1,2,\ldots,p$$

a zatim ih u narednoj iteraciji koristiti za formiranje redukovane korelacione matrice i određivanje njenih karakterističnih vektora. Ovaj iterativni proces ocenjivanja modela faktorske analize nastavljamo do momenta kada promene u sukcesivnim ocenama komunaliteta ne budu zanemarive.

Primer 5.5:

Neka je poznata kovarijaciona matrica Σ vektora $X^T=(X_1, X_2, X_3, X_4)$

$$\Sigma = \begin{bmatrix} 19 & 30 & 2 & 12 \\ & 57 & 5 & 23 \\ & & 38 & 47 \\ & & & 68 \end{bmatrix}$$

i neka se znaju faktorska opterećenja za prvi faktor F_1

$$a_{11} = 4,\ a_{12} = 7,\ a_{13} = -1\ i\ a_{14} = 1$$

kao i doprinos specifičnog faktora ukupnoj varijansi

$$\rho_1^2 = 2,\ \rho_2^2 = 4,\ \rho_3^2 = 1\ i\ \rho_4^2 = 3.$$

Izračunati:

a) Faktorska opterećenja za faktor F_2 ako se pretpostavi da je model sa dva faktora i datim uticajem specifičnog faktora u potpunosti sposoban da objasni kovarijacionu matricu.

b) Proporcije varijabiliteta varijabli X_1, X_2, X_3 i X_4 objašnjene faktorima F_1 i F_2.

c) Promene u reprezentovanju kovarijacione matrice vektora $\boldsymbol{X}$, ako se inicijalno ocenjeni faktori rotiraju matricom $\boldsymbol{G}$ za koju važi da je $\boldsymbol{G^T=G^{-1}}$. Skicirati izgled rotacije od 30° i promene pozicije faktorskih opterećenja.

Rešenje:

a)

$$h_i^2 = \sigma_{ii}^2 - \rho_i, \qquad h_i^2 = \beta_{1i}^2 + \beta_{2i}^2 = a_{1i}^2 + a_{2i}^2$$

$$\begin{array}{llll} h_1^2 = 17 & h_2^2 = 53 & h_3^2 = 37 & h_4^2 = 65 \\ a_{21}^2 = 1 & a_{22}^2 = 4 & a_{23}^2 = 36 & a_{24}^2 = 64 \\ a_{21} = 1 & a_{22} = 2 & a_{23} = 6 & a_{24} = 8 \end{array}$$

Faktorska opterećenja za faktor F_2:

$$F_2\left(a_{21} = 1 \quad a_{22} = 2 \quad a_{23} = 6 \quad a_{24} = 8\right)$$

b) Faktorima F_1 i F_2 je objašnjeno:

za	X_1	0.8947	varijabiliteta,
za	X_2	0.9298	varijabiliteta,
za	X_3	0.9736	varijabiliteta,
za	X_4	0.9559	varijabiliteta.

c) Ako se inicijalno ocenjeni faktori rotiraju ortogonalnom matricom $\boldsymbol{G}$ za koju važi da je $\boldsymbol{G^T=G^{-1}}$, kovarijaciona matrica se neće promeniti. Rotacija za 30°:

$$M = \begin{bmatrix} \cos 30° & \sin 30° \\ -\sin 30° & \cos 30° \end{bmatrix} = \begin{bmatrix} 0.866 & 0.5 \\ -0.5 & 0.866 \end{bmatrix}$$

$$\tilde{\Gamma} = \tilde{B}M$$

$$\tilde{B} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix}, \quad \tilde{\Gamma} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 0.866 & 0.5 \\ -0.5 & 0.866 \end{bmatrix} = \begin{bmatrix} 2.964 & 2.868 \\ 5.062 & 5.236 \\ -3.866 & 4.708 \\ -3.134 & 7.444 \end{bmatrix}.$$

