Analiza glavnih komponenti i faktorska analiza

Vilić, Ana

Sveučilište J.J.Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike

Ana Vilić

Analiza glavnih komponenti i faktorska analiza

Diplomski rad

Osijek, 2020.

Sveučilište J.J.Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike

Ana Vilić

Analiza glavnih komponenti i faktorska analiza

Diplomski rad

Voditelj: izv. prof. dr. sc. Nenad Šuvak

Osijek, 2020.

Sadržaj

Uvod

U ovom ćemo se radu baviti statističkom metodom zvanom analiza glavnih komponenti (engl. principal component analysis ili kraće PCA) i faktorskom analizom. Povijesno gledano, metodu glavnih komponenti osmislio je engleski matematičar K. Pearson 1901. godine, a neovisno o njemu razvio ju je i američki matematičar H. Hotelling 1933. godine. Analiza glavnih komponenti uobičajeno se provodi sa svrhom smanjenja dimenzionalnosti slučajnog uzorka na temelju kojeg zaključujemo. Ilustrativno, pretpostavimo da imamo p slučajnih varijabli, tada glavne komponente definiramo kao linearne kombinacije slučajnih varijabli koje imaju određena svojstva o kojima će biti riječi kasnije. Ideja analize glavnih komponenti je originalne varijable zamijeniti s prvih nekoliko glavnih komponenti uz minimalan gubitak informacija i time smanjiti dimenzionalnost prostora s p na m, $m < p$.

U radu ćemo najprije pokazati kako dolazimo do glavnih komponenti, a zatim ćemo istražiti neka njihova algebarska svojstva. Nakon toga ćemo pokazati neka statistička svojstva vezana za analizu glavnih komponenti. Osim toga, dotaknut ćemo se i uporabe glavnih komponenti u linearnoj regresiji. Glavne komponente se koriste i u sklopu drugih multivarijatnih tehnika kao što su diskriminantna analiza i klaster analiza, no o tome neće biti riječi u ovom radu.

Često se analiza glavnih komponenti smatra dijelom faktorske analize, ali te su dvije metode u suštini različite. Osnovna ideja faktorske analize je da se p opaženih slučajnih varijabli mogu iskazati kao linearne funkcije nekakvih teoretskih slučajnih varijabli (faktora) uz grešku. Stoga je glavna razlika između ovih dviju metoda što se u faktorskoj analizi pretpostavlja postojanje modela. Faktorsku analizu, kakvu je danas poznajemo, osmislio je engleski psiholog C. E. Spearman 1904. godine. U radu ćemo opisati faktorski model te usporediti ovu metodu s analizom glavnih komponenti.

1 Osnovni pojmovi

U ovom ćemo poglavlju navesti neke osnovne pojmove i rezultate iz teorije vjerojatnosti i linearne algebre potrebne za razumijevanje daljnjeg teksta. Najprije ćemo definirati slučajni vektor.

Definicija 1.1. *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}^p$ za koju je $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$, $\forall B \in \mathcal{B}(\mathbb{R}^p)$ je p-dimenzionalan slučajni vektor na $(\Omega, \mathcal{F}, P)$.*

Podsjetimo se i sljedećeg rezultata vezanog za slučajne vektore:

Propozicija 1.1. *Neka je $X : \Omega \to \mathbb{R}^p$, $X = (X_1, X_2, \ldots, X_p)$. Tada je X slučajni vektor ako i samo ako je X_k slučajna varijabla za svaki $k = 1, 2, \ldots p$.*

Dokaz se može pronaći u [21].

Iz Propozicije 1.1. slijedi da je p-dimenzionalan slučajni vektor uređena p-torka slučajnih varijabli.

U radu će od velike važnosti biti matrica kovarijanci i matrica korelacija slučajnog vektora te koreliranost slučajnih varijabli. Stoga ćemo navesti njihove definicije i neke osnovne rezultate vezane uz te pojmove. Najprije uvodimo pojam momenta dvodimenzionalnog slučajnog vektora:

Definicija 1.2. *Neka je $X = (X_1, X_2)$ dvodimenzionalan slučajni vektor. Očekivanje*

$$E[X_1^k X_2^l], \quad k, l \in \mathbb{N}_0$$

slučajne varijable $X_1^k X_2^l$ (ako postoji) zovemo ishodišni moment reda (k, l) slučajnog vektora (X_1, X_2). Očekivanje $E[(X_1 - E[X_1])^k (X_2 - E[X_2])^l]$ (ako postoji) zovemo centralni moment reda (k, l) slučajnog vektora (X_1, X_2). Specijalno, centralni moment reda (1,1) dvodimenzionalnog slučajnog vektora $X = (X_1, X_2)$ tj. $E[(X_1 - E[X_1])(X_2 - E[X_2])]$ zovemo korelacijski moment ili kovarijanca i označavamo ga s $Cov(X_1, X_2)$.

Teorem 1.1. *Neka je $X = (X_1, X_2)$ dvodimenzionalan slučajni vektor za koji postoji EX_1 i EX_2. Ako su slučajne varijable X_1 i X_2 nezavisne, onda je $Cov(X_1, X_2) = 0$.*

Može se pokazati da obrat ovog teorema općenito ne vrijedi. Dokaz teorema se može pronaći u [4].

Definicija 1.3. *Kažemo da su slučajne varijable X_1 i X_2 nekorelirane ako je $Cov(X_1, X_2) = 0$.*

Definicija 1.4. *Neka je $X = (X_1, X_2)$ dvodimenzionalan slučajni vektor i neka slučajne varijable X_1 i X_2 imaju varijance $\sigma_{X_1} > 0$ i $\sigma_{X_2} > 0$. Koeficijent korelacije slučajnog vektora $X = (X_1, X_2)$ definiran je izrazom*

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sigma_{X_1} \sigma_{X_2}}.$$

Teorem 1.2. *Neka je $X = (X_1, X_2)$ slučajni vektor i neka slučajne varijable X_1 i X_2 imaju varijance $\sigma_{X_1} > 0$ i $\sigma_{X_2} > 0$. Veza među komponentama je linearna, tj. postoje realni brojevi a $(a \neq 0)$ i b takvi da je*

$$X_2 = aX_1 + b$$

ako i samo ako je $|\rho(X_1, X_2)| = 1$. Pritom je koeficijent korelacije 1 ako je $a > 0$ odnosno -1 ako je $a < 0$.

Dokaz se može pronaći u [4].

Neka je $X = (X_1, \ldots, X_p)$ p-dimenzionalan slučajni vektor takav da postoji $E[X_i^2]$ za $i = 1, \ldots, p$. Zapišemo li očekivanja slučajnih varijabli $E[X_1], \ldots E[X_p]$ u vektorskom obliku kao $E[X] = [E[X_1], \ldots, E[X_p]]^T$ onda je

$$E[(X - E[X])(X - E[X])^T] = \begin{bmatrix} E[(X_1 - E[X_1])^2] & \ldots & E[(X_1 - E[X_1])(X_p - E[X_p])] \\ \vdots & \ddots & \vdots \\ E[(X_p - E[X_p])(X_1 - E[X_1])] & \ldots & E[(X_p - E[X_p])^2] \end{bmatrix}$$

$$= \begin{bmatrix} VarX_1 & \ldots & Cov(X_1, X_p) \\ \ldots & \ddots & \ldots \\ Cov(X_1, X_p) & \ldots & Var(X_p) \end{bmatrix}.$$

Takvu matricu $Cov(X) = E[(X - E[X])(X - E[X])^T]$ zovemo matrica kovarijanci slučajnog vektora X.

Primijetimo, matrica kovarijanci je simetrična i pozitivno semidefinitna[1]. Pokažimo da je matrica kovarijanci pozitivno semidefinitna.

Neka je $v \in \mathbb{R}^p$ proizvoljan vektor. Vrijedi sljedeće:

$$v^T Cov(X) v = v^T E[(X - E[X])(X - E[X]^T)]v = E(v^T(X - E[X]))^2 \geq 0.$$

Označimo s X_s slučajni vektor dobiven standardizacijom slučajnog vektora X, tj. :

$$X_s = \left[\frac{X_1 - E[X_1]}{\sqrt{VarX_1}}, \ldots, \frac{X_p - E[X_p]}{\sqrt{VarX_p}} \right]^T.$$

Matricu kovarijanci slučajnog vektora X_s zovemo korelacijska matrica slučajnog vektora $X = (X_1, \ldots, X_p)$ i označavamo s $Corr(X)$,

$$Corr(X) = \begin{bmatrix} 1 & \rho_{12} & \ldots & \rho_{1p} \\ \rho_{21} & 1 & \ldots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \ldots & 1 \end{bmatrix},$$

pri čemu je

$$\rho_{ij} = \rho(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}}, \quad \forall i, j \in \{1, \ldots p\}.$$

U nastavku navodimo još neke osnovne rezultate iz linearne algebre:

Definicija 1.5. *Vektor $v \in \mathbb{R}^n$ svojstveni je vektor matrice $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ s pripadnom svojstvenom vrijednošću $\lambda \in \mathbb{R}$ ako je*

$$\boldsymbol{A}v = \lambda v, \quad v \neq 0.$$

Definicija 1.6. *Za matricu $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ definiramo karakteristični polinom od $\boldsymbol{A}$ kao*

$$k_{\boldsymbol{A}}(z) = \det(\boldsymbol{A} - z\boldsymbol{I}).$$

Teorem 1.3. *$\lambda \in \mathbb{R}$ je svojstvena vrijednost matrice $\boldsymbol{A}$ ako i samo ako je $k_{\boldsymbol{A}}(\lambda) = 0$.*

Dokaz se može pronaći u [3] ili [23].

Propozicija 1.2. *Neka je $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ pozitivno semidefinitna matrica. Tada su njene svojstvene vrijednosti nenegativne. Svojstveni vektori koji pripadaju različitim svojstvenim vrijednostima matrice $\boldsymbol{A}$ su ortogonalni.*

Dokaz se može pronaći u [3].

Teorem 1.4. *Svaka simetrična matrica je ortogonalno slična dijagonalnoj matrici.*

Dokaz se može pronaći u [3].

[1] Za simetričnu matricu $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ kažemo da je pozitivno semidefinitna ako za sve $x \in \mathbb{R}^n$ vrijedi da je $x^T \boldsymbol{A} x \geq 0$.

2 Analiza glavnih komponenti

Kako bismo ispitali veze između p koreliranih varijabli, ponekad je korisno pretvoriti taj originalan skup varijabli u novi skup nekoreliranih varijabli koje zovemo glavne komponente. Glavne komponente su međusobno nekorelirane linearne kombinacije originalnih varijabli čije varijance imaju posebno svojstvo. Primjerice, prva glavna komponenta je linearna kombinacija slučajnih varijabli maksimalne varijance. Prva glavna komponenta stoga objašnjava što je više moguće varijacije originalnih podataka. Obično je ideja analize glavnih komponenti vidjeti objašnjava li prvih nekoliko glavnih komponenti većinu varijacije originalnih podataka. Ukoliko je to slučaj, neke od originalnih varijabli su vrlo korelirane i daju nam iste informacije pa je moguće smanjiti dimenzionalnost problema.

Može se pokazati da se izračunavanje glavnih komponenti svodi na računanje svojstvenih vektora i svojstvenih vrijednosti matrice kovarijanci slučajnog uzorka, koja je pozitivno semidefinitna matrica. Dakle, jedna od primjena teorije svojstvenih vrijednosti i svojstvenih vektora je analiza glavnih komponenti.

2.1 Izvod glavnih komponenti

Neka je $X = (X_1, \ldots, X_p)$ p-dimenzionalan slučajni vektor. Prva glavna komponenta linearna je kombinacija komponenti tog slučajnog vektora s maksimalnom varijancom. Drugim riječima, za izvod prve glavne komponente tražimo linearnu funkciju $\alpha_1^T X$, gdje je α_1 p-dimenzionalan vektor konstanti,

$$\alpha_1^T X = \alpha_{11} X_1 + \cdots + \alpha_{1p} X_p$$

koja ima maksimalnu varijancu. Nadalje, tražimo linearnu funkciju $\alpha_2^T X$ nekoreliranu s $\alpha_1^T X$ maksimalne varijance i nju zovemo druga glavna komponenta. K-ta glavna komponenta linearna je funkcija $\alpha_k^T X$ maksimalne varijance nekorelirana s $\alpha_1^T X, \alpha_2^T X, \ldots, \alpha_{k-1}^T X$.

Može se pronaći do p glavnih komponenti, no, cilj analize glavnih komponenti je smanjiti dimenzionalnost problema, stoga očekujemo da će većina varijacije slučajnog vektora X biti opisana s m glavnih komponenti gdje je $m < p$. Motivacija ove metode je da je jednostavnije provoditi nekakvu analizu na manjem uzorku nekoreliranih varijabli nego na većem uzorku koreliranih varijabli. Važno je napomenuti da nema smisla provoditi analizu glavnih komponenti u slučaju da su varijable gotovo nekorelirane. Tada će taj postupak pronaći komponente koje su vrlo slične originalnim varijablama. Sada kada smo pojasnili osnovne ideje analize glavnih komponenti pokazat ćemo kako ih pronaći.

Neka je X p-dimenzionalan slučajni vektor s matricom kovarijanci $\mathbf{\Sigma}$ i vektorom očekivanja $E[X] = 0$ te neka je $\alpha_1 \in \mathbb{R}^p$ p-dimenzionalan vektor. Da bismo dobili prvu glavnu komponentu trebamo maksimizirati varijancu slučajnog vektora $\alpha_1^T X$:

$$Var(\alpha_1^T X) = E[\alpha_1^T X X^T \alpha_1] = \alpha_1^T \mathbf{\Sigma} \alpha_1. \tag{2.1}$$

Očito se maksimalna varijanca neće postići za konačan α_1, stoga uvodimo ograničenje da za α_1 mora vrijediti da je $\alpha_1^T \alpha_1 = 1$. Dakle, treba maksimizirati (2.1) uz uvjet:

$$\alpha_1^T \alpha_1 = 1. \tag{2.2}$$

Klasičan pristup rješavanju ovog problema jest Lagrangeova metoda za određivanje uvjetnih ekstrema. Treba maksimizirati:

$$\phi_1 = \alpha_1^T \mathbf{\Sigma} \alpha_1 - \lambda(\alpha_1^T \alpha_1 - 1),$$

pri čemu je λ Lagrangeov multiplikator.

Parcijalno deriviranje po varijabli α_1 daje da treba vrijediti sljedeće:

$$\frac{\partial \phi_1}{\partial \alpha_1} = 2\boldsymbol{\Sigma}\alpha_1 - 2\lambda\alpha_1 = 0,$$

tj.

$$\begin{aligned}(\boldsymbol{\Sigma} - \lambda \boldsymbol{I}_p)\alpha_1 &= 0,\\ \boldsymbol{\Sigma}\alpha_1 &= \lambda\alpha_1,\end{aligned} \tag{2.3}$$

gdje je $\boldsymbol{I}_p$ jedinična matrica dimenzija $p \times p$.

