Analiza glavnih komponenti

Strmečki, Dolores

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Dolores Strmečki

ANALIZA GLAVNIH KOMPONENTI

Diplomski rad

Voditelj rada:
doc. dr. sc. Hrvoje Planinić

Zagreb, rujan 2023.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Analiza glavnih komponenti (engl. *Principal component analysis*, PCA) statistička je metoda koja se ponajprije koristi za interpretaciju, vizualizaciju te smanjenje dimenzionalnosti velikih skupova podataka uz što veće očuvanje količine informacija sadržane u početnim podacima. Ideja metode potječe od engleskog matematičara Karla Pearsona još iz 1901. godine, ali je za njezinu formulaciju 1933. i daljnji razvitak zaslužan američki statističar Harold Hotelling. Na samim počecima, primjena analize glavnih komponenti nije bila toliko raširena, a procvat doživljava krajem 20. i početkom 21. stoljeća zbog ubrzanog razvoja računalnih tehnologija te sve veće dostupnosti podataka.

Analiza glavnih komponenti nerijetko otkriva veze između varijabli koje nisu uočljive na prvi pogled, čime otvara prostor novom tumačenju podataka te sve raširenu uporabu u različitim područjima znanosti. Primjerice, u strojnom učenju ova se metoda koristi za vizualizaciju podataka jer omogućuje projekciju podataka u niže dimenzije, što olakšava razumijevanje odnosa među promatranim varijablama. S druge strane, u svijetu financija analiza glavnih komponenti pomaže u otkrivanju anomalija kao što su prijevare, a u bioinformatici se koristi za klasifikaciju uzoraka tkiva ili bolesti.

Glavni cilj ovoga rada upoznavanje je s metodom analiziranja glavnih komponenti: od same definicije glavnih pojmova preko teorijske pozadine koja koristi klasične matematičke rezultate iz područja vjerojatnosti i statistike pa sve do same primjene analize glavnih komponenti na stvarnim problemima, odnosno podacima. Izlaganje rezultata i primjera u ovome radu najviše se temelji na Johnsonovoj i Wichernovoj knjizi *Applied Multivariate Statystical Analysis* [6].

Prvo poglavlje bavi se izlaganjem teorijske pozadine metode analize glavnih komponenti. Definicije i teorijski rezultati potrebni za razumijevanje ovog rada navedeni su u podnožjima stranica, a mogu se pronaći u [4] i [1]. Glavne komponente definiraju se kao nekorelirane linearne kombinacije slučajnih varijabli koje reprezentiraju inicijalni skup podataka, čija je varijanca najveća moguća. Opisuju se metode pronalaska glavnih komponenti koristeći svojstvene vrijednosti i svojstvene vektore kovarijacijske matrice inicijalnog skupa podataka za početne, ali i za standardizirane varijable. Ilustriran je i izračun uzoračkih glavnih komponenti, bitan kada je na raspolaganju skup rezultata

nezavisnih realizacija promatranih slučajnih varijabli. Nadalje, raspravlja se o odgovoru na jako bitno pitanje: koliko glavnih komponenti uzeti u obzir prilikom analize kako bi se očuvalo što više informacija sadržanih u početnim podacima, a smanjio teret rada s velikim, višedimenzionalnim skupovima podataka. Na samom kraju poglavlja dana je geometrijska interpretacija glavnih komponenti.
U drugom poglavlju obrađeni su problemi iz stvarnog svijeta te je ilustrirano koliko analiza glavnih komponenti može biti snažan alat za deskriptivnu analizu i bolje razumijevanje podataka, što samo po sebi može dovesti do bitnih poslovnih i znanstvenih zaključaka ili usmjeriti daljnju analizu podataka na pravi način.

Poglavlje 1

Analiza glavnih komponenti

Središnja ideja metode je linearna transformacija inicijalnog skupa podataka u skup manje dimenzije uz što manji gubitak informacija o podacima. Uzmimo kao primjer razinu čovjekovog zadovoljstva njegovim zaposljenjem. Na zadovoljstvo očito utječu brojni parametri kao što su radno okruženje, visina plaće ili udaljenost radnog mjesta od mjesta stanovanja. Također, promjena vrijednosti jednog parametra može utjecati na promjenu vrijednosti drugog, što daljnju analizu takvog skupa podataka čini zahtjevnijom. Iznos zadovoljstva bilo bi znatno lakše ispitati u skupu manje dimenzije, u kojem je dodatno promjena vrijednosti jednog parametra nezavisna od promjene vrijednosti drugog. Kreiranje opisanog skupa uz zadržavanje što više informacija upravo ilustrira metoda analize glavnih komponenti.

1.1 Glavne komponente

Pretpostavimo da je inicijalni skup podataka reprezentiran nizom slučajnih varijabli X_i, $i = 1, ..., p$ pri čemu je $p \in \mathbb{N}$. Neka $Y_1, Y_2, ..., Y_p$ označuju linearne kombinacije slučajnih varijabli X_i, $i = 1, ..., p$.
Preciznije,

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p = \mathbf{a}_1'\mathbf{X}, \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p = \mathbf{a}_2'\mathbf{X}, \\ &\vdots \\ Y_p &= a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p = \mathbf{a}_2'\mathbf{X}, \end{aligned} \tag{1.1}$$

gdje je $\mathbf{X}'$[1] slučajni vektor $[X_1, X_2, ..., X_p]$, a $\mathbf{a_i}'$ su realni vektori $[a_{i1}, a_{i2}, ..., a_{ip}]$ za $i = 1, ..., p$.

[1] $X' \in \mathbb{R}^{1xp}$ označuje transponirani vektor vektora $X \in \mathbb{R}^{px1}$.

Za postizanje što manjeg gubitka informacija o danom skupu podataka bitno je sačuvati što veću varijaciju između početnih varijabli X_i, $i = 1, ..., p$.
Općenito, u uzorku od jedne slučajne varijable, varijanca[2] te slučajne varijable koristi se za opisivanje količine varijacije te varijable. Kada se promatrani uzorak sastoji od $p \in \mathbb{N}$ slučajnih varijabli, varijacija između varijabli opisuje se kovarijacijskom matricom[3].
Glavne komponente su zapravo nekorelirane[4] linearne kombinacije $Y_1, Y_2, ..., Y_p$ dane formulom (1.1) čija je varijanca što je moguće veća.
Ako je $\mathbf{\Sigma}$ kovarijacijska matrica vektora $\mathbf{X}'$, tada je varijanca linearnih kombinacija Y_i dana s

$$\begin{aligned}
\mathrm{Var}(Y_i) &= \mathrm{Var}(\mathbf{a}_i'\mathbf{X}) \\
&= \mathrm{Var}\left[\sum_{j=1}^{p} a_{ij}X_j\right] \\
&= \sum_{j=1}^{p} \mathrm{Var}\left[a_{ij}X_j\right] + 2\sum_{1\le j<k\le p} \mathrm{Cov}\left[a_{ij}X_j, a_{ik}X_k\right] \qquad (1.2) \\
&= \sum_{j=1}^{p} a_{ij}^2 \,\mathrm{Var}\left[X_j\right] + 2\sum_{1\le j<k\le p} a_{ij}a_{ik}\,\mathrm{Cov}\left[X_j, X_k\right] \\
&= \mathbf{a_i'\Sigma a_i},
\end{aligned}$$

za svaki $i = 1, ..., p$.
Iz formule (1.2) je vidljivo kako se varijanca $\mathrm{Var}(Y_i)$ može povećati množenjem vektora $\mathbf{a_i}$ s nekom konstantom. Kako bi se izbjegla takva neodređenost, pretpostavlja se da su vektori $\mathbf{a_i}$ vektori jedinične duljine [5].
Glavne komponente se sada mogu definirati kao:

Prva glavna komponenta $(Y_1) \rightarrow$ linearna kombinacija $\mathbf{a_1}'\mathbf{X}$ koja maksimizira $\mathrm{Var}(\mathbf{a_1}'\mathbf{X})$ uz uvjet $\mathbf{a_1}'\mathbf{a_1} = 1$.

---

[2] **Varijanca** slučajne varijable X definira se kao $\mathrm{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$.

[3] **Kovarijacijska matrica** ili **matrica kovarijanci** slučajnog vektora $\mathbf{Z} = (X_1, X_2, ..., X_p)$ definira se kao

$$\mathbf{Cov}(\mathbf{Z}) = \mathbb{E}\left[(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^T\right] = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_1, X_p) & \mathrm{Cov}(X_2, X_p) & \cdots & \mathrm{Var}(X_p) \end{bmatrix}$$ pri čemu je

$\mathrm{Cov}(X_i, X_j) = \mathbb{E}\left[(\mathbf{X_i} - \mathbb{E}[\mathbf{X_i}])(\mathbf{X_j} - \mathbb{E}[\mathbf{X_j}])\right]$ kovarijanca slučajnih varijabli X_i i X_j za $i, j = 1, ..., p$.

[4] Za slučajne varijable X i Y kažemo da su **nekorelirane** ako je $\mathrm{Cov}(X, Y) = 0$.

[5] Za vektor $\mathbf{a} \in \mathbb{R}^p$ kažemo da je **jedinične duljine** ili normaliziran ako mu je duljina (ili norma) $\|\mathbf{a}\| = 1$. Analogno, za vektor $\mathbf{a} \in \mathbb{R}^p$ kažemo da je jedinične duljine ako vrijedi $\mathbf{a}'\mathbf{a} = 1$.

Druga glavna komponenta (Y_2) $\rightarrow$ linearna kombinacija $\mathbf{a_2}'\mathbf{X}$ koja maksimizira $\text{Var}(\mathbf{a_2}'\mathbf{X})$ uz uvjete $\mathbf{a_2}'\mathbf{a_2} = 1$ i $\text{Cov}(\mathbf{a_1}'\mathbf{X}, \mathbf{a_2}'\mathbf{X}) = 0$.

...

i-ta glavna komponenta (Y_i) $\rightarrow$ linearna kombinacija $\mathbf{a_i}'\mathbf{X}$ koja maksimizira $\text{Var}(\mathbf{a_i}'\mathbf{X})$ uz uvjete $\mathbf{a_i}'\mathbf{a_i} = 1$ i $\text{Cov}(\mathbf{a_k}'\mathbf{X}, \mathbf{a_i}'\mathbf{X}) = 0$ za $k < i$.

Dakle, prvi korak je maksimizirati varijancu danu formulom (1.2) na prostoru jediničnih vektora, tj. pronaći rješenje problema

$$\begin{cases} \mathbf{a_1}'\mathbf{\Sigma}\mathbf{a_1} \rightarrow \max \\ \mathbf{a_1}'\mathbf{a_1} = 1 \end{cases} \tag{1.3}$$

Jedan od načina dolaska do rješenja je uvođenjem Lagrangeove funkcije

$$\mathbf{L}(\mathbf{a_1}, \lambda) = \mathbf{a_1}'\mathbf{\Sigma}\mathbf{a_1} + \lambda(1 - \mathbf{a_1}'\mathbf{a_1}),$$

gdje je $\lambda \in \mathbb{R}$ Lagrangeov multiplikator.
Problem (1.3) sada se svodi na problem rješavanja sustava parcijalnih jednadžbi[6]

$$\frac{\partial \mathbf{L}}{\partial \mathbf{a_1}}(\mathbf{a_1}, \lambda) = 2\mathbf{\Sigma}\mathbf{a_1} - 2\lambda\mathbf{a_1} = 0, \tag{1.4}$$

$$\frac{\partial \mathbf{L}}{\partial \lambda}(\mathbf{a_1}, \lambda) = 1 - \mathbf{a_1}'\mathbf{a_1} = 0. \tag{1.5}$$

Izraz (1.4) jednak je

$$\mathbf{\Sigma}\mathbf{a_1} = \lambda\mathbf{a_1}. \tag{1.6}$$

Odnosno, vektor $\mathbf{a_1}$ i λ su svojstveni vektor i svojstvena vrijednost matrice Σ. Nadalje, množenjem jednakosti (1.6) slijeva s vektorom $\mathbf{a_1}'$ slijedi

$$\mathbf{a_1}'\mathbf{\Sigma}\mathbf{a_1} = \text{Var}(Y_1) = \mathbf{a_1}'\lambda\mathbf{a_1} = \lambda\mathbf{a_1}'\mathbf{a_1} = \lambda,$$

---

[6]Općenito, parcijalna derivacija funkcije $\mathbf{f} : A \rightarrow \mathbb{R}$, $A \subseteq \mathbb{R}^n$ otvoren, po vektoru $\mathbf{v} \in \mathbb{R}^n$ je zapravo vektor parcijalnih derivacija te funkcije po komponentama tog vektora, odnosno $\frac{\partial \mathbf{f}}{\partial \mathbf{v}} = (\frac{\partial \mathbf{f}}{\partial v_1}, \frac{\partial \mathbf{f}}{\partial v_2}, ..., \frac{\partial \mathbf{f}}{\partial v_n})$.

Primjeri nekih standardnih derivacija koje se koriste u nastavku rada:

1. $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}$, $\mathbf{f}(x) = \mathbf{b}'\mathbf{x}$, $\mathbf{b} \in \mathbb{R}^n \Rightarrow \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}}(\sum_{i=1}^n b_i x_i) = (\frac{\partial(\sum_{i=1}^n b_i x_i)}{\partial x_1}, \frac{\partial(\sum_{i=1}^n b_i x_i)}{\partial x_2}, ..., \frac{\partial(\sum_{i=1}^n b_i x_i)}{\partial x_n}) = (b_1, , b_2, ..., b_n) = \mathbf{b}$
2. $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}$, $\mathbf{f}(x) = \mathbf{x}'\mathbf{x} \Rightarrow \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}}(\sum_{i=1}^n x_i^2) = (\frac{\partial(\sum_{i=1}^n x_i^2)}{\partial x_1}, \frac{\partial(\sum_{i=1}^n x_i^2)}{\partial x_2}, ..., \frac{\partial(\sum_{i=1}^n x_i^2)}{\partial x_n}) = (2x_1, , 2x_2, ..., 2x_n) = 2\mathbf{x}$
3. $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}$, $\mathbf{f}(x) = \mathbf{x}'\mathbf{A}\mathbf{x}$, gdje je $\mathbf{A} = (a_{ij}) \in M_n(\mathbb{F})$ simetrična matrica $\Rightarrow \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}}(\sum_{i,j=1}^n x_i x_j a_{ij}) = (\frac{\partial(\sum_{i,j=1}^n x_i x_j a_{ij}}{\partial x_1}, \frac{\partial(\sum_{i,j=1}^n x_i x_j a_{ij}}{\partial x_2}, ..., \frac{\partial(\sum_{i,j=1}^n x_i x_j a_{ij})}{\partial x_n}) = (2x_1 \sum_{j=1}^n a_{1j}, , 2x_2 \sum_{j=1}^n a_{2j}, ..., 2x_n \sum_{j=1}^n a_{nj}) = 2\mathbf{A}\mathbf{x}$

gdje posljednja jednakost vrijedi zbog (1.5).
Dakle, varijanca $\text{Var}(Y_1)$ postiže maksimum za vektor $\mathbf{a_1}$ koji je jednak svojstvenom vektoru matrice Σ koji odgovara najvećoj svojstvenoj vrijednosti λ od matrice Σ. Drugim riječima, prva glavna komponenta Y_1 je linearna kombinacija vektora X i svojstvenog vektora kovarijacijske matrice Σ koji odgovara najvećoj svojstvenoj vrijednosti te matrice.

