Matematički fakultet , Univerzitet u Beogradu

Master rad

Tema:

Analiza glavnih komponenti

Mentor: Vesna Jevremović

Student: Božidarka Zlatić

oktobar, 2011.

Beograd

1. Uvod

Analiza glavnih komponenti (engl. *Principal component analysis*, *PCA*) je metod multivarijacione analize koja se koristi za smanjivanje dimenzije skupa podataka uz istovremeno zadržavanje maksimalno mogućeg varijabiliteta koji je prisutan u tim podacima. Mogućnosti analize glavnih komponenti prvi je opisao Karl Pearson (1901.), ali praktične primene razradio je Hotelling (1933.). Šira primena ove tehnike, zbog kompleksnog računa za veći broj promenljivih, pričekala je dostupnost računara.

Anlaiza glavnih komponenti je jedna od dve najčešće korišćene procedure faktorske analize. Često se faktorska analiza naziva tehnikom analize međuzavisnosti (engl. *analiysis of interdependence*), jer analizira nezavisnost pitanja, promenljivih ili objekata. Druga metoda faktorske analize je analiza skupina.

Tehnike faktorske analize možemo ilustrovati sledećim primerom. Pretpostavimo da želimo da ispitamo kako budući studenti biraju fakultet na kome će studirati. Prvi korak bi bio da odredimo kako studenti vide i ocenjuju te institucije. Studentima bismo mogli postaviti neka konkretna pitanja koja se tiču ovog istraživanja. Na primer: zašto im se dopada neki fakultet, zašto smatraju dva fakulteta sličnim, šta sve utiče na njihov izbor fakulteta (društvo, perspektivnost fakulteta, blizina, veličina fakulteta, dobri profesori..). Rezultat anketiranja studenata bi mogao da sadrži i preko sto stavki, tj. pitanja, pri čemu bismo mi dobili isto toliko i promenljivih. To bi mogle biti: veličina, blizina, dobri profesori, društvo, sportske organizacije, prespektivnost, bezličan, težak, skup...Drugi korak bi bio da pitamo studente koliko im je bitna svaka od ovih navedenih karakteristika fakulteta. U ovom delu bismo mogli naše istraživanje mnogo da zakomplikujemo zbog velikog broja promenljivih ili karakteristika. Mnoge od tih karakteristika bi mogle biti i suvišne jer mere istu stvar. Ukoliko je cilj našeg istraživanja da smanji broj bitnih karakteristika za dalju analizu, tj. da izbacimo one promenljive ili karakteristike koje su slične i mere istu stvar, u tom slučaju koristimo analizu glavnih komponenti. U daljem istraživanju može nas interesovati da identifikujemo grupe studenata prema tome šta očekuju od fakulteta. Možemo pretpostaviti da neku grupu studenata zanima niska cena školarine, drugu grupu da je fakultet što bliži kući, treću kvalitet obrazovanja, četvrtu društveni aspekti, itd. Analiza skupina se koristi za određivanje tih grupa. Ona se koristi za identifikovanje ljudi, objekata ili promenljivih koje formiraju prirodne grupe ili skupine.

U ovom radu, bavićemo se analizom glavnih komponenti, ciljevima i svrhom ove analize kao i njenom metodologijom.

1.1. Ciljevi analize glavnih komponenti

Analiza glavnih komponenti je metod kojim možemo indentifikovati obrazac u dobijenim podacima i predstaviti podatke na takav način da se istaknu njihove sličnosti i razlike.

Obrazac u podacima može biti veoma teško naći u slučaju velikih dimenzija. Tada je analiza glavnih komponenti moćan alat za analizu podataka. Druga glavna prednost ove analize je da nakon pronalaska obrasca u podacima potom smanjimo dimenziju podataka bez gubitka informacija.

Dakle, glavni cilj analize glavnih komponenti je da otkrije skrivenu strukturu skupa podataka. Na taj način možemo biti u stanju da:

▪ smanjimo broj promenljivih na prihvatljiv nivo, ali bez gubitaka informacija

▪ olakšamo interpretaciju originalnog skupa podataka

▪ indentifikujemo suštinski koncept koji leži u osnovi podataka.

1.2. Metodologija

Kao što je već napomenuto, osnovni zadatak metode glavnih komponenti je smanjenje dimenzije skupa podataka. Osnovna ideje za smanjenje broja promenljivih se ostvaruje kroz linearnu kombinaciju originalnih promenljivih. Dimenzije linearnih kombinacija su često lakše za tumačenje i služe kao međukorak u nekim složenijim analizama. Preciznije su one linearne kombinacije koje ostvaruju najveće razlike među vrednostima. Tj. u potrazi smo za linearnom kombinacijom sa najvećom disperzijom.

Uopštenije, zadatak analize glavnih komponenti je određivanje nekoliko linearnih kombinacija originalnih promenljivih koje će, pored toga što imaju maksimalnu disperziju, biti međusobno nekorelisane, gubeći u što je moguće manjoj meri informacije sadržane u skupu originalnih promenljivih. U postupku ove metode originalne promenljive se transformišu u nove promenljive (linearne kominacije) koje nazivamo *glavne komponente* ili *faktori.* Prva glavna komponenta konstruiše se tako da obuvata najveći deo disperzije originalnog skupa podataka, a naredne glavne komponente onaj deo disperzije originalnog skupa podataka koji nije obuhvaćen prethodno izdvojenim glavnim komponentama. Analiza je bazirana na pretpostavci da će nekoliko glavnih komponenti, čiji je broj znatno manji od broja originalnih promenljivih, biti dobra aproksimacija originalnog skupa podataka.

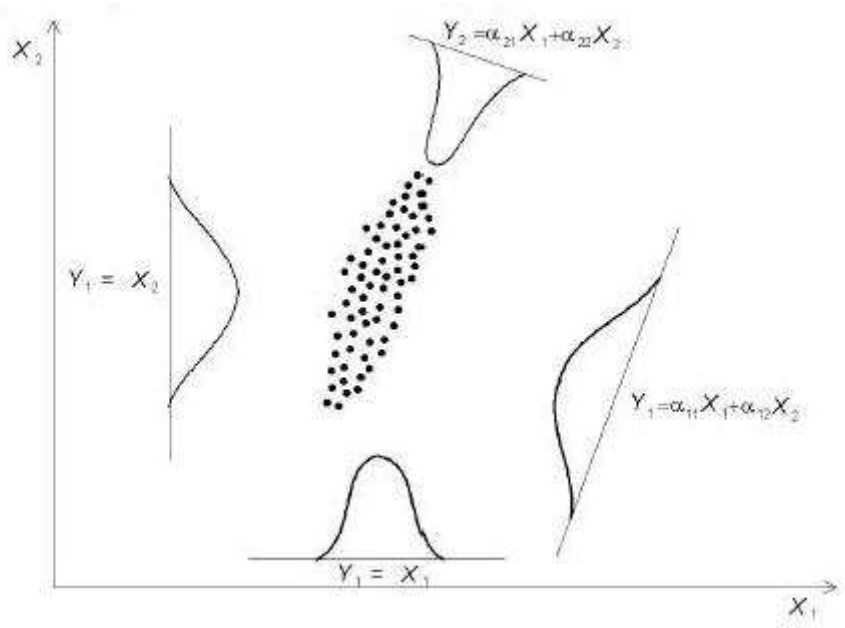
Izvršavajući ove zadatke metode analize glavnih komponenti, mi postižemo ciljeve ove analize. Prvi cilj, smanjenje dimenzije originalnog skupa podataka možemo još i opisati kao „sumiranje“ podataka. Ako je moguće višedimenzionalni skup podataka predstaviti preko manjeg broja linearnih kombinacija, tada ćemo na primer, umesto 10 promenljivih u daljoj analizi koristiti 2 linearne kominacije. Takođe, smanjenjem dimenzije originalnog skupa promenljivih olakšavamo interpretaciju originalne strukture podataka na bazi manjeg broja međusobno nekorelisanih glavnih komponenti koje je u ovom slučaju moguće prikazati i grafički.

1.3. Geometrijsko tumačenje glavnih komponenti

Poslužimo se jednostavnim primerom u kome na dvodimenzionalnom skupu promenljivih ilustrujemo osnovnu ideju metode i način konstruisanja glavnih komponenti. Na slici 1. prikazan je dijagram osipanja promenljivih X_1 i X_2. Linearna kombinacija ove dve promenljive je:

$$Y_1 = \alpha_{11} X_1 + \alpha_{12} X_2,$$

gde smo sa α_j označili koeficijent linearne kombinacije uz $j-$tu promenljivu u prvoj linearnoj kombinaciji.



Slika 1. Projekcije skupa (roja) tačaka

Ako izaberemo koeficijente $\alpha_{11} = 1$, $\alpha_{12} = 0$, odnosno $\alpha_{11} = 0$, $\alpha_{12} = 1$, dobićemo da je prva linearna kombinacija jednaka prvoj promenljivoj, odnosno da je druga linearna kombinacija jednaka drugoj promenljivoj. U geometrijskom smislu ovakvim izborom koeficijenata linearne kominacije dobijamo promenljivu Y_1 koja je formirana na osnovu projekcije skupa tačaka na X_1 i X_2 osu. Raspodela od Y_1 u prvom slučaju predstvalja marginalni raspored od X_1, a u drugom slučaju marginalni raspored od X_2. Ako se zahteva reprezentovanje dvodimenzionalnog skupa samo jednom promenljivom onda bismo izabrali onu koja ima veći varijabilitet. Ovakav izbor se objašnjava time da na osnovu

promenljive sa većim varijabilitetom možemo u većoj meri razlikovati pojedinačne observacije dvodimenzionalnog skupa. U ekstremnom slučaju, kada sve tačke roja (skupa) leže na pravoj upravnoj na X_1 osu, tada je dovoljno u analizi zadržati samo promenljivu X_2 jer ona nosi svu informaciju o varijabilitetu dvodimenzionalnog skupa podataka.

Sa slike 1, vidimo da je promenljiva X_2 kandidat za reprezenta dvodimenzionalnog skupa podataka jer ima veću disperziju od promenljive X_1. Postavlja se pitanje, da li postoji takav izbor koeficijenata linearne kombinacije koji će kao rezultat imati veću disperziju promenljive Y_1 nego što je to u slučaju prethodno navedenog izbora koeficijenata. Naš izbor koeficijenata svodi se na zadatak maksimiziranja disperzije linearne kombinacije uz uslov da je zbir kvadrata koeficijenata linearne kombinacije jednak jedinici. Geometrijski ovaj uslov znači da je vektor koeficijenata linearne kominacije $[\alpha_{11}, \alpha_{12}]^T$ jedinične dužine. Uslov normiranja se uvodi u cilju postizanja jednoznačne definisanosti linearne kombinacije. U geometrijskom smislu izborom koeficijenata menjamo ugao pod kojim projektujemo tačke iz roja (skupa) tačaka na pravu liniju. Biramo one koeficijente koji će dati projekciju tačaka sa najvećom disperzijom. Na prethodnoj slici, to je projekcija označena sa $Y_1 = \alpha_{11} X_1 + \alpha_{12} X_2$, odnosno, ta linearna kombinacija ima najveću disperziju od svih linearnih kombinacija koje se mogu dobiti promenom ugla projekcije. Ovu linearn kombinaciju, koja predstavlja projekciju roja (skupa) tačaka na prvu liniju sa najvećom disperzijom, nazivamo *prva glavna komponenta*.

Ako je za potrebe analize dovoljno izdvojiti jednu glavnu komponentu, onda se na ovom mestu zastavljamo. U suprotnom, formiramo sledeću linearnu kombinaciju $Y_2 = \alpha_{21} X_1 + \alpha_{22} X_2$. Njene koeficijente određujemo tako da joj se maksimizira disperzija uz normirajući uslov kao i kod prve glavne komponente, uz dodatni uslov da su Y_1 i Y_2 međusono nekorelisane. Uslov međusobne nekorelisanosti u geometrijskom smislu zahteva da prave linije na koje se projektuje roj (skup) tačaka kod prve i druge linearne kombinacije budu međusobno normalne.

Kako je određen položaj prve prave linije, druga linija leži pod pravim uglom u odnosu na prvu.

2. Matematička osnova analize glavnih komponenti

Najpre, u ovom odeljku ćemo dati neke osnovne matematičke pojmove koje treba dobro poznavati radi što boljeg razumevanja procesa analize glavnih komponenti, a potom ćemo preći na matematičke osnove samog procesa analize glavnih komponenti. Spomenućemo neke pojmove iz matematičke statistike, kao i deo matrične algebre koji se odnose na sopstvene vektore i sopstvene vrednosti, koje su s jedne strane važne osobine matrica, a sa druge od fundamentalnog značaja za analizu glavnih komponenti. Na kraju ovog poglavlja

biće objašnjen metod Lagranžovog multiplikatora za određivanje ekstremnih vrednosti funkcija više promenljivih.

2.1. Elementi matrične algebre

Matrica predstavlja šemu sastavljenu od p redova (vrsta) i n kolona sa $p \times n$ elemenata:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \left[a_{ij} \right]_{p \times n}.$$

Sa A^T ili A' označavamo transponovanu matricu matrice A, i nju dobijamo uzajamnom zamenom mesta redova i kolona matrice.

Za kvadratnu matricu $A(n \times n)$ kažemo da je simetrična ako je $A = A^T$, odnosno ako je $a_{ij} = a_{ji}$, za svako i, j.

Kvadratna matrica A^{-1} se naziva inverzna matrica date kvadratne matrice A ako je $A^{-1}A = AA^{-1} = I$, gde je I jedinična matrica.

Neka je $A(k \times k)$ kvadratna matrica i $I(k \times k)$ jedinična matrica. Skalari $\lambda_1, \lambda_2, ..., \lambda_k$ koji zadovoljavaju jednačinu:

$$\left| A - \lambda I \right| = 0$$

odnosno

$$\det(A - \lambda I) = 0,$$

nazivaju se karakteristični koreni ili sopstvene vrednosti matrice A. Jednačina $\left| A - \lambda I \right| = 0$ se naziva i karakteristična jednačina. Ako važi da je $\underset{(k \times 1)}{x} \neq 0$ i $Ax = \lambda x$, za $\mathbf{x}$ se kaže da predstavlja karakteristični vektor ili sopstveni vektor matrice A, pridružen karakterističnom korenu λ. Ekvivalentan uslov je:

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

Kada je matrica A simetrična matrica formata $n \times n$ tada postoji n realnih sopstvenih vrednosti $\lambda_1, \lambda_2, ..., \lambda_n$ i n realnih sopstvenih vektora $x_1, x_2, ..., x_n$.

Primer 2.1.1. Neka je $A=\begin{pmatrix}1 & 2\\ 2 & 3\end{pmatrix}$. Sopstvene vrednosti možemo naći na sledeći način:

$$|A-\lambda I|=0$$

$$\begin{vmatrix}1-\lambda & 2\\ 2 & 3-\lambda\end{vmatrix}=(1-\lambda)(3-\lambda)-4=0 \;\Rightarrow\; \lambda_1=2+\sqrt{5} \text{ i } \lambda_2=2-\sqrt{5}$$

Nađimo sopstvene vektore.

Za sopstvenu vrednost $\lambda_1=2+\sqrt{5}$, sopstveni vektor je:

$$\begin{pmatrix}1-2-\sqrt{5} & 2\\ 2 & 3-2-\sqrt{5}\end{pmatrix}\begin{pmatrix}x_{11}\\ x_{21}\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}$$

$$x_1=(x_{11},x_{21})^T=(0.5257,0.8506)^T$$

Za sopstvenu vrednost $\lambda_2=2-\sqrt{5}$, sopstveni vektor je

$$\begin{pmatrix}1-2+\sqrt{5} & 2\\ 2 & 3-2+\sqrt{5}\end{pmatrix}\begin{pmatrix}x_{12}\\ x_{22}\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}$$

$$x_2=(X_{12},X_{22})^T=(0.8506,-0.5257)^T.$$

Oni su ortogonalni jer je $x_1^T x_2=0$.

□

Navedimo neke važne karakteristike sopstvenih vrednosti i sopstvenih vektora i njihovu važnost u slučaju simetričnih matrica.

1. Ako su sopstvene vrednosti simetrične matrice različite, onda su sopstveni vektori ortogonalni. Ova osobina sopstvenih vektora je jako bitna za analizu glavnih komponenti. Ovo tvrđenje se jednostavno dokazuje. Neka su λ_i i λ_j dve različite sopstvene vrednosti i njihovi odgovarajući sopstveni vektori x_i i x_j. Tada je:

   $$Ax_i=\lambda_i x_i \quad \text{i} \quad Ax_j=\lambda_j x_j.$$

   Ako prvu jednakost pomnožimo sa x_j^T a drugu sa x_i^T sa leve strane, a zatim od prve oduzmemo drugu, dobijamo:

$$(\lambda_i - \lambda_j)x_j^T x_i = x_j^T A x_i - x_i^T A x_j = 0$$

jer je A simetrična matrica. Kako smo pretpostavili da su λ_i i λ_j različiti, odatle sledi da su x_i i x_j ortogonalni, odnosno:

$$x_j^T x_i = 0.$$

2. Sopstvene vrednosti simetrične matrice su realne.
3. Neka je A matrica formata $n \times n$ sa sopstvenim vrednostima $\lambda_i, i = 1,2,...n$, onda je determinanta:

$$\det(A) = \prod_{i=1}^{n} \lambda_i .$$

4. Za matrice A formata $n \times m$ i A^T formata $m \times n$, sopstvene vrednosti različite od nule od AA^T i $A^T A$ su iste i imaju istu mnogostrukost. Ako je $\mathbf{x}$ netrivijalan sopstveni vektor od AA^T za sopstvenu vrednost $\lambda \neq 0$, onda $y = A^T x$ je netrivijalan sopstveni vektor od $A^T A$. Ovo je vrlo važno svojstvo koje pokazuje vezu između sopstvenih vektora od AA^T i $A^T A$.