Rotacijom za 30° se dobija matrica

$$\tilde{\Gamma} = \begin{bmatrix} 2.964 & 2.868 \\ 5.062 & 5.236 \\ -3.866 & 4.708 \\ -3.134 & 7.444 \end{bmatrix}.$$

6. KLASTER ANALIZA

Primer 6.1:

U tabeli su dati podaci o proizvodnji belog i crnog vina za pet proizvođača (u hiljadama litara):

redni broj	proizvođač	belo vino	crno vino
1	NAVIP	5	4
2	SMEDEREVKA	2	4
3	VINO ŽUPA	4	3
4	KRAJINA VINO	7	6
5	KOSOVO VINO	10	9

Izračunati apsolutno i kvadratno odstojanje između proizvođača.

$$E = \{NAVIP, SMEDEREVKA, VINO\ ZUPA, KRAJINA\ VINO, KOSOVO\ VINO\}$$

$$E = \{e_1, e_2, e_3, e_4, e_5\}, \quad N = 5$$

$$X = (belo\,vino \quad crno\,vino)$$

$$X = (X_1 \quad X_2), \quad p = 2$$

- apsolutno odstojanje:

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|, \quad k = 1,2$$

$$d_{11} = 0$$

$i = 1, j = 2$:

$$d_{12} = |x_{11} - x_{21}| + |x_{12} - x_{22}| = |5-2| + |4-4| = 3$$

$i = 1, j = 3$:

$$d_{13} = |x_{11} - x_{31}| + |x_{12} - x_{32}| = |5-4| + |4-3| = 2$$

$$d_{14} = |5-7| + |4-6| = 2+2 = 4$$

$$d_{15} = |5-10| + |4-9| = 5+5 = 10$$

$$d_{21} = d_{12}$$

$$d_{22} = 0$$

$i = 2, j = 3$:

$$d_{23} = |x_{21} - x_{31}| + |x_{22} - x_{32}| = |2-4| + |4-3| = 2+1 = 3$$

$$d_{24} = |2-7| + |4-6| = 5+2 = 7$$

$$d_{25} = |2-10| + |4-9| = 8+5 = 13$$

$$d_{34} = |4-7| + |3-6| = 3+3 = 6$$

$$d_{35} = |4-10| + |3-9| = 6+6 = 12$$

$$d_{45} = |7-10| + |6-9| = 3+3 = 6$$

$$D = \begin{bmatrix} 0 & 3 & 2 & 4 & 10 \\ & 0 & 3 & 7 & 13 \\ & & 0 & 6 & 12 \\ & & & 0 & 6 \\ & & & & 0 \end{bmatrix}$$

- kvadratno euklidsko odstojanje:

$$d_{12}^2 = (x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 = (5-2)^2 + (4-4)^2$$

$$d_{13}^2 = (5-4)^2 + (4-3)^2 = 1^2 + 1^2 = 2$$

$$d_{14}^2 = (5-7)^2 + (4-6)^2 = 4+4 = 8$$

...

$$D=\begin{bmatrix} 0 & 9 & 2 & 8 & 50 \\ & 0 & 5 & 29 & 89 \\ & & 0 & 18 & 72 \\ & & & 0 & 18 \\ & & & & 0 \end{bmatrix}$$

Primer 6.2:

Na osnovu podataka iz prethodnog primera formirati hijerarhijsku klasifikaciju korišćenjem metode pojedinačnog povezivanja ***(single linkage)*** a zatim nacrtati odgovarajući ***dendogram.***

$$\begin{array}{ll} 1 & NAVIP \\ 2 & SMEDEREVKA \\ 3 & VINO\ ZUPA \\ 4 & KRAJINA\ VINO \\ 5 & KOSOVO\ VINO \end{array} \begin{bmatrix} 0 & 9 & \textcircled{2} & 8 & 50 \\ & 0 & 5 & 29 & 89 \\ & & 0 & 18 & 72 \\ & & & 0 & 18 \\ & & & & 0 \end{bmatrix} = D^{(1)}$$