Korolar 1.1.1. Neka je Σ kovarijacijska matrica slučajnog vektora $\mathbf{X}' = [X_1, X_2, ..., X_p]$ i neka su $(\lambda_1, \mathbf{v_1}), (\lambda_2, \mathbf{v_2}), ..., (\lambda_p, \mathbf{v_p})$ parovi svojstvenih vrijednosti λ_i i svojstvenih vektora $\mathbf{v_i}$ matrice Σ uz $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Tada je i-ta glavna komponenta dana s

$$Y_i = \mathbf{v_i}'\mathbf{X} = v_{i1}X_1 + v_{i2}X_2 + ... + v_{ip}X_p, \qquad i = 1, ..., p.$$

S tim izborom,

$$\begin{aligned} \text{Var}(Y_i) &= \mathbf{v_i'}\mathbf{\Sigma}\mathbf{v_i} = \lambda_i, \qquad i = 1, ..., p, \\ \text{Cov}(Y_i, Y_k) &= \mathbf{v_i}'\mathbf{\Sigma}\mathbf{v_k} = 0, \qquad i \neq k. \end{aligned}$$

Dokaz. Za $i = 1$ iznad je pokazano da je Y_1 linearna kombinacija svojstvenog vektora v_1 (jer je to svojstveni vektor koji odgovara najvećoj svojstvenoj vrijednosti kovarijacijske matrice Σ) i vektora $\mathbf{X}$ i da je $\text{Var}(Y_1) = \lambda_1$.
Druga glavna komponenta $Y_2 = \mathbf{a_2}'\mathbf{X}$ treba maksimizirati $\text{Var}(\mathbf{a_2}'\mathbf{X})$ i biti nekorelinara s Y_1.
Općenito, kovarijanca varijabli Y_i i Y_k je oblika

$$\begin{aligned} \text{Cov}(Y_i, Y_k) &= \text{Cov}((\mathbf{a_i}'\mathbf{X}), (\mathbf{a_k}'\mathbf{X})) \\ &= E\left[(\mathbf{a_i}'\mathbf{X} - E[\mathbf{a_i}'\mathbf{X}])(\mathbf{a_k}'\mathbf{X} - E[\mathbf{a_k}'\mathbf{X}])\right] \\ &= E\left[\mathbf{a_i}'(\mathbf{X} - E[\mathbf{X}])\mathbf{a_k}'(\mathbf{X} - E[\mathbf{X}])\right] \\ &= \mathbf{a_i}'E\left[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T\right]\mathbf{a_k} \\ &= \mathbf{a_i}'\,\text{Cov}(\mathbf{X})\mathbf{a_k} \\ &= \mathbf{a_i}'\mathbf{\Sigma}\mathbf{a_k}, \end{aligned} \tag{1.7}$$

za svaki $i, k = 1, ..., p$.
Dakle, treba vrijediti

$$\text{Cov}(Y_1, Y_2) = \mathbf{a_1}'\mathbf{\Sigma}\mathbf{a_2} = \mathbf{a_2}'\mathbf{\Sigma}\mathbf{a_1} = \mathbf{a_2}'\lambda_1\mathbf{a_1} = \lambda_1\mathbf{a_2}'\mathbf{a_1} = 0,$$

gdje druga jednakost slijedi iz simetričnosti kovarijacijske matrice Σ, a treća jednakost slijedi iz izraza (1.6).

Ako pretpostavimo da je $\lambda_1 \neq 0$[7], tada je problem pronalaženja druge komponente dan zadaćom

$$\begin{cases} \mathbf{a_2}'\mathbf{\Sigma a_2} \to \max \\ \mathbf{a_2}'\mathbf{a_2} = 1 \\ \mathbf{a_2}'\mathbf{a_1} = 0 \end{cases}$$

Pripadna Lagrangeova funkcija je oblika

$$\mathbf{L}(\mathbf{a_2}, \mu, \phi) = \mathbf{a_2}'\mathbf{\Sigma a_2} + \mu(1 - \mathbf{a_2}'\mathbf{a_2}) + \phi(\mathbf{a_2}'\mathbf{a_1}),$$

uz Lagrangeove multiplikatore $\mu, \phi \in \mathbb{R}$.
Rješavamo sljedeći sustav parcijalnih jednadžbi

$$\frac{\partial \mathbf{L}}{\partial \mathbf{a_2}}(\mathbf{a_2}, \mu, \phi) = 2\mathbf{\Sigma a_2} - 2\mu\mathbf{a_2} + \phi\mathbf{a_1} = 0, \tag{1.8}$$

$$\frac{\partial \mathbf{L}}{\partial \mu}(\mathbf{a_2}, \mu, \phi) = 1 - \mathbf{a_2}'\mathbf{a_2} = 0, \tag{1.9}$$

$$\frac{\partial \mathbf{L}}{\partial \phi}(\mathbf{a_2}, \mu, \phi) = \mathbf{a_2}'\mathbf{a_1} = 0. \tag{1.10}$$

Množenjem jednakosti (1.8) slijeva s $\mathbf{a_1}'$ dobivamo

$$2\mathbf{a_1}'\mathbf{\Sigma a_2} - 2\mu\mathbf{a_1}'\mathbf{a_2} + \phi\mathbf{a_1}'\mathbf{a_1} = 0.$$

Kako je $\mathbf{a_1}'\mathbf{\Sigma a_2} = 0$ (kovarijanca) i $\mathbf{a_1}'\mathbf{a_2} = 0$ (uvjet (1.10)) i kako vrijedi $\mathbf{a_1}'\mathbf{a_1} = 1$, slijedi da je $\phi = 0$.
Koristeći činjenicu da je $\phi = 0$, izraz (1.8) možemo zapisati kao

$$\mathbf{\Sigma a_2} = \mu\mathbf{a_2}. \tag{1.11}$$

Odnosno, vektor $\mathbf{a_2}$ i μ su svojstveni vektor i svojstvena vrijednost matrice Σ. Množenjem izraza (1.11) slijeva s vektorom $\mathbf{a_2}'$ slijedi

$$\mathbf{a_2}'\mathbf{\Sigma a_2} = \mathrm{Var}(Y_2) = \mathbf{a_2}'\mu\mathbf{a_2} = \mu\mathbf{a_2}'\mathbf{a_2} = \mu,$$

gdje posljednja jednakost vrijedi zbog (1.9).
Kao i u slučaju $i = 1$, dobivamo da je varijanca $\mathrm{Var}(Y_2)$ jednaka svojstvenoj vrijednosti matrice Σ. Dakle, tražimo jedinični vektor $\mathbf{a_2}$ (uvjet (1.9)) koji odgovara što većoj svojstvenoj vrijednosti i koji je ortogonalan sa svojstvenim vektorom $\mathbf{a_1}$ (uvjet (1.10)). Rješenje je upravo v_2, odnosno svojstveni vektor druge po redu najveće svojstvene vrijednosti matrice Σ. Također, iz ortogonalnosti svojstvenih vektora[8] slijedi i

[7] U slučaju da su $\lambda_m = ... = \lambda_p = 0$ za neki $m \geq 1$, onda je dovoljno uzeti da su vektori $\mathbf{a_m}, ..., \mathbf{a_p}$ ortonormirana baza za jezgru matrice Σ. Ortonormirana baza je skup vektora jedinične duljine koji su međusobno ortogonalni (okomiti), a jezgra matrice Σ definira se kao $\mathrm{Ker}\,\Sigma = \left\{x \in \mathbb{R}^{px1} : \Sigma x = 0\right\}$.

[8] Ako kao pretpostavku uzmemo da su sve svojstvene vrijednosti različite tada su svi svojstveni vektori međusobno ortogonalni. U slučaju da postoje jednake svojstvene vrijednosti tada se svojstveni vektori mogu izabrati tako da budu ortogonalni.

nekoreliranost varijabli Y_1 i Y_2, tj. $\text{Cov}(Y_1, Y_2) = 0$. Time je dokazana tvrdnja korolara za $i = 2$.
Analognim postupkom dobije se da su treća, četvrta,..., p-ta glavna komponenta linearne kombinacije vektora $\mathbf{X}$ i svojstvenih vektora matrice Σ koji odgovaraju trećoj, četvrtoj,..., p-toj po veličini svojstvenoj vrijednost matrice Σ i da vrijedi $\text{Var}(Y_i) = \lambda_i$ te $\text{Cov}(Y_i, Y_k) = 0$ za $i, k = 1, ..., p$, $i \neq k$. □

Sljedeći rezultat pokazuje kako je i ukupna varijanca polaznog skupa podataka sačuvana u glavnim komponentama.

Korolar 1.1.2. Neka je $\Sigma = (\sigma_{ij}) \in M_p(\mathbb{F})$[9]kovarijacijska matrica slučajnog vektora $\mathbf{X}' = [X_1, X_2, ..., X_p]$ s parovima svojstvenih vrijednosti i svojstvenih vektora $(\lambda_1, \mathbf{v_1})$, $(\lambda_2, \mathbf{v_2})$,..., $(\lambda_p, \mathbf{v_p})$ uz $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ i $p \in \mathbb{N}$. Neka su $Y_1 = \mathbf{v_1}'\mathbf{X}$, $Y_2 = \mathbf{v_2}'\mathbf{X}$,..., $Y_p = \mathbf{v_p}'\mathbf{X}$ glavne komponente. Tada vrijedi

$$\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \sum_{i=1}^{p} \text{Var}(X_i) = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1}^{p} \text{Var}(Y_i).$$

Dokaz. Iz definicije kovarijacijske matrice znamo da je $\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \sum_{i=1}^{p} \text{Var}(X_i) = \text{tr}(\Sigma)$[10]. Nadalje, označimo s Λ dijagonalnu matricu čiji je k-ti dijagonalni element k-ta svojstvena vrijednost λ_k i s $\mathbf{P}$ ortogonalnu matricu[11] $\left[\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_p}\right]$.
Sada po definiciji svojstvenih vrijednosti i svojstvenih vektora vrijedi

$$\Sigma\mathbf{P} = \mathbf{P}\Lambda.$$

Množenjem s $\mathbf{P}'$ zdesna, zbog ortogonalnosti matrice $\mathbf{P}$, slijedi

$$\Sigma = \mathbf{P}\Lambda\mathbf{P}'. \tag{1.12}$$

Koristeći jednakost (1.12) i svojstvo operatora traga za umnožak matrica[12] dobivamo

$$\sum_{i=1}^{p} \text{Var}(X_i) = \text{tr}(\Sigma) = \text{tr}(\mathbf{P}\Lambda\mathbf{P}') = \text{tr}(\Lambda\mathbf{P}'\mathbf{P}) = \text{tr}(\Lambda) = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1}^{p} \text{Var}(Y_i).$$

□

[9]Općenito, za $m, n \in \mathbb{N}$, $M_{mn}(\mathbb{F})$ je oznaka za skup svih matrica s m-redaka i n-stupaca. Za $m = n$, odnosno za kvadratnu matricu, koristi se oznaka $M_n(\mathbb{F})$.

[10]**Trag** matrice $\mathbf{A} = (a_{ij}) \in M_n(\mathbb{F})$ definira se kao $\mathbf{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.

[11]Za kvadratnu matricu $\mathbf{A}$ kažemo da je **ortogonalna** ako je $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$, gdje $\mathbf{I}$ označuje jediničnu matricu. Ortogonalnost matrice $\mathbf{P}$ slijedi iz ortogonalnosti i jedinične duljine svojstvenih vektora $\mathbf{v_i}$, $i = 1, ..., p$.

[12]Za matrice $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{F})$ vrijedi $\mathbf{tr}(\mathbf{AB}) = \mathbf{tr}(\mathbf{BA})$.

Kao posljedica korolara 1.1.1. i 1.1.2, udio ukupne varijance opisane pomoću k-te glavne komponente dan je izrazom

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

gdje su λ_i svojstvene vrijednosti matrice Σ za $i = 1, ..., p$.
Dakle, kada je suma varijanci prvih nekoliko glavnih komponenti iznosom blizu ukupne varijance, te komponente sadrže većinu ključnih informacija o polaznom skupu podataka. "Zamjena" početnih p varijabli s ovim vodećim komponentama pojednostavila bi daljnju analizu bez značajnog gubitka informacija.
Nadalje, promatrajući jedan vektor koeficijenata $\mathbf{v_i}$, za proizvoljni $i \in \{1, ..., p\}$, njegove komponente $v_{i1}, v_{i2}, ..., v_{ip}$ pomažu u interpretaciji odnosa glavne komponente Y_i i početnih varijabli $X_1, X_2, ..., X_p$. Veličina v_{ik} mjeri važnost k-te početne varijable za i-tu glavnu komponentu bez obzira na ostale početne varijable. Dodatno, ona je proporcionalna koeficijentu korelacije između k-te početne varijable i i-te glavne komponente.

Korolar 1.1.3. Neka su $Y_1 = \mathbf{v_1'X}$, $Y_2 = \mathbf{v_2'X}$,..., $Y_p = \mathbf{v_p'X}$ glavne komponente dobivene iz kovarijacijske matrice $\Sigma = (\sigma_{ij}) \in M_p(\mathbb{F})$ s parovima svojstvenih vrijednosti i svojstvenih vektora $(\lambda_1, \mathbf{v_1})$, $(\lambda_2, \mathbf{v_2})$,..., $(\lambda_p, \mathbf{v_p})$. Tada su s

$$\rho_{Y_i,X_k} = \frac{v_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \qquad i, k = 1, ..., p$$

dani koeficijenti korelacije[13] između glavne komponente Y_i i početne varijable X_k.

Dokaz. Neka su $i, k \in \{1, ..., p\}$ proizvoljni. Definirajmo s $\mathbf{e_k}$ p-dimenzionalni vektor koji na k-tom mjestu ima jedinicu, a na ostalim mjestima nule. Tada se X_k može zapisati kao $X_k = \mathbf{e_k}'\mathbf{X}$.
Iz formule (1.7) i simetričnosti matrice Σ slijedi

$$\mathrm{Cov}(Y_i, X_k) = \mathrm{Cov}(\mathbf{v_i}'\mathbf{X}, \mathbf{e_k}'\mathbf{X}) = \mathbf{v_i}'\Sigma\mathbf{e_k} = \mathbf{e_k}'\Sigma\mathbf{v_i}.$$

Kako je $\Sigma\mathbf{v_i} = \lambda_i\mathbf{v_i}$ dobivamo da je $\mathrm{Cov}(Y_i, X_k) = \mathbf{e_k}'\lambda_i\mathbf{v_i} = \lambda_i v_{ik}$. Sada iz korolara 1.1.1. znamo da je $\mathrm{Var}(Y_i) = \lambda_i$, a iz definicije kovarijacijske matrice Σ da je $\mathrm{Var}(X_k) = \sigma_{kk}$ pa slijedi da je koeficijent korelacije jednak

$$\rho_{Y_i,X_k} = \frac{\mathrm{Cov}(Y_i, X_k)}{\sqrt{\mathrm{Var}(Y_i)}\sqrt{\mathrm{Var}(X_k)}} = \frac{\lambda_i v_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{v_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \qquad i, k = 1, ..., p$$

□

[13] **Koeficijent korelacije** između slučajnih varijabli X i Y definira se kao $\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$.

