



UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI
FAKULTET
DEPARTMAN ZA MATEMATIKU I
INFORMATIKU



Tanja Stanić

ANALIZA GLAVNIH KOMPONENTI SA PRIMERIMA U STATISTIČKIM SOFTVERIMA R I STATISTICA

-MASTER RAD-

Mentor

dr Zagorka Lozanov-Crvenković

Novi Sad, 2019.

Sadržaj

Predgovor	6
1 Uvodni pojmovi	7
1.1 Višedimenzionalne slučajne promenljive	7
1.1.1 Numeričke karakteristike slučajnih promenljivih	8
1.1.2 Kovarijansna matrica	9
1.1.3 Korelaciona matrica	11
1.2 Uzorak iz višedimenzionalne raspodele	12
1.2.1 Uzoračka sredina, kovarijansna i korelaciona matrica	13
1.2.2 Geometrijska interpretacija uzorka	15
1.3 Višedimenzionalna normalna raspodela	17
1.4 Uzorak iz višedimenzionalne normalne raspodele	18
1.5 Zaključivanje na osnovu sredine	19
1.5.1 Interval poverenja	19
1.5.2 Testiranje hipoteza	21
1.6 Zaključivanje na osnovu kovarijansne i korelacione matrice	22
2 Analiza glavnih komponenti	23
2.1 Uvod	23
2.2 Definicija glavnih komponenti	25
2.3 Osobine glavnih komponenti	27
2.3.1 Algebarska i statistička svojstva	27
2.3.2 Geometrijska svojstva	31
2.4 Interpretacija glavnih komponenti na osnovu korelacione matrice	32
2.5 Osobine glavnih komponenti čije su varijanse međusobno jednake ili jednake nuli	35
2.6 Uzoračke glavne komponente	36
2.6.1 Definicija i svojstva uzoračkih glavnih komponenti	36
2.6.2 SVD dekompozicija	39
2.6.3 Funkcija raspodele uzoračkih glavnih komponenti	41
2.6.4 Određivanje intervala poverenja	42
2.6.5 Testiranje hipoteza	43
2.7 Izbor broja glavnih komponenti	46
2.8 Rotacija i interpretacija glavnih komponenti	49
2.9 Primena glavnih komponenti	52
3 Primeri u statističkim softverima	54
3.1 Statistica	54
3.2 R	66

Literatura

94

Pregled slika

1.1	<i>Dijagram rastojanja - rotacija koordinantnog sistema [17]</i>	16
2.1	<i>Projekcija skupa tačaka [17]</i>	24
2.2	<i>Scree test</i>	48
2.3	<i>Prikaz ortogonalne i kose rotacije</i>	51
2.4	<i>Primer autlajera</i>	53
3.1	<i>Izgled početnog menija za analizu u paketu Statistica</i>	54
3.2	<i>Definisanje promenljivih za analiziranje u paketu Statistica</i>	55
3.3	<i>Meni za analiziranje promenljivih u paketu Statistica</i>	56
3.4	<i>Scree test u paketu Statistica za podatke Cars.sta</i>	57
3.5	<i>Projekcija promenljivih u novom sistemu za podatke Cars.sta</i>	58
3.6	<i>Meni za analiziranje opservacija u paketu Statistica</i>	59
3.7	<i>Projekcija opservacija u novom sistemu za podatke Cars.sta</i>	59
3.8	<i>Projekcija opservacija bez prikazanog naziva opservacija za podatke Cars.sta</i>	60
3.9	<i>Definisanje promenljivih za podatke Activities.sta</i>	61
3.10	<i>Scree test za podatke Activities.sta</i>	61
3.11	<i>Projekcija promenljivih u novom sistemu za podatke Activities.sta</i>	62
3.12	<i>Projekcija opservacija u novom sistemu za podatke Activities.sta</i>	63
3.13	<i>Prikaz prvih nekoliko vrsta u podacima Mačići</i>	64
3.14	<i>Scree test za primer sa mačićima</i>	64
3.15	<i>Projekcija opservacija u novom sistemu za primer sa mačićima</i>	65
3.16	<i>Prikaz karakterističnih korena i vektora za podatke iris</i>	68
3.17	<i>Barplot objašnjene varijacije podataka za svaku komponentu za podatke iris</i>	69
3.18	<i>Scree test za podatke iris</i>	69
3.19	<i>Prikaz projekcije opservacija u novom sistemu za podatke iris</i>	70
3.20	<i>Prikaz projekcije opservacija u novom sistemu pomoću funkcije ggbiplot za podatke iris</i>	71
3.21	<i>Prikaz strukture objekta myPr</i>	72
3.22	<i>Matrica korelacije za podatke USJudgeRatings</i>	73
3.23	<i>Deskriptivna statistika i varijansa promenljivih za podatke USJudgeRatings</i>	73
3.24	<i>Karakteristični koreni i vektori standardizovanih promenljivih za podatke USJudgeRatings</i>	74
3.25	<i>Matrica koeficijenata glavnih komponenti za podatke USJudgeRatings</i>	75
3.26	<i>Prikaz matrice skorova za podatke USJudgeRatings</i>	75
3.27	<i>Grafički prikaz opservacija u koordinatnom sistemu sa komponentama za podatke USJudgeRatings</i>	76
3.28	<i>Scree test i kumulativni scree test za podatke USJudgeRatings</i>	77
3.29	<i>Prikaz strukture objekta rezult za podatke Diabetes</i>	79
3.30	<i>Karakteristični koreni za podatke Diabetes</i>	80
3.31	<i>Karakteristični koreni za podatke Diabetes</i>	80

3.32	<i>Prikaz dobijenih informacija o varijablama za podatke Diabetes</i>	81
3.33	<i>Koeficijenti promenljivih i grafički prikaz promenljivih za podatke Diabetes</i>	82
3.34	<i>Grafički prikaz kvaliteta reprezentovanih promenljivih pomoću funkcija corrplot i fviz_cos2 za podatke Diabetes</i>	83
3.35	<i>Grafički prikaz promenljivih za podatke Diabetes</i>	83
3.36	<i>Prikaz procenta učešća promenljivih u objašnjenju varijanse komponentama za podatke Diabetes</i>	83
3.37	<i>Grafički prikaz procenta učešća promenljivih u objašnjenju varijanse komponentama za podatke Diabetes</i>	84
3.38	<i>Prikaz procenta učešća promenljivih u objašnjenju varijanse sa prve dve komponente za podatke Diabetes</i>	84
3.39	<i>Prikaz dobijenih informacija o opservacijama za podatke Diabetes</i>	85
3.40	<i>Grafički prikaz skorova za podatke Diabetes</i>	85
3.41	<i>Grafički prikaz skorova u zavisnosti od koeficijenata reprezentativnosti za podatke Diabetes</i>	86
3.42	<i>Grafički prikaz procentualnog učešća skorova u prve dve komponente za podatke Diabetes</i>	86
3.43	<i>Grafički prikaz procentualnog učešća svih skorova u prve dve komponente za podatke Diabetes</i>	86
3.44	<i>Grafički prikaz skorova sa kvalitativnom promenljivom Outcome za podatke Diabetes</i>	87
3.45	<i>Grafički prikaz skorova sa parametrom confidence za podatke Diabetes</i>	88
3.46	<i>Rezultat funkcije dudi.pca za podatke Happiness</i>	89
3.47	<i>Scree plot za podatke Happiness</i>	90
3.48	<i>Grafički prikaz promenljivih za podatke Happiness</i>	90
3.49	<i>Grafički prikaz skorova za podatke Happiness</i>	91
3.50	<i>Grafički prikaz opservacija i promenljivih za podatke Happiness</i>	91
3.51	<i>Prikaz koeficijenta promenljivih za podatke Happiness</i>	92
3.52	<i>Grafički prikaz opservacija u paketu FactoMineR za podatke Happiness</i>	92

Pregled tabela

1.1	<i>Matrica podataka</i>	13
3.1	<i>Koeficijenti korelacije promenljivih za podatke Cars.sta</i>	55
3.2	<i>Karakteristični koreni korelacione matrice i procenat objašnjene varijanse podataka Cars.sta</i>	56
3.3	<i>Karakteristični vektori korelacione matrice za podatke Cars.sta</i>	57
3.4	<i>Koeficijenti zadržanih komponenti u odnosu na promenljive za podatke Cars.sta</i>	58
3.5	<i>Karakteristični koreni korelacione matrice i procenat objašnjene varijanse podataka Activities.sta</i>	61
3.6	<i>Koeficijenti zadržanih komponenti za odgovarajuće promenljive za podatke Cars.sta</i>	62
3.7	<i>Koeficijenti korelacije promenljivih za podatke Activities.sta</i>	63
3.8	<i>Koeficijenti korelacije promenljivih za primer sa mačićima</i>	64
3.9	<i>Prikaz prvih šest vrsta u podacima iris</i>	67
3.10	<i>Deskriptivna statistika za podatke iris</i>	67
3.11	<i>Prikaz koeficijenata glavnih komponenti za podatke iris</i>	67
3.12	<i>Prikaz proporcije objašnjene varijanse glavnim komponentama za podatke iris</i>	68
3.13	<i>Prikaz karakterističnih vektora za podatke iris</i>	68
3.14	<i>Prikaz skorova za prvih pet opservacija za podatke iris</i>	71
3.15	<i>Prikaz prve tri vrste u setu podataka USJudgeRatings</i>	72
3.16	<i>Prikaz proporcije objašnjene varijanse komponentama za podatke USJudgeRatings</i>	77
3.17	<i>Prikaz prve tri vrste Diabetes podataka</i>	78
3.18	<i>Prikaz prve tri vrste Happiness podataka</i>	89

Predgovor

” Ne postoji grana matematike koliko god bila apstraktna da jednog dana ne bi mogla biti primenjena u praksi”

Nikolay Ivanovich Lobachesky

Statistika je oblast matematike koja podrazumeva prikupljanje, prikazivanje, analizu i korišćenje podataka u cilju da se izvedu zaključci i donesu odluke. Primenjuje se skoro u svim oblastima, za davanje procena, istraživanje tendencija, procenu rizika, analiziranje odnosa i faktora koji određuju posmatranu pojavu.

Analiza glavnih komponenti predstavlja statističku analizu redukcije dimenzionalnosti skupa podataka. Predstavlja linearnu transformaciju originalnog, korelisanog skupa podataka u novi skup nekoreliranih promenljivih sa opadajućom vrednošću varijanse. Nove promenljive se nazivaju glavnim komponentama i one predstavljaju linearnu kombinaciju originalnih promenljivih. Primenjuje se kada se više promenljivih u skupu odnosi na istu dimenziju, a da ne pružaju neke dodatne informacije koje nisu već sadržane u nekoj drugoj promenljivoj.

U prvom poglavlju rada predstavljeni su relevantni matematički pojmovi neophodni za definisanje, razumevanje i primenu glavnih komponenti i to iz teorije verovatnoće, statistike i linearne algebre.

U drugom poglavlju se bavimo samim glavnim komponentama, kako populacionim, tako i uzoračkim. Bavimo se njihovim definisanjem, zatim svojstvima i to statističkim i geometrijskim, razmatramo različite načine njihove interpretacije, primenom.

Treće poglavlje je posvećeno primeni glavnih komponenti i implementaciji u statističkim softverima R i Statistica. Kroz primere je prikazan način korišćenja softvera za dobijanje komponenti, navedene su i istumačene naredbe za postupak dobijanja komponenti, za načine na koje je određen broj komponenti, za vizuelne prikaze, a dobijeni rezultati su istumačeni.

Na kraju, želim da izrazim posebnu zahvalnost svojoj mentorki, prof.dr Zagorki Lozanov-Crvenković, na pruženim savetima i podršci tokom pisanja ovog rada, a pre svega na pruženom znanju tokom studiranja.

Takođe, zahvalila bih se članovima komisije prof.dr Ljiljani Gajić i prof.dr Ivani Štajner-Papuga na svom prenetom znanju tokom studiranja.

Najveću zahvalnost dugujem roditeljima i bratu, na ljubavi i podršci koju mi neprestano pružaju.

Tanja Stanić

1 Uvodni pojmovi

1.1 Višedimenzionalne slučajne promenljive

Kao matematička osnova statistike, teorija verovatnoće je od velike važnosti za mnoge aktivnosti koje uključuju analizu velikih skupova podataka. U ovom delu ćemo objasniti neke osnovne pojmove iz teorije verovatnoće, preko kojih ćemo iskazati osnovne pokazatelje kako populacije tako i uzorka, koji će nam služiti za dalji rad.[25]

Definicija 1.1. *Prostor verovatnoća je trojka $(\Omega, \mathcal{F}, P)$, određena slučajnim eksperimentom, gde je Ω skup svih mogućih ishoda slučajnog eksperimenta, $\mathcal{F}$ σ -algebra, a P funkcija koja svakom skupu $A \in \mathcal{F}$ dodeljuje broj $P(A)$, koji se naziva verovatnoća da se desi događaj A .*

Definicija 1.2. *Preslikavanje $X : \Omega \rightarrow \mathbb{R}^n$, se zove n -dimenzionalna slučajna promenljiva nad prostorom verovatnoća $(\Omega, \mathcal{F}, P)$, ako važi da je X $\mathcal{F}$ -merljivo, tj. $X^{-1}(S) \in \mathcal{F}$, za svaki Borelov skup S .*

Označimo sa $X_1, X_2, \dots, X_p$ p jednodimenzionalnih slučajnih promenljivih. Skup ovih slučajnih promenljivih zapisujemo kao slučajan vektor $\mathbf{X}$ dimenzija $p \times 1$ odnosno

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

Slučajan vektor je vektor čiji su elementi slučajne promenljive, a matrica čije su kolone slučajni vektori se naziva slučajna matrica.

Definicija 1.3. *Funkcija raspodele n -dimenzionalne slučajne promenljive $\mathbf{X} = (X_1, \dots, X_n)$ je*

$$F_{\mathbf{X}}(x) = F_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = P(X_1 < x_1, \dots, X_n < x_n)$$

za $(x_1, \dots, x_n) \in \mathbb{R}^n$.

Definicija 1.4. *Marginalna funkcija raspodele za k -tu slučajnu promenljivu X_k , $k = 1, \dots, n$ u n -dimenzionalnoj slučajnoj promenljivoj $(X_1, X_2, \dots, X_p)$ je*

$$F_{X_k}(x_k) = F_{X_1, \dots, X_n}(\infty, \dots, \infty, x_k, \infty, \dots, \infty)$$

Definicija 1.5. *Slučajna promenljiva $\mathbf{X} = (X_1, \dots, X_n)$ je n -dimenzionalna slučajna promenljiva neprekidnog tipa ako postoji integrabilna funkcija $\varphi_{\mathbf{X}}(x_1, \dots, x_n) \geq 0$, $(x_1, \dots, x_n) \in \mathbb{R}^n$, takva da je, za svaki Borelov skup S*

$$P\{(X_1, \dots, X_n) \in S\} = \int_S \dots \int \varphi_{\mathbf{X}}(x_1, \dots, x_n) dx_1 \dots dx_n$$

Funkcija $\varphi_{\mathbf{X}}(x)$ se naziva gustina raspodele slučajne promenljive $\mathbf{X}$.

Specijalno, ako izaberemo $S = \{(u_1, \dots, u_n) \in \mathbb{R}^n, u_i < x_i\}$ važi:

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_n} \varphi_{\mathbf{X}}(u_1, \dots, u_n) du_n \dots du_1$$

1.1.1 Numeričke karakteristike slučajnih promenljivih

U ovom delu ćemo ukazati na pojedine parametre koji do izvesnog stepena karakterišu bitne crte raspodele slučajne promenljive. Očekivanje slučajne promenljive reprezentuje centar rasturanja vrednosti slučajne promenljive, a disperzija i standardna devijacija mere to rasturanje oko centra rasturanja.[25]

Očekivanje slučajne promenljive nazivamo još očekivana vrednost, srednja vrednost, sredina, prosek.

Definicija 1.6. Neka je X apsolutno neprekidna slučajna promenljiva sa gustinom $\varphi_X(x)$. Očekivanje slučajne promenljive X dato je sa

$$E(X) = \int_{-\infty}^{+\infty} x\varphi_X(x)dx$$

i ono postoji ako i samo $\int_{-\infty}^{+\infty} |x|\varphi_X(x)dx < +\infty$.

Teorema 1.1. Neka je X apsolutno neprekidna slučajna promenljiva sa gustinom $\varphi_X(x)$. Neka je $f : \mathbb{R} \rightarrow \mathbb{R}$ Borelova funkcija. Tada je:

$$E(f(X)) = \int_{-\infty}^{+\infty} f(x)\varphi_X(x)dx$$

Varijansa ili disperzija slučajne promenljive predstavlja meru odstupanja od srednje vrednosti.

Definicija 1.7. Momenat reda k , $k \in \mathbb{N}$, slučajne promenljive X je $E(X^k)$. Centralni momenat reda k slučajne promenljive X je $E((X - E(X))^k)$.

Definicija 1.8. Centralni momenat reda 2 slučajne promenljive X se zove varijansa (disperzija) slučajne promenljive X i označava se sa $D(X)$ ili $\sigma^2(X)$ ili $\text{Var}(X)$, tj.

$$D(X) = E((X - E(X))^2)$$

Definicija 1.9. Standardna devijacija (standardno odstupanje, prosečno odstupanje) slučajne promenljive X se definiše kao

$$\sigma(X) = \sqrt{D(X)}$$

Definicija 1.10. Neka je X slučajna promenljiva sa očekivanjem $E(X)$ i varijansom $D(X)$. Standardizovana (normalizovana) slučajna promenljiva X^* je

$$X^* = \frac{X - E(X)}{\sqrt{D(X)}} = \frac{X - E(X)}{\sigma(X)}$$

Teorema 1.2. Neka je X^* standardizovana slučajna promenljiva. Tada je

$$E(X^*) = 0 \quad i \quad D(X^*) = 1$$

1.1.2 Kovarijansna matrica

Neka je $\mathbf{X}$ vektor dimenzija $p \times 1$ čiji svaki element predstavlja slučajnu promenljivu sa svojom marginalnom raspodelom. Sredinu tih promenljivih označavamo sa $\mu_j = E(X_j)$, a varijansu sa $\sigma_j = E(X_j - \mu_j)^2$ za $j = 1, \dots, p$, koristimo i oznake σ_{jj} i $Var(X_j)$ za varijansu. Sredina slučajnog vektora $\mathbf{X}$ je

$$\boldsymbol{\mu} = E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

Za svake dve promenljive X_j i X_k definišemo kovarijansu $\sigma_{jk} = E[(X_j - \mu_j)(X_k - \mu_k)]$, koju još označavamo sa $Cov(X_j, X_k)$. Na osnovu definicije nam važi da je $Cov(X_j, X_j) = Var(X_j)$ i $Cov(X_j, X_k) = Cov(X_k, X_j) = \sigma_{kj} = \sigma_{jk}$.

Sada definišemo simetričnu matricu dimenzija $p \times p$ kod koje je j -ti dijagonalni element $\sigma_{jj} = Var(X_j)$, a element u j -toj vrsti i k -toj koloni je $\sigma_{jk} = Cov(X_j, X_k)$. Tu matricu nazivamo kovarijansna matrica i označavamo je sa $\boldsymbol{\Sigma}$, tj.

$$\begin{aligned} Cov(\mathbf{X}) = \boldsymbol{\Sigma} &= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix} \\ &= \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_p) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_p, X_1) & Cov(X_p, X_2) & \dots & Var(X_p) \end{bmatrix} \\ &= \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_p) \\ Cov(X_1, X_2) & Var(X_2) & \dots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_p) & Cov(X_2, X_p) & \dots & Var(X_p) \end{bmatrix} \end{aligned}$$

Kovarijansnu matricu možemo iskazati kao očekivanu vrednost slučajne matrice odnosno važi $E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] = \boldsymbol{\Sigma}$.

Sada navodimo nekoliko važnih rezultata za osobine sredine slučajnog vektora $\mathbf{X}$ i kovarijansnu matricu.

Na osnovu definicije kovarijanse sledi da je za $a, b, c, d \in \mathbb{R}$

$$Cov(cX_j + a, dX_k + b) = cdCov(X_j, X_k)$$

Sada ćemo ovaj slučaj da uopštimo na linearnu kombinaciju p slučajnih promenljivih iz slučajnog vektora $\mathbf{X}$, sa sredinom $\boldsymbol{\mu}$ i kovarijansnom matricom $\boldsymbol{\Sigma}$. Novu slučajnu promenljivu

Y definišemo pomoću linearne kombinacije $Y = a_1X_1 + a_2X_2 + \dots + a_pX_p = \mathbf{a}'\mathbf{X}$, za dati vektor koeficijenata linearne kombinacije $\mathbf{a}' = [a_1, \dots, a_p]$. Njena očekivana vrednost je

$$\mu_Y = E(Y) = E(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\boldsymbol{\mu}$$

a varijansa je

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

Varijansa slučajne promenljive Y je data kao kvadratna forma i u potpunosti je određena kovarijansnom matricom $\boldsymbol{\Sigma}$ i koeficijentima $a_1, \dots, a_p$.

Sada ćemo razmotriti opštiji slučaj, odnosno posmatramo q linearnih kombinacija p slučajnih promenljivih:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p$$

$$\vdots$$

$$Y_q = a_{q1}X_1 + a_{q2}X_2 + \dots + a_{qp}X_p$$

Možemo ih zapisati u matričnom obliku kao $\mathbf{Y} = \mathbf{A}\mathbf{X}$, gde su $\mathbf{Y}$ i $\mathbf{A}$ vektor dimenzija $q \times 1$ i matrica dimenzija $q \times p$ koeficijenata linearne kombinacije, respektivno. Sredina slučajnog vektora $\mathbf{Y}$ je

$$\boldsymbol{\mu}_Y = E(\mathbf{Y}) = E(\mathbf{A}\mathbf{X}) = \mathbf{A}\boldsymbol{\mu}_X$$

a kovarijansna matrica

$$\boldsymbol{\Sigma}_Y = \text{Cov}(\mathbf{Y}) = \text{Cov}(\mathbf{A}\mathbf{X}) = \mathbf{A}\boldsymbol{\Sigma}_X\mathbf{A}'$$

gde su $\boldsymbol{\mu}_X$ i $\boldsymbol{\Sigma}_X$ sredina i kovarijansna matrica vektora $\mathbf{X}$.

Neka su date matrice $\mathbf{A}$ dimenzija $q \times p$ i $\mathbf{B}$ dimenzija $s \times r$ čiji elementi su slučajni brojevi. Formiramo q linearnih kombinacija slučajnih promenljivih p odnosno $\mathbf{Y} = \mathbf{A}\mathbf{X}$ i formiramo s linearnih kombinacija r slučajnih promenljivih $\mathbf{Z}' = [Z_1, \dots, Z_r]$ odnosno $\mathbf{W} = \mathbf{B}\mathbf{Z}$. Ako sa $\boldsymbol{\Sigma}_{XZ}$ označimo kovarijansnu matricu dimenzija $p \times r$ između promenljivih X_j i Z_k , tada važi

$$\text{Cov}(\mathbf{Y}, \mathbf{W}) = \text{Cov}(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{Z}) = \mathbf{A}\boldsymbol{\Sigma}_{XZ}\mathbf{B}'$$

Na osnovu ovog najopštijeg rezultata prethodne rezultate tretiramo kao specijalne slučajeve.

1.1.3 Korelaciona matrica

Koeficijent korelacije između slučajnih promenljivih X_j i X_k definišemo na sledeći način

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}} \quad (1.1)$$

i on predstavlja normalizovanu kovarijansu između X_j i X_k . Koeficijent korelacije uzima vrednost iz intervala $[-1, 1]$. Ako koeficijent korelacije uzima pozitivne vrednosti, tada je korelacija direktna ili pozitivna, a ako uzima negativne vrednosti onda je korelacija inverzna ili negativna. Ukoliko je $\rho_{jk} = -1$ onda postoji perfektna linearna veza između X_j i X_k sa negativnim predznakom, a ukoliko je $\rho_{jk} = 1$ onda je ta veza sa pozitivnim predznakom. Kada je $\rho_{jk} = 0$ to tumačimo kao odsustvo korelacije.

Korelacionu matricu ρ možemo dobiti na osnovu kovarijansne matrice, njen j -ti element je definisan izrazom 1.1. U matricnoj notaciji veza između korelacione i kovarijansne matrice data je sledećim izrazom

$$\begin{aligned} \rho &= (\mathbf{D}^{\frac{1}{2}})^{-1} \Sigma (\mathbf{D}^{\frac{1}{2}})^{-1} \\ &= \begin{bmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\sigma_{pp}}} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\sigma_{pp}}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{bmatrix} \end{aligned}$$

gde smo sa $\mathbf{D}$ označili dijagonalnu matricu koja sadrži elemente kovarijansne matrice na glavnoj dijagonali. Na osnovu prethodne jednakosti, kovarijansnu matricu možemo zapisati na sledeći način $\Sigma = \mathbf{D}^{\frac{1}{2}} \rho \mathbf{D}^{\frac{1}{2}}$ [17].

1.2 Uzorak iz višedimenzionalne raspodele

Skup svih elemenata na kojima se proučava neka pojava naziva se statistički skup ili populacija. Osobine po kojima se elementi populacije razlikuju se nazivaju obeležja. Iz populacije se na određen način uzima jedan njen podskup koji se naziva uzorak. Slučajan uzorak se dobija postupkom kod kojeg svaki element populacije ima istu šansu da bude izabran u uzorak i tim se obezbeđuje reprezentativnost uzorka.

Populaciju ϵ možemo posmatrati kao skup svih mogućih ishoda nekog eksperimenta, a obeležje X kao slučajnu promenljivu $X : \epsilon \rightarrow \mathbb{R}$. Kada smo izabrali uzorak, na svakom elementu iz uzorka posmatra se obeležje X . Niz dobijenih vrednosti se naziva realizovani uzorak. [19]

Definicija 1.11. *Neka se na populaciji ϵ posmatra obeležje X . Prost slučajan uzorak obima n za obeležje X je n -torka nezavisnih slučajnih promenljivih $(X_1, X_2, \dots, X_n)$ od kojih svaka ima istu raspodelu kao i obeležje X .*

Definicija 1.12. *Neka je $(X_1, X_2, \dots, X_n)$ prost slučajan uzorak obima n za obeležje i $f : \mathbb{R}^n \rightarrow \mathbb{R}^s$ Borelova funkcija. Slučajna promenljiva $Y = f(X_1, X_2, \dots, X_n)$ je statistika ako u njoj ne figuriše nepoznati parametar.*

Statistike koje ćemo koristiti su:

1. uzoračka sredina

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

2. uzoračka varijansa

- ako je poznata sredina μ

$$\hat{S}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

- ako nije poznata sredina μ

$$\bar{S}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

3. popravljena uzoračka varijansa

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

4. standardna devijacija uzorka

$$\hat{S}_n = \sqrt{\hat{S}_n^2}, \quad \bar{S}_n = \sqrt{\bar{S}_n^2}, \quad S = \sqrt{S^2}$$

5. uzorački koeficijent korelacije

$$R_{XY} = \frac{\frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X}_n \bar{Y}_n}{\bar{S}_X \bar{S}_Y}$$

Pretpostavimo da smo tokom merenja prikupili podatke za i objekata $i = 1, \dots, n$ o njihovih j osobina $j = 1, \dots, p$. Podatke prezentujemo u vidu matrice podataka. Pošto smo pretpostavili da imamo n opservacija (objekata) odnosno vrsta i p slučajnih promenljivih (obeležja) odnosno kolona, matrica podataka nam ovako izgleda

	Promenljiva 1	Promenljiva 2	...	Promenljiva j	...	Promenljiva p
Opservacija 1	X_{11}	X_{12}	...	X_{1j}	...	X_{1p}
Opservacija 2	X_{21}	X_{22}	...	X_{2j}	...	X_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Opservacija i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ip}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Opservacija n	X_{n1}	X_{n2}	...	X_{nj}	...	X_{np}

Tabela 1.1: *Matrica podataka*

Matricu podataka $\mathbf{X}^1$ dimenzija $n \times p$ možemo tretirati dvojako u zavisnosti da li je posmatramo po vrstama ili po kolonama. Ako je posmatramo po vrstama, onda svaki red predstavlja jednu višedimenzionalnu opservaciju slučajnih promenljivih $X_1, X_2, \dots, X_p$ i u tom slučaju matrica podataka sadrži n opservacija višedimenzionalne slučajne promenljive. Ako je posmatramo po kolonama, onda svaka kolona predstavlja n opservacija jedne od p promenljivih.

1.2.1 Uzoračka sredina, kovarijansna i korelaciona matrica

Za slučajan uzorak uzet iz višedimenzionalne raspodele u mogućnosti smo da izračunamo pokazatelje uzorka: sredinu, kovarijansnu i korelacionu matricu. Ove pokazatelje koristimo u deskriptivne svrhe.

Uzoračka sredina, koju označavamo sa $\bar{\mathbf{X}}$, definišemo kao slučajan vektor dimenzija $p \times 1$ čiji su elementi uzoračke sredine promenljivih tj.

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

Uzoračku matricu uzajamnih proizvoda odstupanja od sredine, u oznaci $\mathbf{X}^{**'}\mathbf{X}^*$, defini-

¹Za matricu podataka i slučajan vektor koristimo istu oznaku, ali je iz konteksta jasno na šta se odnosi oznaka.

šemo kao simetričnu slučajnu matricu dimenzija $p \times p$ i to na sledeći način

$$\mathbf{X}^{*'}\mathbf{X}^* = \begin{bmatrix} \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 & \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) & \dots & \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{ip} - \bar{X}_p) \\ \sum_{i=1}^n (X_{i2} - \bar{X}_2)(X_{i1} - \bar{X}_1) & \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2 & \dots & \sum_{i=1}^n (X_{i2} - \bar{X}_2)(X_{ip} - \bar{X}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n (X_{ip} - \bar{X}_p)(X_{i1} - \bar{X}_1) & \sum_{i=1}^n (X_{ip} - \bar{X}_p)(X_{i2} - \bar{X}_2) & \dots & \sum_{i=1}^n (X_{ip} - \bar{X}_p)^2 \end{bmatrix}$$

gde je $\mathbf{X}^*$ matrica centriranih podataka

$$\mathbf{X}^* = \begin{bmatrix} X_{11} - \bar{X}_1 & X_{12} - \bar{X}_2 & \dots & X_{1p} - \bar{X}_p \\ X_{21} - \bar{X}_1 & X_{22} - \bar{X}_2 & \dots & X_{2p} - \bar{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} - \bar{X}_1 & X_{n2} - \bar{X}_2 & \dots & X_{np} - \bar{X}_p \end{bmatrix}$$

Preko matrice $\mathbf{X}^{*'}\mathbf{X}^*$ definišemo uzoračku kovarijansnu matricu, u oznaci $\mathbf{S}$, kao simetričnu slučajnu matricu dimenzija $p \times p$

$$\mathbf{S} = \frac{1}{n-1} \mathbf{X}^{*'}\mathbf{X}^* = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$$

gde vektor $(\mathbf{X}_i - \bar{\mathbf{X}})$ dimenzija $p \times 1$ za elementi ima članove i -te vrste matrice $\mathbf{X}^*$.

Na osnovu elementa u j -toj vrsti i k -toj koloni u uzoračkoj kovarijansnoj matrici

$$S_{jk} = \frac{1}{n-1} (\mathbf{X}_{ij} - \bar{X}_j)(\mathbf{X}_{ik} - \bar{X}_k) \quad j, k = 1, \dots, p$$

možemo odrediti uzoračku korelacionu matricu čiji je element u j -toj vrsti i k -toj koloni

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}}\sqrt{S_{kk}}} \quad j, k = 1, \dots, p$$

a korelaciona matrica ovako izgleda

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1p} \\ r_{12} & 1 & r_{23} & \dots & r_{2p} \\ r_{13} & r_{23} & 1 & \dots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & r_{3p} & \dots & 1 \end{bmatrix}$$

Do uzoračke korelacione matrice mogli smo doći koristeći relaciju koju smo uspostavili između populacione kovarijansne i korelacione matrice. Kod uzoračke kovarijansne i korelacione matrice imamo relaciju $\mathbf{R} = (\mathbf{D}^{\frac{1}{2}})^{-1} \mathbf{S} (\mathbf{D}^{\frac{1}{2}})^{-1}$.

Do uzoračke korelacione matrice možemo doći preko matrice standardizovanih opservacija $\tilde{\mathbf{X}}$ dimenzija $n \times p$, gde je

$$\tilde{\mathbf{X}} = \begin{bmatrix} \frac{X_{11}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{12}-\bar{X}_2}{\sqrt{S_{22}}} & \dots & \frac{X_{1p}-\bar{X}_p}{\sqrt{S_{pp}}} \\ \frac{X_{21}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{22}-\bar{X}_2}{\sqrt{S_{22}}} & \dots & \frac{X_{2p}-\bar{X}_p}{\sqrt{S_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{X_{n1}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{n2}-\bar{X}_2}{\sqrt{S_{22}}} & \dots & \frac{X_{np}-\bar{X}_p}{\sqrt{S_{pp}}} \end{bmatrix}$$

pa je uzoračka korelaciona matrica data sledećom formulom

$$\mathbf{R} = \frac{1}{n-1} \tilde{\mathbf{X}}' \tilde{\mathbf{X}}$$

1.2.2 Geometrijska interpretacija uzorka

Ako matricu podataka posmatramo po vrstama, tada svaka vrsta predstavlja jednu tačku u p -dimenzionalnom vektorskom prostoru. Stepem rasprostranjenosti n tačaka u ovom prostoru zavisi od sredine i varijanse promenljivih, kao i njihove uzajamne povezanosti. Prirodan način poređenja p realizovanih opservacija u dve vrste matrice podataka, jeste izračunavanje njihovog međusobnog rastojanja tzv. Euklidskog rastojanja. Često se sredina koristi kao tačka u odnosu na koju posmatramo odstojanje ostalih tačaka iz uzorka i predstavlja kvadrat Euklidskog rastojanja vektora $\mathbf{x}_r$ u odnosu na sredinu, u oznaci d_r^2 , data sledećim izrazom

$$d_r^2 = (\mathbf{x}_r - \bar{\mathbf{x}})'(\mathbf{x}_r - \bar{\mathbf{x}}) = \sum_{j=1}^p (x_{rj} - \bar{x}_j)^2$$

Kada smo orijentisani ka vrstama u matrici podataka, onda nam je osnovna veličina u analizi rastojanje. U statističkim analizama Euklidsko rastojanje ne daje najbolje rezultate. Zbog toga posmatramo statistička rastojanja odnosno rastojanja koja uzimaju u obzir varijansu i kovarijansu promenljivih.

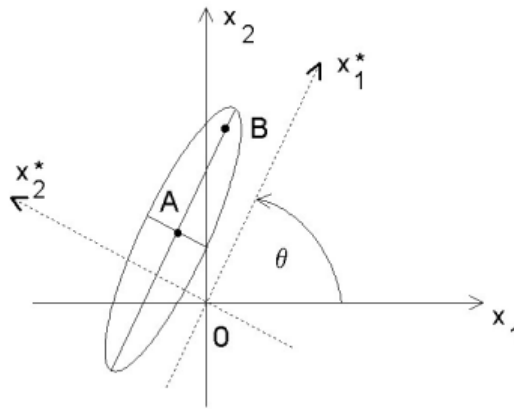
Ideju konstruisanja statističkog rastojanja ćemo izložiti pomoću slike 1.1. Na slici je elipsom predstavljen skup podataka. Pozitivna orijentisanost skupa u odnosu na osu x_1 tj. u odnosu na ugao θ , ukazuje na pozitivnu korelisanost promenljivih. Prilikom određivanja rastojanja, koordinatni sistem ćemo rotirati upravo za ugao θ . Zatim ćemo koordinatni početak tog rotiranog sistema pomeriti da se poklopi sa težištem skupa odnosno sa tačkom A, tim postizemo centriranje opservacija u novom koordinatnom sistemu. Posmatrajući skup tačaka u odnosu na novi koordinatni početak $x_1^* A x_2^*$, vidimo da svojim oblikom ukazuje na nekorelisanost promenljivih. Na ovaj način smo rotacijom eliminisali korelaciju promenljivih. Ono što treba da uzmemo u obzir je različit varijabilitet promenljivih. Statističko rastojanje ma koje tačke uzorka od težišta skupa tačaka predstavlja Euklidsko rastojanje na osnovu standardizovanih koordinata tačaka u novom koordinatnom sistemu. Može se pokazati da kvadrat rastojanja preko originalnih koordinata ima formu u matričnom obliku

$$d^2 = [x_1 - \bar{x}_1 \ x_2 - \bar{x}_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \end{bmatrix}$$

gde su koeficijenti a_{11}, a_{12}, a_{22} takvi da je d nenegativno za bilo koji par vrednosti x_1 i x_2 . Ovi koeficijenti zavise od ugla θ i kovarijansne matrice S_{11}, S_{12}, S_{22} . Sve tačke čije je rastojanje od težišta skupa tačaka konstantno zadovoljavaju ovu prethodnu jednakost. Ona predstavlja jednačinu elipse sa centrom u tački A. Ako na mesto matrice koeficijenata $[a_{ij}]$ stavimo inverznu uzoračku kovarijansnu matricu, tada definišemo Mahalanobisovo rastojanje (eng. Mahalanobis distance)[20]. Kvadrat Mahalanobisovog rastojanja, odnosno uopštenje standardnog Euklidskog rastojanja, dat je sledećim izrazom

$$d^2 = [x_1 - \bar{x}_1 \ x_2 - \bar{x}_2] \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \end{bmatrix}$$

Ovaj izraz sa jedne strane predstavlja jednačinu elipse, a sa druge strane predstavlja kvadrat Mahalanobisovog rastojanja proizvoljne tačke (x_1, x_2) od centra $(\bar{x}_1, \bar{x}_2)$.



Slika 1.1: *Dijagram rastojanja - rotacija koordinantnog sistema [17]*

Do sada smo razmatrali dvodimenzionalni slučaj, pa ćemo sad dati uopštenje za p -dimenzionalni slučaj. Za tačku $\mathbf{x}' = [x_1, x_2, \dots, x_p]$ definišemo njeno Mahalanobisovo rastojanje u odnosu na sredinu $\bar{\mathbf{x}}' = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p]$ dato sa d sledećim izrazom

$$d^2 = (\mathbf{x} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$$

Ovim izrazom je istovremeno opisan elipsoid u p -dimenzionalnom prostoru za fiksnu vrednost Mahalanobisovog rastojanja.[17]

1.3 Višedimenzionalna normalna raspodela

Jednodimenzionalna normalna (Gausova) raspodela ima najveći značaj među raspodelama i od posebnog je značaja u statističkom zaključivanju. Mnoge metode zaključivanja polaze od pretpostavke da je raspodela normalna, a to omogućava jednostavnije određivanje osobina i raspodela ocena nepoznatih parametara ili test statistika, jer je u potpunosti određena sa dva parametra. Takođe, predstavlja dobru aproksimaciju raspodela uzoračkih statistika, bez obzira na raspodelu populacije odakle je uzet uzorak.

Ulogu i značaj koji ima jednodimenzionalna normalna raspodela u analizi slučajne promenljive, ima i višedimenzionalna normalna raspodela kod slučajnog vektora. Pretpostavka koja nam omogućava definisanje postupaka statističkog zaključivanja je višedimenzionalna normalna raspodela.

Slučajna promenljiva X ima normalnu $\mathcal{N}(\mu, \sigma^2)$ raspodelu, $\mu \in \mathbb{R}$ i $\sigma > 0$, ako je njena gustina raspodele

$$\varphi_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

gde μ predstavlja sredinu, a σ^2 varijansu slučajne promenljive X .

Izraz u eksponatu možemo zapisati na sledeći način

$$-\frac{(x-\mu)^2}{2\sigma^2} = -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 = -\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)$$

i tada kvadrat odstupanja $\mathbf{x}$ od $\boldsymbol{\mu}$ u jedinicama standardne devijacije u p -dimenzionalnom slučaju pišemo kao

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

odnosno reč je o uopštenom odstojanju od $\mathbf{x}$ od $\boldsymbol{\mu}$, gde smo sa $\boldsymbol{\mu}$ označili vektor očekivane vrednosti slučajnog vektora $\mathbf{X}$, a sa $\boldsymbol{\Sigma}$ njegovu kovarijansnu matricu.

Funkcija gustine višedimenzionalne normalne slučajne promenljive $\mathbf{X}' = [X_1, X_2, \dots, X_p]$ data je sa

$$\varphi_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}, \quad x_j \in \mathbb{R}, \quad j = 1, \dots, p$$

i označavamo je sa $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Funkcija gustine će biti konstantna ako je $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$, c je konstanta, a navedeni izraz predstavlja kvadrat Mahalanobisovog rastojanja. Ovaj izraz nam kaže da smo funkciju gustine presekli na visini c^2 . Projekcija tog preseka u ravni je elipsa.

U opštem slučaju dobijamo da je kvadratom Mahalanobisovog rastojanja između $\boldsymbol{\mu}$ i $\mathbf{x}$ definisan omotač elipsoida čiji je centar u $\boldsymbol{\mu}$. Funkcija gustine slučajne promenljive $\mathbf{X}$ je konstantna po omotaču elipsoida. Taj omotač nazivamo kontura konstantne gustine verovatnoće.

Pravac prve ose elipsoida određen je normalizovanim karakterističnim vektorom, koji je pridružen najvećem karakterističnom korenu kovarijansne matrice. Pravac druge ose elipsoida određen je karakterističnim vektorom koji je pridružen drugom po veličini karakterističnom korenu matrice $\boldsymbol{\Sigma}$. Ukoliko su svi karakteristični koreni matrice različiti i pozitivni, položaj osa elipsoida je jednoznačno određen, a ose su ortogonalne međusobno. Ako su dva karakteristična korena jednaka, tada je presek elipsoida u ravni koja je generisana odgovarajućim

karakterističnim vektorima, kružnica. U tom slučaju položaj osa nije jednoznačno određen, jer tad postoji beskonačno mnogo položaja dve međusobno ortogonalne ose.