   Ako jednačinu $AA^T x = \lambda x$ pomnožimo sa obe strane sa A^T dobijamo:

   $A^T AA^T x = \lambda A^T x$.

   Ako zamenimo u jednačinu $y = A^T x$, dobijamo $A^T Ay = \lambda y$. Ako vektor x nije jednak nuli, nije onda ni $y = A^T x$, pa odatle sledi da je y sopstveni vektor.

Još jedna bitna stvar koju treba znati je da kada nađemo sopstvene vektore, želimo da njihova dužina bude tačno jedan. To je zato što dužina vektora ne utiče na to da li je sopstveni vektor ili ne. Tako da, kada god nađemo spstveni vektor, obično ga skaliramo na dužinu jedan, tako da svi sopstveni vektori imaju istu dužinu. Vektor možemo normirati na sledeći način:

$$\frac{\vec{x}}{|\vec{x}|}, \vec{x} \neq \vec{0}.$$

Nažalost, pronalaženje sopstvenih vektora je lako samo ako imamo matricu malih dimenzija, ne veću od 3×3. Za matrice većih dimenzija, pronalaženje sopstvenih vektora se malo komplikuje. Uobičajen način za pronaleženje sosptvenih vektora za matrice većeg fomata svodi se na iterativene metode. Postoji više numeričkih metoda za oređivanje

sopstvenih vrednosti i spostvenih vektora. Neke od njih su: metod stepenovanja, Jakobijeva metoda, itd.

Neka je A simetrična matrica formata $n \times n$. Onda za sopstvene vrednosti λ_i i odgovarajuće sopstvene vektore v_i, imamo:

$$Av_i = \lambda_i v_i \ , \ i = 1,...n .$$

Ovih n jednačina možemo zapisati matričnoj formi na sledeći način:

$$AV = V\Lambda$$

gde je $V = [v_1, v_2, ..., v_n]$, odnosno, V je matrica čije su kolone sopstveni vektori. A $\Lambda = diag(\lambda_1, ..., \lambda_n)$. Kako smo pretpostavili da je A simetrična, iz svojstva 1, v_i su ortogonalni. Kako je V ortogonalna matrica, množenjem obe strane sa V^T, i koristeći $VV^T = I$, dobijamo da je:

$$A = V\Lambda V^T .$$

Ovo se zove *dekompozicija sopstvenih vrednosti* od A. Bilo koju simetričnu matricu možemo rastaviti na ovaj način. Kako bismo odredili A, dovoljne su nam samo sopstvene vrednosti i sopstveni vektori. Ako su sopstvene vrednosti različite, dekompozicija je jedinstvena.

Osobine matrica:

Neka su A i B kvadratne matrice formata $k \times k$. Tada važi:

1. $|A| = |A^T|$
2. Ako je A nesingularna matrica, tada je $|A||A^{-1}| = 1$
3. $|AB| = |A||B|$

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

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

2.2. Pojmovi matematičke statistike

2.2.1. Standardna devijacija i disperzija

Neka je dat uzorak $X = (X_1, X_2, ..., X_n)$. Možemo izračunati srednju vrednost uzorka na sledeći način

$$\overline{X} = \frac{1}{n}\sum_{i=}^{n} X_i .$$

Međutim, srednja vrednost uzorka nam ne govori puno o uzorku. Na primer, možemo imati dva različita uzorka, a njihove srednje vrednosti mogu biti jednake: $(0,8,12,20)$ i $(8,9,11,12)$. Šta je različito kod ova dva uzorka? To je standardna devijacija koja predstavlja meru širenja podataka. Standardnu devijaciju možemo još definisati i kao prosečnu udaljenost od sredine skupa podataka. Standadnu devijaciju računamo na sledeći način:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} .$$

Pa ako uporedimo prethodna dva uzorka i njihove standardne devijacije, videćemo da je standardna devijacija prvog ($S = 8.3266$) veća od standardne devijacije drugog ($S = 1.825$), zbog činjenice da su podaci mnogo širu u odnosu na srednju vrednost.

Disperzija je još jedna mera širenja podataka u odnosu na srednju vrednost. Računamo je na sledeći način:

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

Prednost korišćenja standardne devijacije u odnosu na disperziju kao mere rasipanja jeste da je izražena u istim mernim jedinicama kao i srednja vrednost.

2.2.2.Kovarijansa

Neka su X i Y slučajne veličine, tada kovarijansa pokazuje kako se one zajedno menjaju, odnosno kovarijansa je mera zavisnosti između dve promenljive. Formula za računanje kovarijanse između dve promenljive je:

$$\mathrm{cov}(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y}).$$

Iz prethodnog izraza vidi se da kovarijansa predstavlja aritmetričku sredinu proizvoda odstupanja vrednosti promenljive X od njene aritmetričke sredine.

Imamo rezultate ispitivanja studenata koje smo pitali koliko sati ukupno troše na učenje i koju ocenu dobiju na kraju. Dakle, imamo dve promenljive. Prva je X - vreme učenja, a druga je Y - ocena koju dobiju. Ako je vrednost kovarijanse pozitivna ona nam govori da se obe promenljive povećavaju zajedno, tj. ako se povećava broj sati učenja onda se povećava i ocena. Ako je vrednost kovarijanse negativna, onda se jedna promenljiva povećava a druga smanjuje, npr. broj časova učenja se povećava a konačna ocena se smanjuje. Ako su sve vrednosti (barem jedne promenljive) međusobno jednake kovarijansa je jednaka nulu. Kovarijansa za realne promenljive (a o takvim promenljivim je ovde reč) je simetrična, odnosno važi:

$$\mathrm{cov}(X,Y)=\mathrm{cov}(Y,X).$$

2.2.3. Kovarijansna matrica

Podsetimo da se kovarijansa uvek meri između dve promenljive. Ako imamo skup podataka koji je veći od dva, onda imamo više merenja kovarijanse. Npr. ako imao trodimenzionalni skup podataka (X,Y,Z), tada računamo $\mathrm{cov}(X,Y)$, $\mathrm{cov}(X,Z)$ i $\mathrm{cov}(Y,Z)$. Kada imamo $n-$dimenzionalni skup podataka, korisno je sve dobijene kovarijanse između različitih promenljivih staviti u matricu. Naprimer, definicija kovarijansne matrice za skup podataka (X,Y,Z) je:

$$C=\begin{pmatrix}\mathrm{cov}(X,X) & \mathrm{cov}(X,Y) & \mathrm{cov}(X,Z)\\ \mathrm{cov}(Y,X) & \mathrm{cov}(Y,Y) & \mathrm{cov}(Y,Z)\\ \mathrm{cov}(Z,X) & \mathrm{cov}(Z,Y) & \mathrm{cov}(Z,Z)\end{pmatrix}.$$

Na glavnoj dijagonali kovarijansne vrednosti su između jedne promenljive, pa su one jednake disperziji tih promenljivih.

2.2.4. Višedimenzionalne slučajne promenljive

Definicija

Neka je dato p jednodimenzionalnih slučajnih promenljivih $X_1, X_2, ..., X_p$. Skup ovih slučajnih promenljivih pišemo kao $(p \times 1)$ *slučajni vektor* $\mathbf{X}$,

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

Dakle, X je slučajan vektor čiji su elementi slučajne promenljive. Kolekciju slučajnih vektora možemo predstaviti u vidu matrice. Takvu matricu čiji su elementi slučajne promenljive nazivamo *slučajna matrica*.

Funkcija raspodele $(p \times 1)$ slučajnog vektora $\mathbf{X}$ definiše se na sledeći način:

$$F_{\mathbf{X}}(\mathbf{x}) = F_{\mathbf{X}}(x_1, x_2, ..., x_p) = P\{\mathbf{X} \le \mathbf{x}\} = P\{X_1 \le x_1, X_2 \le x_2, ..., X_p \le x_p\}.$$

Funkcija raspodele neprekidnog slučajnog vektora $\mathbf{X}$ može da se definiše se i na sledeći način:

$$F_{\mathbf{X}}(\mathbf{X}) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_3} f_X(\mathbf{X}) dx_1 dx_2 \ldots dx_p,$$

gde je $f_{\mathbf{X}}(\mathbf{X}) = f_{\mathbf{X}}(x_1, x_2, ..., x_p)$ višedimenzionalna funkcija gustine od $\mathbf{X}$. Do višedimenzionalne funkcije gustine slučajnog vektora dolazimo diferencirajući višedimenzionalnu funkciju raspodele $F_{\mathbf{X}}(\mathbf{X})$. Višedimenzionalna funkcija gustine $f_{\mathbf{X}}(\mathbf{X}) = f_{\mathbf{X}}(x_1, x_2, ..., x_p)$ ima sledeće osobine: 1.) $f_{\mathbf{X}}(\mathbf{X}) \ge 0$, za svako $x \in R^p$ i 2.) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_X(\mathbf{X}) dx_1 dx_2 \ldots dx_p = 1.$

Marginalna raspodela

Funkcija raspodele ili gustina raspodele jedne slučajne promenljive dobijene na osnovu višedimenzionalne funkcije raspodele ili višedimenzionalne funkcije gustine nazivamo marginalna jednodimenzionalna funkcija raspodele, odnosno marginalna jednodimenzionalna funkcija gustine. Pretpostavimo da je poznata funkcija raspodele, $F_{\mathbf{X}}(\mathbf{X})$. Tada do marginalne funkcije raspodele slučajne promenljive X_1 dolazimo na osnovu izraza

$$F_{X_1}(x_1) = F_{\mathbf{X}}(x_1, \infty, \infty, ..., \infty),$$

a ako nam je poznata funkcija gustine, $f_X(\mathbf{x})$, tada marginalnu fnkciju gustine od X_1 dobijamo na osnovu izraza

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x_1, x_2, ..., x_p) dx_2 \ldots dx_p .$$

Srednja vrednost i kovarijaciona matrica

Neka je $\mathbf{X}(p \times 1)$ slučajan vektor, čiji svaki element predstavlja jednodimenzionalnu slučajnu promenljivu sa svojom marginalnom raspodelom. Za svaku jednodimenzionalnu slučajnu promenljivu možemo odrediti sredinu $\mu_j = E(X_j)$ i disperziju $\sigma_j^2 = E(X_j - \mu_j)^2$, koju još označavamo i sa $D(X_j)$. Sredina slučajnog vektora $\mathbf{X}$ je $(p \times 1)$ vektor čiji su elementi $\mu_j = E(X_j)$, $j = 1, ..., p$, i označavamo ga sa $\boldsymbol{\mu}$

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

Kao što smo već napomenuli, za ma koji par slučajnih promenljivih X_j i X_k definišemo kovarijansu na sledeći način:

$$\sigma_{jk} = E\left[(X_j - \mu_j)(X_k - \mu_k)\right] = Cov(X_j, X_k)$$

$$\sigma_{ii} = \sigma_i^2 = Cov(X_i, X_i) = D(X_i)$$

$$Cov(X_i, X_k) = Cov(X_k, X_j) = \sigma_{kj} = \sigma_{jk}$$

Za slučajan vektor $\mathbf{X}$ definišemo $p \times p$ simetričnu matricu kod koje je $j-$ti dijagonalni element $\sigma_{ii} = D(X_i)$, a čiji je $(j,k)-$element $\sigma_{jk} = Cov(X_j, X_k)$, $j \neq k$. Ovu matricu nazivamo kovarijaciona matrica (kovarijansna matrica) od $\mathbf{X}$ i označavamo je sa $D(\mathbf{X})$ ili $Cov(\mathbf{X})$, odnosno $\mathbf{\Sigma}$. Tako je

$$Cov(\mathbf{X}) = \mathbf{\Sigma} = \left[\sigma_{jk}\right] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} = \begin{bmatrix} D(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_p) \\ Cov(X_1, X_2) & D(X_2) & \cdots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_p) & Cov(X_2, X_p) & \cdots & D(X_p) \end{bmatrix}$$

Kovarijacionu matricu možemo iskazati i kao očekivanu vrednost slučajne matrice.

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

Ova slučajna matrica proizvod je slučajnih vektora odstupanja od sredine, tj. $(\mathbf{X} - \mathbf{\mu})(\mathbf{X} - \mathbf{\mu})^T$, pa je njena očekivana vrednost:

$$E\left[(\mathbf{X} - \mathbf{\mu})(\mathbf{X} - \mathbf{\mu})^T\right] = \mathbf{\Sigma}.$$

Navedimo nekoliko važnih rezultata za osobine sredine slučajnog vektora $\mathbf{X}$ i njegove kovarijacione matrice. Pre toga navedimo poznato svojstvo kovarijanse. Neka su slučajne promenljive X_j i X_k linearno transformisane. To znači da su definisane nove slučajne promenljive $cX_j + a$ i $dX_j + b$, gde su a, b, c i d realne konstante. Na osnovu definicije kovarijanse, sledi da je

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

Uopštimo ovo na slučaj linearne kombinacije p slučajnih promenljivih iz slučajnog vektora $\mathbf{X}$, sa sredinom $\mathbf{\mu}$ i kovarijacionom matricom $\mathbf{\Sigma}$. Linearnom kombinacijom $Y = a_1X_1 + a_2X_2 + \ldots + a_pX_p = \mathbf{a^T X}$, za dati vektor koeficijenata linearne kombinacije: $\mathbf{a}^T = \left[a_1, a_2, \ldots, a_p\right]$, definišemo novu slučajnu promenljivu Y čija funkcija gustine $f_Y(y)$ zavisi $f_X(x)$. Njena očekivana vrednost je

$$\mu_Y = E(Y) = E(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\boldsymbol{\mu},$$

i disperzija

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

Znači da je disperzija od Y data kao kvadratna forma i u potpunosti je određena kovarijacionom matricom $\boldsymbol{\Sigma}$ slučajnog vektora $\mathbf{X}$ i koeficijentima $a_1, a_2, \ldots, a_p$.

Razmotrimo opšti slučaj q linearnih kombinacija p slučajnih promenljivih:

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

Ove linearne kombinacije u matričnoj formi možemo zapisati na sledeći način: $\mathbf{Y} = \mathbf{AX}$, gde je $\mathbf{Y}$ vektor formata $q \times 1$ i $\mathbf{A}(q \times p)$ matrica koeficijenata linearnih kombinacija. Sredina slučajnog vektora $\mathbf{Y}$ je

$$\boldsymbol{\mu}_{\mathbf{Y}} = E(\mathbf{Y}) = E(\mathbf{AX}) = \mathbf{A}\boldsymbol{\mu}_{\mathbf{X}}$$

A kovarijaciona matrica

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = Cov(\mathbf{Y}) = Cov(\mathbf{AX}) = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{A}^T,$$

gde je $\boldsymbol{\mu}_{\mathbf{X}}$ sredina slučajnog vektora $\mathbf{X}$, i $\boldsymbol{\Sigma}_{\mathbf{X}}$ njegova kovarijaciona matrica.

Korelaciona matrica

Koeficijent korelacije između dve slučajne promenljive X_j i X_k definišemo kao

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}},$$

što predstavlja normalizovanu kovarijansu između X_j i X_k. Koeficijent korelacije uzima vrednsot iz intervala [-1 , 1]. Ukoliko ρ_{jk} ima vrednost blisku donjoj ili gornjoj granici intervala, tada možemo reći da postoji mogućnost dobre linearne veze između X_j i X_k, i to sa negativnim, odnosno pozitivnim predznakom.

Korelacionu matricu $\boldsymbol{\rho}$ možemo dakle dobiti na osnovu poznate kovarijacione matrice. U matričnom zapisu, veza između korelacione i kovarijacione matrice data je sa:

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

$$= \begin{bmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\sigma_{pp}}} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\sigma_{pp}}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

gde smo sa $\mathbf{D}$ označili dijagonalnu matricu koja sadrži elemente na glavnoj dijagonali kovarijacione matrice $\boldsymbol{\Sigma}$. Na osnovu uspostavljene relacije između korelacione i kovarijacione matrice, imamo da važi i $\boldsymbol{\Sigma} = \mathbf{D}^{1/2}\boldsymbol{\rho}\mathbf{D}^{1/2}$.

Generalizovana disperzija

Jedan od najčešće korišćenih pokazatelja stepena raspršenosti podataka u slučaju jedne promenljive jeste disperzija ili standardna devijacija. Za promenljive čija je dimenzija $p \geq 2$ definisana je kovarijaciona matrica koja pruža informaciju o disperziji i kovarijansi promenljivih. Interes nam je da u višedimenzionalnom slučaju definišemo sintetički pokazatelj koji će na osnovu jednog broja iskazati stepen varijabiliteta p – dimenzionalnog skupa podataka. Koriste se dve definicije generalizovane disperzije. Prema prvoj, češće korišćenoj definiciji, ona je u uzorku jednaka determinanti uzoračke kovarijacione matrice, u oznaci $|S|$, a prema drugoj, generalizovana varijansa je jednaka tragu uzoračke kovarijacione matrice (zbir elemenata na glavnoj dijagonali matrice), u oznaci $tr(S)$.