Općenito, koeficijenti korelacije između glavne komponente i početnih varijabli doprinose u tumačenju značaja te glavne komponente. Na primjer, ako je neka početna varijabla izrazito korelirana s promatranom glavnom komponentom, ona ima značajan utjecaj na tu glavnu komponentu. Tada bi ta komponenta mogla biti važna u tumačenju varijabilnosti podataka povezane s tom početnom varijablom. Međutim, treba naglasiti kako koeficijent korelacije mjeri isključivo doprinos pojedinačne početne varijable na neku glavnu komponentu. On ne daje informaciju o važnosti te početne varijable u odnosu na druge početne varijable u kontekstu te glavne komponente.

1.1.1 Glavne komponente dobivene iz standardiziranih varijabli

Promotrimo sada glavne komponente dobivene iz standardiziranih slučajnih varijabli. Pretpostavljamo da je $[X_1, X_2, ..., X_p] = \mathbf{X}' \sim N_p(\mu, \mathbf{\Sigma})$ pri čemu je $p \in \mathbb{N}$, tj. vektor reprezentanata inicijalnog skupa podataka je normalno distribuiran s vektorom očekivanja μ i kovarijacijskom matricom $\mathbf{\Sigma} = (\sigma_{ij})$ za $i, j = 1, ..., p$.
Nadalje, neka vektor $\mathbf{Z}' = [Z_1, Z_2, ..., Z_p]$ označuje vektor dobiven postupkom standardizacije[14] inicijalnog normalnog slučajnog vektora $\mathbf{X}'$. Dakle, komponente $Z_1, Z_2, ..., Z_p$ dane su s:

$$\begin{aligned} Z_1 &= \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}, \\ Z_2 &= \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}}, \\ &\vdots \\ Z_p &= \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}. \end{aligned} \tag{1.13}$$

Tada je iz definicije standardizirane slučajne varijable vektor $\mathbf{Z}$ normalan slučajan vektor, a pripadna kovarijacijska matrica $\boldsymbol{\rho}$ jednaka je korelacijskoj matrici [15] vektora $\mathbf{X}$.

[14] Neka je X normalna slučajna varijabla s očekivanjem μ i varijancom σ^2, tj. $X \sim N(\mu, \sigma^2)$. Njoj pridružena **standardizirana slučajna varijabla** definira se kao $Z = \frac{X-\mu}{\sigma}$ i za nju vrijedi $Z \sim N(0, 1)$.

[15] **Korelacijska matrica** ili **matrica korelacije** slučajnog vektora $\mathbf{X} = (X_1, X_2, ..., X_p)$ je kovarijacijska matrica njemu pridruženog standardiziranog slučajnog vektora $\mathbf{Z} = (\frac{X_1-\mu_1}{\sqrt{\sigma_{11}}}, \frac{X_2-\mu_2}{\sqrt{\sigma_{22}}}, ..., \frac{X_p-\mu_p}{\sqrt{\sigma_{pp}}})$, gdje je $\mu_i = \mathbb{E}(X_i)$ i $\sigma_{ii} = \mathrm{Var}(X_i)$ za $i = 1, ..., p$. Definira se kao:

$$\mathbf{Corr}(\mathbf{X}) = \begin{bmatrix} 1 & \rho_{X_1,X_2} & \cdots & \rho_{X_1,X_p} \\ \rho_{X_1,X_2} & 1 & \cdots & \rho_{X_2,X_p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_1,X_p} & \rho_{X_2,X_p} & \cdots & 1 \end{bmatrix},$$

pri čemu je ρ_{X_i,X_j} koeficijent korelacije između slučajnih varijabli X_i i X_j za $i, j = 1, ..., p$.

Glavne komponente $Y_1, Y_2, ..., Y_p$ sada se računaju kao linearne kombinacije standardiziranih slučajnih varijabli Z_i, $i = 1, ..., p$.
Sljedeći rezultat dobiven je kao jednostavna primjena korolara 1.1.1 - 1.1.3. na slučajan vektor $\mathbf{Z}$, pa se dokaz izostavlja.

Korolar 1.1.4. Neka je $\boldsymbol{\rho}$ kovarijacijska matrica standardiziranog slučajnog vektora $\mathbf{Z}' = [Z_1, Z_2, ..., Z_p]$ i neka su $(\lambda_1, \mathbf{v_1}), (\lambda_2, \mathbf{v_2}), ..., (\lambda_p, \mathbf{v_p})$ parovi svojstvenih vrijednosti λ_i i svojstvenih vektora $\mathbf{v_i}$ matrice ρ uz $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Tada je i-ta glavna komponenta dana s

$$Y_i = \mathbf{v_i}'\mathbf{Z} = v_{i1}Z_1 + v_{i2}Z_2 + ... + v_{ip}Z_p, \qquad i = 1, ..., p.$$

S tim izborom,

$$\begin{aligned} \mathrm{Var}(Y_i) &= \mathbf{v_i'}\boldsymbol{\rho}\mathbf{v_i} = \lambda_i, \qquad i = 1, ..., p \\ \mathrm{Cov}(Y_i, Y_k) &= \mathbf{v_i}'\boldsymbol{\rho}\mathbf{v_k} = 0, \qquad i \neq k. \end{aligned}$$

Također,

$$\sum_{i=1}^{p} \mathrm{Var}(Z_i) = \sum_{i=1}^{p} \mathrm{Var}(Y_i)$$

i vrijedi

$$\rho_{Y_i,Z_k} = v_{ik}\sqrt{\lambda_i}, \qquad i, k = 1, ..., p.$$

Uočimo, kako su $Z_i \sim N(0, 1)$, slijedi da je $\sum_{i=1}^{p} \mathrm{Var}(Z_i) = p$. Dakle, ukupna varijaca standardiziranih varijabli je jednaka broju reprezentanata inicijalnog skupa podataka. Posljedično, udio standardizirane ukupne varijance opisane pomoću k-te glavne komponente dan je izrazom $\frac{\lambda_k}{p}$, gdje je λ_k svojstvena vrijednost matrice ρ za $k = 1, ..., p$. Time se, kao i kod nestandardiziranih varijabli, lako može izračunati koliki udio standardizirane ukupne varijance čuva prvih nekoliko glavnih komponenti.
Treba naglasiti da glavne komponente dobivene iz standardiziranih varijabli, odnosno iz matrice korelacije, nisu jednake glavnim komponentama dobivenim iz inicijalnih varijabli, odnosno iz matrice kovarijance. Razlog tomu je da općenito svojstveni parovi (λ_i, v_i) nisu jednaki za kovarijacijsku matricu Σ i korelacijsku matricu ρ slučajnog vektora $\mathbf{X}$. Može se činiti kako se glavne komponente iz standardiziranih varijabli mogu jednostavnim transformacijama dovesti do glavnih komponenti iz inicijalnih varijabli jer su standardizirane varijable nastale jednostavnim trasformacijama inicijalnih varijabli, ali ni to ovdje nije slučaj.
Naime, glavne komponente su invarijantne na ortogonalne transformacije[16] vektora $\mathbf{X}$, a

[16]Ortogonalne transformacije čuvaju duljinu vektora i kut između dva vektora. Primjeri takvih transformacija su rotacija ili zrcaljenje. Ako primijenimo ortogonalnu transformaciju na inicijalni skup podataka, mijenjamo samo njihovu reprezentaciju u prostoru, struktura odnosa između podataka ostaje nepromijenjena pa se time i glavne komponente ne mijenjaju.

iz definicije vektora $\mathbf{Z}$ (1.13) vidljivo je kako on nije ortogonalna transformacija vektora $\mathbf{X}$. Zbog toga vrijedi i da glavne komponente iz standardiziranih varijabli ne daju iste informacije kao i glavne komponente iz inicijalnih varijabli. Oba pristupa imaju prednosti i nedostataka. Na primjer, prednost standardizacije inicijalnih varijabli je što se rezultati analize za različite skupove podataka, mjerenih različitim mjernim jedinicama, mogu direktno usporediti. Kod glavnih komponenti dobivenih iz matrice kovarijance nedostatak je ta osjetljivost komponenti na mjerne jedinice inicijalnih varijabli. Ako npr. postoji velika razlika u varijancama inicijalnih varijabli, one s najvećom varijancom će dominirati u prvih nekoliko glavnih komponenti, što nije mjerodavno ako se mjerne jedinice tih varijabli razlikuju. S druge strane, nedostatak glavnih komponenti dobivenih iz matrice korelacije je pretpostavka da su inicijalne varijable normalno distribuirane jer to ne vrijedi tako često u praksi. Ipak, analiza glavnih komponenti se rijetko koristi u statističkom zaključivanju, odnosno u inferencijalnoj statistici, ona se danas puno više primjenjuje kao deskriptivni alat pa se tom nedostatku ne pridaje velika važnost.

1.2 Uzoračke glavne komponente

U ovom odjeljku opisuje se izračun glavnih komponenti kada je inicijalni skup podataka skup od $n \in \mathbb{N}$ nezavisnih mjerenja odabranih $p \in \mathbb{N}$ varijabli. Neka vektori $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}$ reprezentiraju tih n nezavisnih mjerenja p-dimenzionalne populacije $X = (X_1, X_2, ..., X_p)$, pri čemu su slučajne varijable X_i reprezentanti populacije za $i = 1, .., p$.
Označimo s $\mathbf{X} = (x_{ji})$ matricu dimenzije $n \times p$ u kojoj (j, i)-ti element predstavlja j-to mjerenje i-te varijable, tj. $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}]'$.
Nadalje, neka $\bar{\mathbf{x}}$ označuje uzoračku sredinu[17], a $\mathbf{S}$ uzoračku kovarijacijsku matricu[18]
Sada, za svaki $i \in \{1, ..., p\}$, tražimo p-dimenzionalni realni vektor $\mathbf{a_i} = \left[a_{i1}, a_{i2}, ..., a_{ip}\right]'$ takav da slučajna varijabla $\mathbf{a_i'}X$ ima maksimalnu uzoračku varijancu.
Neka je $i \in \{1, ..., p\}$ proizvoljan, uzoračka sredina slučajne varijable $\mathbf{a_i'}X$ jednaka je

$$\frac{1}{n}\sum_{j=1}^{n} \mathbf{a_i'x_j} = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{p} a_{ik}x_{jk} = \sum_{k=1}^{p} a_{ik}\left(\frac{1}{n}\sum_{j=1}^{n} x_{jk}\right) = \sum_{k=1}^{p} a_{ik}\bar{x_k} = \mathbf{a_i'}\bar{\mathbf{x}}.$$

[17] Za n nezavisnih mjerenja populacije $(X_1, X_2, ..., X_p)$, **uzoračka sredina** $\bar{\mathbf{x}}$ je p-dimenzionalni vektor čije se komponenente definiraju kao $\bar{x_i} = \frac{1}{n}\sum_{j=1}^{n} x_{ji}$, pri čemu x_{ji} je j-to mjerenje i-te varijable X_i.

[18] Za n nezavisnih mjerenja populacije $(X_1, X_2, ..., X_p)$, **uzoračka kovarijacijska matrica S** definira se kao $\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix} = \begin{bmatrix} \frac{1}{n-1}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2 & \cdots & \frac{1}{n-1}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)(x_{jp}-\bar{x}_p) \\ \vdots & \ddots & \vdots \\ \frac{1}{n-1}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)(x_{jp}-\bar{x}_p) & \cdots & \frac{1}{n-1}\sum_{j=1}^{n}(x_{jp}-\bar{x}_p)^2 \end{bmatrix}$. Dijagonalne elementi matrice $\mathbf{S}$ nazivamo **uzoračke varijance**, a elemente van dijagonale **uzoračke kovarijance**.

Time je uzoračka varijanca dana s

$$\begin{aligned}
\frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{a_i'x_j}-\mathbf{a_i'}\bar{\mathbf{x}})^2 &= \frac{1}{n-1}\sum_{j=1}^{n}\left(\sum_{k=1}^{p}a_{ik}x_{jk}-\sum_{k=1}^{p}a_{ik}\bar{x_k}\right)^2\\
&= \frac{1}{n-1}\sum_{j=1}^{n}\left(\sum_{k=1}^{p}a_{ik}(x_{jk}-\bar{x_k})\right)^2\\
&= \frac{1}{n-1}\sum_{j=1}^{n}\left(\sum_{k=1}^{p}a_{ik}^2(x_{jk}-\bar{x_k})^2+2\sum_{1\le k<l\le p}a_{ik}(x_{jk}-\bar{x_k})a_{il}(x_{jl}-\bar{x_l})\right)\\
&= \sum_{k=1}^{p}a_{ik}^2\frac{1}{n-1}\sum_{j=1}^{n}(x_{jk}-\bar{x_k})^2+2\sum_{1\le k<l\le p}a_{ik}a_{il}\frac{1}{n-1}\sum_{j=1}^{n}(x_{jk}-\bar{x_k})(x_{jl}-\bar{x_l})\\
&= \mathbf{a_i'Sa_i},
\end{aligned}$$

gdje je **S** uzoračka kovarijacijska matrica.
Uočavamo da je, kao i u prethodnim odjeljcima, prirodno pretpostaviti da su vektori koeficijenata $\mathbf{a_i}$ jedinične duljine.
Dakle, ako imamo $n \in \mathbb{N}$ mjerenja $(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n})$ populacije X reprezentirane varijablama $X_1, X_2, ..., X_p$, i-ta glavna komponenta $\hat{y}_i$ se definira kao linearna kombinacija $\mathbf{a_i}'X$ koja maksimizira uzoračku varijancu $\mathbf{a_i'Sa_i}$, pri čemu je $\mathbf{a_i}$ jedinični vektor, i za svaki $k < i$ je uzoračka kovarijanca para $(\mathbf{a_k}'X, \mathbf{a_i}'X)$ jednaka nuli.
Općenito, za $k < i$, uzoračka kovarijanca para $(\hat{y_k}, \hat{y_i}) = (\mathbf{a_k'}X, \mathbf{a_i'}X)$ dana je formulom

$$\begin{aligned}
&\frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{a_k'x_j}-\mathbf{a_k'}\bar{\mathbf{x}})(\mathbf{a_i'x_j}-\mathbf{a_i'}\bar{\mathbf{x}})\\
&= \frac{1}{n-1}\sum_{j=1}^{n}\left(\sum_{l=1}^{p}a_{kl}x_{jl}-\sum_{l=1}^{p}a_{kl}\bar{x}_l\right)\left(\sum_{l=1}^{p}a_{il}x_{jl}-\sum_{l=1}^{p}a_{il}\bar{x}_l\right)\\
&= \frac{1}{n-1}\sum_{j=1}^{n}\left(\sum_{l=1}^{p}a_{kl}(x_{jl}-\bar{x}_l)\right)\left(\sum_{l=1}^{p}a_{il}(x_{jl}-\bar{x}_l)\right)\\
&= \frac{1}{n-1}\sum_{j=1}^{n}\left(\sum_{l=1}^{p}a_{kl}a_{il}(x_{jl}-\bar{x}_l)^2+2\sum_{1\le l<m\le p}a_{kl}a_{im}(x_{jl}-\bar{x}_l)(x_{jm}-\bar{x_m})\right)\\
&= \sum_{l=1}^{p}a_{kl}a_{il}\frac{1}{n-1}\sum_{j=1}^{n}(x_{jl}-\bar{x}_l)^2+2\sum_{1\le k<l\le p}a_{kl}a_{im}\frac{1}{n-1}\sum_{j=1}^{n}(x_{jl}-\bar{x}_l)(x_{jm}-\bar{x_m})\\
&= \mathbf{a_k'Sa_i}.
\end{aligned}$$