Prvo se povezuju ona dva elementa koja imaju najmanje odstojanje

$$\min d_{ij} = 2 = d_{13}$$

Znači, 1 NAVIP i 3 VINO ŽUPA čine jedan klaster. Sada primenjujemo algoritam klasifikovanja

$$D(e_i, e_j) = \min d_{sl}$$

$$\begin{array}{ll} 1 \cup 3 & NAVIP \cup VINO\ ZUPA \\ 2 & SMEDEREVKA \\ 4 & KRAJINA\ VINO \\ 5 & KOSOVO\ VINO \end{array} \begin{bmatrix} 0 & \textcircled{5} & 8 & 50 \\ & 0 & 29 & 89 \\ & & 0 & 18 \\ & & & 0 \end{bmatrix} = D^{(2)}$$

$$\min d_{ij} = 5$$

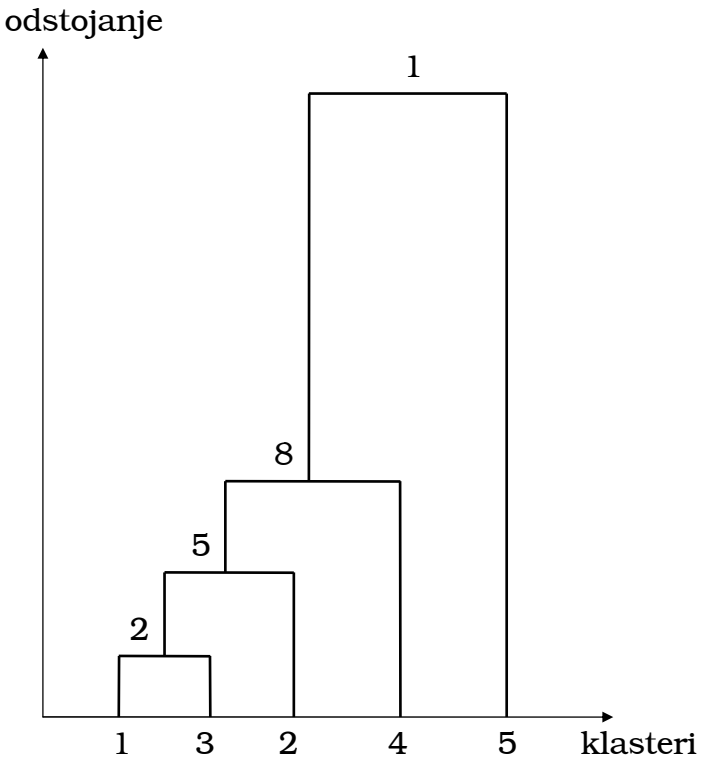
Klasteru NAVIP **U** VINO ŽUPA se pridružuje klaster 2 SMEDEREVKA.

$$\begin{array}{ll} 1 \cup 3 \cup 2 & \\ 4 & KRAJINA\ VINO \\ 5 & KOSOVO\ VINO \end{array} \begin{bmatrix} 0 & \textcircled{8} & 50 \\ & 0 & 18 \\ & & 0 \end{bmatrix} = D^{(3)}$$

Klasteru 1 **U** 3 **U** 2 se pridružuje klaster 4 KRAJINA VINO.

$$\begin{matrix} 1 \cup 3 \cup 2 \cup 4 \\ 5 \; KOSOVO\; VINO \end{matrix} \begin{bmatrix} 0 & 18 \\ & 0 \end{bmatrix}$$

Dendogram:



$$D^{(h)} = \begin{bmatrix} 0 & 5 & 2 & 8 & 18 \\ & 0 & 5 & 8 & 18 \\ & & 0 & 8 & 18 \\ & & & 0 & 18 \\ & & & & 0 \end{bmatrix}$$

Primer 6.3:

Neka su dati podaci o indeksu rezultata i snage sedam desetobojaca:

takmičar	rang snage	rang rezultata
A	10	8
B	2	2
V	2	5
G	4	2
D	2	3
E	8	9
Z	8	7

K-mean algoritmom podeliti skup desetobojaca na dve grupe sa početnim rešenjima za prvi centroid *A*, a za drugi centroid *D* i izračunati međusobno odstojanje konačno formiranih klasa.