Očito je λ svojstvena vrijednost matrice kovarijanci $\boldsymbol{\Sigma}$, a α_1 pripadajući svojstveni vektor (vidi Definicija 1.5.). Primijetimo sada da za varijancu slučajnog vektora $\alpha_1^T X$ vrijedi sljedeće:

$$Var(\alpha_1^T X) \overset{(2.1)}{=} \alpha_1^T \boldsymbol{\Sigma} \alpha_1 \overset{(2.3)}{=} \alpha_1^T \lambda \alpha_1 = \lambda \alpha_1^T \alpha_1 \overset{(2.2)}{=} \lambda \tag{2.4}$$

stoga za maksimizaciju varijance treba maksimizirati λ.

Iz ovoga slijedi da je λ najveća svojstvena vrijednost matrice kovarijanci $\boldsymbol{\Sigma}$. Zaključujemo da je prva glavna komponenta slučajnog vektora X jednaka $\alpha_1^T X$ pri čemu je α_1 svojstveni vektor koji pripada najvećoj svojstvenoj vrijednosti matrice kovarijanci $\boldsymbol{\Sigma}$. Kako je $\boldsymbol{\Sigma}$ pozitivno semidefinitna matrica, sve njene svojstvene vrijednosti su veće ili jednake nuli. Za sada ćemo pretpostaviti da su sve svojstvene vrijednosti matrice $\boldsymbol{\Sigma}$ međusobno različite i veće od 0, tj. da su svojstvene vrijednosti od $\boldsymbol{\Sigma}$: $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$. Dakle, $Var(\alpha_1^T X) = \lambda_1$.

Za izračunavanje druge glavne komponente $\alpha_2^T X$ vektora X treba maksimizirati $Var(\alpha_2^T X) = \alpha_2^T \boldsymbol{\Sigma} \alpha_2$ uz uvjete $Cov(\alpha_1^T X, \alpha_2^T X) = 0$ i $\alpha_2^T \alpha_2 = 1$. Za matricu kovarijanci $Cov(\alpha_1^T X, \alpha_2^T X)$ vrijedi sljedeće:

$$\begin{aligned}Cov(\alpha_1^T X, \alpha_2^T X) &= E[(\alpha_1^T X - E[\alpha_1^T X])(\alpha_2^T X - E[\alpha_2^T X])^T]\\ &= E[\alpha_1^T (X - E[X])(X - E[X])]^T \alpha_2]\\ &= \alpha_1^T E[(X - E[X])(X - E[X])^T]\alpha_2\\ &= \alpha_1^T \boldsymbol{\Sigma} \alpha_2\\ &= \alpha_2^T \boldsymbol{\Sigma} \alpha_1\\ &= \alpha_2^T \lambda_1 \alpha_1\\ &= \lambda_1 \alpha_1^T \alpha_2.\end{aligned}$$

Iz ovoga slijedi da uvjet $Cov(\alpha_1^T X, \alpha_2^T X) = 0$ možemo zapisati na različite načine:

$$\alpha_1^T \boldsymbol{\Sigma} \alpha_2 = 0, \;\; \alpha_2^T \boldsymbol{\Sigma} \alpha_1 = 0 \tag{2.5}$$

$$\alpha_1^T \alpha_2 = 0, \;\; \alpha_2^T \alpha_1 = 0. \tag{2.6}$$

Odabir zapisa uvjeta je proizvoljan. Uz uvjet da je $\alpha_2^T \alpha_2 = 1$ i npr. $\alpha_2^T \alpha_1 = 0$ dolazimo do sljedećeg izraza kojeg treba maksimizirati:

$$\phi_2 = \alpha_2^T \boldsymbol{\Sigma} \alpha_2 - \lambda(\alpha_2^T \alpha_2 - 1) - \nu \alpha_2^T \alpha_1,$$

gdje su λ i ν Lagrangeovi multiplikatori.

Parcijalnim deriviranjem po α_2 dobivamo:

$$\frac{\partial \phi_1}{\partial \alpha_2} = 2\boldsymbol{\Sigma}\alpha_2 - 2\lambda\alpha_2 - \nu\alpha_1 = 0. \tag{2.7}$$

Množenjem ovog izraza slijeva s α_1^T daje:

$$2\alpha_1^T\mathbf{\Sigma}\alpha_2 - 2\alpha_1^T\lambda\alpha_2 - \nu\alpha_1^T\alpha_1 = 0,$$

a prva dva izraza su jednaka nuli zbog (2.5) i (2.6), stoga slijedi da je:

$$\nu\alpha_1^T\alpha_1 = 0,$$

tj. zbog $\alpha_1^T\alpha_1 = 1$:

$$\nu = 0. \tag{2.8}$$

Uvrštavanjem (2.8) u (2.7) dobivamo:

$$\mathbf{\Sigma}\alpha_2 - \lambda\alpha_2 = 0,$$

iz čega zaključujemo da je λ svojstvena vrijednost matrice $\mathbf{\Sigma}$, a α_2 pripadajući svojstveni vektor.

Slično kao i za varijancu prve glavne komponente (vidi (2.4)), može se pokazati da za varijancu druge glavne komponente vrijedi:

$$Var(\alpha_2^T X) = \alpha_2^T\mathbf{\Sigma}\alpha_2 = \lambda.$$

S obzirom na to da želimo maksimizirati varijancu vektora $\alpha_2^T X$, λ treba biti što veći. Kako smo pretpostavili da $\mathbf{\Sigma}$ nema dvije jednake svojstvene vrijednosti, slijedi da je $\lambda \neq \lambda_1$. Kada bi vrijedilo da je $\lambda = \lambda_1$ to bi značilo da je $\lambda_2 = \lambda_1$ što bi bilo u kontradikciji s uvjetom $\alpha_2^T\alpha_1 = 0$. Stoga za λ uzimamo drugu najveću svojstvenu vrijednost od $\mathbf{\Sigma}$, tj. $\lambda = \lambda_2$, a α_2 je pripadajući svojstveni vektor.

Slično se može pokazati da je k-ta glavna komponenta slučajnog vektora X jednaka $\alpha_k^T X$ i $Var(\alpha_k^T X) = \lambda_k$ pri čemu je λ_k k-ta najveća svojstvena vrijednost matrice $\mathbf{\Sigma}$, a α_k pripadajući svojstveni vektor za $k = 3, \ldots, p$. Dokaz se može pronaći u [1].

U ovom će se radu termin glavne komponente odnositi na izvedene varijable $\alpha_k X$, a α_k ćemo zvati vektor koeficijenata ili vektor težina k-te glavne komponente, $k = 1, 2, \ldots, p$.

Za ilustraciju računanja glavnih komponenti na podacima vidi Primjer 2.1.

2.2 Svojstva glavnih komponenti

U ovom ćemo dijelu rada razmotriti neka matematička i statistička svojstva glavnih komponenti matrice kovarijanci $\mathbf{\Sigma}$ slučajnog vektora X. Do sada smo definirali i izveli glavne komponente, no to nije jedini način na koji ih možemo definirati i izvesti. Glavne komponente su "optimalne" linearne funkcije od X s obzirom na različite kriterije, stoga ih možemo izvesti i preko nekog algebarskog ili geometrijskog kriterija. U nastavku će biti navedeni neki od kriterija preko kojih također možemo definirati glavne komponente.

Neka je Z p-dimenzionalni vektor čiji je k-ti element, Z_k, k-ta glavna komponenta vektora X, za $k = 1, \ldots, p$. Tada je:

$$Z = \boldsymbol{A}^T X \tag{2.9}$$

pri čemu je $\boldsymbol{A}$ ortogonalna matrica čiji je k-ti stupac, α_k, k-ti svojstveni vektor matrice kovarijanci $\mathbf{\Sigma}$.

Time smo definirali glavne komponente ortogonalnom linearnom transformacijom slučajnog vektora X. Kako su stupci matrice $\boldsymbol{A}$ svojstveni vektori matrice $\mathbf{\Sigma}$ vrijedi i sljedeće:

$$\mathbf{\Sigma}\boldsymbol{A} = \boldsymbol{A}\mathbf{\Lambda}, \tag{2.10}$$

pri čemu je $\boldsymbol{\Lambda}$ dijagonalna matrica čiji je dijagonalni element, λ_k, k-ta svojstvena vrijednost matrice $\boldsymbol{\Sigma}$ te $\lambda_k = Var(\alpha_k^T X) = Var(Z_k)$.

Zbog ortogonalnosti matrice $\boldsymbol{A}$ slijedi da se izraz (2.10) može zapisati i na sljedeće načine:

$$\boldsymbol{A}^T \boldsymbol{\Sigma} \boldsymbol{A} = \boldsymbol{\Lambda} \tag{2.11}$$

$$\boldsymbol{\Sigma} = \boldsymbol{A} \boldsymbol{\Lambda} \boldsymbol{A}^T. \tag{2.12}$$

U nastavku poglavlja koristit ćemo oznake koje smo upravo uveli.

Teorem 2.1. *Neka je $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ ortogonalna matrica i $X = (X_1, \ldots, X_p)$ slučajni vektor. Tada ortogonalna transformacija $Y = \boldsymbol{A}X$ slučajnog vektora X i sam slučajni vektor X imaju jednaku generaliziranu varijancu te im je zbroj varijanci komponenti jednak.*

Dokaz. Neka je $E[X] = 0$ i $E[XX^T] = \boldsymbol{\Sigma}$. Tada vrijedi da je $E[Y] = 0$ i $E[YY^T] = E[\boldsymbol{A}X(\boldsymbol{A}X)^T] = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$. Generalizirana varijanca vektora Y je:

$$\det(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T) = \det(\boldsymbol{A}) \det(\boldsymbol{\Sigma}) \det(\boldsymbol{A}^T) = \det(\boldsymbol{\Sigma}) \det(\boldsymbol{A}\boldsymbol{A}^T) = \det(\boldsymbol{\Sigma})$$

što je jednako generaliziranoj varijanci slučajnog vektora X.

Suma varijanci komponenti vektora Y je:

$$\sum_{i=1}^{p} E[Y_i^2] = tr(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T) = tr(\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{A}) = tr(\boldsymbol{\Sigma}\boldsymbol{I}) = \sum_{i=1}^{p} E[X_i^2].$$

□

Kako smo upravo pokazali da vektor glavnih komponenti možemo zapisati kao ortogonalnu transformaciju vektora X vrijedi:

Propozicija 2.1. *Generalizirana varijanca vektora glavnih komponenti jednaka je generaliziranoj varijanci originalnog vektora i suma varijanci glavnih komponenti jednaka je sumi varijanci originalnih slučajnih varijabli slučajnog vektora.*

Kako je suma varijanci glavnih komponenti jednaka sumi varijanci originalnih slučajnih varijabli slučajnog vektora, tj.

$$\sum_{i=1}^{p} Var(Z_i) = \sum_{i=1}^{p} \lambda_i = tr(\Lambda) = \sum_{i=1}^{p} Var(X_i)$$

možemo reći da i-ta glavna komponenta objašnjava $\lambda_i / \sum_{j=1}^{p} \lambda_j$ varijacije originalnih podataka. Slično, prvih m glavnih komponenti objašnjava $\sum_{j=1}^{m} \lambda_j / \sum_{j=1}^{p} \lambda_j$ varijacije originalnih podataka.

Sada ćemo razmotriti neka optimizacijska svojstva glavnih komponenti koja se mogu pronaći u [16].

Svojstvo 1. *Neka je $q \in \mathbb{N}$, $q \leq p$, i $\boldsymbol{B}$ ortonormirana matrica dimenzija $p \times q$. Tada je za linearnu transformaciju*

$$Y = \boldsymbol{B}^T X,$$

gdje je Y q-dimenzionalan slučajni vektor, trag matrice kovarijanci slučajnog vektora Y, $\boldsymbol{\Sigma}_y = \boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B}$, maksimiziran za $\boldsymbol{B} = \boldsymbol{A}_q$, pri čemu je $\boldsymbol{A}_q$ matrica koja se sastoji od prvih q redaka matrice $\boldsymbol{A}$, čiji su stupci svojstveni vektori matrice $\boldsymbol{\Sigma}$.

Dokaz. Označimo s β_k stupce matrice $\boldsymbol{B}$. Stupci matrice $\boldsymbol{A}$ svojstveni su vektori matrice kovarijance $\boldsymbol{\Sigma}$ i tvore bazu za p-dimenzionalan prostor. Stupce β_k stoga možemo zapisati u sljedećem obliku:

$$\beta_k = \sum_{j=1}^{p} c_{jk}\alpha_j, \quad k = 1, 2, \ldots, q,$$

gdje su c_{jk}, za $j = 1, 2, \ldots, p$ i $k = 1, \ldots, q$, odgovarajuće konstante.

Matricu $\boldsymbol{B}$ onda možemo zapisati kao

$$\boldsymbol{B} = \boldsymbol{AC}$$

pri čemu je $\boldsymbol{C}$ matrica dimenzija $p \times q$ i $[\boldsymbol{C}]_{jk} = c_{jk}$.

Kako je matrica $\boldsymbol{A}$ ortogonalna, a stupci matrice $\boldsymbol{B}$ ortonormirani vrijedi i:

$$\boldsymbol{C} = \boldsymbol{A}^T\boldsymbol{B}, \tag{2.13}$$

$$\boldsymbol{C}^T\boldsymbol{C} = \boldsymbol{B}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{B} = \boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{I}_q, \tag{2.14}$$

gdje je $\boldsymbol{I}_q$ jedinična matrica dimenzija $q \times q$.

Dakle, slijedi da su stupci matrice $\boldsymbol{C}$ također ortonormirani i:

$$\sum_{j=1}^{p}\sum_{k=1}^{q} c_{jk}^2 = q. \tag{2.15}$$

Matricu $\boldsymbol{C}$ možemo shvatiti kao prvih q stupaca ortogonalne matrice $\boldsymbol{D}$ dimenzije $p \times p$. Kako je $\boldsymbol{D}$ ortogonalna matrica, redci joj zadovoljavaju svojstvo $d_j^T d_j = 1$, za $j = 1, \ldots, p$. Kako su redci matrice $\boldsymbol{C}$ sastavljeni od prvih q elemenata redaka matrice $\boldsymbol{D}$ slijedi da je $c_j^T c_j \leq 1$, za $j = 1, \ldots, p$, tj.

$$\sum_{k=1}^{q} c_{jk}^2 \leq 1. \tag{2.16}$$

Za matricu kovarijanci $\boldsymbol{\Sigma}_y$ slučajnog vektora Y, $\boldsymbol{\Sigma}_y = \boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B}$, vrijedi sljedeće:

$$\begin{aligned}\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{C}^T\boldsymbol{A}^T\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{C} &= \boldsymbol{C}^T\boldsymbol{\Lambda}\boldsymbol{C}, \quad \textit{koristeći (2.11)} \\ &= \sum_{j=1}^{p} \lambda_j c_j c_j^T\end{aligned}$$

pri čemu je c_j^T j-ti redak matrice $\boldsymbol{C}$.