Korolar 1.2.1. Neka je $\mathbf{S} = (s_{ij})$ uzoračka kovarijacijska matrica dimenzije $p \times p$ za $p \in \mathbb{N}$ i neka su $(\hat{\lambda_1}, \hat{\mathbf{v_1}}), (\hat{\lambda_2}, \hat{\mathbf{v_2}}), ..., (\hat{\lambda_p}, \hat{\mathbf{v_p}})$ parovi svojstvenih vrijednosti $\hat{\lambda_i}$ i svojstvenih vektora $\hat{\mathbf{v}}_\mathbf{i}$ matrice $\mathbf{S}$ uz $\hat{\lambda_1} \geq \hat{\lambda_2} \geq ... \geq \hat{\lambda_p} \geq 0$. Tada je i-ta glavna komponenta dana s

$$\hat{y_i} = \hat{\mathbf{v}_\mathbf{i}}'\mathbf{x} = \hat{v_{i1}}x_1 + \hat{v_{i2}}x_2 + ... + \hat{v_{ip}}x_p, \qquad i = 1, ..., p,$$

gdje je $\mathbf{x}$ proizvoljna opservacija populacije $X = (X_1, X_2, ..., X_p)$.
S tim izborom,

$$\text{uzoračka varijanca } (\hat{y_i}) = \hat{\lambda_i}, \qquad i = 1, ..., p,$$
$$\text{uzoračka kovarijanca } (\hat{y_i}, \hat{y_k}) = 0, \qquad i \neq k.$$

Također,

$$\text{ukupna uzoračka varijanca } = \sum_{i=1}^{p} s_{ii} = \hat{\lambda_1} + \hat{\lambda_2} + ... + \hat{\lambda_p}$$

i vrijedi

$$\text{uzorački koeficijent korelacije} = r_{\hat{y}_i, x_k} = \frac{\hat{v_{ik}}\sqrt{\hat{\lambda_i}}}{\sqrt{s_{kk}}}, \qquad i, k = 1, ..., p.$$

Dokaz. Dokaz slijedi analognim računom kao u dokazima korolara 1.1.1., 1.1.2. i 1.1.3., uz zamjenu varijance, kovarijance, kovarijacijske matrice Σ i koeficijenta korelacije ρ s uzoračkom varijancom, uzoračkom kovarijancom, uzoračkom kovarijacijskom matricom $\mathbf{S}$ i uzoračkim koeficijentom korelacije[19] r, redom. □

1.2.1 Uzoračke glavne komponente dobivene iz standardiziranog uzorka

U slučaju kad su reprezentanti $X_1, X_2, ..., X_p$ populacije X mjereni u različitim mjernim jedinicima, te u dobivenom uzorku postoji značajna razlika u rasponu vrijednosti, često se kao prvi korak u analizi takvog skupa podataka provodi postupak standardizacije.
Neka je standardizirani uzorak reprezentiran matricom $\mathbf{Z}$[20].

[19] Za n mjerenja slučajnih varijabli X i Y **uzorački koeficijent korelacije** ili **Pearsonov koeficijent korelacije** $r_{X,Y}$ definira se kao $r_{x,y} = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$, gdje je s_{xy} uzoračka kovarijanca, a s_{xx} i s_{yy} uzoračke varijance slučajnih varijabli X i Y.

[20] Za uzorak od n nezavisnih mjerenja populacije $(X_1, X_2, ..., X_p)$, **standardizirani uzorak** reprezentiran matricom $\mathbf{Z} = (z_{ji})$ dimenzije $n \times p$ definira se kao $\mathbf{Z} = \begin{bmatrix} \frac{x_{11}-\bar{x_1}}{\sqrt{s_{11}}} & \frac{x_{12}-\bar{x_2}}{\sqrt{s_{22}}} & \dots & \frac{x_{1p}-\bar{x_p}}{\sqrt{s_{pp}}} \\ \frac{x_{21}-\bar{x_1}}{\sqrt{s_{11}}} & \frac{x_{22}-\bar{x_2}}{\sqrt{s_{22}}} & \dots & \frac{x_{2p}-\bar{x_p}}{\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x_1}}{\sqrt{s_{11}}} & \frac{x_{n2}-\bar{x_2}}{\sqrt{s_{22}}} & \dots & \frac{x_{np}-\bar{x_p}}{\sqrt{s_{pp}}} \end{bmatrix}$, pri čemu je x_{ji} j-to mjerenje i-te varijable X_i, $\bar{x_i}$ uzoračka sredina, a s_{ii} uzoračka varijanca varijable X_i. Za svaki z_{ji} vrijedi da je uzoračka sredina jednka nuli i uzoračka varijanca jednaka jedan.

Iz definicije standardiziranog uzorka znamo da je pripadni vektor uzoračke sredine jednak nul-vektoru i da su sve uzoračke varijance jednake jedan. Za definiciju uzoračke kovarijacijske matrice preostaje još izvesti formulu za uzoračke kovarijance.
Neka su $i, j \in \{1, ..., p\}$ proizvoljni. Uzoračka kovarijanca između X_i i X_j, na uzorku reprezentiranim matricom $\mathbf{Z}$, definira se kao

$$\frac{1}{n-1}\sum_{k=1}^{n}\left(\frac{x_{ki}-\bar{x}_i}{\sqrt{s_{ii}}}-0\right)\left(\frac{x_{kj}-\bar{x}_j}{\sqrt{s_{jj}}}-0\right) = \frac{1}{(n-1)\sqrt{s_{ii}}\sqrt{s_{jj}}}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{ki}-\bar{x}_i)$$
$$= \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}},$$

gdje posljednja jednakost slijedi iz definicije uzoračke kovarijacijske matrice $\mathbf{S}$.
Sada vidimo da je uzoračka kovarijacijska matrica standardizirianog uzorka upravo uzoračka korelacijska matrica $\mathbf{R}$[21] inicijalnog uzorka.
Time je i-ta glavna komponenta $\hat{y}_i$ definirana kao linearna kombinacija jediničnog realnog vektora $\mathbf{a_i}$ i slučajnih varijabli $X_1, X_2, ..., X_p$ koja maksimizira uzoračku varijancu $\mathbf{a_i'Ra_i}$ i koja je nekorelirana s ostalih $p-1$ glavnih komponenti, odnosno za koju vrijedi da je $\mathbf{a_k'Ra_i} = 0$ za sve $k < i$.
Posljedica primjene korolora 1.2.1. na standardizirani uzorak $\mathbf{Z}$ dana je sljedećim rezultatom.

Korolar 1.2.2. Neka je $\mathbf{R}$ uzoračka kovarijacijska matrica standardiziranih opservacija $\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_n}$ i neka su $(\hat{\lambda}_1, \hat{\mathbf{v}}_\mathbf{1}), (\hat{\lambda}_2, \hat{\mathbf{v}}_\mathbf{2}), ..., (\hat{\lambda}_p, \hat{\mathbf{v}}_\mathbf{p})$ parovi svojstvenih vrijednosti $\hat{\lambda}_i$ i svojstvenih vektora $\hat{\mathbf{v}}_\mathbf{i}$ matrice $\mathbf{R}$ uz $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p \geq 0$. Tada je i-ta glavna komponenta dana s

$$\hat{y}_i = \hat{\mathbf{v}}_\mathbf{i}'\mathbf{z} = \hat{v_{i1}}z_1 + \hat{v_{i2}}z_2 + ... + \hat{v_{ip}}z_p, \qquad i = 1, ..., p,$$

gdje je $\mathbf{z}$ proizvoljna standardizirana opservacija populacije $X = (X_1, X_2, ..., X_p)$.
S tim izborom,

$$\text{uzoračka varijanca } (\hat{y}_i) = \hat{\lambda}_i, \qquad i = 1, ..., p,$$
$$\text{uzoračka kovarijanca } (\hat{y}_i, \hat{y}_k) = 0, \qquad i \neq k.$$

[21] Za n nezavisnih mjerenja populacije $(X_1, X_2, ..., X_p)$, **uzoračka korelacijska matrica** definira se kao $\mathbf{R} = \begin{bmatrix} 1 & r_{X_1,X_2} & \cdots & r_{X_1,X_p} \\ r_{X_1,X_2} & 1 & \cdots & r_{X_2,X_p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{X_1,X_p} & r_{X_2,X_p} & \cdots & 1 \end{bmatrix}$, pri čemu je r_{X_i,X_j} je uzorački koeficijent korelacije između slučajnih varijabli X_i i X_j za $i, j = 1..., p$.

Također,

$$\text{ukupna uzoračka varijanca } = p = \hat{\lambda}_1 + \hat{\lambda}_2 + ... + \hat{\lambda}_p$$

i vrijedi

$$\text{uzorački koeficijent korelacije} = r_{\hat{y}_i, z_k} = \hat{v_{ik}} \sqrt{\hat{\lambda}_i}, \qquad i, k = 1, ..., p.$$

1.2.2 Ilustrativni primjer

Prikažimo opisani postupak izračuna uzoračkih glavnih komponenti pomoću primjera. U primjeru ćemo analizirati uzoračku korelacijsku matricu pet varijabli koje se odnose na osobno zadovoljstvo članova američkih oružanih snaga. Dana matrica temelji se na podacima prikupljenim iz istraživanja koje obuhvaća 9,147 američkih vojnika.
Primjer je obrađen u [3], a svi rezultati su dobiveni koristeći programski jezik R.
Neka je $\mathbf{R}$ uzoračka matrica korelacije za slučajni vektor $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)$:

$$\mathbf{R} = \begin{bmatrix} 1 & 0.451 & 0.511 & 0.197 & 0.162 \\ 0.451 & 1 & 0.445 & 0.252 & 0.238 \\ 0.511 & 0.445 & 1 & 0.301 & 0.227 \\ 0.197 & 0.252 & 0.301 & 1 & 0.620 \\ 0.162 & 0.238 & 0.227 & 0.620 & 1 \end{bmatrix}.$$

Slučajne varijable X_1, X_2, X_3, X_4, X_5 redom reprezentiraju zadovoljstvo američkog vojnika poslom, obukom, uvjetima rada, medicinskom skrbi te stomatološkom skrbi.
Kako imamo pet inicijalnih varijabli, možemo odrediti pet glavnih komponenti. Parovi svojstvenih vrijednosti i svojstvenih vektora matrice $\mathbf{R}$ dani su tablicom:

Svojstveni vektori ($\hat{v_i}$)					
Varijable zadovoljstva	$\hat{v_1}$	$\hat{v_2}$	$\hat{v_3}$	$\hat{v_4}$	$\hat{v_5}$
poslom	0.443	0.439	0.302	-0.718	0.064
obukom	0.456	0.296	-0.829	0.119	0.059
uvjetima rada	0.480	0.306	0.454	0.658	-0.192
medicinskom skrbi	0.441	-0.533	0.112	0.056	0.711
stomatološkom skrbi	0.412	-0.585	-0.056	-0.186	-0.671
Svojstvene vrijednosti ($\hat{\lambda}_i$)	2.365	1.205	0.574	0.484	0.372

Tablica 1.1: Svojstveni vektori i svojstvene vrijednosti uzoračke matrice korelacije $\mathbf{R}$

Lako se provjeri da je svaki svojstveni vektor jedinične duljine i da su svi oni međusobno ortogonalni.

Iz tablice 1.1. redom čitamo glavne komponente:

$$Y_1 = 0.443X_1 + 0.456X_2 + 0.480X_3 + 0.441X_4 + 0.412X_5$$
$$Y_2 = 0.439X_1 + 0.296X_2 + 0.306X_3 - 0.533X_4 - 0.585X_5$$
$$Y_3 = 0.302X_1 - 0.829X_2 + 0.454X_3 + 0.112X_4 - 0.056X_5$$
$$Y_4 = -0.718X_1 + 0.119X_2 + 0.658X_3 + 0.056X_4 - 0.186X_5$$
$$Y_5 = 0.064X_1 + 0.059X_2 - 0.192X_3 + 0.711X_4 - 0.671X_5$$

Koristeći formulu za uzoračku varijancu glavnih komponenti dobivamo da je uzoračka varijanca prve glavne komponente jednaka:

$$\begin{aligned}\mathrm{Var}(Y_1) &= \hat{v_1}'\mathbf{R}\hat{v_1} \\ &= [1.048, 1.079, 1.136, 1.044, 0.975]'\hat{v_1} \\ &= 2.365 \\ &= \hat{\lambda_1}.\end{aligned}$$

Analognim računom slijedi da su i uzoračke varijance od Y_2, Y_3, Y_4 i Y_5 redom jednake $\hat{\lambda_2}, \hat{\lambda_3}, \hat{\lambda_4}$ i $\hat{\lambda_5}$.
Također, iz formule za uzoračku kovarijancu glavnih komponenti, slijedi da je kovarijanca Y_1 i Y_2 jednaka:

$$\begin{aligned}\mathrm{Cov}(Y_1, Y_2) &= \hat{v_1}'\mathbf{R}\hat{v_2} \\ &= [1.048, 1.079, 1.136, 1.044, 0.975]'\hat{v_2} \\ &= 0,\end{aligned}$$

tj. Y_1 i Y_2 su nekolinearne. Analognim računom slijedi i nekolinearnost između ostalih glavnih komponenti.
Ako promotrimo prvu glavnu komponentu Y_1, vidimo da su koeficijenti uz svaku varijablu X_i približno jednaki, tj. svaka varijabla je otprilike jednako reprezentirana prvom komponentom. Zbog toga, možemo prvu glavnu komponentu interpretirati kao mjeru generalnog zadovoljstva.
Nadalje, uočimo da vrijedi

$$\hat{\lambda_1} + \hat{\lambda_2} + \hat{\lambda_3} + \hat{\lambda_4} + \hat{\lambda_5} = 2.365 + 1.205 + 0.574 + 0.484 + 0.372 = 5 = \sum_{i=1}^{5} \mathrm{Var}(X_i),$$

odnosno, ukupna varijanca je sačuvana u glavnim komponenta. Time udio varijance koju opisuje prva glavna komponenta iznosi

$$\frac{\hat{\lambda_1}}{5} = \frac{2.365}{5} = 0.473 = 47.3\%.$$

Ako promotrimo drugu glavnu komponentu Y_2, vidimo da su koeficijenti uz varijable koje opisuju zadovoljstvo vezano uz posao pozitivni, a koeficijenti uz dvije varijable koje opisuju zadovoljstvo vezano uz zdravstvenu skrb negativni, pa ju možemo interpretirati kao odnos između poslovnog i zdravstvenog zadovoljstva. Udio varijance koju opisuje Y_2 jednaka je 24.1%.
Dakle, prve dvije glavne komponente opisuju 71.4% ukupne varijance.
Ostale glavne komponente Y_3, Y_4 i Y_5 redom opisuju 11.5%, 9.7% i 7.4% ukupne varijance.
Pogledajmo sada uzoračke koeficijente korelacije između inicijalnih varijabli X_i i glavnih komponenata Y_i.
Koristeći formulu za uzorački koeficijent korelacije između inicijalne varijable i glavne komponente, danu korolarom 1.2.2., dobivamo

Uzorački koeficijenti korelacije					
Varijable zadovoljstva	Y_1	Y_2	Y_3	Y_4	Y_5
poslom	0.681	0.482	0.229	-0.5	0.039
obukom	0.702	0.325	-0.628	0.083	0.036
uvjetima rada	0.739	0.336	0.344	0.457	-0.117
medicinskom skrbi	0.679	-0.586	0.085	0.039	0.433
stomatološkom skrbi	0.634	-0.642	-0.043	-0.13	-0.409

Tablica 1.2: Koeficijenti korelacije između inicijalnih varijabli i glavnih komponenti

I iz uzoračkih koeficijenata korelacije vidimo da je svaka inicijalna varijabla približno jednako korelirana s prvom glavnom komponentom, te da su u slučaju druge glavne komponente Y_2, varijable koje opisuju zadovoljstvo vezano uz posao pozitivno korelirane s Y_2, a one koje opisuju zadovoljstvo vezano uz zdravstvenu skrb negativno korelirane s Y_2. Time, možemo reći da bi interpetacija glavnih komponenti pomoću uzoračkih koeficijenta korelacija odgovarala interpretaciji glavnih komponenti pomoću svojstvenih vektora uzoračke matrice korelacije.