Pokazano je da ako slučajna promenljiva $\mathbf{X}$ ima normalnu raspodelu odnosno $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ onda $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ ima χ^2 raspodelu sa p stepeni slobode.

Verovatnoća da tačka $\mathbf{X}$ pripada elipsoidu je

$$P(\{\mathbf{X} : (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq \chi_{p;\alpha}^2\}) = 1 - \alpha$$

gde smo sa $\chi_{p;\alpha}^2$ označili gornji 100α percentil χ^2 raspodele sa p stepeni slobode. Ovaj rezultat je našao primenu kod provere normalnosti raspodele.[17]

1.4 Uzorak iz višedimenzionalne normalne raspodele

U prethodnom delu smo pretpostavili da su nam parametri višedimenzionalne normalne raspodele poznati. U stvarnosti to nije slučaj, te je potrebno na osnovu uzorka oceniti parametre višedimenzionalne normalne raspodele.

Pretpostavimo da je $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ slučajan uzorak sa n slučajnih promenljivih $\mathbf{X}$ sa p -dimenzionalnom raspodelom čije su sredina $\boldsymbol{\mu}$ i kovarijansna matrica $\boldsymbol{\Sigma}$ nepoznate. To znači da elementi slučajnog uzorka imaju jednaku raspodelu tj. $\mathbf{X}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ i da su međusobno nezavisni. Na osnovu definicije slučajnog uzorka njegova zajednička raspodela je jednaka proizvodu marginalnih raspodela. Ova funkcija se za fiksni uzorak naziva funkcija verodostojnosti za uzorak i predstavlja funkciju nepoznatih parametara.

Definicija 1.13. Funkcija verodostojnosti za prost slučajni uzorak $(X_1, X_2, \dots, X_n)$ na osnovu realizovanog uzorka $(x_1, x_2, \dots, x_n)$ obima n je

$$L(\theta, x_1, x_2, \dots, x_n) = \varphi(x_1, \theta) \varphi(x_2, \theta) \cdots \varphi(x_n, \theta)$$

za neprekidnu raspodelu, gde je θ nepoznati parametar.

Ako obeležje ima normalnu raspodelu sa parametrima $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ onda je funkcija verodostojnosti

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} e^{-\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}$$

Definicija 1.14. Neka je $L(\theta) = L(\theta, x_1, \dots, x_n)$ funkcija verodostojnosti za prost slučajni uzorak $(X_1, \dots, X_n)$ na osnovu realizovanog uzorka $(x_1, \dots, x_n)$. Ako $\hat{\theta} = u(x_1, \dots, x_n)$ maksimizira $L(\theta)$, onda je statistika $\hat{\theta} = u(X_1, \dots, X_n)$ ocena nepoznatog parametra θ dobijena metodom maksimalne verodostojnosti.

Maksimiziranjem funkcije verodostojnosti $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ po $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ dobijamo ocenu najveće verodostojnosti nepoznatih parametara. Tada nam važi da je ocena najveće verodostojnosti sredine

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$$

a ocena najveće verodostojnosti kovarijansne matrice

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})^2 = \bar{\mathbf{S}}^2 = \frac{n-1}{n} \mathbf{S}$$

Teorema 1.3. Neka je $(X_1, X_2, \dots, X_n)$ prost slučajan uzorak za obeležje X sa sredinom $E(X) = \mu$ i varijansom $Var(X) = \sigma^2$. Tada važi da je

$$E(\bar{X}_n) = E(X), \quad Var(\bar{X}_n) = \frac{Var(X)}{n}$$

Teorema 1.4. Neka je $(X_1, X_2, \dots, X_n)$ prost slučajan uzorak za obeležje X sa normalnom $N(\mu, \sigma^2)$ raspodelom. Tada

- statistika $\bar{X}_n$ ima $N(\mu, \frac{\sigma^2}{n})$ raspodelu, a statistika $\frac{\bar{X}_n - \mu}{\sigma} \sqrt{n}$ ima standardizovanu normalnu raspodelu tj. $N(0, 1)$
- statistika $\frac{n\hat{S}_n^2}{\sigma^2}$ ima χ_n^2 raspodelu, a statistika $\frac{n\bar{S}_n^2}{\sigma^2}$ ima χ_{n-1}^2 raspodelu, gde su $\hat{S}_n^2$ i $\bar{S}_n^2$ uzoračke varijanse, ako je poznata i ako nije poznata sredina, respektivno.

Na osnovu ovih teorema i činjenice da je raspodela linearne kombinacije normalno raspodeljenih slučajnih promenljivih normalna, važi da je sredina slučajnog uzorka obima n uzetog iz višedimenzionalne normalne raspodele, takođe, normalno raspodeljena. Imamo da je $\mu = \bar{X} \sim N_p(\mu, \frac{\Sigma}{n})$.

Znamo da suma kvadrata nezavisnih slučajnih promenljivih sa standardizovanom normalnom raspodelom ima χ^2 raspodelu. Na osnovu prethodne teoreme znamo da važi $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$. Analogno, u višedimenzionalnom slučaju imamo na osnovu matrice podataka $\mathbf{X}$ dimenzije $n \times p$ da slučajna matrica $\mathbf{X}'\mathbf{X}$ ima p -dimenzionalnu Višartovu raspodelu (eng. Wishart distribution) u oznaci $W_p(\Sigma, n)$, gde n predstavlja broj stepeni slobode.

Višartova raspodela predstavlja uopštenje χ^2 raspodele, jer se za $p = 1$ i $\Sigma = 1$ svodi na χ^2 raspodelu. Na osnovu uzoračke kovarijanske matrice $\mathbf{S}$ statistika $(n-1)\mathbf{S}$ ima Višartovu raspodelu sa $n-1$ stepeni slobode, u oznaci $(n-1)\mathbf{X} \sim W_p(\Sigma, n-1)$, a koristimo je u statističkom zaključivanju.

Sada navodimo dve osobine Višartove raspodele. Ako su matrice A_1 i A_2 međusobno nezavisne i imaju Višartovu raspodelu odnosno $A_i \sim W_p(\Sigma, n_i)$, $i = 1, 2$, onda i matrica zbira ima Višartovu raspodelu odnosno $A_1 + A_2 \sim W_p(\Sigma, n_1 + n_2)$. Prema drugoj osobini, ako matrica A ima Višartovu raspodelu odnosno $A \sim W_p(\Sigma, n)$, onda i matrica BAB' ima Višartovu raspodelu odnosno $BAB' \sim W_p(B\Sigma B', n)$ gde je B nesingularna matrica dimenzija $p \times p$. [29]

1.5 Zaključivanje na osnovu sredine

Na osnovu slučajnog uzorka uzetog iz populacije koja ima normalnu raspodelu definisali smo brojnu ocenu nepoznatog parametra, sredinu μ . Umesto da izvodimo zaključak o stvarnoj vrednosti parametra, možemo pokušati da zaključimo u kom se intervalu nalazi stvarna vrednost parametra. Definisaćemo intervalnu ocenu sredine μ i pri tome razlikovati slučaj kad nam je varijansa populacije poznata i slučaj kad varijansa populacije nije poznata. [19]

1.5.1 Interval poverenja

Definicija 1.15. Neka je dat prost slučajan uzorak $(X_1, X_2, \dots, X_n)$ za obeležje X i neka su $U_1 = u_1(X_1, X_2, \dots, X_n)$ i $U_2 = u_2(X_1, X_2, \dots, X_n)$ dve statistike takve da važi

$$P\{U_1 \leq \theta \leq U_2\} = \alpha$$

gde α ne zavisi od θ . Tada se interval (U_1, U_2) naziva interval poverenja za parametar θ na nivou poverenja α ili $100\alpha\%$ interval poverenja za θ .

Jedna od statistika U_1, U_2 može biti konstantna i tada govorimo o jednostranom intervalu poverenja. Uobičajena vrednost za α je 0.9, 0.95, 0.99.

Pretpostavimo da obeležje X ima normalnu $\mathcal{N}(\mu, \sigma^2)$ raspodelu. Određujemo interval poverenja za sredinu ako je poznata varijansa σ^2 populacije.

Kako slučajna promenljiva X ima normalnu raspodelu, onda slučajna promenljiva $\frac{\bar{X}_n - \mu}{\sigma} \sqrt{n}$ ima normalnu $\mathcal{N}(0, 1)$ raspodelu pa se za uzorak obima n i nivo poverenja α može odrediti broj $z_{\alpha/2}$ tako da

$$P\left\{\frac{|\bar{X}_n - \mu|}{\sigma} \sqrt{n} \leq z_{\alpha/2}\right\} = \alpha$$

Broj $z_{\alpha/2}$ predstavlja gornji $100\frac{\alpha}{2}$ percentil standardizovane normalne raspodele.

Dalje, sledi da je

$$P\left\{\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right\} = \alpha$$

Odavde dobijamo da je $100(1 - \alpha)\%$ interval poverenja sredine

$$\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}}\right)$$

U slučaju kad nam varijansa σ^2 populacije nije poznata, interval poverenja određujemo na sledeći način.

Tada nam slučajna promenljiva $\frac{\bar{X}_n - \mu}{\bar{S}_n} \sqrt{n-1}$ ima Studentovu t_{n-1} raspodelu, pa se $t_{n-1;\alpha/2}$ može odrediti tako da važi

$$P\left\{\frac{|\bar{X}_n - \mu|}{\bar{S}_n} \sqrt{n-1} \leq t_{n-1;\alpha/2}\right\} = \alpha$$

Broj $t_{n-1;\alpha/2}$ predstavlja gornji $100\frac{\alpha}{2}$ percentil Studentove raspodele sa $n - 1$ stepeni slobode. Odavde dobijamo da je $100(1 - \alpha)\%$ interval poverenja za sredinu slučajne promenljive

$$\left(\bar{X}_n - t_{n-1;\alpha/2} \frac{\bar{S}_n}{\sqrt{n-1}}, \bar{X}_n + t_{n-1;\alpha/2} \frac{\bar{S}_n}{\sqrt{n-1}}\right)$$

Interval poverenja se može odrediti i ako koristimo popravljenu uzoračku varijansu S^2 . U ovom slučaju promenljiva $\frac{\bar{X}_n - \mu}{S} \sqrt{n}$ ima Studentovu t_{n-1} raspodelu sa $n - 1$ stepeni slobode i dobijamo da je $100(1 - \alpha)\%$ interval poverenja za sredinu slučajne promenljive

$$\left(\bar{X}_n - t_{n-1;\alpha/2} \frac{S}{\sqrt{n}}, \bar{X}_n + t_{n-1;\alpha/2} \frac{S}{\sqrt{n}}\right)$$

I u ovom slučaju broj $t_{n-1;\alpha/2}$ predstavlja gornji $100\frac{\alpha}{2}$ percentil Studentove raspodele sa $n - 1$ stepeni slobode.

Pre nego što smo uzeli uzorak, interval poverenja predstavlja slučajan interval jer mu granice zavise od slučajnih promenljivih $\bar{X}_n$ i S . Verovatnoća da taj interval sadrži stvarnu vrednost sredine iznosi $1 - \alpha$, što znači da od velikog broja takvih nezavisnih intervala njih $100(1 - \alpha)\%$ sadržeće stvarnu vrednost sredine.

1.5.2 Testiranje hipoteza

Prilikom testiranja hipoteza koristi se test statistika $T = f(X_1, X_2, \dots, X_n)$. Statistički test se sprovodi tako što se odredi odgovarajuća kritična oblast C vezana za statistiku T i odluka se donosi na sledeći način

- ako registrovana vrednost statistike T ne pripada kritičnoj oblasti C nemamo razloga da odbacimo osnovnu hipotezu
- ako registrovana vrednost statistike T pripada kritičnoj oblasti C osnovna hipoteza se odbacuje

Statistički test se može sprovesti korišćenjem p-vrednosti, gde je p-vrednost veličina kritične oblasti čija je granica registrovana vrednost t_{reg} test statistike. Ako je oblast odbacivanja osnovne hipoteze oblika $\{T > c\}$ tada se verovatnoća $p = P_{H_0}\{T > t_{reg}\}$ naziva p-vrednost. Odluka se donosi na sledeći način

- ako je $p < \alpha$, H_0 se odbacuje
- ako je $p > \alpha$, H_0 se ne odbacuje

Veliki broj testova se zasniva na pretpostavci da obeležje koje posmatramo ima normalnu $\mathcal{N}(\mu, \sigma^2)$ raspodelu. Pod tom pretpostavkom moguće je odrediti raspodelu test statistike i na osnovu te raspodele izvesti zaključke.

Kada nam je poznata varijansa σ^2 populacije za testiranje osnovne hipoteze $H_0 : \mu = \mu_0$ protiv alternativne hipoteze $H_1 : \mu \neq \mu_0$ koristimo statistiku $\bar{X}_n$ koja ima normalnu $\mathcal{N}(\mu_0, \frac{\sigma^2}{n})$ ili statistiku $\frac{\bar{X}_n - \mu_0}{\sigma} \sqrt{n}$ koja ima normalnu $\mathcal{N}(0, 1)$ raspodelu.

Na osnovu uzetog uzorka prihvatamo osnovnu hipotezu ako je apsolutna vrednost test statistike manja od kritične vrednosti $z_{\alpha/2}$, gde je α nivo značajnosti testa, u suprotnom je odbacujemo.

Kada nam je nepoznata varijansa σ^2 populacije za testiranje osnovne hipoteze $H_0 : \mu = \mu_0$ protiv alternativne hipoteze $H_1 : \mu \neq \mu_0$ koristimo statistiku $\frac{\bar{X}_n - \mu_0}{S} \sqrt{n}$ koja ima Studentovu t_{n-1} raspodelu sa $n - 1$ stepeni slobode.

Na osnovu uzetog uzorka prihvatamo osnovnu hipotezu ako je apsolutna vrednost test statistike manja od kritične vrednosti $t_{n-1; \alpha/2}$, gde je α nivo značajnosti testa, u suprotnom je odbacujemo. Alternativan način provere gornje hipoteze zasnovan je na izračunatom intervalu poverenja sredine. Ukoliko je realizovana vrednost sredine unutar izračunatog intervala poverenja, tada prihvatamo nultu hipotezu na izabranom nivou značajnosti, a u suprotnom je odbacujemo.

1.6 Zaključivanje na osnovu kovarijanske i korelacione matrice

Neke pretpostavke za primenu analize su bazirane na osobinama kovarijanske i korelacione matrice. Npr. dijagonalnost kovarijanske matrice nam sugeriše nezavisnost višedimenzionalnih slučajnih promenljivih sa normalnom raspodelom.

Sada razmatramo testiranje nulte hipoteze $H_0 : \Sigma = \sigma^2 \mathbf{I}$ protiv alternativne $H_1 : \Sigma \neq \sigma^2 \mathbf{I}$, gde je $\mathbf{I}$ jedinična matrica. Nulta hipoteza nam kaže da su vandijagonalni elementi kovarijanske matrice jednaki nuli, što znači da su p -dimenzionalne normalno raspodeljene slučajne promenljive slučajnog vektora X međusobno nezavisne. A njihove varijanse su jednake međusobno i jednake su zajedničkoj varijansi σ^2 . Sa ovakvom strukturom kovarijanske matrice, elipsoid konstantne gustine verovatnoće $(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = d^2$ svodi se na $(\mathbf{x} - \boldsymbol{\mu})' (\mathbf{x} - \boldsymbol{\mu}) = \sigma^2 d^2$, što u $p = 2$ predstavlja jednačinu kružnice, a u $p = 3$ jednačinu lopte, u opštem slučaju jednačinu sfere sa centrom u tački $\boldsymbol{\mu}$. Zbog toga osnovu (nultu) hipotezu nazivamo hipoteza o sferičnosti.

Konstruišući test na principu količnika verodostojnosti, sledi da je test statistika za testiranje nulte hipoteze $H_0 : \Sigma = \sigma^2 \mathbf{I}$ data na sledeći način

$$np \ln \frac{\frac{1}{p} \sum_{j=1}^p \lambda_j}{\left(\prod_{j=1}^p \lambda_j \right)^{\frac{1}{p}}}$$

Ovaj razlomak nam predstavlja količnik aritmetičke i geometrijske sredine karakterističnih korena $\lambda_1, \lambda_2, \dots, \lambda_p$ uzoračke kovarijanske matrice $\mathbf{S}$. Ako je tačna nulta hipoteza, ova test statistika ima χ^2 asimptotsku raspodelu sa $\frac{1}{2}(p+2)(p-1)$ stepeni slobode. Nulta hipoteza se prihvata na izabranom nivou značajnosti ako je izračunata vrednost statistike manja od kritične vrednosti.

Korišćenjem principa količnika verodostojnosti možemo da konstruišemo test za testiranje hipoteze o strukturi korelacione matrice populacije.

Nulta hipoteza $H_0 : \boldsymbol{\rho} = \mathbf{I}$ protiv alternativne hipoteze $H_1 : \boldsymbol{\rho} \neq \mathbf{I}$ nam sugeriše da su p -dimenzionalne normalno raspodeljene slučajne promenljive slučajnog vektora X međusobno nezavisne. Količnik verodostojnosti se svodi na $|\mathbf{R}|^{\frac{n}{2}}$, gde je $\mathbf{R}$ uzoračka korelaciona matrica. Umesto uobičajene transformacije ovog količnika verodostojnosti, koristimo Bartletovu test statistiku (eng. Bartlett's test) koja bolje aproksimira χ^2 raspodelu

$$-\left(n - 1 - \frac{2p+5}{6}\right) \ln |\mathbf{R}|$$

odnosno

$$-\left(n - 1 - \frac{2p+5}{6}\right) \sum_{j=1}^p \ln \lambda_j$$

gde su λ_j karakteristični koreni uzoračke korelacione matrice. Test statistika ima asimptotsku χ^2 raspodelu sa $\frac{1}{2}p(p-1)$ stepeni slobode. Nulta hipoteza se prihvata na izabranom nivou značajnosti ako je izračunata vrednost statistike manja od kritične vrednosti.[17]

2 Analiza glavnih komponenti

2.1 Uvod

Multivarijantna analiza je poslednjih godina našla veliku primenu u skoro svim naučnim oblastima. U procesu naučnog objašnjenja prirode nekog fenomena polaznu osnovu čine podaci koji se odnose na jedan ili više skupova objekata. Ti objekti mogu biti: pojedinci, ljudske zajednice, prirodne pojave, predmeti, pojave koje su proizvod ljudske delatnosti. Dešava se da ne umemo da sagledamo kompleksnu prirodu objekata u potpunosti. Imamo mogućnost obuhvatanja različitih karakteristika jedne višedimenzionalne pojave. Te karakteristike, obeležja ili kako ih još zovemo promenljive su predmet našeg merenja. Ispitujemo prirodu objekata istovremenim merenjem većeg broja promenljivih na svakoj jedinici posmatranja iz jednog ili više skupova objekata.

Multivarijantna analiza predstavlja skup statističkih metoda koje simultantno analiziraju višedimenzionalna merenja dobijena za svaku jedinicu posmatranje iz skupa objekata koji ispituje.

Klasifikacije metoda multivarijantne analize zasnovane su na različitim kriterijumima. Prema jednoj od klasifikacija metode delimo u dve grupe: metode zavisnosti i metode međuzavisnosti. Ukoliko u istraživanju imamo jedan skup koji predstavlja zavisne promenljive, a drugi skup predstavlja nezavisne promenljive onda koristimo metode zavisnosti. Ako skup svih promenljivih ne možemo da podelimo na zavisne i nezavisne, tada koristimo metode međuzavisnosti. Kod metoda zavisnosti pokušavamo da objasnimo jednu ili više zavisnih promenljivih pomoću skupa nezavisnih promenljivih. Metodi međuzavisnosti nisu prediktivni, njima pokušavamo da pojednostavimo kompleksnu strukturu podata i to redukcijom podataka.

Metode zavisnosti su: regresija, kanonička korelaciona analiza, diskriminaciona analiza, analiza varijanse, logit analiza. Metode međuzavisnosti su: analiza glavnih komponenti, faktorska analiza, višedimenzionalno proporcionalno prikazivanje, loglinearni modeli.

Analiza glavnih komponenti (eng.principal component analysis - PCA) je jedna od najjednostavnijih multivarijantnih tehnika.

Ovu tehniku je prvi put opisao Karl Pearson 1901.godine po analogiji sa glavnim osama u mehanici[21], ali je opis izračunavanja i naziv pod kojim je metod danas poznat dao Harold Hotelling 1933.godine. Izračunavanja su bila previše komplikovana kad je trebalo napraviti analizu sa više promenljivih. Široka primena ove metode je počela sa pojavom računara koji su omogućili da ostvari svoj potencijal za rešavanje problema sa velikim brojem promenljivih.

Analiza glavnih komponenti predstavlja statističku analizu redukcije dimenzionalnosti skupa podataka na način da bude obuhvaćena što veća količina varijanse podataka. Primenjuje se kad je veliki broj promenljivih u skupu redundantan odnosno kada se više promenljivih odnosi na istu dimenziju a da ne pružaju neke dodatne informacije koje nisu već sadržane u nekoj drugoj promenljivoj.

Zadatak metode glavnih komponenti je određivanje linearnih kombinacija originalnih promenljivih koje će imati maksimalnu varijansu i biti međusobno nekorelisane, pri tom gubeći u najmanjoj mogućoj meri informacije sadržane u originalnom skupu podataka.

Originalne promenljive transformišemo u nove promenljive odnosno linearne kombinacije, koje nazivamo glavne komponente (eng.principal component).

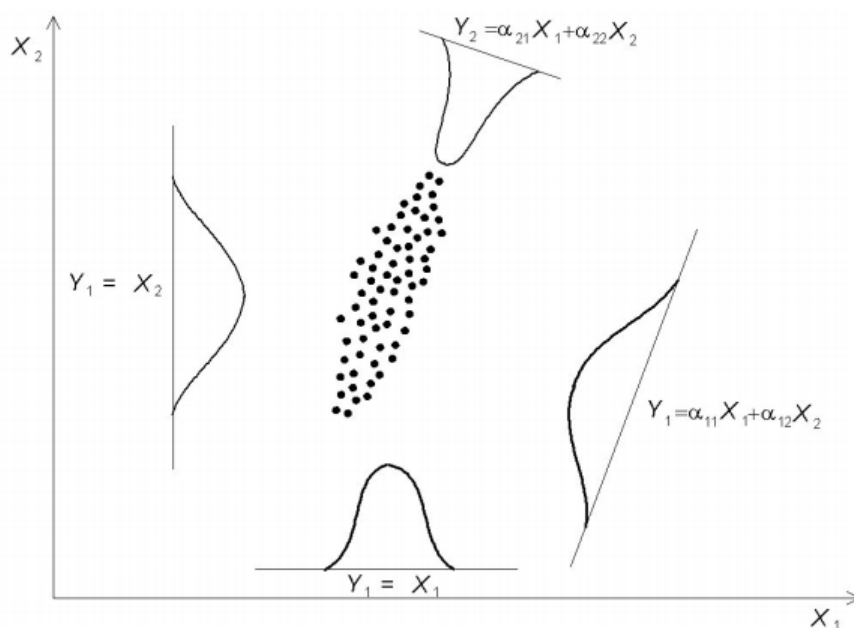
Prva glavna komponenta je konstruisana da obuhvata najveći deo varijanse originalnog skupa podataka, a naredne komponente obuhvataju onaj deo varijanse koji nije obuhvaćen prethodno izdvojenim komponentama.

Postupkom se originalni korelisani skup podataka transformiše u skup nekorelisanih promenljivih sa opadajućim vrednostima varijanse. Želimo da varijansa većeg broja glavnih komponenti bude zanemarljivo mala i da se kovarijansna struktura originalnog skupa podataka može aproksimirati sa svega nekoliko komponenti.

Ovim postizemo izvestan stepen uštede jer se broj promenljivih u daljoj analizi značajno smanjuje, a samim tim lakše generišemo hipoteze o nekom proučavanom fenomenu, te bolje razumemo i lakše interpretiramo njegovu strukturu.

Analiza glavnih komponenti ne postiže uvek smanjenje velikog broja originalnih promenljivih na mali broj komponenti. Ako originalne promenljive nisu u korelaciji analiza neće dati dobar rezultat, a najbolji rezultat je moguće postići kada su originalne promenljive u visokoj korelaciji bilo pozitivnoj ili negativnoj. Ukoliko su promenljive u visokoj korelaciji može se očekivati da će skup od na primer 20 originalnih promenljivih biti redukovano na svega dve ili tri komponente odnosno linearne kombinacije, samim tim je otkriven visok stepen redundantnosti kod originalnih promenljivih.

Sledećim primerom se pokazuje osnovna ideja metode kao i način konstruisanja glavnih komponenti. Na slici 2.1 prikazan je dijagram rasturanja promenljivih X_1 i X_2 . Linearna kombinacija ove dve promenljive je $Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2$, gde je sa α_{1i} označen koeficijent linearne kombinacije uz i -tu promenljivu u prvoj linearnoj kombinaciji. Ako izaberemo da su $\alpha_{11} = 1$ i $\alpha_{12} = 0$ odnosno $\alpha_{11} = 0$ i $\alpha_{12} = 1$ dobijamo da je prva linearna kombinacija jednaka prvoj promenljivoj odnosno drugoj promenljivoj, respektivno. Ovakvim izborom, geometrijski gledano, dobijamo promenljivu Y_1 čija je raspodela formirana na osnovu projekcije skupa tačaka na X_1 i X_2 osu. Raspodela promenljive Y_1 za prvi izbor koeficijenata predstavlja marginalnu raspodelu promenljive X_1 , a za drugi izbor marginalnu raspodelu promenljive X_2 . Ako je potrebno dvodimenzionalni skup reprezentovati jednom promenljivom biramo onu koja ima veću varijansu, jer tad možemo u većoj meri razlikovati pojedinačne opservacije dvodimenzionalnog skupa.



Slika 2.1: Projekcija skupa tačaka [17]

Na slici 2.1 vidimo da je promenljiva X_2 bolji kandidat za reprezentovanje dvodimenzionalnog skupa podataka, jer ima veću varijansu od promenljive X_1 . Prirodno se nameće

pitanje postoji li takav izbor koeficijenata linearne kombinacije koji će dati veću varijansu promenljive Y_1 . Izbor koeficijenata predstavlja maksimiziranje varijanse linearne kombinacije uz uslov da je zbir kvadrata koeficijenata linearne kombinacije jednak jedinici. Geometrijski ovaj uslov znači da je vektor koeficijenata linearne kombinacije jedinične dužine. Uslov uvodimo da bismo postigli jednoznačnu definisanost linearne kombinacije. Geometrijski gledano izborom koeficijenata menjamo ugao pod kojim projektujemo tačke iz skupa tačaka na pravu liniju. Biramo koeficijente koji će nam dati projekciju tačaka sa najvećom varijansom. Na slici 2.1 najveću varijansu, od svih linearnih kombinacija koje se mogu dobiti promenom ugla projekcije, ima linearna kombinacija $Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2$. Tu linearnu kombinaciju nazivamo prva glavna komponenta.

Ukoliko je za našu analizu potrebno, možemo da formiramo narednu linearnu kombinaciju $Y_2 = \alpha_{21}X_1 + \alpha_{22}X_2$. Njene koeficijente određujemo isto kao koeficijente prve komponente uz dodatni uslov da su Y_1 i Y_2 međusobno nekorelisane. Ovaj uslov, geometrijski posmatrano, znači da prave na koje se projektuje skup tačaka kod prve i druge linearne kombinacije treba da budu normalne². Linearna kombinacija $Y_2 = \alpha_{21}X_1 + \alpha_{22}X_2$ je druga glavna komponenta. Na slici vidimo da su prva i druga komponenta paralelne sa glavnim i sporednom osom elipse koja se može nacrtati oko skupa tačaka odnosno elipse konstantne gustine verovatnoće.[17]

2.2 Definicija glavnih komponenti

U ovom delu se bavimo teorijskim modelom glavnih komponenti. U literaturi se još naziva populacioni model glavnih komponenti.

Neka je $\mathbf{x}$ vektor koji sadrži p slučajnih promenljivih i posmatramo strukturu varijanse ovih promenljivih te njihovu kovarijansu ili korelaciju. Komplikovano je posmatrati svih p varijansi i $\frac{p(p-1)}{2}$ kovarijansi, osim u slučaju kad je p mali broj. Lakši pristup je naći nekoliko izvedenih promenljivih koje očuvavaju najveći deo informacija sadržan u tim varijansama i kovarijansama.

U nastavku je prikazan postupak kako se dobijaju glavne komponente uz pomoć kovarijanske matrice Σ .

Neka je $y_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + \dots + \alpha_{1p}x_p = \alpha_1'\mathbf{x}$ linearna kombinacija elemenata slučajnog vektora $\mathbf{x}$, gde su $\alpha_{11}, \alpha_{12}, \dots, \alpha_{1p}$ koeficijenti linearne kombinacije. Naš zadatak je da odredimo vektor koeficijenata α_1 tako da se maksimizira varijansa od y_1 . Iz prethodnog dela znamo da važi $Var(y_1) = Var(\alpha_1'\mathbf{x}) = \alpha_1'\Sigma\alpha_1$. Kako se $Var(y_1) = \alpha_1'\Sigma\alpha_1$ može proizvoljno povećati množenjem vektora α_1 proizvoljnim skalarom, uvodi se ograničenje da je vektor koeficijenata jedinične dužine odnosno da važi $\alpha_1'\alpha_1 = 1$.

Problem maksimizacije $\alpha_1'\Sigma\alpha_1$ uz ograničenje $\alpha_1'\alpha_1 = 1$ rešavamo korišćenjem metoda Lagranžovih množitelja. Prvo formiramo Lagranžovu funkciju

$$\mathcal{L}(\alpha_1; \lambda) = \alpha_1'\Sigma\alpha_1 - \lambda(\alpha_1'\alpha_1 - 1) \quad (2.1)$$

gde λ predstavlja Lagranžov množitelj. Zatim diferenciramo Lagranžovu funkciju po koeficijentu α_1 i dobijeni izraz izjednačimo sa nulom, pa dobijamo

$$\Sigma\alpha_1 - \lambda\alpha_1 = 0 \quad (2.2)$$

²Kad odredimo prvu glavnu komponentu, druga treba da bude pod uglom od 90°.

odnosno

$$(\Sigma - \lambda I)\alpha_1 = 0 \quad (2.3)$$

gde I predstavlja jediničnu matricu. Da bismo dobili netrivialno rešenje za α_1 determinanta $|\Sigma - \lambda I|$ mora biti jednaka nuli, a to dalje znači da λ mora biti jedan od karakterističnih korena kovarijanske matrice Σ . Ako izraz 2.2 pomnožimo sa leve strane sa α_1' dobijamo

$$\alpha_1'\Sigma\alpha_1 - \lambda\alpha_1'\alpha_1 = 0 \quad (2.4)$$

odnosno

$$\alpha_1'\Sigma\alpha_1 = \lambda\alpha_1'\alpha_1 = \lambda$$

jer važi $\alpha_1'\alpha_1 = 1$, pa zbog ovog biramo jedan karakteristični koren. Pošto želimo da maksimiziramo varijansu za λ , biramo najveći karakteristični koren, recimo λ_1 . Na osnovu uslova 2.3 sledi da je α_1 odgovarajući karakteristični vektor pridružen karakterističnom korenu λ_1 . Njegovim normiranjem tj. uslovom da je $\alpha_1'\alpha_1 = 1$ dobijamo traženi vektor α_1 .

Ako treba da odredimo više od jedne linearne kombinacije, postupamo kao u slučaju određivanja prve glave komponente, uz dodatni uslov da kovarijansa između prve i druge glavne komponente bude jednaka nuli. Neka je $y_2 = \alpha_{21}x_1 + \alpha_{22}x_2 + \dots + \alpha_{2p}x_p = \alpha_2'x$ linearna kombinacija gde su $\alpha_{21}, \alpha_{22}, \dots, \alpha_{2p}$ koeficijenti koje treba odrediti uz normirajući uslov $\alpha_2'\alpha_2 = 1$ i uz uslov nekorelisanosti prve i druge komponente odnosno uz uslov $\alpha_2'\alpha_1 = 0$. Dodatni uslov dobijamo na sledeći način

$$\text{Cov}(y_2, y_1) = \text{Cov}(\alpha_2'x, \alpha_1'x) = \alpha_2'\Sigma\alpha_1 = \alpha_1'\Sigma\alpha_2 = \alpha_2'\alpha_1\lambda_1 = \alpha_1'\alpha_2\lambda_1 \quad (2.5)$$

jer je $\Sigma\alpha_1 = \alpha_1\lambda_1$, pa je $\alpha_2'\alpha_1\lambda_1 = 0$ samo kad je $\alpha_2'\alpha_1 = 0$.

Svaka od jednačina

$$\alpha_2'\Sigma\alpha_1 = 0$$

$$\alpha_1'\Sigma\alpha_2 = 0$$

$$\alpha_2'\alpha_1 = 0$$

$$\alpha_1'\alpha_2 = 0$$

može se koristiti da se opiše nekorelisanost glavnih komponenti y_1 i y_2 , a dalje u radu proizvoljno biramo poslednju.

Sada formiramo Lagranžovu funkciju sa dva množitelja

$$\mathcal{L}(\alpha_1, \alpha_2; \lambda, \phi) = \alpha_2'\Sigma\alpha_2 - \lambda(\alpha_2'\alpha_2 - 1) - \phi\alpha_2'\alpha_1 \quad (2.6)$$

gde su λ i ϕ Lagranžovi množitelji. Diferenciramo po α_2 , a zatim dobijeni izraz izjednačimo sa nulom i dobijemo

$$\Sigma\alpha_2 - \lambda\alpha_2 - \phi\alpha_1 = 0 \quad (2.7)$$

Dobijeni izraz pomnožimo sa leve strane sa α_1' i dobijamo sledeće

$$\alpha_1'\Sigma\alpha_2 - \lambda\alpha_1'\alpha_2 - \phi\alpha_1'\alpha_1 = 0. \quad (2.8)$$

Poznato nam je da su prva dva člana u ovom izrazu jednaka nuli, a da je $\alpha_1' \alpha_1 = 1$, pa zbog toga sledi da je $\phi = 0$. Dalje, dobijamo da je $\Sigma \alpha_2 - \lambda \alpha_2 = 0$, tj. $\Sigma \alpha_2 = \lambda \alpha_2$, a to znači da je λ karakteristični koren matrice Σ , dok je α_2 odgovarajući karakteristični vektor. Za λ biramo što je moguće veću vrednost. Drugi po veličini karakteristični koren označavamo sa λ_2 , njemu odgovarajući karakteristični vektor je α_2 , a linearna kombinacija $y_2 = \alpha_2' x$ se naziva druga glavna komponenta.

Pod pretpostavkom da su svi karakteristični koreni matrice Σ međusobno različiti, λ ne može biti jednaka λ_1 , jer bi u tom slučaju važio da su $\alpha_1 = \alpha_2$, a to je kontradiktorno sa uslovom $\alpha_1' \alpha_2 = 0$.

Glavnih komponenti ima onoliko koliko ima karakterističnih korena i do njih se dolazi ponavljanjem već opisanog postupka. Ako su svi karakteristični koreni kovarijansne matrice Σ , koje smo poredali u opadajućem nizu tj. $\lambda_1 > \lambda_2 > \dots > \lambda_p \geq 0$, međusobno različiti, tada postoji p glavnih komponenti $y_1, y_2, \dots, y_p$ takvih da je $y_i = \alpha_i' x$, $i = 1, 2, \dots, p$. Vektori koeficijenata $\alpha_1, \alpha_2, \dots, \alpha_p$ predstavljaju karakteristične vektore matrice Σ koji su pridruženi karakterističnim korenima λ_i .

Treba napomenuti da su neki autori za vektore α_i koristili naziv glavne komponente, što može biti konfuzno. Termin glavne komponente je rezervisan za promenljive $\alpha_i' x$, a vektori α_i predstavljaju vektore koeficijenata.

2.3 Osobine glavnih komponenti

2.3.1 Algebarska i statistička svojstva

Definišući glavne komponente ukazali smo na njihove sledeće osobine koje slede direktno na osnovu definicije:

- $E(y_i) = 0$
- $Var(y_i) = \lambda_i$
- $Cov(y_i, y_j) = 0, i \neq j$
- $Var(y_1) > Var(y_2) > \dots > Var(y_p) \geq 0$

Neka je y vektor koji se sastoji od p glavnih komponenti y_i tj. $y' = [y_1, y_2, \dots, y_p]$. U matričnoj notaciji vektor y je

$$y = Ax \quad (2.9)$$

gde je A matrica dimenzija $p \times p$ čije su vrste karakteristični vektori kovarijansne matrice Σ odnosno $\alpha_1, \alpha_2, \dots, \alpha_p$, a x vektor originalnih slučajnih promenljivih.

Matrica A ima osobinu da je $A' = A^{-1}$ ³, pa se $y = Ax$ naziva ortogonalna transformacija ili rotacija, a matrica A ortogonalna matrica. Jedna od osobina matrice A je $|A| = \pm 1$.

Transformacija se naziva ortogonalna jer se vrši rotacija koordinatnih osa za izvestan ugao, pri čemu ose ostaju normalne jedna u odnosu na drugu, a ugao između ma koja dva vektora ostaje isti.

Jednakost $\Sigma \alpha_i = \lambda_i \alpha_i$, $i = 1, 2, \dots, p$ u matričnom zapisu glasi

$$\Sigma A = A \Lambda \quad (2.10)$$

³Na osnovu osobina karakterističnih vektora, jer važi $\alpha_i' \alpha_i = 1$ i $\alpha_i' \alpha_j = 0, i \neq j$

gde je Λ matrica na čijoj dijagonali se nalazi p karakterističnih korena $\lambda_i, i = 1, 2, \dots, p$.
Već smo rekli da je matrica A ortogonalna matrica, pa za nju važi jednakost

$$A'A = AA' = I$$

Zbog ove osobine lako možemo izraziti matrice Λ i Σ na sledeći način

$$A'\Sigma A = \Lambda \quad (2.11)$$

$$\Sigma = A\Sigma A' \quad (2.12)$$

U nastavku su navedena neka od svojstava ortonormirane linearne transformacije definisane sa 2.9.

Teorema 2.1. *Neka je data ortonormirana linearna transformacija*

$$y = B'x$$

gde je y vektor koji sadrži q elemenata i B' matrica dimenzija $q \times p$, za bilo koji ceo broj q , $1 \leq q \leq p$. Neka je $\Sigma_y = B'\Sigma B$ kovarijansna matrica vektora y . Tada važi:

- $tr(\Sigma_y)$ ima maksimalnu vrednost ukoliko je $B = A_q$, gde matrica A_q sadrži prvih q kolona matrice A
- $tr(\Sigma_y)$ ima minimalnu vrednost ukoliko je $B = A_q^*$, gde matrica A_q^* sadrži poslednjih q kolona matrice A

Dokaz. Neka je β_k k -ta kolona matrice B jednaka linearnoj kombinaciji kolona matrice A u p -dimenzionalnom prostoru, tj.

$$\beta_k = \sum_{i=1}^p c_{ik} \alpha_i, \quad k = 1, 2, \dots, q$$

gde je c_{ik} , $i = 1, 2, \dots, p$, $k = 1, 2, \dots, q$, konstanta. Zbog toga važi $B = AC$, gde je C matrica dimenzija $p \times q$ sa elementom c_{ik} u i -toj vrsti i k -toj koloni. Tada važi:

$$B'\Sigma B = C'A'\Sigma AC = C'\Lambda C = \sum_{i=1}^p \lambda_i c_i c_i'$$

gde je c_i' i -ta kolona matrice C i gde smo koristili jednakost 2.11. Dalje, važi

$$tr(B'\Sigma B) = \sum_{i=1}^p \lambda_i tr(c_i c_i') = \sum_{i=1}^p \lambda_i tr(c_i' c_i) = \sum_{i=1}^p \lambda_i c_i' c_i = \sum_{i=1}^p \sum_{k=1}^q \lambda_i c_{ik}^2 \quad (2.13)$$

Pošto je $C = A'B$ onda važi sledeće

$$C'C = B'AA'B = B'B = I_q$$

jer znamo da je matrica A ortogonalna i da su kolone matrice B ortonormirane. Zbog toga važi

$$\sum_{i=1}^p \sum_{k=1}^q c_{ik}^2 = q \quad (2.14)$$

kao i da su kolone matrice $\mathbf{C}$ ortonormirane.

Matricu $\mathbf{C}$ možemo posmatrati kao prvih q kolona ortogonalne matrice $\mathbf{D}$ čije su dimenzije $p \times p$. Vrste matrice $\mathbf{D}$ su ortonormirane, pa važi $\mathbf{d}_i' \mathbf{d}_i = 1$ $i = 1, 2, \dots, p$. Kako matrica $\mathbf{C}$ sadrži prvih q elemenata vrste matrice $\mathbf{D}$ sledi da je $\mathbf{c}_i' \mathbf{c}_i \leq 1$, $i = 1, 2, \dots, p$ pa je

$$\sum_{k=1}^q c_{ik}^2 \leq 1 \quad (2.15)$$

U jednakosti 2.13 koeficijenti uz λ_i su $\sum_{k=1}^q c_{ik}^2$. Suma ovih koeficijenata jednaka je q i ni jedan od koeficijenata ne može prekoračiti 1, što zaključujemo iz jednakosti 2.14 i 2.15, respektivno. Zbog toga što smo λ_i , $i = 1, \dots, p$ poređali u opadajući niz, jasno je da ćemo $\sum_{i=1}^p (\sum_{k=1}^q c_{ik}^2) \lambda_i$ maksimizovati ako nađemo c_{ik} za koje važi

$$\sum_{k=1}^q c_{ik}^2 = \begin{cases} 1, & i = 1, \dots, q \\ 0, & i = q + 1, \dots, p \end{cases} \quad (2.16)$$

Ukoliko je $\mathbf{B}' = \mathbf{A}'_q$, tada

$$c_{ik} = \begin{cases} 1, & 1 \leq i = k \leq q \\ 0, & \text{ostalo} \end{cases}$$

zadovoljava jednakost 2.16. Zbog toga $\text{tr}(\mathbf{\Sigma}_y)$ dostiže maksimalnu vrednost kad je $\mathbf{B}' = \mathbf{A}'_q$.