Za trag od $\boldsymbol{\Sigma}_y$ vrijedi:

$$\begin{aligned}tr(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B}) &= \sum_{j=1}^{p} \lambda_j tr(c_j c_j^T) \\ &= \sum_{j=1}^{p} \lambda_j tr(c_j^T c_j) \\ &= \sum_{j=1}^{p} \lambda_j c_j^T c_j \\ &= \sum_{j=1}^{p}\sum_{k=1}^{q} \lambda_j c_{jk}^2. \end{aligned} \tag{2.17}$$

Želimo maksimizirati izraz (2.17) kojeg možemo zapisati i ovako:

$$\sum_{j=1}^{p}\left(\sum_{k=1}^{q} c_{jk}^2\right)\lambda_j. \tag{2.18}$$

Uočimo da se u tom izrazu pojavljuje suma $\sum_{k=1}^{q} c_{jk}^2$ za koju smo pokazali da je manja ili jednaka od 1 (2.16). Osim toga smo pokazali i da je $\sum_{j=1}^{p}\sum_{k=1}^{q} c_{jk}^2$ jednako q (2.15). Kako smo pretpostavili da je $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ jasno je da će izraz (2.18) biti maksimiziran za c_{jk} takve da vrijedi:

$$\sum_{k=1}^{q} c_{jk}^2 = \begin{cases} 1, & j = 1, \ldots, q \\ 0, & j = q+1, \ldots, n. \end{cases} \tag{2.19}$$

Ako stavimo da je $\boldsymbol{B}^T = \boldsymbol{A}_q^T$ onda je:

$$c_{jk} = \begin{cases} 1, & 1 \leq j = k \leq q \\ 0, & \textit{inače.} \end{cases}$$

čime smo zadovolji (2.19). Time smo pokazali da je $tr(\boldsymbol{\Sigma}_y)$ maksimalne vrijednosti za $\boldsymbol{B}^T = \boldsymbol{A}_q^T$. □

Svojstvo 2. *Neka je $q \in \mathbb{N}$, $q \leq p$ i $\boldsymbol{B}$ ortonormirana matrica dimenzija $p \times q$. Tada je za linearnu transformaciju:*

$$Y = \boldsymbol{B}^T X,$$

gdje je Y q-dimenzionalan slučajni vektor, trag matrice kovarijanci slučajnog vektora Y, $\boldsymbol{\Sigma}_y = \boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B}$, minimiziran za $\boldsymbol{B} = \boldsymbol{A}_q^*$ gdje je $\boldsymbol{A}_q^*$ matrica koja se sastoji od posljednjih q redaka matrice $\boldsymbol{A}$, čiji su stupci svojstveni vektori matrice $\boldsymbol{\Sigma}$..

Ovo se svojstvo može dokazati na sličan način kao i prethodno svojstvo, Svojstvo 1., stoga taj dokaz nećemo navoditi. Statistička implikacija ovog svojstva je da zadnje glavne komponente nisu samo ostaci analize glavnih komponenti koje treba zanemariti. Pokazuje se da su zadnje glavne komponente korisne u pronalaženju gotovo linearnih veza između elemenata vektora X i pri odabiru podskupa varijabli vektora X radi smanjenja dimenzionalnosti.

Svojstvo 3. (Spektralna dekompozicija matrice kovarijanci $\boldsymbol{\Sigma}$)
Za matricu kovarijanci $\boldsymbol{\Sigma}$ vrijedi:

$$\boldsymbol{\Sigma} = \lambda_1\alpha_1\alpha_1^T + \lambda_2\alpha_2\alpha_2^T + \cdots + \lambda_p\alpha_p\alpha_p^T, \tag{2.20}$$

pri čemu su $\lambda_1 > \lambda_2 > \cdots > \lambda_p \geq 0$ svojstvene vrijednosti od $\boldsymbol{\Sigma}$, a $\alpha_1, \alpha_2, \ldots, \alpha_p$ pripadajući svojstveni vektori.

Dokaz. Iz (2.12):

$$\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^T$$

raspisivanjem desne strane jednadžbe dobivamo da je

$$\boldsymbol{\Sigma} = \sum_{k=1}^{p} \lambda_k\alpha_k\alpha_k^T,$$

čime smo pokazali (2.20). □

Iz ovog svojstva možemo uočiti i da je:

$$Var(X_j) = \sum_{k=1}^{p} \lambda_k\alpha_{kj}^2.$$

Statistička implikacija ovog svojstva je da varijance komponenti vektora X možemo rastaviti na opadajuće doprinose glavnih komponenti, ali i cijelu matricu kovarijanci možemo rastaviti na doprinose $\lambda_k\alpha_k\alpha_k^T$ svake

od glavnih komponenti. Svojstvo 1. govori da glavne komponente, redom, objašnjavaju najviše moguće $\mathrm{tr}(\boldsymbol{\Sigma})$, a trenutno svojstvo govori, intuitivno, da dobro objašnjavaju i ne-dijagonalne elemente matrice $\boldsymbol{\Sigma}$.

Iz (2.20) je jasno da se matrica kovarijanci može egzaktno rekonstruirati uz pomoć koeficijenata i varijanci prvih r glavnih komponenti ukoliko je rang te matrice jednak r.

Svojstvo 4. *Neka je $q \in \mathbb{N}$, $q \leq p$, i $\boldsymbol{B}$ ortonormirana matrica dimenzija $p \times q$. Tada je za linearnu transformaciju:*

$$Y = \boldsymbol{B}^T X,$$

gdje je Y q-dimenzionalan slučajni vektor, determinanta matrice kovarijanci slučajnog vektora Y, $\det(\boldsymbol{\Sigma}_y) = \det(\boldsymbol{B}^T \boldsymbol{\Sigma} \boldsymbol{B})$, maksimizirana za $\boldsymbol{B} = \boldsymbol{A}_q$ gdje je $\boldsymbol{A}_q$ matrica koja se sastoji od prvih q redaka matrice $\boldsymbol{A}$, čiji su stupci svojstveni vektori matrice $\boldsymbol{\Sigma}$.

Dokaz ovog svojstva se može pronaći u [16].

Svojstvo 5. *Pretpostavimo da želimo napraviti predikciju slučajnih varijabli X_j slučajnog vektora X na temelju linearne funkcije $Y = \boldsymbol{B}^T X$ koju smo definirali ranije. Označimo sa σ_j^2 rezidualnu varijancu dobivenu prediktiranjem X_j na temelju Y. Tada je suma rezidualnih varijanci, $\sum_{j=1}^{p} \sigma_j^2$, minimizirana za $\boldsymbol{B} = \boldsymbol{A}_q$.*

Dokaz. Vidi [16]. □

Statistička implikacija ovog rezultata je da je najbolji linearni prediktor od X u q-dimenzionalnom prostoru, u smislu minimiziranja sume rezidualnih varijanci, definiran s prvih q glavnih komponenti.

Do sada smo naveli nekoliko algebarskih svojstva glavnih komponenti, a sada ćemo se osvrnuti i na neka njihova geometrijska svojstva.

Svojstvo 6. *Promotrimo familiju p-dimenzionalnih elipsoida*

$$X^T \boldsymbol{\Sigma}^{-1} X = const. \tag{2.21}$$

Glavne komponente definiraju osi tih elipsoida.

Dokaz. Glavne komponente su definirane transformacijom $Z = \boldsymbol{A}^T X$ (2.9). Matrica $\boldsymbol{A}$ je ortogonalna stoga vrijedi da je $X = \boldsymbol{A} Z$.

Uvrštavanjem $X = \boldsymbol{A} Z$ u (2.21) dobivamo:

$$(\boldsymbol{A} Z)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{A} Z) = const = Z^T \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A} Z. \tag{2.22}$$

Općenito vrijedi da su svojstveni vektori matrice $\boldsymbol{\Sigma}^{-1}$ jednaki svojstvenim vektorima matrice $\boldsymbol{\Sigma}$, a svojstvene vrijednosti matrice $\boldsymbol{\Sigma}^{-1}$ jednake su recipročnim svojstvenim vrijednostima matrice $\boldsymbol{\Sigma}$ uz pretpostavku da su sve pozitivne.

Iz (2.11) je $\boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A} = \boldsymbol{\Lambda}^{-1}$ stoga za (2.22) slijedi:

$$Z^T \boldsymbol{\Lambda}^{-1} Z = const.$$

Zadnja jednadžba može biti zapisana kao:

$$\sum_{k=1}^{p} \frac{Z_k^2}{\lambda_k} = const,$$

što je jednadžba elipsoida. Jednadžba također implicira da su dužine poluosi elipsoida jednake $\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_p^{1/2}$.

□

Ovaj rezultat je važan u statističkom smislu kada vektor X ima višedimenzionalnu normalnu distribuciju. U tom slučaju elipsoidi dani s (2.22) definiraju konture na kojima je funkcija distribucije vektora X jednaka konstanti. Rezultat daje geometrijsku interpretaciju algebarske definicije glavnih komponenti koju smo dali u prvom poglavlju. Naime, prva (najveća) os elipsoida definira smjer u kojemu je statistička varijacija najveća. Druga os maksimizira statističku varijaciju uz uvjet da je ortogonalna na prvu os, itd.

Svojstvo 7. *Neka su X_1 i X_2 nezavisni jednako distribuirani slučajni vektori podvrgnuti istoj linearnoj transformaciji:*

$$Y_i = \boldsymbol{B}^T X_i, \quad i = 1, 2,$$

pri čemu je $\boldsymbol{B}$ ortonormirana matrica dimenzija $p \times q$. Vrijednost $E[(Y_1 - Y_2)^T(Y_1 - Y_2)]$ maksimizirana je za $\boldsymbol{B} = \boldsymbol{A}_q$, gdje je $\boldsymbol{A}_q$ matrica koja se sastoji od prvih q redaka matrice $\boldsymbol{A}$, čiji su stupci svojstveni vektori matrice $\boldsymbol{\Sigma}$.

Dokaz. S obzirom na to da su X_1 i X_2 jednako distribuirani imaju jednaka očekivanja μ i matricu kovarijanci $\boldsymbol{\Sigma}$. Stoga, Y_1 i Y_2 također imaju jednako očekivanje, $\boldsymbol{B}^T\mu$, i jednaku matricu kovarijanci, $\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B}$.

$$\begin{aligned}
&E[(Y_1 - Y_2)^T(Y_1 - Y_2)] \\
&= E[\{(Y_1 - \boldsymbol{B}^T\mu) - (Y_2 - \boldsymbol{B}^T\mu)\}^T\{(Y_1 - \boldsymbol{B}^T\mu) - (Y_2 - \boldsymbol{B}^T\mu)\}] \\
&= E[(Y_1 - \boldsymbol{B}^T\mu)^T(Y_1 - \boldsymbol{B}^T\mu)] + E[(Y_2 - \boldsymbol{B}^T\mu)^T(Y_2 - \boldsymbol{B}^T\mu)].
\end{aligned}$$

Međuprodukti se ponište zbog nezavisnosti Y_1 i Y_2.

Sada, za $i = 1, 2$ imamo:

$$\begin{aligned}
E[(Y_i - \boldsymbol{B}^T\mu)^T(Y_i - \boldsymbol{B}^T\mu)] &= E[tr\{(Y_i - \boldsymbol{B}^T\mu)^T(Y_i - \boldsymbol{B}^T\mu)\}] \\
&= E[tr\{(Y_i - \boldsymbol{B}^T\mu)(Y_i - \boldsymbol{B}^T\mu)^T\}] \\
&= tr\{E[(Y_i - \boldsymbol{B}^T\mu)(Y_i - \boldsymbol{B}^T\mu)^T]\} \\
&= tr(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B}),
\end{aligned}$$

a iz Svojstva 1. je trag $tr(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})$ maksimiziran za $\boldsymbol{B} = \boldsymbol{A}_q$ pa je svojstvo dokazano. □

Geometrijska interpretacija ovog svojstva je da je očekivana kvadratna euklidska udaljenost u q-dimenzionalnom potprostoru između dva p-dimenzionalna vektora iste distribucije maksimizirana ukoliko je potprostor definiran s prvih q glavnih komponenti.

2.3 Glavne komponente jednakih i nul-svojstvenih vrijednosti

Do sada smo pretpostavljali da su svojstvene vrijednosti matrice kovarijanci sve međusobno različite i da nijedna nije jednaka nuli. U praksi je vrlo rijetko da su neke od svojstvenih vrijednosti jednake ili da su jednake nuli.

U slučaju jednakosti q svojstvenih vrijednosti pripadajući svojstveni vektori razapinju q-dimenzionalan prostor, no, unutar tog prostora su proizvoljni. Geometrijska interpretacija ovog slučaja je da za npr. $q = 2$ ili $q = 3$ osi kružnice ili sfere nisu jedinstveno definirane (vidi Svojstvo 6.), a do sličnog problema dolazi i za $q > 3$, kod hipersferi. Stoga glavne komponente koje pripadaju svojstvenim vrijednostima koje se ponavljaju nisu jedinstveno definirane.

Drugi slučaj, da je neka od svojstvenih vrijednosti jednaka nuli pojavljuje se češće u praksi, no i taj je slučaj vrlo rijedak. Ukoliko je q svojstvenih vrijednosti matrice $\boldsymbol{\Sigma}$ jednako nuli, onda je rang matrice $\boldsymbol{\Sigma}$ jednak $p-q$ umjesto q. Glavna komponenta čija je varijanca jednaka 0 definira konstantnu linearnu vezu između elemenata vektora X. Ako ovakva veza postoji, to implicira da se vrijednost jedne varijable može točno izračunati pomoću drugih varijabli. Stoga bismo u ovom slučaju mogli smanjiti broj varijabli s p na $p-q$ bez gubitka informacija te ponovno provesti izračun glavnih komponenti.

2.4 Glavne komponente uzorka

U prethodnim smo poglavljima razmatrali glavne komponente na razini populacije. U praksi nam je dostupan uzorak na temelju kojeg procjenjujemo glavne komponente. U ovom ćemo poglavlju opisati kako procjenjujemo glavne komponente na temelju podataka i navesti neke rezultate vezane za uzoračke glavne komponente.

Pretpostavimo da imamo realizacije $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ slučajnog uzorka $((X_1^1, \ldots, X_p^1), \ldots, (X_1^n, \ldots, X_p^n))$.