1.3 Broj glavnih komponenti

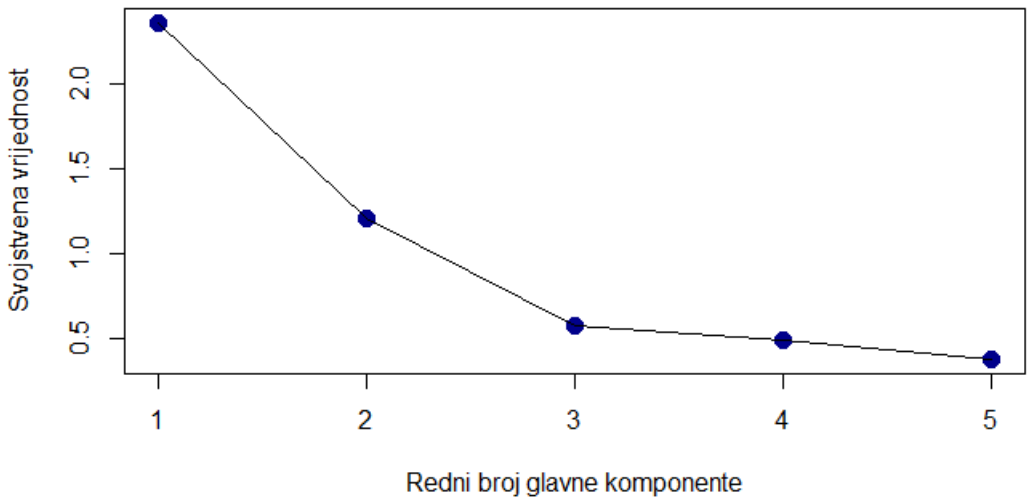
Odabir odgovarajućeg broja glavnih komponenti neizostavan je korak u primjeni analize glavnih komponenti. Iako je u prethodnim odjeljcima opisano kako izračunati svih p glavnih komponenti za promatranu p-dimenzionalnu populaciju, cilj je transformirati inicijalni skup podataka u skup manje dimenzije, bez gubitka bitnih informacija. U tom kontekstu, prirodno se postavlja pitanje koliko glavnih komponenti zadržati.

Pretpostavimo da želimo zadržati barem 80% ukupne varijacije u podacima. Iz prethodnih rezultata, bez obzira radi li se o populacijskim ili uzoračkim (standardiziranim) glavnim komponentama, dobivamo da se broj glavnih komponenti koje zadovoljavaju traženi uvjet određuje pomoću kumulativne sume udjela ukupne varijance koju opisuje svaka glavna komponenta.
Preciznije, tražimo najmanji $m \in \mathbb{N}$, $m < p$, za koji vrijedi

$$\sum_{i=1}^{m} \frac{\mathrm{Var}(Y_i)}{\sum_{j=1}^{p} \mathrm{Var}(X_j)} \geq 0.8,$$

gdje je $\frac{\mathrm{Var}(Y_i)}{\sum_{j=1}^{p} \mathrm{Var}(X_j)}$ udio ukupne varijance opisane pomoći i-te glavne komponente Y_i.
U slučaju kada glavne komponente računamo iz (uzoračke) korelacijske matrice, znamo da je ukupna varijanca standardiziranog skupa podataka jednaka p, odnosno broju varijabli koje reprezentiraju promatranu populaciju. Postoji pravilo, poznato još pod nazivom *Kaiserovo pravilo* (Kaiser, 1960.), koje kaže kako bi se trebale zadržati samo one glavne komponente čije je varijanca veća od jedan. Ideja se krije iza primjera u kojem su sve inicijalne varijable X_i nezavisne. Tada je pripadna korelacijska matrica jednaka jediničnoj matrici, pa su glavne komponente jednake originalnim varijablama s varijancom jedan. Zbog toga se smatra da svaka glavna komponenta koja ima varijancu manju od jedan sadrži manje informacija nego početna varijabla X_i, pa ona nije vrijedna zadržavanja. Ipak, navedeno pravilo nema čvrstu teoretsku podlogu zbog čega bi ga se trebalo primjenjivati s oprezom[22].
Također, za određivanje broja glavnih komponenti koristan je i tzv. graf opadajuće krivulje (eng. *scree plot*). Riječ je o grafu koji prikazuje redni broj glavne komponente u odnosu na njenu varijancu. Kako su glavne komponente poredane po veličini varijance, graf je opadajući. Traži se točka u kojoj se naglo smanjuje varijanca glavnih komponenti i nakon koje je razlika u varijancama glavnih komponenti relativno mala, odnosno traži se tzv. "lakat" grafa. Prva koordinata te točke označuje redni broj posljednje glavne komponente koju treba zadržati.
Prikažimo graf opadajuće krivulje za Primjer 1.1. iz odjeljka 1.1.3.

[22]Navedimo primjer u koje ne bi bilo razborito slijepo koristi *Kaiserovo pravilo*. Neka je koeficijent uz neku inicijalnu varijablu X_i mali u $(p-1)$ glavnih komponenti, a u jednoj glavnoj komponenti, čija je varijanca blizu jedan, ali manja, neka ta varijabla dominira. Očito varijabla X_i pruža informaciju neovisnu od informacija koje pružaju ostale varijable pa ne bi bilo preporučljivo izbrisati glavnu komponentu u kojoj ona dominira. Iz tog razloga postoje razne teorije koju granicu, umjesto jedan, bi bilo najbolje uzeti.



Slika 1.1: Graf opadajuće krivulje za Primjer 1.1.

Na grafu uočavamo izraženi lakat u točki $(3, \lambda_3)$. Kada bi inicijalne varijable X_1, X_2, X_3, X_4 i X_5 zamijenili s prve tri glavne komponente Y_1, Y_2 i Y_3, po izračunu navedenom u primjeru 1.1., opisali bi 82.9% ukupne varijance. Ipak, i u točki $(2, \lambda_2)$ se može uočiti lakat na grafu. Ako bi inicijalni skup podataka reprezentirali s prve dvije glavne komponente Y_1 i Y_2, opisali bi 71.4% ukupne varijance. Po Kaiserovom pravilu odabrali bi samo prve dvije glavne komponente jer je varijance treće glavne komponente jednaka 0.574, što je manje od jedan.
Krajnji odabir broja glavnih komponenti ovisi o namjeni analize skupa podataka, graf opadajuće krivulje i Kaiserovo pravilo samo su alati koji pomažu pri odabiru.

1.4 Geometrijska interpretacija glavnih komponenti

Promotrimo sada glavne komponente iz geometrijske perspektive. U prethodnim odjeljcima vidjeli smo da se glavne komponente definiraju kao linearna kombinacija inicijalnih varijabli $X_1, X_2, ..., X_p$, gdje $X_1, X_2, ..., X_p$ reprezentiraju neki skup podataka, te da su vektori koeficijenata u tim linearnim kombinacija upravo svojstveni vektori pripadne kovarijacijske (ili korelacijske) matrice. Općenito, svojstveni vektori matrice definiraju smjerove koji su ”invarijantni” ili ”stabilni” pri transformaciji s tom matricom, odnosno, to su vektori koji ne mijenjaju smjer prilikom množenja s pripadnom matricom,

samo se skaliraju svojstvenom vrijednosti. Kada promatramo kovarijacijsku (ili korelacijsku) matricu, njihovi svojstveni vektori definiraju smjerove maksimalne varijabilnosti u podacima. Također, što je po apsolutnoj vrijednosti svojstvena vrijednost veća, pripadni svojstveni vektor ukazuje na smjer veće varijabilnosti i time čuva više informacija o danom skupu podataka. Iz tog razloga je svojstveni vektor kojem pripada najveća svojstvena vrijednost vektor koeficijenata za prvu glavnu komponentu, svojstveni vektor s drugom po redu najvećom svojstvenom vrijednosti je vektor koeficijenata druge glavne komponente, itd.
Geometrijski, dani skup podataka inicijalno je reprezentiran u p-dimenzionalnom kooordinatnom sustavu u kojem svaki reprezentant X_i predstavlja jednu koordinatnu os. Glavne komponente su osi novog kooordinatnog sustava iz kojeg se mogu prepoznati odnosi između inicijalnih varijabli koji ranije nisu bili uočljivi. Također, jasnije se vidi koje dimenzije se smiju zanemariti, a da se pri tome ne izgube bitne informacije o danom skupu podataka. Za jednostavniju vizualizaciju novog prostora razapetog glavnim komponentama navodimo sljedeći primjer.

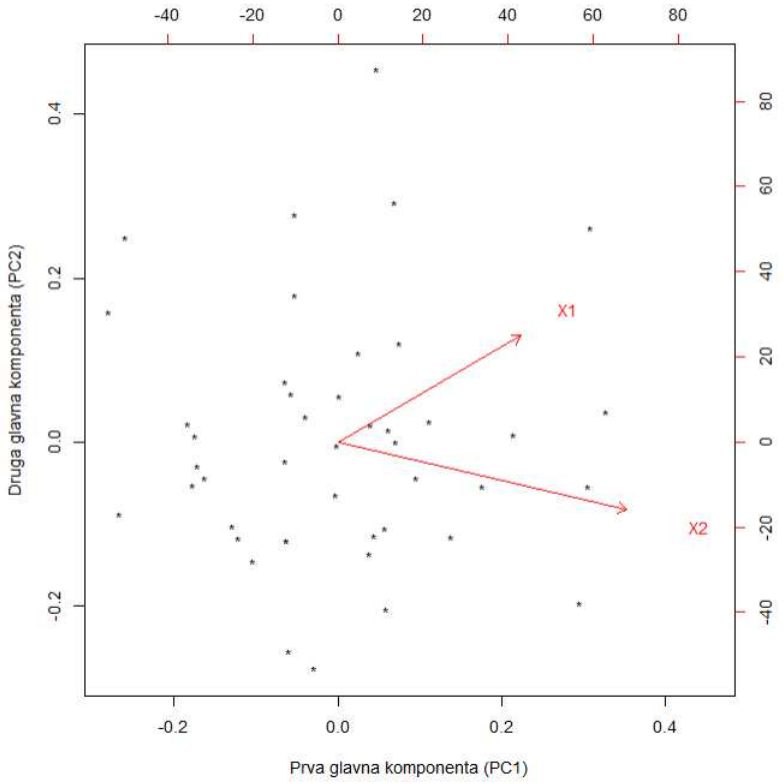
Primjer 1.4.1. Na uzorku od 45 muških Kukastih zmajeva (vrsta ptice grabljivice, lat. *Chondrohierax uncinatus*) u milimetrima je mjerena duljina repa ($\mathbf{X_1}$) i duljina krila ($\mathbf{X_2}$). Podaci su dani u tablici 1.3[23].

	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_1}$	$\mathbf{X_2}$
1.	180	278	185	282	284	277
2.	186	277	195	285	176	281
3.	206	308	183	276	185	287
4.	184	290	202	308	191	295
5.	177	273	177	254	177	267
6.	177	284	177	268	197	310
7.	176	267	170	260	199	299
8.	200	281	186	274	190	273
9.	191	287	177	272	180	278
10.	193	271	178	266	189	280
11.	212	302	192	281	194	290
12.	181	254	204	276	186	287
13.	195	297	191	290	191	286
14.	187	281	178	265	187	288
15.	190	284	177	275	186	275

Tablica 1.3: Izmjerene duljine repa ($\mathbf{X_1}$) i duljine krila ($\mathbf{X_2}$) Kukastih zmajeva

[23] Izvor tablice podataka je [6].

Uočimo iz tablice 1.3 da su duljine repa i krila približno jednake i obje duljine izražene su u milimetrima, zbog toga u ovom primjeru nema značajne potrebe za standardizacijom danog uzorka. Prije nego prikažemo uzorak u prostoru razapetom glavnim komponentama primijetimo da u uzorku postoji jedno mjerenje koje se značajno razlikuje od ostalih. Riječ je o 31. Kukastom zmaju koji ima značajno dulji rep od ostalih jedinki, dok mu je duljina krila približno jednaka duljini krila ostalih jedinki. Kako se takva pojava dogodila u samo jednom mjerenju od njih 45, u svrhu donošenja relevantnih zaključaka, to mjerenje ćemo nadalje zanemariti.
Napomena: Glavne komponente i sljedeći graf dobiveni su pomoću programskog jezika R i u njemu dostupnih funkcija `prcomp()` *i* `biplot()`.



Slika 1.2: Grafički prikaz projekcije uzorka na prostor glavnih komponenti

Na slici 1.2 svaka zvjezdica (”*”) označuje projekciju mjerenja repa i krila jednog muškog Kukastog zmaja na prostor glavnih komponenti. Iz lijeve i donje koordinatne osi očitavaju se vrijednosti glavnih komponenti za svako mjerenje u danom uzorku. Zvjezdice koje su blizu imaju slična obilježja, u ovom slučaju bliske zvjezdice ukazuju na jedinke koje imaju slične duljine repa i krila. Nadalje, vektori daju informacije o

inicijalnim varijablama i njihovom utjecaju na prve dvije glavne komponente. Duljinu i smjer vektora određuju svojstveni vektori, odnosno vektori koeficijanata glavnih komponenti. Na primjer, vektor X_1, koji reprezentira varijablu duljinu repa ($\mathbf{X_1}$), ima početak u ishodištu novog koordinatnog sustava, a kraj u točki čija je prva koordinata prvi element prvog svojstvenog vektora, a druga koordinata prvi element drugog svojstvenog vekora. Dakle, općenito bi za i-tu inicijalnu varijablu uzimali i-te elemente prva dva svojstvena vektora i oni bi označavali koordinate točke do koje bi "povukli" vektor koji reprezentira tu i-tu inicijalnu varijablu. Uočimo da smjer vektora ukazuje na koreliranost inicijalne varijable i glavne komponente. To proizlazi iz činjenice da je koeficijent korelacije između *i*-te glavne komponente i *j*-te inicijalne varijable proporcionalan *j*-tom elementu *i*-tog svojstvenog vektora. U ovom primjeru su i duljina repa i duljina krila pozitivno korelirane s prvom glavnom komponentom. S drugom glavnom komponentom je duljina repa pozitivno korelirana, a duljina krila je negativno korelirana. Koordinate vrhova vektora možemo očitati iz gornje i desne koordinatne osi. Dakle, projekcijom vektora na te koordinatne osi isčitavamo koeficijente prvih dviju glavnih komponenti, odnosno isčitavamo jačinu utjecaja inicijalnih varijabli na prve dvije glavne komponente. Na primjer, vidimo da duljina krila ima veću težinu, odnosno jači utjecaj na prvu glavnu komponentu, nego što ima na drugu. Nadalje, znamo da vrijedi da ako dva vektora zatvaraju mali kut, da je tada kosinus tog kuta blizu jedan. Kako je kosinus kuta između dva vektora proporcionalan njihovom skalarnom produktu, a skalarni produkt jednak je njihovoj kovarijanci, vrijedi da što je kosinus kuta veći da je kovarijanca, a time i korelacija veća. Dakle, vrijedi da vektori koji zatvaraju mali kut ukazuju na varijable koje su korelirane. Ako zatvaraju kut od oko 90°, te varijable najčešće nisu korelirane, a u slučaju da zatvaraju kut blizu 180°, tada bismo rekli da su te varijable negativno korelirane. U ovom primjeru, vidimo da su varijable duljina repa i duljina krila korelirane varijable.