$$A^0 = (10,8), B^0 = (2,3)$$

$$\begin{aligned}
&d\left(e_1, A^0\right) = 14 \quad d\left(e_1, B^0\right) = 1 \quad e_1 \to B' \\
&d\left(e_2, A^0\right) = 11 \quad d\left(e_2, B^0\right) = 2 \quad e_2 \to B' \\
&d\left(e_3, A^0\right) = 12 \quad d\left(e_3, B^0\right) = 3 \quad e_3 \to B' \\
&d\left(e_4, A^0\right) = 3 \quad d\left(e_4, B^0\right) = 12 \quad e_4 \to A' \\
&d\left(e_5, A^0\right) = 3 \quad d\left(e_5, B^0\right) = 10 \quad e_5 \to A'
\end{aligned}$$

$$A' = (8.67,8), B' = (2.5,3)$$

$$\begin{aligned}
&d\left(e_1, A'\right) = 12.67 \quad d\left(e_1, B'\right) = 1.5 \quad e_1 \to B'' \\
&d\left(e_2, A'\right) = 9.67 \quad d\left(e_2, B'\right) = 2.5 \quad e_2 \to B'' \\
&d\left(e_3, A'\right) = 10.67 \quad d\left(e_3, B'\right) = 2.5 \quad e_3 \to B'' \\
&d\left(e_4, A'\right) = 1.67 \quad d\left(e_4, B'\right) = 11.5 \quad e_4 \to A'' \\
&d\left(e_5, A'\right) = 1.67 \quad d\left(e_5, B'\right) = 9.5 \quad e_5 \to A''
\end{aligned}$$

Konačno rešenje:

$$A''\{A, E, Z\} \quad B''\{B, V, G, D\}.$$

Međusobno odstojanje klasa:

$$d_{I,II} = 11.17.$$

7. REŠENI ZADACI

Zadatak 1.

Neka je data kovarijaciona matrica

$$\Sigma = \begin{bmatrix} 13 & -4 & 2 \\ & 13 & -2 \\ & & 10 \end{bmatrix}$$

i neka su poznata dva korena jednačine $|\Sigma - \lambda I| = 0$

$$\lambda_1 = 9 \;\; i \;\; \lambda_2 = 9 .$$

a) Odrediti varijabilitet prve glavne komponente.
b) Odrediti analitički oblik druge glavne komponente.
c) Koliko treba uzeti glavnih komponenata da bi bilo obuhvaćeno barem 70% ukupnog varijabiliteta vektora ***X***?

a)

$$tr(\Sigma) = tr(\Lambda) \Rightarrow 13+13+10 = \lambda_1 + \lambda_2 + \lambda_3$$
$$36 = 18 + \lambda_3 \Rightarrow \lambda_3 = 36 - 18 = 18$$

Varijabilitet prve glavne komponente je

$$\max \lambda_i = 18 .$$

b)

$$\Sigma P_2 = \lambda_2 P_2 \Leftrightarrow \begin{bmatrix} 13 & -4 & 2 \\ -4 & 13 & -2 \\ 2 & -2 & 10 \end{bmatrix} \begin{bmatrix} p_{12} \\ p_{22} \\ p_{32} \end{bmatrix} = 9 \begin{bmatrix} p_{12} \\ p_{22} \\ p_{32} \end{bmatrix}$$

$$\begin{aligned} 13p_{12} - 4p_{22} + 2p_{32} &= 9p_{12} \\ -4p_{12} + 13p_{22} - 2p_{32} &= 9p_{22} \\ 2p_{12} - 2p_{22} + 10p_{32} &= 9p_{32} \end{aligned} \Leftrightarrow \begin{aligned} 4p_{12} - 4p_{22} + 2p_{32} &= 0 \\ -4p_{12} + 4p_{22} - 2p_{32} &= 0 \\ 2p_{12} - 2p_{22} + p_{32} &= 0 \end{aligned}$$