Definiramo vrijednosti

$$\tilde{z}_{i1} = a_1^T \boldsymbol{x}_i, \quad i = 1, 2, \ldots, n,$$

pri čemu je a_1^T vektor koeficijenata koji maksimizira uzoračku varijancu

$$\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{z}_{i1} - \overline{z}_1)^2, \quad \overline{z}_1 = \frac{1}{n}\sum_{i=1}^{n}\tilde{z}_{i1},$$

uz uvjet da je $a_1^T a_1 = 1$.

Nadalje, neka je

$$\tilde{z}_{i2} = a_2^T \boldsymbol{x}_i, \quad i = 1, 2, \ldots, n,$$

i a_2^T odabran na način da maksimizira uzoračku varijancu od $\tilde{z}_{i2}$ uz uvjete da je

1. $a_2^T a_2 = 1$,
2. $\tilde{z}_{i2}$ nekoreliran s $\tilde{z}_{i1}$ u uzorku.

Nastavljajući ovaj postupak dobivamo uzoračku verziju definicije glavnih komponenti. K-ta uzoračka glavna komponenta je $Z_k = a_k^T X$, za $k = 1, 2, \ldots, p$ i $X = [X_1, X_2, \ldots, X_p]^T$. Vrijednost $\tilde{z}_{ik}$ je skor i-tog podataka k-te glavne komponente.

Izvod analogan onom iz Poglavlja 2.1 se može provesti uz uzoračku varijancu i uzoračku matricu kovarijanci. Tada je varijanca k-te glavne komponente jednaka k-toj najvećoj svojstvenoj vrijednosti uzoračke matrice kovarijanci $\boldsymbol{S}$, a težine komponente su definirane pripadajućim svojstvenim vektorima za $k = 1, 2, \ldots, p$. U nastavku uvodimo neke oznake koje će nam biti od koristi kasnije.

Označimo svojstvene vrijednosti matrice $\boldsymbol{S}$ sa l_k i pripadne svojstvene vektore s a_k, za $k = 1.2, \ldots, p$.

Neka je $\tilde{\boldsymbol{X}}$ matrica dimenzije $n \times p$ i $[\tilde{\boldsymbol{X}}]_{ik} = \tilde{x}_{ik}$, pri čemu je $\tilde{x}_{ik}$ k-ti element od $\boldsymbol{x}_i$. Neka je $\tilde{\boldsymbol{Z}}$ također matrica dimenzije $n \times p$ i $[\tilde{\boldsymbol{Z}}]_{ik} = \tilde{z}_{ik}$. Tada vrijedi

$$\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{X}}\boldsymbol{A}$$

pri čemu je $\boldsymbol{A}$ ortogonalna matrica dimenzije $p \times p$ čiji je k-ti stupac jednak a_k.

Za uzoračku matricu kovarijanci, $\boldsymbol{S}$ je:

$$[\boldsymbol{S}]_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_{ij} - \overline{x}_j)(\tilde{x}_{ik} - \overline{x}_k),$$

pri čemu je

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{ij}, \quad j = 1, 2, \ldots, p.$$

Stoga, matricu $\boldsymbol{S}$ možemo zapisati kao

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X} \tag{2.23}$$

pri čemu je $\boldsymbol{X}$ matrica dimenzija $n \times p$ i $[\boldsymbol{X}]_{ij} = (\tilde{x}_{ij} - \overline{x}_j)$.

Zadnju oznaku koju uvodimo je druga definicija za matricu skorova glavnih komponenti:

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A}.$$

Skorovi glavnih komponenti ove matrice imaju jednake varijance i kovarijance kao i matrica $\tilde{\boldsymbol{Z}}$, no aritmetičke sredine stupaca te matrice su nula, a ne $\overline{z}_k$, za $k = 1, 2, \ldots, p$.

Svojstva koja smo naveli u poglavlju (2.2) vrijede i za uzoračke glavne komponente uz neke preinake. Navodimo još spektralnu dekompoziciju uzoračke matrice kovarijanci:

$$\boldsymbol{S} = l_1 a_1 a_1^T + l_2 a_2 a_2^T + \ldots l_p a_p a_p^T. \tag{2.24}$$

Primjer 2.1. *Ilustrirat ćemo tehniku izračunavanja glavnih komponenti na bazi podataka "trees" ugrađenoj u R-u. Baza podataka "trees" sadrži mjerenja za promjer, visinu i obujam trideset i jednog posječenog stabla crne trešnje. Promjer drveta mjeren je u inčima, visina drveta u stopama i volumen drveta u kubičnim stopama. Označimo varijable Promjer s X_1, Visinu s X_2 i Volumen s X_3.*

X_1	X_2	X_3
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
⋮	⋮	⋮

Tablica 2.1: Dio baze podataka "trees"

Procijenjena matrica kovarijanci za taj skup podataka je:

$$\begin{bmatrix} 9.847914 & 10.38333 & 49.88812 \\ 10.383333 & 40.60000 & 62.66000 \\ 49.888118 & 62.66000 & 270.20280 \end{bmatrix} \tag{2.25}$$

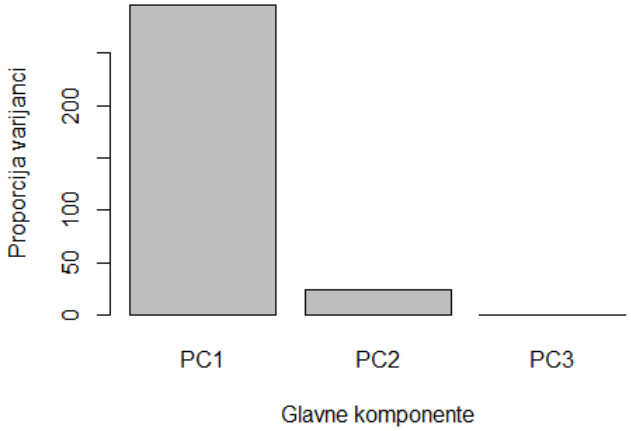
Svojstveni vektori i pripadajuće svojstvene vrijednosti matrice kovarijanci su:

$$\begin{aligned} a_1 &= [-0.1756, -0.2420, -0.9542 \ ]^T, \ \ l_1 = 295.2630 \\ a_2 &= [\ \ 0.0909, -0.9691, \ \ 0.2290 \ ]^T, \ \ l_2 = 24.8204 \\ a_3 &= [\ \ 0.9802, \ \ 0.0465, -0.1922 \ ]^T, \ \ l_3 = 0.5604. \end{aligned}$$

Glavne komponente su dane izrazom:

$$Z_i = a_i^T X,$$

a njihove varijance su l_i za $i = 1, 2, 3$.



Slika 2.1: Proporcije varijanci glavnih komponenti

	Z_1	Z_2	Z_3
Standardna devijacija	17.1834	4.98200	0.74859
Proporcija varijance	0.9208	0.07741	0.00175
Kumulativna proporcija	0.9208	0.99825	1.00000

Tablica 2.2: Važnost komponenti

Iz Tablice 2.2 iščitavamo da prva glavna komponenta objašnjava 92.08% ukupne varijacije podataka, druga komponenta objašnjava 7.741% ukupne varijacije podataka, a treća glavna komponenta objašnjava 0.175% ukupne varijacije.

2.4.1 Dekompozicija matrice na singularne vrijednosti

U ovom ćemo dijelu opisati poznat rezultat iz linearne algebre koji je bitan i u kontekstu analize glavnih komponenti - dekompoziciju matrice na singularne vrijednosti.

Neka je $\boldsymbol{X}$ proizvoljna matrica dimenzija $n \times p$. Tada $\boldsymbol{X}$ možemo faktorizirati kao:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{A}^T, \tag{2.26}$$

pri čemu je

1. $\boldsymbol{U}$ matrica dimenzije $n \times r$, $\boldsymbol{A}$ matrica dimenzije $p \times r$, a stupci matrica $\boldsymbol{A}$ i $\boldsymbol{U}$ su ortonormirani,
2. $\boldsymbol{L}$ dijagonalna matrica dimenzija $r \times r$,
3. rang matrice $\boldsymbol{X}$ jednak r.

Pretpostavimo da imamo realizacije $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ jednostavnog slučajnog uzorka $((X_1^1, X_2^1, \ldots, X_p^1), \ldots, (X_1^n, X_2^n, \ldots, X_p^n))$. Te podatke zapisujemo u matricu $\tilde{\boldsymbol{X}}$ dimenzija $n \times p$ na način da je $[\tilde{\boldsymbol{X}}]_{ik} = \tilde{x}_{ik}$ pri čemu je $\tilde{x}_{ik}$ k-ti element od $\boldsymbol{x}_i$.
Dokaz postojanja spektralne dekompozicije za proizvoljnu matricu $\boldsymbol{X}$ dimenzije $n \times p$ provest ćemo na matrici $\boldsymbol{X}$ čiji su elementi $[\boldsymbol{X}]_{ij} = (\tilde{x}_{ij} - \overline{x}_j)$.

Iz (2.23) i (2.24) znamo da je spektralna dekompozicija matrice $\boldsymbol{X}^T\boldsymbol{X}$:

$$(n-1)\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X} = l_1^* a_1 a_1^T + l_2^* a_2 a_2^T + \cdots + l_p^* a_p a_p^T \tag{2.27}$$

pri čemu je $l_i^* = (n-1)l_i$ za $i = 1, 2, \ldots, p$.

Iz pretpostavke (3) da je rang matrice $\boldsymbol{X}$ jednak r slijedi da je i rang matrice $\boldsymbol{X}^T\boldsymbol{X}$ jednak r. Stoga su zadnjih $(p-r)$ svojstvenih vrijednosti tih matrica jednake nuli. Nadalje, onda je i zadnjih $(p-r)$ izraza u (2.27) jednako nuli. Dakle, vrijedi sljedeće:

$$(n-1)\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X} = l_1^* a_1 a_1^T + l_2^* a_2 a_2^T + \cdots + l_r^* a_r a_r^T.$$

Definirajmo matricu $\boldsymbol{A}$ kao matricu dimenzija $p \times r$ čiji je k-ti stupac:

$$a_k^* = (n-1)a_k, \quad k = 1, 2, \ldots, r,$$

tj. a_k^* je svojstveni vektor svojstvene vrijednosti l_k^*.

Neka je $\boldsymbol{U}$ matrica dimenzija $n \times r$ čiji je k-ti stupac:

$$u_k = {l_k^*}^{-1/2} \boldsymbol{X} a_k^*, \quad k = 1, 2, \ldots, r.$$

Konačno, neka je $\boldsymbol{L}$ dijagonalna matricu dimenzija $r \times r$ čiji je k-ti dijagonalni element jednak $l_k^{-1/2^*}$.

Ovime smo definirali $\boldsymbol{U}$, $\boldsymbol{L}$ i $\boldsymbol{A}$ na način da zadovoljavaju uvjete (1) i (2), trebamo još pokazati da vrijedi:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{A}^T.$$

Dakle, imamo:

$$\begin{aligned}
\boldsymbol{U}\boldsymbol{L}\boldsymbol{A}^T &= \boldsymbol{U} \begin{bmatrix} {l_1^*}^{1/2} {a_1^*}^T \\ {l_2^*}^{1/2} {a_2^*}^T \\ \vdots \\ {l_r^*}^{1/2} {a_r^*}^T \end{bmatrix} \\
&= \sum_{k=1}^{r} {l_k^*}^{-1/2} \boldsymbol{X} {a_k^*}^{1/2} {l_k^*}^{1/2} {a_k^*}^T \\
&= \sum_{k=1}^{r} \boldsymbol{X} a_k^* {a_k^*}^T \\
&= \sum_{k=1}^{p} \boldsymbol{X} a_k^* {a_k^*}^T.
\end{aligned}$$

Zadnja jednakost slijedi iz toga što su a_k^* za $k = (r+1), (r+2), \ldots, p$ svojstveni vektori od $\boldsymbol{X}^T\boldsymbol{X}$ koji pripadaju svojstvenim vrijednostima jednakim nuli. Naime, vektor $\boldsymbol{X}a_k^* = Z_k$ je vektor skorova k-te glavne komponente, a $\boldsymbol{X}a_k^* = 0$ za $k = (r+1), (r+2), \ldots, p$, zbog centriranja stupaca matrice $\boldsymbol{X}$ i nul-varijance zadnjih $(p-r)$ glavnih komponenti.

Stoga je

$$\boldsymbol{U}\boldsymbol{L}\boldsymbol{A}^T = \boldsymbol{X} \sum_{k=1}^{p} a_k^* {a_k^*}^T = \boldsymbol{X}$$

jer je $p \times p$ matrica čiji su stupci a_k^* ortogonalna te su njeni redci ortonormirani.

Drugačiji pristup ovom dokazu se može pronaći u [23].

Dekompozicija matrice na svojstvene vrijednosti daje efikasan način za izračunavanje glavnih komponenti. Jasno je da pronalaskom matrica $\boldsymbol{U}$, $\boldsymbol{L}$ i $\boldsymbol{A}$ koje zadovoljavaju (2.26) dolazimo do svojstvenih vektora (iz matrice $\boldsymbol{A}$) i korijena svojstvenih vrijednosti (iz matrice $\boldsymbol{L}$) matrice $\boldsymbol{X}^T\boldsymbol{X}$, a time i do koeficijenata i standardnih devijacija glavnih komponenti za uzoračku matricu kovarijanci $\boldsymbol{S}$.

2.4.2 Distribucijski rezultati za glavne komponente

U ovom ćemo poglavlju navesti neke poznate distribucijske rezultate koji se tiču analize glavnih komponenti. Nedostaci rezultata koje ćemo navesti su da pretpostavljaju da skup originalnih varijabli ima višedimenzionalnu normalnu distribuciju, da vrijede asimptotski i da nisu jednostavni za izvesti.

Definicija 2.1. *Neka je $\tilde{\boldsymbol{X}}$ matrica n podataka iz $N_p(0, \boldsymbol{\Sigma})$ distribucije, dimenzija $n \times p$ i $\boldsymbol{M} = \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}$. Tada kažemo da $\boldsymbol{M}$ ima Wishartovu distribuciju s parametrom $\boldsymbol{\Sigma}$ i n stupnjeva slobode. Oznaka $\boldsymbol{M} \sim W_p(\boldsymbol{\Sigma}, n)$. U slučaju da je $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ kažemo da je distribucija u standardnoj formi.*

Teorem 2.2. *Neka je $\tilde{\boldsymbol{X}}$ matrica n podataka iz $N_p(0, \boldsymbol{\Sigma})$ distribucije dimenzija $n \times p$ s elementima $[\tilde{\boldsymbol{X}}]_{ij} = \tilde{x}_{ij}$ te neka je $\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_{ij}$, za $j = 1, \ldots, p$. Tada za uzoračku matricu kovarijanci $\boldsymbol{S}$ čiji su elementi*

$$[\boldsymbol{S}]_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{x}_{ij} - \overline{x}_j)(\tilde{x}_{ik} - \overline{x}_k), \;\; j = 1, \ldots, p, \; k = 1, \ldots, p$$

vrijedi

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

tj. $(n-1)\boldsymbol{S}$ ima Wishartovu distribuciju s parametrom $\boldsymbol{\Sigma}$ i stupnjem slobode $(n-1)$.