Drugi deo tvrđenja se pokazuje na sličan način.

□

Ova teorema govori da je prvih q ($q \leq p$) glavnih komponenti značajno jer one opisuju najveću varijansu, dok poslednjih nekoliko glavnih komponenti ima veoma malu varijansu i donekle nisu korisne za uklanjanje u model i opis problema. Ali, mogu da budu korisne za otkrivanje linearne povezanosti između elemenata vektora $\mathbf{x}$.

Teorema 2.2 (Spektralna dekompozicija matrice $\mathbf{\Sigma}$). Razlaganje matrice $\mathbf{\Sigma}$ dato na sledeći način

$$\mathbf{\Sigma} = \lambda_1 \alpha_1 \alpha_1' + \lambda_2 \alpha_2 \alpha_2' + \dots + \lambda_p \alpha_p \alpha_p' \quad (2.17)$$

se naziva spektralna dekompozicija.

Dokaz. Iz jednačine 2.12 imamo

$$\mathbf{\Sigma} = \mathbf{A} \mathbf{\Sigma} \mathbf{A}'$$

Raspisujući matrični proizvod sa desne strane jednačine dobijamo da je $\mathbf{\Sigma}$ jednaka

$$\sum_{i=1}^p \lambda_i \alpha_i \alpha_i'$$

što je trebalo pokazati.

□

Teorema 2.3. Neka je data ortonormirana linearna transformacija $\mathbf{y} = \mathbf{B}'\mathbf{x}$ gde je $\mathbf{y}$ vektor koji sadrži q elemenata i $\mathbf{B}'$ matrica dimenzija $q \times p$, za bilo koji ceo broj q , $1 \leq q \leq p$. Ako je $\det(\Sigma_{\mathbf{y}})$ determinanta kovarijanske matrice onda $\det(\Sigma_{\mathbf{y}})$ ima maksimalnu vrednost ukoliko je $\mathbf{B} = \mathbf{A}_q$, gde matrica $\mathbf{A}_q$ sadrži prvih q kolona matrice $\mathbf{A}$.

Ovu teoremu navodimo bez dokaza.⁴

Trag i determinanta kovarijanske matrice se nazivaju generalizovane varijanse. U nastavku ćemo pokazati da su generalizovane varijanse glavnih komponenti jednake generalizovanim varijansama originalnih promenljivih.

Kako je vektor glavnih komponenti $\mathbf{y} = \mathbf{A}\mathbf{x}$, njegova kovarijanska matrica je $Var(\mathbf{y}) = \mathbf{A}\Sigma\mathbf{A}'$. Ako sada zamenimo Σ dobijamo $Var(\mathbf{y}) = \mathbf{A}(\mathbf{A}'\mathbf{\Lambda}\mathbf{A})\mathbf{A}' = \mathbf{\Lambda}$, jer je $\mathbf{A}$ ortogonalna matrica. Dobili smo da je kovarijanska matrica glavnih komponenti $\mathbf{\Lambda}$.

Njena determinanta je $|\mathbf{\Lambda}|$ i jednaka je proizvodu karakterističnih korena λ_i . Na osnovu jednakosti 2.11 važi sledeće $|\mathbf{\Lambda}| = |\mathbf{A}'\Sigma\mathbf{A}| = |\mathbf{A}'||\Sigma||\mathbf{A}| = |\Sigma|$ ⁵, jer je $|\mathbf{A}| = \pm 1$. Dobili smo da su generalizovane varijanse (determinante kovarijansnih matrica) originalnog i transformisanog skupa međusobno jednake.

Trag kovarijanske matrice glavnih komponenti jednak je zbiru karakterističnih korena λ_i .

Na osnovu jednakosti 2.11 važi sledeće $tr(\mathbf{\Lambda}) = tr(\mathbf{A}'\Sigma\mathbf{A}) = tr(\mathbf{A}'\mathbf{A}\Sigma) = tr(\Sigma)$ ⁶, jer je $\mathbf{A}$ ortogonalna matrica. Odnosno dobili smo da su generalizovane varijanse (tragovi kovarijansnih matrica) originalnog i transformisanog skupa međusobno jednake.

Posmatrajući dijagonalne elemente kovarijanske matrice vidimo da je $Var(x_i) = \sum_{k=1}^p \lambda_k \alpha_{ki}^2$.

Relativan udeo i -te glavne komponente u objašnjenju ukupne varijanse određujemo stavljajući u međusobni odnos karakteristični koren λ_i i generalizovanu varijansu (trag matrice), odnosno određujemo na osnovu sledećeg izraza

$$\frac{\lambda_i}{\sum_{k=1}^p \lambda_k}, i = 1, \dots, p$$

Ukoliko u analizi glavnih komponenti dobijemo relativno visok doprinos jedne ili više glavnih komponenti ukupnoj varijansi, moguće je dalju analizu zasnovati na njima umesto na svim glavnim komponentama. Dalje, zaključujemo da možemo razložiti celu kovarijansnu matricu na doprinose $\lambda_k \alpha_k \alpha_k'$ svake glavne komponente. Iako se ne smanjuju striktno, elementi $\lambda_k \alpha_k \alpha_k'$ teže smanjenju kako se k povećava, jer će se λ_k smanjivati kako se povećava k , dok elementi α_k teže da ostanu isti zbog uslova normalnosti odnosno $\alpha_k' \alpha_k = 1$, $k = 1, \dots, p$. Izraz za kovarijansnu matricu $\Sigma = \mathbf{A}'\mathbf{\Lambda}\mathbf{A}$ možemo raspisati

$$\begin{aligned} \Sigma &= [\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_p] \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} \begin{bmatrix} \alpha_1' \\ \alpha_2' \\ \vdots \\ \alpha_p' \end{bmatrix} \\ &= \lambda_1 \alpha_1 \alpha_1' + \lambda_2 \alpha_2 \alpha_2' + \dots + \lambda_p \alpha_p \alpha_p' = \sum_{i=1}^p \lambda_i \alpha_i \alpha_i' \end{aligned}$$

⁴Dokaz se nalazi u [15, str.15].

⁵Determinanta proizvoda dve matrice jednaka je proizvodu njihovih determinanti.

⁶Korištena osobina traga matrice $tr(AB) = tr(BA)$.

Znači da je doprinos i -te glavne komponente kovarijansnoj matrici Σ dat sa $\lambda_i \alpha_i \alpha_i'$. Zadržavajući broj glavnih komponenti manji od p , kovarijansnu matricu aproksimiramo zbirom doprinosa zadržanih glavnih komponenti. Ukoliko doprinos ukupnoj varijansi naših zadržanih glavnih komponenti prelazi neku unapred fiksiranu vrednost, za očekivati je da aproksimacija kovarijansne matrice Σ dobro reprezentuje kovarijansnu strukturu originalnog skupa podataka. Ovo nije glavni rezultat analize, nego tek prateći. Glavne komponente su skoncentrisane na aproksimaciju varijanse, a ne kovarijanse promenljivih.

Prvi deo tvrđenja 2.1 nam kaže da glavne komponente objašnjavaju što je moguće bolje dijagonalne elemente kovarijansne matrice, ali i da dobro objašnjavaju vandijagonalne elemente kovarijansne matrice.

Iz jednakosti 2.17 je jasno da kovarijansnu matricu možemo konstruisati koristeći koeficijente i varijansu prvih r glavnih komponenti, gde je r rang kovarijansne matrice.

2.3.2 Geometrijska svojstva

U ovom delu ćemo pokazati dva geometrijska svojstva.

Teorema 2.4. *Neka je data familija p -dimenzionalnih elipsoida*

$$\mathbf{x}'\Sigma^{-1}\mathbf{x} = \text{const.} \quad (2.18)$$

Tada glavne komponente definišu ose ovog elipsoida.

Dokaz. *Glavne komponente su definisane transformacijom odnosno $\mathbf{y} = \mathbf{A}'\mathbf{x}$ ⁷, a kako je $\mathbf{A}$ ortogonalna matrica važi da je $\mathbf{x} = \mathbf{A}\mathbf{y}$. Zamenjujući dobijeno u jednakosti 2.18, dobijamo*

$$(\mathbf{A}\mathbf{y})'\Sigma^{-1}(\mathbf{A}\mathbf{y}) = \text{const} = \mathbf{y}'\mathbf{A}'\Sigma^{-1}\mathbf{A}\mathbf{y}$$

Poznato je da su karakteristični vektori matrice Σ^{-1} recipročni u odnosu na karakteristične vektore matrice Σ , uz pretpostavku da su elementi vektora striktno pozitivni. Kad u odgovarajuću jednakost 2.11 ubacimo poznato, važi da je $\mathbf{A}'\Sigma^{-1}\mathbf{A} = \Lambda^{-1}$, pa dalje važi da je

$$\mathbf{y}'\Lambda^{-1}\mathbf{y} = \text{const.}$$

Poslednju jednakost možemo zapisati na sledeći način

$$\sum_{k=1}^p \frac{y_k^2}{\lambda_k} = \text{const} \quad (2.19)$$

Jednačina 2.19 predstavlja jednačinu elipsoida u odnosu na njegove ose.

□

Ovaj rezultat je statistički značajan ako slučajni vektor $\mathbf{x}$ ima višedimenzionalnu normalnu raspodelu. U to slučaju, elipsoid dat jednačinom 2.18 definiše konture funkcije konstantne gustine verovatnoće za raspodelu od $\mathbf{x}$. Prva (i najveća) osa tj. glavna osa tog elipsoida definiše pravac u kom je varijacija najveća, što se slaže sa ranije datom definicijom prve glavne komponente. Druga osa maksimizuje varijaciju, i to tako da je ortogonalna prvoj osi, što se, takođe, slaže sa definicijom datom ranije. Ovu geometrijsku interpretaciju glavnih komponenti pomoću osa elipsoida konstantne gustine verovatnoće dao je Hotteling.

⁷Matrica $\mathbf{A}'$ je ortogonalna matrica dimenzija $p \times p$ čije su kolone karakteristični vektori kovarijansne matrice

Teorema 2.5. *Pretpostavimo da su $\mathbf{x}_1$ i $\mathbf{x}_2$ nezavisni slučajni vektori sa istom raspodelom i pretpostavimo da im odgovara ista linearna transformacija*

$$\mathbf{y}_i = \mathbf{B}'\mathbf{x}_i, i = 1, 2$$

Ako je matrica $\mathbf{B}$ matrica dimenzija $p \times q$ sa ortonormiranim kolonama takva da maksimizira $E[(\mathbf{y}_1 - \mathbf{y}_2)'(\mathbf{y}_1 - \mathbf{y}_2)]$, tada je $\mathbf{B} = \mathbf{A}_q$, gde matrica $\mathbf{A}_q$ sadrži prvih q kolona matrice $\mathbf{A}$.

Dokaz. Vidimo da $\mathbf{x}_1$ i $\mathbf{x}_2$ imaju isto očekivanje $\boldsymbol{\mu}$ i kovarijansnu matricu $\boldsymbol{\Sigma}$.

Stoga, $\mathbf{y}_1$ i $\mathbf{y}_2$ imaju istu sredinu $\mathbf{B}'\boldsymbol{\mu}$ i kovarijansnu matricu $\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}$. Dalje,

$$\begin{aligned} E[(\mathbf{y}_1 - \mathbf{y}_2)'(\mathbf{y}_1 - \mathbf{y}_2)] &= E\{[(\mathbf{y}_1 - \mathbf{B}'\boldsymbol{\mu}) - (\mathbf{y}_2 - \mathbf{B}'\boldsymbol{\mu})]'[(\mathbf{y}_1 - \mathbf{B}'\boldsymbol{\mu}) - (\mathbf{y}_2 - \mathbf{B}'\boldsymbol{\mu})]\} \\ &= E[(\mathbf{y}_1 - \mathbf{B}'\boldsymbol{\mu})'(\mathbf{y}_1 - \mathbf{B}'\boldsymbol{\mu})] + E[(\mathbf{y}_2 - \mathbf{B}'\boldsymbol{\mu})'(\mathbf{y}_2 - \mathbf{B}'\boldsymbol{\mu})] \end{aligned}$$

Unakrsni proizvod je jednak nuli, jer su $\mathbf{y}_1$ i $\mathbf{y}_2$ nezavisni.

Dalje, za $i=1,2$ važi

$$\begin{aligned} E[(\mathbf{y}_i - \mathbf{B}'\boldsymbol{\mu})'(\mathbf{y}_i - \mathbf{B}'\boldsymbol{\mu})] &= E\{tr[(\mathbf{y}_i - \mathbf{B}'\boldsymbol{\mu})'(\mathbf{y}_i - \mathbf{B}'\boldsymbol{\mu})]\} = E\{tr[(\mathbf{y}_i - \mathbf{B}'\boldsymbol{\mu})(\mathbf{y}_i - \mathbf{B}'\boldsymbol{\mu})']\} \\ &= tr\{E[(\mathbf{y}_i - \mathbf{B}'\boldsymbol{\mu})(\mathbf{y}_i - \mathbf{B}'\boldsymbol{\mu})']\} = tr(\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}) \end{aligned}$$

Dobijamo da je $E[(\mathbf{y}_i - \mathbf{B}'\boldsymbol{\mu})'(\mathbf{y}_i - \mathbf{B}'\boldsymbol{\mu})] = 2tr(\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B})$.

Iz teoreme 2.1 znamo da je $tr(\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B})$ maksimizovan kad je $\mathbf{B} = \mathbf{A}_q$, pa je teorema dokazana.

□

Ova teorema može da se posmatra kao algebarsko svojstvo glavnih komponenti, sam dokaz je algebarski. Ali, ovo svojstvo ima geometrijsku interpretaciju. Očekivano kvadratno Euklidsko rastojanje, u q -dimenzionalnom potprostoru, između dva vektora sa p slučajnih promenljivih koje imaju istu raspodelu, je najveće moguće ako je potprostor definisan prvom glavnom komponentom.

2.4 Interpretacija glavnih komponenti na osnovu korelacione matrice

Iako su glavne komponente proizvod mehaničkog postupka transformacije jednog skupa podataka u drugi, pokušaj njihove smislene interpretacije vrlo često može biti bezuspešan. Pokušaj njihove interpretacije možemo zasnovati na apsolutnoj veličini i predznaku koeficijenata linearne kombinacije. U toj interpretaciji koeficijenti bliski nuli nam sugerišu da odgovarajuća originalna promenljiva ne učestvuje značajno u formiranju te glavne komponente čije koeficijente posmatramo.

Definisanje i osobine glavnih komponenti smo zasnovali na osnovu karakterističnih korena i vektora kovarijansne matrice $\boldsymbol{\Sigma}$. Problem koji se javlja u interpretaciji glavnih komponenti je posledica njihove osetljivosti na različite merne skale originalnih promenljivih. Problem nastaje u praksi kada se dogodi da sve promenljive imaju različite jedinice mere, npr. jedna promenljiva predstavlja temperaturu, druga težinu, treća dužinu... Takođe, ako jedna od originalnih promenljivih ima znatno veću varijansu od ostalih, tad će ta promenljiva dominirati prvom glavnom komponentom bez obzira na korelacionu strukturu podataka.

Umesto da direktno koristimo koeficijente linearne kombinacije u cilju interpretacije glavnih komponenti, analizu možemo zasnovati na:

- koeficijentima korelacije originalnih promenljivih i glavnih komponenti
- korelacionoj matrici
- kovarijansi od $\frac{x_i}{\omega_i}$, gde je x_i i-ti elemenat vektora $\mathbf{x}$, a težine ω_i biramo u odnosu na važnost promenljivih

Prvi pristup je korišćenje koeficijenata korelacije originalnih promenljivih i glavnih komponenti. Sada ćemo da odredimo koeficijente korelacije između originalnih promenljivih i glavnih komponenti. Ranije smo odredili kovarijansnu matricu glavnih komponenti i originalnih promenljivih, odnosno $Var(\mathbf{y}) = \mathbf{\Lambda}$ i $Var(\mathbf{x}) = \mathbf{\Sigma}$, respektivno. Kovarijansa između $\mathbf{x}$ i $\mathbf{y}$ je

$$Cov(\mathbf{x}, \mathbf{y}) = Cov(\mathbf{x}, \mathbf{A}\mathbf{x}) = \mathbf{\Sigma}\mathbf{A}' = (\mathbf{A}'\mathbf{\Lambda}\mathbf{A})\mathbf{A}' = \mathbf{A}'\mathbf{\Lambda} = [\alpha_1\lambda_1, \alpha_2\lambda_2, \dots, \alpha_p\lambda_p]$$

Koeficijent korelacije između k-te originalne promenljive i i-te glavne komponente je dat sledećim izrazom

$$\rho_{x_k, y_i} = \frac{Cov(x_k, y_i)}{\sqrt{Var(x_k)}\sqrt{Var(y_i)}} = \frac{\lambda_i \alpha_{ik}}{\sqrt{\sigma_{kk}}\sqrt{\lambda_i}} = \alpha_{ik} \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i, k = 1, \dots, p$$

Iz prethodne jednakosti vidimo da se koeficijent uz k-tu promenljivu u i-toj glavnoj komponenti množi količnikom njihovih standardnih devijacija. U matričnom zapisu korelaciona matrica između vektora originalnih promenljivih $\mathbf{x}$ i vektora glavnih komponenti $\mathbf{y}$ je data

$$\rho_{\mathbf{x}, \mathbf{y}} = \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$$

gde smo sa $\mathbf{D}$ označili dijagonalnu matricu čiji su elementi varijanse originalnih promenljivih.

Drugi pristup je korišćenje korelacione matrice originalnih promenljivih umesto njihove kovarijansne matrice. Glavne komponente možemo da definišemo na sledeći način

$$\mathbf{y} = \mathbf{A}'\mathbf{x}^*$$

gde je $\mathbf{A}$ matrica čije kolone predstavljaju karakteristični vektori korelacione matrice, a $\mathbf{x}^*$ je standardizovani vektor, odnosno i-ti elemenat vektora $\mathbf{x}^*$ je $\frac{x_i}{\sigma_{ii}^{1/2}}$, $i=1, \dots, p$, x_i je i-ti element vektora $\mathbf{x}$ i σ_{ii} je varijansa elementa x_i . Zato korelacionu matricu možemo posmatrati kao kovarijansnu matricu standardizovanih promenljivih.

Sva svojstva koja su važila kod kovarijansne matrice važe i za korelacionu matricu. Dodajemo da je ukupna varijansa merena generalizovanom varijansom (tragom) jednaka p , odnosno dimenziji korelacione matrice, a da je koeficijent korelacije između k-te originalne promenljive i i-te glavne komponente jednak $\alpha_{ik}\sqrt{\lambda_i}$. U matričnom zapisu korelaciona matrica je data izrazom $\mathbf{A}\mathbf{\Lambda}^{\frac{1}{2}}$.

Izgleda kao da glavne komponente koje dobijemo pomoću korelacione matrice možemo vrlo lako dobiti na osnovu kovarijansne matrice, ali to nije tačno. Karakteristični koreni i vektori korelacione matrice nisu u jednostavnoj vezi sa odgovarajućim korenima i vektorima kovarijansne matrice. Glavne komponente dobijene iz korelacione matrice putem transformacije iz $\mathbf{x}^*$ u $\mathbf{x}$ mogu da budu kao komponente dobijene iz kovarijansne matrice $\mathbf{\Sigma}$, ali samo u specifičnim slučajevima. To objašnjavamo činjenicom da su glavne komponente invarijantne u odnosu na ortogonalnu transformaciju vektora $\mathbf{x}$, ali nisu invarijantne u odnosu na neke

druge transformacije. Naša transformacija iz $\mathbf{x}$ u $\mathbf{x}^*$ nije ortogonalna.

Zaključak je da se glavne komponente dobijene iz kovarijanske matrice razlikuju od glavnih komponenti dobijenih iz korelacione matrice i da se ne mogu izraziti jedna preko druge, ali postupak dobijanja komponenti je identičan u oba slučaja.

Još jedan problem koji se javlja kad za dobijanje glavnih komponenti koristimo kovarijansnu matricu umesto korelacione je to što je mnogo teže upoređivati dobijene rezultate sa rezultatima dobijenim nekim drugim analizama. Zatim, varijansa glavnih komponenti dobijenih iz različitih korelacionih matrica istih dimenzija je ista, što nije slučaj sa kovarijansnom matricom. Takođe, upoređivanje koeficijenata glavnih komponenti dobijenih različitim korelacionim matricama u cilju pronalaženja dve korelacione matrice koje daju iste komponente je mnogo lakše nego što bi bio slučaj sa kovarijansnim matricama.

Korišćenje kovarijanske matrice ima prednost u dva slučaja. Prva prednost je da je lakše izvoditi statističke zaključke o populacionim komponentama u odnosu na uzoračke komponente⁸ pomoću kovarijanske matrice nego pomoću korelacione. Druga prednost za korišćenje kovarijanske matrice naspram korelacione nam se ukazuje kada su svi elementi vektora $\mathbf{x}$ u istim jedinicama. Ali, kad su elementi $\mathbf{x}$ u različitim jedinicama koristimo korelacionu matricu.

Naposletku, ukažimo na iznos varijanse originalnih promenljivih koji je objašnjen zadržanim skupom glavnih komponenti. On pokazuje u kom stepenu zadržane glavne komponente dobro aproksimiraju varijansu svake originalne promenljive ponaosob. Na osnovu jednakosti 2.12 imamo da je varijansa k -te promenljive

$$\sigma_{kk}^2 = \sum_{i=1}^p \lambda_i \alpha_{ik}^2, \quad k = 1, \dots, p$$

Doprinos svake glavne komponente varijansi k -te promenljive jednak je kvadratu koeficijenta korelacije odnosno glavne komponente i te originalne promenljive. Doprinos svih glavnih komponenti računamo na osnovu korelacione matrice $\mathbf{A}\mathbf{\Lambda}^{\frac{1}{2}}$ tako što saberemo kvadrate elemenata u njenoj k -toj vrsti. Ukoliko smo zadržali nekoliko prvih glavnih komponenti, tad stavljanjem u odnos dobijene sume i odgovarajuće varijanse originalne promenljive dobijamo proporciju varijanse te promenljive koja je objašnjena zadržanim glavnim komponentama. Ta proporcija se naziva komunalitet promenljive i predstavlja procenat objašnjenja varijanse originalnih promenljivih zadržanim komponentama. Korišćenjem korelacione umesto kovarijanske matrice originalnih promenljivih odmah dobijamo proporciju varijanse originalne promenljive objašnjene zadržanim glavnim komponentama, jer je standardizacijom promenljivih vrednost varijanse jednaka jedinici.

Treći pristup je korišćenje kovarijansi od $\frac{x_i}{\omega_i}$, gde je x_i i -ti elemenat vektora $\mathbf{x}$, a težine ω_i biramo u odnosu na važnost promenljivih za analizu.

U specijalnom slučaju kad je $\omega_i = \sigma_{ii}^{\frac{1}{2}}$ i kad je $\omega_i = 1$ glavne komponente dobijamo pomoću korelacione i kovarijanske matrice, respektivno.

Neki autori su mišljenja da je izbor različitih vrednosti za ω_i bolji u primeni. Mada, u praksi je relativno neobično biranje vrednosti za ω_i .

⁸Definisane su u daljem radu.

2.5 Osobine glavnih komponenti čije su varijanse međusobno jednake ili jednake nuli

U ovom delu ćemo razmatrati dva problema koja se pojavljuju u teoriji, a relativno su neuobičajeni u praksi. Do sad smo pretpostavljali da su karakteristični koreni kovarijanske ili korelacione matrice svi međusobno različiti i da ni jedan od njih nije jednak nuli.

Jednakost karakterističnih korena i stoga jednakost varijanse glavnih promenljivih se javlja kod nekih određenih "patern matrica"⁹. Posledica ove pojave je da za q jednakih karakterističnih korena odgovarajući karakteristični vektori čine jedinstven q -dimenzionalni prostor koji nije ortogonalan nekom proizvoljnom prostoru. Geometrijski, za $q = 2$ i za $q = 3$ elipsa i sfera, respektivno, ne mogu biti jedinstveno definisane. Dodatni problem kod jednakosti varijansi glavnih promenljivih je taj što statistički zaključci postaju mnogo komplikovaniji.

Drugi problem odnosno kad je varijansa jednaka nuli se češće pojavljuje, ali je i dalje prilično neuobičajen. Ako je q karakterističnih korena jednako nuli, onda je rang matrice Σ jednak $p - q$ umesto p , kao što smo ranije imali. Zbog ovog rezultata bismo morali izvršiti modifikacije dokaza koje smo izveli za svojstva glavnih komponenti.

Svaka glavna komponenta sa nula varijansom definiše postojanje linearne veze između elemenata vektora x . Ako ova veza postoji, onda ona implicira da je jedna promenljiva redundantna, odnosno njena vrednost može biti određena iz vrednosti drugih promenljivih. Zbog toga možemo da smanjimo broj promenljivih sa p na $p - q$ bez gubljenja informacija. Idealno bi bilo da se linearna veza uoči pre određivanja glavnih komponenti i da se, prema tome, redukuje broj promenljivih.

Broj karakterističnih korena kovarijanske matrice koji su jednaki nuli je isti kao i broj karakterističnih korena korelacione matrice koji su jednaki nuli, jer postojanje linearne veze između elemenata vektora x implicira postojanje linearne veze između standardizovanih promenljivih i obrnuto. Jednakost nekih karakterističnih korena kovarijanske (korelacione) matrice sa odgovarajućim karakterističnim korenima korelacione (kovarijanske) matrice ne implicira da su svi koreni jedne matrice jednaki korenima druge matrice. Ako je $p > 2$, tada je poslednjih $p - 1$ karakterističnih korena kovarijanske matrice jednako. Ovo, u opštem slučaju, ne važi za korelacionu matricu.

⁹"Patern matrice" (eng. patterned matrices) su matrice koje imaju određenu strukturu ili obrazac. Više o njima u [16].

2.6 Uzoračke glavne komponente

Do sada smo analizu glavnih komponenti zasnivali na populacionoj kovarijansnoj ili korelacionoj matrici. U ovom delu se bavimo analizom glavnih komponenti dobijenih iz uzoračke kovarijansne ili korelacione matrice odnosno analizu zasnivamo na osnovu reprezentativnog uzorka.

2.6.1 Definicija i svojstva uzoračkih glavnih komponenti

Analizu glavnih komponenti zasnivamo na uzoračkoj kovarijansnoj matrici S ili na uzoračkoj korelacionoj matrici R . Umesto uzoračke kovarijansne matrice S , koja predstavlja nepristrasnu ocenu kovarijansne matrice Σ , možemo koristiti $\hat{\Sigma} = \frac{n-1}{n}S$ koja predstavlja ocenu najveće verodostojnosti kovarijansne matrice Σ . Bez obzira koja od ocena se koristi, glavne komponente su identične, kao i proporcija objašnjene varijanse. Matrice S i $\hat{\Sigma}$ daju istu uzoračku korelacionu matricu R .

Neka je uzet uzorak od n elemenata $x_1, x_2, \dots, x_n$ slučajnog vektora x iz p -dimenzionalne populacije sa sredinom μ i kovarijansnom matricom Σ .

Neka je $\hat{y}_{i1} = \mathbf{a}_1' \mathbf{x}_i$, $i = 1, 2, \dots, n$. Biramo vektor koeficijenata $\mathbf{a}_1'$ tako da maksimiziramo uzoračku varijansu

$$\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_{i1} - \bar{y}_1)^2$$

uz uslov normalnosti, odnosno da vazi $\mathbf{a}_1' \mathbf{a}_1 = 1$.

Dalje, neka je $\hat{y}_{i2} = \mathbf{a}_2' \mathbf{x}_i$, $i = 1, 2, \dots, n$ i neka biramo vektor koeficijenata $\mathbf{a}_2$ tako da maksimiziramo uzoračku varijansu, uz uslov normalnosti $\mathbf{a}_2' \mathbf{a}_2 = 1$ i uslov da su $\hat{y}_{i1}$ i $\hat{y}_{i2}$ međusobno nekorelisane. Nastavljajući dalje ovaj proces dobijamo uzoračke glavne komponente. Sa $\mathbf{a}_k' \mathbf{x}$ za $k = 1, 2, \dots, p$ definišemo k -tu uzoračku komponentu, a $\hat{y}_{ik}$ predstavlja skor i -te opservacije za k -tu uzoračku komponentu. Skorove glavnih komponenti koristimo u grafičkom prikazu opservacija u dvodimenzionalnom prostoru generisanom parovima glavnih komponenti.

U izrazu $\mathbf{a}_k' \mathbf{x}$ smo sa $\mathbf{a}_k$ označili ocenu vektora koeficijenata originalnih promenljivih $x_1, x_2, \dots, x_p$ za k -tu glavnu komponentu, i tu ocenu vektora računamo za svaki element uzorka.

Prateći izvođenje populacioni glavnih komponenti, ispostavlja se da je uzoračka varijansa k -te uzoračke glavne komponente l_k odnosno najveći karakteristični koren uzoračke kovarijansne matrice S za $x_1, x_2, \dots, x_n$ i $\mathbf{a}_k$ je odgovarajući karakteristični vektor.

Definišimo matricu $\hat{X}$ dimenzija $n \times p$ takvu da joj je u i -toj vrsti i k -toj koloni element $\hat{x}_{ik}$ tj. k -ti element od x_i . Zatim, definišimo matricu $\hat{Y}$ dimenzija $n \times p$ takvu da joj je u i -toj vrsti i k -toj koloni element $\hat{y}_{ik}$. Tada važi da je $\hat{Y} = \hat{X}A$, gde je A ortogonalna matrica dimenzija $p \times p$ u čijoj k -toj koloni se nalazi vektor $\mathbf{a}_k$.

Ako je sredina svakog elementa vektora x jednaka nuli, onda je kovarijansna matrica jednaka $S = \frac{1}{n} \hat{X}' \hat{X}$. Češće se desava da sredina elemenata vektora x nepoznata i u tom slučaju element u j -toj vrsti i k -toj koloni matrice S je

$$\frac{1}{n-1} \sum_{i=1}^n (\hat{x}_{ij} - \bar{x}_j)(\hat{x}_{ik} - \bar{x}_k)$$

gde $\bar{x}_j$ predstavlja uzoračku sredinu odnosno

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n \hat{x}_{ij}, \quad j = 1, 2, \dots, p$$

Stoga, matrica S može da se zapiše na sledeći način

$$S = \frac{1}{n-1} X'X \quad (2.20)$$

gde je X matrica dimenzija $n \times p$ sa elementom $\hat{x}_{ij} - \bar{x}_j$ u i -toj vrsti i j -toj koloni. Koristićemo oznaku x_{ij} za element u i -toj vrsti i j -toj koloni matrice X . Zatim, definišemo matricu skorova kao $Y = XA$.

Važno je napomenuti da su karakteristični vektori matrice $\frac{1}{n-1}X'X$ i $X'X$ identični i da su karakteristični koreni matrice $\frac{1}{n-1}X'X$ jednaki $\frac{1}{n-1}$ (karakteristični koreni matrice $X'X$). Nekad je zgodnije raditi sa karakterističnim korenima i vektorima matrice $X'X$ nego direktno sa S .

Sada definišemo algebarska i geometrijska svojstva uzoračkih glavnih komponenti. Osvrćući se na algebarska svojstva populacionih glavnih komponenti, definišimo

$$y_i = B'x_i, \quad i = 1, 2, \dots, n$$

gde B , kao u prethodno definisanim teoremama, predstavlja matricu dimenzija $p \times q$ čije kolone su ortonormirane. Teoreme nam i dalje važe, ali uzoračka kovarijansna matrica opservacija y_i zamenjuje Σ_y , a matrica A je matrica čija je k -ta kolona a_k , dok A_q i A_q^* sadrže prvih i poslednjih q kolona, respektivno. Dokazi su slični kao kod populacionih glavnih komponenti.

Spektralna dekompozicija nam sad izgleda ovako

$$S = l_1 a_1 a_1' + l_2 a_2 a_2' + \dots + l_p a_p a_p' \quad (2.21)$$

Statistički zaključci dobijeni za uzoračke komponente se ne razlikuju od zaključaka dobijenih za populacione komponente, osim što sve moramo da posmatramo u kontekstu uzorka, a ne populacije.

Dalje, navodimo svojstvo koje nam kaže kako možemo koristiti uzoračke glavne komponente u regresiji¹⁰.

Teorema 2.6. *Pretpostavimo da se X , koji smo prethodno definisali, sastoji od n opservacija p prediktivnih promenljivih x merenih njihovom uzoračkom sredinom i da je odgovarajuća regresiona jednačina*

$$y = X\beta + \varepsilon$$

gde je y vektor n opservacija zavisnih promenljivih merenih njihovom uzoračkom sredinom¹¹. Pretpostavimo da je X transformisana pomoću jednakosti $Z = XB$, gde je B ortogonalna matrica dimenzija $p \times p$. Regresijsku jednačinu možemo zapisati na sledeći način

$$y = Z\gamma + \varepsilon$$

¹⁰Regresija je statistička metoda koja analizira zavisnost jedne promenljive (zavisna promenljiva) u odnosu na skup drugih promenljivih (nezavisne promenljive).

¹¹Oznaka y za nezavisne promenljive, koja se standardno koristi u regresiji, nema veze sa oznakom y koju smo do sad koristili u radu.

gde $\gamma = B^{-1}\beta$. Ocenjivač za γ je $\hat{\gamma} = (Z'Z)^{-1}Z'y$. Elementi ocenjivača $\hat{\gamma}$ imaju najmanju moguću varijansu ako je $B = A$, matrica A je matrica čija je k -ta kolona k -ti karakteristični vektor matrice $X'X$ tj. uzoračke kovarijanske matrice S . Zbog toga Z sadrži vrednosti uzoračkih glavnih komponenti vektora x .

Dokaz ovog svojstva izostavljamo¹².

Ovo svojstvo nam kaže da je dobra ideja zameniti prediktivne promenljive u regresionoj jednačini sa prvih nekoliko glavnih komponenti. Mana ovog tvrđenja što ne isključuje vezu između zavisne promenljive y i elemenata od x , kao ni između y i glavnih komponenti.

Geometrijska svojstva uzoračkih komponenti, kao i algebarska, su relevantna geometrijskim svojstvima populacionih komponenti, uz male modifikacije.

Svojstvo iz teoreme 2.4 nam i dalje važi ako kovarijansnu matricu Σ zamenimo uzoračkom kovarijansnom matricom S . Elipsoid dat jednačinom $x'S^{-1}x = const$ sada ne definiše konture funkcije konstantne gustine verovatnoće, nego ocenjuje konture od $x_1, x_2, \dots, x_n$ sa višedimenzionalnom normalnom raspodelom. Uvođenjem nenula uzoračke sredine, elipsoid dat jednačinom

$$(x - \bar{x})'S^{-1}(x - \bar{x}) = const.$$

definiše konture fiksnog Mahalanobisovog rastojanja (eng. Mahalanobis distance) od uzoračke sredine $\bar{x}$. Neki autori su interpretirali glavne komponente kao traženje ortogonalnih pravaca za koje je Mahalanobisovo rastojanje od originalnih podataka do hipersfere¹³ minimalno.

Svojstvo iz teoreme 2.5 populacionih komponenti ćemo preneti na uzoračke glavne komponente.

Teorema 2.7. *Neka su opservacije $x_1, x_2, \dots, x_n$ transformisane pomoću*

$$y_i = B'x_i, \quad i = 1, 2, \dots, n$$

gde je matrica B matrica dimenzija $p \times q$ sa ortonormiranim kolonama, tako da važi su $y_1, \dots, y_n$ projekcije $x_1, x_2, \dots, x_n$ na q -dimenzionalni potprostor. Tada je

$$\sum_{h=1}^n \sum_{i=1}^n (y_h - y_i)'(y_h - y_i)$$

maksimizovano za $B = A_q$, a minimizovano za $B = A_q^*$.

Ovo svojstvo znači da ako je n opservacija projektovano na q -dimenzionalni potprostor, onda je suma kvadrata Euklidskog rastojanja između svakog para opservacija maksimalna ako je potprostor definisan pomoću prvih q glavnih komponenti ili minimalna ako je potprostor definisan pomoću poslednjih q komponenti.

Dokaz je sličan dokazu odgovarajućeg svojstva populacionih komponenti, pa ga nećemo navoditi.

¹²Dokaz se nalazi u [15, str.32].

¹³Hipersfera predstavlja generalizaciju pojma sfere u prostoru proizvoljne dimenzije. Kao i sfera, hipersfera je površ, a može se definisati kao skup svih tačaka koje se nalaze na datom rastojanju od date tačke.

Argumenti za i protiv korišćenja uzoračke korelacione matrice umesto uzoračke kovarijansne matrice su identični onim koje smo naveli kod populacionih komponenti. Takođe, ne postoji veza između uzoračkih glavnih komponenti dobijenih pomoću korelacione matrice i onih dobijenih pomoću odgovarajuće kovarijansne matrice.

Zatim, problemi koji su se pojavljivali kada su karakteristični koreni populacione kovarijansne matrice jednaki nuli ili su međusobno jednaki se pojavljuju i kod uzoračkih komponenti.

U praksi jednakost korena je retka. Uzorkovanje nam obezbeđuje nejednakost karakterističnih korena dobijenih iz uzoračkih matrica.

Takođe, sa pažljivo odabranim promenljivima, glavne komponente sa nula varijansom su relativno retke. Ukoliko se pojavi q karakterističnih korena jednakih nuli, to znači da tačke $x_1, \dots, x_n$ leže u $(p - q)$ -dimenzionalnom potprostoru p -dimenzionalnog prostora. To znači da q različitih linearnih funkcija p originalnih promenljivih imaju konstantne vrednosti za svaku opservaciju. Idealno bi bilo da se veze između promenljivih uoče pre pronalaska komponenti i da se redukuje broj promenljivih kako bismo izbegli ove veze.

2.6.2 SVD dekompozicija

SVD (eng. singular value decomposition) ili dekompozicija pomoću singularnih vrednosti je dekompozicija koja se koristi u mnogim algoritmima numeričke algebre. Takođe, ima i teorijski značaj u linearnoj algebri.[6]

SVD dekompozicija je značajna za analizu glavnih komponenti u nekoliko aspekata.

Neka je proizvoljna matrica X dimenzija $n \times p$ koja predstavlja matricu sa n opservacija i p promenljivih sa njihovim sredinama. Matricu X možemo zapisati na sledeći način

$$X = ULA' \quad (2.22)$$

gde su

1. U i A su matrice dimenzija $n \times r$ i $p \times r$, respektivno, i koje imaju ortonormirane kolone pa važi $U'U = I_r$ i $A'A = I_r$
2. L je dijagonalna matrica dimenzija $r \times r$
3. r je rang matrice X

Posmatrajmo spektralnu dekompoziciju matrice $X'X$.

Poslednjih $p - r$ članova u jednakosti 2.21 u odgovarajućem izrazu za $X'X$ su jednaki nuli, jer je poslednjih $p - r$ karakterističnih korena matrice X jednako nuli i zbog toga matrica $X'X$ ima rang jednak r . Prema tome važi sledeće

$$(n - 1)S = X'X = l_1 \mathbf{a}_1 \mathbf{a}_1' + l_2 \mathbf{a}_2 \mathbf{a}_2' + \dots + l_r \mathbf{a}_r \mathbf{a}_r'$$

Matricu A definišemo kao matricu dimenzija $p \times r$ čija je k -ta kolona $\mathbf{a}_k$, a matricu U kao matricu dimenzija $n \times r$ čija je k -ta kolona

$$\mathbf{u}_k = l_k^{-\frac{1}{2}} X \mathbf{a}_k \quad \text{za } k = 1, \dots, r$$

Matricu $\mathbf{L}$ definišemo kao dijagonalnu matricu dimenzija $r \times r$ sa k -tim dijagonalnim elementom $l_k^{\frac{1}{2}}$. Matrice izabrane na ovaj način zadovoljavaju uslove 1.i 2. pa sada možemo pokazati da važi $\mathbf{X} = \mathbf{U}\mathbf{L}\mathbf{A}'$.

$$\begin{aligned} \mathbf{U}\mathbf{L}\mathbf{A}' &= \mathbf{U} \begin{bmatrix} l_1^{1/2} \mathbf{a}_1' \\ l_2^{1/2} \mathbf{a}_2' \\ \vdots \\ l_r^{1/2} \mathbf{a}_r' \end{bmatrix} \\ &= \sum_{k=1}^r l_k^{-\frac{1}{2}} \mathbf{X} \mathbf{a}_k l_k^{\frac{1}{2}} \mathbf{a}_k' = \sum_{k=1}^r \mathbf{X} \mathbf{a}_k \mathbf{a}_k' \\ &= \sum_{k=1}^p \mathbf{X} \mathbf{a}_k \mathbf{a}_k' \end{aligned}$$

Poslednja jednakost važi jer je $\mathbf{a}_k$, $k = r + 1, r + 2, \dots, p$ karakteristični vektor za odgovarajuće nula karakteristične korene matrice $\mathbf{X}'\mathbf{X}$. Vektor $\mathbf{X}\mathbf{a}_k$ je vektor skorova k -te glavne komponente, nula varijansa poslednjih $p - r$ glavnih komponenti implicira da je $\mathbf{X}\mathbf{a}_k = 0$ za $k = r + 1, r + 2, \dots, p$. Prema tome, važi

$$\mathbf{U}\mathbf{L}\mathbf{A}' = \mathbf{X} \sum_{k=1}^p \mathbf{a}_k \mathbf{a}_k' = \mathbf{X}$$

kao što je trebalo pokazati.