Korolar 1.4.1. Neka je $\mathbf{X}$ p-dimenzionalni slučajni vektor i neka je Σ pripadna kovarijacijska matrica. Tada za familiju p-dimenzionalnih elipsoida

$$\mathbf{X}'\Sigma^{-1}\mathbf{X} = c^2, \qquad c \in \mathbb{R}, \tag{1.14}$$

vrijedi da glavne komponente definiraju osi tih elipsoida[24].

[24]P-dimenzionalni **elipsoid** je skup točaka $\mathbf{x} = (x_1, x_2, ..., x_p)$, u prostoru dimenzije $p \in \mathbb{N}$, koje zadovoljavaju sljedeću jednakost

$$\frac{(x_1 - c_1)^2}{a_1^2} + \frac{(x_2 - c_2)^2}{a_2^2} + ... + \frac{(x_p - c_p)^2}{a_p^2} = 1,$$

pri čemu je $(c_1, c_2, ..., c_p)$ središte elispoida, a $(a_1, a_2, ..., a_p)$ su duljine poluosi elipsoida.

Dokaz. Za $i \in \{1, ..., p\}$, glavne komponente Y_i definirane su kao $Y_i = \mathbf{v_i}'\mathbf{X}$, gdje je $\mathbf{v_i}$ svojstveni vektor matrice Σ. Neka je $\mathbf{Y} = [Y_1, Y_2, ..., Y_p]'$ vektor glavnih komponenti i $\mathbf{V} = [\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_p}]$ matrica svojstvenih vektora. Tada vrijedi $\mathbf{Y} = \mathbf{V}'\mathbf{X}$. Kako je matrica $\mathbf{V}$ ortogonalna, slijedi $\mathbf{X} = \mathbf{VY}$.
Sada se jednakost (1.14) može zapisati kao

$$(\mathbf{VY})'\Sigma^{-1}(\mathbf{VY}) = \mathbf{Y}'\mathbf{V}'\Sigma^{-1}\mathbf{VY} = c^2 \tag{1.15}$$

Kako su svojstveni vektori matrice Σ^{-1} jednaki svojstvenim vektorima matrice Σ, a svojstvene vrijednosti matrice Σ^{-1} su recipročne svojstvene vrijednosti matrice Σ[25], vrijedi

$$\mathbf{V}'\Sigma^{-1}\mathbf{V} = \Lambda^{-1},$$

gdje je $\Lambda^{-1} = [\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, ..., \frac{1}{\lambda_p}]$ vektor recipročnih svojstvenih vrijednosti matrice Σ.
Time je jednakost (1.15) jednaka

$$\mathbf{Y}'\Lambda^{-1}\mathbf{Y} = \sum_{i=1}^{p} \frac{Y_i^2}{\lambda_i} = c^2.$$

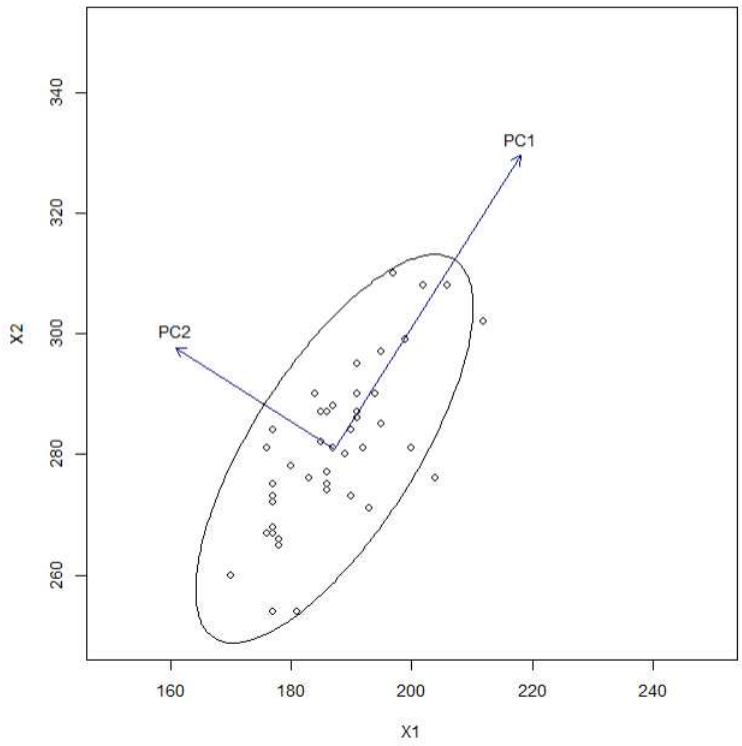
Dakle, elipsoide definirane jednakosti (1.15) možemo promatrati unutar koordinatnog sustava razapetog glavnim komponentama. Tada osi elipsoida leže na pravcima $Y_1, Y_2, ..., Y_p$ orijentiranim u smjeru vektora $\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_p}$, redom, i vrijedi da su duljine poluosi elipsoida redom jednake $c\sqrt{\lambda_1}, c\sqrt{\lambda_2}, ..., c\sqrt{\lambda_p}$. □

Dani rezultat je od posebnog interesa kada je slučajni vektor $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$. Naime, tada su elipsoidi $(\mathbf{X} - \mu)'\Sigma^{-1}(\mathbf{X} - \mu) = c^2, c \in \mathbb{R}$, upravo konture konstantne gustoće p-dimenzionalne normalne distribucije. Pa ako promatramo kooordinatni sustav čije je ishodište središte elipsoida i u kojemu su osi tih elipsoida upravo koordinatne osi, drugim riječima, promatramo sustav razapet glavnim komponentama, interpretacija polaznog skupa podataka je olakšana. Razlog tomu je što su koordinatne osi usklađene s pravcima maksimalne varijabilnosti podataka, time su i uočljivije one glavne komponente koje čuvaju najveću varijabilnost u podacima i one koje bi se u daljnoj analizi mogle zanemariti.
Kod uzoračkih glavnih komponenti, odnosno kada promatramo realizacije slučajnog vektora $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$; μ aproksimiramo s $\bar{\mathbf{x}}$, a Σ s pozitivno definitnom matricom $\mathbf{S}$[26].

[25] Općenito, za ortogonalnu matricu $\mathbf{A}$ i odgovarajući par svojstvene vrijednosti i svojstvenog vektora $(\lambda, \mathbf{v})$, možemo primijetiti da vrijedi: $\mathbf{v} = \mathbf{A^{-1}}\lambda\mathbf{v}$ (zbog ortogonalnosti matrice). Odnosno, $\frac{1}{\lambda}\mathbf{v} = \mathbf{A^{-1}v}$, pa zaključujemo da matrice $\mathbf{A}$ i $\mathbf{A^{-1}}$ imaju jednake svojstvene vektore i recipročne svojstvene vrijednosti.

[26] Za simetričnu matricu $\mathbf{A} \in \mathbf{M_p}(\mathbb{F})$ kažemo da je **pozitivno definitna** ako za svaki $x \in \mathbb{R}^p$, $x \neq 0$, vrijedi $x'\mathbf{A}x > 0$. Odnosno, sve njene svojstvene vrijednosti su pozitivne. Općenito, kovarijacijske matrice višedimenzionalne normalnih slučajnih varijabli su uvijek pozitivno definitne; to svojstvo je nužno kako bi se osiguralo normalno ponašanje distribucije.

Tada elipsoid $(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = c^2$ aproksimira elipsoid $(\mathbf{X} - \mu)'\Sigma^{-1}(\mathbf{X} - \mu) = c^2$, za $c \in \mathbb{R}$. Dakle, novi kooordinatni sustav ima središte $\bar{\mathbf{x}}$ i razapet je uzoračkim glavnim komponentama $\hat{y}_i$, čiji smjerovi odgovaraju smjerovima maksimalne uzoračke varijance. U slučaju kada distribucija uzorka odstupa od normalne razdiobe te graf uzorka odstupa od eliptičnog oblika, i dalje se mogu iz uzoračke kovarijacijske (ili korelacijske) matrice izdvojiti svojstvene vrijednosti i izračunati glavne komponente. Dobivene glavne komponente ponovno definiraju osi elipsoida $(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = c^2$, tj. osi novog koordinatnog sustava. Taj novi koordinatni sustav je dobiven pomicanjem ishodišta inicijalnog koordinatnog sustava u $\bar{\mathbf{x}}$ i rotacijom inicijalnih koordinatnih osi dok ne dođu u smjer maksimalne varijance.
Za uzorak iz primjera 1.4.1., elipsoid i pripadne osi elipsoida prikazane su na sljedećoj slici:



Slika 1.3: Grafički prikaz odnosa inicijalnog koordinatnog sustava i koordinatnog sustava razapetog glavnim komponentama

Na slici 1.3 uočljivo je kako su podaci najviše raspršeni upravo u smjerovima plavih strelica, odnosno u smjerovima glavnih komponenti. Smjer prve glavne komponente određuje smjer najveće raspršenosti u podacima, a smjer druge glavne komponente je

okomit na smjer prve i određuje smjer druge po redu najveće raspršenosti u podacima. Zbog toga, možemo reći da je ovaj novi koordinatni sustav, razapet glavnim komponentama, pogodniji za daljnju analizu uzorka,
Kada je riječ o inicijalnom sustavu dimenzije veće od dva, projekcijama uzorka na dvodimenzionalne koordinatne sustave razapete glavnim komponenta lako se uočava koje glavne komponente definiraju smjerove najveće raspršenosti. Također, uočavaju se i one komponente koje se mogu izostaviti, a da se skup bez velikog gubitka informacija reprezentira manjim brojem varijabli.

Poglavlje 2

Primjena analize glavnih komponenti

U ovom poglavlju opisat ćemo primjenu metode analize glavnih komponenti kroz dva primjera. Svaki od primjera ilustrira snagu metode u rješavanju praktičnih problema i kako ona može biti vrijedan alat u istraživanju i analizi podataka za postizanje relevantih spoznaja. Rezultati su dobiveni pomoću programskog jezika R i u njemu dostupnih paketa. U prvom primjeru kategoriziramo robne marke na temelju simuliranih ocjena potrošača. Primjer je obrađen u [2]. Drugim primjerom ilustriramo primjenu metode na podacima iz stvarnog svijeta. U uzorku su bilježene karakteristike automobila, pri čemu su karakteristike mjerene na modelima automobila iz 2004. godine. Drugi primjer je obrađen u [5].

2.1 Robne marke

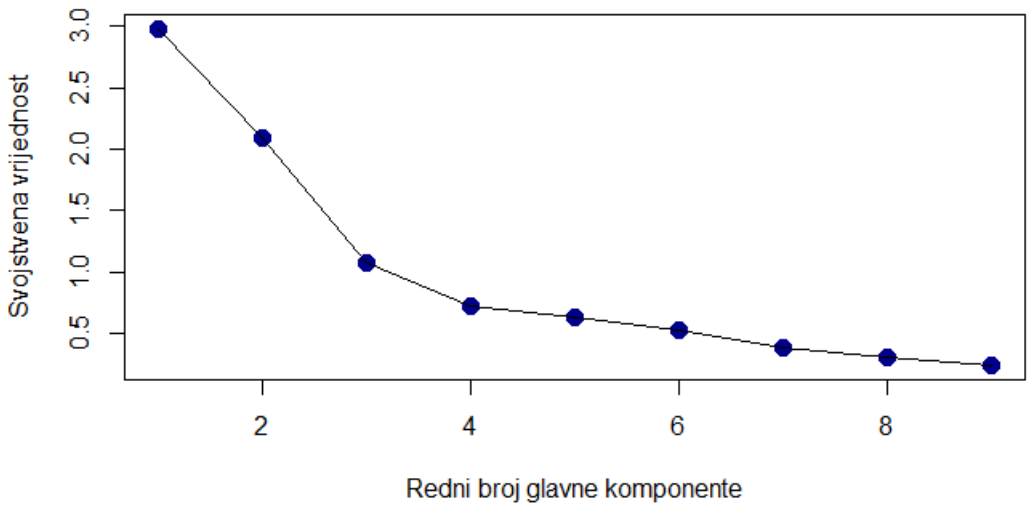
Svaki od 100 ispitanih potrošača je dodijelio ocjenu opažajnim pridjevima vezanim uz 10 vrsta robnih marki. Preciznije, ispitanik je dodjeljivao ocjene od 1 do 10 svakom od 9 opažajnih pridjeva za svaki od 10 robnih marki. Dakle, na ljestvici od 1 do 10, gdje 1 označuje najmanje, a 10 najviše, je ispitanik određivao koliko bi neki pridjev dodijelio određenoj robnoj marki. Na primjer, ako je pridjev ”skup”, a robna marka je ”Coca-Cola” ispitanik ocjenom od 1 do 10 izražava koliko je po njegovom mišljenju skupa marka ”Coca-Cola”. U ovom primjeru nećemo navoditi nazive robnih marki, već ćemo ih označiti slovima od **a** do **j**. Opažajni pridjevi su redom kvalitetan, vodeći na tržištu, najnoviji, zabavan, ozbiljan, povoljan, vrijedan, u trendu i poželjan za ponovnu kupnju. Podaci su dani u matrici tako da prvih devet stupaca odgovara opažajnim pridjevi, a posljednji, deseti stupac, odgovara oznaci robne marke.
Za početak, promotrimo odnose opažajnih pridjeva. Dakle, zanemarujemo posljednji stupac, odnosno oznake robnih markih i pridjeve uzimamo kao reprezentante danog uzorka. Primijetimo, imamo ukupnih 1000 mjerenja (ocjena) za svaki pridjev (svaki od

100 ispitanika je ocijenio svaku od 10 robnih marki). Prvi korak je centriranje uzorka, ono je učinjeno pomoću R funkcije `scale()`, pri čemu je parametar *center* postavljen na vrijednost true, a parametar scale na vrijednost false. Standardizacija, odnosno centriranje i skaliranje, u ovom primjeru nije potrebna jer su sve inicijalne varijable na istoj skali (sve ocjene su od 1 do 10).
Izračunajmo sada uzoračke glavne komponente pomoću R funcije `prcomp()`. Uzoračke glavne komponente i pripadne svojstvene vrijednosti uzoračke korelacijske matrice dane su u sljedećoj tablici:

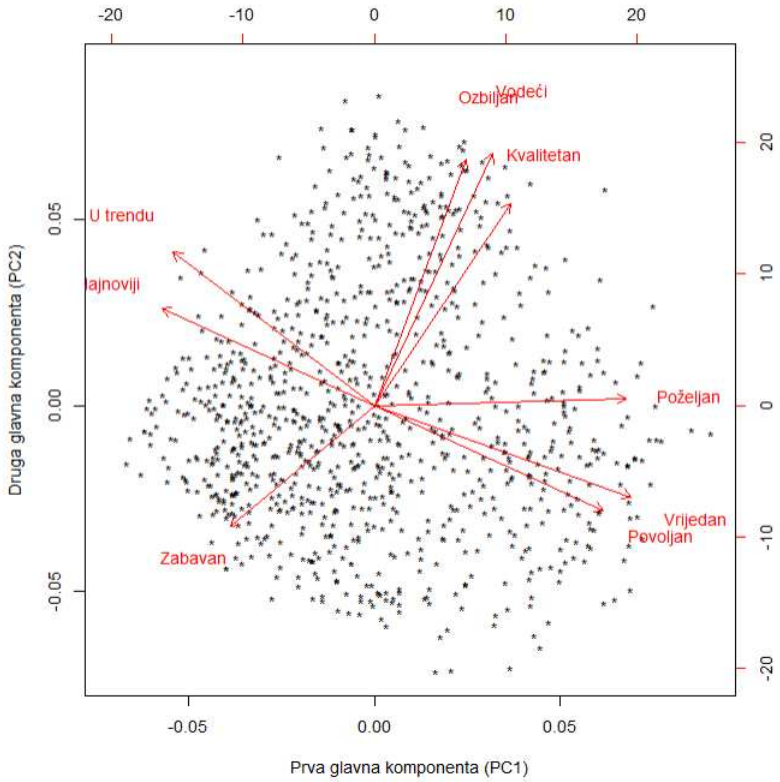
Uzoračke glavne komponente ($\hat{y}_i$)									
Pridjevi	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$	$\hat{y}_5$	$\hat{y}_6$	$\hat{y}_7$	$\hat{y}_8$	$\hat{y}_9$
Kvalitetan	0.237	0.420	0.039	-0.526	0.468	-0.337	0.364	-0.144	0.052
Vodeći	0.206	0.524	-0.095	-0.089	-0.295	-0.297	-0.614	0.288	-0.179
Najnoviji	-0.370	0.201	-0.533	0.214	0.106	-0.174	-0.185	-0.643	0.058
Zabavan	-0.251	-0.250	-0.418	-0.751	-0.331	0.141	-0.007	0.075	0.032
Ozbiljan	0.158	0.510	-0.041	0.099	-0.555	0.392	0.445	-0.184	0.091
Povoljan	0.399	-0.218	-0.490	0.167	-0.013	-0.139	0.288	0.058	-0.647
Vrijedan	0.447	-0.190	-0.369	0.151	-0.063	-0.220	0.017	0.148	0.728
U trendu	-0.351	0.318	-0.371	0.168	0.366	0.266	0.154	0.615	0.059
Poželjan za ponovnu kupnju	0.439	0.015	-0.125	-0.130	0.356	0.675	-0.389	-0.202	-0.017
Svojstvene vrijednosti ($\hat{\lambda}_i$)	2.979	2.097	1.079	0.727	0.638	0.535	0.390	0.312	0.243

Tablica 2.1: Uzoračke glavne komponente i pripadne svojstvene vrijednosti

Iz tablice vidimo da su prve tri svojstvene veće od jedan, pa bi po *Kaiserovom pravilu* bile dovoljne prve tri glavne komponente za reprezentaciju danog skupa podataka.
Nadalje, iz grafa opadajuće krivulje, prikazanom na slici 2.1., uočavamo dva tzv. ”lakta”. Prvi se postiže u trećoj glavnoj komponenti, a drugi u četvrtoj. Kako se traži točka nakon koje je razlika u svojstvenim vrijednostima relativno mala, pravilan izbor može biti uzimanje i tri i četiri glavne komponente. Krajnji odabir broja glavnih komponenti ovisi o željenom cilju udjela varijance podataka koje one opisuju. Udio ukupne varijance koju opisuje prva glavna komponenta je 33.1%. Druga, treća i četvrta redom opisuju 23.3%, 12% i 8.1% varijance. Ako npr. želimo zadržati barem 80% ukupne varijance, uzeli bi prve četiri glavne komponente.
Za lakšu interpretaciju glavnih komponenti i tumačenje odnosa između pridjeva projicirajmo uzorak na prostor prve dvije glavne komponente. Graf projekcije dobiven je u R-u pomoću funkcije *biplot()* i prikazan je na slici 2.2.



Slika 2.1: Graf opadajuće krivulje



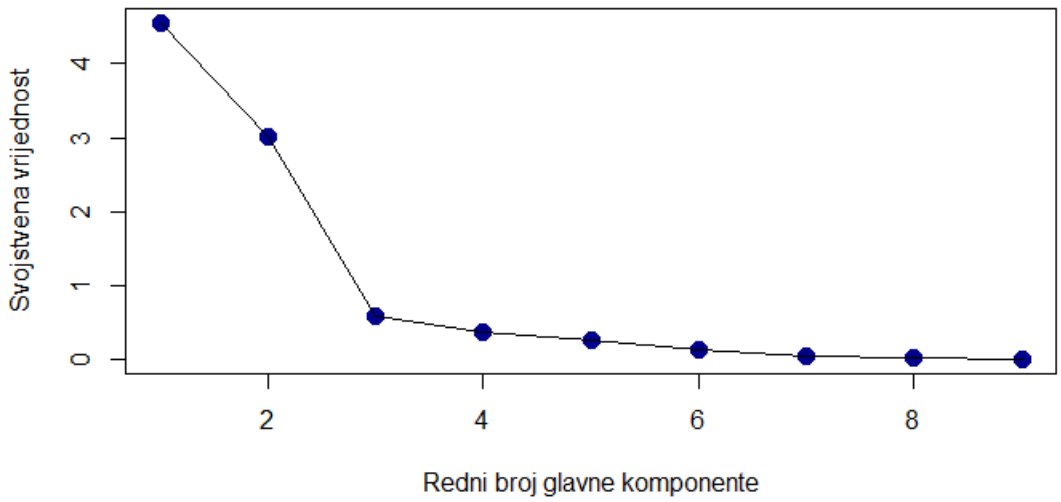
Slika 2.2: Grafički prikaz projekcije uzorka na prostor prvih dviju glavnih komponenti

Iz slike 2.2. uočavamo da se opažajni pridjevi po koreliranosti dijele u četiri kategorije. Vektori ”U trendu” i ”Najnoviji” zatvaraju mali kut, pa možemo reći da su ti pridjevi korelirani i čine jednu kategoriju, nazovimo ju suvremenost. Dakle, ispitanici su uglavnom slično ocjenjivali ta dva pridjeva. Nadalje, pridjeve ”Ozbiljan”, ”Vodeći” i ”Kvalitetan” čine drugu kategoriju, kategoriju vodstvo. Pridjev ”Zabavan” čini zasebnu kategoriju zabava, jer vidimo da on nije izrazito koreliran s niti jednim preostalim pridjevom. Na kraju, posljednju kategoriju čine ”Poželjan za ponovnu kupnju”, ”Vrijedan” i ”Povoljan”, nazovimo tu kategoriju vrijednost. Uočimo nadalje da kategorije vodstvo i zabava, te suvremenost i vrijednost zatvaraju kutove iznosa blizu 180°. Dakle, možemo reći da su te kategorije negativno korelirane. Taj negativni utjecaj jedne kategorije na drugu upravo reprezentiraju prve dvije glavne komponente. Naime, vidimo da na prvu glavnu komponentu najveći utjecaj imaju kategorije suvremenost i vrijednost. Vrijednost je pozitivno korelirana s prvom glavnom komponentom, a suvremenost negativno. Time prvu glavnu komponentu možemo interpretirati kao odnos između pridjeva vezanih za suvremenost i pridjeva vezanih za vrijednost neke robne marke. Na drugu glavnu koponentu najveći utjecaj imaju kategorije vodstvo i zabava. Kako je ponovno jedna kategorija pozitivno korelirana s drugom glavnom komponentom, a druga negativno, i za nju možemo reći da opisuje odnos između pridjeva ”Zabavan” i pridjeva vezanih uz vodstvo robne marke na tržištu.
Sada, nakon što smo kategorizirali opažajne pridjeve, pogledajmo koje kategorije najviše opisuju pojedinu robnu marku. Za svaku robnu marku, svakom pridjevu pridružit ćemo ocjenu jednaku srednjoj ocjeni 100 ispitanika. Dakle, svaki pridjev imat će jednu ocjenu za svaku različitu robnu marku. Nakon pripreme podataka, ponovno pozivom R funkcije `prcomp()` dobivamo glavne komponetne. U sljedećoj tablici navodimo svojstvene vrijednosti uzoračke korelacijske matrice:

Svojstvene vrijednosti ($\hat{\lambda}_i$)								
4.556	3.010	0.591	0.378	0.260	0.134	0.046	0.021	0.002

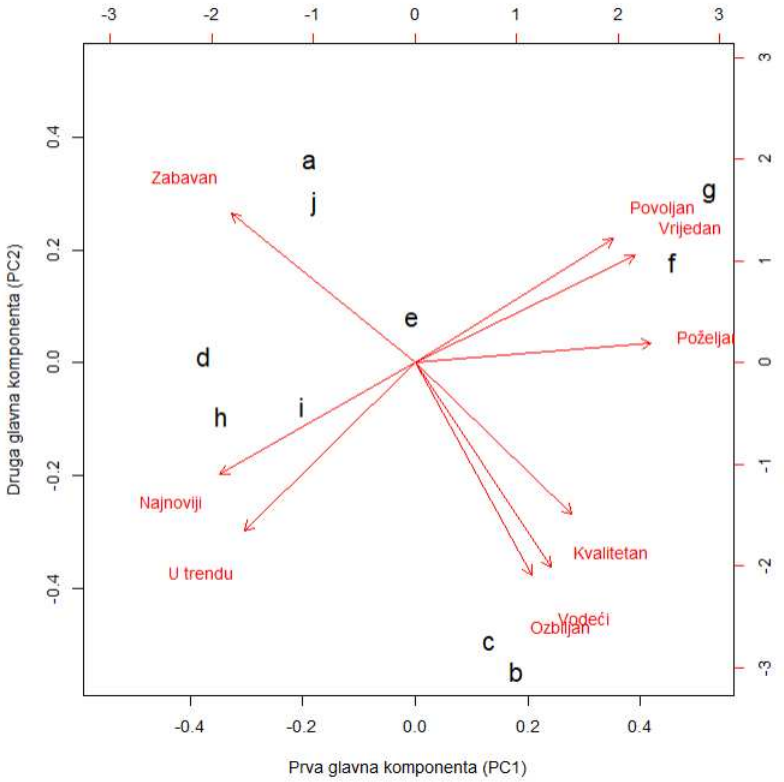
Tablica 2.2: Svojstvene vrijednosti uzoračke matrice korelacije

Iz tablice 2.2. vidljivo je kako su prve dvije svojstvene vrijednosti dominantne. Ako pogledamo graf opadajuće krivulje:
Vidimo da je na grafu opadajuće krivulje, prikazan slikom 2.3., izražajan je tzv. ”lakat” u trećoj glavnoj komponenti. Dakle, ako bismo zaključak donosili samo na temelju grafa opadajuće krivulje, podatke bi reprezentirali s prve tri glavne komponente. S druge strane, kada bi se vodili *Kaiserovim pravilom* prve dvije glavne komponente bi bile dovoljne za interpretaciju danog skupa podataka. Izbor ponovno ovisi o udjelu varijance kojeg želimo opisati s glavnim komponentama. U ovom slučaju bi prva glavna komponente opisivala 50.5%, druga 33.4%, a treća 6.6%. Dakle, ako nam je dovoljno da glavne komponente



Slika 2.3: Graf opadajuće krivulje

opisuju barem 80% ukupne varijance, dovoljno bi bilo uzeti prve dvije glavne komponente. Prikažimo sada projekciju uzorka na prostor razapet prvim dvjema glavnim komponentama.
Iz projekcije, pikazane slikom 2.4., vidimo da pridjeve možemo ponovno kategorizirati u navedene četiri kategorije. Dakle, uzimanjem srednjih ocjena za svaki pojedini pridjev nije imalo veliki utjecaj na korelaciju između njih. Jedino što je različito u odnosu na sliku 2.2. je da su vektori ”Zabavan” te ”Kvalitetan”, ”Ozbiljan” i ”Vodeći” zamijenili smjerove, ali tu je samo riječ o predzaku vektora koeficijenata druge glavne komponente. Promotrimo nadalje položaje robnih marki na grafu. Vidimo da su npr. robne marke **f** i **g** visoko pozicinirane kada je u pitanju kategorija vrijednost. S druge strane, robne marke **d**, **h** i **i** prednjače kada je u pitanju suvremenost, **a** i **j** kada je u pitanju zabava, a **c** i **b** kada je riječ o vodstvu robnih marki. Uočimo da se robna marka **e** nalazi oko središta, odnosno oko izvora vektora pridjeva. Dakle, robna marka **e** nije diferencirana niti u jednom smjeru vektora. To za robnu marku može biti dobro, a i loše, ovisi o preferencijama i strateškim ciljevima vlasnika. Ako npr. vlasnik robne marke **e** želi imati sigurnu marku, marku koja se sviđa mnogim potrošačima, njena sadašnja pozicija bila bi poželjna. S druge strane, ako vlasnik želi da mu je marka jaka u nekoj kategoriji pridjeva, ovaj položaj je tada nepoželjan. Recimo da želimo pomaknuti robnu marku **e** u smjeru neke kategorije. Tada možemo gledati odnose srednjih ocjena robne marke **e** i marke koja je jaka u toj kategoriji. Ako na primjer želimo da je marka **e** na glasu zabavne robne marke na tržištu,



Slika 2.4: Grafički prikaz projekcije uzorka na prostor prvih dviju glavnih komponenti

mogli bismo usporediti srednje ocjene pridjeva marki **a** i **j** sa srednjom ocjenom pridjeva marke **e** i na temelju usporedbe donijeti buduće strateške odluke. Ako nam je cilj samo maknuti robnu marku **e** iz grupe marki "koju svi vole i koja se ne ističe", dobra strategija bila bi tražiti mjesto na mapi koje već nije popunjwno nekom drugom robnom markom. Na primjer, vidimo da između marki **c** i **b** te **f** i **g** postoji puno praznog prostora, dakle ne postoji robna marka koja u bliskoj mjeri obuhvaća kategorije i vodstvo i vrijednost. Time bismo mogli uzeti srednju vrijednost ocjena pridjeva za marke **b**,**c**, **f** i **g**, i od nje oduzeti srednju vrijednost ocjena pridjeva za marku **e**. Rezultati su dani u sljedećoj tablici:

Razlika srednjih vrijednosti								
Kvalitetan	Vodeći	Najnoviji	Zabavan	Ozbiljan	Povoljan	Vrijedan	U trendu	Poželjan
1.175	0.391	-0.937	-0.934	0.573	-0.250	0.079	-0.470	0.669

Tablica 2.3: Razlika srednjih vrijednosti između robnih marki **b**, **c**, **f** i **g** te robne marke **e**

Dobiveni brojevi u tablici 2.3. sugeriraju da ako želimo da se robna marka pomakne iz središte u kojem je sad, da bi bilo dobro da povećamo naglasak na izvedbu kvalitete robne marke, a da možda smanjimo naglasak na izvedbe ”najnoviji” i ”zabavan”.
Dakle, ako općenito želimo uspoređivati više robnih marki u visokodimenzionalnom prostoru, vidimo da nam od velike pomoći može biti analiza glavnih komponenti. Dani uzorak uspjeli smo interpretirati pomoću dvije glavne komponente pri čemu smo obuhvatili preko 80% varijance u podacima. Ipak, ovo je samo dio analize koja se u praksi provodi nad nekim skupom podataka. Treba biti oprezan pri donošenju zaključaka jer mogli smo, na primjer, umjesto srednje vrijednosti ocjena pridjeva koristiti medijan ocjena pridjeva. Dobra praksa bila bi agregirati uzorak na više načina i usporediti rezultate. Nadalje, treba imati na umu da su odnosi pridjeva strogo povezani s robnim markama koje promatramo. Da smo npr. imali druge robne marke možda pridjeve ne bi mogli kategorizirati u četiri kategorije, kao što smo u ovom uzorku. Isto tako, možda jedna robna marka ima veliki utjecaj na kotegoriziranje pridjeva, pa bi možda dodavanje neke nove robne marke ili izbacivanje jedne od mjerenih promijenilo odnose pridjeva (tj. njihov položaj na grafu projekcije uzorka na prostor glavnih komponenti). Također, ako promatramo projekciju uzorka na prostor glavnih komponenti ne bi bilo ispravno donositi zaključke o jačini robnih marki u smjeru jednog opažajnog pridjeva. Na primjer, iz slike 2.4. bi se moglo reći da su marke **b** i **c** slabije u smjeru pridjeva ”najnoviji”, nego što su to marke **d**, **h** i **i**. Ali zapravo, vrijedi da **b** ima veću prosječnu ocjenu za pridjev ”najnoviji” nego bilo koja druga marka. Općenito, marke **b** i **c** su slične u odnosu na glavne komponente koje u obzir uzimaju sve pridjeve, ali nisu nužno sliče po pojedinačnim pridjevima. Dakle, kada koristimo analizu glavnih komponenti i fokusiramo se na prvih nekoliko dimenzija, gledamo zapravo najznačajnije sličnosti između varijabli koje su istaknute u tim dimenzijama, a manje značajne razlike koje se ne izražavaju u tih prvih nekoliko dimenzija se mogu prikriti ili teško zamijetiti.