2.3. Metod Lagranževog multiplikatora

Ekstrem funkcije $f(x,y)$ uz dato ograničenje $\varphi(x,y)=0$ naziva se uslovnim ili vezanim, a tako dobijeni maksimum ili minimum uslovnim maksimumom ili minimumom. Tada funkcija $f(x,y)$ zavisi samo od jednog argumenta, a zadatak je moguće svesti na problem ekstrema funkcije sa jednim argumentom. Zbog teškoća koje se mogu javljati prilikom izračunavanja promenljive x ili y iz jednačine $\varphi(x,y)=0$ i zamene u funkciju $f(x,y)$, za traženje ekstrema u ovakvim slučajevima koristi se metod koji se naziva metod Lagranžovog multiplikatora.

Prvo postavljamo Lagranžovu funkciju:

$$F(x,y)=f(x,y)+\lambda(x,y),$$

gdje je λ konstanta koja se naziva Lagranžov multiplikator. Potreban uslov za postojanje ekstrema funkcije $f(x,y)$, u tački (x_0,y_0) je da su parcijalni izvodi prvog reda Lagranžove funkcije u toj tački jednaki nuli, tj.

$$\frac{\partial F}{\partial x}=\frac{\partial f(x_0,y_0)}{\partial x}+\lambda\frac{\partial \varphi(x_0,y_0)}{\partial x}=0,$$

$$\frac{\partial F}{\partial y}=\frac{\partial f(x_0,y_0)}{\partial y}+\lambda\frac{\partial \varphi(x_0,y_0)}{\partial y}=0,$$

$$\frac{\partial F}{\partial y}=\varphi(x_0,y_0)=0.$$

Dovoljan uslov za postojanje ekstrema funkcije $f(x,y)$ u tački (x_0,y_0) obuhvata ispitivanje potrebnog uslova i

$$d^2F(x_0,y_0,\lambda_0)=\frac{\partial^2F(x_0,y_0,\lambda_0)}{\partial x^2}dx^2+2\frac{\partial^2F(x_0,y_0,\lambda_0)}{\partial x\partial y}dxdy+\frac{\partial^2F(x_0,y_0,\lambda_0)}{\partial y^2}dy^2$$

gde je $\frac{\partial\varphi}{\partial x}dx+\frac{\partial\varphi}{\partial y}dy=0$.

3. Analiza glavnih komponenti

Nakon uvodnog poglavlja u kome iznosimo osnovne zadatke i ciljeve metode glavnih komponenti, i matematičke osnove ove analize, u okviru ovog poglavlja dajemo formalnu definiciju i osnovne osobine glavnih komponenti kao i način njihove interpretacije. Nakon toga, bavimo se testiranjem značajnosti glavnih komponenti, određivanjem broja glavnih komponenti kao i njihovom rotacijom i značajem te rotacije.

3.1. Glavne komponente

3.1.1. Definicija glavnih komponenti

Pretpostavimo da je $\mathbf{X}$ $p-$dimenzionalni slučajni vektor sa kovarijacionom matricom $\mathbf{\Sigma}$. Neka je

$$Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \ldots + \alpha_{1p}X_p = \boldsymbol{\alpha}_1^T \mathbf{X}$$

linearna kombinacija elemenata slučajnog vektora $\mathbf{X}$, gde su $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p}$ koeficijenti linearne kombinacije. U poglavlju matematičke osnove, već smo pokazali da je

$$D(Y_1) = D(\alpha_1^T X) = \alpha_1^T \Sigma \alpha_1 .$$

Naš zadatak je da odredimo vektor koeficijenata $\boldsymbol{\alpha}_1$ tako da se maksimizira disperzija od Y_1. Kako se $D(Y_1) = \boldsymbol{\alpha}_1^{\mathrm{T}} \mathbf{\Sigma} \boldsymbol{\alpha}_1$ može proizvoljno povećati množenjem vektora $\boldsymbol{\alpha}_1$ proizvoljnim skalarom, time uvodimo ograničenje da je vektor koeficijenata jedinične dužine, tj. da je

$$\boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{\alpha}_1 = 1 .$$

Problem maksimiziranja $D(Y_1) = \boldsymbol{\alpha}_1^{\mathrm{T}} \mathbf{\Sigma} \boldsymbol{\alpha}_1$ uz ograničenje da je $\boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{\alpha}_1 = 1$ rešavamo korišćenjem Lagranžovog multiplikatora tako što ćemo maksimizirati Lagranžovu funkciju

$$\boldsymbol{\alpha}_1^{\mathrm{T}} \mathbf{\Sigma} \boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{\alpha}_1 - 1)$$

gde je λ Lagranžov multiplikator. Diferenciranjem Lagranžove funkcije po koeficijentima $\boldsymbol{\alpha}_1$, i izjednačavanjem dobijenog izraza sa nulom, dobijamo

$$\mathbf{\Sigma} \boldsymbol{\alpha}_1 - \lambda \boldsymbol{\alpha}_1 = 0$$

ili

$$(\boldsymbol{\Sigma} - \lambda \mathbf{I})\boldsymbol{\alpha}_1 = 0,$$

gde je $\mathbf{I}$ jedinična matrica formata $p \times p$. Da bi se dobilo netrivijalno rešenje za $\boldsymbol{\alpha}_1$ determinanta $|\boldsymbol{\Sigma} - \lambda \mathbf{I}|$ mora biti jednaka nuli. To znači da λ mora biti jedan od karakterističnih korena kovarijacione matrice $\boldsymbol{\Sigma}$. Odluka o izboru jednog od karakterističnih korena donosimo na sledeći način. Ako pomnožimo s leve strane izraz $\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1 = 0$ sa $\boldsymbol{\alpha}_1^T$ dobićemo

$$\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_1 = 0.$$

Kako težimo da maksimiziramo disperzuju, za λ ćemo uzeti najveću sopstvenu vrednost, recimo λ_1. Na osnovu uslova $(\boldsymbol{\Sigma} - \lambda \mathbf{I})\boldsymbol{\alpha}_1 = 0$, sledi da je $\boldsymbol{\alpha}_1$ odgovarajući sopstveni vektor pridružen sopstvenoj vrednosti λ_1. Njegovim normiranjem ($\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_1 = 1$) dobićemo traženi vektor $\boldsymbol{\alpha}_1$.

Ako želimo da odredimo više od jedne linearne kombinacije tada postupamo kao u slučaju određivanja prve glavne komponente uz uslov da kovarijansa prve i druge glavne komponente bude jednaka nuli. Neka je druga linearna kombinacija:

$$Y_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + \ldots + \alpha_{2p}X_p = \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}$$

čije koeficijente $\alpha_{21}, \alpha_{22}, \ldots, \alpha_{2p}$ treba odrediti uz uslov $\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_2 = 1$, pri čemu se uslov nekorelisanosti prve i druge glavne komponente svodi na $\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_1 = 0$. Ova činjenica sledi iz toga što je

$$\mathrm{cov}(Y_2, Y_1) = \mathrm{cov}(\boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X}) = \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_1\lambda_1 = \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_2\lambda_1,$$

pošto je $\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1\lambda_1$, a $\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_1\lambda_1 = 0$ samo kada je $\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_1 = 0$. Formiramo Lagranžovu funkciju sa dva množitelja

$$\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda(\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_2 - 1) - \phi\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_1$$

gde su λ i ϕ Lagranžovi množitelji. Diferenciranjem po $\boldsymbol{\alpha}_2$, a zatim izjednačavanjem dobijenog izraza sa nulom dobijamo

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1 = 0.$$

Ako pomnožimo dobijeni izraz sa leve strane sa $\boldsymbol{\alpha}_1^T$ dobijamo

$$\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_1 = 0.$$

Kako su prva dva člana u prethodnom izrazu jednaka nuli a $\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_1 = 0$, sledi da je $\phi = 0$. Prema tome je

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 = 0,$$

odnosno

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I})\boldsymbol{\alpha}_2 = 0\,.$$

Tada je λ karakteristični koren kovarijacione matrice $\boldsymbol{\Sigma}$, a $\boldsymbol{\alpha}_2$ odgovarajući karakteristični vektor. Kao i u slučaju prve glavne komponente biramo za λ što je moguće veću vrednost, jer je

$$\lambda = \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2\,.$$

Drugu po veličini sopstvenu vrednost označavamo sa λ_2 , a njen ogovarajući sopstveni vektor je $\boldsymbol{\alpha}_2$, a linearna kombinacija $Y_2 = \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}$ predstavlja drugu glavnu komponentu.

Na ovaj način možemo doći do svih glavnih komponenti kojih može biti koliko ima različitih sopstvenih vrednosti kovarijacione matrice. Ako su sve sopstvene vrednosti matrice $\boldsymbol{\Sigma}$ međusobno različite, i neka su uređene u opadajući niz $\lambda_1 > \lambda_2 > \ldots > \lambda_p \geq 0$, tada postoji p glavnih komponenti: $Y_1, Y_2, \ldots, Y_p$. Vektori koeficijenata $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p$ predstavljaju sopstvene vektore matrice $\boldsymbol{\Sigma}$ koji su pridruženi sopstvenim vrednostima $\lambda_1, \lambda_2, \ldots, \lambda_p$. Očekivane vrednosti glavnih komponenti su $E(Y_j) = 0$, $j = 1, \ldots, p$, disperzije $D(Y_j) = \lambda_j$, $j = 1, \ldots, p$, a kovarijansa svakog para glavnih komponenti je jednaka nuli.

Primer 3.1.1. Neka je data sledeća kovarijaciona matrica:

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

Odrediti sve glavne komponente.

Rešenje:

Odredimo prvo sopstevene vrednosti. Njih određujemo na osnovu jednačine:

$$\left|\boldsymbol{\Sigma} - \lambda \mathbf{I}\right| = 0.$$

Imamo tri sopstvene vrednosti, i to su: $\lambda_1 = 5.2361$, $\lambda_2 = 4$, $\lambda_3 = 0.7639$. Sopstvene vektore određujemo iz jednačine $(\boldsymbol{\Sigma} - \lambda \mathbf{I})\boldsymbol{\alpha}^T = 0$. Imamo tri sopstvena vektora:

$$\boldsymbol{\alpha}_1^T = \begin{bmatrix} 0 & -0.5257 & 0.8507 \end{bmatrix}$$

$$\boldsymbol{\alpha}_2^T = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{\alpha}_3^T = \begin{bmatrix} 0 & -0.8507 & -0.5257 \end{bmatrix}.$$

Glavne komponente određujemo iz formule: $Y = \boldsymbol{\alpha}^{\mathbf{T}}\mathbf{X}$. Pa su tri glavne komponente sledeće:

$$Y_1 = \boldsymbol{\alpha}_{\mathbf{1}}^{\mathbf{T}}\mathbf{X} = -0.5257 X_2 + 0.8507 X_3$$

$$Y_2 = \boldsymbol{\alpha}_{\mathbf{2}}^{\mathbf{T}}\mathbf{X} = X_1$$

$$Y_3 = \boldsymbol{\alpha}_3^{\mathbf{T}}\mathbf{X} = -0.8507 X_2 - 0.5257 X_3$$

Pokažimo još da su disperzije glavnih komponenti jednake odgovarajućim sopstvenim vrednostima kovarijacione matrice. Pokažimo to na primeru prve glavne komponente, za ostale dve je analogno.

$$D(Y_1) = D(\alpha_1^T X) = D(-0.5257 X_2 + 0.8507 X_3) = (-0.5257)^2 D(X_2) + (0.8507)^2 D(X_3) +$$
$$2(-0.5257)(0.8507)\operatorname{cov}(X_2, X_3) = 0.2764 \cdot 2 + 0.7236 \cdot 4 - 0.8944(-2) = 5.2361 = \lambda_1$$

Pokažimo i nekorelisanost glavnih komponenata na primeru prve i druge glavne komponente, a za ostale se pokazuje analogno.

$$\operatorname{cov}(Y_1, Y_2) = \operatorname{cov}(-0.5257 X_2 + 0.8507 X_3, X_1) = -0.5257 \operatorname{cov}(X_2, X_1) + 0.8507 \operatorname{cov}(X_3, X_1) =$$
$$= -0.5257 \cdot 0 + 0.8507 \cdot 0 = 0.$$

□

3.1.2. Osobine glavnih komponenti

Osobine glavnih komponenti koje slede na osnovu definicije su:

$$E(Y_j) = 0,\ D(Y_j) = \lambda_j,\ \operatorname{cov}(Y_i, Y_j) = 0,\ i \neq j$$

$$D(Y_1) \geq D(Y_2) \geq \ldots \geq D(Y_p) \geq 0$$

Sada ćemo pokazati da su generalizovane disperzije glavnih komponenti jednake generalizovanim disperzijama originalnog skupa promenljivih. Pre toga ćemo se podsetiti na rezultate matrične algebre koji su nam potrebni za ovaj dokaz.

Neka je $\mathbf{Y}$ vektor glavnih komponenti takav da je $\mathbf{Y}^T = (Y_1, Y_2, \ldots, Y_p)$. Sada se transformacija originalnog skupa promenljivih sadržanog u vektoru $\mathbf{X}$ može zapisati na sledeći način:

$$\mathbf{Y} = \mathbf{AX},$$

gde je $\mathbf{A}$ matrica formata $p \times p$ čiji su redovi sopstveni vektori kovarijacione matrice $\mathbf{\Sigma}$, odnosno $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p$, pridruženi odgovarajućim sopstvenim vrednostima $\lambda_1, \lambda_2, \ldots, \lambda_p$. Na osnovu osobina sopstvenih vektora ($\boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j = 1$ i $\boldsymbol{\alpha}_i^T \boldsymbol{\alpha}_j = 0, \ i \neq j$) matrica $\mathbf{A}$ ima osobinu da je $\mathbf{A}^T = \mathbf{A}^{-1}$, pa se $\mathbf{Y} = \mathbf{AX}$ naziva *ortogonalna transformacija* ili *rotacija*, a sama matrica $\mathbf{A}$ ortogonalna matrica. Njena osobina je i da je $|\mathbf{A}| = \pm 1$. Transformacija se naziva ortogonalna jer se sa njom vrši rotacija koordinatnih osa za izvestan ugao, pri čemu ose ostaju upravne jedna na drugu, a ugao između ma koja dva vektora ostaje isti nakon transformacije.

Korišćenjem ortogonalne matrice $\mathbf{A}$ možemo izvršiti ortogonalnu dekompoziciju kvadratne simetrične matrice čiji su koreni različiti. Imamo da je $\mathbf{\Sigma} = \mathbf{A}^T \mathbf{\Lambda A}$, gde je $\mathbf{\Lambda}$ dijagonalna matrica čiji su elementi sopstvene vrednosti matrice $\mathbf{\Sigma}$, a matrica $\mathbf{A}$ je ortogonalna matrica čiji su redovi sopstveni vektori kovarijacione matrice $\mathbf{\Sigma}$. Kako je vektor glavnih komponenti $\mathbf{Y} = \mathbf{AX}$, to je njegova kovarijaciona matrica $D(\mathbf{Y}) = \mathbf{A\Sigma}A^T$. Ako sada zamenimo $\mathbf{\Sigma}$ dobićemo $D(\mathbf{Y}) = \mathbf{A}(\mathbf{A}^T \mathbf{\Lambda A})\mathbf{A}^T = \mathbf{\Lambda}$, pošto je $\mathbf{A}$ ortogonalna matrica za koju važi $\mathbf{A}^T \mathbf{A} = \mathbf{I}$, čime smo na drugačiji način izveli glavne komponente.

Na osnovu dobijenog rezultata možemo odrediti generalizovanu disperziju vektora $\mathbf{Y}$. Prema prvoj definiciji generalizovana disperzija je jednaka determinanti kovarijacione matrice. Kovarijaciona matrica glavnih komponenti je $\mathbf{\Lambda}$. Njena determinanta je $|\mathbf{\Lambda}|$ i jednaka je proizvodu sopstvenih vrednosti λ_j. Na osnovu izraza ortogonalne dekompozicije matrice $\mathbf{\Sigma}$ dobijamo da je $\mathbf{\Lambda} = \mathbf{A\Sigma}A^T$. Prema osobini determinante, imamo da je $|\mathbf{\Lambda}| = |\mathbf{A\Sigma}A^T| = |\mathbf{A}||\mathbf{\Sigma}||\mathbf{A}^T| = |\mathbf{\Sigma}|$, jer je $|\mathbf{A}| = \pm 1$. To nam pokazuje da su prema prvoj definiciji generalizovane disperzije originalnog i transformisanog skupa podataka međusobom jednake.