SVD dekompozicija je važna za analizu glavnih komponenti u dva slučaja.

Kao prvo, obezbeđuje nam efikasnu metodu pronalaska glavnih komponenti. Jasno je, ako pronađemo matrice $\mathbf{U}, \mathbf{L}, \mathbf{A}$ takve da zadovoljavaju jednakost 2.22 matrice $\mathbf{A}$ i $\mathbf{L}$ će sadržati karakteristične vektore i kvadratne korene svojstvenih vrednosti matrice $\mathbf{X}'\mathbf{X}$, tj. koeficijente i standardne devijacije glavnih komponenti dobijenih na osnovu uzoračke kovarijanske matrice $\mathbf{S}$. U matrici $\mathbf{U}$ dobijamo skalirane skorove. Kada jednakost 2.22 pomnožimo sa desne strane matricom $\mathbf{A}$, dobijamo $\mathbf{X}\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{A}'\mathbf{A} = \mathbf{U}\mathbf{L}$, jer je $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$. Matrica $\mathbf{X}\mathbf{A}$ je dimenzija $n \times r$ i njena k -ta kolona sadrži skorove k -te glavne komponente. Skorovi su dati sledećom formulom

$$y_{ik} = u_{ik} l_k^{\frac{1}{2}} \quad \text{za } i = 1, \dots, n \text{ i } k = 1, \dots, r$$

ili u matričnom zapisu $\mathbf{Y} = \mathbf{U}\mathbf{L}$.

Drugo, obezbeđuje nam dodatni uvid kako glavne komponente rade. Daju nam algebarski i grafički način za predstavljanje i tumačenje rezultata glavnih komponenti.

2.6.3 Funkcija raspodele uzoračkih glavnih komponenti

Matematičari su uložili mnogo truda u izvođenje funkcije raspodele, uglavnom asimptotske, za koeficijente i varijansu uzoračkih glavnih komponenti odnosno za karakteristične korene i vektore uzoračke kovarijansne matrice. Rezultati koje su dobili imaju neke mane. Prva mana je što se do rezultati dolazi komplikovanim matematičkim računom, zatim, što su rezultati uglavnom asimptotski i što su zasnovani na pretpostavci da originalne promenljive imaju višedimenzionalnu normalnu raspodelu.

Pretpostavimo da x ima p -dimenzionalnu normalnu raspodelu sa sredinom μ i kovarijansnom matricom Σ tj. $x \sim N(\mu, \Sigma)$. Pretpostavljamo da znamo Σ , a da μ nije dato. Tada

$$(n-1)S \sim W_p(\Sigma, n-1)$$

odnosno $(n-1)S$ ima Višartovu raspodelu (eng. Wishart distribution)¹⁴ sa parametrima $\Sigma, n-1$. Funkcija gustine za matricu B koja ima $W_p(\Sigma, n-1)$ raspodelu izgleda ovako

$$c|B|^{\frac{n-p-2}{2}} e^{-\frac{1}{2}tr(\Sigma^{-1}B)}$$

gde je

$$c^{-1} = 2^{\frac{p(n-1)}{2}} \pi^{\frac{p(1-p)}{4}} |\Sigma|^{\frac{n-1}{2}} \prod_{j=1}^p \Gamma\left(\frac{n-j}{2}\right)$$

a sa Γ smo označili Gama funkciju¹⁵.

Neka su, za $k = 1, 2, \dots, p$, $l_k, \mathbf{a}_k$ karakteristični koreni i vektori matrice S , respektivno, a λ_k, α_k matrice Σ , respektivno. Neka su l, λ vektori koji sadrže elemente l_k i λ_k , respektivno, a j -ti elementi vektora $\mathbf{a}_k$ i α_k su a_{kj} i α_{kj} , respektivno. Pretpostavljamo, iako pretpostavka nije realna, da su $\lambda_1 > \lambda_2 > \dots > \lambda_p > 0$ tj. da su populacioni karakteristični koreni pozitivni i različiti. Sledeći rezultati su asiptionski:

1. elementi l_k su nezavisni od elemenata $\mathbf{a}_k$
2. l i $\mathbf{a}_k$ imaju zajedničku normalnu raspodelu

3.

$$E(l) = \lambda, E(\mathbf{a}_k) = \alpha_k, k = 1, \dots, p \quad (2.23)$$

4.

$$cov(l_k, l_{k'}) = \begin{cases} \frac{2\lambda_k^2}{n-1}, & k = k' \\ 0, & k \neq k' \end{cases} \quad (2.24)$$

5.

$$cov(a_{kj}, a_{k'j'}) = \begin{cases} \frac{\lambda_k}{n-1} \sum_{i=1}^p \frac{\lambda_i \alpha_{ij} \alpha_{ij'}}{(\lambda_i - \lambda_k)^2}, & k = k' \\ \frac{\lambda_k \cdot \lambda_{k'} \cdot \alpha_{kj} \cdot \alpha_{k'j'}}{(n-1)(\lambda_k - \lambda_{k'})^2}, & k \neq k' \end{cases} \quad (2.25)$$

Treba istaći da su ovi rezultati asimptotski i samim tim aproksimiraju samo konačne uzorke.

¹⁴Višartova raspodela je, u statistici, generalizacija χ^2 raspodele, a u slučaju da stepeni slobode nisu celi brojevi onda je uopštenje Gama raspodele.[29]

¹⁵Gama funkcija je definisana nesvojstvenim integralom $\Gamma(z) = \int_0^{\infty} e^{-t} t^{z-1} dt$ i važi $\Gamma(n+1) = n!$

Zaključci izvedeni na osnovu uzoračkih glavnih komponenti

Rezultate, koje smo dobili u prethodnom odeljku, možemo koristiti za donošenje zaključaka o populacionim glavnim komponentama na osnovu datih uzoračkih glavnih komponenti, ako su sve potrebne pretpostavke ispunjene. Glavna pretpostavka je da x ima višedimenzionalnu normalnu raspodelu, a ta pretpostavka često nije zadovoljena i zbog toga su rezultati ograničeni u praksi. Može se protumačiti da se analiza glavnih komponenti izvodi samo u slučaju da podaci, bar asimptotski, imaju višedimenzionalnu normalnu raspodelu i da samo u tom slučaju imamo ispravne zaključke. To je poprilično uzak pogled na ono što analiza glavnih komponenti može, jer nam ona može obezbediti informacije za različite setove podataka, bilo da su promenljive neprekidne i normalno raspoređene ili ne. Većinski se koristi kao deskriptivna tehnika, a manje za izvođenje zaključaka. Mnogi analizu glavnih komponenti tretiraju kao čisto deskriptivnu tehniku. Ideje zaključivanja mogu biti vrlo korisne, pa ćemo sada da se bavimo njima.

2.6.4 Određivanje intervala poverenja

Asimptotska marginalna raspodela za λ_k i α_{kj} , data u prethodnom delu, se može koristiti za konstrukciju intervala poverenja za λ_k i α_{kj} , respektivno. Marginalna raspodela za λ_k , koristeći jednakosti 2.23 i 2.24, je aproksimativno

$$l_k \sim N\left(\lambda_k, \frac{2\lambda_k^2}{n-1}\right) \quad (2.26)$$

odnosno

$$\frac{l_k - \lambda_k}{\lambda_k \left(\frac{2}{n-1}\right)^{\frac{1}{2}}} \sim N(0, 1)$$

što nas dovodi do intervala poverenja, sa koeficijentom poverenja $1 - \alpha$ za λ_k , u sledećem obliku

$$\frac{l_k}{\left(1 + \frac{2}{n-1} z_{\alpha/2}\right)^{\frac{1}{2}}} < \lambda_k < \frac{l_k}{\left(1 - \frac{2}{n-1} z_{\alpha/2}\right)^{\frac{1}{2}}} \quad (2.27)$$

gde je $z_{\alpha/2}$ gornji $100 \cdot \frac{\alpha}{2}$ percentil standardne normalne raspodele. U izvođenju intervala poverenja se pretpostavlja da je n dovoljno veliko da važi $\frac{2}{n-1} z_{\alpha/2} < 1$. Napominjemo da dobijeni interval poverenja može biti veoma širok, čak i za velike vrednosti n , tj. veličine uzorka. Iz tog razloga postupak određivanja glavnih komponenti koje treba zadržati u daljoj analizi ne možemo zasnivati na osnovu dobijene intervalne ocene karakterističnog korena. Drugi način da odredimo interval poverenja je da posmatramo raspodelu od $\ln(l_k)$. Na osnovu 2.26 sledi da je aproksimativno

$$\ln(l_k) \sim N\left(\ln(\lambda_k), \frac{2}{n-1}\right)$$

i tako otklanjamo zavisnost nepoznatog parametra λ_k u varijansi. Aproksimativni interval poverenja za $\ln(l_k)$, sa koeficijentom poverenja $1 - \alpha$ je $\ln(l_k) \pm \frac{2}{n-1} z_{\alpha/2}$. Transformacijom nazad u λ_k , dobijamo da je aproksimativni interval poverenja sledećeg oblika

$$l_k e^{-\frac{2}{n-1} z_{\alpha/2}} \leq \lambda_k \leq l_k e^{\frac{2}{n-1} z_{\alpha/2}} \quad (2.28)$$

Sada je l_k asimptotski nezavisan, a zajedničke oblasti poverenja za nekoliko λ_k su dobijene jednostavnim kombinovanjem intervala datih u 2.27 i 2.28 i biranjem koeficijenata poverenja da bismo dobili željeni nivo poverenja.

Aproksimativni interval poverenja za α_{kj} možemo odrediti pomoću marginalne raspodele za a_{kj} čije su sredina i varijansa date u 2.23 i 2.25 na sličan način kao i za λ_k , samo što izraz postane komplikovaniji. Još komplikovaniji postane kada hoćemo da odredimo oblast poverenja za nekoliko α_{kj} i to zbog zavisnosti pojedinačnih a_{kj} . Posmatrajmo $\mathbf{a}_k$. Iz 2.23 i 2.25 sledi da je, aproksimativno,

$$\mathbf{a}_k \sim N(\boldsymbol{\alpha}_k, \mathbf{T}_k)$$

gde je

$$\mathbf{T}_k = \frac{\lambda_k}{n-1} \sum_{\substack{l=1 \\ l \neq k}}^p \frac{\lambda_l}{(\lambda_l - \lambda_k)^2} \boldsymbol{\alpha}_l \boldsymbol{\alpha}_l'$$

Rang ove matrice $\mathbf{T}_k$ je $p-1$. Može se pokazati da, aproksimativno, važi

$$(n-1)(\mathbf{a}_k - \boldsymbol{\alpha}_k)'(l_k \mathbf{S}^{-1} + l_k^{-1} \mathbf{S} - 2\mathbf{I}_p)(\mathbf{a}_k - \boldsymbol{\alpha}_k) \sim \chi_{p-1}^2 \quad (2.29)$$

Kako je $\mathbf{a}_k$ karakteristični vektor matrice $\mathbf{S}$, a l_k karakteristični koren, sledi da je $l_k^{-1} \mathbf{S} \mathbf{a}_k = l_k^{-1} l_k \mathbf{a}_k = \mathbf{a}_k$, $l_k \mathbf{S}^{-1} \mathbf{a}_k = l_k l_k^{-1} \mathbf{a}_k = \mathbf{a}_k$ i $(l_k \mathbf{S}^{-1} + l_k^{-1} \mathbf{S} - 2\mathbf{I}_p) \mathbf{a}_k = \mathbf{a}_k + \mathbf{a}_k - 2\mathbf{a}_k = \mathbf{0}$ pa se 2.29 pretvara u

$$(n-1)\boldsymbol{\alpha}_k'(l_k \mathbf{S}^{-1} + l_k^{-1} \mathbf{S} - 2\mathbf{I}_p)\boldsymbol{\alpha}_k \sim \chi_{p-1}^2 \quad (2.30)$$

Iz 2.30 dobijamo aproksimativnu oblast poverenja za $\boldsymbol{\alpha}_k$, sa koeficijentom poverenja $1 - \lambda$, koja ima sledeći oblik

$$(n-1)\boldsymbol{\alpha}_k'(l_k \mathbf{S}^{-1} + l_k^{-1} \mathbf{S} - 2\mathbf{I}_p)\boldsymbol{\alpha}_k \leq \chi_{p-1;\alpha}^2$$

2.6.5 Testiranje hipoteza

Analiza komponenti predstavlja redukciju podataka, ali nam nije sugerisano sa koliko glavnih komponenti je potrebno vršiti analizu. Zainteresovani smo za one glavne komponente koje imaju najveće karakteristične korene i zbog toga kad govorimo o značajnosti glavnih komponenti imamo u vidu testove karakterističnih korena.

Rezultate dobijene iz 2.23, 2.24 i 2.25 koristimo za testiranje hipoteza. Ako testiramo osnovnu hipotezu $H_0 : \lambda_k = \lambda_{k_0}$ protiv alternativne hipoteze $H_1 : \lambda_k \neq \lambda_{k_0}$, tada je pogodna test statistika

$$\frac{l_k - \lambda_{k_0}}{\frac{2}{n-1} \lambda_{k_0}}$$

koji, aproksimativno, ima normalnu $N(0, 1)$ raspodelu za H_0 , pa pretpostavku H_0 odbacujemo sa nivoom značajnosti α ako je

$$\left| \frac{l_k - \lambda_{k_0}}{\frac{2}{n-1} \lambda_{k_0}} \right| \geq z_{\alpha/2}$$

Slično, rezultat iz 2.30 možemo iskoristiti da testiramo $H_0 : \boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k_0}$ protiv $H_1 : \boldsymbol{\alpha}_k \neq \boldsymbol{\alpha}_{k_0}$. Hipotezu H_0 odbacujemo sa nivoom značajnosti α ako je

$$(n-1)\boldsymbol{\alpha}_{k_0}'(l_k \mathbf{S}^{-1} + l_k^{-1} \mathbf{S} - 2\mathbf{I}_p)\boldsymbol{\alpha}_{k_0} \geq \chi_{p-1;\alpha}^2$$

Ovo je aproksimativna test statistika, a pomoću nekih modifikacija možemo poboljšati χ^2 aproksimaciju.

Za karakteristične korene kovarijansne matrice najpoznatiji je Bartletov test (eng. Bartlett's test), koji koristimo za testiranje hipoteze da su poslednja $p - q$ korena međusobno jednaka. Dakle, nulta hipoteza je $H_0 : \lambda_{q+1} = \lambda_{q+2} = \dots = \lambda_p$ i testiramo je protiv alternativne hipoteze H_1 da su barem dva, od poslednjih $p - q$ karakterističnih korena, međusobno različita. Ako prihvatimo nultu hipotezu, tada u analizi koristimo prvih q glavnih komponenti jer za njih pretpostavljamo da obuhvataju relevantan iznos ukupne varijanse, a da poslednjih $p - q$ glavnih komponenti sa jednakom varijansom mere samo šum u podacima.

Test statistika, koju koristimo za testiranje hipoteze H_0 protiv H_1 , zasnovana na pretpostavci o normalnosti, a konstruisana na osnovu principa količnika najveće verodostojnosti (eng. likelihood ratio-LR)¹⁶ je

$$LR = \left(\frac{\prod_{k=q+1}^p l_k}{\left[\sum_{k=q+1}^p \frac{l_k}{p-q} \right]^{p-q}} \right)^{\frac{n}{2}}$$

Zaključujemo da je test zasnovan na poređenju geometrijske i aritmetičke sredine poslednjih $p - q$ karakterističnih korena. Poznato je da se geometrijska sredina koristi ako u podacima ima ekstremnih vrednosti, jer izravnavu proporcionalne promene podataka, a i da je uvek manja ili jednaka aritmetičkoj sredini. Ukoliko je tačna nulta hipoteza, vrednost LR test statistike je jednaka nuli. Udaljavajući se od nulte hipoteze razlika između geometrijske i aritmetičke sredine biće sve veća, što znači da će vrednost test statistike LR biti sve manja i u tom slučaju bismo odbacili nultu hipotezu o jednakosti poslednjih $p - q$ korena.

Raspodela test statistike LR je komplikovana, ali možemo koristiti statistiku $-2 \ln(LR)$ koja, aproksimativno, ima χ^2 raspodelu (ako je tačna nulta hipoteza) sa $v = \frac{1}{2}(p - q + 2)(p - q - 1)$ stepeni slobode, odnosno

$$n \left[(p - q) \ln(\bar{l}) - \sum_{k=q+1}^p \ln(l_k) \right] \sim \chi_v^2 \quad (2.31)$$

gde je $\bar{l}$ uzoračka sredina poslednjih $p - q$ korena, odnosno

$$\bar{l} = \sum_{k=q+1}^p \frac{l_k}{p - q}$$

Pri tome može biti izvršena modifikacija aproksimacije u cilju njenog poboljšanja i to na sledeći način

$$\left[n - \frac{2p + 11}{6} \right] \left[(p - q) \ln(\bar{l}) - \sum_{k=q+1}^p \ln(l_k) \right] \sim \chi_{v;\alpha}^2$$

Znači da se nulta hipoteza o jednakosti poslednjih $p - q$ karakterističnih korena odbacuje na nivou značajnosti α , ako važi

$$\left[n - \frac{2p + 11}{6} \right] \left[(p - q) \ln(\bar{l}) - \sum_{k=q+1}^p \ln(l_k) \right] \geq \chi_{v;\alpha}^2$$

Postoje modifikacije aproksimacije takve da se testira jednakost svakog od podskupa $p - q$ uzastopnih karakterističnih korena, ne nužno najmanjih ili da se testira da li poslednjih $p - q$ korena prati linearni trend.

¹⁶Otuda oznaka za testove ovog tipa - LR testovi.

Testovi karakterističnih korena mogu biti od pomoći u izboru broja glavnih komponenti. Međutim, i oni kao i intervali poverenja sugerišu značajnost velikog broja karakterističnih korena, a samim tim i glavnih komponenti. Stoga se pri izboru broja komponenti češće biraju drugi kriterijumi, o kojim će kasnije biti više rečeno.

Svi rezultati dobijeni u ovom i prethodnom delu su zasnovana na kovarijansnoj, a ne korelacionoj, matrici. Generalno, zaključci na osnovu korelacione matrice su mnogo komplikovaniji. Na primer, asimptotska raspodela test statistike iz 2.31, za korelacionu matricu, je χ^2 raspodela sa ograničenjima.

Ako analizu glavnih komponenti zasnivamo na korelacionoj matrici, tada izložena test statistika za testiranje jednakosti poslednjih $p - q$ karakterističnih korena kovarijansne matrice nema svog parnjaka. U ovom slučaju koristimo test o nezavisnosti promenljivih, pa je nulta hipoteza $H_0 : \rho = I$, gdje je ρ korelaciona matrica originalnih promenljivih. Drugim rečima, testiramo hipotezu da su karakteristični koreni međusobno jednaki. Test statistika

$$-\left[n - 1 - \frac{2p + 5}{6}\right] \ln |S|$$

ima asimptotsku χ^2 raspodelu sa $\frac{1}{2p(p-1)}$ stepeni slobode, gde je sa S označena uzoračka korelaciona matrica.

Praktični postupak primene navedenog testa je sledeći. Prvo se testira hipoteza da su svi karakteristični koreni jednaki međusobno odnosno $q = 0$. Ovo je poznata hipoteza sferičnosti. Ako se odbaci ova hipoteza, postavlja se nova hipoteza prema kojoj su svi karakteristični koreni, osim prvog, međusobno jednaki odnosno $q = 1$. Ako se odbaci ova hipoteza, postupak testiranja nastavljamo, ali sada nulta hipoteza sugeriše da su svi karakteristični koreni, osim prva dva, međusobno jednaki tj. $q = 2$. Odbacivanjem ove nulte hipoteze nastavljamo postupak testiranja sve dok se ne prihvati hipoteza o jednakosti poslednjih $p - q$ karakterističnih korena.

Svi ovi testovi imaju relativna ograničenja kad treba da se koriste u praksi. Ne samo što je većina testova asimptotska i aproksimativna, nego se često oslanjamo na pretpostavku o višedimenzionalnoj normalnosti. Takođe, često nije moguće definisati hipotezu za određeni test. Zbog toga se analiza glavnih komponenti više koristi za analizu i opis podataka, nego za testiranje i verifikovanje hipoteza.

2.7 Izbor broja glavnih komponenti

Jedan od ciljeva analize glavnih komponenti je redukcija početnog skupa podataka. Umesto velikog broja promenljivih p u daljoj analizi koristimo manji broj glavnih komponenti m koje u najvećoj mogućoj meri treba da obuhvate varijansu polaznog skupa podataka. Vrlo je važno odrediti broj glavnih komponenti bez velikog gubitka informacija. Postoje mnogi pristupi za određivanje broja glavnih komponenti koje ćemo zadržati u daljoj analizi. Sada izložimo manje formalne postupke, koji su našli mesto u brojnim primenama ovog metoda. Ti postupci su:

- kumulativni procenat totalne varijanse
- veličina varijanse - Kajzerov kriterijum
- grafički prikaz vrednosti karakterističnih korena - "scree test"
- geometrijska sredina karakterističnih korena

Prvi, možda najočigledniji, pristup polazi od fiksiranja kumulativnog procenta ukupne varijanse objašnjene izdvojenim skupom glavnih komponenti. U zavisnosti od problema se bira kumulativni procenat, npr. 80% ili 90% ukupne varijanse, pa se povećava broj glavnih komponenti dok se ne postigne unapred zadata granična vrednost kriterijuma. Broj glavnih komponenti koje zadržavamo je najmanja vrednost m za koju je fiksirani postotak prekoračen. Glavne komponente su birane tako da imaju najveću moguću varijansu i varijansa k -te glavne komponente je l_k . Dalje, $\sum_{k=1}^p l_k = \sum_{j=1}^p s_{jj}$ odnosno suma varijanse glavnih komponenti je jednaka sumi varijanse elemenata vektora x . Definicija postotka varijanse objašnjene sa prvih m glavnih komponenti je

$$t_m = 100 \frac{\sum_{k=1}^m l_k}{\sum_{j=1}^p s_{jj}} = 100 \frac{\sum_{k=1}^m l_k}{\sum_{k=1}^p l_k}$$

što postaje

$$t_m = \frac{100}{p} \sum_{k=1}^m l_k$$

u slučaju da imamo korelacionu matricu.

Birajući t^* negde između 70% i 90% i zadržavajući m komponenti, gde je m najmanji broj za koji važi $t_m > t^*$, obezbeđujemo da u praksi sa prvih m glavnih komponenti ne gubimo mnogo informacija iz vektora x . Vrednost t^* postaje manja kako se povećava p ili broj opservacija n . Najbolja vrednost za t^* je između 70% i 90%. Vrednost veća od 90% se koristi kada jedna ili dve glavne komponente reprezentuju očigledan izvor varijacije. Kada je p veoma veliko biranje vrednosti m prema fiksiranoj vrednosti od 70% nam daje veliku vrednost za m , što je nepraktično.

Košćenje ovog pravila je, na neki način, ekvivalentno traženju spektralne dekompozicije kovarijansne ili korelacione matrice S ili SVD dekompozicije za datu matricu X , kojim smo se bavili u prethodnim poglavljima. U oba slučaja, odlučivanje koliko članova uključujemo u dekompoziciju da bismo dobili dobru aproksimaciju za S ili X je blisko povezano sa određivanjem t_m . Bilo je mnogo pokušaja da se odredi raspodela za t_m i, samim tim, dobije formalna procedura za izbor glavnih komponenti m , zasnovanih na t_m .

Očigledna je subjektivnost ovog načina određivanja broja glavnih komponenti, jer se on određuje na osnovu proizvoljno fiksirane vrednosti kriterijuma kumulativnog procenta objašnjene varijanse.

Prethodni princip se koristi za i glavne komponente dobijene na osnovu kovarijansne matrice i za one dobijene na osnovu korelacione matrice. Drugi princip se koristi za glavne komponente dobijene pomoću korelacione matrice, ali može da se prilagodi za neke tipove kovarijansnih matrica.

Ideja ovog principa, u slučaju korelacione matrice, je da su svi elementi vektora x nezavisni i tada su glavne komponente iste kao originalne promenljive i imaju jediničnu varijansu. Stoga, svaka komponenta koja ima varijansu manju od 1 sadrži manje informacija nego originalne promenljive, pa je ne treba zadržati u daljoj analizi. Ovo pravilo se naziva Kajzerov kriterijum (eng. Kaiser's rule). Pomoću njega zadržavamo samo one glavne komponente čija varijansa l_k je veća od 1.

Ako imamo podatke sa grupama promenljivih takvim da je korelacija unutar grupa visoka, a između grupa niska, onda postoji jedna glavna komponenta povezana sa svakom grupom promenljivih čija varijansa je veća od 1, dok je svaka druga komponenta povezana sa grupama promenljivih čija varijansa je manja od 1. Po ovom principu ćemo zadržati jednu i samo jednu glavnu komponentu povezanu sa svakom grupom promenljivih tog tipa.

Osim ovog intuitivnog objašnjenja za princip, postoji još razloga zašto je granica za prihvatanje glavnih komponenti u daljoj analizi jednaka $l_k = 1$.

Posmatrajmo promenljivu koja je manje ili više nezavisna od ostalih promenljivih u populaciji. Sa takvom promenljivom u uzorku će $p - 1$ glavna komponenta imati male koeficijente, a biće dominantna jedna komponenta čija varijansa će biti blizu jedinice ako posmatramo korelacionu matricu. Kao promenljivu koja obezbeđuje nezavisne informacije od ostalih promenljivih, ne bi bilo pametno isključiti je iz dalje analize. Ako koristimo Krajzerov kriterijum ili ako je $l_k < 1$ onda ćemo isključiti tu promenljivu iz dalje analize. Zbog toga je preporučljivo koristiti novu granicu l^* manju od jedan. Neki autori smatraju da je $l^* = 0.7$ bolja kao granica za odbacivanje i prihvatanje glavnih komponenti.

Ovo pravilo važi za korelacione matrice, ali se lako može prilagoditi kovarijansnim matricama i to tako što će granica za zadržavanje glavnih komponenti biti prosečna vrednost karakterističnih korena $\bar{l}$ ili $l^* = 0.7\bar{l}$.

Alternativni način za određivanje granice za ovaj princip je sledeći model. Imamo štap jedinične dužine, koji na slučajan način lomimo na p delova. Tada se može pokazati da je očekivana dužina k -tog najvećeg dela

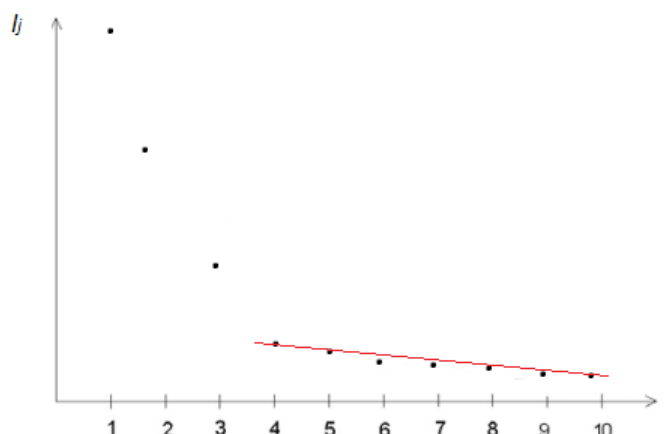
$$l_k^* = \frac{1}{p} \sum_{j=k}^p \frac{1}{j}$$

Jedan način da odlučimo da li je proporcija varijanse objašnjena k -tom glavnom komponentom dovoljno velika da zadržimo tu komponentu je da je uporedimo sa proporcijom u l_k^* . Glavna komponenta za koju proporcija prevazilazi l_k^* se zadržava, inače se odbacuje.

Treći pristup se zasniva na grafičkom prikazu vrednosti karakterističnih korena prema njihovom rednom broju. Ovaj dijagram se zove "scree test". Predložio ga je Catell 1966. godine. Prelom ili lakat na krivoj se određuje tako što se postavi lenjir uz poslednje vrednosti karakterističnih korena proveravajući da li one leže na pravoj liniji. Broj glavnih komponenti određujemo tako što uočavamo tačku nakon koje pomenuta prava linija ima prelom, pri čemu se krećemo od većeg ka manjem rednom broju glavne komponente. Broj glavnih komponenti

predstavlja upravo redni broj glavne komponente čija vrednost karakterističnog korena kao poslednja leži na pravoj liniji. Scree test nije od pomoći ukoliko na grafikonu nema očiglednog preloma ili ih ima više od jednog.

Na slici 2.2 je prikazan primer kad je $p=10$, četvrta glavna komponenta leži na crvenoj manje više pravoj liniji, ostale glavne komponente (treća, druga i prva) imaju karakteristične korene koji ne leže na crvenoj pravoj. To znači da scree test sugerise da se zadrže prve četiri komponente.



Slika 2.2: Scree test

Ovaj princip, primenjen na glavnim komponentama, koristi kao kriterijum razliku $l_{k-1} - l_k$ odnosno formalna numerička granica nije $l_{k-1} - l_k$, nego posmatramo kada razlika $l_{k-1} - l_k$ prestaje biti velika, što zavisi od apsolutne vrednosti $l_{k-1} - l_k$ tj. posmatramo relativne vrednosti od $l_{k-1} - l_k$ i $l_k - l_{k+1}$. Za ovaj princip je teško je definisati formalno numeričko pravilo i zbog toga se koristi isključivo grafički.

U opštem slučaju se pokazalo da je za $p \geq 20$ Kajzerov kriterijum restriktivan tj. uključuje suviše mali broj glavnih komponenti. A nasuprot njemu, scree test zadržava veliki broj glavnih komponenti. Zato se sugerise nalaženje kompromisa za broj glavnih komponenti koje treba uključiti u dalju analizu na osnovu ovih principa.

Prema sledećem pristupu kao kriterijum izbora koristi se geometrijska sredina. Generalizovana varijansa (determinanta matrice) jednaka je proizvodu karakterističnih korena odnosno $\prod_{j=1}^p l_j$. Ako dobijenu vrednost dignemo na stepen $\frac{1}{p}$ dobićemo geometrijsku sredinu karakterističnih korena. Dakle, prosečna generalizovana varijansa je data geometrijskom sredinom karakterističnih korena, pa prema ovom pristupu zadržavamo one glavne komponente čiji karakteristični koren je veći od geometrijske sredine svih karakterističnih korena.

Odgovor na pitanje o broju glavnih komponenti koje ćemo zadržati u analizi možemo potražiti u formalnom postupku testiranja značajnosti poslednjih karakterističnih korena.

U delu kada smo testirali hipoteze smo koristili Bartlettov test. Nulta hipoteza ovog testa glasi $H_{0,q} : \lambda_{q+1} = \lambda_{q+2} = \dots = \lambda_p$ i testiramo je protiv alternativne hipoteze da su bar dva, od poslednjih $p - q$, korena različita. Koristeći ovaj test za različite vrednosti q , dobijamo glavne komponente koje objašnjavaju znatnu količinu varijacije i one koje su šum. Ako je m broj glavnih komponenti koje zadržavamo u daljoj analizi i koje nisu šum, tada sekvencijalno koristimo ovaj test da nađemo m . Hipotezu $H_{0,p-2}$ prvo testiramo, tj. testiramo da li važi

da je $\lambda_{p-1} = \lambda_p$, i ako ova hipoteza nije odbačena onda testiramo hipotezu $H_{0,p-3}$. Ako ne odbacimo hipotezu $H_{0,p-3}$, onda testiramo hipotezu $H_{0,p-4}$ i tako nastavljamo sve dok se ne odbaci neka hipoteza $H_{0,q}$ tj. kad dobijemo da ne važi $q = q^*$. Tada je broj m jednak sa $q^* + 1$. Ovaj pristup ima nekoliko mana. Prva je da je test statistika zasnovana na pretpostavci normalnosti za vektor x i da su, čak i tad, rezultati aproksimativni. Druga mana je da ukupni nivo značajnosti testova niza nije isti kao idividualni nivo svakog testa, a ni nivo značajnosti se ne može lako utvrditi jer broj testova koje treba da sprovedemo nije fiksiran, a testovi nisu međusobno nezavisni. Najveća mana ovog pristupa se ogleda u zadržavanju suviše velikog broja glavnih komponenti, pa nam on nije dobar u praksi.

Postoji još jedan pristup za određivanje koliko komponenti treba da zadržimo u analizi, koji se naziva parcijalna korelacija. Za glavne komponente dobijene na osnovu korelacione matrice, parcijalna korelacija između p promenljivih nam daje prvih nekoliko glavnih komponenti koje treba da zadržimo u daljoj analizi. Kriterijum zasnovan na prosečnoj vrednosti kvadrata parcijalne korelacije odnosno

$$V = \sum_{\substack{i=1 \\ i \neq j}}^p \sum_{j=1}^p \frac{(r_{ij}^*)^2}{p(p-1)}$$

gde je r_{ij} parcijalna korelacija između i -te i j -te promenljive, nam daje prvih m glavnih komponenti. Statistički r_{ij} meri jačinu linearne veze između i -te i j -te promenljive. Optimalna vrednost za m odgovara minimalnoj vrednosti kriterijuma.

2.8 Rotacija i interpretacija glavnih komponenti

Kao što je ranije rečeno, glavne komponente su korisne ako mogu jednostavno da se interpretiraju. Definisali smo ih kao linearnu kombinaciju svih originalnih promenljivih, a to znači da interpretacija nije uvek laka. Jedan od načina pomoću kojih možemo olakšati interpretaciju je rotacija komponenti. Ovaj pristup nam može obezbediti jednostavnije tumačenje i to u nekim slučajevima. Treba napomenuti da ovaj pristup ima mnogo nedostataka.

Glavne komponente predstavljaju linearnu funkciju svih p originalnih promenljivih. Ako su koeficijenti glavnih komponenti slične veličine ili je nekoliko koeficijenata veliko, a ostali su mali, onda je lako interpretirati komponente.

Pretpostavimo da prvih m komponenti opisuje većinu varijacije p -dimenzionalnog skupa podataka. Važnije je jednostavno interpretirati m -dimenzionalni prostor definisan komponentama m , nego svaku komponentu pojedinačno. Način na koji to postizemo je rotiranjem osa u m -dimenzionalnom prostoru tako da što više moguće pojednostavimo interpretaciju osa.

Pretpostavimo da je matrica A_m dimenzija $p \times m$ čija je k -ta kolona vektor koeficijenata k -te glavne komponente. Koeficijenti ortogonalno rotiranih glavnih komponenti predstavljaju kolone matrice B_m , gde je $B_m = A_m T$, a matrica T je ortogonalna matrica dimenzija $m \times m$. Primenjeni postupak ortogonalne transformacije matrice, u geometrijskom smislu, predstavlja rotaciju koordinatnih osa za izvestan ugao, pri čemu ose i dalje ostaju ortogonalne. Ortogonalnu transformaciju matrice A_m i tim impliciranu ortogonalnu transformaciju komponenti nazivamo ortogonalna rotacija glavnih komponenti. Napuštanjem zahteva da rotirane komponente moraju biti međusobno ortogonalne nastala je kosa rotacija glavnih komponenti. U opštem slučaju, postupak rotacije komponenti primenjujemo u cilju dobijanja matrice B_m koja će olakšati interpretaciju komponenti. Pomoću kriterijuma koji se naziva jednostavna struktura se opredeljujemo za izbor ugla kojim ćemo rotirati faktore.

Jednostavna struktura predstavlja jedan od kriterijuma koji se koristi u analizi da bismo lakše interpretirali glavne komponente. Kriterijum je definisao Thurstone 1945.godine. Kriterijumom pokušavamo da postignemo mali broj velikih vrednosti i veliki broj malih vrednosti u matrici $\mathbf{B}_m$. Za matricu $\mathbf{B}_m$ kažemo da poseduje jednostavnu strukturu, ako važi sledeće:

- svaka njena vrsta odnosno promenljiva ima barem jednu nultu vrednost koeficijenta
- svaka njena kolona odnosno komponenta ima bar m nultih vrednosti koeficijenata
- svaki dve kolone sadrže nekoliko vrsta čiji koeficijenti su jednaki nuli u jednoj od posmatranih kolona, a različite od nule u ostalim kolonama
- kada je $m \geq 4$ svake dve kolone sadrže veći broj vrsta sa nultim koeficijenata u obe posmatrane kolone
- svake dve kolone imaju mali broj koeficijenata koji su različiti od nule u svakoj od posmatranih kolona

Rotacijom komponenti želimo da dobijemo novu, rotiranu matricu $\mathbf{B}_m$ sa navedenim osobinama jednostavne strukture, jer se tim olakšava interpretacija.

Svi metodi ortogonalne rotacije koriste ortogonalnu matricu kojom transformišu matricu koeficijenata glavnih komponenti $\mathbf{A}_m$. Novodobijena matrica koeficijenata komponenti $\mathbf{B}_m$ bi trebalo da u većoj meri izađe u susret zahtevanim osobinama jednostavne strukture. Metodi ortogonalne rotacije se međusobno razlikuju prema načinu definisanja analitičke mere jednostavnosti strukture. Određivanje ortogonalne matrice se vrši maksimiziranjem izabrane analitičke mere. Neke od ortogonalnih metoda su:

- varimax - maksimizuje sumu varijanse kvadrata koeficijenata
- equamax - unapređuje varimax metod
- quartimax - maksimizuje sumu koeficijenata podignutih na četvrti stepen
- orthomax

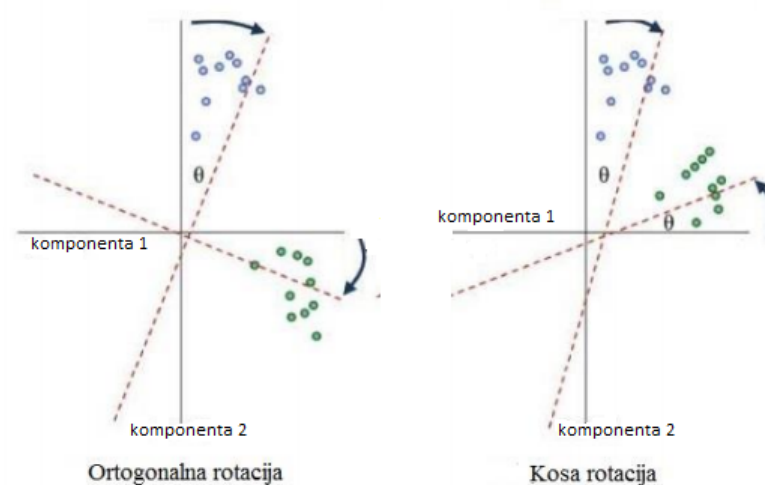
Najčešće korišćeni metod je Kajzerov varimax. U većini statističkih softvera je varimax kriterijum podešen kao default kriterijum.

Napuštanjem zahteva za ortogonalnošću komponenti dolazimo do metoda koje pri rotaciji dozvoljavaju da rotirani faktori zaklapaju ugao različit od 90° . Metode kose rotacije su zasnovane na različitim kriterijumima. Njihovim korišćenjem se dobijaju rešenja koja korespondiraju jednostavnoj strukturi, a pri tome koeficijenti nisu suviše korelisani međusobno. Lista metoda kose rotacije je sledeća:

binormamin, biquartimin, covarimin, direct oblimin, indirect oblimin, maxplane, oblinorm, oblimax, obliquimax, optres, orthoblique, orthotran, promax, quartimin i tandem criteria.

Dve najvažnije kose metode su promax i oblimin. Promax predstavlja proširenje metoda varimax. Oblimin je veoma fleksibilan zbog svog parametra koji kad je jednak nuli daje najviše resenja, a kad je jednak jedinici postaje covarimin metod.

Na slici 2.3 je prikazano kako izgleda ortogonalna i kosa rotacija glavnih komponenti.



Slika 2.3: Prikaz ortogonalne i kose rotacije

U kontekstu analize glavnih komponenti, rotacija ima mnogo nedostataka i to:

- Postoji mnogo različitih metoda rotacije, čak, lista od 19 metoda. Najčešće ih biramo proizvoljno. U kompjuterskim paketima je kao default vrednost postavljen varimax metod. Različiti metodi daju različite rezultate, što nije dobro za analizu.
- Analiza glavnih komponenti uspešno maksimizira varijansu. Kada izvršimo rotaciju, ukupna varijansa u rotiranom m -dimenzionalnom potprostoru ostaje nepromenjena odnosno i dalje je maksimum koji se može postići, ali je ravnomerno raspoređena među rotiranim komponentama. To znači da će informacije o potencijalno dominantnoj komponenti biti izgubljene.
- Izbor broja komponenti m može imati veliki efekat na rezultate posle rotacije.

2.9 Primena glavnih komponenti

Primena glavnih komponenti je veoma široka. Primenjuju se u biologiji, farmaciji, biomedicini, finansijama, agroekonomiji, ekologiji, zdravstvu.

Dobijene glavne komponente najčešće predstavljaju tek polazni podatak za druge metode multivarijantne analize, kao što su regresija, diskriminaciona analiza, kanonička korelaciona analiza, klaster analiza.

U istraživanjima u kojim broj promenljivih prevazilazi broj opservacija dolazi do problema sa singularitetima, pa je nužna redukcija skupa promenljivih koju vršimo korišćenjem glavnih komponenti. Npr. u diskriminacionoj analizi¹⁷ umesto originalnog skupa podataka koristimo izvestan broj glavnih komponenti da bismo formirali funkcije za razdvajanje grupa. Ukoliko prve dve komponente imaju visok udeo u varijansi originalnog skupa, mogu biti korišćene u dvodimenzionalnom grafičkom prikazu podataka da bismo ilustrovali međusobno razdvajanje grupa. Problem koji se javlja kad koristimo glavne komponente u diskriminacionoj analizi je taj što prve glavne komponente ne moraju biti dobri diskriminatori između grupa, tj. u opštem slučaju nema garancije da će razdvajanje između grupa biti u pravcu glavnih komponenti sa visokom varijansom.