Dobili smo tri jednačine sa tri nepoznate ali su prve dve redudantne. Za P_{21} i P_{22} odgovara bilo koje rešenje, npr $P_{21}=1$ i $P_{22}=1$, Iz treće jednačine dobijamo da je $P_{32}=0$. Vektor

$$\begin{pmatrix} 1 & 1 & 0 \end{pmatrix}^T$$

zadovoljava sistem jednačina ali ne i uslov normiranosti karakterističnih vektora. Da bi se to postiglo, rešenja se dele sa korenom zbira kvadrata, tj. sa

$$\sqrt{1^2 + 1^2 + 0^2} = \sqrt{2}$$

pa je konačno

$$P_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{pmatrix}^T.$$

c)

$$\frac{\lambda_1}{tr(\Lambda)} = 0.5,\ \frac{\lambda_2}{tr(\Lambda)} = 0.25,\ \frac{\lambda_3}{tr(\Lambda)} = 0.25,\ tr(\Sigma) = tr(\Lambda) = 36$$

$$V(P_1^T X) < 0.7 \cdot 36$$

$$V(P_1^T X) + V(P_2^T X) > 0.7 \cdot 36$$

Odavde zaključujemo da treba uzeti prve dve glavne komponente da bi se obuhvatilo najmanje 70% ukupnog varijabiliteta.

Zadatak 2.

Ako su poznati varijabiliteti prve dve glavne komponente $V(Y_1)=18$ i $V(Y_2)=9$ i ako se zna ukupni varijabilitet vektora $\boldsymbol{X}$

$$tr(\Sigma)=36$$

pri čemu su glavne komponente date sa

$$Y_1=\frac{2}{3}X_1-\frac{2}{3}X_2+\frac{1}{3}X_3$$
$$Y_2=\frac{1}{\sqrt{2}}X_1+\frac{1}{\sqrt{2}}X_2$$
$$Y_3=\frac{1}{\sqrt{18}}X_1-\frac{1}{\sqrt{18}}X_2-\frac{4}{\sqrt{18}}X_3$$

a) odrediti procenat varijabiliteta koji nosi treća komponenta,
b) odrediti kovarijacionu matricu vektora $\boldsymbol{X}$.

a)

$$V(Y_3)=\lambda_3=tr(\lambda)-\left(\lambda_1+\lambda_2\right)=tr(\lambda)-\lambda_1-\lambda_2=36-27=9$$

$$\frac{\lambda_3}{tr(\Lambda)}\cdot 100=\frac{9}{36}\cdot 100=25\%\ \text{ varijabiliteta.}$$

b)

$$V(X)=\Sigma=P\Lambda P^T=P\begin{bmatrix}\lambda_1 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_3\end{bmatrix}P^T=\lambda_1P_1P_1^T+\lambda_2P_2P_2^T+\lambda_3P_3P_3^T=$$

$$=18\cdot\begin{bmatrix}\frac{2}{3}\\ -\frac{2}{3}\\ \frac{1}{3}\end{bmatrix}\begin{bmatrix}\frac{2}{3} & -\frac{2}{3} & \frac{1}{3}\end{bmatrix}+9\cdot\begin{bmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\\ 0\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0\end{bmatrix}+9\cdot\begin{bmatrix}\frac{1}{\sqrt{18}}\\ -\frac{1}{\sqrt{18}}\\ -\frac{4}{\sqrt{18}}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{18}} & -\frac{1}{\sqrt{18}} & -\frac{4}{\sqrt{18}}\end{bmatrix}=$$