Prema drugoj definiciji generalizovana disperzija jednaka je tragu kovarijacione matrice. Trag kovarijacione matrice jednak je zbiru sopstvenih vrednosti λ_j. Prema izrazu ortogonalne dekompozicije matrice $\mathbf{\Sigma} = \mathbf{A}^T \mathbf{\Lambda A}$, dobijamo da je $\mathbf{\Lambda} = \mathbf{A\Sigma}A^T$. Ako primenimo

osobinu traga matrice da je $tr(\mathbf{BC}) = tr(\mathbf{CB})$, imamo da je $tr(\mathbf{\Lambda}) = tr(\mathbf{A\Sigma}A^T) = tr(\mathbf{A}^T\mathbf{A\Sigma}) = tr(\mathbf{\Sigma})$, pošto je $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. Na osnovu ovoga možemo zaključiti da su i prema drugoj definiciji generalizovane disperzije originalnog i transformisanog skupa podataka međusobno jednake.

Kako su disperzije glavnih komponenti jednake sopstvenim vrednostima, možemo govoriti o *relativanom doprinosu* (engl. *the proportion of variance explained*) $j-$ te glavne komponente u objašnjenju ukupne disperzije, i određujemo ga na sledeći način:

$$\frac{\lambda_j}{\sum_{k=1}^{p}\lambda_k}, \quad j = 1,2,\ldots, p.$$

Primer 3.1.2. Na osnovu podataka koji su dati u primeru 2, odrediti generalizovane disperzije originalnog i transformisanog skupa podataka, kao i relativan doprinos svake od glavnih komponenti.

Determinante kovarijacionih matrica su jednake:

$$|\mathbf{\Sigma}| = 16 \text{ i } |\mathbf{\Lambda}| = 5.2361 \cdot 4 \cdot 0.7639 = 16.$$

Tragovi kovarijacionih matrica su jednaki:

$$tr(\mathbf{\Sigma}) = 4 + 2 + 4 = 10 \text{ i } tr(\mathbf{\Lambda}) = 5.2361 + 4 + 0.7639 = 10.$$

Pa možemo zaključiti da su generalizovane varijanse vektora $\mathbf{X}$ i $\mathbf{Y}$ međusono jednake.

Relativan doprinos prve glavne komponente ukupnom varijabilitetu je:

$$\frac{\lambda_1}{tr(\mathbf{\Lambda})} = \frac{5.2361}{10} = 0.5236, \text{ a to je } 52.36\%,$$

druge glavne komponente:

$$\frac{\lambda_2}{tr(\mathbf{\Lambda})} = \frac{4}{10} = 0.4, \text{ a to je } 40\%,$$

i treće glavne komponente

$$\frac{\lambda_3}{tr(\mathbf{\Lambda})} = \frac{0.7639}{10} = 0.0764 \text{ , a to je } 7.64\%$$

□

Ukoliko u ovoj analizi dobijemo relativno visok doprinos jedne ili nekoliko prvih glavnih komponenti, tada je moguće dalju analizu zasnovati na njima, a ne na svim glavnim komponentama. Ovo je jedan od kriterijuma odabira broja glavnih komponenti, o kojima će kasnije biti reči.

Izraz za kovarijacionu matricu $\mathbf{\Sigma} = \mathbf{A}^T \mathbf{\Lambda} \mathbf{A}$ pišemo u razvijenom obliku

$$\mathbf{\Sigma} = \begin{bmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \cdots & \boldsymbol{\alpha}_p \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_1^T \\ \boldsymbol{\alpha}_2^T \\ \vdots \\ \boldsymbol{\alpha}_p^T \end{bmatrix}$$

$$= \lambda_1 \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1^T + \lambda_2 \boldsymbol{\alpha}_2 \boldsymbol{\alpha}_2^T + \ldots + \lambda_p \boldsymbol{\alpha}_p \boldsymbol{\alpha}_p^T = \sum_{j=1}^{p} \lambda_j \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j^T .$$

Znači da je doprinos $j-$ te glavne komponente kovarijacionoj matrici $\mathbf{\Sigma}$ dat matricom $\lambda_j \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j^T$. Zadržavajući manji broj glavnih komponenti od p, kovarijacionu matricu $\mathbf{\Sigma}$ aproksimiramo zbirom matrica doprinosa zadržanih glavnih komponenti. Ukoliko doprinos ukupne disperzije zadržanih komponenti prelazi neku unapred fiksiranu vrednost, na primer 80%, tada je za očekivati da će ta aproksimacija kovarijacione matrice $\mathbf{\Sigma}$ relativno dobro reprezentovati kovarijacionu strukturu originalnog skupa podataka.

3.2. Interpretacija glavnih komponenti

Do sada smo analizu glavnih komponenti bazirali na kovarijacionoj matrici $\mathbf{\Sigma}$. Problem koji se javlja u interpretaciji glavnih komponenti posledica je njihove osetljivosti na različite merne skale originalnih promenljivih. Ako u analizi jedna od promenljivih ima znatno veću disperziju od ostalih, tada će ta promenljiva dominirati prvom glavnom komponentom bez obzira na korelacionu strukturu podataka. Jedna mogućnosti za rešavanje ovog problema je da u tom slučaju ne koristimo direktno koeficijente linearne kombinacije u cilju interpretacije glavnih komponenti, nego da analizu zasnivamo na koeficijentima korelacije originalnih promenljivih i glavnih komponenti. Druga mogućnost je da celu analizu baziramo na korelacionoj, a ne kovarijacionoj matrici originalnih podataka.

Odredimo koeficijente korelacije između originalnih promenljivih i glavnih komponenti. U radu smo već odredili kovarijacione matrice od $\mathbf{Y}$ i $\mathbf{X}$, i važi $D(\mathbf{Y}) = \mathbf{\Lambda}$ i $D(\mathbf{X}) = \mathbf{\Sigma}$. Kovarijansa između $\mathbf{X}$ i $\mathbf{Y}$ je

$$\operatorname{cov}(\mathbf{X}, \mathbf{Y}) = \operatorname{cov}(\mathbf{X}, \mathbf{AX}) = \mathbf{\Sigma A}^T = (\mathbf{A}^T \mathbf{\Lambda A})\mathbf{A}^T = \mathbf{A}^T \mathbf{\Lambda} = \left[\boldsymbol{\alpha}_1 \lambda_1, \boldsymbol{\alpha}_2 \lambda_2, ..., \boldsymbol{\alpha}_p \lambda_p\right].$$

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

$$\rho_{X_k Y_j} = \frac{\operatorname{cov}(X_k, Y_j)}{\sqrt{D(X_k)}\sqrt{D(Y_j)}} = \frac{\lambda_j \alpha_{jk}}{\sqrt{\sigma_{kk}}\sqrt{\lambda_j}} = \alpha_{jk} \frac{\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}, \quad j,k = 1,2,..., p .$$

Znači da se koeficijent linearne kombinacije uz $k-$ tu promenljivu u $j-$ toj glavnoj komponenti množi količnikom njihovih standardnih devijacija. U matričnom zapisu korelaciona matrica između vektora originalnih promenljivih $\mathbf{X}$ i vektora glavnih komponenti $\mathbf{Y}$ data je sledećim izrazom

$$\boldsymbol{\rho}_{\mathbf{XY}} = \mathbf{\Lambda}^{1/2} \mathbf{A} \mathbf{D}^{-1/2},$$

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

Ukažimo sada na iznos disperzije originalnih promenljivih koji je objašnjen zadržanim skupom glavnih komponenti. On pokazuje u kom stepenu zadržane glavne komponente dobro aproksimiraju disperziju svake originalne promenljive ponaosob. Na osnovu izraza za ortogonalnu dekompoziciju kovarijacione matrice imamo da je disperzija $k-$ te promenljive

$$\sigma_{kk}^2 = \sum_{j=1}^{p} \lambda_j \alpha_{jk}^2 , \quad k = 1,2,..., p .$$

To znači da je doprinos svake glavne komponente disperziji $k-$te promenljive jednak kvadratu koeficijenata korelacije odnosne glavne komponente i te originalne promenljive. Doprinos svih glavnih komponenti računamo na osnovu korelacione matrice $\mathbf{A\Lambda}^{1/2}$ tako što ćemo sabrati kvadrate u njenom $k-$tom redu. Ukoliko u našoj analizi zadržimo nekoliko prvih glavnih komponenti tada, stavljanjem u odnos dobijene sume i odgovarajuće disperzije originalne promenljive, dobijamo proporciju disperzije te promenljive koja je objašnjena zadržanim glavnim komponentama. Ova proporcija u analizi glavnih komponenti se naziva *komunalitet* promenljive. Možemo još reći da komunalitet predstavlja procenat „objašnjenja“ disperzije originalne promenljive zadržanim glavnim komponentama. Korišćenjem korelacione matrice umesto kovarijacione matrice originalnih promenljivih odmah dobijamo proporciju disperzije originalne promenljive

„objašnjene“ zadržanim glavnim komponentama jer je standardizacijom promenljivih vrednost disperzije jednaka jedan.

U slučaju kad, ako koristimo korelacionu matricu umesto kovarijacione matrice, odmah dobijamo komunalitet. Korelaciona matrica se još naziva i matrica strukture glavnih komponenti. Matrica strukture glavnih komponenti sadrži *opterećenja* glavnih komponenti koja predstavlja koeficijente korelacije između izabranih glavnih komponenti i promenljivih. Opterećenja ukazuju na važnost svake promenljive za pojedinu glavnu komponentu.

Pokažimo ovo na jednom primeru. Sledeća tabela predstavlja matricu strukture glavnih komponenti za 15 promenljivih i pripadajući kumunalitet za svaku promenljivu.

Faktori				Komunaliteti
1.	2.	3.	4.	
0.64319	-0.27850	-0.01546	-0.43327	0.67921
0.35104	-0.56413	0.43600	0.10311	0.64221
0.26408	0.43097	0.66299	-0.04377	0.69695
0.72500	-0.12430	0.04455	-0.30775	0.63777
0.58146	0.29653	-0.42246	0.29621	0.69223
0.35493	-0.54134	0.28165	0.47889	0.72769
0.61848	-0.10206	0.02069	-0.31120	0.49021
0.45699	0.45126	0.33137	0.20879	0.56587
0.51394	0.52278	0.38543	0.03866	0.68748
0.34656	0.26936	0.02975	-0.26011	0.26120
0.80752	-0.16924	-0.14125	-0.15609	0.72504
0.76550	0.14951	-0.21480	0.00593	0.65451
0.60592	-0.35453	0.09643	0.23064	0.55532
0.65052	-0.10479	-0.15877	0.10270	0.46992
0.69189	0.25042	-0.36724	0.31396	0.77486

Tabela 1. Opterećenja glavnih komponenti

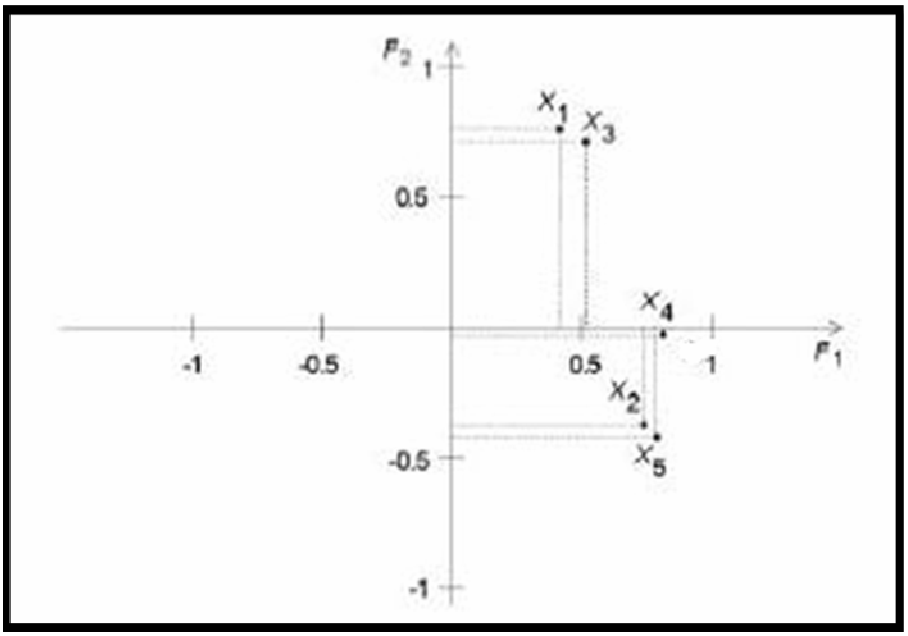
3.3. Rotacija glavnih komponenti

Glavne komponente su prvobitno dobijene rotacijom osa, nakon čega nove promenljive su nekorelisane i imaju maksimalnu disperziju. Ako se dobijene komponente ne mogu na zadovoljavajući način tumačiti, onda se komponente dodatno rotiraju, tražeći adekvatnu interpretaciju komponenti.

Analizirajući promenljive iz prethodne tabele, uočavamo da pojedine promenljive su korelisane sa nekoliko glavnih komponenti, a analizirajući faktore, možemo uočiti da je prvi faktor definisan viskim opterećenjima većeg broja glavnih komponenti. Odnosno,

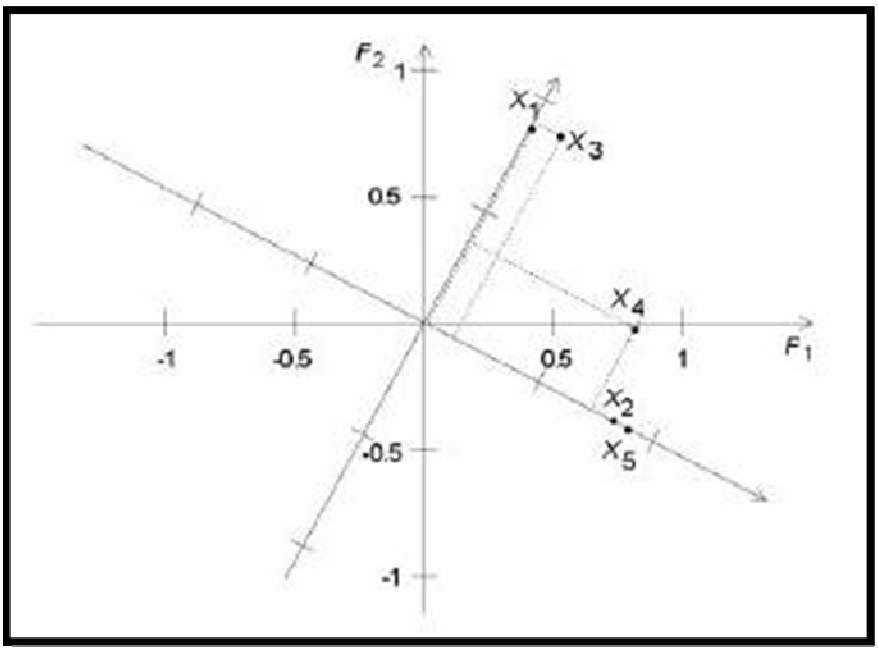
uočavamo da se jedna promenljiva javlja na više glavnih komponenti. Da bi se to izbeglo, radimo transformacijuu glavnih komponenti, odnosno *rotaciju glavnih komponenti*. Cilj rotacije jeste dobijanje jednostavne strukture u kojoj faktori treba da budu što nezavisniji. Odnosno, jedna glavna komponenta treba da bude određena jednim skupom promenljivih, druga drugim skupom promenljivih, itd. i pri tom da bude što manje promenljivih koje bi bile zajedničke većem broju glavnih komponenti.

Ilustrujmo potrebu za rotacijom grafičkim prikazom. Neka su na slici 2. za promenljive X_1, X_2, X_3, X_4 i X_5 prikazana opterećenja glavnih komponenti u dvodimenzionalnom prostoru koji grade dve glavne komponente F_1 i F_2. Na osnovu grafika je očigledno da sem kod četvrte, kod svih ostalih promenljivih imamo relativno visoke vrednosti opterećenja kod obe glavne komponente (projekcije tačaka na prvu i drugu osu glavnih komponenti prikazane su isprekidanim tačkama). To nam otežava interpretaciju dobijenog rešenja pomoću glavnih komponenti, jer nismo u mogućnosti nedvosmisleno zaključiti koje promenljive određuju prvu, odnosno drugu komponentu. Takođe, na slici se jasno uočava grupisanje promenljivih u dve grupe. Prvu grupu čine promenljive X_2, X_4, X_5, a drugu preostale dve promenljive: X_1, X_3.



Slika 2. Nerotirana opterećenja glavnih komponenti

Primenom ortogonalne transformacije matrice opterećenja glavnih komponenti, rotiramo ose glavnih komponenti tako da one u svom novom položaju prolaze sto bliže tačkama koje predstavlaju opterećenja pet originalnih originalnih komponenti. Na sledećoj slici prikazane su ose glavnih komponenti, odnosno rotirana opterećenja.



Slika 3. Rotirana opterećenja glavnih komponenti

Na slici 3. vidimo da druga i peta promenljiva određuju prvu glavnu promenljivu, a prva i treća određuju drugi faktor. Za četvrtu promenljivu se može reći da je bliža prvoj nego drugoj promenljivoj. Međusobni položaj pet tačaka na slici 3. nije se promenio nakon rotacije glavnih komponenti, nego se promenio samo referentni koordinatni sistem u odnosu na koji te tačke posmatramo. Dakle, promenom ugla gledanja na opterećenja, odnosno rotacijom glavnih komponenti, jasnije sagledavamo prirodu komponenti. Projekcija tačaka na rotirane ose glavnih komponenti ukazuje na promenjene vrednosti opterećenja svake promenljive u odnosu na prvu i drugu glavnu komponentu. Tako na primer, prvobitno visoka vrednost oba opterećenja kod treće promenljive, nakon rotacije ukazuje na visoku vrednost opterećenja na drugoj i nisku vrednost na prvoj glavnoj komponenti.