Dakle, iz ovog teorema slijedi da se proučavanje distribucijskih svojstava koeficijenata i varijance glavnih komponenti svodi na proučavanje svojstava svojstvenih vrijednosti i vektora od Wishartovih slučajnih varijabli. Detaljnija analiza Wishartove distribucije i dokaz ovog teorema se može se pronaći u [1] i [17].

Prije nego navedemo neke asimptotske rezultate za distribucije koje se tiču analize glavnih komponenti, podsjetit ćemo se nekih oznaka i uvesti još neke nove oznake.

Pretpostavimo da je dano n podataka za p-dimenzionalan slučajni vektor $X \sim N_p(0, \boldsymbol{\Sigma})$. Tada za njegovu uzoračku matricu kovarijanci vrijedi $(n-1)\boldsymbol{S} \sim W_p(\boldsymbol{\Sigma}, n-1)$. Neka su l_k svojstvene vrijednosti matrice $\boldsymbol{S}$ i a_k pripadajući svojstveni vektori, a λ_k svojstvene vrijednosti matrice $\boldsymbol{\Sigma}$ te α_k odgovarajući svojstveni vektori za $k = 1, 2, \ldots, p$. Nadalje, neka je l p-dimenzionalni vektor čiji su elementi l_k i λ p-dimenzionalan vektor čiji su elementi λ_k. Neka su j-ti elementi od a_k i α_k redom a_{kj} i α_{kj}. Pretpostavimo da su $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$, tj. da su svojstvene vrijednosti matrice $\boldsymbol{\Sigma}$ pozitivne i međusobno različite. Tada sljedeći rezultati vrijede asimptotski, tj. aproksimativni su za konačne uzorke:

1. svi l_k nezavisni su od svih a_k
2. l i a_k imaju zajedničku normalnu distribuciju
3. $$E[l] = \lambda, \; E[a_k] = \alpha_k, \;\; k = 1, 2, \ldots p. \tag{2.28}$$
4. $$Cov(l_k, l_{k'}) = \begin{cases} \dfrac{2\lambda_k^2}{n-1}, & k = k' \\ 0, & k \neq k'. \end{cases} \tag{2.29}$$

$$Cov(a_{kj}, a_{k'j'}) = \begin{cases} \dfrac{\lambda_k}{n-1} \displaystyle\sum_{\substack{l=1 \\ l \neq k}}^{p} \dfrac{\lambda_j \alpha_{lj} \alpha_{lj'}}{(\lambda_l - \lambda_k)^2}, & k = k' \\ -\dfrac{\lambda_k \lambda_{k'} \alpha_{kj} \alpha_{k'j'}}{(n-1)(\lambda_k - \lambda_k')^2}, & k \neq k'. \end{cases} \tag{2.30}$$

2.4.3 Točkovna i intervalna procjena parametara

U ovom ćemo dijelu kratko opisati procjenu parametara glavnih komponenti. Prvo ćemo se osvrnuti na točkovnu procjenu parametara, a zatim na intervalnu procjenu gdje će nam od koristi biti rezultati iz prethodnog poglavlja. Oznake koje koristimo jednake su onim iz prethodnog poglavlja, tj. Poglavlja 2.4.2.

Procjenitelj maksimalne vjerodostojnosti (engl. maximum likelihood estimator, MLE) za matricu kovarijanci multivarijatne normalne distribucije, $\boldsymbol{\Sigma}$, je $\frac{n-1}{n}\boldsymbol{S}$, vidi npr. [19].

Procjenitelji maksimalne vjerodostojnosti za λ i α_k mogu se izvesti iz vrijednosti MLE za $\boldsymbol{\Sigma}$. MLE za vektor svojstvenih vrijednosti matrice $\boldsymbol{\Sigma}$, $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_p]^T$, $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$, iznosi $\hat{\lambda} = \frac{n-1}{n}l$, pri čemu je $l = [l_1, l_2, \ldots, l_p]^T$ vektor svojstvenih vrijednosti matrice $\boldsymbol{S}$. Uz uvjet da su elementi od λ pozitivni i međusobno različiti, MLE za svojstvene vektore matrice $\boldsymbol{\Sigma}$, α_k, dan je s $\hat{\alpha}_k = a_k$, pri čemu su a_k svojstveni vektori matrice $\boldsymbol{S}$, $k = 1, 2, \ldots, p$. U slučaju da su neke svojstvene vrijednosti λ_k jednake onda je MLE jednak prosjeku pripadajućih l_k pomnožen s $\frac{n-1}{n}$. MLE za α_k koji pripadaju jednakim λ_k nisu jedinstveni: za dobivanje novog skupa MLE dovoljno je matricu čiji su stupci MLE za α_k dimenzija $p \times q$ pomnožiti s bilo kojom $q \times q$ ortogonalnom matricom, pri čemu je q kratnost svojstvenih vrijednosti. Za dokaz navedenih rezultata vidi [2] i [17].

Za točkovne procjene λ i α_k se najčešće uzimaju l i a_k.

Za konstruiranje intervalnih procjena za λ_k i α_k koristimo asimptotske distribucije od l_k i a_k o kojima je bila riječ ranije. Za l_k je marginalna distribucija iz (2.28) i (2.29):

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

tj.

$$\frac{l_k - \lambda_k}{\lambda_k[2/(n-1)]^{1/2}} \sim N(0,1).$$

Pouzdani interval pouzdanosti $(1-\alpha)$ za λ_k je onda sljedećeg oblika:

$$\frac{l_k}{[1+\tau z_{\alpha/2}]^{1/2}} < \lambda_k < \frac{l_k}{[1-\tau z_{\alpha/2}]^{1/2}}, \tag{2.31}$$

pri čemu je $\tau^2 = 2/(n-1)$, a $z_{\alpha/2}$ je $(1-\alpha/2)$ kvantil standardne normalne distribucije $N(0,1)$.

Za a_k iz (2.28) i (2.30) slijedi da je:

$$a_k \sim N(\alpha_k, T_k),$$

pri čemu je

$$T_k = \frac{\lambda_k}{(n-1)} \sum_{\substack{l=1 \\ l \neq k}}^{p} \frac{\lambda_l}{(\lambda_l - \lambda_k)^2} \alpha_l \alpha_l^T.$$

Matrica T_k ima rang $(p-1)$ jer je svojstvena vrijednost koja pripada svojstvenom vektoru α_k jednaka nuli. Može se pokazati da približno vrijedi:

$$(n-1)\alpha_k^T(l_k \boldsymbol{S}^{-1} + l_k^{-1}\boldsymbol{S} - 2I_p)\alpha_k \sim \chi^2_{(p-1)}. \tag{2.32}$$

Iz (2.32) slijedi da je područje pouzdanosti s pouzdanošću $(1-\alpha)$ oblika

$$(n-1)\alpha_k^T(l_k \boldsymbol{S}^{-1} + l_k^{-1}\boldsymbol{S} - 2I_p)\alpha_k \leq h_\alpha,$$

pri čemu h_α predstavlja $(1-\alpha)$ kvantil hi-kvadrat distribucije sa stupnjem slobode $(p-1)$.

2.4.4 Testiranje hipoteza

Rezultate koje smo koristili da bismo definirali pouzdane intervale za l_k i a_{kj} korisni su i kod konstruiranja test statistika za hipoteze. Primjerice, ukoliko želimo testirati:

$$H_0 : \lambda_k = \lambda_{k0}$$
$$H_1 : \lambda_k \neq \lambda_{k0},$$

prikladna test statistika je

$$\frac{l_k - \lambda_{k0}}{\tau\lambda_{k0}}$$

pri čemu je $\tau^2 = 2/(n-1)$ i l_k k-ta svojstvena vrijednost uzoračke matrice kovarijanci $\boldsymbol{S}$.

Navedena test statistika ima približno $N(0,1)$ distribuciju ukoliko je hipoteza H_0 istinita. Hipotezu H_0 odbacili bismo sa statističkom značajnosti α u slučaju:

$$\left|\frac{l_k - \lambda_{k0}}{\tau\lambda_k 0}\right| \geq z_{\alpha/2}.$$

Slično, možemo testirati hipoteze

$$H_0 : \alpha_k = \alpha_{k0}$$
$$H_1 : \alpha_k \neq \alpha_{k0}.$$

Hipotezu H_0 bismo odbacili na razini značajnosti α u slučaju da je:

$$(n-1)\alpha_{k0}^T(l_k\boldsymbol{S}^{-1} + l_k^{-1}\boldsymbol{S} - 2I_p)\alpha_{k0} \geq h_\alpha.$$

za h_α $(1-\alpha)$ kvantil hi-kvadrat distribucije sa stupnjem slobode $(p-1)$. Postoje različiti testovi koji se tiču oblika matrice $\boldsymbol{\Sigma}$ i njezinih svojstvenih vrijednosti te svojstvenih vektora. Najpoznatiji je Bartlettov test koji testira hipotezu $H_{0q} : \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_p$, tj. slučaj da su zadnjih $(p-q)$ svojstvenih vrijednosti jednake, nasuprot hipoteze H_{1q}, slučaj da postoje bar dvije različite svojstvene vrijednosti među zadnjih $(p-q)$ svojstvenih vrijednosti. Odluka o tome koliko glavnih komponenti treba zadržati može se odrediti testiranjem H_{0q} za različite vrijednosti q. Testna statistika za ove hipoteze se može pronaći ukoliko pretpostavimo da l_k imaju zajedničku normalnu distribuciju i konstruiramo kvocijent vjerodostojnosti. Testna statistika sljedećeg je oblika:

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

a $-2ln(Q)$ ima približno χ^2 distribuciju s ν stupnjeva slobode.

Računanje stupnjeva slobode netrivijalan je zadatak, no može se pokazati da je $\nu = \frac{1}{2}(p-q+2)(p-q-1)$, stoga je uz pretpostavku istinitosti hipoteze H_{0q}

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

pri čemu je:

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

2.5 Problem skaliranja u analizi glavnih komponenti

Važno je napomenuti da glavne komponente skupa varijabli ovise o mjernim jedinicama u kojima su iskazane te varijable. Primjerice, prisjetimo se baze podataka iz Primjera 2.1, na raspolaganju su nam 31 podatak o promjeru, visini i volumenu stabala crne trešnje. Konkretno, dane su realizacije slučajnog vektora $X = (X_1, X_2, X_3)$ pri čemu X_1 modelira promjer stabla u inčima, X_2 modelira visinu stabla u stopama i X_3 modelira volumen stabla u kubičnim stopama. Pretpostavimo da želimo te podatke imati u drugim mjernim jedinicama, tj. definiramo slučajni vektor $W = (W_1, W_2, W_3)$ pri čemu je W_1 slučajna varijabla koja modelira promjer stabla u centimetrima, W_2 modelira visinu stabla u metrima, a W_3 modelira volumen stabla u kubičnim metrima. Tada je

$$W = \boldsymbol{K}X,$$

pri čemu je $\boldsymbol{K}$ dijagonalna matrica oblika:

$$\boldsymbol{K} = \begin{bmatrix} 2.54 & 0 & 0 \\ 0 & 1/3.281 & 0 \\ 0 & 0 & 1/35.315 \end{bmatrix}.$$

Ako uzoračku matricu kovarijanci od X označimo sa $\boldsymbol{S}$ onda je matrica kovarijanci za novi vektor dana izrazom:

$$\boldsymbol{S}_w = \boldsymbol{K}\boldsymbol{S}\boldsymbol{K} \tag{2.33}$$

jer je $\boldsymbol{K}^T = \boldsymbol{K}$.

U nastavku ćemo govoriti o općenitom slučaju za X i W p-dimenzionalne slučajne vektore. Označimo svojstvene vrijednosti matrice $\boldsymbol{S}_w$ s λ_k^* i pripadne svojstvene vektore s α_k^*.

Postavlja se pitanje hoće li glavne komponente koje smo izračunali iz matrice $\boldsymbol{S}_w$, $\alpha_k^* W$, biti jednake glavnim komponentama koje smo izračunali iz $\boldsymbol{S}$, $\alpha_k X$, ukoliko vektor W izrazimo u terminima vektora X. Prema [5], može se pokazati da je odgovor u pravilu ne, osim u sljedeća dva slučaja:

- Svi dijagonalni elementi matrice $\boldsymbol{K}$ su jednaki, tj. vrijedi da je $\boldsymbol{K} = c\boldsymbol{I}$, pri čemu je c skalar. Ovo bi značilo da je svaka varijabla skalirana na isti način.
- Varijable čiji su pripadajući dijagonalni elementi međusobno različiti u matrici $\boldsymbol{K}$ su nekorelirane. U slučaju da su svi dijagonalni elementi matrice $\boldsymbol{K}$ različiti, glavne komponente se neće promijeniti samo ako je matrica $\boldsymbol{S}_w$ dijagonalna matrica, a tada nema smisla raditi analizu glavnih komponenti jer su sve varijable međusobno nekorelirane.

Dokaz. Za težine prve glavne komponente vektora X, α_1, vrijedi:

$$\boldsymbol{S}\alpha_1 = \lambda_1\alpha_1. \tag{2.34}$$

Za težine prve glavne komponente vektora W vrijedi:

$$\boldsymbol{S}_w\alpha_1^* = \lambda_1^*\alpha_1^*$$

tj. iz (2.33)

$$\boldsymbol{K}\boldsymbol{S}\boldsymbol{K}\alpha_1^* = \lambda_1^*\alpha_1^* \tag{2.35}$$

Množenjem izraza (2.34) s $\boldsymbol{K}$ daje:

$$\boldsymbol{K}\boldsymbol{S}\alpha_1 = \lambda_1\boldsymbol{K}\alpha_1.$$

Sada, ako je prva glavna komponenta od $\boldsymbol{S}_w$ jednaka prvoj glavnoj komponenti od $\boldsymbol{S}$ u terminima originalnih varijabli, tada $\alpha_1^{*T} W_1^* = \alpha_1^{*T}(\boldsymbol{K}X)$ treba biti razmjerno s $\alpha_1^T X$, tj. treba vrijediti:

$$\alpha_1^{*T}\boldsymbol{K} = c\alpha_1^T,$$

za neku konstantu c, ili ekvivalentno:

$$\alpha_1^* = c\boldsymbol{K}^{-1}\alpha_1. \tag{2.36}$$

Uvrštavanjem (2.36) u (2.35) dobivamo:

$$\boldsymbol{K}\boldsymbol{S}\alpha_1 = \lambda_1^*\boldsymbol{K}^{-1}\alpha_1$$

ili

$$(\lambda_1\boldsymbol{K}^2 - \lambda_1^*\boldsymbol{I})\alpha_1 = 0, \tag{2.37}$$

$$(\boldsymbol{K}^2 - \frac{\lambda_1^*}{\lambda_1}\boldsymbol{I})\alpha_1 = 0. \tag{2.38}$$

Sličan rezultat se može dobiti i za ostale glavne komponente.