U kanoničkoj korelacionoj analizi¹⁸ određujemo glavne komponente zasebno za svaki skup originalnih promenljivih i potom se analiza obavi na osnovu dobijenih glavnih komponenti dveju grupa. Ako se zadrže sve glavne komponente rezultati analize kanoničke korelacije se lakše mogu interpretirati. Ukoliko u analizi zadržimo nekoliko glavnih komponenti, javlja se problem i to da linearna kombinacija jednog skupa promenljivih koje su visoko korelisane sa linearnom kombinacijom drugog skupa promenljivih ne moraju nužno biti sadržane u prvim nekoliko glavnih komponenti tog originalnog skupa promenljivih.

Primena komponenti u regresiji

Pojava visoke linearne zavisnosti između regresora¹⁹ u višestrukoj regresiji se naziva multikolinearnost. Parametri regresije su neprecizno ocenjeni kada je problem multikolinearnosti znatno izražen. Ovaj problem možemo rešiti korišćenjem glavnih komponenti u regresijskoj analizi.

Na osnovu skupa nezavisnih promenljivih formiraju se glavne komponente, a zatim se obavlja regresiona analiza korišćenjem glavnih komponenti kao regresora. Ako u regresionom modelu zadržimo sve glavne komponente, onda nismo rešili problem multikolinearnosti, iz tog razloga želimo da zadržimo manji broj glavnih komponenti. Korišćenje manjeg broja glavnih komponenti od broja originalnih regresora u regresionom modelu zahteva da se prethodno izvrši izbor glavnih komponenti koje ćemo koristiti u analizi. Do sada smo izbor komponenti zasnivali na visini varijanse tj. birali smo one sa najvećom varijansom. Ovde biramo na osnovu visine korelacije glavnih promenljivih i zavisne promenljive. U regresioni model uključujemo one komponente koje imaju najveću korelaciju sa zavisnom promenljivom, jer je cilj regresije da se objasni varijacija zavisne promenljive. Koristimo iste kriterijume pri

¹⁷Diskriminaciona analiza se bavi problemom razdvajanja grupa i alokacijom opservacija u ranije definisanim grupama. Zavisna promenljiva je kategorička, a nezavisne promenljive su numeričke.

¹⁸Kanonička korelaciona analiza predstavlja uopštenje višestruke regresije. Kod izračunavanja kanoničke korelacije formiramo dve linearne kombinacije, jednu za skup nezavisnih, a drugu za skup zavisnih promenljivih. Koeficijente ovih linearnih kombinacija određujemo tako da koeficijent korelacije između njih bude maksimalan.

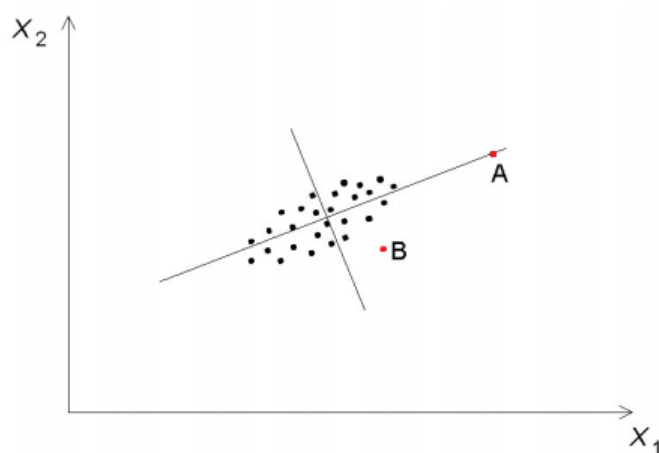
¹⁹U regresiji nezavisne promenljive još zovemo regresori.

izboru komponenti, koje smo ranije naveli u radu, ali sada ispitujeemo korelaciju glavnih promenljivih sa zavisnom promenljivom. Ocene modela u koji smo uključili manji broj glavnih komponenti su pristrasne ocene. Korišćenjem regresije glavnih komponenti očekujemo da ćemo znatno preciznije oceniti parametre modela nego pomoću regresionog modela originalnih promenljivih dobijenih metodom najmanjih kvadrata, po cenu da su te ocene u izvesnom stepenu pristrasne.

Primena komponenti u identifikaciji autlajera

Autlajeri (eng.outlier) nemaju preciznu definiciju. Pod autlajerima se podrazumevaju one opservacije koje na neki način nisu konzistentne sa preostalim delom podataka. Autlajeri mogu imati, ne nužno, drastične i disproporcionalne efekte na rezultate različitih analiza. Postupci za utvrđivanje prisustva ekstremnih vrednosti u jednodimenzionalnom slučaju mogu poslužiti za identifikaciju višedimenzionalnih autlajera. Umesto originalnih promenljivih u tim postupcima koristimo glavnu komponentu, kao jednodimenzionalnu promenljivu.

Na slici 2.4 smo prikazali dva karakteristična primera autlajera. Prikazani su kao tačke A i B . Tačkom A smo predstavili autlajer čije prisustvo možemo utvrditi na osnovu raspodele originalnih promenljivih x_1 i x_2 . Koordinate tačke A su izvan intervala varijacije ostalih tačaka tj.vrednosti prve i druge originalne promenljive. Ali je tačka A saglasna sa korelacionom strukturom ovog dvodimenzionalnog skupa podataka. U slučaju tačke B , jednodimenzionalna raspodela originalnih promenljivih ne sugerise da tačka predstavlja autlajer, ali zato tačka nije saglasna sa korelacionom strukturom podataka, pa zato predstavlja autlajer.



Slika 2.4: Primer autlajera

U ovom primeru važi, a važi i u opštem slučaju, da će prva glavna komponenta imati visoku vrednost varijanse i težiće da bude u vezi sa promenljivima koje imaju veliku varijansu i međusobno su korelisane. Autlajeri sa stanovišta prvih glavnih komponenti mogu biti otkriveni na osnovu raspodele originalnih promenljivih, kao što je slučaj sa tačkom A . Poslednje glavne komponente imaju najmanju varijansu i osetljive su na opservacije koje nisu saglasne sa korelacionom strukturom podataka, te se one ne mogu otkriti na osnovu raspodele originalnih promenljivih. U našem primeru, prva glavna komponenta otkriva autlajere tipa A , a poslednja odnosno druga komponenta autlajere tipa B . U primeni glavnih komponenti postoji interes za korišćenje poslednjih glavnih komponenti, radi identifikacije autlajera.[17]

3 Primeri u statističkim softverima

U ovom delu se posvećujemo primeni analize glavnih komponenti kroz obradu primera pomoću statističkih softverskih paketa. Razlika između statističkih paketa kada je u pitanju analiza glavnih komponenti je u tome što se kod nekih programa analiza glavnih komponenti nalazi u modulu za faktorsku analizu kao podtip faktorske analize, dok je kod drugih programa smeštena u poseban modul, kao što je slučaj kod programa Statistica. Druga razlika je u tome što su promenjeni predznaci koeficijenata u jednačinama za glavne komponente, ali to ne utiče na rezultate analize.

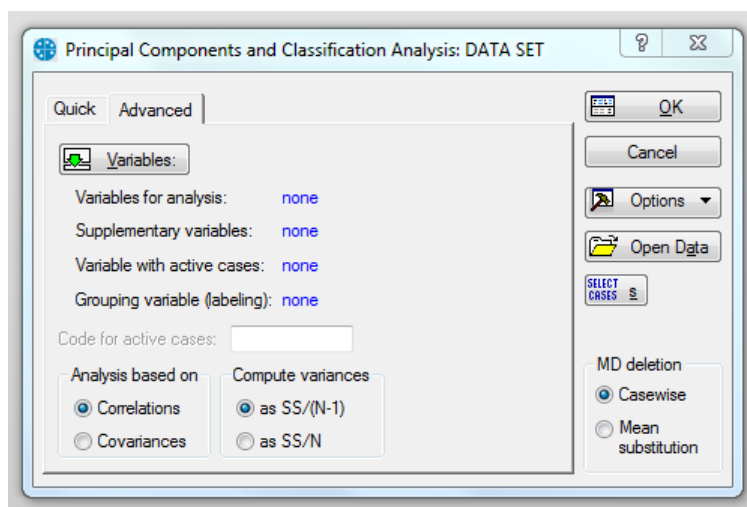
3.1 Statistica

Statistički softver Statistica je razvijen sredinom osamdesetih godina prošlog veka. Poslednja verzija Statistica 13.0 je objavljena u septembru 2015.godine. Statistica nam pruža mogućnost obrade, upravljanja, pretrage, statističke obrade i vizualizacije podataka, kao i mašinsko učenje.

Koraci za pokretanje analize u programu su sledeći:

Statistics → Multivariate\Exploratory → Principal Components & Classification

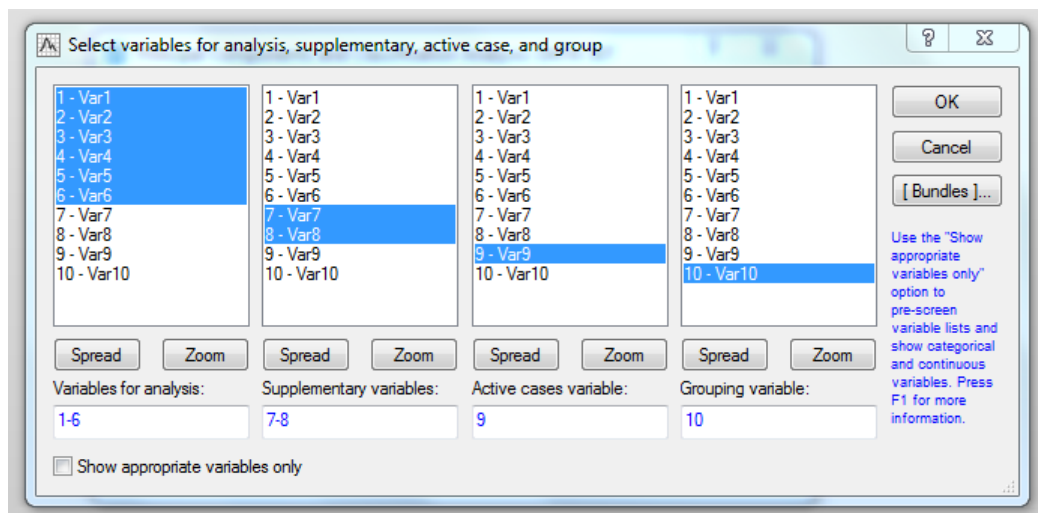
Dobijamo sledeći meni:



Slika 3.1: Izgled početnog menija za analizu u paketu Statistica

U kartici *Advanced* definišemo promenljive klikom na *Variables*. Tada nam se otvara prozor na slici 3.2. U odnosu na potrebe analize biramo kako ćemo klasifikovati promenljive. Možemo izabrati sve promenljive (na slici 3.2 deo *Variables for analysis*) i primeniti na njima tehniku ili, kao kao što je prikazano na slici, neke promenljive posmatrati kao dodatne (*Supplementary variables*), zatim možemo koristiti promenljivu pomoću koje ćemo grupisati podatke (*Grouping variable*), ili možemo odabrati da neka bude aktivna u odnosu na ostale (*Active cases variable*). Dodatne promenljive su kvantitativne promenljive i mogu da daju dodatne informacije o promenljivima korištenim u analizi. Npr. ako je neka promenljiva predstavlja rang u nekoj aktivnosti (prvo mesto, drugo mesto), postavimo je za dodatnu promenljivu i možemo da uporedimo na osnovu rangiranja rezultate za opservacije. Aktivne promenljive su kvalitativne i ako je za aktivnu promenljivu izabrana npr. promenljiva pol

(gender), onda dobijene rezultate možemo da razlikujemo u odnosu na muškarce (male) i žene (female). Tada moramo izabrati koji pol posmatramo po defaultu. To radimo posle klika na dugme OK na slici 3.2, tako što u polju *Code for active cases* upišemo npr.female. Kao što smo ranije rekli, analizu možemo zasnovati na kovarijansnoj ili korelacionoj matrici. Na slici 3.1 biramo u delu *Analysis based on* koju matricu ćemo koristiti (Correlations ili Covariances).



Slika 3.2: Defnisanje promenljivih za analiziranje u paketu Statistica

Da bismo opisali ostale korake u sprovođenju analize koristićemo podatke *Cars.sta* koji se nalaze u bazi podataka u Statistici. Predstavljaju performans, ekonomičnost potrošnje goriva i cene za različite vrste automobila. Različite vrste automobila (Audi, BMW...), ukupno njih 22, predstavljaju opservacije, a promenljive su price-cena, acceleration-ubrzanje, braking-kočenje, handling-upravljanje i mileage-kilometraža.

Da bismo dobili koeficijente korelacije, prikazane u tabeli 3.1, pratimo sledeće korake Statistics → Basic Statistics → Correlation matrices → One variable list → Select variables → OK

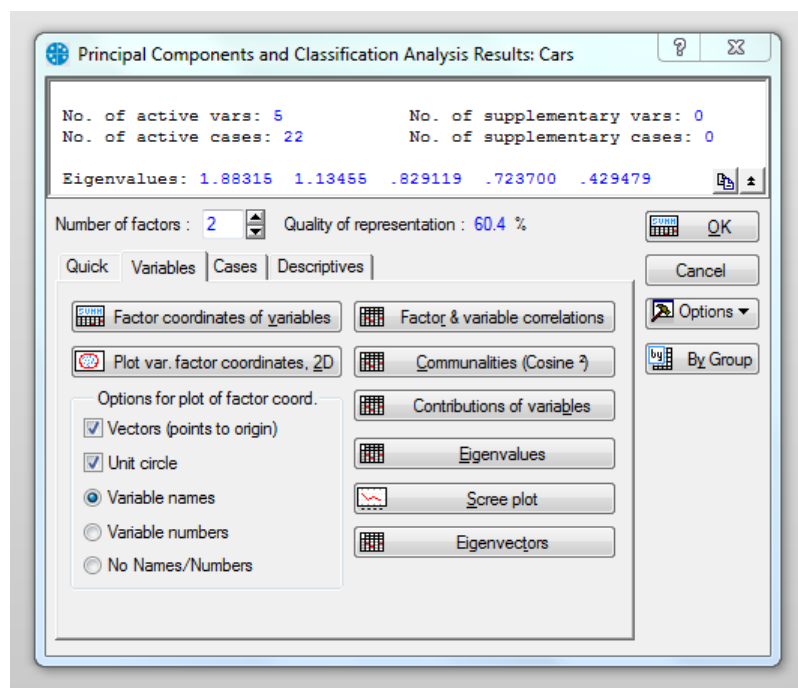
Variable	PRICE	ACCELERATION	BRAKING	HANDLING	MILEAGE
PRICE	1.000	-0.485	-0.044	0.269	-0.374
ACCELERATION	-0.485	1.000	-0.156	-0.228	0.123
BRAKING	0.044	-0.156	1.000	0.037	0.108
HANDLING	0.269	-0.228	0.037	1.000	-0.242
MILEAGE	-0.374	0.123	0.108	-0.242	1.000

Tabela 3.1: Koeficijenti korelacije promenljivih za podatke *Cars.sta*

Zbir dijagonalnih elemenata u korelacionoj matrici je jednak broju promenljivih i jednak je zbiru karakterističnih korena (eigenvalue), u našem primeru imamo 5 promenljivih i zbir dijagonalnih elemenata jednak je pet.

Kada smo svih pet promenljivih odabrali kao *Variables for analysis* dobijamo sledeći meni prikazan na slici 3.3. U gornjem delu menija možemo da vidimo broj promenljivih i broj opservacija, kao i vrednosti karakterističnih korena. Kada odlučimo koliko komponenti ćemo zadržati u daljoj analizi, onda taj broj navedemo u *Number of factors*, u našem primeru zadržavamo dve komponente pa je unesen broj dva. Odmah možemo da vidimo koliki postotak

varijacije podataka smo objasnili izborom zadržavanja tih komponenti. Taj broj vidimo u *Quality of representation*, koji u ovom primeru iznosi 60.4%. A da bismo odlučili koje komponente zadržavamo u daljoj analizi, u kartici *Variables* biramo opciju *Eigenvalue* i dobijamo tabelu 3.2. Vrednost u drugoj koloni tabele predstavlja koliko je svaka komponenta doprinela objašnjenju varijanse podataka, a u poslednjoj koloni koliko su komponente kumulativno doprinele objašnjenju varijanse podataka (prva komponenta je objasnila 37.663% varijacije, prve dve komponente su objasnile 60.354% varijacije, dok su prve tri komponente objasnile 76.936% varijacije podataka). Pomoću te tabele možemo doneti odluku koliko komponenti ćemo zadržati u daljoj analizi i to korišćenjem Kajzerovog kriterijuma koji smo ranije definisali, a koji nam kaže da zadržimo onoliko komponenti koliko ima karakterističnih korena većih od jedinice (prvi karakteristični koren iznosi 1.883, a drugi 1.134).



Slika 3.3: Meni za analiziranje promenljivih u paketu Statistica

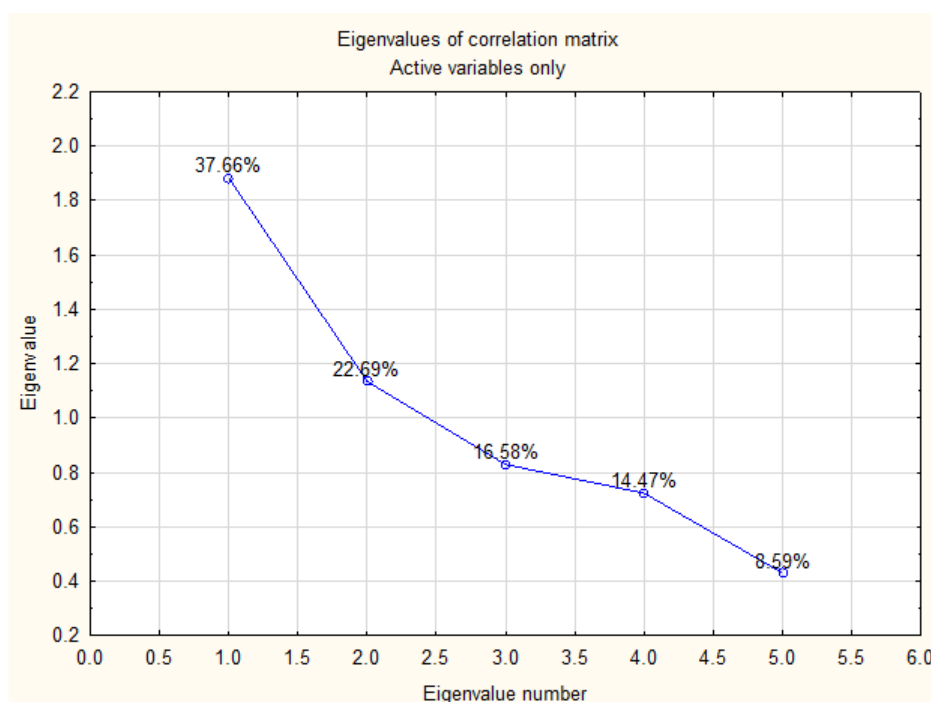
Value number	Eigenvalue	% Total Variance	Cumulative Eigenvalue	Cumulative %
1	1.883	37.663	1.883	37.663
2	1.134	22.690	3.017	60.354
3	0.829	16.582	3.846	76.936
4	0.723	14.474	4.570	91.410
5	0.429	8.589	5.000	100.000

Tabela 3.2: Karakteristični koreni korelacione matrice i procenat objašnjene varijanse podataka Cars.sta

Za donošenje odluke o broju komponenti koje ćemo zadržati, koristimo i scree test koji smo ranije definisali. Scree test dobijamo u kartici *Variables* tako što kliknemo na dugme *Scree plot*, prikazano na slici 3.4. Na slici vidimo da se prelom na grafiku nalazi u trećem karakterističnom korenu, a to nam sugeriše da treba da koristimo prve dve komponente u daljoj analizi. Koristeći Kajzerov kriterijum i scree test dobili smo isti zaključak odnosno da u daljoj analizi koristimo dve komponente.

U tabeli 3.3 se nalaze karakteristični vektori korelacione matrice²⁰, a dobijamo ih klikom na dugme *Eigenvectors* u kartici *Variables*.

Klikom na dugme *Factor coordinates of variables* u kartici *Variables* dobijamo koeficijente komponenti za odgovarajuće promenljive i to za prve dve komponente koje smo zadržali, a to je prikazano u tabeli 3.4. Kako smo analizu zasnovali na korelacionoj matrici, koeficijente možemo posmatrati kao koeficijente korelacije između odgovarajuće promenljive i komponente. Zaključujemo da nam je prva osa, koja odgovara vrednosti karakterističnog vektora 1.883, negativno korelisana sa promenljivim cena (price) i upravljanje (handling), a sa promenljivim ubrzanje (acceleration) i kilometražom (mileage) pozitivno korelisana. Druga osa, koja odgovara vrednosti karakterističnog korena 1.134, je negativno korelisana sa promenljivom kočenje (braking).



Slika 3.4: Scree test u paketu Statistica za podatke Cars.sta

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
PRICE	-0.597	0.050	0.348	0.047	0.718
ACCELERATION	0.507	0.386	-0.439	0.215	0.594
BRAKING	-0.023	-0.814	-0.305	0.466	0.155
HANDLING	-0.043	-0.008	-0.730	-0.524	0.027
MILEAGE	0.441	-0.428	0.240	-0.676	0.324

Tabela 3.3: Karakteristični vektori korelacione matrice za podatke Cars.sta

Grafički prikaz nam olakšava interpretaciju. Da bismo dobili projekciju promenljivih u novom koordinatnom sistemu (gde su ose komponente), potrebno je kliknuti na dugme *Plot var.factor coordinates.2D* u kartici *Variables* na slici 3.3. Projekcija promenljivih u novom sistemu je prikazana na slici 3.5. Primetimo da će po defaultu ovaj grafik da prikazuje jediničnu kružnicu (čekirano polje Unit circle), kao i da svi vektori moraju da leže unutar kružnice (Vectors - points to origin). Najveći vektor je jednak jedinici, u opštem slučaju, a

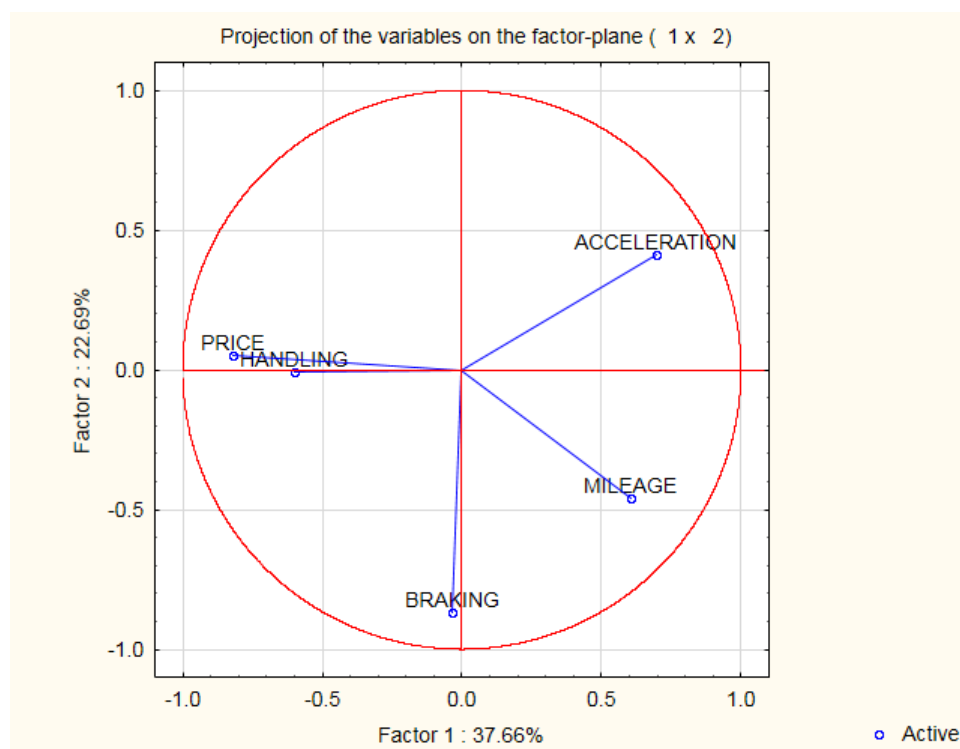
²⁰U tabeli se komponente nazivaju faktori jer paket Statistica sam generiše taj naziv.

Variable	Factor 1	Factor 2
PRICE	-0.820	0.053
ACCELERATION	0.696	0.411
BRAKING	-0.032	-0.868
HANDLING	-0.598	-0.009
MILEAGE	0.605	-0.456

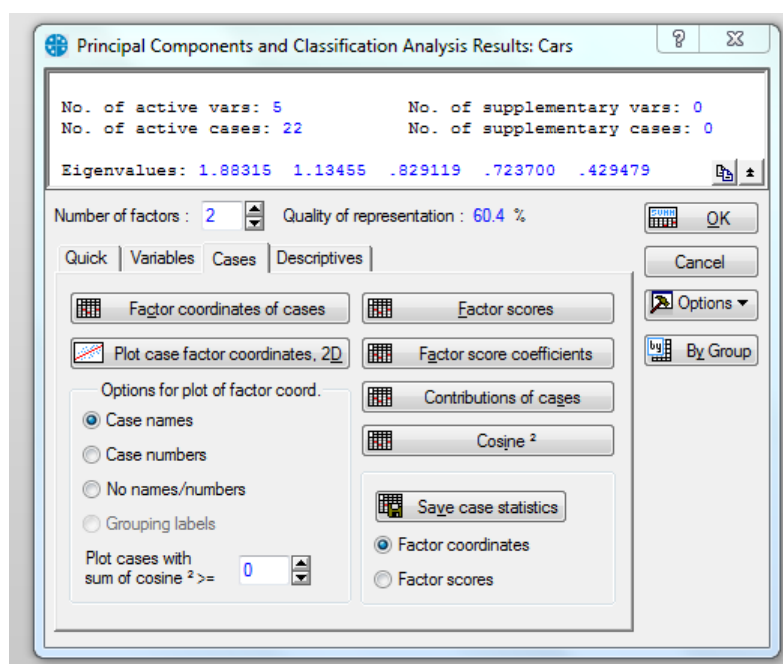
Tabela 3.4: Koeficijenti zadržanih komponenti u odnosu na promenljive za podatke Cars.sta

suma kvadrata svih vektora ne može preći jedinicu. Što su promenljive bliže kružnici, to su bitnije za posmatrane podatke.

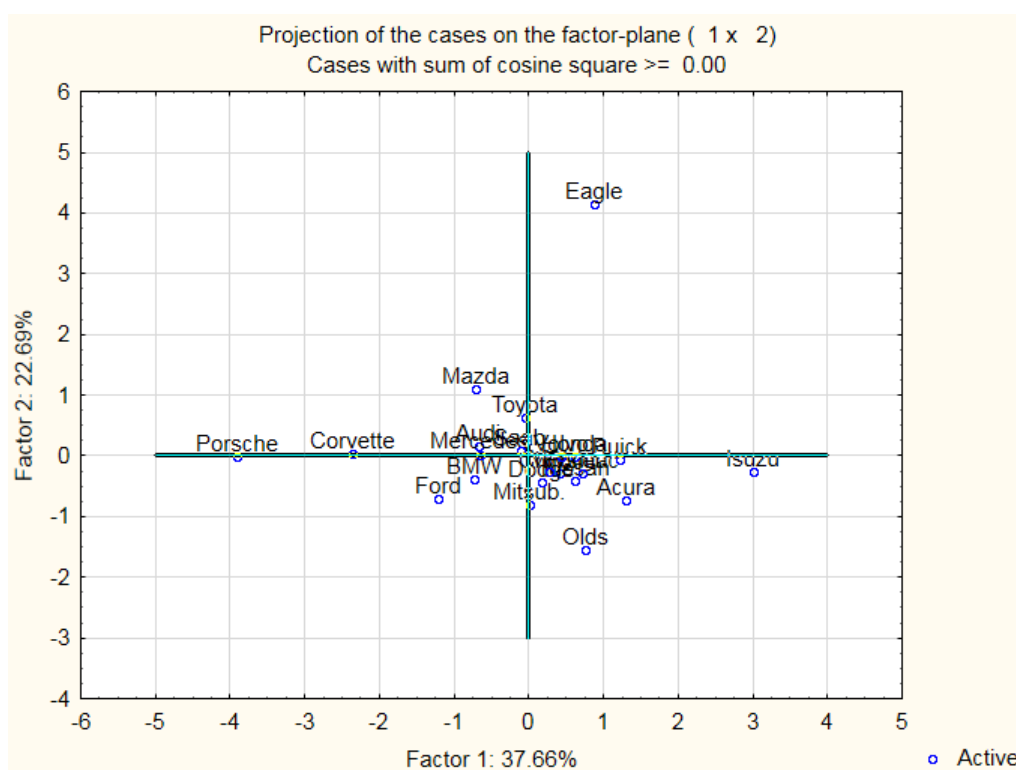
Otvaranjem kartice *Cases* na slici 3.3 dobijamo meni za analiziranje opservacija prikazan na slici 3.6. Možemo prikazati koeficijente opservacija za izabrane komponente i to klikom na *Factor coordinates of cases* na slici 3.6. Kao što smo već rekli, grafički prikaz nam olakšava interpretaciju i tumačenje, pa klikom na dugme *Plot case factor coordinates.2D* dobijamo grafik za opservacije, prikazan na slici 3.7. Na grafiku apscisa predstavlja prvu komponentu, dok druga komponenta predstavlja ordinatu. Takođe, na grafiku je prikazan naziv opservacija, u našem slučaju različite marke automobila, što često može da bude nečitljivo. Opcija prikaza naziva opservacija može se isključiti u delu *Options for plot of factor coord.*, slika 3.6, klikom na radio button *No names/numbers* i tada ćemo na grafiku imati samo tačke, a klikom na radio button *Case numbers* imaćemo tačke i redne brojeve opservacija.



Slika 3.5: Projekcija promenljivih u novom sistemu za podatke Cars.sta



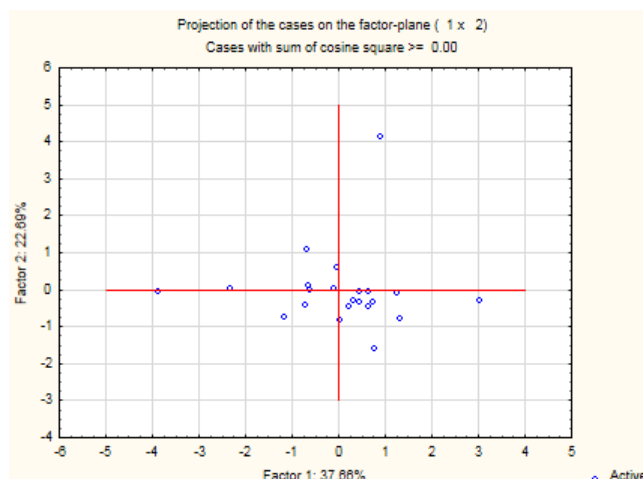
Slika 3.6: Meni za analiziranje opservacija u paketu Statistica



Slika 3.7: Projekcija opservacija u novom sistemu za podatke Cars.sta

Ovim grafičkim prikazom na slici 3.7 je obuhvaćeno 60.4% varijacije podataka. Na desnoj strani prve komponente (x-ose) se nalazi najviše opservacija odnosno različitih marki automobila. Na desnoj strani se nalaze pozitivne vrednosti koordinata pa u našem primeru promenljive ubrzanje i kilometraža definišu desnu stranu. To znači da je za većinu marki automobila bitna ekonomičnost potrošnje goriva. Sa druge strane npr. za marke Porsche i Corvette je bitna cena i performans, jer je leva strana definisana promenljivim cena i upravljanje. Za automobil marke Olds je bitna ekonomičnost potrošnje goriva podjednako kao i sistem kočenja. Za automobil marke Eagle je najbitnija brzina i ubrzanje. Na slici 3.8

je prikazana projekcija opservacija, samo ovaj put bez prikazanog naziva, da bi grafik bio jasniji i čitljiviji.



Slika 3.8: Projekcija opservacija bez prikazanog naziva opservacija za podatke *Cars.sta*

U kartici *Descriptives* na slici 3.3 možemo da izračunamo korelacionu i kovarijansnu matricu, kao i deskriptivnu statistiku. Zatim, možemo da prikazemo razne dijagrame, kao što su box & whisker, histogram, 2D i 3D scatterplots, u zavisnosti šta nam je potrebno za neku dalju analizu ili grafičke prikaze.

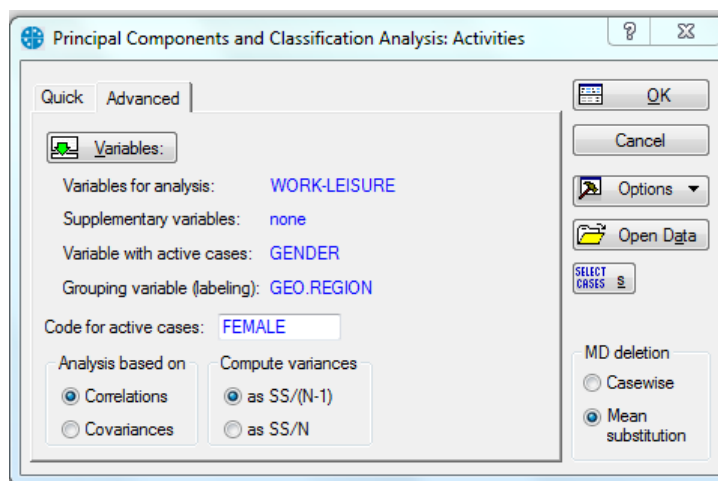
U sledećem primeru ćemo koristiti podatke *Activities.sta* koji se nalaze u bazi podataka u Statistici. Podaci predstavljaju provedene sate u okviru deset različitih aktivnosti. Aktivnosti predstavljaju naše promenljive i navedene su redom: posao (work), prevoz (transport), domaćinstvo (household), deca (children), kupovina (shopping), lična nega (personal care), obroci (meal), spavanje (sleep), gledanje tv-a (tv), slobodno vreme (leisure). Imamo još dve promenljive pol (gender) i region (geom.region), koje predstavljaju kvalitativne promenljive. Ukupno imamo 28 opservacija. Neki od podataka nedostaju (missing value), ali oni će biti zamenjeni odgovarajućim sredinama.

Kada otvorimo podatke i izaberemo analizu glavnih komponenti kako smo prethodno opisali, definišemo promenljive na sledeći način koji je prikazan na slici 3.9. Svih deset promenljivih koje predstavljaju aktivnost smo odabrali za analizu, promenljivu pol smo definisali kao promenljivu sa aktivnim opservacijama, a promenljivu region smo definisali kao promenljivu pomoću koje grupišemo podatke. To znači da posmatramo aktivnosti žena i muškaraca u odnosu na različit geografski region. Analizu smo zasnovali na korelacionoj matrici. U delu *MD deletion* klikom na *Mean substitution* zamenjujemo podatke koje nedostaju odgovarajućim sredinama.

Karakteristični koreni su prikazani u tabeli 3.5.

U daljoj analizi ćemo zadržati prve dve glavne komponente, koje ukupno objašnjavaju 71.93% varijacije podataka. U ovom primeru, po Kajzerovom kriterijumu je preporučljivo zadržavanje prve tri komponente, ali koristeći scree test vidi se jasnije da možemo koristiti dve komponente u daljoj analizi. Scree test je prikazan na slici 3.10.

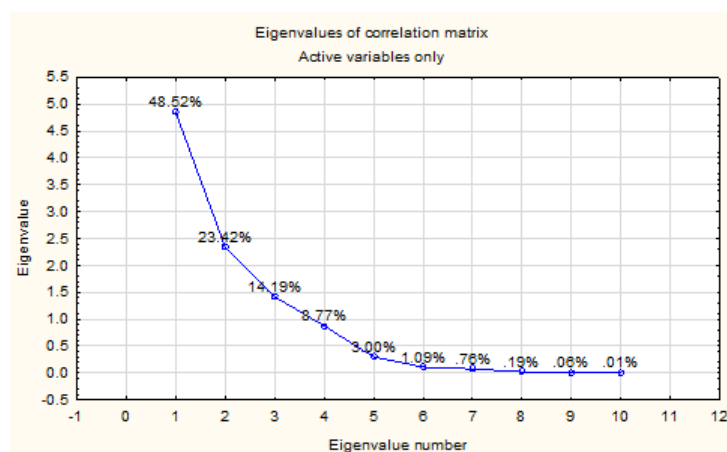
U tabeli 3.6 smo prikazali koeficijente za prve dve zadržane komponente u odnosu na odgovarajuće promenljive. Zaključujemo da nam je prva osa, koja odgovara vrednosti karakterističnog vektora 4.851, pozitivno korelisana sa promenljivim posao (work) i prevoz (transport), a sa promenljivima domaćinstvo (household), deca (children), obroci (meal) i



Slika 3.9: Definisanje promenljivih za podatke Activities.sta

Value number	Eigenvalue	% Total Variance	Cumulative Eigenvalue	Cumulative %
1	4.851	48.515	4.851	48.515
2	2.342	23.421	7.193	71.936
3	1.418	14.185	8.612	86.122
4	0.877	8.771	9.489	94.893
5	0.299	2.998	9.789	97.892
6	0.108	1.088	9.898	98.980
7	0.076	0.763	9.974	99.743
8	0.019	0.191	9.993	99.934
9	0.005	0.056	9.999	99.991
10	0.000	0.008	10.000	100.000

Tabela 3.5: Karakteristični koreni korelacione matrice i procenat objašnjene varijanse podataka Activities.sta



Slika 3.10: Scree test za podatke Activities.sta

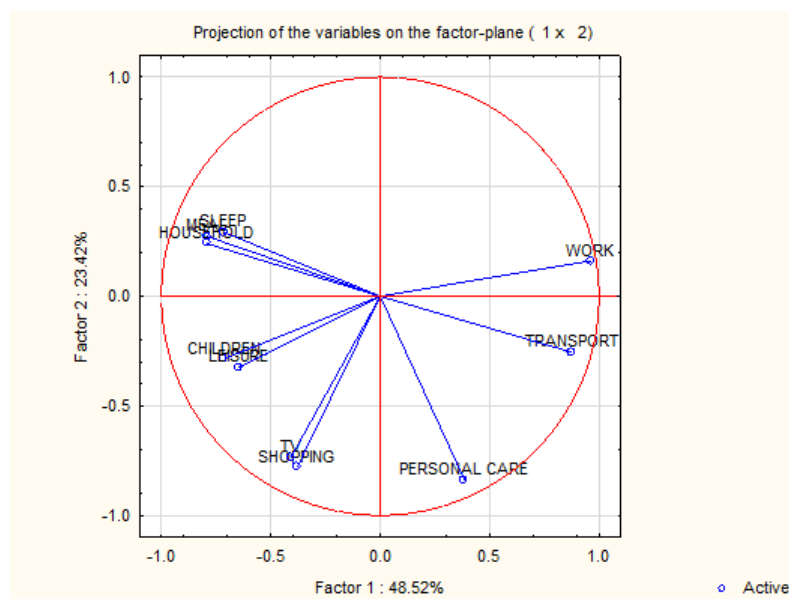
slobodno vreme (leisure) negativno korelisana. Prva komponenta u ovom primeru predstavlja aktivnosti vezane za provođenje vremena u kući i na poslu. Druga osa, koja odgovara vrednosti karakterističnog korena 2.342, je negativno korelisana sa promenljivima lična nega (personal care), kupovina (shopping) i gledanje tv-a (tv). Druga komponenta se odnosi na aktivnosti u kojima vreme trošimo na relaksaciju, opuštanje i uživanje. Zaključak je da deset

dnevnih aktivnosti možemo da posmatramo kao dve aktivnosti i to obaveze vezane za posao i kuću, te vreme potrošeno na odmor i relaksaciju.

Korelaciju promenljivih sa glavnim komponentama smo grafički prikazali na slici 3.11.

Variable	Factor 1	Factor 2
WORK	0.963	0.162
TRANSPORT	0.877	-0.256
HOUSEHOLD	-0.793	0.240
CHILDREN	-0.713	-0.258
SHOPPING	-0.378	-0.778
PERSONAL CARE	0.383	-0.836
MEAL	-0.792	0.273
SLEEP	-0.716	0.293
TV	-0.407	-0.733
LEISURE	-0.645	-0.325

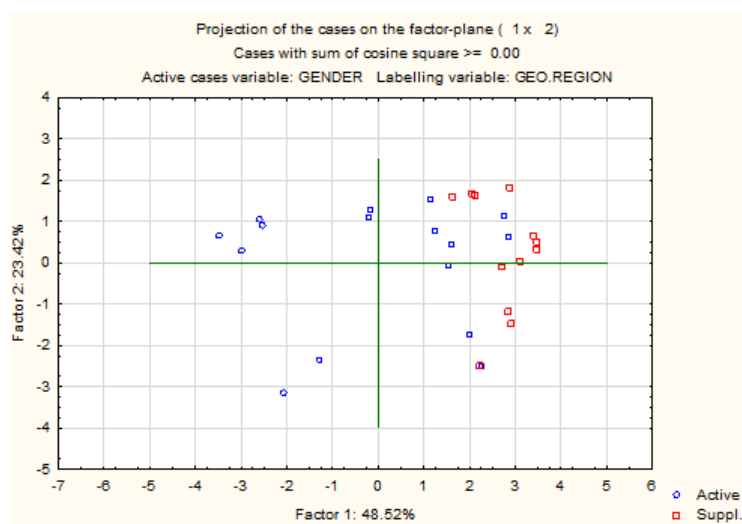
Tabela 3.6: Koeficijenti zadržanih komponenti za odgovarajuće promenljive za podatke Cars.sta



Slika 3.11: Projekcija promenljivih u novom sistemu za podatke Activities.sta

Na slici 3.12 ćemo prikazati projekciju opservacija u novom sistemu sa glavnim komponentama kao osama. Ovim grafičkim prikazom je obuhvaćeno 71.93% varijacije podataka. Plavom bojom su označene aktivne opservacije odnosno žene, a crvenom muškarci. Vidimo da su sve opservacije označene crvenom bojom nalaze na desnoj strani u odnosu na prvu osu - prvu glavnu komponentu. Promenljive posao (work) i prevoz (transport) su pozitivno korelisane sa prvom komponentom, odnosno one definišu desnu stranu ose. To znači da su dnevni aktivnosti muškaraca uglavnom vezane za posao. Kod žena su dnevne aktivnosti vezane ili za posao ili za domaćinstvo.