3.3.1. Metod ortogonalne rotacije

Ortogonalna rotacija glavnih komponenti ne menja međusobni odnos faktorskih osa, one su i dalje ortogonalne. One se po tome razlikuju od neortogonalne rotacije kod koje nema tog ograničenja jer se ose glavnih komponenti rotiraju nezavisno jedna od druge. Nakon neortogonalne rotacije u opštem slučaju zaklapaju međusobom ugao različit od 90^0. Metod ortogonalne rotacije koristi ortogonalnu matricu kojom transformiše matricu opterećenja.

Postoji više metoda ortogonalne rotacije. Neki od njih su varimax, quartimax, equimax. Ali najčešće korišćen od njih je varimax kriterijum. U varimax rotaciji, svaka glavna komponenta teži da postigne veliko opterećenje (1 ili skoro 1) za manji broj promenljivih i malo opterećenje (blizu nuli) za ostale promenljive, kako bi se lakše interpretirali rezultati.

Varijabilitet objašnjen svakom glavnom komponentom pre rotacije se ponovo aranžira rotacijom. Ukupna objašnjena disperzija ostaje ista i nakon rotacije. Međutim, prva rotirana glavna komponenta ne mora objašnjavati maksimum disperzije. Količina disperzije, koju svaka glavna komponenta objašnjava mora se ponovo računati.

3.3.2. Metod neortogonalne rotacije

Izostavljanjem zahteva za ortogonalnošću glavnih komponenti dolazimo do metoda koji pri rotaciji dozvoljava mogućnost da rotitane glavne komponente zaklapaju različit ugao od 90^0.

3.4. Testiranje značajnosti glavnih komponenti

Analiza glavnih komponenti predstavlja metod za redukciju podataka i kao takva nije zasnovana na teorijskom modelu. U postupku redukcije podataka nije naglašeno sa koliko glavnih komponenata je potrebno izvršiti analizu da bi se obuhvatio značajan iznos ukupne disperzije. Iz dosadašnjeg izlaganja možemo zaključiti da smo u analizi zainteresovani za one glavne komponente koje imaju najveće sopstvene vrednosti. Odatle ne sledi da su manje interesantne glavne komponente sa manjim sopstvenim vrednostima. Zbog svega toga, možemo govoriti i o testiranju značajnosti glavnih komponenti u vidu testova sopstvenih vrednosti. Možemo reći da je testiranje značajnosti još jedan od načina odabira glavnih komponenti koje su bitne za analizu, pored načina odabira glavnih komponenti prema veličini njihovih sopstvenih vrednosti. Pored ova dva, navešćemo i još neke načine za odabir glavnih komponenti.

Za sopstvene vrednosti kovarijacione matrice najpoznatiji je test koji se pripisuje Bartlettu, za testiranje hipoteze da su poslednje $(p-k)$ sopstvene vrednosti međusobno jednake. Odnosno, postavljamo nultu hipotezu

$$H_0 : \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p ,$$

protiv alternativne hipoteze H_1, da su barem dve od poslednjih $(p-k)$ sopstvenih vrednosti različite među sobom. Ako kao rezultat testa dobijemo da prihvatamo nultu hipotezu, tada u analizi koristimo samo prvih k glavnih komponenti jer za njih pretpostavljamo da obuhvataju značajan iznos ukupne disperzije, a da poslednje $(p-k)$ glavne komponente mere samo „šum“ u podacima.

Test statistika za testiranje navedene hipoteze konstruisana je uz pretpostavku o normalnosti, a zasnovana je na korišćenju principa količnika maksimalne verodostojnosti (eng. Likelihood Ratio test),

$$LR = \left(\frac{\prod_{j=k+1}^{p} \hat{\lambda}_j}{\left[\frac{1}{p-k} \sum_{j=k+1}^{p} \hat{\lambda}_j \right]^{p-k}} \right)^{\frac{n}{2}},$$

gde su $\hat{\lambda}_j$ karakteristični koreni uzoračke kovarijacione matrice. Prema navedenom izrazu zaključujemo da je test zasnovan na poređenju geometrijske i aritmetičke sredine poslednja $(p-k)$ karakteristična korena. Ukoliko je tačna nulta hipoteza, LR statistika ima vrednost jednaku nuli. U suprotnom, udaljavajući se od nulte hipoteze razlika između aritmetičke i geometrijske sredine postajaće sve veća, što znači da će vrednost LR statistike biti sve manja. U tom slučaju odbacili bismo nultu hipotezu o jednakosti poslednja $(p-k)$ karakteristična korena. U primeni ovog testa koristimo asimptotski raspored statistike $-2\ln LR$ koja ima asimptotski χ^2-raspored (ako je tačna nulta hipoteza) sa $v = \frac{1}{2}(p-k+2)(p-k-1)$ stepeni slobode, pri čemu je izvršena modifikacija prvobitne aproksimacije u cilju njenog poboljšanja

$$\left[n - \frac{1}{6}(2p+11) \right] \left[(p-k)\ln \overline{\lambda}_{p-k} - \sum_{j=k+1}^{p} \ln \hat{\lambda}_j \right] \sim \chi_v^2,$$

gde je $\overline{\lambda}_{p-k}$ aritmetička sredina poslednjih $(p-k)$ karakterističnih korena uzoračke kovarijacione matrice. Znači da se nulta hipoteza o jednakosti poslednja $(p-k)$ karakteristična korena odbacuje na nivou značajnosti
α, ako je izračunata vrednost test statistike veća ili jednaka kritičnoj vrednosti $\chi^2_{v;\alpha}$.

Praktičan postupak primena ovog testa za testiranje jednakosti poslednje $(p-k)$ sopstvene vrednosti kovarijacione matrice je sledeći. Prvo se testira hipoteza da su svi karakteristični koreni jednaki među sobom, $k=0$. Ako se odbaci ova hipoteza, postavlja se nova prema kojoj su sve sopstvene vrednosti, osim prve, međusono jednake, $k=1$. Ako se odbaci ova hipoteza, postupak testiranja nastavljamo, ali sada testiramo nultu hipotezu da su sve sopstvene vrednosti međusobom jednake osim prve dve, $k=2$. Ovaj postupak nastavljamo sve dok se ne prihvati hipoteza o jednakosti poslednjih $(p-k)$ sopstvenih vrednosti.

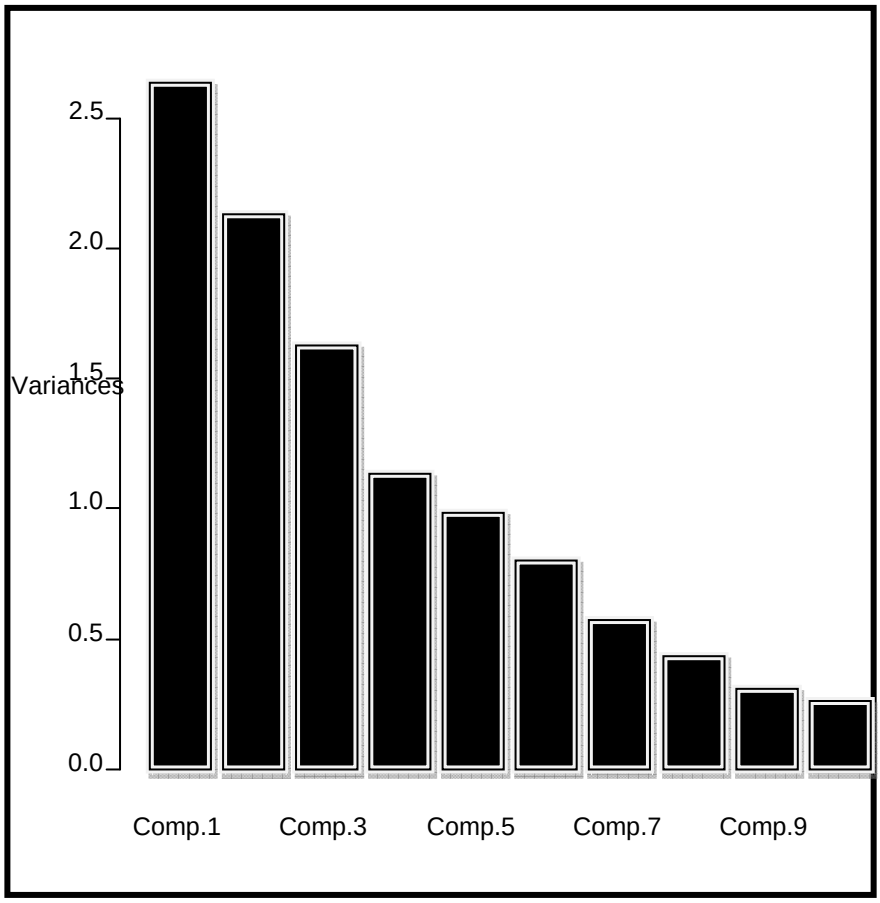
Testovi sopstvenih vrednosti mogu biti od pomoći u izboru broja glavnih komponenti. Međutim, oni mogu sugerisati značajnost velikog broja sopstvenih vrednosti, a time i glavnih komponenti. Pa se zbog toga u praksi češće koriste neki drugi kriterijumi.

3.5. Izbor broja glavnih komponenti

Kao što smo već napomenuli, jedan od ciljeva analize glavnih komponenti jeste smanjenje dimenzije početnog skupa podataka. Nameće se logično pitanje, koliko glavnih komponenti treba uključiti u analizu? Maksimalan broj glavnih komponenti je jednak broju originalnih promenljivih, međutim, time ne bismo rešili jedan od osnovnih zadataka analize glavnih komponenti.

Kriterijum karakteristične vrednosti ili kriterijum sopstvenih vrednosti (engl. *eigenvalue criterium*). Karakteristična vrednost predstavlja iznos disperzije u originalnim promenljivim koji je povezan sa određenom glavnom komponentom. Po ovom kriterijumu zadržavaju se one glavne komponente čija je karakteristična vrednost veća od 1, a ostali faktori se ne uključuju u model. Faktor sa karakterističnom vrednošću manjom od 1 nije ništa bolji od originalne promenljive jer usled standardizacije, svaka promenljiva ima disperziju jednaku 1. Pa zbog toga, faktor treba da objasni barem onaj iznos varijabiliteta koji daje jedna promenljiva, jer inače, bolje je koristiti originalnu promenljivu. Ovaj kriterijum je u širokoj primeni u analizi glavnih komponenti i poznat je još i pod nazivom Kaiserov kriterijum (Kaiser, 1958.)

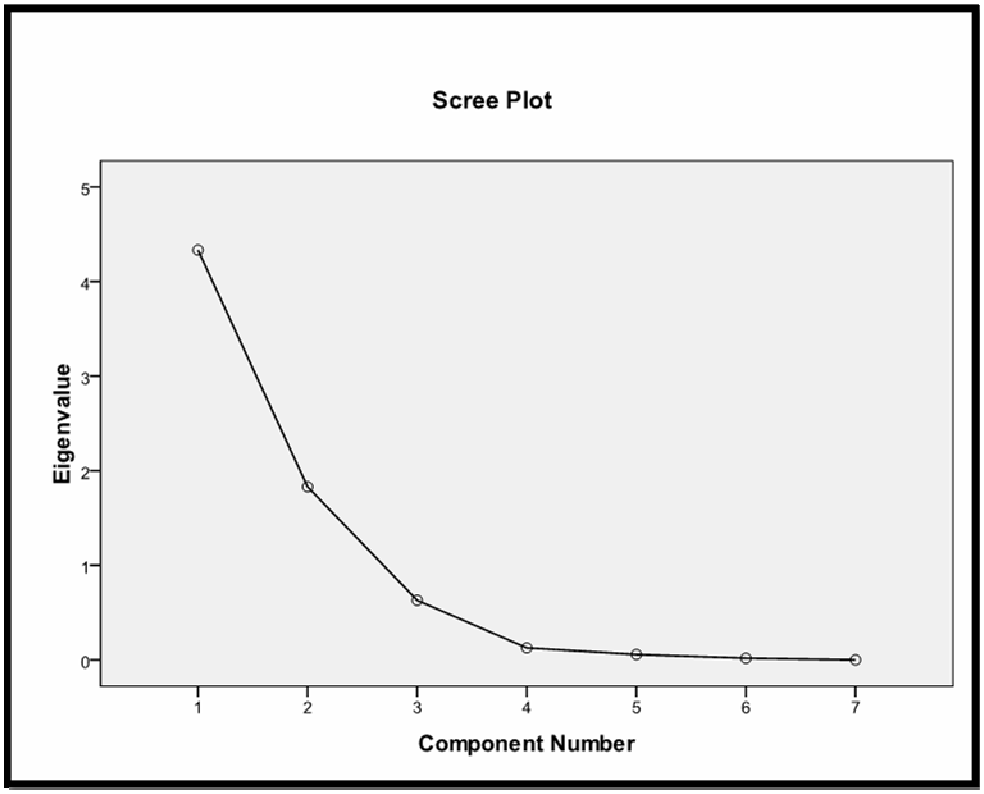
Dijagram procentualnog učešća varijabiliteta (engl. *percentage of variance criteria*). Kod ovog pristupa broj formalnih glavnih komponenti se određuje tako da kumulativno procentualno učešće varijabiliteta koji oni objašnjavaju dostiže neki zadovoljavajući nivo. Nivo varijabiliteta koji je zadovoljavajući zavisi od vrste problema. Neretko se koristi kriterijum baziran na 70% objašnjenog varijabiliteta. Pa se u tom slučaju zadržavaju glavne komponente dok se ne postigne ovaj unapred zadat nivo. Očita je subjektivnost ovog načina određivanja broja glavnih komponenti jer se on određuje na bazi proizvoljno fiksirane vrednosti kriterijuma kumulativne proporcije objašnjene varijanse. Dijagram procentualnog učešća prikazan je na sledećoj slici 4.



Slika 4. Dijagram procentualnog učešća varijabiliteta

Kriterijum testa značajnosti (engl. *significance test criteria*). Moguće je odrediti statističku značajnost disperzija različitih komponenti i zadržati samo one faktore čije disperzije su statistički značajne. Mana ovog pristupa je što će za velike uzorke mnoge glavne komponente biti statistički značajni.

Kriterijum dijagram osipanja. Ovaj kriterijum zasniva se na grafičkom prikazu sopstvenih vrednosti prema njihovom rednom broju. Ovaj kriterijum se još i naziva „scree test", a predložio ga je Cattell (1966.). Bira se broj komponenti koji se nalazi u prelomnoj tački. Prelom na krivoj se određuje tako što se prisloni lenjir uz poslednje sopstvene vrednosti proveravajući da li one leže na pravoj liniji. Broj glavnih komponenti određujemo tako što uočavamo tačku nakon koje spomenuta 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 sopstvena vrednost kao poslednja leži na pravoj liniji. Na sledećoj slici 5. prikazan je primer dijagrama osipanja, na kome možemo videti da bi u tom slučaju imali dve glavne komponente, jer postoje dve sopstvene vrednosti koje leže na pravoj. Ovaj kriterijum nije od pomoći ukoliko na grafiku nema očiglednih preloma, ili ukoliko ih ima više od jednog.



Slika 5. Dijagram osipanja

U praksi se primenjuje još neki kriterijumi za određivanje broja glavnih komponenti. Na sledećem primeru pokazaćemo kako koristimo ove kriterijume. Posmatrajmo sledeću tabelu.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,755	55,092	55,092	2,755	55,092	55,092	2,736	54,711	54,711
2	1,775	35,497	90,589	1,775	35,497	90,589	1,794	35,878	90,589
3	,377	7,542	98,131						
4	,065	1,299	99,431						
5	,028	,569	100,000						

Extraction Method: Principal Component Analysis.