$$=18\cdot\begin{bmatrix}\frac{4}{9} & -\frac{4}{9} & \frac{2}{9}\\ -\frac{4}{9} & \frac{4}{9} & -\frac{2}{9}\\ \frac{2}{9} & -\frac{2}{9} & \frac{1}{9}\end{bmatrix}+9\cdot\begin{bmatrix}\frac{1}{2} & \frac{1}{2} & 0\\ \frac{1}{2} & \frac{1}{2} & 0\\ 0 & 0 & 0\end{bmatrix}+9\cdot\begin{bmatrix}\frac{1}{18} & -\frac{1}{18} & -\frac{4}{18}\\ -\frac{1}{18} & \frac{1}{18} & \frac{4}{18}\\ -\frac{4}{18} & \frac{4}{18} & \frac{16}{18}\end{bmatrix}=$$

$$= \begin{bmatrix} 13 & -4 & 2 \\ -4 & 13 & -2 \\ 2 & -2 & 10 \end{bmatrix} = \Sigma .$$

Zadatak 3.

Neka je dat vektor

$$X = \begin{pmatrix} X_1 & X_2 & X_3 \end{pmatrix}^T$$

i njegova kovarijaciona matrica

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

sa karakterističnim vektorima

$$P_1^T = \begin{pmatrix} 0.383 & -0.924 & 0 \end{pmatrix}$$
$$P_2^T = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$$
$$P_3^T = \begin{pmatrix} 0.924 & 0.383 & 0 \end{pmatrix}$$

a) Izračunati glavne komponente Y_1, Y_2 i Y_3 i njihove varijanse.
b) Izračunati kovarijansu i korelaciju prve glavne komponente Y_1 sa varijablom X_1.

a)

$$Y_1 = P_1^T X = 0.383X_1 - 0.924X_2$$
$$Y_2 = P_2^T X = X_3$$
$$Y_3 = P_3^T X = 0.924X_1 - 0.383X_2$$

$$V(Y_1) = V(0.383X_1 - 0.924X_2) = 0.383^2 V(X_1) + 0.924^2 V(X_2) +$$
$$+ 2 \cdot 0.383 \cdot (-0.924) Cov(X_1, X_2) = 0.147 \cdot 1 + 0.854 \cdot 5 - 0.708 \cdot (-2) =$$
$$= 5.83 = \lambda_1$$

$$V(Y_2) = V(X_3) = 2 = \lambda_2$$

$$V(Y_3) = tr(\Sigma) - (\lambda_1 + \lambda_2) = 8 - 7.83 = 0.17$$

b) X_1 se može napisati kao

$$\begin{pmatrix}1 & 0 & 0\end{pmatrix}\cdot\begin{bmatrix}X_1\\X_2\\X_3\end{bmatrix}=L_1^T X$$

$$Cov(X_1,Y_1)=Cov(L_1^T X,P_1^T X)=L_1^T Cov(X,X)P_1 =$$
$$=L_1^T V(X)P_1=L_1^T\Sigma P_1=L_1^T\lambda_1 P_1$$

jer je $\Sigma P_1=\lambda_1 P_1$. Stoga je

$$Cov(X_1,Y_1)=\begin{pmatrix}1 & 0 & 0\end{pmatrix}\cdot 5.83\cdot\begin{bmatrix}0.383\\-0.924\\0\end{bmatrix}=5.83\cdot 0.383=2.23$$

$$Cor(X_1,Y_1)=\frac{Cov(X_1,Y_1)}{\sqrt{V(X_1)V(Y_1)}}=\frac{2.23}{\sqrt{1\cdot 5.83}}=0.925.$$

Zadatak 4.

Neka je data korelaciona matrica

$$R=\begin{bmatrix}1 & 0.63 & 0.45\\ & 1 & 0.35\\ & & 1\end{bmatrix}$$

a) Pokazati da se $\boldsymbol{R}$ može generisati modelom faktorske analize sa jednim faktorom tipa

$$X_1^*=0.9F_1+\varepsilon_1$$
$$X_2^*=0.7F_1+\varepsilon_2$$
$$X_3^*=0.5F_1+\varepsilon_3$$

pri čemu je

$$X^* = \frac{X - \mu}{\sigma}$$

$$V(F_1) = 1$$

$$Cov(F_1, \varepsilon_i) = 0, \quad i = 1,2,3$$

$$V(\varepsilon) = \begin{bmatrix} 0.19 & & \\ & 0.51 & \\ & & 0.75 \end{bmatrix}$$

b) Izračunati komunalitete.