U ovom primeru smo deset promenljivih, deset različitih aktivnosti uspeali da svedemo na dve promenljive i to prvu koja predstavlja aktivnosti posao-kuća, a druga promenljiva

Slika 3.12: Projekcija opservacija u novom sistemu za podatke *Activities.sta*

predstavlja aktivnosti vezane za moderan način života, kupovinu-relaksaciju.

Bitno je napomenuti da su u ovom primeru promenljive korelisane više nego u prethodnom, što doprinosi boljem rezultatu analize, lakšoj interpretaciji. Koeficijenti korelacije za promenljive u primeru sa podacima *Activities.sta* su prikazani u sledećoj tabeli 3.7.

Variable	WORK	TRANSPORT	HOUSEHOLD	CHILDREN	SHOPPING	PERSONAL CARE	MEAL	SLEEP	TV	LEISURE
WORK	1.000	-	-	-	-	-	-	-	-	-
TRANSPORT	0.933	1.000	-	-	-	-	-	-	-	-
HOUSEHOLD	-0.910	-0.866	1.000	-	-	-	-	-	-	-
CHILDREN	-0.834	-0.766	0.836	1.000	-	-	-	-	-	-
SHOPPING	-0.659	-0.488	0.549	0.535	1.000	-	-	-	-	-
PERSONAL CARE	-0.032	-0.044	-0.076	-0.044	0.496	1.000	-	-	-	-
MEAL	-0.456	-0.596	0.300	0.315	-0.199	-0.247	1.000	-	-	-
SLEEP	-0.630	-0.772	0.452	0.307	0.080	-0.089	0.810	1.000	-	-
TV	0.128	0.166	-0.403	-0.110	0.051	0.115	0.265	-0.008	1.000	-
LEISURE	-0.452	-0.380	0.160	0.076	0.248	0.055	0.336	0.525	0.008	1.000

Tabela 3.7: Koeficijenti korelacije promenljivih za podatke *Activities.sta*

U sledećem primeru koristimo podatke²¹ sa jedne farme gde je popisano pet različitih telesnih dimenzija mačića. Mačići su podeljeni u dve grupe. Prvoj grupi pripadaju mačići koji su otporni na bolest (od rednog broja 1 do 10), dok drugoj grupi pripadaju oni koji nisu otporni (od rednog broja 11 do 25). Na slici 3.13 je prikazano nekoliko prvih vrsta u podacima Mačići. Prva promenljiva pod nazivom grupa predstavlja da li su mačići otporni na bolesti ili ne (0 - otporni, 1 - neotporni), a ostale promenljive predstavljaju različite telesne dimenzije.

Koeficijenti korelacije su prikazani u tabeli 3.8, jasno vidimo da su poprilično visoki, što nam sugeriše da treba da odredimo glavne komponente i izvršimo analizu.

Kada otvorimo dijalog za analiziranje komponenti treba da definišemo promenljive. Promenljive koje predstavljaju telesne dimenzije mačića definišemo kao *Variables for analysis*, a promenljivu koja razdvaja mačiće u dve grupe definišemo kao *Grouping variables*. Analizu smo zasnovali na korelacionoj matrici.

Dobijamo da su vrednosti karakterističnih korena sledeće 3.414, 1.267, 0.21, 0.093, 0.016. Prvom komponentom objašnjavamo 68.28% varijacije podataka, dok drugom objašnjavamo

²¹Podaci su iz ličnog arhiva.

Data: podaciDimenzijeMacici* (6v by 49c)

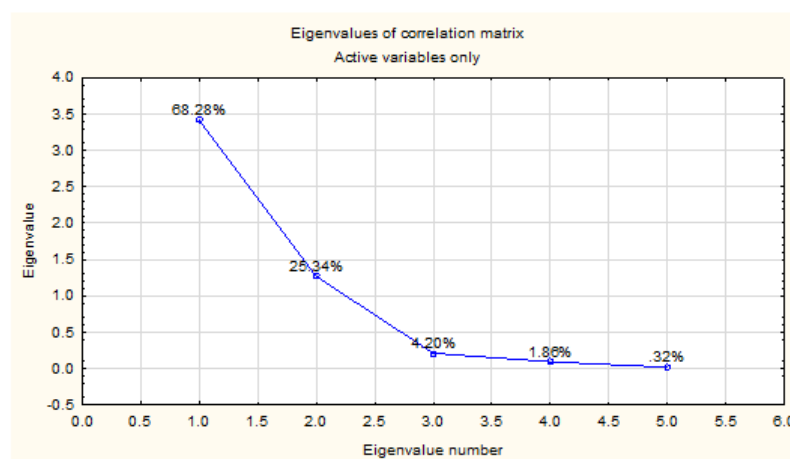
	1 grupa	2 dimenzija1	3 dimenzija2	4 dimenzija3	5 dimenzija4	6 dimenzija5				
1	0	156	245	31.6	18.5	20.5				
2	0	154	240	30.4	17.9	19.6				
3	0	153	240	31	18.4	20.6				
4	0	153	236	30.9	17.7	20.2				
5	0	155	243	31.5	18.6	20.3				
6	0	163	247	32	19	20.9				
7	0	157	238	30.9	18.4	20.2				
8	0	155	239	32.8	18.6	21.2				
9	0	164	248	32.7	19.1	21.1				
10	0	158	238	31	18.8	22				
11	0	158	240	31.3	18.6	22				
12	0	160	244	31.1	18.6	20.5				
13	0	161	246	32.3	19.3	21.8				
14	0	157	245	32	19.1	20				
15	0	157	235	31.5	18.1	19.8				
16	0	156	237	30.9	18	20.3				

Slika 3.13: Prikaz prvih nekoliko vrsta u podacima Mačići

Variable	DIMENZIJA1	DIMENZIJA2	DIMENZIJA3	DIMENZIJA4	DIMENZIJA5
DIMENZIJA1	1.000	-	-	-	-
DIMENZIJA2	0.765	1.000	-	-	-
DIMENZIJA3	0.550	0.376	1.000	-	-
DIMENZIJA4	0.437	0.249	0.925	1.000	-
DIMENZIJA5	0.443	0.222	0.901	0.980	1.000

Tabela 3.8: Koeficijenti korelacije promenljivih za primer sa mačićima

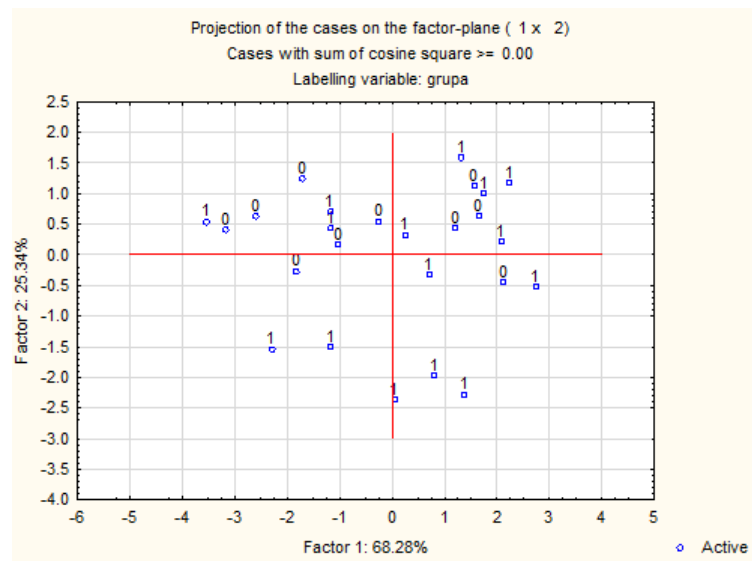
25.34% varijacije. Kajzerov kriterijum, kao i scree test prikazan na slici 3.14, sugerišu zadržavanje prve dve komponente.



Slika 3.14: Scree test za primer sa mačićima

Slika 3.15 pokazuje raspodelu dve grupe mačića (mačići iz prve grupe koji su otporni na bolesti su označeni sa 0) u koordinatnom sistemu čije ose su prve dve glavne komponente. Ovim grafičkim prikazom je obuhvaćeno 93.62% varijacije originalnih podataka. Na slici se jasno vidi kako su mačići sa ekstremnom veličinom određenih delova tela podložniji bolesti. Ovaj primer je specifičan jer prikazuje kako u nekim slučajevima glavne komponente iako nisu definisane šta tačno predstavljaju (nije jasno određene, predstavljaju na neki način zbir telesnih dimenzija), mogu da pomognu u smanjenju redundantnosti i prikazu podataka u

dvodimenzionalnom prostoru. Pomoću komponenti smo uspeli da grupišemo mačiće otporne na bolest i neotporne na bolest u odnosu na telesne dimenzije.



Slika 3.15: Projekcija opservacija u novom sistemu za primer sa mačićima

3.2 R

R je besplatan programski jezik i softversko okruženje za statističku obradu podataka. Razvijen je krajem devedesetih godina prošlog veka, a poslednjih godina je u širokoj upotrebi. R obezbeđuje širok izbor statističkih tehnika i to klasične statičke testove, linearne i nelinearne modele, analizu vremenskih serija, klaster analizu, klasifikaciju itd, kao i grafičke tehnike. Projektovan je kao programski jezik što korisnicima omogućava definisanje novih funkcija. Takođe, može biti proširen za specifične oblasti proučavanja, kroz različite pakete obezbeđene od strane korisnika.

U R-u postoji nekoliko funkcija iz različitih paketa pomoću kojih je moguće odrediti glavne komponente i to:

- `prcomp()` iz paketa `stats`
- `princomp()` iz paketa `stats`
- `PCA()` iz paketa `FactoMineR`
- `dudi.pca()` iz paketa `ade4`
- `epPCA()` iz paketa `ExPosition`
- `acp()` iz paketa `amap`

Bez obzira koju funkciju odlučili da koristimo, rezultati u analizi glavnih komponenti treba da sadrže vrednosti karakterističnih korena, tabelu glavnih komponenti, kao i tabelu koeficijenata korelacije između promenljivih i glavnih komponenti.

Funkcija `prcomp()` se nalazi u default paketu `stats`, što znači da nije potrebno ništa dodatno instalirati odnosno taj paket je po defaultu instaliran. Funkciju `prcomp()` ćemo primeniti na skup podata *Iris flower data set* koji se nalazi u R-u. Skup podataka se sastoji od 50 opservacija za svaku od tri vrste iris cvetova (*Iris setosa*, *Iris virginica*, *Iris versicolor*) odnosno ukupno 150 opservacija. Imamo četiri promenljive koje mere dužinu (`length`) i širinu (`width`) čašičnih listića (`sepals`) i latica (`petals`) u centrimetrima.

Prvi korak analizi je učitavanje podataka naredbom

```
data("iris")
```

Zatim, ukoliko želimo da prikažemo prvih nekoliko vrsta podataka, to omogućava sledeća naredba

```
head(iris)
```

Prikaz tih podataka vidimo u sledećoj tabeli:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Length	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Tabela 3.9: Prikaz prvih šest vrsta u podacima iris

Deskriptivnu statistiku, prikazanu u tabeli 3.10, dobijamo naredbom `summary(iris)`

Sepal.Length	Sepal.Width	Petal.Length	Petal.Length	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199	
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	

Tabela 3.10: Deskriptivna statistika za podatke iris

Naredba za dobijanje glavnih komponenti je `myPr = prcomp(iris[,1:4], scale=TRUE)`

Sa `myPr` smo označili objekat (dali mu naziv) u kom su sačuvani koeficijenti glavnih komponenti.

Funkcija `prcomp` predstavlja funkciju za izračunavanje komponenti, kao što smo ranije rekli. Funkcija ima dva argumenta. Prvi argument `iris[,1:4]` predstavlja skup podataka, u našem primeru četiri kolone numeričkih promenljivih, a drugi argument `scale=TRUE` služi za standardizaciju promenljivih, a postavljen je na `true` jer je default vrednost `false`.

Pozivanjem objekta `myPr` dobijamo koeficijente komponenti prikazane u tabeli 3.11.

Vidimo da je prva komponenta pozitivno korelisana sa promenljivima dužina čašičnog listića (`Sepal.Length`), dužina latica (`Petal.Length`) i širinom latica (`Petal.Width`). Druga komponenta je negativno korelisana sa promenljivom širina čašičnog lista (`Sepal.Width`).

	PC1	PC2	PC3	PC4
Sepal.Length	0.521	-0.377	0.719	0.261
Sepal.Width	-0.269	-0.923	-0.244	-0.123
Petal.Length	0.580	-0.024	-0.142	-0.801
Petal.Length	0.564	-0.066	-0.634	0.523

Tabela 3.11: Prikaz koeficijenata glavnih komponenti za podatke iris

Naredbom `summary(myPr)` dobijamo tabelu 3.12 u kojoj vidimo u kojem postotku su komponente pojedinačno i kumulativno objasnile varijaciju podataka. Prva komponenta je objasnila 72.9% varijacije podataka, dok je druga komponenta objasnila 22.8% varijacije podataka, a ukupno su prve dve komponente obuhvatile 95.8% varijacije podataka.

	PC1	PC2	PC3	PC4
Standard deviation	1.708	0.956	0.383	0.143
Proportion of Variance	0.729	0.228	0.036	0.005
Cumulative Proportion	0.729	0.958	0.994	1.000

Tabela 3.12: Prikaz proporcije objašnjene varijanse glavnim komponentama za podatke iris

Da bismo dobili karakteristične korene i vektore koristimo sledećih nekoliko naredbi:

```
scaled_data <- apply(iris[,1:4], 2, scale)
head(scaled_data)

iris.cov <- cov(scaled_data)
iris.eigen <- eigen(iris.cov)
str(iris.eigen)
```

Funkcija `apply(iris[,1:4], 2, scale)` služi za standardizaciju promenljivih, a standardizovane podatke čuvamo u objektu pod nazivom `scaled_data`.

Funkciju `head(scaled_data)` koristimo za prikaz nekoliko prvih vrsta u standardizovanim podacima.

Funkcija `cov(scaled_data)` koristimo da bismo izračunali matricu kovarijansi, koju čuvamo u objektu pod nazivom `iris.cov`.

Pomoću funkcije `eigen(iris.cov)` dobijamo karakteristične korene i vektore, koje čuvamo u objektu `iris.eigen`.

Da bismo prikazali karakteristične korene i vektore koristimo funkciju `str(iris.eigen)`, rezultat koji smo dobili je prikazan na slici 3.16. Karakteristične vektore možemo da ubacimo u matricu i to naredbom (`matrix_a <- iris.eigen$vectors[,1:2]`), a prikazani su u tabeli 3.13.

```
> str(iris.eigen)
List of 2
 $ values : num [1:4] 2.9185 0.914 0.1468 0.0207
 $ vectors: num [1:4, 1:4] 0.521 -0.269 0.58 0.565 -0.377 ...
 - attr(*, "class")= chr "eigen"
```

Slika 3.16: Prikaz karakterističnih korena i vektora za podatke iris

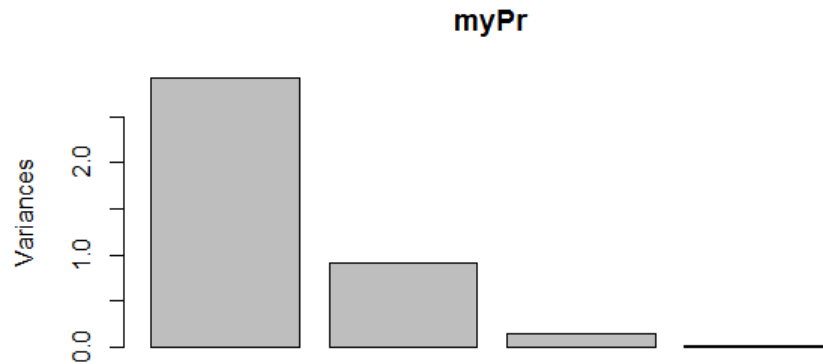
	[, 1]	[, 2]
[1,]	0.521	-0.377
[2,]	-0.269	-0.923
[3,]	0.580	-0.024
[4,]	0.564	-0.066

Tabela 3.13: Prikaz karakterističnih vektora za podatke iris

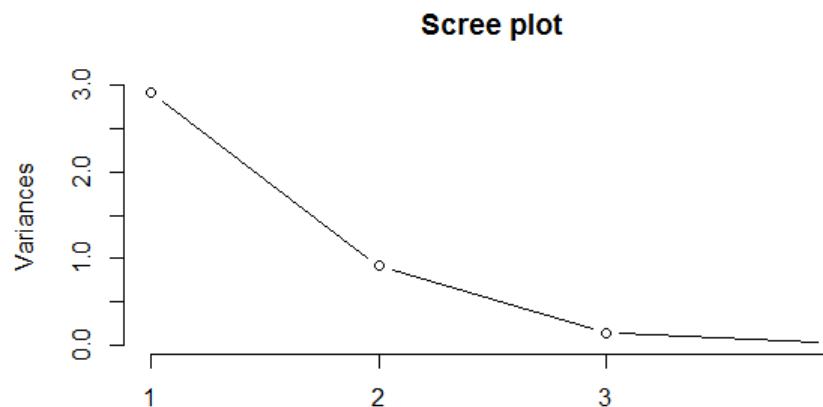
Napomena: Oznaka `[, 1]` se u R-u odnosi na prvu kolonu, a oznaka `[1, ]` na prvu vrstu.

Naredbom `plot(myPr)` dobijamo barplot koji je prikazan na slici 3.17. Sa prve dve komponente smo objasnili 95.8% varijacije podataka. Na dijagramu je grafički prikazano koliko svaka komponenta objašnjava varijaciju podataka. Jasno se vidi da prve dve komponente dominiraju. Kada u naredbu dodamo dva argumenta `plot(myPr, main = "Scree test", type = "l")`, dobijamo linije na grafiku odnosno scree test, što je prikazano na slici 3.18.

Naredbom `biplot(myPr, scale = 0)` prikazujemo projekciju opservacija u novom sistemu čije su koordinatne ose prve dve glavne komponente, što je prikazano na slici 3.19. Prvi argument naredbe su podaci, a drugi argument nam olakšava čitanje podataka. Na slici su brojevima označene opservacije, a crveni vektori predstavljaju karakteristične vektore. Vidimo da su promenljive dužina čašičnog listića, dužina i širina latica pozitivno korelisane u odnosu na prvu glavnu komponentu, a da je širina čašičnog listića negativno korelisana. U odnosu na drugu glavnu komponentu sve promenljive su negativno korelisane.



Slika 3.17: Barplot objašnjene varijacije podataka za svaku komponentu za podatke iris



Slika 3.18: Scree test za podatke iris

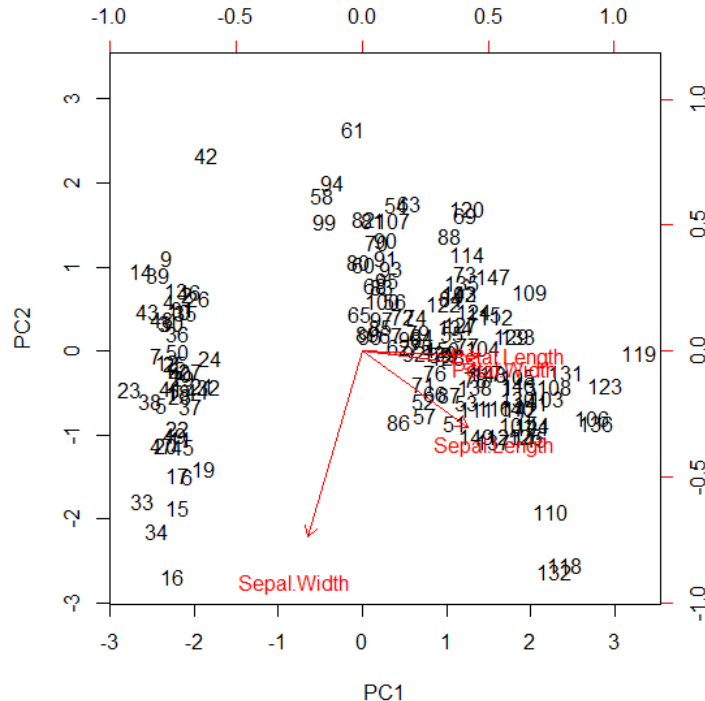
Osim naredbom `biplot(myPr, scale = 0)`, grafički prikaz projekcije opservacija u novom sistemu sa prve dve glavne komponente možemo dobiti pomoću sledećih naredbi:

```
ir.species <- iris[,5]
install.packages("devtools")
library(devtools)
install_github("ggbiplot", "vqv")

g <- ggbiplot(myPr, obs.scale = 1, var.scale = 1,
```

```
groups = ir.species, ellipse = TRUE, circle = TRUE)
g <- g + scale_color_discrete(name = "")
g <- g + theme(legend.direction = 'horizontal', legend.position = 'top')

print(g)
```



Slika 3.19: Prikaz projekcije opservacija u novom sistemu za podatke iris

Naredbom `ir.species <- iris[,5]` smo napravili objekat koji predstavlja petu promenljivu u podacima, vrste iris cvetova (setosa, versicolor, virginica).

Zatim smo instalirali paket `devtool`, a onda ga učitali.

Naredbom `install_github("ggbiplot", "vqv")` smo direktno sa GitHub-a instalirali pakete `ggbiplot` i `vqv`.

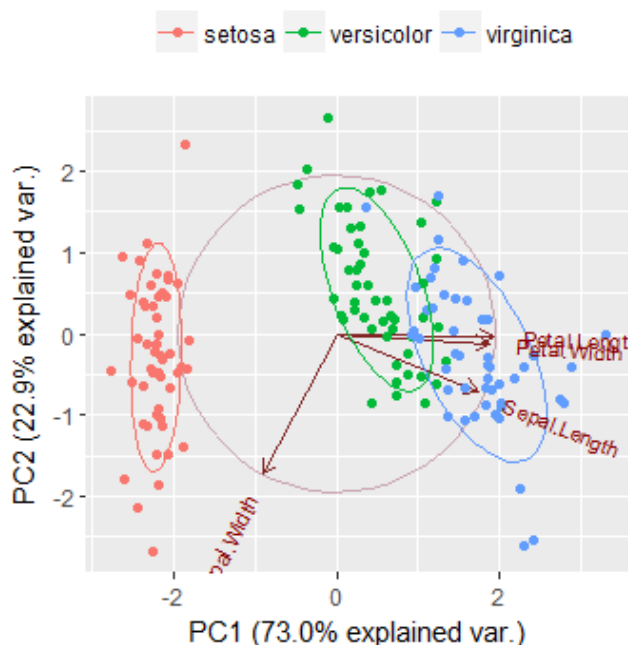
Naredbom `g <- ggbiplot(myPr, obs.scale = 1, var.scale = 1, groups = ir.species, ellipse = TRUE, circle = TRUE)` smo kreirali objekat koji smo nazvali `g`. Funkcija `ggbiplot` služi za dobijanje grafičkog prikaza. Njen prvi parametar su podaci, naše komponente, drugi parametar predstavlja standardizaciju opservacija, dok parametar `var.scale=1` predstavlja standardizaciju promenljivih. Pomoću parametra `groups=ir.species` grupišemo opservacije po promenljivoj `ir.species` koja nam predstavlja vrste iris cvetova. Parametar `ellipse=TRUE` iscrtava elipse za svaku grupu opservacija, u našem primeru za tri vrste cvetova, postavljamo ga na vrednost `TRUE` jer je default vrednost `FALSE`. Parametar `circle=TRUE` možemo da primenimo samo ako smo prethodno standardizovali promenljive i opservacije, služi da iscrtava krug u kom su podaci korelisani, a njegova default vrednost je `FALSE`.

Naredbom `g <- g + scale_color_discrete(name = "")` smo definisali boje po grupama, u našem slučaju po vrstama cvetova. A naredbom `g <- g + theme(legend.direction = 'horizontal', legend.position = 'top')` smo definisali gde će se nalaziti legenda na dijagramu.

Na kraju naredba `print(g)` služi da se prikaže objekat `g` koji smo definisali odnosno dobijamo grafički prikaz opservacija koji možemo da vidimo na slici 3.20.

Svaka tačka iz originalnog skupa podataka je prikazana u novom sistemu sa glavnim kompo-

mentama koje predstavljaju linearnu kombinaciju originalnih promenljivih. Pomoću glavnih komponenti smo uspeali da razdvojimo tri vrste iris cvetova. Vrsta setosa se posebno izdvojila, dok vrste versicolor i virginica imaju preklapanja jer imaju više sličnosti međusobno nego sa vrstom setosa. Elipse oko vrsta cvetova predstavljaju 95% interval poverenja varijacije podataka za svaku od vrsta.



Slika 3.20: Prikaz projekcije opservacija u novom sistemu pomoću funkcije ggbiplot za podatke iris

Da napomenemo, projekcije opservacija u novom sistemu nazivamo skorovima. Predstavljaju koeficijente glavnih komponenti za svaku opservaciju. Da bismo ih prikazali u tabeli koristimo sledeće naredbe:

```
str(myPr)
```

```
myPr$x
```

Funkcija str služi da prikaže strukturu proizvoljnog objekta, u našem slučaju glavnih komponenti. Kada je pozovemo dobijemo sledeći rezultat prikazan na slici 3.22. Softver sam generiše naziv za skorove, označeni su kao x. Da bismo ih prikazali u tabeli koristimo naredbu myPr\$x. U tabeli 3.14 smo prikazali prvih pet vrsta iz tabele sa skorovima koju dobijemo u softveru.

	PC1	PC2	PC3	PC4
[1,]	-2.257	-0.478	0.127	0.024
[2,]	-2.074	0.671	0.233	0.102
[3,]	-2.356	0.340	-0.044	0.028
[4,]	-2.291	0.595	-0.090	-0.065
[5,]	-2.381	-0.644	-0.015	-0.035

Tabela 3.14: Prikaz skorova za prvih pet opservacija za podatke iris

Napomena: Podaci o iris cvetovima se često koriste u literaturi kao primeri i datiraju dugo, a u ovom radu je odabran kao primer zbog slikovitog prikaza primene glavnih komponenti.


```

> str(myPr)
List of 5
 $ sdev      : num [1:4] 1.708 0.956 0.383 0.144
 $ rotation: num [1:4, 1:4] 0.521 -0.269 0.58 0.565 -0.377 ...
 .. attr(*, "dimnames")=List of 2
 .. ..$ : chr [1:4] "Sepal.Length" "Sepal.width" "Petal.Length" "Petal.width"
 .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
 $ center   : Named num [1:4] 5.84 3.06 3.76 1.2
 .. attr(*, "names")= chr [1:4] "Sepal.Length" "Sepal.width" "Petal.Length" "Petal.width"
 $ scale     : Named num [1:4] 0.828 0.436 1.765 0.762
 .. attr(*, "names")= chr [1:4] "Sepal.Length" "Sepal.width" "Petal.Length" "Petal.width"
 $ x         : num [1:150, 1:4] -2.26 -2.07 -2.36 -2.29 -2.38 ...
 .. attr(*, "dimnames")=List of 2
 .. ..$ : NULL
 .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
 - attr(*, "class")= chr "prcomp"
>

```

Slika 3.21: Prikaz strukture objekta myPr

Za sledeći primer odabran je set podataka pod nazivom *Lawyers' ratings of state judges in the US Superior Court* koji se nalazi u R-u. Skup podataka se sastoji od 43 opservacije i 12 promenljivih, opservacije predstavljaju sudije, a promenljive predstavljaju rejting državnih sudija Vrhovnog Suda SAD-a u različitim oblastima. Promenljive su sledeće: CONT-broj kontakta advokata sa sudijom, INTG-integritet, DMNR-ponašanje, DILG-marljivost, CFMG-upravljanje predmetima, DECI-brzo odlučivanje, PREP-priprema za suđenje, FAMI-poznavanje zakona, ORAL-usmeno izrečene presude, WRIT-pisano izrečene presude, PSY-psihička sposobnost, RTEN-dostojno zadržavanje.

Kao i u prethodnom primeru, učitavamo podatke i prikazujemo prve 3 vrste sledećim naredbama

```

data("USJudgeRatings")
head(USJudgeRatings)

```

Prikaz tih podataka vidimo u sledećoj tabeli:

	COUNT	INTG	DMNR	DILG	CFMG	DECI	PREP	FAMI	ORAL	WRIT	PSY	RTEN
AARONSON,L.H.	5.7	7.9	7.7	7.3	7.1	7.4	7.1	7.1	7.1	7.0	8.3	7.8
ALEXANDER,J.M.	6.8	8.9	8.8	8.5	7.8	8.1	8.0	8.0	7.8	7.9	8.5	8.7
ARMENTANO,A.J.	7.2	8.1	7.8	7.8	7.5	7.6	7.5	7.5	7.3	7.4	7.9	7.8

Tabela 3.15: Prikaz prve tri vrste u setu podataka USJudgeRatings

Na sledećoj slici 3.22 prikazana je matrica korelacije podataka i to pomoću sledećih naredbi:

```

rez.corr <- cor(USJudgeRatings[1:12])
rez.corr
round(rez.corr, 2)

```

Funkciju `cor(USJudgeRatings[1:12])` koristimo da bismo izračunali korelacionu matricu, a matricu odnosno koeficijente korelacije čuvamo u objektu pod nazivom `rez.corr`. Pomoću funkcije `round(rez.corr, 2)` dobijamo matricu sa koeficijentima korelacije zaokružanim na dve decimale.

Na osnovu matrice korelacije zaključujemo da su podaci pozitivno korelisani i to visoko, što nam daje još bolje rezultate prilikom analiziranja glavnih komponenti.

```

Console ~/
> round(rez.corr,2)
      CONT  INTG  DMNR  DILG  CFMG  DECI  PREP  FAMI  ORAL  WRIT  PHYS  RTEN
CONT  1.00 -0.13 -0.15  0.01  0.14  0.09  0.01 -0.03 -0.01 -0.04  0.05 -0.03
INTG -0.13  1.00  0.96  0.87  0.81  0.80  0.88  0.87  0.91  0.91  0.74  0.94
DMNR -0.15  0.96  1.00  0.84  0.81  0.80  0.86  0.84  0.91  0.89  0.79  0.94
DILG  0.01  0.87  0.84  1.00  0.96  0.96  0.98  0.96  0.95  0.96  0.81  0.93
CFMG  0.14  0.81  0.81  0.96  1.00  0.98  0.96  0.94  0.95  0.94  0.88  0.93
DECI  0.09  0.80  0.80  0.96  0.98  1.00  0.96  0.94  0.95  0.95  0.87  0.92
PREP  0.01  0.88  0.86  0.98  0.96  0.96  1.00  0.99  0.98  0.99  0.85  0.95
FAMI -0.03  0.87  0.84  0.96  0.94  0.94  0.99  1.00  0.98  0.99  0.84  0.94
ORAL -0.01  0.91  0.91  0.95  0.95  0.95  0.98  0.98  1.00  0.99  0.89  0.98
WRIT -0.04  0.91  0.89  0.96  0.94  0.95  0.99  0.99  0.99  1.00  0.86  0.97
PHYS  0.05  0.74  0.79  0.81  0.88  0.87  0.85  0.84  0.89  0.86  1.00  0.91
RTEN -0.03  0.94  0.94  0.93  0.93  0.92  0.95  0.94  0.98  0.97  0.91  1.00
>

```

Slika 3.22: Matrica korelacije za podatke USJudgeRatings

Deskriptivna statistika i varijansa za svaku od promenljivih je prikazana na sledećoj slici, a dobijene su naredbama:

```
summary(USJudgeRatings)
```

```
apply(USJudgeRatings[1:12], 2, var)
```

Pomoću funkcije `apply(USJudgeRatings[1:12], 2, var)` dobijamo varijansu za svaku od promenljivih.

```

Console ~/
> summary(USJudgeRatings)
      CONT      INTG      DMNR      DILG      CFMG
Min.   :5.700   Min.   :5.900   Min.   :4.300   Min.   :5.100   Min.   :5.400
1st Qu.:6.850   1st Qu.:7.550   1st Qu.:6.900   1st Qu.:7.150   1st Qu.:7.000
Median :7.300   Median :8.100   Median :7.700   Median :7.800   Median :7.600
Mean   :7.437   Mean   :8.021   Mean   :7.516   Mean   :7.693   Mean   :7.479
3rd Qu.:7.900   3rd Qu.:8.550   3rd Qu.:8.350   3rd Qu.:8.450   3rd Qu.:8.050
Max.   :10.600  Max.   :9.200   Max.   :9.000   Max.   :9.000   Max.   :8.700

      DECI      PREP      FAMI      ORAL      WRIT
Min.   :5.700   Min.   :4.800   Min.   :5.100   Min.   :4.700   Min.   :4.900
1st Qu.:7.100   1st Qu.:6.900   1st Qu.:6.950   1st Qu.:6.850   1st Qu.:6.900
Median :7.700   Median :7.700   Median :7.600   Median :7.500   Median :7.600
Mean   :7.565   Mean   :7.467   Mean   :7.488   Mean   :7.293   Mean   :7.384
3rd Qu.:8.150   3rd Qu.:8.200   3rd Qu.:8.250   3rd Qu.:8.000   3rd Qu.:8.050
Max.   :8.800   Max.   :9.100   Max.   :9.100   Max.   :8.900   Max.   :9.000

      PHYS      RTEN
Min.   :4.700   Min.   :4.800
1st Qu.:7.700   1st Qu.:7.150
Median :8.100   Median :7.800
Mean   :7.935   Mean   :7.602
3rd Qu.:8.500   3rd Qu.:8.250
Max.   :9.100   Max.   :9.200

> apply(USJudgeRatings[1:12], 2, var)
      CONT      INTG      DMNR      DILG      CFMG      DECI      PREP      FAMI      ORAL
0.8852492 0.5931229 1.3080620 0.8116168 0.7397896 0.6447065 0.9089147 0.9005759 1.0201883
      WRIT      PHYS      RTEN
0.9237763 0.8828018 1.2121373
>

```

Slika 3.23: Deskriptivna statistika i varijansa promenljivih za podatke USJudgeRatings

Dalje, prvo standardizujemo promenljive iz skupa podataka, a zatim odredimo karakteristične korene i vektore i to sledećim naredbama:

```
scaled_sudije <- apply(USJudgeRatings[1:12], 2, scale)
```

```
sudije.cov <- cov(scaled_sudije)
```

```
sudije.eigen <- eigen(sudije.cov)
```

```
str(sudije.eigen)
```

```
koreni <- sudije.eigen$values[1:12]
```

```
koreni
```

```
vektori <- sudije.eigen$vectors[1:12,1:12]
```

```
vektori
```

Na slici 3.24 su prikazani karakteristični koreni i vektori dobijeni u programu. Vidimo da je prvi karakteristični koren iznosi 10.13, drugi 1.10, treći 0.33, itd.

```
Source
Console ~/
> scaled_sudije <- apply(USJudgeRatings[1:12], 2, scale)
> sudije.cov <- cov(scaled_sudije)
> sudije.eigen <- eigen(sudije.cov)
> str(sudije.eigen)
List of 2
 $ values : num [1:12] 10.1335 1.1041 0.3329 0.2538 0.0845 ...
 $ vectors: num [1:12, 1:12] 0.00308 -0.28855 -0.28688 -0.30435 -0.30257 ...
 - attr(*, "class")= chr "eigen"
>
> koreni <- sudije.eigen$values[1:12]
> koreni
[1] 10.133503726 1.104146980 0.332901600 0.253847001 0.084452786 0.037286058
[7] 0.019682813 0.015415288 0.007833472 0.005611727 0.003258175 0.002060374
> vektori <- sudije.eigen$vectors[1:12,1:12]
> vektori
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 0.003075143 0.932890644 -0.334756548 -0.058576867 -0.093438368 -0.004064432
[2,] -0.288550775 -0.182040993 -0.549360126 -0.173977074 0.014543880 0.369937339
[3,] -0.286884206 -0.197565743 -0.556490386 0.124412022 0.228832817 -0.394724667
[4,] -0.304354091 0.036304667 0.163629910 -0.321395544 0.301936920 0.598676072
[5,] -0.302572733 0.168393523 0.207341904 -0.012949223 0.448430522 -0.085728870
[6,] -0.301891969 0.127877299 0.297902771 -0.030491779 0.424003128 -0.392609484
[7,] -0.309406446 0.032230248 0.151869345 -0.213656069 -0.202853400 0.083216652
[8,] -0.306679527 -0.001315183 0.195290454 -0.200651140 -0.507470003 -0.101538704
[9,] -0.312708348 -0.003625720 0.002150634 0.007441042 -0.246059421 -0.150272440
[10,] -0.311061231 -0.031378756 0.056045596 -0.137104995 -0.305562842 -0.238172386
[11,] -0.280723624 0.089037698 0.154000444 0.841266046 -0.118424976 0.299281534
[12,] -0.309790218 -0.039381306 -0.172869757 0.184223629 -0.006717911 0.036126847
      [,7]      [,8]      [,9]      [,10]      [,11]      [,12]
[1,] 0.005214784 -6.006597e-02 0.02514533 0.03038881 -0.0145329260 0.007940919
[2,] -0.449810741 3.341645e-01 0.27537794 -0.10897641 0.1125535650 -0.009848658
[3,] 0.466747889 -2.470974e-01 0.19910004 0.07241282 -0.1343234234 -0.059121657
[4,] 0.209710731 -3.548587e-01 -0.03977180 0.38339165 -0.0709517642 -0.053790339
[5,] 0.246903359 7.135261e-01 -0.14342471 -0.09850310 -0.1658680927 -0.025082947
[6,] -0.536429933 -3.024227e-01 0.25823773 -0.06743847 0.1284999526 -0.044141604
[7,] 0.335390036 -1.536754e-01 0.10876864 -0.67986284 0.3187612119 0.273286884
[8,] -0.036378004 2.038889e-02 0.22306628 -0.04004599 -0.5733628652 -0.421739844
[9,] 0.057580177 9.062990e-02 -0.29951714 0.25599455 0.6386061655 -0.494391025
[10,] -0.060899994 1.261203e-01 -0.02497324 0.47478254 -0.0004056397 0.696107204
```

Slika 3.24: Karakteristični koreni i vektori standardizovanih promenljivih za podatke USJudgeRatings

Sada pomoću sledeće naredbe dobijamo matricu koeficijenata glavnih komponenti (opciono je prikazano prvih 5):

```
mca_koef <- sudije.eigen$vectors[1:12,1:5]
```

Kako su u R-u po defaultu karakteristični vektori postavljeni u negativnom pravcu, matricu koeficijenata množimo sa -1 da bismo dobili pozitivan pravac i to sledećom naredbom.

```
mca_koef <- -mca_koef
```

Kako bi bilo preglednije i uočljivije, dodeljujemo nazive promenljivih u prvoj koloni i nazive kolona koji predstavljaju komponente, te prikazujemo tako dobijenu matricu koeficijenata sledećim naredbama i na slici 3.25 :

```
row.names(mca_koef) <- c("CONT", "INTG", "DMNR", "DILG", "CFMG", "DECI",
"PREP", "FAMI", "ORAL", "WRIT", "PSY", "RTEN")
```

```
colnames(mca_koef) <- c("PC1", "PC2", "PC3", "PC4", "PC5")
```

```
mca_koef
```

Ovim je prikazano da je na više načina moguće izračunati glavne komponente, ne samo koristeći postojeće funkcije u programu.