Tabela 2. Određivanje broja faktora

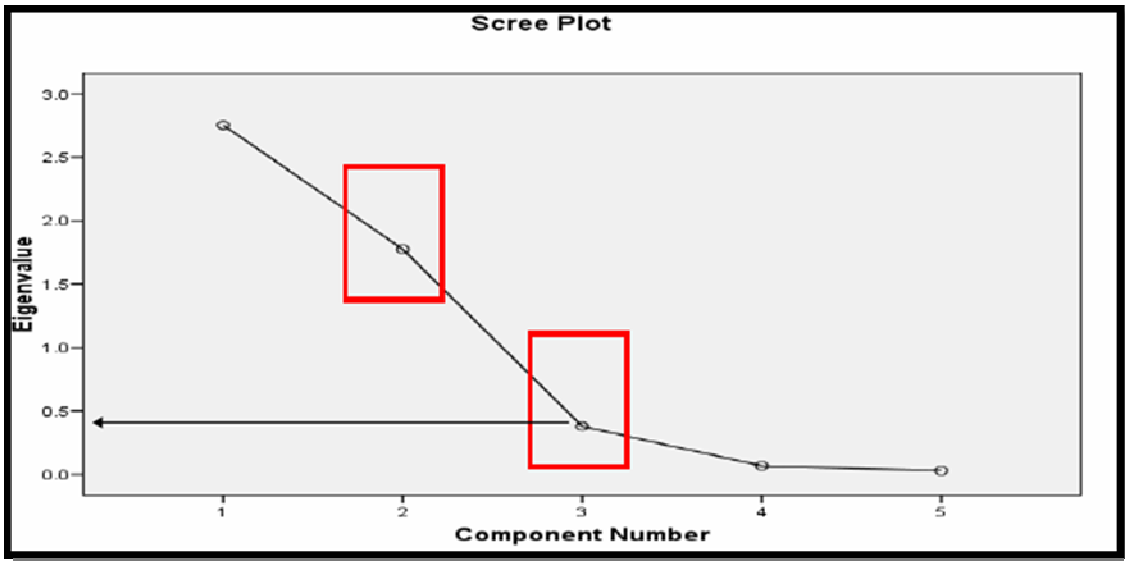
Kao što smo već napomenuli, broj faktora koji će se koristiti u modelu može se odrediti na više načina:

Kriterijum karakteristične vrednosti. Za ovaj kriterijum smo već napomenuli da se zadržavaju samo one glavne komponente čija je karakteristična vrednost veća od 1. U tabeli

možemo primetiti da samo prve dve komponente imaju vrednosti veće od 1, i to su 2.755 i 1.775.

Kriterijum procentualnog učešća varijabiliteta. Neretko se kao kriterijum za kumulativno procentualno učešće varijabiliteta koji glavne komponente treba da objasne uzima 70%. U našem primeru , vidimo da prva komponenta objašnjava 55%, tako da ona sama nije dovoljana. Ako uzmemo i drugu komponentu za dalju analizu, onda ta dva faktora objašnjavaju čak 90% ukupnog varijabiliteta.

Dijagram osipanja. Bira se broj faktora koji se nalazi u prelomnoj tački ali se gleda i nivo varijabiliteta odnosno karakteristične vrednosti (u ovom slučaju biramo 2 faktora). Vidi sliku 6.



Slika 6. Dijagram osipanja

Pored ovih navedenih kriterijuma, prilikom praktične primene analize glavnih komponenti, možemo se rukovoditi još nekim kriterijumima.

Iskustveno pravilo. Sve uključene komponnete moraju da objasne bar onoliko varijabiliteta koliko jedna „prosečna promenljiva". Ako imamo pet promenljivih, tada svaka komponenta mora da objasni više od 20% ukupnog varijabiliteta. Vođeni ovim pravilom možemo zaključiti da samo prve dve komponente objašnjavaju više od 20% varijabiliteta, čak više od 55% i 35%.

Iskustveno pavilo. Ne uzimamo komponentu gde dolazi do značajnog pada u količini varijabiliteta koji ona objašnjavaju. Možemo primetiti da kod treće komponente dolazi do značajnog pada u količini varijabiliteta. Druga komponenta objašnjava 35% a treća samo 7%

4. Analiza glavnih komponenti u praksi

U ovom delu daćemo nekoliko paktičnih primera analize glavnih komponenti. Ovde ćemo pokušati da kroz konkretan primer objasnimo ovu metodu i objasnimo dobijene rezultate.

Primer 4.1. Za analizu su korišćeni statistički programi R i SPSS. U cilju bolje prezentacije i objašnjenja, autor je koristio neke već gotove grafike iz literature [1]. Preuzeti grafici su naznačeni u tekstu, a preostali deo primera je obradio autor.

Dati su podaci o prosečnom trošku nekoliko različitih tipova Francuskih porodica na hranu. Tipovi francuskih porodica su sledeci:

MA-manual workers (manuelni radnici)
EM-employees (poslodavci)
CA-menagers (menadžeri)

Posmatrane su porodice sa različitim brojem dece: dvoje, troje, četvoro i petoro.

		hleb (bread)	povrće (vegetables)	voće (fruits)	meso (meat)	živina (poultry)	mleko (milk)	vino (wine)
1	MA2	332	428	354	1437	526	247	427
2	EM2	293	559	388	1527	567	239	258
3	CA2	372	767	562	1948	927	235	433
4	MA3	406	563	341	1507	544	324	407
5	EM3	386	608	396	1501	558	319	363
6	CA3	438	843	689	2345	1148	243	341
7	MA4	534	660	367	1620	638	414	407
8	EM4	460	699	484	1856	762	400	416
9	CA4	385	789	621	2366	1149	304	282
10	MA5	655	776	423	1848	759	495	486
11	EM5	584	995	548	2056	893	518	319
12	CA5	515	1097	887	2630	1167	561	284

Tabela 3 . Podaci iz baze french.food

Naš cilj je da smanjimo dimenziju podataka, i otkrijemo suštinski koncept koji leži u osnovi ovih podataka. Prvi korak u analizi glavnih komponenti je formiranje glavnih komponenti i određivanje broja glavnih komponenti. Glavne komponente možemo izabrati uz pomoć različitih kriterijuma o kojima je već bilo reči.

Total Variance Explained									
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	**4,333**	**61,903**	**61,903**	**4,333**	**61,903**	**61,903**	**3,835**	**54,780**	**54,780**
2	**1,830**	**26,147**	**88,050**	**1,830**	**26,147**	**88,050**	**2,329**	**33,271**	**88,050**
3	,631	9,012	97,062						
4	,128	1,833	98,896						
5	,058	,822	99,718						
6	,019	,269	99,987						
7	,001	,013	100,000						
Extraction Method: Principal Component Analysis.									

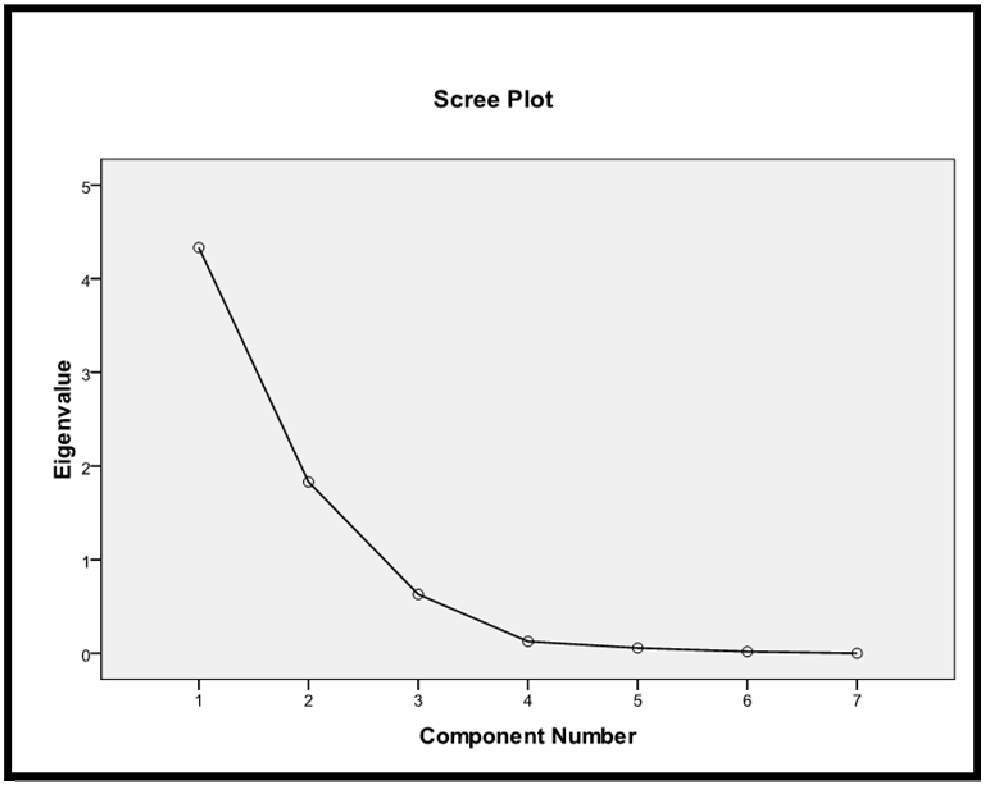
Tabela 3. Objašnjen varijabilitet

Određivanje broja glavnih komponenti

1.**Kriterijum karakteristične vrednosti.** Možemo uočiti u prethodnoj tabeli da imamo dve sopstvene vrednosti koje su veće od 1, to su 4.333 i 1.830. Tako da po ovom kriterujumu imamo dve glavne komponente.

2.Kriterijum procentualnog učešća. Možemo imati dva faktora, a i više. Ovakav zaključak iznpsimo jer prva glavna komponena objašnjava čak 61% ukupnog varijabiliteta, a druga 26%.

3.Dijagram osipanja. Prelomna tačka je kod trećeg faktora, tako da po ovom kriterijumu imamo dve glavne komponente. Pogledati sledeću sliku 7.



Slika 7. Dijagram osipanja

4.Iskustveno pravilo. U ovom primeru imamo 7 promenljivih, tako da nam to govori da svaki faktor mora da objasni najmanje 14% ukupnog varijabiliteta. Možemo primetiti da samo prva dva faktora objašnjavaju više od 14% ukupnog varijabiliteta, prvi čak 62% a drugi 26%. Tako da i po ovom kriterijumu imamo dva faktora.

5.Iskustveno pravilo. Komponenta gde dolazi do značajnog pada u količini varijabiliteta koji oni objašnjavaju. Možemo primetiti da kod treće glavne komponente dolazi do značajnog pada u količini varijabiliteta. Druga glavna komponenta objašnjava 26% a treći samo 9%.

Objašnjen varijabilitet

Prva glavna komponenta objašnjava 61,9% ukupnog varijabiliteta ovih sedam promenljivih u analizi, druga glavna komponenta dodatnih 26,1% varijabiliteta.

Primenom različitih kriterijuma možemo reći da smo izabrali dve komponente, odnosno dva faktora i da one objašnjavaju ukupno skoro 88% ukupnog varijabiliteta originalnih promenljivih. Ovo znači da se za buduće analize mogu koristiti kao promenljive ova dve komponente umesto 7 originalnih promenljnivih uz gubitak informacija od 12%.

Rotacija

Cilj rotacije jeste dobijanje jednostavne strukture u kojoj glavne komponente treba da budu što nezavisnije. Odnosno, jedna glavna komponenta treba da bude određen jednim skupom promenljivih, druga drugim skupom promenljivih, itd. i pri tom da bude što manje promenljivih koje bi bile zajedničke većem broju glavnih komponenti.

Prilikom rotacije zadržava se objašnjen procenat varijabiliteta pomoću glavnih komponenti ali se varijabilitet raspoređuje na izabrane komponente odnosno faktore. Velike promene u koeficijentima ukazuju da se faktori lakše tumače.

Total Variance Explained

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,333	61,903	61,903	3,835	54,780	54,780
2	1,830	26,147	88,050	2,329	33,271	88,050

Tabela 4. Objašnjen varijabilitet

Prilikom rotacije se menjaju opterećenja i trebalo bi da opterećenja u tabeli sa rotacijom daju jasniju interpretaciju. U sledećoj tabeli su data opterećenja glavnih komponenti nakon rotacije.

Rotated Component Matrix[a]

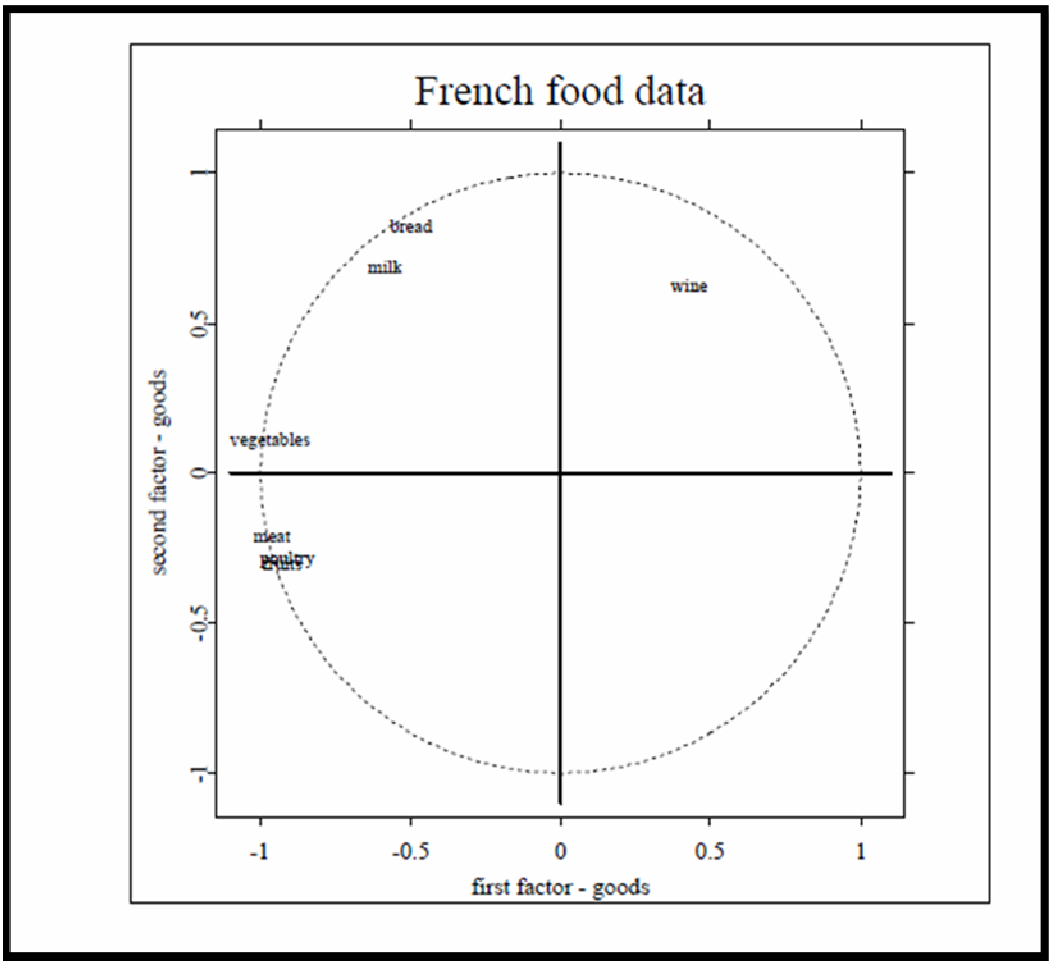
	Component	
	1	2
hleb	,071	**,976**
povrce	**,808**	,552
voce	**,955**	,166
meso	**,946**	,258
zivina	**,934**	,169
mleko	,207	**,893**
vino	-,672	,389

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with

Tabela 5. Opterećenja glavnih komponenti nakon rotacije

Kako je kvalitet reprezentacije dobar za sve promenljive (osim za promenljivu vino), imamo lepu sliku o korelaciji između originalnih promenljivih i glavnih komponenti. Iz tabele posmatramo samo ona opterećenja koja su veća od 0.7 ili manja od -0.7. Možemo primetiti da je vino negativno korelisano sa grupom promenljivih koju čine meso, povrće, voće i živina i koje su međusobno pozitivno korelisane. Ove navedene promenljive se grupišu oko prve glavne komponente. Potom, mleko i hleb su pozitivno korelisani, ali su zato slabo korelisane sa mesom, voćem i živinom.

Međutim, prethodna reprezentacija može biti i bolja. Posmatrajmo sledeću tabelu korelacija sa faktorima. Možemo posmatrati i sledeću sliku. Naših sedam promenljivih smo projektovali na podprostor dimenzije dva.



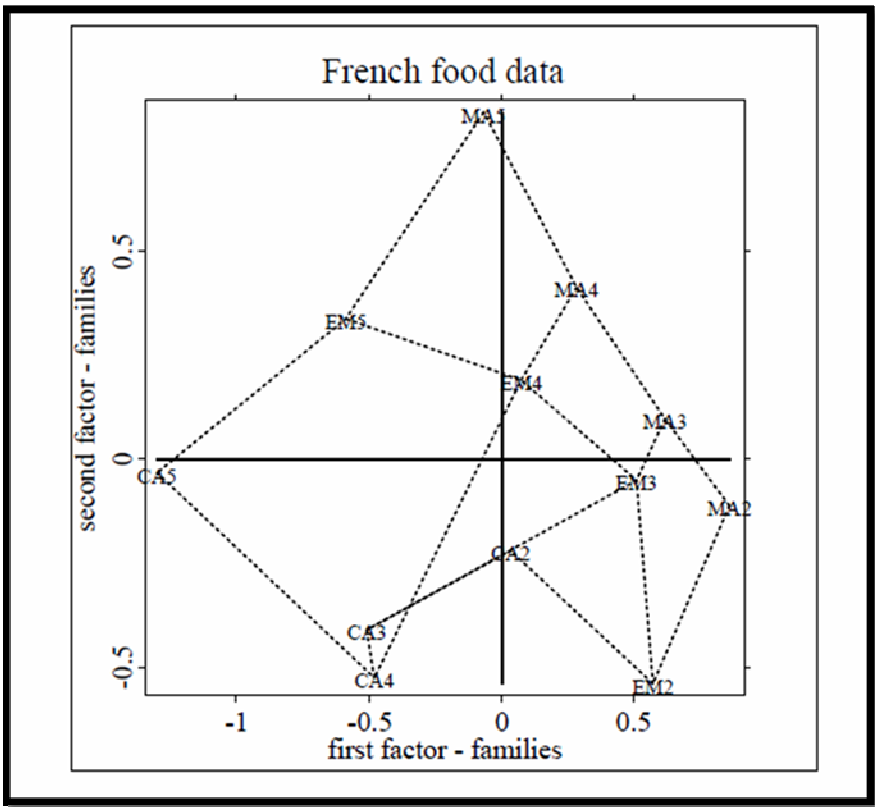
Slika 8 . Prva i druga glavna komponenta, iz literature [1]

	$r_{X_i Z_1}$	$r_{X_i Z_2}$	$r^2_{X_i Z_1} + r^2_{X_i Z_2}$
hleb	-0.499	0.842	0.957
povrce	-0.970	0.133	0.958
voce	-0.929	-0.278	0.941
meso	-0.962	-0.191	0.962
zivina	-0.911	-0.266	0.901
mleko	-0.584	0.707	0.841
vino	0.428	0.648	0.604

Tabela 6. Korelacija sa faktorima

Sa prethodne slike i tabele možemo primetiti da prvi faktor čine povrće, meso, živina i voće (sa pozitivnim znakom), a drugi faktor mleko, hleb i vino (sa negativnim znakom). Na slici možemo primetiti kako se grupišu promenljive, oko kog faktora.