a) Model je $X = AF_1 + \varepsilon$, a ***R*** se ocenjuje sa

$$\tilde{R} = AA^T + V(\varepsilon) = \begin{bmatrix} 0.9 \\ 0.7 \\ 0.5 \end{bmatrix} \begin{bmatrix} 0.9 & 0.7 & 0.5 \end{bmatrix} + \begin{bmatrix} 0.19 & & \\ & 0.51 & \\ & & 0.75 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.81 & 0.63 & 0.45 \\ 0.63 & 0.49 & 0.35 \\ 0.45 & 0.35 & 0.25 \end{bmatrix} + \begin{bmatrix} 0.19 & & \\ & 0.51 & \\ & & 0.75 \end{bmatrix} = R$$

Kako je $R - \tilde{R} = 0$ zaključujemo da model generiše ***R***.

b)

$$h_i^2 = 1 - \varepsilon_i \Rightarrow h_1^2 = 0.81,\ h_2^2 = 0.49 \ \ i \ \ h_3^2 = 0.25.$$

Zadatak 5.

Neka je data matrica podataka

	X_1	X_2
e_1	5	3
e_2	-1	1
e_3	1	-2
e_4	-3	-2

a) Podeliti skup $E = \{e_1, e_2, e_3, e_4\}$ na dva skupa ***Quick cluster*** algoritmom korišćenjem euklidskog odstojanja $d_{ij} = \sum_{k=1}^{2} (X_{ik} - X_{jk})^2$ uz dato početno rešenje da je

$$A^0 = \{e_1, e_2\} \ i \ B^0 = \{e_3, e_4\}.$$

b) Izračunati na isti način distancu između konačno formiranih klasa.

a) Centroidi početnog rešenja:

	X_1	X_2
A^0	$\frac{5-1}{2} = 2$	$\frac{3+1}{2} = 2$
B^0	$\frac{1+3}{2} = 2$	$\frac{-2-2}{2} = -2$

$$d(e_1, A^0) = (5-2)^2 + (3-2)^2 = 10$$
$$d(e_1, B^0) = (5-1)^2 + (3+2)^2 = 61 \Rightarrow d(e_1, A^0) < d(e_1, B^0) \Rightarrow e_1 \to A'$$
$$d(e_2, A^0) = 10$$
$$d(e_2, B^0) = 9 \Rightarrow e_2 \to B'$$
$$d(e_3, A^0) = 17$$
$$d(e_3, B^0) = 1 \Rightarrow e_3 \to B'$$
$$d(e_4, A^0) = 41$$
$$d(e_4, B^0) = 25 \Rightarrow e_4 \to B'$$

$$A' = \{e_1\} \ i \ B' = \{e_2, e_3, e_4\}$$

pa su centroidi

$$A'(5,3) \ i \ B'(-1,-1)$$

$$d(e_1, A') = 0$$
$$d(e_1, B') = 52 \Rightarrow e_1 \to A''$$
$$d(e_2, A') = 40$$
$$d(e_2, B') = 0 \Rightarrow e_2 \to B''$$
$$d(e_3, A') > d(e_3, B') \Rightarrow e_3 \to B''$$
$$d(e_4, A') > d(e_4, B') \Rightarrow e_4 \to B''$$

Kako je ***A'=A"*** i ***B'=B"*** konačna klasifikacija je:

$$A = \{e_1\} \ i \ B = \{e_2, e_3, e_4\}$$

b) $d(A,B) = (\overline{X}_{1A} - \overline{X}_{1B})^2 + (\overline{X}_{2A} - \overline{X}_{2B})^2 = 52.$

8. LITERATURA

1. Kovačić Zlatko: *"Multivarijaciona analiza"*