Dalje su prikazane naredbe pomoću kojih dobijamo skorove odnosno koeficijente glavnih komponenti za svaku opservaciju:

```
PC1 <- as.matrix(scaled_sudije) %*% mca_koef[,1]
```

```
PC2 <- as.matrix(scaled_sudije) %*% mca_koef[,2]
```

Zatim te podatke prikazujemo kao matricu sledećom naredbom, što je prikazano na slici 3.26.

```
PC <- data.frame(USJudgeRatings[,0], PC1, PC2)
```

```
head(PC)
```

```

> mca_koef <- -mca_koef
> row.names(mca_koef) <- c("CONT", "INTG", "DMNR", "DILG", "CFMG",
+ "DECI", "PREP", "FAMI", "ORAL", "WRIT", "PSY", "RTEN")
> colnames(mca_koef) <- c("PC1", "PC2", "PC3", "PC4", "PC5")
> mca_koef

```

	PC1	PC2	PC3	PC4	PC5
CONT	-0.003075143	-0.932890644	0.334756548	0.058576867	0.093438368
INTG	0.288550775	0.182040993	0.549360126	0.173977074	-0.014543880
DMNR	0.286884206	0.197565743	0.556490386	-0.124412022	-0.228832817
DILG	0.304354091	-0.036304667	-0.163629910	0.321395544	-0.301936920
CFMG	0.302572733	-0.168393523	-0.207341904	0.012949223	-0.448430522
DECI	0.301891969	-0.127877299	-0.297902771	0.030491779	-0.424003128
PREP	0.309406446	-0.032230248	-0.151869345	0.213656069	0.202853400
FAMI	0.306679527	0.001315183	-0.195290454	0.200651140	0.507470003
ORAL	0.312708348	0.003625720	-0.002150634	-0.007441042	0.246059421
WRIT	0.311061231	0.031378756	-0.056045596	0.137104995	0.305562842
PSY	0.280723624	-0.089037698	-0.154000444	-0.841266046	0.118424976
RTEN	0.309790218	0.039381306	0.172869757	-0.184223629	0.006717911

Slika 3.25: Matrica koeficijenata glavnih komponenti za podatke USJudgeRatings

Da bismo dobili skorove glavnih komponenti odnosno prikaz opservacija u novom koordinatnom sistemu gde su ose glavne komponente, koristili smo matrični proizvod matrice sa standardizovanim podacima naših opservacija i prve kolone, respektivno druge, matrice koeficijenata koju smo u prethodnim koracima izračunali. Te matrične proizvode smo sačuvali u objektima pod nazivom PC1 i PC2, a zatim smo jednostavnim naredbom `data.frame` spojili podatke u matricu, tako da je u prvoj koloni prikazan naziv svakog od sudija, slika 3.26.

```

> PC <- data.frame(USJudgeRatings[,0] , PC1, PC2)
> head(PC)

```

	PC1	PC2
AARONSON, L. H.	-0.5857508	1.8130995
ALEXANDER, J. M.	2.3800115	0.8678932
ARMENTANO, A. J.	0.2251390	0.2987803
BERDON, R. I.	3.6178993	0.5893672
BRACKEN, J. J.	-6.8711315	-0.1357495
BURNS, E. B.	2.4454791	1.4057858

Slika 3.26: Prikaz matrice skorova za podatke USJudgeRatings

Dalje će vizuelno biti prikazana projekcija opservacija u novom koordinatnom sistemu, kao i zbog čega su ostavljene samo 2 komponente prilikom analiziranja.

Da bismo mogli da prikažemo vizuelno podatke odnosno da imamo više opcija prilikom manipulacije i prikaza, instalirani su 3 paketa, a zatim i učitana, sledećim naredbama:

```

install.packages("tidyverse")
library(tidyverse)
install.packages("gridExtra")
library(gridExtra)
install.packages("ggplot2")
library(gridExtra)

```

Grafički, na slici 3.27, su prikazani skorovi i to pomoću sledećih naredbi:

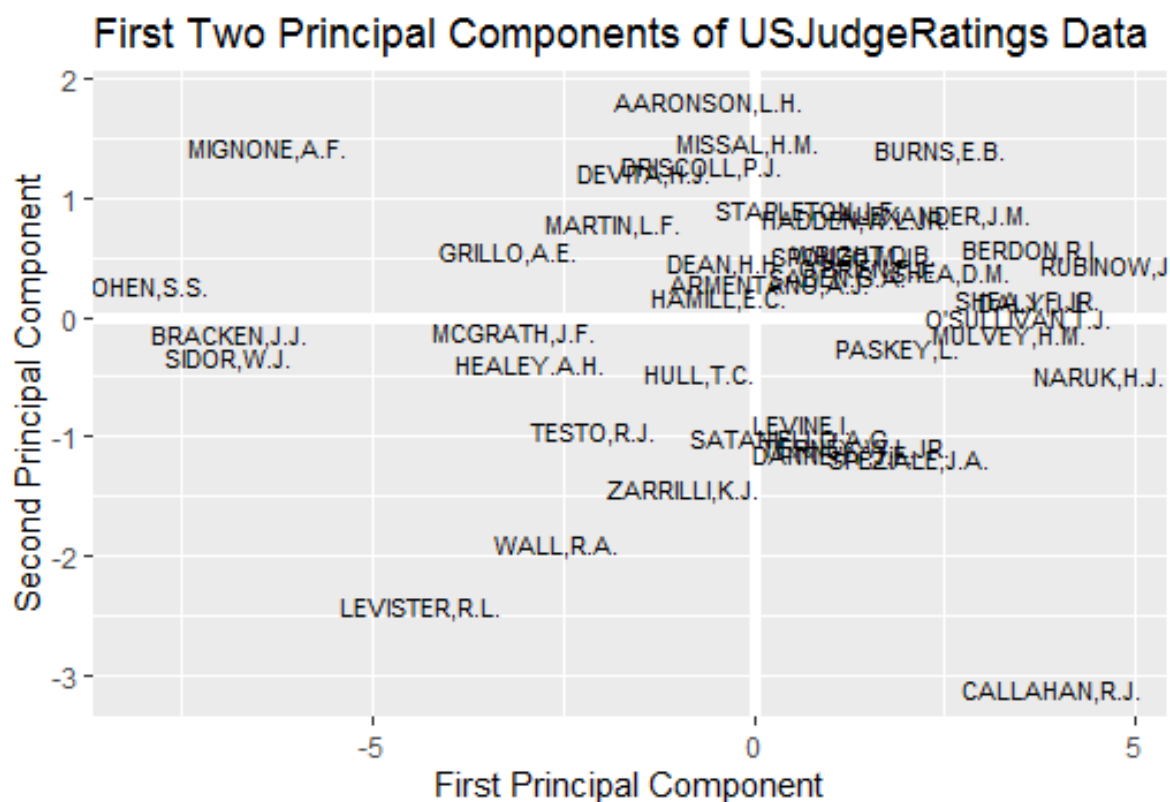
```

ggplot(PC, aes(PC1, PC2)) +
  modelr::geom_ref_line(h = 0) +
  modelr::geom_ref_line(v = 0) +
  geom_text(aes(label = Judges), size = 3) +
  xlab("First Principal Component") +
  ylab("Second Principal Component") +
  ggtitle("First Two Principal Components of USJudgeRatings Data")

```

Funkcija `ggplot` služi za grafički prikaz. Njen prvi parametar je skup podataka, u našem

primeru matrica PC, drugi parametar su ose koje smo prethodno definisali kao vektore PC1 i PC2. Dalje smo naredbama `modelr::geom_ref_line(h = 0)` i `modelr::geom_ref_line(v = 0)` dodali na grafiku bele linije koje predstavljaju ose. Ose su imenovane kao First i Second Principal Component naredbama `xlab("First Principal Component")` i `ylab("Second Principal Component")`, a pomoću naredbe `ggtitle("First Two Principal Components of USJudgeRatings Data")` dat je opis slike.



Slika 3.27: Grafički prikaz opservacija u koordinatnom sistemu sa komponentama za podatke USJudgeRatings

Potrebno je napomenuti da na ovaj način analiziranja pomoću glavnih komponenti nije moguće izvesti određene predikcije, nego je moguće naći povezanost među promenljivim, kao i izvesti zaključak kako to utiče na opservacije.

U ovom primeru je prikazan drugi način za dobijanje scree testa pomoću naredbi navedenih u nastavku. Prvo računamo proporciju objašnjene varijanse glavnim komponentama i to sledećim naredbama:

```
PVE <- sudije.eigen$values / sum(sudije.eigen$values)
round(PVE, 2)
```

U objektu koji smo nazvali PVE (proportion of variance explained) čuvamo vrednosti. Proporciju objašnjene varijanse računamo kao količnik određenog karakterističnog korena i ukupnog broja glavnih komponenti, što je urađeno naredbom `sudije.eigen$values / sum(sudije.eigen$values)`. Da bi rezultat bio pregledniji, cifre su zaokružene na dve decimale naredbom `round(PVE,2)`. Rezultat je dat u tabeli 3.16. Iz tabele vidimo da su prva i druga komponenta objasnile ukupno 93% varijanse podataka, što je dovoljan podatak da se u daljoj analizi odlučimo za zadržavanje prve 2 komponente. U nastavku prikazujemo i vizuelno, pomoću scree testa i kumulativnog scree testa, sledećim naredbama, a na slici 3.28:

```
PVEplot <- qplot(c(1:12), PVE) +
```


PVE	0.84	0.09	0.03	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
------------	------	------	------	------	------	------	------	------	------	------	------	------

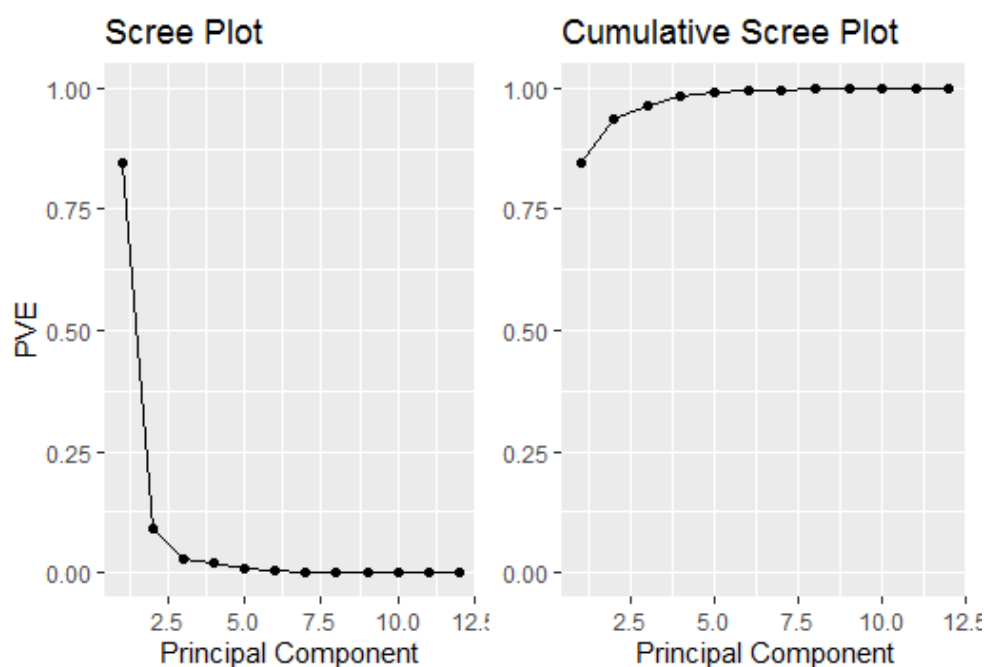
Tabela 3.16: Prikaz proporcije objašnjene varijanse komponentama za podatke USJudgeRatings

```
geom_line() +
xlab("Principal Component") +
ylab("PVE") +
ggtitle("Scree Plot") +
ylim(0, 1)

cumPVE <- qplot(c(1:12), cumsum(PVE)) +
geom_line() +
xlab("Principal Component") +
ylab(NULL) +
ggtitle("Cumulative Scree Plot") +
ylim(0,1)
```

```
grid.arrange(PVEplot, cumPVE, ncol = 2)
```

Naredbe su slične kao one korištene za prethodnu sliku, samo umesto funkcije ggplot je korištena naredba qplot, koja takođe služi za grafički prikaz. Funkciji qplot je prosleđen objekat PVE za scree plot, a za kumulativni scree plot je prosleđena funkcija cumsum, na objektu PVE, koja vraća vektor čiji su elementi kumulativna suma.



Slika 3.28: Scree test i kumulativni scree test za podatke USJudgeRatings

Jasno se vidi na slici 3.28 da je za dalju analizu dovoljno posmatrati prve dve komponente (desni deo - Scree Plot), na levom delu je prikazano kumulativno objasnjenje varijanse za sve vrednosti karakterističnih korena odnosno sve komponente. Jasno se vidi da je prvom komponentom objašnjeno 84%, sa prve dve 93%, sa tri 96%, dok je sa prvih 6 komponenti objašnjeno 100% varijanse.

Kako je prva komponenta pozitivno korelisana sa promenljivima ORAL, WRIT, PREP

i FAMI može se protumačiti da predstavlja pripremljenost, poznavanje pravosuđa i iskustvo (zbog irečenih presuda), dok je druga komponenta negativno korelisana sa promenljivom CONT i ona predstavlja posvećenost sudija slučajevima i advokatima.

Sudija Callahan je sudija koji ima najveće ocene odnosno najbolji rejting među sudijama i zato se nalazi u donjem desnom uglu na slici 3.27, jer je i dobro pripremljen i poznaje pravosuđe i ima iskustvo i posvećen je radu na raznim slučajevima, a imao je i najviše kontakata sa advokatima. Najviše je onih sudija sa iskustvom koji su pripremljeni i poznaju sistem, ali sa manjim brojem kontakata sa advokatima na određenim slučajevima i oni imaju visok rejting. Sudije Cohen i Mignone imaju najmanje ocene i nalaze se u gornjem levom uglu slike 3.27.

Za sledeći primer odabran je set podataka pod nazivom *Pima Indians Diabetes Database* - file *Diabetes.csv* koji se nalazi na internet stranici [18]. Originalni podaci su dobijeni od Nacionalnog instituta za dijabetes, probavne i bolesti bubrega, merenja su izvršena na ženama domorodskih naroda Arizone starijim od 21 godine, a cilj prikupljanja tih podataka je bio da se predvidi, na osnovu određenih merenja, da li pacijent ima dijabetes ili ne.

Skup podataka se sastoji od 768 opservacije i 9 promenljivih, ali u ovom primeru posmatramo 100 opservacija i 9 promenljivih, zbog jednostavnijeg prikaza. Promenljive su sledeće: Pregnancies - broj trudnoća, Glucose - koncentracija glukoze u plazmi 2 sata tokom oralnog testa tolerancije na glukozu, BloodPressure - dijastolički krvi pritisak, SkinThickness - debljina nabora kože tricepsa, Insulin - nivo insulina, BMI - Body mass index, DiabetesPedigreeFunction - podatak da li se iz generacije u generaciju bolest prenosi genetski, Age - broj godina, Outcome - kvalitativna promenljiva sa vrednošću 0 i 1, koja kaže da li osoba ima dijabetes ili ne.

Naredba za učitavanje podataka u program je:

```
dijabetes <- read.csv(choose.files())
print(dijabetes)
dijabetesNovi <- dijabetes[1:100, 1:8]
head(dijabetesNovi, 5)
```

U sledećoj tabeli su prikazane prve 3 vrste podataka koje koristimo u ovom primeru.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
1	6	148	72	35	0	33.6	0.627	50
2	1	85	66	29	0	26.6	0.351	31
3	8	183	64	0	0	23.3	0.672	32

Tabela 3.17: Prikaz prve tri vrste *Diabetes* podataka

Kako se podaci nalaze u csv²² formatu, naredbom read.csv se učitavaju u program, a u ovom primeru su sačuvani u objektu pod nazivom dijabetes. Kako je gore navedeno, radi preglednosti, biće korišteno prvih 100 opservacija i to smo postigli naredbom dijabetes[1:100, 1:8], takođe, odabrano je 8 promenljivih (ne posmatra se promenljiva Outcome jer nije numeričkog tipa odnosno kvantitativna nego kategorička odnosno kvalitativna).

U ovom primeru, za dobijanje glavnih komponenti koristimo funkcije iz paketa FactoMineR. Prvo je potrebno instalirati paket i učitati ga, a takođe učitavamo paket factoextra, koji služi za vizualizaciju podataka. Takođe, potrebno je instalirati i pakete devtools i ggplot2, ukoliko ne postoje već instalirani. Sledeće naredbe služe za učitavanje i instaliranje:

```
install.packages("FactoMineR")
library("FactoMineR")
```

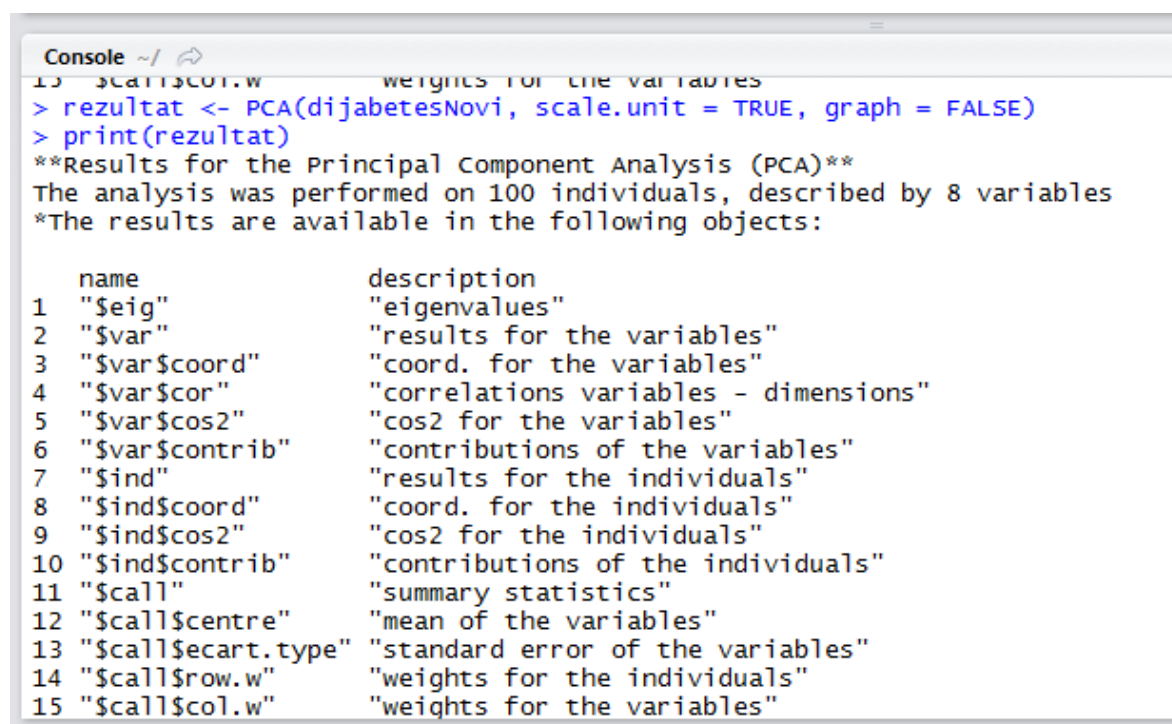
²²CSV je skraćenica sa značenjem comma separate value, a to znači da je skup podataka razdvojen zarezima.

```
if(!require(devtools)) install.packages("devtools")
devtools::install_github("kassambara/factoextra")
install.packages("factoextra")
library("factoextra")
install.packages("ggplot2")
```

Pomoću funkcije PCA u okviru paketa dobijamo glavne komponente, kao što je navedeno na početku poglavlja o programskom jeziku R-u, a u ovom primeru naredba izgleda ovako:

```
rezultat <- PCA(dijabetesNovi, scale.unit = TRUE, graph = FALSE)
print(rezultat)
```

U objektu pod nazivom *rezultat* smo sačuvali dobijene komponente i sve relevantne informacije koje nam ovaj paket nudi, što je prikazano na slici 3.29, a dobijeno pomoću naredbe `print(rezultat)`. Funkciji PCA je kao prvi parametar prosleđen skup podatak gde posmatramo prvih 100 opservacija originalnih podataka, drugim parametrom smo standardizovali podatke, sa trećim parametrom postoji mogućnost da se iscrta grafik ukoliko se postavi vrednost parametra na TRUE.



Slika 3.29: Prikaz strukture objekta *rezult* za podatke *Diabetes*

Na slici vidimo kratak opis na kojim podacima je urađena analiza, a zatim vidimo u kojim objektima su sačuvani rezultati analize. Rezultati su karakteristični koreni, varijansa, skorovi, učešće pojedinačnih promenljivih u komponentama, učešće opservacija u komponentama, itd. U nastavku će biti prikazani podaci koji su dobijeni pomoću funkcije PCA. Karakteristične korene dobijamo na sledeći način, a prikazani su na slici 3.30:

```
koreni <- get_eigenvalue(rezultat)
koreni
```

Na slici 3.30, u prvoj koloni, se vidi da prvi karakteristični koren iznosi 2.51, drugi 1.65, treći 1.05, četvrti 0.95, itd. U drugoj koloni se vidi procenat objašnjene varijanse određenim korenom, a u trećoj kumulativni procenat. Prvim korenom je objašnjeno 31% varijanse,

drugim 20%, trećim 13%, a ukupno sa prva tri 65%.

Grafički prikaz postotka objašnjene varijanse odnosno scree plot moguće je dobiti na dva načina, sledećim naredbama:

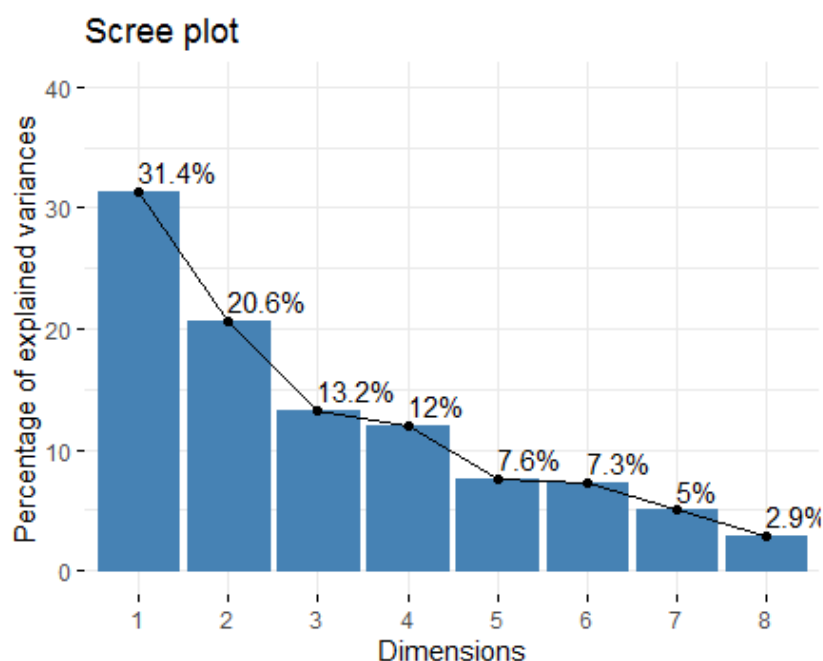
```
fviz_eig(rezultat, addlabels = TRUE, ylim = c(0, 40))
```

```
fviz_screplot(rezultat, addlabels = TRUE, ylim = c(0, 40))
```

Funkcija fviz služi za grafički prikaz, drugi parametar predstavlja prikaz postotka na grafiku ukoliko se postavi na logičku vrednost true, a treći parametar predstavlja interval koji se posmatra na y osi, što je prikazano na slici 3.31.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.5107108	31.383885	31.38389
Dim.2	1.6519065	20.648831	52.03272
Dim.3	1.0589364	13.236705	65.26942
Dim.4	0.9571349	11.964187	77.23361
Dim.5	0.6089234	7.611542	84.84515
Dim.6	0.5833016	7.291271	92.13642
Dim.7	0.3994216	4.992770	97.12919
Dim.8	0.2296647	2.870808	100.00000

Slika 3.30: Karakteristični koreni za podatke Diabetes



Slika 3.31: Karakteristični koreni za podatke Diabetes

Da bismo na jednom mestu imali sve informacije o varijablama, što je prikazano na slici 3.32, pozivamo funkciju get_pca_var i to na sledeći način:

```
varijable <- get_pca_var(rezultat)
```

```
varijable
```

Rezultat funkcije su koeficijenti promenljivih u odnosu na komponente, korelisanost između promenljivih i komponenti, učešće određene promenljive u glavnim komponentama iskazano u procentima.

```
> varijable <- get_pca_var(rezultat)
> varijable
Principal Component Analysis Results for variables
=====
      Name      Description
1 "$coord"    "coordinates for the variables"
2 "$cor"      "Correlations between variables and dimensions"
3 "$cos2"     "Cos2 for the variables"
4 "$contrib"  "contributions of the variables"
```

Slika 3.32: Prikaz dobijenih informacija o varijablama za podatke Diabetes

Koeficijente promenljivih dobijamo pomoću funkcije navedene u nastavku `head(varijable$coord, 8)`, a grafički prikaz promenljivih je dobijen funkcijom `fviz_pca_var(rezultat, col.var = "black")`, što je prikazano na slici 3.33. Na slici vidimo da su pozitivno korelisane promenljive `SkinThickness`, `DiabetesPedigreeFunction`, `Insulin`, `BMI` u odnosu na prvu komponentu, a promenljive `BloodPressure`, `Glucose`, `Age`, `Pregnacies` su negativno korelisane u odnosu na drugu komponentu, dok je promenljiva `SkinThickness` pozitivno korelisana (što vidimo i u tabeli koeficijenata iznad grafičkog prikaza). Grafički se može prikazati kvalitet reprezentovanih promenljivih sledećim naredbama, što je prikazano na slici 3.34:

```
corrplot(varijable$cos2, is.corr=FALSE)
fviz_cos2(rezultat, choice = "var", axes = 1:2)
```

Da bi bilo moguće koristiti funkciju `corrplot`, potrebno je instalirati paket pod nazivom `corrplot` i učitati ga. Kod funkcije `fviz_cos2` prvi parametar predstavlja matricu podataka, drugi parametar izbor da li će biti posmatrane vrste ili kolone ili, u slučaju glavnih komponenti varijable ili opservacije, dok treći parametar predstavlja dimenzionalnost odnosno broj zadržanih komponenti.

Na slici 3.34 vidimo da su promenljive `Pregnancies` i `Ages` najbolje reprezentuju prve dve komponente odnosno da su najviše zastupljene u prve dve glavne komponente, odnosno da broj godina i ukupan broj trudnoća najviše utiču na pojavu dijabetesa. Promenljiva `DiabetesPedigreeFunction` je najmanje zastupljena u prve dve komponente, što se vidi na slici 3.34. Na to nam ukazuje i dužina vektora na slici 3.33, odnosno što je promenljiva bliža centru kruga to je njena zastupljenost u komponentama niža. Da bi bilo uočljivije, na slici 3.35 je različitim bojama prikazana zastupljenost promenljivih (crvenom bojom prikazane najzastupljenije, a plavom najmanje zastupljene), a naredba kojom dobijamo grafik je sledeća: `fviz_pca_var(rezultat, col.var = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE, title = "First Two PC of Diabetes Data")`

Funkciji za grafički prikaz prosleđujemo u prvom parametru podatke, drugim parametrom određujemo da li će biti obojene promenljive ili opservacije, da li u odnosu na koeficijente ili u odnosu na učešće, trećim parametrom određujemo boje, navedena naredba generiše sama paletu boja, četvrti parametar služi za otklanjanje preklapanja teksta na grafiku kad se postavi na logičku vrednost `true`, petim parametrom je dodeljen naziv grafiku.

Dalje su prikazana naredbe za prikaz procenta učešća promenljivih u objašnjavanju varijanse u dobijenim komponentama, prikazani na slici 3.36, kao i za grafički prikaz bar plotu učešća promenljivih u odnosu na prvu komponentu, drugu komponentu i obe komponente, respektivno, slike 3.37 i 3.38.

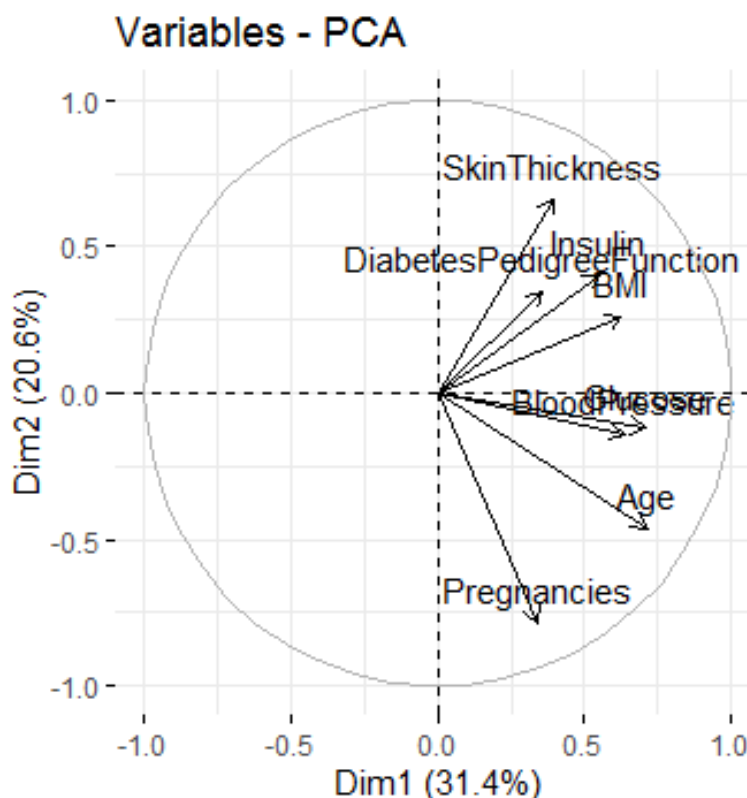
```
head(varijable$contrib, 8)
fviz_contrib(rezultat, choice = "var", axes = 1, top = 8)
fviz_contrib(rezultat, choice = "var", axes = 2, top = 8)
fviz_contrib(rezultat, choice = "var", axes = 1:2, top = 8)
```

```

Console ~/
> head(variable$coord,8)

```

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Pregnancies	0.3425262	-0.7828351	-0.03033985	-0.18203627	0.35688483
Glucose	0.7079353	-0.1195786	-0.31879458	0.34414744	-0.27568011
BloodPressure	0.6364014	-0.1387818	0.38643348	-0.30893668	-0.38921132
SkinThickness	0.3918192	0.6639087	-0.06166000	-0.41716468	0.37910092
Insulin	0.5495108	0.4046348	-0.65228916	-0.01912484	-0.06325242
BMI	0.6255546	0.2615122	0.44819178	-0.20926798	-0.05487160
DiabetesPedigreeFunction	0.3591036	0.3468700	0.41323363	0.69434662	0.23525433
Age	0.7151566	-0.4607655	-0.07832032	0.09907547	0.21908388



Slika 3.33: Koeficijenti promenljivih i grafički prikaz promenljivih za podatke Diabetes

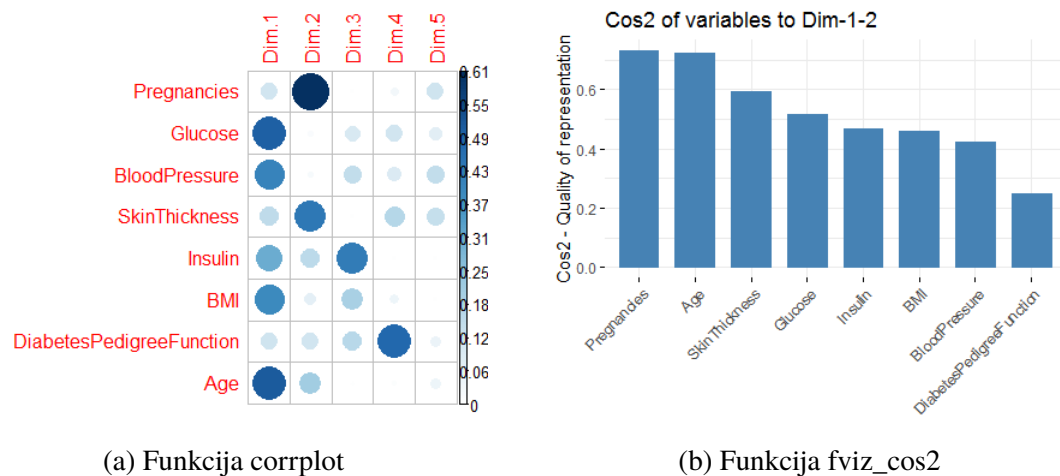
Funkciji za grafički prikaz učešća `fviz_contrib` u prvom parametru su prosledeni podaci, drugim parametrom je određeno da se radi o promenljivim, trećim parametrom je određena komponenta u odnosu na koju se posmatra, a zatim i broj promenljivih parametrom `top`, ukoliko je velik broj promenljivih da se odredi koliko je dovoljno za posmatranje.

Napomena: U ovom paketu program komponente naziva Dim1, Dim2, Dim3, itd.

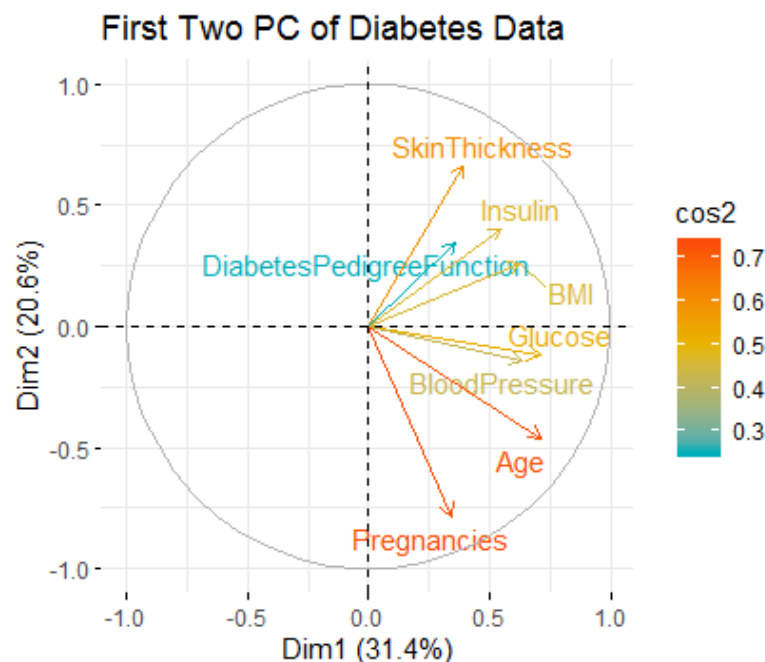
Crvena linija na graficima predstavlja očekivano prosečno učešće promenljivih.

Kod prve komponente se vidi da najveće učešće imaju promenljive Ages, Glucose, Blood-Pressure, BMI, dok su kod druge komponente to promenljive Pregnancies, SkinThickness.

Na slici 3.38 se vidi da, prilikom zadržavanja prve dve komponente, najveći postotak varijanse objašnjavaju promenljive Pregnancies, Ages i SkinThickness, odnosno da broj godina i broj trudnoća kod žene najviše utiče na pojavu dijabetesa.



Slika 3.34: Grafički prikaz kvaliteta reprezentovanih promenljivih pomoću funkcija corrplot i fviz_cos2 za podatke Diabetes



Slika 3.35: Grafički prikaz promenljivih za podatke Diabetes

```
> head(variable$contrib, 8)
```

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Pregnancies	4.672948	37.0983901	0.08692743	3.46212458	20.9167174
Glucose	19.961375	0.8656083	9.59736428	12.37416509	12.4809669
BloodPressure	16.131160	1.1659487	14.10196415	9.97162170	24.8775889
SkinThickness	6.114695	26.6827918	0.35903526	18.18201041	23.6019036
Insulin	12.026958	9.9115355	40.18004710	0.03821398	0.6570398
BMI	15.585969	4.1399834	18.96958796	4.57543501	0.4944616
DiabetesPedigreeFunction	5.136210	7.2836342	16.12580648	50.37087390	9.0889267
Age	20.370686	12.8521079	0.57926736	1.02555533	7.8823950

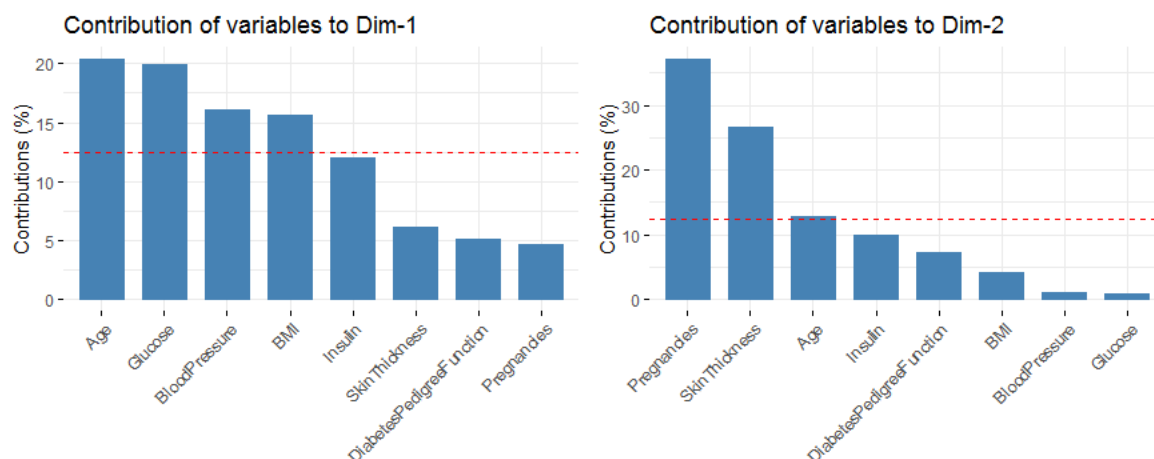
Slika 3.36: Prikaz procenta učešća promenljivih u objašnjenju varijanse komponentama za podatke Diabetes

U nastavku se navodi funkcija koja daje strukturu za opservacije, kako bi na jednom mestu imali sve potrebne informacije, što je prikazano na slici 3.39, a funkcija glasi:

```
opserv <- get_pca_ind(rezultat)
```

```
opserv
```

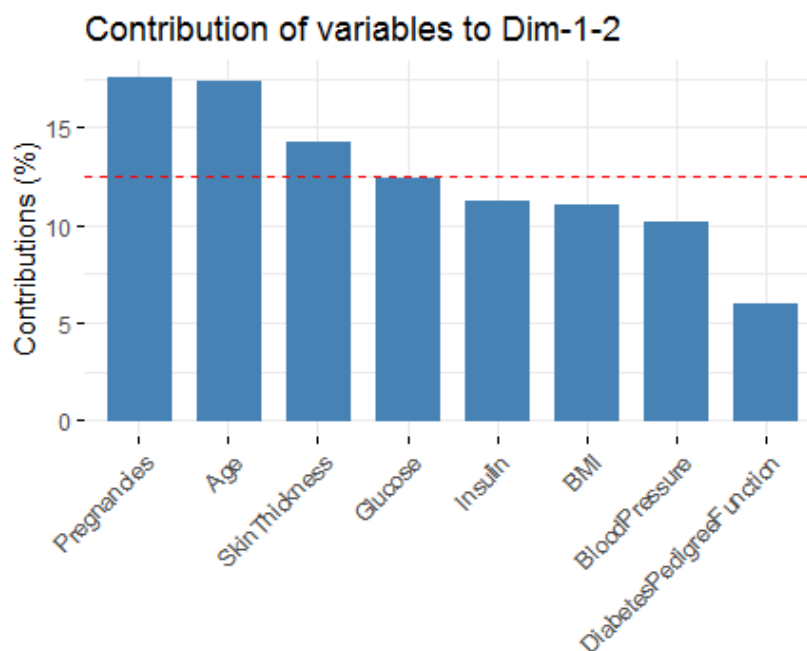
Rezultat funkcije su koeficijenti opservacija u odnosu na komponente, koeficijent reprezen-



(a) U odnosu na prvu komponentu

(b) U odnosu na drugu komponentu

Slika 3.37: Grafički prikaz procenta učešća promjenljivih u objašnjenu varijanse komponentama za podatke Diabetes



Slika 3.38: Prikaz procenta učešća promjenljivih u objašnjenju varijanse sa prve dve komponente za podatke Diabetes

tivnosti opservacija u odnosu na odabrane komponente, učešće pojedinačnih opservacija u glavnim komponentama iskazano u procentima.