Posmatrajmo sledeći grafik.



Slika 9. Reprezentacija francuskih porodica, iz literature [1.]

Povučene linije na prethodnoj slici povezuju porodice sa istim brojem dece i porodice sa istim profesijama. Ako posmatramo ovu sliku, i grafik na kome su predstavljene glavne komponente hrane možemo uočiti da su troškovi za hleb, mleko i vino slični za MA-manual workers i EM-employees. Porodice menadzera karakterišu veći troškovi za povrće, voće, meso i živinu.

Primer 4.2.

Posmatramo rezultate dobijene merenjem tragova 10 hemijskih elemenata u sedimentima na različitim lokacijama u dve sezone. Za analizu su korišćeni statistički programi SPSS i R. Rezultati merenja dati su sledećoj tabeli.

Jesen 2005	Fe	Mn	Zn	Ni	Pb	Cu	Co	As	Cd	Hg
Sv.Stasije	9263.32	208.709	25.058	18.208	7.000	6.582	17.464	4.88	0.746	0.064
Kukuljina	11711.49	325.927	45.220	74.509	9.561	14.363	10.239	5.21	0.541	0.098
H.Novi	6090.023	772.347	23.801	32.335	3.722	11.855	9.028	3.77	0.869	0.084
Žanjice	10507.45	497.121	19.814	16.382	3.937	7.676	6.579	19.75	0.773	0.027
Mamula	1591.182	282.293	7.7628	12.772	5.1417	4.7211	5.046	4.55	0.4146	0.029
Bigova	713.712	183.645	4.0234	10.481	1.276	5.6095	13.867	3.46	0.3003	0.028
Budva	1243.18	132.370	5.117	2.668	2.634	3.235	15.992	2.58	0.063	0.003
Bar	6216.407	657.072	22.399	15.774	5.208	14.720	11.396	3.09	0.0682	0.032
Rt Đeran	34637.16	729.066	62.376	336.14	6.421	24.784	14.521	17.71	0.401	0.063
Ada Bojana	21222.72	754.783	46.166	228.617	3.598	20.724	26.163	<0.1	0.4368	0.000

Jesen 2006	Fe	Mn	Zn	Ni	Pb	Cu	Co	As	Cd	Hg
Sv.Stasije	4966.76	174.7	9.06	20.65	0.28	<5.0	<1.0	4.39	<0.05	0.037
Kukuljina	12252.71	409.5	46.19	71.73	0.40	6.98	<5.0	4.45	<0.05	0.093
H.Novi	8978.49	369.1	25.15	46.14	4.79	8.81	<1.0	1.75	0.247	0.040
Žanjice	1994.86	155.4	11.60	26.44	0.09	<5.0	<5.0	1.12	<0.05	0.017
Mamula	1829.69	168.2	5.05	16.87	1.56	<5.0	<1.0	1.85	0.074	0.014
Bigova	1355.14	220.5	4.37	17.09	1.33	<5.0	<5.0	1.93	0.093	0.012
Budva	2749.12	135.1	27.81	16.87	0.19	<5.0	<1.0	2.17	0.051	0.009
Bar	5110.64	406.9	10.10	26.60	1.73	6.05	<5.0	1.29	0.17	0.025
Rt Đeran	40866.61	943.3	67.21	334.9	1.95	23.24	16.86	7.49	0.17	0.031
Ada Bojana	40306.22	983.8	52.77	267.1	1.32	20.67	14.65	3.86	0.077	0.021

Tabela 7 . Tragovi elemenata dobijeni iz uzoraka sedimenata

Naš cilj je da smanjimo dimenziju podataka, i otkrijemo suštinki koncept koji leži u osnovi ovih podataka. Prvi korak u analizi glavnih komponenti je formiranje glavnih komponenti i određivanje broja glavnih komponenti. Glavne komponente možemo izabrati uz pomoć više kriterijuma o kojima je već bilo reči.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	**5.216**	**52.164**	**52.164**	5.216	52.164	52.164	4.806	48.059	48.059
2	**2.081**	**20.806**	**72.970**	2.081	20.806	72.970	1.927	19.274	67.333
3	.949	9.493	82.463	.949	9.493	82.463	1.513	15.130	82.463
4	.770	7.702	90.165						
5	.407	4.069	94.234						
6	.297	2.975	97.209						
7	.148	1.480	98.690						
8	.067	.672	99.362						
9	.053	.531	99.893						
10	.011	.107	100.000						

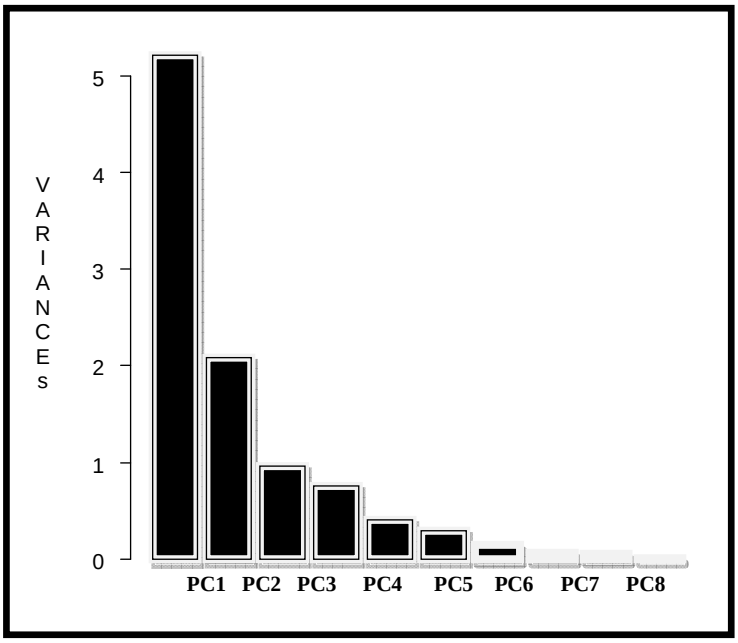
Extraction Method: Principal Component Analysis.

Tabela 8. Objašnjen varijabilitet

Određivanje broja glavnih komponenti

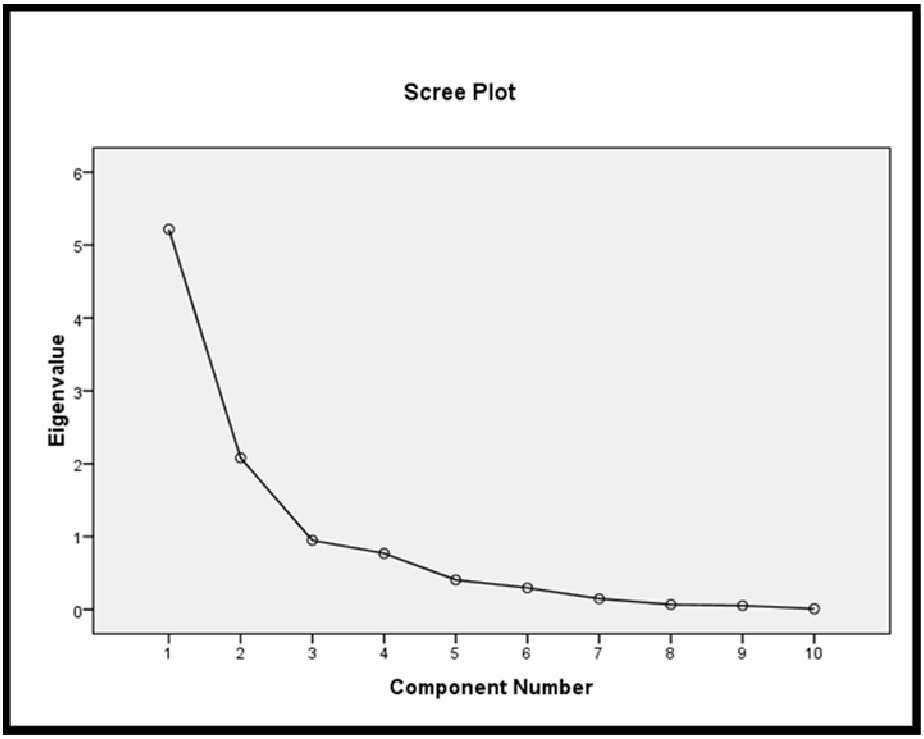
1.**Kriterijum karakteristične vrednosti.** Možemo uočiti u prethodnoj tabeli da imamo dve sopstvene vrednosti koje su veće od 1, to su 5.216 i 2.081. Tako da po ovom kriterujumu imamo dva faktora.

2.Kriterijum procentualnog učešća. Na ovom grafiku možemo primetiti da prve dve komponente objašnjavaju najviše ukupni varijabilitet. Od toga, prva glavna komponenta objašnjava preko 50% ukupnog varijabiliteta, dok druga glavna komponenta objašnjava preko 20% ukupnog varijabiliteta, a svih ostalih 8 komponenti objašnjavaju zajedno ostatak varijabiliteta, što je negde oko 27%.



Slika 10 . Kriterijum procentualnog učešća

3. Dijagram osipanja.



Slika 11. Dijagram osipanja

Prelomna tačka je kod trećeg faktora, pa možemo reći da imamo dve glavne komponente.

4. Iskustveno pravilo. U ovom primeru imamo 10 promenljivih, tako da nam to govori da svaki faktor mora da objasni najmanje 10% ukupnog varijabiliteta. Možemo primetiti da samo prva dva faktora objašnjavaju više od 10% ukupnog varijabiliteta, prvi čak 52,1% a drugi 20,8%. Tako da i po ovom kriterijumu imamo dva faktora.

Objašnjen varijabilitet

Prvi faktor objašnjava 52.1% ukupnog varijabiliteta ovih sedam promenljivih u analizi, drugi faktor dodatnih 20,1% varijabiliteta. To možemo lepo primetiti i na slici 10.

Primenom različitih kriterijuma možemo reći da smo izabrali dva faktora, odnosno dve komponente i da one objašnjavaju ukupno skoro 73% ukupnog varijabiliteta originalnih promenljivih. Ovo znači da se za buduće analize mogu koristiti kao promenljive ova dva faktora umesto 10 originalnih promenljnivih uz gubitak informacija od 27%. Ako bismo se ipak odlučili za tri faktora, gubitak informacija bi bio 17,5% jer bi tada ta tri faktora objasnila čak 82,5% ukupnog varijabiliteta originalnih promenljivih.

Rotacija

Cilj rotacije jeste dobijanje jednostavne strukture u kojoj glavne komponente treba da budu što nezavisnije. Odnosno, jedna glavna komponenta treba da bude određen jednim skupom promenljivih, druga drugim skupom promenljivih, itd. i pri tom da bude što manje promenljivih koje bi bile zajedničke većem broju glavnih komponenti

Prilikom rotacije zadržava se objašnjen procenat varijabiliteta pomoću glavnih komponenti ali se varijabilitet raspoređuje na izabrane komponente odnosno faktore. Velike promene u koeficijentima ukazuju da se faktori lakše tumače. Korišćena je varimax rotacija.

Total Variance Explained

Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
5.216	52.164	52.164	4.806	48.059	48.059
2.081	20.806	72.970	1.927	19.274	67.333
.949	9.493	82.463	1.513	15.130	82.463

Tabela 9. Objašnjen varijabilitet

Interpretacija glavnih komponenti se bazira na opterećenjima**,** koja predstavljaju korelaciju između faktora i originalnih promenljivih. Opterećenja pokazuju koje su originalne promenljive korelirane sa svakom glavnom komponentom i veličinu te korelcije. U sledećoj tabeli su data opterećenja koja nam pokazuju u kolikoj su meri povezane glavne komponente sa originalnim promenljivim.

Rotated Component Matrix[a]

	Component	
	1	2
Fe	**.969**	.047
Mn	**.878**	.132
Zn	**.894**	.219
Ni	**.971**	-.020
Pb	.125	**.840**
Cu	**.952**	.196
Co	.626	.203
As	.312	.537
Cd	.020	**.872**
Hg	.060	**.726**

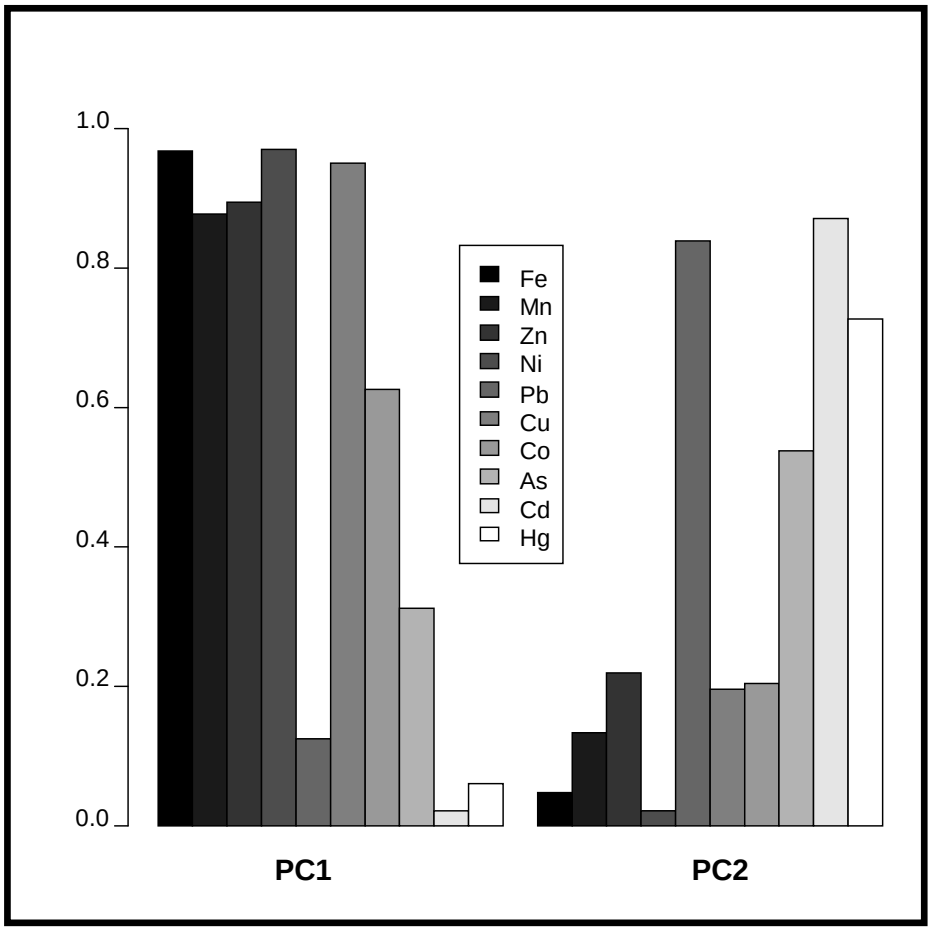
Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Tabela 10 . Opterećenja glavnih komponenti nakon rotacije

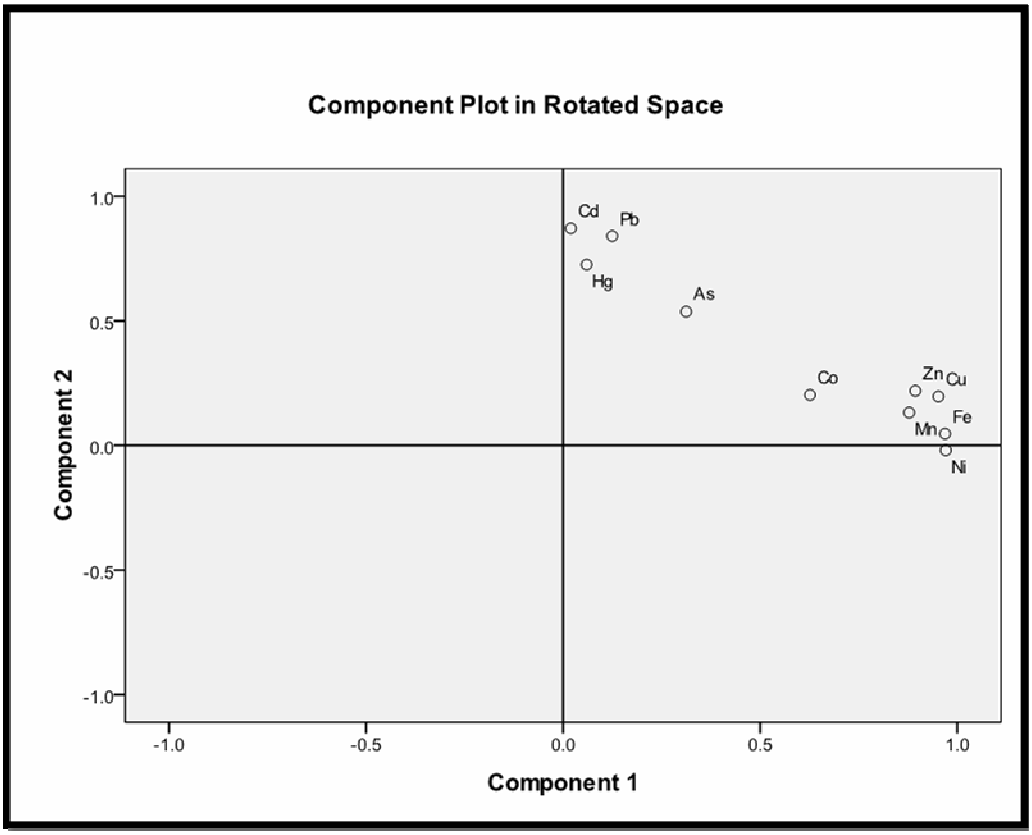
Vrednosti u prethodnoj tabeli su opterećenja koja predstavljaju korelaciju između faktora i originalnih promenljivih. Opterećenja pokazuju koje su originalne varijabile korelirane sa svakim od faktora i veličinu te korelacije. U tabeli posmatramo vrednosti čija su opterećenja veća od 0.7 ili manja od -0.7 (ove vrednosti u tabeli dodatno označene). Kao što možemo primetiti, prva glavna komponenta je u korelaciji sa Fe, Mn, Zn, Ni, Cu . Takođe možemo reći da ova prva glavna komponenta u velikoj meri objašnjava i Co . Druga glavna komponenta je u korelaciji sa Pb, Cd i Hg . Takođe možemo primetiti i da je opterećenje za As veće od 0.5, pa možemo reći da druga glavna komponenta objašnjava i As, mada ne u meri kao i Pb, Cd i Hg .