2.2 Automobili

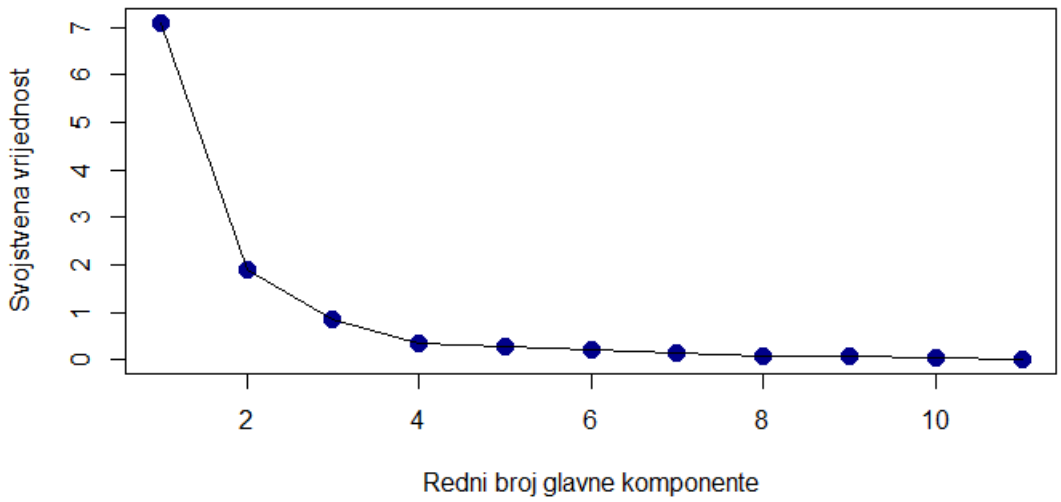
U ovom primjeru promatramo mjerenja osamnaest karakteristika 388 modela automobila iz 2004. godine. Od tih osamnaest karakteristika, njih sedam poprima vrijednost nula ili jedan. Riječ je o tipu modela, odnosno je li automobil sportski, SUV, karavan, jednovolumen ili pick-up te koristi li pogon na sva četiri kotača ili na samo dva. Preostalih jedanaest karakteristika su redom cijena za privatnog kupca, cijena za poslovnog kupca, obujam motora, broj cilindara u motoru, snaga motora, potrošnja u gradu, potrošnja na autoputu, težina vozila, međuosovinski razmak, ukupna duljina vozila i ukupna širina vozila. U metodi analize glavnih komponenti koriste se varijable numeričkog tipa, a kako je prvih sedam karatkeristika logičkog tipa (istina ili laž predstavljeni jedinicom ili nulom, respektivno), te karakterstike izostavljamo. Primijetimo, karakteristike koje ćemo

analizirati mjerene su u različitim mjernim jedinicama. Na primjer, obujam motora izražen je u litrama, težina vozila u kilogramima, duljina i širina u centimetrima, potrošnje su mjerene u miljama po galonu, odnosno koliko milja automobil može prijeći koristeći jedan galon goriva, itd. Iz tog razloga prije izračuna glavnih komponenti standardiziramo podatke pomoću R funkcije `scale()`.
U tablici 2.4 navedimo prvih pet uzoračkih glavnih komponenti i prvih pet pripadnih svojstvenih vrijednosti dobivenih pomoću R funkcije `prcomp()`.

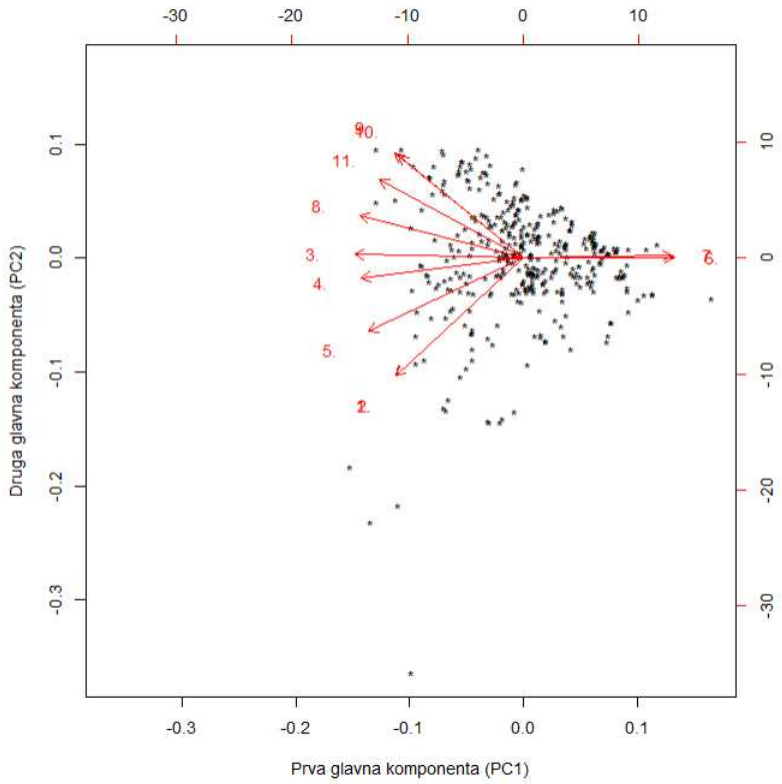
Uzoračke glavne komponente ($\hat{y}_i$)					
Karakteristike	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$	$\hat{y}_5$
1. Cijena za privatnog kupca	-0.264	-0.469	-0.255	0.280	-0.050
2. Cijena za poslovnog kupca	-0.262	-0.470	-0.257	0.288	-0.037
3. Obujam motora	-0.347	0.015	-0.047	-0.525	-0.052
4. Broj cilindara	-0.334	-0.078	-0.081	-0.640	0.126
5. Snaga motora	-0.319	-0.292	-0.076	-0.058	0.120
6. Potrošnja u gradu	0.310	0.003	-0.535	-0.186	-0.326
7. Potrošnja na autoputu	0.307	0.011	-0.599	-0.126	-0.040
8. Težina vozila	-0.336	0.167	0.112	0.120	-0.397
9. Međuosovinski razmak	-0.266	0.418	-0.264	0.221	0.225
10. Duljina vozila	-0.257	0.408	-0.345	0.168	0.456
11. Širina vozila	-0.296	0.313	-0.088	0.091	-0.663
Svojstvene vrijednosti ($\hat{\lambda}_i$)	7.105	1.884	0.850	0.357	0.275

Tablica 2.4: Prvih pet uzoračkih glavnih komponenti i pripadnih pet svojstvenih vrijednosti

Po *Kaiserovom pravilu*, uzeli bi prve dvije uzoračke glavne komponente, jer vidimo iz tablice 2.4. da je treća svojstvena vrijednost manja od jedan. Ipak, treća svojstvena vrijednost iznosi 0.85, što je relativno blizu jedan. Postoje teorije i da je 0.8 dovoljna veličina svojstvene vrijednosti da se pripadna glavna komponenta uzme u obzir.
Iz grafa opadajuće krivulje, prikazanog na slici 2.5, vidimo da se tzv. "lakat" postiže u drugoj i trećoj glavnoj komponenti. Dakle, ako bi zaključak o odabiru dovoljnog broja glavnih komponenti donosili na temelju grafa, polazni skup podataka reprezentirali bi s dvije ili tri glavne komponente. Izbor ovisi o udjelu varijance koje želimo opisati s glavnim komponentama. Prva glavna komponenta opisuje 64.6% ukupne varijance, druga 17.1%, a treća 7.7%. Ako bismo htjeli opisati barem 80% ukupne varijance, vidimo da bi prve dvije glavne komponenti bile dovoljne jer bismo tada u kumulativnoj sumi udjela varijanci imali iznos veći od 80%.
Projicirajmo sada uzorak na prostor razapet prvim dvjema glavnim komponentama.



Slika 2.5: Graf opadajuće krivulje



Slika 2.6: Grafički prikaz projekcije uzorka na prostor prvih dviju glavnih komponenti

Za lakšu čitljivost grafa projekcije, prikazanog na slici 2.6., vektore koji reprezentiraju karakteristike automobila označili smo rednim brojevima po uzoru na tablicu 2.4. Vidimo da vektori koji reprezentiraju varijable potrošnje u gradu i na autoputu, su međusobno jako blizu, skoro da se poklapaju, ali sa svim ostalim vektorima zatvaraju kutove veće od 90°. Dakle, uočavamo negativnu koreliranost potrošnje automobila s ostalim karakteristikama automobila. Ako pogledamo projekciju vektora na prvu glavnu komponentu, jedino su karakteristike potrošnje pozitivno korelirane s prvom glavnom komponentom, ostale karakteristike su sve negativno korelirane s njom. Najveći negativni utjecaj na prvu glavnu komponentu imaju karakteristike obujam motora, težina vozila, broj cilindara u motoru i snaga motora. Dakle, prva glavna komponenta nam govori je li riječ o automobilu koji ima snažan motor, koji je velik i troši dosta benzina ili je riječ o manjem, slabijem i štedljivijem automobilu. Ako pogledamo projekcije na drugu glavnu komponentu, vidimo da potrošnje u gradu i na autocesti nemaju skoro nikakav utjecaj na nju. Druga glavna komponenta opisuje kontrast između veličine automobila te cijene automobila i snage motora. Karakteristike koje opisuju veličinu automobila su pozitivno korelirane s drugom glavnom komponentom, a cijene, snaga motora i broj cilindara u motoru su negativno korelirane s drugom glavnom komponentom. Time, možemo reći da ona odvaja sportske automobile, oni su mali, skupi i imaju jake motore, od automobila poput terenska vozila ili jednovolumeni, koji nisu toliko skupi, veći su te nemaju tako jake motore.

Bibliografija

[1] Lj. Arambašić, *Linearna algebra*, Element, 2022.

[2] Chris Chapman i Elea McDonnell Feit, *R for marketing research and analytics*, sv. 67, Springer, 2015.

[3] George H Dunteman, *Principal components analysis*, sv. 69, Sage, 1989.

[4] Miljenko Huzak, *Vjerojatnost i matematicka statistika*, Predavanja (skripta), dostupno na http://aktuari. math. pmf. unizg. hr/docs/vms. pdf (2006).

[5] Cosma Rohilla Shalizi, *Advanced Data Analysis from an Elementary Point of View*, dostupno na https://www.stat.cmu.edu/ cshalizi/ADAfaEPoV/ADAfaEPoV.pdf (2021).

[6] Richard Johnson Dean Wichern, *Pearson New International Edition*.

Sažetak

U ovome radu predstavljena je analiza glavnih komponenti (engl. *Principal component analysis*, PCA) kao vrijedan statistički alat u deskriptivnoj analizi te jasnijem razumijevanju velikih, višedimenzionalnih skupova podataka. Na samom početku definirane su glavne komponente i načini na koje se one izračunavaju, zajedno s pretpostavkama koje trebaju biti ispunjene kako bi sama analiza bila valjana. Pokazalo se kako se željenu glavnu komponentu može definirati kao linearnu kombinaciju inicijalnih slučajnih varijabli i odgovarajućeg svojstvenog vektora kovarijacijske matrice početnog skupa podataka. Ključna prednost analize glavnih komponenti smanjenje je dimenzionalnosti podataka što se postiže zadržavanjem onih glavnih komponenti koje čine najveći udio u ukupnoj varijanci podataka, a samim time sadrže i većinu ključnih informacija o početnom skupu podataka. Ilustrirana je primjena uzoračke analize glavnih komponenti te njihova geometrijska interpretacija. Na samom kraju rada kroz dva primjera i uz pomoć programskog jezika R detaljno su opisani primjena na simuliranim i stvarnim skupovima podataka te donošenje zaključaka na temelju analize glavnih komponenti.

Summary

This thesis presents *Principal component analysis*, also known as PCA, as an invaluable statistical method used for descriptive analysis and better interpretation of large, multidimensional datasets. At the beginning, principal components are defined and all the ways of calculating principal components are given, backed with mathematical results and conditions that need to be fullfilled in order for the whole analysis to be reliable. It has been shown that a principal component can be defined as a linear combination of the initial random variables and the right eigenvector of the covariance matrix of the original dataset. Furthermoer, the main advantage of PCA is considered to be reducing the dimensionality of the initial data, which is achieved by focusing on those principal components that make up to the biggest part in the total data variance. Exactly those principal components contain the majority of key information about the initial dataset. Moreover, application of sample PCA and the geometrical interpretation of the method are also illustrated. The ending of this thesis brings the reader to two examples where programming language R is used for a very detailed description of PCA application on simulated and real-life datasets, as well as for illustration on how to inference based on PCA can help with real business problems.

Životopis

Ime mi je Dolores Strmečki, a rođena sam 11. rujna 1998. godine u Karlovcu. Osnovno obrazovanje stekla sam u Osnovnoj školi Vladimira Nazora u Topuskom, nakon čega sam nastavila srednjoškolsko obrazovanje u općoj gimnaziji Srednje škole Glina.
Godine 2017. upisala sam sveučilišni prijediplomski studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu, a završila sam ga 2021. godine. Iste godine, upisala sam sveučilišni diplomski studij Financijska i poslovna matematika na istom fakultetu.