Jednadžba (2.37) vrijedi ukoliko je $(\lambda_1\boldsymbol{K}^2 - \lambda_1^*\boldsymbol{I})$ nul-matrica ili ako je $\frac{\lambda_1^*}{\lambda_1}$ svojstvena vrijednost matrice $\boldsymbol{K}^2$ iz (2.38).

1. U slučaju da je $(\lambda_1\boldsymbol{K}^2 - \lambda_1^*\boldsymbol{I})$ nul-matrica onda je $\lambda_1\boldsymbol{K}^2 = \lambda_1^*\boldsymbol{I}$ te je $\boldsymbol{K}^2$ (a time i $\boldsymbol{K}$) proporcionalna jediničnoj matrici $\boldsymbol{I}$. U ovom slučaju će se svojstvene vrijednosti povećati za jednak omjer, a svojstveni vektori matrica $\boldsymbol{S}$ i $\boldsymbol{S}_w$ će biti jednaki.

2. Pogledajmo sada drugi slučaj, kada je $\frac{\lambda_1^*}{\lambda_1}$ svojstvena vrijednost matrice $\boldsymbol{K}^2$. Matrica $\boldsymbol{K}^2$ je dijagonalna te su njene svojstvene vrijednosti elementi koji se nalaze na dijagonali.

   - Ako su svi dijagonalni elementi matrice $\boldsymbol{K}^2$ (a time i matrice $\boldsymbol{K}$) različiti, svojstveni vektori od $\boldsymbol{K}^2$ su vektori standardne baze, a to su ujedno i svojstveni vektori matrice $\boldsymbol{S}$ jedino ako je matrica $\boldsymbol{S}$ dijagonalna te su u tom slučaju varijable nekorelirane.
   - Ukoliko su neki od dijagonalnih elemenata matrice $\boldsymbol{K}^2$ (a time i matrice $\boldsymbol{K}$) jednaki, svojstveni vektori tih elemenata mogu biti odabrani kao ortonormirani skup vektora iz prikladnog vektorskog potprostora. Iz jednadžbe (2.37) vidimo da možemo odabrati te vektore na način da su to ujedno i svojstveni vektori matrice $\boldsymbol{S}$.

     Varijable koje pripadaju jednakim svojstvenim vrijednostima matrice $\boldsymbol{K}$ mogu biti međusobno korelirane, ali moraju biti nekorelirane s ostalim varijablama.

□

Rezultat koji smo upravo pokazali nam govori da se glavne komponente mijenjaju ovisno o skali, tj. da one nisu jedinstvena karakteristika podataka.

Najveći nedostatak glavnih komponenti koje su izračunate na temelju matrice kovarijanci je upravo njihova osjetljivost na mjerne jedinice. Primjerice, ukoliko jedna varijabla ima puno veću varijancu od ostalih varijabli ta će varijabla dominirati u prvoj glavnoj komponenti. S druge strane, ukoliko skaliramo te varijable na način da sve imaju jednaku varijancu utjecaj varijable na prvu glavnu komponentu će se

promijeniti. Zbog ovoga, nema smisla provoditi analizu glavnih komponenti na matrici kovarijanci ukoliko varijable nemaju "približno slične" vrijednosti varijanci.

Ilustrativno, pretpostavimo da varijabla X_1 modelira neku dužinu koja može biti mjerena u centimetrima ili milimetrima, a X_2 neka modelira primjerice težinu u gramima. Neka je $\mathbf{\Sigma_1}$ matrica kovarijanci od $X = (X_1, X_2)$ u slučaju da se dužina mjeri u centimetrima, a $\mathbf{\Sigma_2}$ u slučaju da se dužina mjeri u milimetrima:

$$\mathbf{\Sigma_1} = \begin{bmatrix} 80 & 40 \\ 44 & 80 \end{bmatrix}, \quad \mathbf{\Sigma_2} = \begin{bmatrix} 8000 & 440 \\ 440 & 80 \end{bmatrix}.$$

Prva glavna komponenta za $\mathbf{\Sigma_1}$ je $0.707x_1 + 0.707x_2$, a za $\mathbf{\Sigma_2}$ je $0.998x_1 + 0.055x_2$. Promjenom mjerne jedinice drastično smo utjecali na glavnu komponentu. Glavna komponenta od $\mathbf{\Sigma_1}$ daje jednaku važnost varijablama X_1 i X_2, a glavna komponenta od $\mathbf{\Sigma_2}$ je gotovo u potpunosti dominirana varijablom X_1. Dakle, ukoliko postoje velike razlike u varijancama komponenti vektora X tada će varijable s većim varijancama dominirati u prvim glavnim komponentama.

Problem skaliranja najčešće se rješava na način da se analiza glavnih komponenti provodi na matrici korelacija umjesto na matrici kovarijanci. Stoga u praksi, glavne komponente češće definiramo kao

$$Z = \boldsymbol{A}^T X^*$$

pri čemu je $\boldsymbol{A}$ matrica čiji su stupci svojstveni vektori korelacijske matrice vektora X, a X^* je standardizirana verzija slučajnog vektora X.

Na ovaj način smo osigurali da su varijable u nekom smislu od jednake važnosti pri računanju glavnih komponenti.

Glavne komponente korelacijske matrice neće biti jednake glavnim komponentama matrice kovarijanci ukoliko standardizirane varijable izrazimo u terminima originalnih varijabli osim u specijalnim slučajevima koje smo naveli. Štoviše, glavne komponente koje smo izrazili u terminima originalnih varijabli neće biti niti ortogonalne. Ovo je posljedica toga što linearna kombinacija ortogonalnih pravaca u p-dimenzionalnom prostoru generalno neće dati ortogonalne pravce.

2.6 Glavne komponente u linearnoj regresiji

U ovom poglavlju ćemo se dotaknuti uporabe glavnih komponenti u linearnoj regresiji. U linearnoj regresiji često se javlja problem multikolinearnosti. Problem multikolinearnosti pojavljuje se kada postoje visoke korelacije između dva ili više prediktora što implicira da postoje linearne veze među njima. U tom slučaju, osim što je interpretacija modela otežana, varijance nekih procijenjenih koeficijenata za prediktore mogu biti vrlo velike što dovodi do loših procjena i numerički nestabilnih rješenja.

Jedna mogućnost za rješavanje problema multikolinearnosti je odabrati manji podskup originalnih varijabli koji ne sadrži multikolinearnost. Drugi pristup ovom problemu je umjesto originalnih prediktora koristiti glavne komponente. Glavne komponente su nekorelirane pa se među njima ne javlja problem multikolinearnosti, a uporabom glavnih komponenti možemo smanjiti i broj potrebnih prediktora za objašnjavanje zavisne varijable.

Ukoliko želimo imati model u terminima originalnih varijabli, iz procijenjenih regresijskih koeficijenata za glavne komponente možemo dobiti regresijske koeficijente za originalne prediktore. Ukoliko sve glavne komponente uključimo u regresijski model, model je ekvialentan modelu koji koristi originalne varijable, stoga su velike varijance regresijskih koeficijenata uzrokovane multikolinearnošću i dalje prisutne. S druge

strane, ako neke glavne komponente izostavimo iz regresijskog modela i izračunamo regresijske koeficijente za originalne varijable onda možemo dobiti regresijske koeficijente manje varijance. Nedostatak ove metode je da novi procjenitelj regresijskih koeficijenata za originalne varijable nije nepristran.

2.6.1 Model linearne regresije

Pretpostavimo da na temelju slučajnog vektora $X = (X_1, \ldots, X_p)$ (nezavisne varijable, prediktori) želimo zaključivati o varijabli Y (zavisna varijabla). Jedan od načina kako modelirati tu vezu je statističkim modelom linearne regresije:

$$Y = X^T\beta + \varepsilon$$

pri čemu je β vektor p nepoznatih parametara koje treba procijeniti i ε slučajna varijabla koja predstavlja grešku modela.

Uobičajene pretpostavke kod modela linearne regresije su:

1. $E[\varepsilon] = 0$, tj. $E[Y] = E[X^T\beta]$.
2. $E[\varepsilon|X] = E[\varepsilon]$.

Procjenu za β možemo dobiti na temelju realizacija $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$, $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{ip}]^T$, $i = 1, 2, \ldots, n$ slučajnog uzorka $((Y^1, X_1^1, X_2^1, \ldots, X_p^1), \ldots, (Y^n, X_1^n, X_2^n, \ldots, X_p^n))$, metodom najmanjih kvadrata. Pišemo:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i, \quad i = 1, 2, \ldots, n.$$

Ideja metode najmanjih kvadrata je minimizirati vrijednost:

$$S(\beta) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T\beta).$$

Rješenje tog minimizacijskog problema je:

$$\hat{\beta} = \sum_{i=1}^{n} (\boldsymbol{x}_i\boldsymbol{x}_i^T)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i y_i. \tag{2.39}$$

Izvod ovog rezultata i detaljnija analiza modela linearne regresije se može pronaći u [15].

Uz oznake $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^T$ i $\boldsymbol{X} = [\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_n^T]$ izraz (2.39) možemo matrično zapisati kao:

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

2.6.2 Definicija modela linearne regresije preko glavnih komponenti

Pretpostavimo da imamo n sparenih nezavisnih podataka za varijablu Y i p-dimenzionalan vektor X, $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$, $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{ip}]^T$ za $i = 1, 2 \ldots, n$. Bez smanjenja općenitosti, pretpostavit ćemo da su podaci centrirani.

Neka je dan standardni regresijski model za predviđanje Y na temelju prediktora $X_1, X_2, \ldots, X_p$:

$$\boldsymbol{y} = \boldsymbol{X}\beta + e \tag{2.40}$$

gdje je $\boldsymbol{y}$ vektor n podataka zavisne varijable Y, $\boldsymbol{X}$ matrica dimenzija $n \times p$ za koju je $[\boldsymbol{X}]_{ij} = x_{ij}$, β vektor p regresijskih koeficijenata i e vektor grešaka.

Vrijednosti glavnih komponenti za svaki podatak dani su s:

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A} \tag{2.41}$$

gdje je $[\boldsymbol{Z}]_{ik}$ vrijednost k-te glavne komponente, Z_k, za i-ti podatak, $\boldsymbol{x}_i$, a $\boldsymbol{A}$ matrica dimenzija $p \times p$ čiji je k-ti stupac k-ti svojstveni vektor od $\boldsymbol{X}^T\boldsymbol{X}$.

Kako je matrica $\boldsymbol{A}$ ortogonalna, $\boldsymbol{X}\beta$ možemo zapisati kao:

$$\boldsymbol{X}\beta = \boldsymbol{X}(\boldsymbol{A}\boldsymbol{A}^T)\beta = (\boldsymbol{X}\boldsymbol{A})(\boldsymbol{A}^T\beta) \overset{(2.41)}{=} \boldsymbol{Z}\gamma, \text{ za } \gamma = \boldsymbol{A}^T\beta. \tag{2.42}$$

Izraz (2.40) onda možemo zapisati kao:

$$\boldsymbol{y} = \boldsymbol{Z}\gamma + e, \tag{2.43}$$

čime smo zamijenili prediktore s njihovim glavnim komponentama u regresijskom modelu.

Regresija pomoću glavnih komponenti može biti definirana kao model (2.43) ili kao reducirani model:

$$y = \boldsymbol{Z}_m\gamma_m + e_m \tag{2.44}$$

pri čemu je γ_m vektor m elemenata vektora γ, $\boldsymbol{Z}_m$ matrica čiji su stupci odgovarajući podskup stupaca matrice $\boldsymbol{Z}$, a e_m odgovarajuća greška.

Procijeniti γ metodom najmanjih kvadrata u (2.43) i onda pronaći procjenu za β iz

$$\hat{\beta} = \boldsymbol{A}\hat{\gamma} \tag{2.45}$$

jednako je kao izračunati $\hat{\beta}$ metodom najmanjih kvadrata direktno iz (2.40). No, računanje $\hat{\gamma}$ iz (2.43) jednostavnije je od računanja $\hat{\beta}$ iz (2.40) jer su stupci matrice $\boldsymbol{Z}$ ortogonalni. Vektor $\hat{\gamma}$ je:

$$\hat{\gamma} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T y = (\boldsymbol{X}\boldsymbol{A})^T\boldsymbol{L}^{-2}\boldsymbol{Z}^T\boldsymbol{y} \tag{2.46}$$

pri čemu je $\boldsymbol{L}$ dijagonalna matrica čiji je k-ti dijagonalni element jednak $l_k^{1/2}$, a l_k k-ta najveća svojstvena vrijednost od $\boldsymbol{X}^T\boldsymbol{X}$.

Pogledajmo sada jedan rezultat vezan za glavne komponente u regresiji:

Svojstvo 8. *Uz oznake iz ovog poglavlja, neka je dana regresijska jednadžba (2.40):*

$$\boldsymbol{y} = \boldsymbol{X}\beta + e$$

Pretpostavimo da na $\boldsymbol{X}$ djeluje $\boldsymbol{B}$ tj. $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B}$ pri čemu je $\boldsymbol{B}$ ortogonalna matrica dimenzija $p \times p$. Regresijsku jednadžbu u tom slučaju možemo zapisati kao:

$$\boldsymbol{y} = \boldsymbol{Z}\gamma + e$$

pri čemu je $\gamma = \boldsymbol{B}^{-1}\beta$, vidi (2.42). Procjenitelj dobiven metodom najmanjih kvadrata za γ dan je s $\hat{\gamma} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}$. Pod ovim pretpostavkama, elementi $\hat{\gamma}$ imaju redom najmanju moguću varijancu kada je $\boldsymbol{B}$ matrica čiji je k-ti stupac jednak svojstvenom vektoru matrice $\boldsymbol{X}^T\boldsymbol{X}$ pa time i k-ti svojstveni vektor matrice S. Tada je $\boldsymbol{Z}$ sastavljen od vrijednosti uzoračkih glavnih komponenti za $\boldsymbol{X}$.

Dokaz. Za matricu kovarijanci procjenitelja najmanjih kvadrata $\hat{\gamma}$ vrijedi:

$$\begin{aligned}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} &= (\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B})^{-1} \\ &= \boldsymbol{B}^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{B}^T)^{-1} \\ &= \boldsymbol{B}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{B},\end{aligned}$$

jer je $\boldsymbol{B}$ ortogonalna.