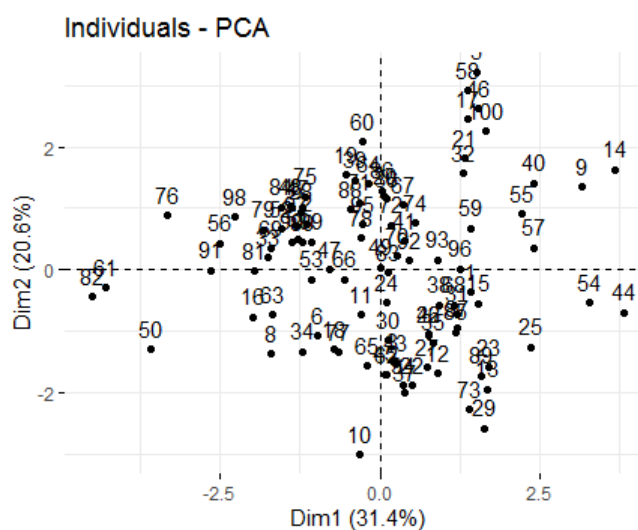
Koeficijente opservacija odnosno skorove dobijamo pomoću funkcije navedene u nastavku `head(opserv$coord,8)`, a grafički prikaz skorova je dobijen funkcijom `fviz_pca_ind(rezultat)`, što je prikazano na slici 3.40. Sledećom naredbom je moguć grafički prikaz opservacija u različitim bojama, na slici 3.41 pod a), u zavisnosti od koeficijenta reprezentativnosti opservacija u odnosu na komponente:

```
fviz_pca_ind(rezultat, col.ind = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = FALSE)
```

Kada se kao argument funkcije `repel` postavi na logičku vrednost `TRUE`, slika 3.41 pod b), onda dobijamo grafički prikaz bez preklapanja tačaka, što je korisno kada se ne posmatra veliki broj opservacija. Ukoliko se posmatra veći broj opservacija, potrebno je sačekati par sekundi dok se ne obradi naredba u programu.

```
> opserv <- get_pca_ind(rezultat)
> opserv
Principal Component Analysis Results for individuals
=====
  Name      Description
1 "$coord"  "Coordinates for the individuals"
2 "$cos2"   "Cos2 for the individuals"
3 "$contrib" "contributions of the individuals"
```

Slika 3.39: Prikaz dobijenih informacija o opservacijama za podatke *Diabetes*

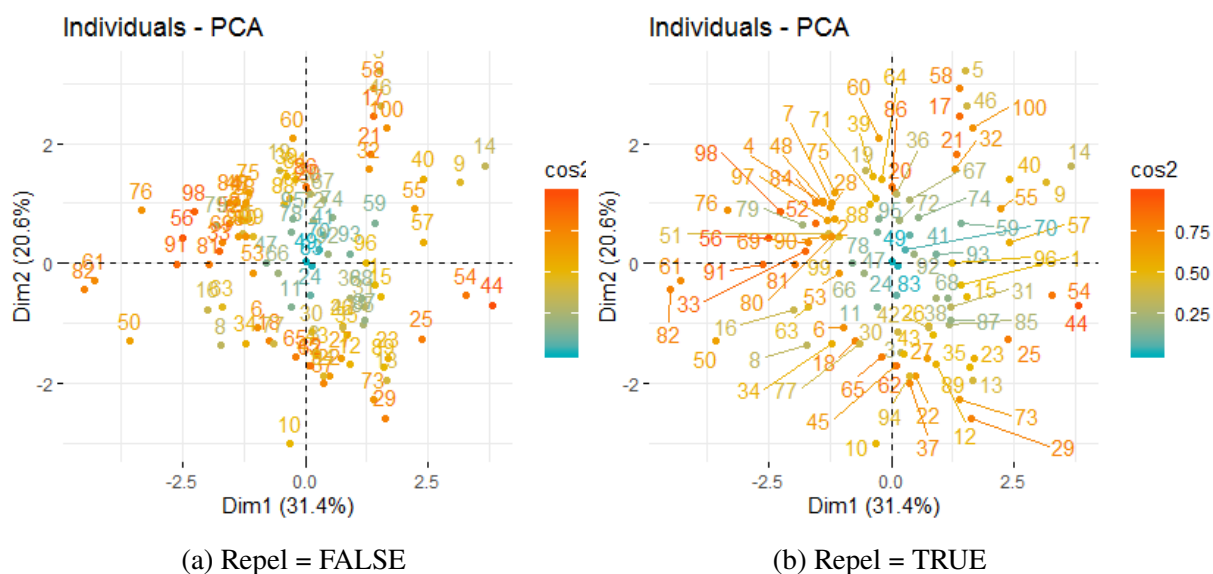


Slika 3.40: Grafički prikaz skorova za podatke *Diabetes*

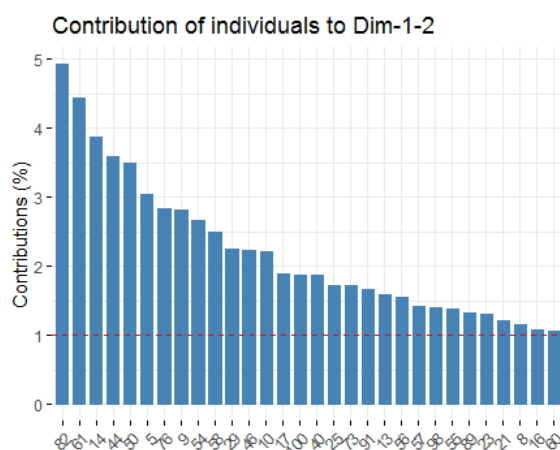
Ukoliko je potrebno grafički prikazati procenat učešća opservacija za prve dve komponente, to je moguće pomoću sledeće naredbe:

```
fviz_contrib(rezultat, choice = "ind", axes = 1:2, top = 30)
```

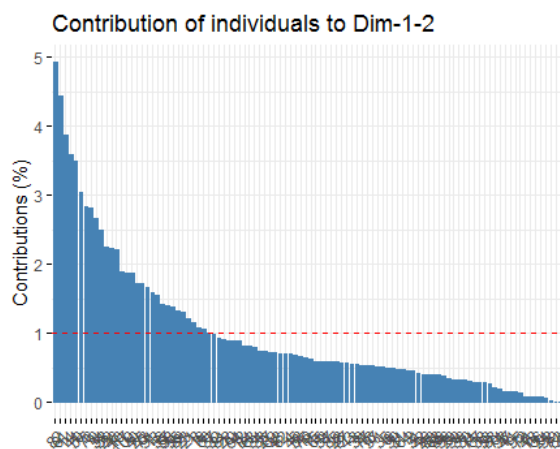
Drugim parametrom je specificirano da posmatramo opservacije, trećim parametrom da posmatramo prve dve komponente, a četvrtim da posmatramo 30 opservacija sa najvećim učešćem (određeno je 30 radi preglednosti na grafiku, moguće je posmatrati sve opservacije), prvim kao i obično učitavamo podatke. Na slici 3.42 je prikazano učešće 30 opservacija, a na slici 3.43 svih 100 opservacija. Crvena linija na graphicima prikazuje nivo prosečnog očekivanog učešća.



Slika 3.41: Grafički prikaz skorova u zavisnosti od koeficijenta reprezentativnosti za podatke Diabetes



Slika 3.42: Grafički prikaz procentualnog učešća skorova u prve dve komponente za podatke Diabetes



Slika 3.43: Grafički prikaz procentualnog učešća svih skorova u prve dve komponente za podatke Diabetes

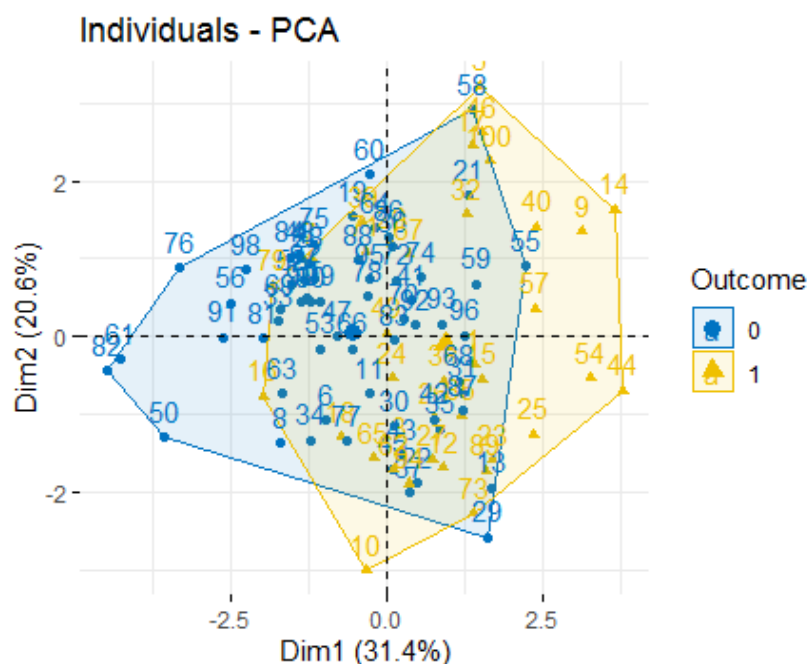
U nastavku u analizu uključujemo 9. promenljivu iz originalnog skupa odnosno kvalitativnu promenljivu pod nazivom Outcome. Naredba za to je sledeća:

```
rezultat2 <- PCA(diabetes, quali.sup = 9, graph=FALSE)
```

U objektu pod nazivom rezultat2 smo sačuvali rezultat primene funkcije PCA odnosno izračunali smo komponente tako što smo postavili promenljivu Outcome kao kvalitativnu promenljivu. Ono što je dalje zanimljivo da se prikaže je grafički prikaz opservacija za prve dve komponente sa kvalitativnom promenljivom (vrednost 0 promenljive Outcome znači da posmatrana osoba nema oboljenje, a vrednost 1 da ima), koji dobijamo sledećom naredbom: `fviz_pca_ind(rezultat2, habillage = 9, addEllipses = TRUE, ellipse.type = "convex", palette = "jco", repel = FALSE)`

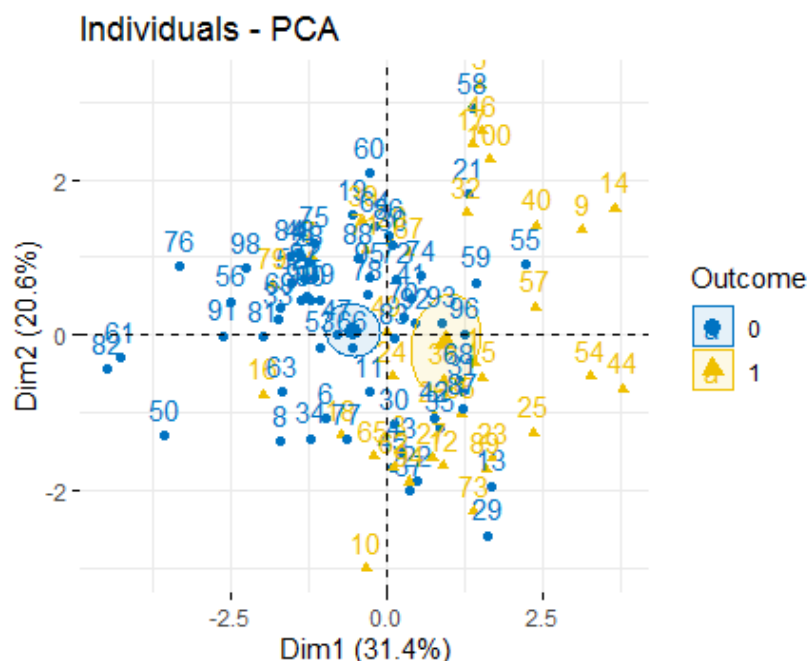
Funkciji je prosleđen objekat sa kvalitativnom promenljivom u prvom parametru, u drugom parametru je specificirano koja je kvalitativna promenljiva, trećim parametrom omogućavamo geometrijski prikaz rasprostranjenosti jedne od vrednosti te promenljive, četvrtim biramo način geometrijskog prikaza (ukoliko se postavi na vrednost convex dobijamo mnogougao kojim obuhvatamo sve tačke koje pripadaju određenoj vrednosti, ukoliko na vrednost confidence dobijamo elipse koje predstavljaju interval poverenja za očekivane vrednosti, ukoliko na vrednost t pretpostavljamo multivarijantnu t raspodelu, a sa vrednošću norm pretpostavljamo normalnu raspodelu), petim parametrom biramo boje prikaza, a sa parametrom repel kad se postavi na vrednost true dobijamo prikaz opservacija bez preklapanja. Kada se primeni gore navedena funkcija, dobija se sledeća slika 3.44.

Na slici 3.45 je prikazano kako izgleda kada se za četvrti parametar odabere vrednost confidence odnosno kada na grafiku dobijemo elipse koje predstavljaju interval poverenja za očekivane vrednosti.



Slika 3.44: Grafički prikaz skorova sa kvalitativnom promenljivom Outcome za podatke Diabetes

Kako je prva komponenta pozitivno korelisana sa promenljivima SkinThickness, DiabetesPedigreeFunction, Insulin i BMI može da se protumači da predstavlja fizičke predispozicije i nasledni faktor, a druga komponenta je negativno korelisana sa Pregnacies, Age, Glucose i BloodPressure, a pozitivno sa SkinThickness može da se protumači da predstavlja zrelost i starosnu i polnu. Na slici 3.44 se vidi da osobe sa lošijim fizičkim predispozicijama, naslednim



Slika 3.45: Grafički prikaz skorova sa parametrom confidence za podatke Diabetes

faktorom i u zrelijim godina imaju oboljenje (žuto obojene tačke na slici) odnosno imaju veće predispozicije ka oboljenju, dok mlađe žene sa boljim predispozicijama i naslednim faktorom nemaju oboljenje odnosno imaju manje predispozicije ka oboljenju. Na slici 3.45 je prikazan interval poverenja za očekivane vrednosti obolelih, odnosno u plavoj elipsi je očekivano da se pojavljuju zdrave žene, dok u žutoj elipsi treba da se pojavljuju žene sa oboljenjem.

Napomena: U ovom primeru smo zadržali samo prve dve glavne komponente zbog jednostavnije analize i lakšeg prikaza dobijenih rezultata. Bolji rezultat bi se postigao zadržavanje prve 3 komponente, jer sa prve dve komponente je objašnjenje ukupno 52.03% varijanse, a sa prve tri 65.26%. Takođe, ovim podacima nije moguće predvideti da li će neko dobiti oboljenje ili ne, moguće je samo analizirati postojeće podatke i izvesti određene zaključke.

Za sledeći primer odabran je set podataka pod nazivom *World Happiness Report 2019.csv* koji se nalazi na internet stranici [18], dalje u tekstu podatke nazivamo Happiness. Svetski izveštaj o sreći predstavlja značaj pregled stanja globalne sreće koji razvrstava 155 zemalja po tome u kom stepenu njihovi građani sebe vidi srećne. Ovogodišnji Izveštaj o sreći sveta usmeren je na sreću i zajednicu: kako se sreća razvijala u proteklih desetak godina, sa fokusom na tehnologije, društvene norme, sukobe i vladine politike koje su pokrenule te promene [18]. Skup podataka se sastoji od 155 opservacija i 10 promenljivih, opservacije predstavljaju različite države u kojima su mereni indikatori sreće potrebni za izveštaj. Promenljive, u ovom primeru indikatori sreće, su sledeće: Ladder - idikator životnog zadovoljstva, SD of Ladder - standardna devijacija (odstupanje) od mere životnog zadovoljstva, Positive affect - indikator pozitivnih emocija, Negative affect - indikator negativnih emocija, Social support - indikator u kojoj je meri socijalna podrška doprinela u izračunavanju ocene sreće, Freedom - indikator koliko je stepen slobode doprineo u izračunavanju ocene sreće, Corruption - indikator koliko je percepcija o korupciji doprinela u izračunavanju ocene sreće, Generosity - indikator koliko je velikodušnost doprinela u izračunavanju ocene sreće, Log og GDP per capita - indikator

stepena do kojeg BDP²³ po glavi stanovnika doprinosi izračunavanju ocene sreće, Healthy life expectancy - indikator stepena do kojeg je očekivani životni vek doprineo u izračunavanju ocene sreće.

Podatke, koje čuvamo u objektu pod nazivom happy, učitavamo i prikazujemo prve 3 vrste naredbama, podaci su prikazani u sledećoj tabeli 3.18 :

```
happy <- read.csv(choose.files())
head(happy, 3)
```

	Ladder	SD of Ladder	Positive affect	Negative affect	Social support	Freedom	Corruption	Generosity	Log of GDP per capita	Healthy life expectancy
Finland	1	4	41	10	2	5	4	47	22	27
Denmark	2	13	24	26	4	6	3	22	14	23
Norway	3	8	16	29	3	3	8	11	7	12

Tabela 3.18: Prikaz prve tri vrste Happiness podataka

U ovom primeru će biti prikazane naredbe za dobijanje glavnih komponenti u okviru paketa ade4. Prvo je potrebno instalirati i učitati podatke i to naredbama

```
install.packages("ade4")
install.packages("factoextra")
install.packages("magrittr")
library(ade4)
library(factoextra)
library(magrittr)
```

Funkcija za dobijanje komponenti je sledeća, a na slici 3.46 je prikazan rezultat pozivanja funkcije dudi.pca:

```
rez <- dudi.pca(happy[,2:11], scannf = FALSE, nf = 5)
```

Prvi parametar funkcije su podaci, drugim parametrom sa logičkom vrednošću true biramo da se scree plot prikazuje, dok sa trećim parametrom biramo koliko komponenti će biti prikazano u rezultatu (odabrano je 5).

Scree plot je moguće dobiti naredbom, a prikazan je na slici 3.47:

```
screeplot(rez, main = "Screeplot")
```

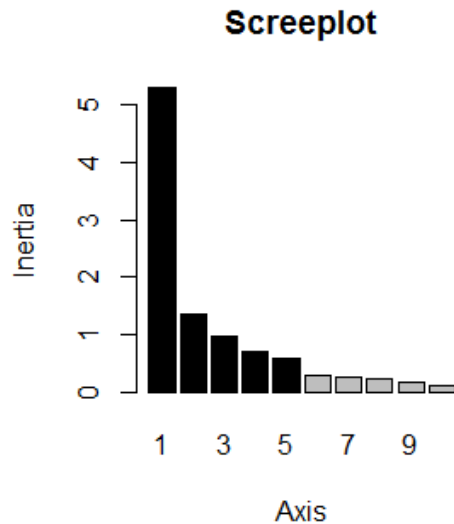
```
> rez <- dudi.pca(happy[,2:11], scannf = FALSE, nf = 5)
> print(rez)
Duality diagramm
class: pca dudi
$call: dudi.pca(df = happy[, 2:11], scannf = FALSE, nf = 5)

$nf: 5 axis-components saved
$rank: 10
eigen values: 5.295 1.363 0.9656 0.7022 0.5788 ...
  vector length mode   content
1 $cw     10      numeric column weights
2 $lw    155      numeric row weights
3 $eig     10      numeric eigen values

  data.frame nrow ncol content
1 $tab      155   10 modified array
2 $li       155    5 row coordinates
3 $l1       155    5 row normed scores
4 $co       10    5 column coordinates
5 $c1       10    5 column normed scores
other elements: cent norm
```

Slika 3.46: Rezultat funkcije dudi.pca za podatke Happiness

²³BDP je skraćenica za Bruto Domaći Proizvod.



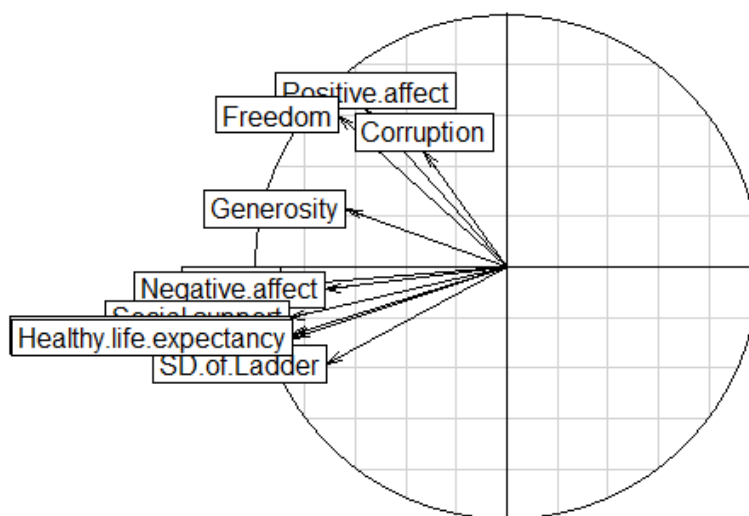
Slika 3.47: Scree plot za podatke Happiness

Kako vidimo na slici 3.46 prva dva karakteristična korena iznose 5.295 i 1.363, pa je zaključak, na osnovu Kajzerovog kriterijuma a i scree plota, da je u daljoj analizi dovoljno posmatrati prve dve komponente. Na slici vidimo i ostale informacije vezane za dobijenu strukturu.

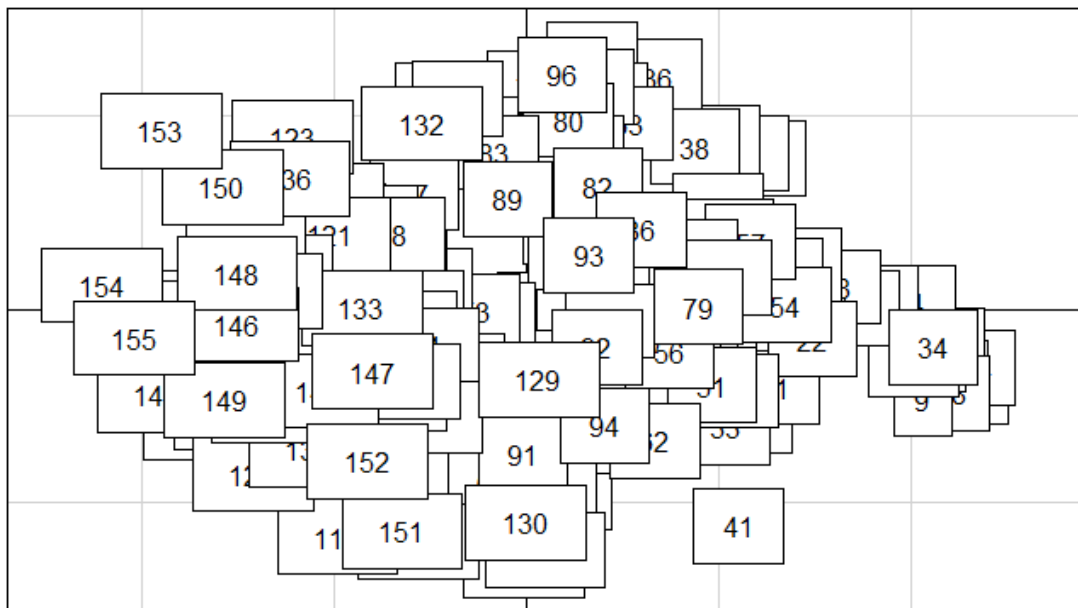
Grafički prikaz promenljivih, skorova, kao i biplota za promenljive i opservacije dobijamo sledećim naredbama, a prikazani su na slikama 3.48, 3.49 i 3.50, respektivno:

```
s.corcircle(rez$sco)
s.label(rez$li, xax = 1, yax = 2)
scatter(rez, posieig = "none", clab.row = 0 )
```

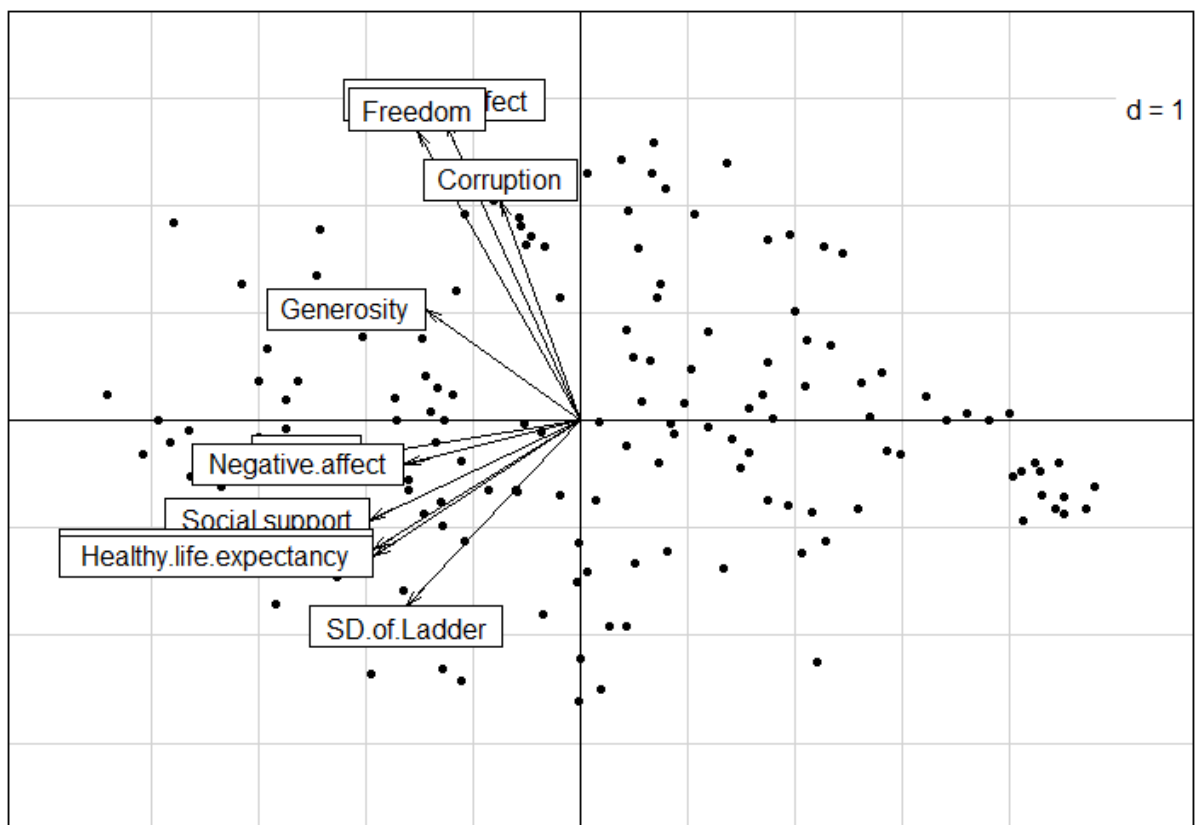
Funkcija `s.corcircle` služi za dobijanje grafičkog prikaza promenljivih, funkcija `s.label` služi za prikaz skorova, njeni drugi i treći parametar predstavljaju odabir prve dve komponente za x i y osu, respektivno, dok funkcija `scatter` služi za iscrtavanje biplota, njen drugi parametar se koriste da se ne prikazuju scree plot, a treći da se opservacije prikažu kao tačkice.



Slika 3.48: Grafički prikaz promenljivih za podatke Happiness



Slika 3.49: Grafički prikaz skorova za podatke Happiness

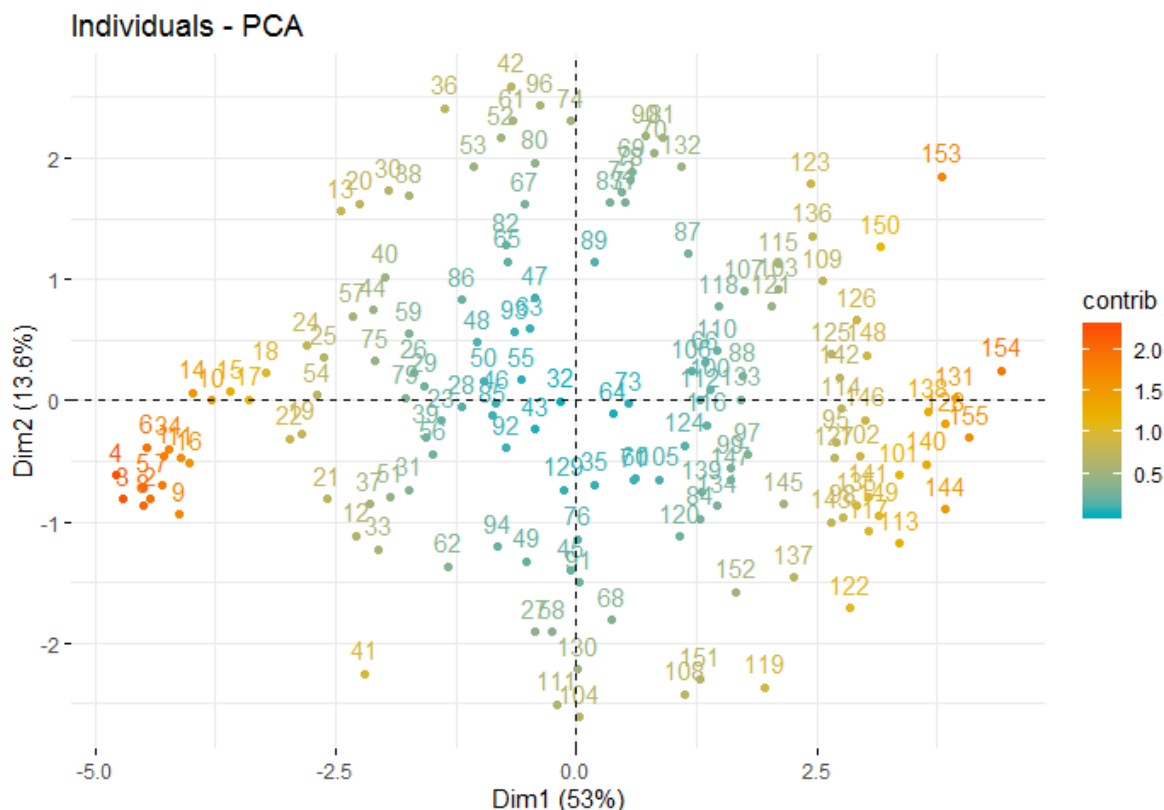


Slika 3.50: Grafički prikaz opservacija i promenljivih za podatke Happiness

Na osnovu slike 3.48 vidimo da su promenljive Positive affect, Freedom, Ladder, SD of Ladder, Healthy life expectancy i Log of GDP per capita one koje najviše utiču na objašnjene varijanse odnosno one najbolje predstavljaju originalan skup. Slika 3.49 grafički prikazuje opservacije, ali nije dobar način za tumačenje dobijenih podataka. Na slici 3.50 je prikazan biplot za promenljive i opservacije, on predstavlja mnogo jasniji način za tumačenje dobijenih podataka. Gore navedenim naredbama i slikama je prikazano kako se koriste funkcije iz paketa ade4. Na osnovu slika 3.51 i 3.52 će biti tumačeni dobijeni podaci.

	Dim.1	Dim.2	Dim.3	Dim.4
Ladder	-0.8971172	0.07254496	0.172376888	0.160470158
SD.of.Ladder	-0.7139434	0.38279451	-0.375110035	-0.210504720
Positive.affect	-0.5573487	-0.62259988	0.398600051	-0.154609252
Negative.affect	-0.7219363	0.08996600	-0.033352373	-0.599223083
Social.support	-0.8638888	0.20470042	0.136397537	-0.034937746
Freedom	-0.6672908	-0.60188936	0.095677714	-0.050154305
Corruption	-0.3276237	-0.45637568	-0.754217962	0.007996383
Generosity	-0.6375198	-0.23037977	-0.168184432	0.417105597
Log.of.GDP.per.capita	-0.8517960	0.26703461	0.002165833	0.175818999
Healthy.life.expectancy	-0.8474706	0.28013898	0.101340042	0.201133976

Slika 3.51: Prikaz koeficijenta promenljivih za podatke Happiness



Slika 3.52: Grafički prikaz opservacija u paketu FactoMineR za podatke Happiness

Prva komponenta je negativno korelisana sa promenljivim Ladder, SD of Ladder, Negative affect, Social support, Log of GDP per capita i Healthy life expectancy, dok je druga promenljiva negativno korelisana sa promenljivima Positive affect, Freedom i Corruption. Prva komponenta predstavlja dužinu trajanja i kvalitet života, kao i ekonomski i socijalni standard, dok druga predstavlja pozitivnost i slobodu življenja i svest o korupciji, a to znači da ove osobine u najvećoj meri utiču na sreću građana.

Na osnovu dobijenih glavnih komponenti, države u kojima građani sebe vide srećne su Finska, Danska Norveška, Island, Holandija, Švajcarska, Švedska, Novi Zeland i Kanada (na slici 3.52 na levoj strani grafika su označene tamnijom crvenom bojom). Na slici je pod brojem 41 država Uzbekistan, njeni građani su srećni u najvećoj meri zbog pozitivnosti i slobode življenja, kao i svesti o kriminalu, a zatim i zbog dužinu trajanja i kvaliteta života. Građani Italije i severnog Kipra se vide kao prosečno srećne (opservacije 37 i 63). Građani južnog Sudana i Centralne Afričke Republike, po ovoj analizi, ne vide sebe kao srećne (opservacije 154 i 155). Građani Srbije (opservacija 70) i BiH (opservacija 78) se ne vide kao srećni, na njihovo zadovoljstvo više utiču dužina trajanja i kvalitet života, dok su građani Hrvatske (opservacija 75) za koju nijansku srećniji, na osnovu dobijenih rezultata analize.

Napomena: U originalnom skupu je 156 država, ali je u ovom primeru isključena opservacija Kosovo. Takođe, u ovom primeru je bilo ukupno 15 missing value i to je rešeno zamenom vrednosti sa odgovarajućim srednjim vrednostima, zbog malog broja nedostajućih vrednosti. Ukoliko je broj missing value velik, u R-u postoji način i naredbe preko kojih se prevazilazi problem u sklopu paketa missMDA, više u [11].

Literatura

- [1] Anderson TW. *An introduction to multivariate statistical analysis*. New Delhi: Wiley Eastern Private Limited. 1972.
- [2] Brown JD. *Choosing the right number of components or factors in PCA and EFA*. Shiken: JALT Testing & Evaluation SIG News letter. Vol.13. No.2. 2009. [online] Dostupno na: http://hosted.jalt.org/test/bro_30.htm
- [3] Brown JD. *Choosing the right type of rotation in PCA and EFA*. Shiken: JALT Testing & Evaluation SIG News letter.13(3). 2009. [online] Dostupno na: <http://hosted.jalt.org/test/PDF/Brown31.pdf>
- [4] Chatfield C. Collins AJ. *Introduction to multivariate analysis*. Chapman & Hall, London and New York. 1980.
- [5] Critchley F. *Influence in principal component analysis*. Biometrika. Vol.72. 1985.
- [6] Cvetković Lj. Numela I. *Univerzitet u Novom Sadu. Fakultet tehničkih nauka u Novom Sadu*. 2015.
- [7] Everitt BS. Dunn G. *Applied multivariate data analysis*. London: Arnold. 2001.
- [8] Flury B. *Common principal components and related multivariate models*. New York. 1988.
- [9] Fulgosi A. *Faktorska analiza. Treće dopunjeno izdanje*. Školska knjiga Zagreb. 1988.
- [10] Hotelling H. *Analysis of a complex of statistical variables into principal components*. *Journal of Educational Psychology*, volume 24, 1933.
- [11] Husson F. Le S. Pages J. *Exploratory Multivariate Analysis by Example Using R*. 2nd edition. Boca Raton, Florida: Chapman & Hall/CRC, 2017. [online] Dostupno na: <http://factominer.free.fr/bookV2/index.html>
- [12] Jeffers JNR. *Two case studies in the application of principal component analysis*. *Applied Statistics*. 1967.
- [13] Johnstone IM. *On the distribution of the largest eigenvalue in principal component analysis*. *The Annals of Statistics*. 2001.
- [14] Johson RA. Wichern DW. *Applied multivariate statistical analysis*. 2nd edition. Englewood Cliffs. NJ: Prentice-Hall. 1988.
- [15] Jolliffe IT. *Principal component analysis*. 2nd edition. New York: Springer-Verlag New York. 2002.
- [16] Jong TP. *Some properties of a certain patterned matrix*. *J. Appl. Math.& Computing* Vol.15. 2004. [online] Dostupno na www.mathnet.or.kr/mathnet/kms_tex/982326.pdf
- [17] Kovačić Z.J. *Multivarijaciona analiza*. Univerzitet u Beogradu. Ekonomski fakultet u Beogradu. 1994.
- [18] Kaggle (2019). *Datasets*. [online] Dostupno na: <https://www.kaggle.com>

- [19] Lozanov-Crvenković Z. *Statistika. Univerzitet u Novom Sadu. Prirodno-matematički fakultet u Novom Sadu. 2012.*
- [20] McCormick C. *Mahalanobis distance. 2014. [online] Dostupno na: <http://mccormickml.com/2014/07/21/mahalanobis-distance/>*
- [21] Pearson K. *On lines and planes of closest fit to systems of points in space. Philosophical Magazine. Series 6. vol. 2. 1901.*
- [22] Pecina M. *Metode multivarijatne analize. Sveučilište u Zagrebu. Agronomski fakultet. 2006. [online] Dostupno na: http://www.agr.unizg.hr/multimedia/pdf/ds1905_metode_mva_osnove_2006.pdf*
- [23] Rao CR. *The use and interpretation of principal component analysis in applied research. Sankhya: The Indian Journal of Statistics. Volume 26. 1964.*
- [24] Raiche G. Riopel M. Blais JG. *Non graphical solutions for the Cattell's scree test. Montreal. 2006. [online] Dostupno na: https://www.researchgate.net/publication/235338705_Non-Graphical_Solutions_for_Cattell's_Scree_Test*
- [25] Rajter-Ćirić D. *Verovatnoća. Univerzitet u Novom Sadu. Prirodno-matematički fakultet u Novom Sadu. 2008.*
- [26] Richman MB. *Rotation of principal components. Journal of climatology. Vol.6. 1986.*
- [27] Tipping ME. Bishop CM. *Probabilistic principal component analysis. Journal of the Royal Statistical Society. Series B. Vol.61. 1999.*
- [28] Shlens J. *A tutorial on principal component analysis. Version 3.02. 2014 [online] Dostupno na : <https://arxiv.org/pdf/1404.1100.pdf>*
- [29] Sungkyu J. *The Wishart distribution. notes. [online] Dostupno na: <http://www.stat.pitt.edu/sungkyu/course/2221Spring15/lec2.pdf>*
- [30] Stojaković Z. Bošnjak I. *Elementi linearne algebre. Novi Sad: Symbol. 2010.*
- [31] Tabachnick BG. Fidell LS. *Using multivariate statistics. 5th edition. Upper Saddle River. NJ: Pearson Allyn & Bacon. 2007.*
- [32] Wall ME. Rechtsteiner A. Rocha LM. *Singular value decomposition and principal component analysis. AIChE Journal. 1995.*

Kratka biografija



Tanja Stanić je rođena 21. avgusta 1992. godine u Prnjavoru u Republici Srpskoj. Osnovnu školu "Nikola Tesla" završila je u Derventi 2007. godine. Gimnaziju "Gimnazija sa tehničkim školama", opšti smer, završila je u Derventi 2011. godine. Sve četiri godine bila je odličan učenik. Nakon toga, upisala je Prirodno-matematički fakultet, smer Primeњena matematika, modul Matematika finansija, gde je 2015. godine završila osnovne studije. Iste godine upisala je i master studije na Prirodno-matematičkom fakultetu. Položila je sve ispite predviđene nastavnim planom i programom master studija, zaključno sa septembarskim rokom 2017. godine. Od maja 2018. godine je zaposlena u Credit Agricole Srbija banci, prvo u okviru Digitalnog programa, a zatim na poziciji Stručni saradnik za analizu i unapređenje poslovnih procesa.

Novi Sad, 2019.

Tanja Stanić

**UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA**

Redni broj:

RBR

Identifikacioni broj:

IBR

Tip dokumentacije: *monografska dokumentacija*

TD

Tip zapisa: *tekstualni štampani materijal*

TZ

Vrsta rada: *master rad*

VR

Autor: *Tanja Stanić*

AU

Mentor: *dr Zagorka Lozanov-Crvenković*

MN

Naslov rada: *Analiza glavnih komponenti sa primerima u statističkim softverima R i Statistica*

NR

Jezik publikacije: *srpski (latinica)*

JP

Jezik izvoda: *s/e*

JI

Zemlja publikovanja: *Republika Srbija*

ZP

Uže geografsko područje: *Vojvodina*

UGP

Godina: *2019.*

GO

Izdavač: *autorski reprint*

IZ

Mesto i adresa: *Novi Sad, Trg Dositeja Obradovića 4*

MA

Fizički opis rada: *3 poglavlja, 100 strana, 32 lit. citat, 18 tabela, 52 grafika*

FO

Naučna oblast: *matematika*

NO

Naučna disciplina: *primenjena matematika*

ND

Ključne reči: *glavne komponente, kovarijansna matrica, karakteristični koreni, karakteristični vektori, korelaciona matrica, R, Statistica*

PO

UDK

Čuva se: u biblioteci Departmana za matematiku i informatiku, Prirodno-matematičkog fakulteta, u Novom Sadu

ČU

Važna napomena:

VN

Izvod: *Analiza glavnih komponenti omogućava objedinjenje i prikaz informacija za skup podataka koji sadrži opservacije odnosno opažanja opisan kvantitativnim promenljivima koje su međusobno korelisane. Svaka promenljiva se može posmatrati kao različita dimenzija. Analiza glavnih komponenti koristi se za dobijanje bitnih informacija iz multivarijantnog skupa podataka i za prikaz tih informacija u vidu skupa novih promenljivih koje se nazivaju glavne komponente. Te nove promenljive predstavljaju linearnu kombinaciju originalnih promenljivih. Broj glavnih komponenti je manji ili jednak broju originalnih promenljivih. Cilj ove metode je identifikacija pravaca odnosno glavnih komponenti duž kojih je varijansa maksimalna. Drugim rečima, ovom metodom smanjuje se dimenzionalnost multivarijantnih podataka na dve ili tri glavne komponente, koje se mogu grafički vizualizovati i to uz minimalan gubitak podataka. Ova metoda je ilustrovana kroz nekoliko primera u statičkim softverima R i Statistica.*

IZ

Datum prihvatanja teme od strane NN veća: 18. septembar 2019.

DP

Datum odbrane:

DO

Članovi komisije:

KO

Predsednik: *dr Ljiljana Gajić, redovni profesor*

Član: *dr Zagorka Lozanov-Crvenković, redovni profesor*

Član: *dr Ivana Štajner-Papuga, redovni profesor*

**UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCE
KEY WORDS DOCUMENTATION**

Accession number:

ANO

Identification number:

INO

Document type: *monograph type*

DT

Type of record: *printed text*

TR

Contents code: *master thesis*

CC

Author: *Tanja Stanić*

AU

Mentor: *dr Zagorka Lozanov-Crvenković*

MN

Title: *Principal component analysis and examples in statistical softwares R and Statistica*

XI

Language of text: *serbian (latin)*

LT

Language of abstract: *s/e*

LA

Country of publication: *Republic of Serbia*

CP

Locality of publication: *Vojvodina*

LP

Publication year: *2019.*

PY

Publisher: *author's reprint*

PU

Publ. place: *Novi Sad, Trg Dositeja Obradovića 4*

PP

Physical description: *3 sections, 100 pages, 32 references, 18 tables, 52 graphs*

PD

Scientific field: *mathematics*

SF

Scientific discipline: *applied mathematics*

SD

Key words: *principal component, covariance matrix, eigenvalue, eigenvector, correlation matrix, R, Statistica*

UC

Holding data: *Department of Mathematics and Informatics' Library, Faculty of Sciences, Novi Sad*

HD

Note:

N

Abstract: Principal component analysis (PCA) allows us to summarize and to visualize the information in a data set containing individuals/observations described by multiple inter-correlated quantitative variables. Each variable could be considered as a different dimension. Principal component analysis is used to extract the important information from a multivariate data table and to express this information as a set of few new variables called principal components. These new variables correspond to a linear combination of the originals. The number of principal components is less than or equal to the number of original variables. The goal of PCA is to identify directions (or principal components) along which the variation in the data is maximal. In other words, PCA reduces the dimensionality of a multivariate data to two or three principal components, that can be visualized graphically, with minimal loss of information. This method is illustrated by the few examples using statistical softwares R and Statistica.

AB

Accepted by the Scientific Board on: *18. September 2019*

ASB

Defended:

DE

Thesis defend board:

DB

President: *dr Ljiljana Gajić, full professor*

Member: *dr Zagorka Lozanov-Crvenković, full professor*

Member: *dr Ivana Štajner-Papuga, full professor*