Sledeći greafik predstavlja grafički koji su elementi u korelaciji sa kojom komponentom. Predstavljeno je svih 10 elemenata preko prve dve glavne komponente.



Slika 12 . Korelacija originalnih promenljivih i glavnih komponenti

Na garfiku možemo primetiti isto što smo uočili posmatrajući opterećenja glavnih komponenti. Prva glavna komponenta je u korelaciji sa Fe , Mn , Zn , Ni , Cu . Druga glavna komponenta je u korelaciji sa Pb , Cd i Hg .



Slika 13 . Korelacija originalnih promenljivih i faktora

Sa prethodne slike možemo takođe primetiti kako se elementi grupisu. Možemo primetiti da se oko prve komponente grupisu Zn, Mn, Fe, Cu i da im je veoma blizu Co. Njihove vrednosti su na prvoj komponenti najbliže jedinici. Zatim možemo primetiti da se oko druge komponete grupišu *Pb*, *Cd*, *Hg* i da im je veoma blizu *As,* kao i da su njihova opterećenja najveća na drugoj komponenti.

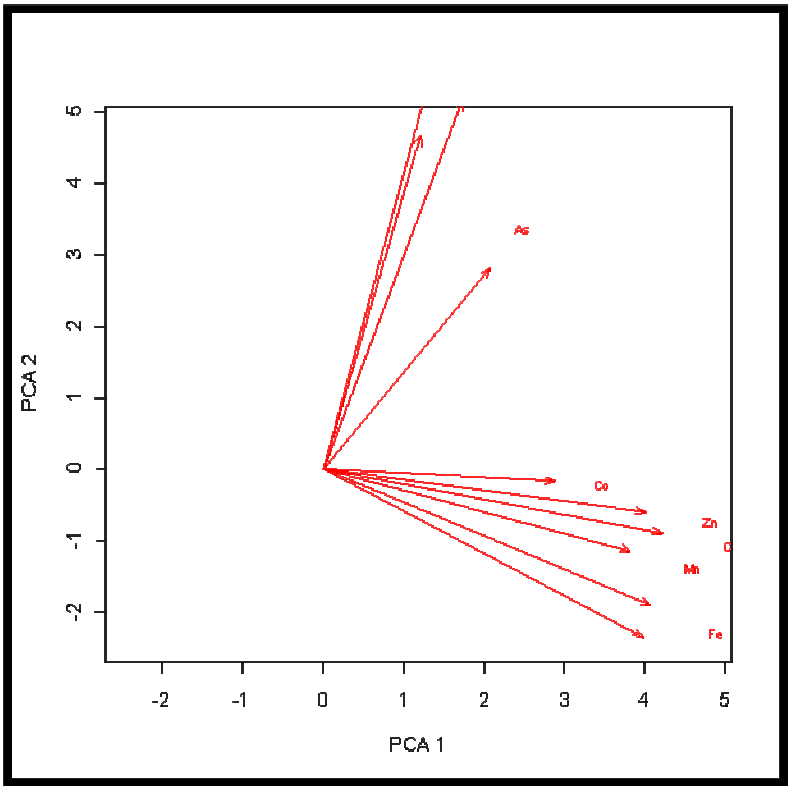
Jedan od ciljeva analize glavnih komponenti je otkrivanje suštinskog koncepta koji leži u osnovi podataka. Iz analize možemo primetiti da su elementi koji se grupišu: Fe, Mn, Zn, Ni, Cu . Oni čine prvu glavnu komponentu. Ovu glavnu komponentu možemo nazvati i „prirodna komponenta“. Ovo objašnjavamo činjenicom da elementi koji se grupišu oko „prirodne komponente“ predstavljaju hemijske elemente koji čine sedimente. Odnosno, oni čine hemijsku strukturu sedimenata. Takođe, možemo reći i da Co pripada ovoj komponenti. As iako više određuje drugu komponentu, možemo reći da u maloj meri određuje i prvu komponentu. To objašnjavamo činjenicom što se u prirodi As može naći kao pratilac gvožnja Fe, i to u rudi $FeAsS$.

Drugu glavnu komponentu možemo nazvati „antropogena komponenta“. Ona je u korelaciji sa elementima Pb, Cd i Hg . Ovi elementi predstavljaju teške toksične metale, i njihove tragove u sedimentima nalazimo isključivo zbog ljudskog uticaja. As više određuje drugu

glavnu komponentu. Iako se može u prirodi naći često i kao pratilac gvožnja (*Fe*), u ovom slučaju njegovo određivanje druge glavne komponente nam može nagovestiti da se u uzorcima sedimenata nalazi u većoj količini zbog antropogenog uticaja. To možemo objasniti time što se ostaci *As* u sedimentima mogu naći kao nataloženi ostaci sagorelih naftnih derivata, uglja, itd. koji se koriste za brodove. *Cd* se koristi kao dodatak u veštačkim đubrivima.

Ukoliko bismo dobijene rezultate hteli dalje da analiziramo, nekom drugom metodom, mogli bismo da koristimo sada dve umesto deset polaznih promenljivih. Ovim rezultatom smo postigli cilj analize glavnih komponenti, i smanjili polaznu dimenziju naših podataka. Otkrili smo i suštinski koncept koji leži osnovi naših podataka, a to je da se elementi *Pb* , *Cd* , *Hg* i *As* nalaze u sedimentima u većoj količini kao rezultat ljudskog uticaja, i kako i sami predstavljaju jako toksične metale, imaju velikog uticaja na zagađivanje životne sredine.

Isto možemo videti i na sledećoj slici.



Slika 14. Korelacija originalnih promenljivih i faktora

Primer 3. Indentifikacija lica primenom analize glavnih komponenti

Čovek je u stanju kroz ceo život prepoznavati na hiljade već viđenih lica, bez obzira na broj godina koje su prošle nakon poslednjeg susreta sa tim licima, uprkos mnogim promenama, starenju ili drugim smetnjama poput naočara, promene frizure ili slično. Prepoznavanje lica nije samo zanimljivo iz teorijskih razloga već i iz praktične primene. Prepoznavanje lica može se primeniti na niz problema iz stvarnog sveta, kao što su: indentifikacija zločinaca, obrada slika i filmova, itd. Međutim, razvoj algoritma za raspoznavanje lica je prilično težak zbog mnogobrojnih karakteristika lica.

Upravo taj problem se može rešiti metodom analize glavnih komponenti. Tom metodom možemo izdvojiti samo one karakteristike lica koje su nam bitne za prepoznavanje lica i kodirati ih što je efikasnije moguće. Tako bismo prilikom dobijanja nove slike mogli upoređivati tu sliku samo po izdvojenim karakteristikama sa ostalim slikama iz baze podataka i prema tome videti kojem licu nova slika pripada.

Konkretno, analiza glavnih komponenti omogućuje smanjenje dimenzije prostora značajnosti, tj. eliminaciju redundantnih podataka iz skupa. Ili matematički rečeno, želimo naći sopstvene vektore kovarijacione matrice skupa slika lica po kojima ćemo prepoznavati nove slike lica. Pri tome sliku tretiramo kao vektor.

Primer 4. Primena metode glavnih komponenata u redukciji dimenzije podataka prilikom obrade slika

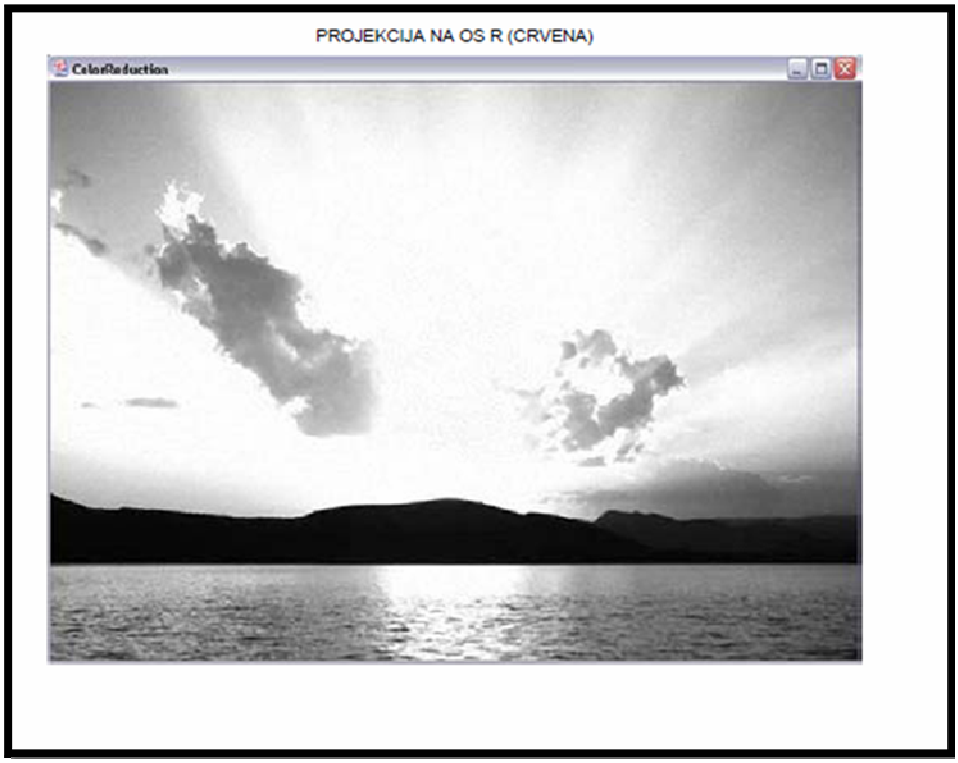
Boja u RGB zapisu je predstavljena u trodimenzionalnom prostoru čiju bazu čine vektori: R,G i B, koji odgovaraju crvenoj, plavoj i zelenoj boji. Dakle, slikovni element (piksel) je jedan vektor u prostoru koji obrazuju vektori R,G,B.

Slika je skup 3-dimenzionalnih podataka. Želimo da sliku u boji, pretvorimo u crno belu sliku, odnosno u nijanse sive. Ovaj postupak možemo posmatrati kao projekciju elemenata skupa iz 3-dimenzionalnog (R,G,B) prostora u 1-dimenizionalni prostor.

Slika 15. Originalna slika

Analiza glavnih komponenti određuje smer u kojem će projekcija imati najveću disperziju, odnosno određuje crno-belu projekciju slike koja će zadržati najviše informacija originalne slike.

Kako imamo sliku koja ima najviše nijansi crvene boje, bolje će izgledati slika koja ima projekciju svih poksela na osu R(crvena boja), nego projekcija na osu G (zelena boja) ili osu B (plava).



Slika 16. Projekcija piksela na osu R

Slika 17. Projekcija piksela na osu B

Posmatrajući prethodne dve slike, možemo zaista potvrditi da projekcija piksela na osu R, daje verniju sliku o originalu. Međutim projekcija na prvu glavnu komponentu, daje najverniju crno-belu sliku originalne slike, što se može uočiti na sledećoj slici.

Slika 18. Projekcija piksela na glavnu komponentu

Zaključak

Analiza glavnih komponenti je vrlo koristan i primenljiv metod u obradi statističkih podataka. Pored navedenih primena, analiza glavnih komponenti se koristi i u analizi tržista. Prikupe se podaci o karakteristikama nekog proizvoda, npr. kafe. Sprovede se pilot istraživanje, i dobijeni su podaci sa 15 karakteristika (promenljivih) kafe: ukus, užitak, buđenje, koncentracija, pauza, odmor... Nakon sprovođenja analize glavnih komponenti, broj promenljivih se svodi sa 15 na 4. Četiri glavne komponente koje možemo nazvati: „opuštanje“, „zavisnost“, „koncentracija“ i „ukus-miris“, zamenjaju svih 15 polaznih promenljivih. Ovim putem smo dobili najosnovnije karakteristike proizvoda što može dalje biti korisno za razvijanje novog proizvoda, bolje razumevanje kupaca, trendova na tržistu.

Analiza glnih komponenti se može sprovoditi kod raznih upitnika, istraživanja. Brojni su primeri primene u ispitivanju tržista, u medicinskim istraživanjima, bankarskim anketama, itd.

Komponente dobijene analizom glavnih komponenti mogu najčešće predstavljati tek polazne podatke za druge metode multivarijacione analize. U istraživanjima u kojima broj promenljivih prevazilazi broj observacija, može doći do probleme, pa je nužna redukcija dimenzije skupa promenljivih pomoću analize glavnih komponenti. Tako na primer, u diskriminacionoj analizi umesto originalnog skupa podataka koristimo izvestan broj glavnih komponenti u cilju formiranja diskriminacionih funkcija za razdvajanje grupa.

Međutim analiza glavnih komponenti ima i neka ograničenja. Najveće ograničenje analize glavne komponente je to što je ona veoma subjektivan proces. Svi elementi, određivanje broja glavnih komponenti, njihovo tumačenje i rotiranje (ukoliko jedan skup komponenti nije zadovoljavajući, možemo ponovo vršiti rotaciju), podrazumevaju donošenje subjektivnih odluka.

Ograničenje je i to što se ne koriste statistički testovi. Stoga je često teško znati da li je dobijeni rezultat slučajan ili stvarno ima smisla. Da bismo izbegli ovo dvoumljenje, trebalo bi uzorak na slučajan način podeliti na dve ili više grupa, i da se analiza glavnih komponenti vrši za svaku grupu nezavisno. Ako se u svakoj analizi pojavljuju iste komponente, tada bismo sa više sigrnosti mogli da tvrdimo da rezultat ne predstavlja statističku slučajnost.

Literatura

1. Härdle Wolfgang, Simar Léopold, *Applied multivariate statistical analysis*, Springer, Berlin 2003.

2. Kovačić Zlatko , *Multivarijaciona analiza*, Ekonomski fakultet Univerziteta u Beogradu, Beograd 1994.

3. Aker Dejvid A., Kumar, Dej Džordž S.,*Marketinško istraživanje*, John Wiley & Sons, Inc, USA, 2007 (knjiga prevedena sa engleskog jezika)

4. Radojičić Zoran, *Linearni statistički modeli*, materijal za predavanja, Ekonomski fakultet Univerziteta u Beogradu 1999.

5. Bogunović N., Bašić Dalbelo B., *Otkrivanje znanja o skupovima podataka, multivarijaciona analiza*, materijal za predavanja, Fakultet elektrotehnike i računarstva, Zagreb 2003/2004.

6. Butković Marijan, *Ispitivanje točnosti prepoznavanja lica primjenom analize glavnih komponenti*, Diplomski rad, Fakultet elektrotehnike i računarstva, Zagreb 2010.

7. Rencher Alvin C., *Methods of multivariate analysis*, John Wiley & Sons, Inc, USA, Brigham Young Uneversity 2002.

SADRŽAJ