Dakle, trebamo minimizirati $tr(\boldsymbol{B}_q^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{B}_q)$, za $q = 1, 2, \ldots, p$ pri čemu se $\boldsymbol{B}_q$ sastoji od prvih q stupaca matrice $\boldsymbol{B}$. Zamijenimo li $\boldsymbol{\Sigma}$ s $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ u Svojstvu 2. iz Poglavlja 2.2, pokazuje se da je $\boldsymbol{B}_q$ matrica koja se sastoji od zadnjih q stupaca matrice čiji je k-ti stupac k-ti svojstveni vektor od $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Nadalje, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ ima jednake svojstvene vektore kao i $\boldsymbol{X}^T\boldsymbol{X}$ osim što im je redoslijed zamijenjen. Stoga zaključujemo da $\boldsymbol{B}_q$ mora imati stupce koji su jednaki prvim q svojstvenim vektorima matrice $\boldsymbol{X}^T\boldsymbol{X}$. Ovo vrijedi za $q = 1, 2, \ldots, p$ pa je svojstvo dokazano. □

Ovo svojstvo govori zašto ima smisla zamijeniti originalne prediktore s prvih nekoliko glavnih komponenti. Naime, posljednje glavne komponente koje bismo izostavili u modelu su one čiji procijenjeni koeficijenti imaju veliku varijancu. Nedostatak ovog svojstva je da nam ne govori ništa o vezi između zavisne varijable i glavnih komponenti. Velika varijanca za koeficijent γ_k ne isključuje postojanje jake veze između k-te glavne komponente z_k i zavisne varijable y.

Jedna od prednosti korištenja glavnih komponenti umjesto originalnih prediktora je lakša interpretacija utjecaja glavnih komponenti na zavisnu varijablu. Naime, zbog nekoreliranosti glavnih komponenti, utjecaj na zavisnu varijablu i procijenjeni koeficijent za glavnu komponentu neovisan je o drugim glavnim komponentama. S druge strane, kod originalnih varijabli se utjecaj na zavisnu varijablu i procijenjeni koeficijent mogu drastično promijeniti ukoliko se neke od varijabli uklone ili dodaju.

Treba napomenuti da unatoč tome što je interpretacija pojedinačnih utjecaja jednostavnija ukoliko koristimo glavne komponente, otežana je ako glavne komponente nemaju jasno značenje.

Glavna prednost korištenja glavnih komponenti u regresiji je u slučaju da je prisutna multikolinearnost originalnih varijabli. Tada, uklanjanjem podskupa glavnih komponenti koje imaju malu varijancu možemo dobiti stabilniju procjenu za β. Da bismo se uvjerili u to, najprije ćemo uvrstiti (2.46) u (2.45):

$$\begin{aligned}\hat{\beta} &= \boldsymbol{A}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y} \\ &= \boldsymbol{A}\boldsymbol{L}^{-2}\boldsymbol{Z}^T\boldsymbol{y} \\ &= \boldsymbol{A}\boldsymbol{L}^{-2}\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{y} \\ &= \sum_{k=1}^{p} l_k^{-1} a_k a_k^T \boldsymbol{X}^T\boldsymbol{y} \end{aligned} \tag{2.47}$$

za l_k k-ti dijagonalni element od $\boldsymbol{L}^2$ i a_k k-ti stupac matrice $\boldsymbol{A}$.

Pretpostavljamo da je $\boldsymbol{y}$ vektor nezavisnih realizacija slučajne varijable Y čija je varijanca σ^2, tj. pretpostavljamo da su y_i nezavisni i dolaze iz iste distribucije. Za matricu kovarijanci od $\hat{\beta}$ vrijedi:

$$\begin{aligned}\sigma^2\left[\boldsymbol{A}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\right]\left[\boldsymbol{A}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\right]^T &= \sigma^2\left[\boldsymbol{A}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\right]\left[\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{A}^T\right] \\ &= \sigma^2\boldsymbol{A}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{A}^T \\ &= \sigma^2\boldsymbol{A}L^{-2}\boldsymbol{A}^T \\ &= \sigma^2\sum_{k=1}^{p} l_k^{-1} a_k a_k^T. \end{aligned} \tag{2.48}$$

Ovaj rezultat objašnjava kako multikolinearnost daje velike varijance za elemente vektora $\hat{\beta}$. Ako je multikolinearnost prisutna onda se ona pojavljuje kao glavna komponenti male varijance. Drugim riječima, posljednje glavne komponente imaju male vrijednosti za l_k i time velike vrijednosti za l_k^{-1}.

Jedan način na koji možemo smanjiti taj efekt je brisanjem izraza koji odgovaraju malim vrijednostima l_k iz (2.47). Na taj način dobivamo sljedeći procjenitelj:

$$\tilde{\beta} = \sum_{k=1}^{m} l_k^{-1} a_k a_k^T \boldsymbol{X}^T \boldsymbol{y}$$

gdje su $l_{m+1}, l_{m+2}, \ldots, l_p$ svojstvene vrijednosti male vrijednosti.

Tada je matrica kovarijanci $V(\tilde{\beta})$ za $\tilde{\beta}$ jednaka:

$$\sigma^2 \sum_{j=1}^{m} l_j^{-1} a_j a_j^T \boldsymbol{X}^T \boldsymbol{X} \sum_{k=1}^{m} l_k^{-1} a_k a_k^T.$$

Uvrstimo li

$$\boldsymbol{X}^T \boldsymbol{X} = \sum_{h=1}^{p} l_h a_h a_h^T$$

dobivamo:

$$V(\tilde{\beta}) = \sigma^2 \sum_{h=1}^{p} \sum_{j=1}^{m} \sum_{k=1}^{m} l_h l_j^{-1} l_k^{-1} a_j a_j^T a_h a_h^T a_k a_k^T.$$

Vektori a_h, $h = 1, 2, \ldots, p$ su ortonormirani, stoga su jedini ne-nul elementi u trostrukoj sumaciji za $h = j = k$, dakle:

$$V(\tilde{\beta}) = \sigma^2 \sum_{k=1}^{m} l_k^{-1} a_k a_k^T.$$

Procjenitelj $\tilde{\beta}$ ima manju varijancu od procjenitelja $\hat{\beta}$, ali nije nepristran. Ovo slijedi iz:

$$\tilde{\beta} = \hat{\beta} - \sum_{k=m+1}^{p} l_k^{-1} a_k a_k^T \boldsymbol{X}^T \boldsymbol{y}, \quad E[\hat{\beta}] = \beta$$

i

$$\begin{aligned} E\left[\sum_{k=m+1}^{p} l_k^{-1} a_k a_k^T \boldsymbol{X}^T y\right] &= \sum_{k=m+1}^{p} l_k^{-1} a_k a_k^T \boldsymbol{X}^T \boldsymbol{X} \beta \\ &= \sum_{k=m+1}^{p} a_k a_k^T \beta. \end{aligned}$$

Zadnji izraz u pravilu nije jednak nula, stoga je $E[\tilde{\beta}] \neq \beta$.

Ukoliko je multikolinearnost velik problem, smanjenje u varijanci će biti značajno, a pristranost može biti vrlo mala. Štoviše, u slučaju da su elementi vektora γ koji odgovaraju uklonjenim glavnim komponentama zaista jednaki nuli onda će procjenitelj biti nepristran.

Regresija provedena na glavnim komponentama u (2.43) i (2.44) ekvivalentna je regresiji (2.40) u slučaju da β procjenjujemo s:

$$\tilde{\beta} = \sum_{M} l_k^{-1} a_k a_k^T \boldsymbol{X}^T \boldsymbol{y} \tag{2.49}$$

pri čemu je M neki podskup od $\{1, 2 \ldots, p\}$.

2.6.3 Odabir glavnih komponenti u linearnoj regresiji

Kod odabira skupa M u (2.49) u obzir treba uzeti dvije stvari: za eliminaciju velikih varijanci uzrokovanih multikolinearnošću treba ukloniti glavne komponente čije su varijance male, a nepoželjno je brisati komponente koje su jako korelirane sa zavisnom varijablom Y.

Jedna strategija za odabir M je obrisati sve komponente čija je varijanca manja od l^*, gdje je l^* neka proizvoljna granična vrijednost.

Drugi način za odabir skupa M je promatrati vektor inflacije varijance (engl. variance inflation factors, VIF) prediktora. VIF j-te standardizirane varijable definira se kao c_{jj}/σ^2 pri čemu je c_{jj} varijanca j-tog elementa procjenitelja $\hat{\beta}$ dobivenog metodom najmanjih kvadrata. Ako su sve varijable nekorelirane onda su svi VIF-ovi jednaki 1. U slučaju da je prisutna velika multikolinearnost onda će VIF-ovi za elemente $\hat{\beta}$ biti visoki za one varijable koje stvaraju multikolinearnost. Stoga, primjerice, ukoliko uklonimo posljednjih nekoliko izraza iz (2.47) VIF-ovi novog pristranog procjenitelja će se smanjiti. Uklanjanje izraza nastavljamo dok ne dobijemo željenu razinu VIF-ova.

Originalni VIF za varijablu povezan je s koeficijentom determinacije R^2 između te varijable i drugih $(p-1)$ varijabli izrazom $VIF = (1 - R^2)^{-1}$. Vrijednost $VIF > 10$ odgovara vrijednosti $R^2 > 0.90$, a $VIF > 4$ je ekvivalentno $R^2 > 0.75$, stoga možemo odabrati određenu graničnu vrijednost za R^2 koja posljedično utječe na veličinu VIF-a.

Nedostatak ove metode, uklanjanja glavnih komponenti s niskom vrijednosti varijance, je što niska varijanca glavne komponente ne implicira da je ona nevažna u regresijskom modelu. Stoga, zadržavanje komponenti koje imaju visoku prediktivnu moć i uklanjanja komponenti niske varijance ponekad nije moguće istovremeno provesti.

Selekcija varijabli se može provesti i na način da se gleda njihov doprinos zavisnoj varijabli u regresijskom modelu. Dakle, selekcija se u tom slučaju bazira na vrijednosti t-statistike koja mjeri (nezavisni) doprinos svake glavne komponente u regresijskoj jednadžbi. Nedostatak ove strategije je što t-test za glavne komponente čija je varijanca male vrijednosti nije jednako dobar kao za komponente veće varijance pa je stoga manje vjerojatno da će biti odabrane. Kompromis između ova dva tipa selekcije je uklanjati varijable manjih varijanci čije su t-vrijednosti statistički neznačajne.

Drugačiji pristup regresiji koji također koristi glavne komponente je latentna regresija (engl. latent root regression). Glavna razlika između regresije pomoću glavnih komponenti i latentne regresije je da se u latentnoj regresiji izračun glavnih komponenti provodi na $(p+1)$ varijabli, na p prediktora i zavisnoj varijabli. Ideja je analizirati glavne komponente čije su svojstvene vrijednosti male. U slučaju da je njihov koeficijent za zavisnu varijablu također male vrijednosti onda ih kategoriziramo kao neprediktivna multikolinearnost te ih uklanjamo iz modela. Više o tome se može pročitati u [13].

3 Faktorska analiza

Faktorska analiza statistička je metoda koja, slično kao i analiza glavnih komponenti, ima za cilj smanjiti dimenzionalnost skupa varijabli. U faktorskoj analizi varijable $X_1, X_1, \ldots, X_p$ želimo prikazati kao linearnu kombinaciju slučajnih varijabli $f_1, f_2, \ldots, f_m$ koje zovemo faktori, $m < p$. Ukoliko postoje velike koreliranosti između varijabli $X_1, X_2, \ldots, X_p$ tada se dimenzionalnost prostora može smanjiti uz uporabu faktora.

U nastavku ćemo definirati model faktorske analize, objasniti kao procjenjujemo parametre modela i napraviti usporedbu faktorske analize i analize glavnih komponenti.

3.1 Defincija modela i pretpostavke

Osnovna ideja faktorske analize je da p varijabli koje opažamo, $X_1, X_2, \ldots, X_p$, mogu biti iskazane kao linearne funkcije $m < p$ teoretskih neopaženih varijabli (faktora) uz grešku. Za varijable $X_1, X_2, \ldots, X_p$ i faktore $f_1, f_2, \ldots, f_m$ definiramo faktorski model:

$$\begin{aligned} X_1 - \mu_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1m} f_m + \varepsilon_1 \\ X_2 - \mu_2 &= \lambda_{21} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1m} f_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{pm} f_m + \varepsilon_p, \end{aligned} \tag{3.1}$$

pri čemu su λ_{ij} konstantne koje ćemo zvati faktorski koeficijenti (engl. factor loadings), $\mu_i = E[X_i]$ su očekivanja, a ε_i su greške, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, m$.

Koeficijent λ_{ij} govori o važnosti j-tog faktora f_j za i-tu varijablu X_i te se može koristiti u interpretaciji od f_j. Primjerice, faktor f_2 možemo opisati ili interpretirati na osnovu koeficijenata $\lambda_{12}, \lambda_{22}, \ldots, \lambda_{p2}$. Greške ε_i ponekad zovemo i specifični faktori (engl. specific factors) jer ε_i vežemo uz X_i.

Model (3.1) možemo zapisati matrično na sljedeći način:

$$X - \mu = \mathbf{\Lambda} f + \varepsilon, \tag{3.2}$$

pri čemu je $X = (X_1, X_2, \ldots, X_p)$, $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)$, $f = (f_1, f_2, \ldots, f_m)$ i $[\mathbf{\Lambda}]_{ij} = \lambda_{ij}$ za $i = 1, 2, \ldots, p$ i $j = 1, 2, \ldots, m$.

Sljedeće pretpostavke su uobičajene kod faktorske analize:

1. $E[\varepsilon] = 0, \;\; E[f] = 0,$
2. $E[\varepsilon\varepsilon^T] = \mathbf{\Phi}, \;\; E[f\varepsilon^T] = \mathbf{0}, \;\; E[ff^T] = \boldsymbol{I}_m,$

pri čemu je u pretpostavci 2. $\mathbf{\Phi}$ dijagonalna matrica, $\mathbf{0}$ matrica nula i $\boldsymbol{I}_m$ jedinična matrica.

Pretpostavka $E[\varepsilon\varepsilon^T] = \mathbf{\Phi}$ govori da su greške međusobno nekorelirane što je osnovna pretpostavka faktorskog modela. Ovom pretpostavkom smo konstatirali da je sav X koji se može pripisati nekim zajedničkim utjecajima sadržan u $\mathbf{\Lambda} f$. Nadalje, pretpostavka $E[f\varepsilon^T] = 0$ govori da su zajednički faktori nekorelirani s greškama što je također jedna od osnovnih pretpostavki faktorskog modela. Zadnja navedena pretpostavka $E[ff^T] = \boldsymbol{I}_m$, govori da su zajednički faktori međusobno nekorelirani, no ova pretpostavka nije nužna za faktorski model.

Uz pretpostavke koje smo dosad naveli, za matricu kovarijanci $\mathbf{\Sigma}$ vektora X vrijedi:

$$\begin{aligned}\mathbf{\Sigma} = Cov(X) &= Cov(\mathbf{\Lambda} f + \varepsilon) \\ &= Cov(\mathbf{\Lambda} f) + Cov(\varepsilon) \\ &= \mathbf{\Lambda} \boldsymbol{I} \mathbf{\Lambda}^T + \mathbf{\Phi} \\ &= \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Phi}. \end{aligned} \tag{3.3}$$

Iz (3.3) varijancu od X_i možemo zapisati na sljedeći način:

$$Var(X_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \phi_i. \tag{3.4}$$

Rastavili smo varijancu od X_i na dva dijela: varijanca koja je prisutna zbog zajedničkih faktora koju ćemo zvati zajednička varijanca (engl. communality ili common variance) i varijanca koja je jedinstvena za svaki od X_i koju ćemo zvati specifična varijanca (ϕ_i) (engl. specificity, unique variance, residual variance). Zajedničku varijancu $\lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2$ ćemo u nastavku označavati s h_i^2.

Iz (3.3) vidimo da su kovarijance vektora X modelirane samo s elementima matrice $\mathbf{\Lambda}$ jer je $\mathbf{\Phi}$ dijagonalna matrica. Primjerice, ako je $m = 2$ onda imamo model:

$$\begin{aligned} X_1 - \mu_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \varepsilon_1 \\ X_1 - \mu_2 &= \lambda_{21} f_1 + \lambda_{12} f_2 + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \varepsilon_p. \end{aligned}$$

Matrica $\mathbf{\Lambda}$ je tada sljedećeg oblika:

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \vdots & \vdots \\ \lambda_{p1} & \lambda_{p2} \end{bmatrix}$$

Kovarijanca od npr. X_1 i X_2 iz (3.3) je tada:

$$Cov(X_1, X_2) = \lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22}.$$

Ako su X_1 i X_2 slične, tj. ako imaju puno zajedničkih karakteristika, imat će slične koeficijente uz f_1 i f_2, tj. redci $(\lambda_{11}, \lambda_{12})$ i $(\lambda_{21}, \lambda_{22})$ će biti slični. U tom je slučaju vjerojatno da će $\lambda_{11}\lambda_{21}$ ili $\lambda_{12}\lambda_{22}$ imati veliku vrijednost.

Osim toga, možemo i kovarijance između X_i i f_j izraziti u terminima λ_{ij}. Pogledajmo primjerice, $Cov(X_1, f_2)$:

$$\begin{aligned} Cov(X_1, f_2) &= E[(X_1 - \mu_1)(f_2 - \mu_{f_2})] \\ &= E[(\lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1m} f_m + \varepsilon_1) f_2] \\ &= E[\lambda_{11} f_1 f_2 + \lambda_{12} f_2^2 + \cdots + \lambda_{1m} f_m f_2 + \varepsilon_1 f_2] \\ &= \lambda_{11} Cov(f_1, f_2) + \lambda_{12} Var(f_2) + \cdots + \lambda_{1m} Cov(f_m, f_2) \\ &= \lambda_{12} \end{aligned}$$

jer je $\mu_{f_2} = 0$, f_2 nekorelirano s ostalim f_j i $Var(f_2) = 1$. Dakle, težine λ_{ij} predstavljaju kovarijancu varijabli s faktorima. Generalno,

$$Cov(X_i, f_j) = \lambda_{ij}, \quad i = 1, 2, \ldots, p, j = 1, 2, \ldots, m,$$

a to možemo matrično zapisati kao:

$$Cov(X, f) = \mathbf{\Lambda}.$$

Rijetke se matrice kovarijanci $\mathbf{\Sigma}$ mogu izraziti egzaktno kao $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Phi}$ gdje je $\mathbf{\Phi}$ dijagonalna matrica, a $\mathbf{\Lambda}$ dimenzija $p \times m$ pri čemu je $m << p$. No, držimo se pretpostavke da je $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Phi}$ jer je ona vrlo bitna za procjenu matrice $\mathbf{\Lambda}$.

Jedna prednost faktorskog modela je da je iz $\mathbf{\Lambda}$ odmah jasno odgovara li taj model podacima. U slučaju da takav model ne odgovara podacima javljaju se dva problema u procjeni:

1. nije jasno koliko treba biti faktora,

2. nije jasno što su faktori.

3.2 Nejedinstvenost faktorskih koeficijenata

Težine $\mathbf{\Lambda}$ u modelu (3.2) nisu jedinstveno određene. Naime, one mogu biti pomnožene s ortogonalnom matricom $\boldsymbol{T}$ bez da se naruši jednakost $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Phi}$. Zbog ovoga, nakon prvotne procjene faktorskog modela se istražuju i druga moguća rješenja za $\mathbf{\Lambda}$ koja mogu biti od koristi pri interpretaciji modela.

Neka matrice $\mathbf{\Lambda}$ i $\mathbf{\Phi}$ zadovoljavaju jednadžbu (3.3) te neka je $\boldsymbol{T}$ ortogonalna matrica. Zbog ortogonalnosti matrice $\boldsymbol{T}$ vrijedi $\boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{I}$ te $\boldsymbol{T}\boldsymbol{T}^T$ možemo uvrstiti u model (3.2) na sljedeći način:

$$X - \mu = \mathbf{\Lambda}\boldsymbol{T}\boldsymbol{T}^T f + \varepsilon.$$

Na taj način smo dobili sljedeći model:

$$X - \mu = \mathbf{\Lambda}^* f^* + \varepsilon,$$

pri čemu je

$$\begin{aligned} \mathbf{\Lambda}^* &= \mathbf{\Lambda}\boldsymbol{T}, \\ f^* &= \boldsymbol{T}^T f. \end{aligned}$$

Ako umjesto $\mathbf{\Lambda}$ uvrstimo $\mathbf{\Lambda}^* = \mathbf{\Lambda}\boldsymbol{T}$ u $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Phi}$ dobivamo sljedeće:

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{\Lambda}^* \mathbf{\Lambda}^{*T} + \mathbf{\Phi} \\ &= \mathbf{\Lambda}\boldsymbol{T}(\mathbf{\Lambda}\boldsymbol{T})^T + \mathbf{\Phi} \\ &= \mathbf{\Lambda}\boldsymbol{T}\boldsymbol{T}^T\mathbf{\Lambda} + \mathbf{\Phi} \\ &= \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Phi}, \end{aligned}$$

jer je $\boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{I}$. Stoga nove težine $\mathbf{\Lambda}^* = \mathbf{\Lambda}\boldsymbol{T}$ također reproduciraju $\mathbf{\Sigma}$ kao i $\mathbf{\Lambda}$:

$$\mathbf{\Sigma} = \mathbf{\Lambda}^* \mathbf{\Lambda}^{*T} + \mathbf{\Phi} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Phi}.$$

Faktori $f^* = \boldsymbol{T}^T f$ također zadovoljavaju pretpostavke (1) i (2), tj.: $E[f^*] = 0$, $Cov(f^*) = \mathbf{0}$ i $Cov(f^*, e) = \mathbf{0}$.

Zajednička varijanca h_i^2, za $i = 1, 2, \ldots, p$ također ostaje nepromijenjena transformacijom $\mathbf{\Lambda}^* = \mathbf{\Lambda}\boldsymbol{T}$. Zajednička varijanca h_i^2 je suma kvadrata i-tog retka matrice $\mathbf{\Lambda}$. Ako označimo i-ti redak matrice $\mathbf{\Lambda}$ s λ_i^T onda je suma kvadrata u vektorskoj notaciji jednaka $h_i^2 = \lambda_i^T \lambda_i$. i-ti redak matrice $\mathbf{\Lambda}^*$ jednak je ${\lambda_i}^{*T} = \lambda_i^T \boldsymbol{T}$, a pripadajuća zajednička varijanca je:

$$h_i^{*2} = \lambda_i^{*T} \lambda_i^* = \lambda_i^T \boldsymbol{T}\boldsymbol{T}^T \lambda_i = \lambda_i^T \lambda_i = h_i^2.$$

3.3 Procjena parametara faktorskog modela

Kod procjene parametara modela, uobičajeno se najprije procjenjuju parametri $\mathbf{\Lambda}$ i $\mathbf{\Phi}$, a procjene za faktore f nalazimo kasnije. Opisat ćemo dvije metode za procjenu faktorskih koeficijenata $\mathbf{\Lambda}$ i matrice $\mathbf{\Phi}$. Konkretno, opisat ćemo metodu zvanu metoda glavnih komponenti i metodu maksimalne vjerodostojnosti. Opis ovih i drugih metoda za procjenu tih matrica mogu se pronaći u [20]. Prvo ćemo opisati metodu glavnih komponenti za procjenu parametara u faktorskom modelu. Treba naglasiti da metoda glavnih komponenti nije povezana s analizom glavnih komponenti, a u nastavku ćemo komentirati zašto metoda nosi ovo ime.

U praksi nam nije poznata matrica kovarijanci nego ju procjenjujemo. Dakle, neka su $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ realizacije slučajnog uzorka $((X_1^1, X_2^1, \ldots, X_p^1), \ldots, (X_1^n, X_2^n, \ldots, X_p^n))$ pomoću kojeg smo izračunali uzoračku matricu kovarijanci $\boldsymbol{S}$. Trebamo pronaći procjenu za $\mathbf{\Lambda}$, $\hat{\mathbf{\Lambda}}$, koja aproksimira model (3.3) pri čemu je $\mathbf{\Sigma}$ procijenjena sa $\boldsymbol{S}$:

$$\boldsymbol{S} \cong \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Phi}}.$$

Kod metode glavnih komponenti prvotno zanemarujemo $\hat{\mathbf{\Phi}}$ i faktoriziramo $\boldsymbol{S}$ kao $\boldsymbol{S} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T$. Kako bismo faktorizirali $\boldsymbol{S}$ koristimo spektralnu dekompoziciju:

$$\boldsymbol{S} = \boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^T$$

pri čemu je $\boldsymbol{C}$ ortogonalna matrica čiji su stupci normirani svojstveni vektori, a $\boldsymbol{D}$ je dijagonalna matrica dimenzija $p \times p$ na čijoj su dijagonali svojstvene vrijednosti matrice $\boldsymbol{S}$:

$$\boldsymbol{D} = diag(\theta_1, \theta_2, \ldots, \theta_p).$$

Kako je $\boldsymbol{S}$ pozitivno semidefinitna matrica sve njene svojstvene vrijednosti su veće ili jednake od nula, stoga $\boldsymbol{D}$ možemo zapisati na sljedeći način:

$$\boldsymbol{D} = \boldsymbol{D}^{1/2}\boldsymbol{D}^{1/2}.$$

Sada je:

$$\begin{aligned} \boldsymbol{S} &= \boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^T = \boldsymbol{C}\boldsymbol{D}^{1/2}\boldsymbol{D}^{1/2}\boldsymbol{C}^T \\ &= (\boldsymbol{C}\boldsymbol{D}^{1/2})(\boldsymbol{C}\boldsymbol{D}^{1/2})^T. \end{aligned} \tag{3.5}$$

Izraz (3.5) je oblika

$$\boldsymbol{S} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T, \;\; \hat{\mathbf{\Lambda}} = \boldsymbol{C}\boldsymbol{D}^{1/2},$$

no, ne definiramo $\hat{\mathbf{\Lambda}} = \boldsymbol{C}\boldsymbol{D}^{1/2}$ jer je $\boldsymbol{C}\boldsymbol{D}^{1/2}$ matrica dimenzije $p \times p$.

Tražimo matricu $\hat{\mathbf{\Lambda}}$ dimenzije $p \times m$ pri čemu je $m < p$.

Stoga definiramo matricu $\boldsymbol{D}_1 = diag(\theta_1, \theta_2, \ldots, \theta_m)$ kao dijagonalnu matricu s prvih m najvećih svojstvenih vrijednosti matrice $\boldsymbol{S}$ na dijagonali $\theta_1 > \theta_2 > \cdots > \theta_m$ i matricu $\boldsymbol{C}_1 = (c_1, c_2, \ldots, c_m)$ s pripadajućim svojstvenim vektorima. Dakle, procjenjujemo $\mathbf{\Lambda}$ s prvih m stupaca matrice $\boldsymbol{C}\boldsymbol{D}^{1/2}$:

$$\hat{\mathbf{\Lambda}} = \boldsymbol{C}_1\boldsymbol{D}_1^{1/2} = (\sqrt{\theta_1}c_1, \sqrt{\theta_2}c_2, \ldots, \sqrt{\theta_m}c_m)$$

pri čemu je $\hat{\mathbf{\Lambda}}$ matrica dimenzije $p \times m$, $\boldsymbol{C}_1$ dimenzija $p \times m$ i $\boldsymbol{D}_1^{1/2}$ dimenzije $m \times m$.

Ime metode, metoda glavnih komponenti, dolazi od toga što su redci matrice $\hat{\mathbf{\Lambda}}$ proporcionalni svojstvenim vektorima matrice $\boldsymbol{S}$. To znači da su faktorski koeficijenti j-tog faktora proporcionalni koeficijentima j-te glavne komponente. No, ovo nije konačna verzija faktorskih koeficijenata, najčešće se koeficijenti još

rotiraju množenjem ortogonalnom matricom, stoga je interpretacija faktorskih koeficijenta drugačija od interpretacije koeficijenata glavnih komponenti.

Uočimo, i-ti dijagonalni element matrice $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T$ jednak je sumi kvadrata i-tog retka $\hat{\mathbf{\Lambda}}$, tj. $\hat{\lambda}_i^T\hat{\lambda}_i = \sum_{j=1}^m \hat{\lambda}_{ij}^2$. Dijagonalne elemente matrice $\hat{\mathbf{\Phi}}$, $[\hat{\mathbf{\Phi}}]_{ii} = \hat{\phi}_i$, definiramo na sljedeći način:

$$\hat{\phi}_i = s_{ii} - \sum_{j=1}^{m} \hat{\lambda}_{ij}^2.$$

Konačno, $\boldsymbol{S}$ aproksimiramo na sljedeći način:

$$\boldsymbol{S} \cong \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Phi}} \tag{3.6}$$

pri čemu je $\hat{\mathbf{\Phi}} = diag(\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p)$.

Dijagonalni elementi matrice $\boldsymbol{S}$, tj. uzoračke varijance od X, modelirane su egzaktno s (3.6), a nedijagonalni elementi od $\boldsymbol{S}$ su aproksimativni.

Kod ove metode je procjena za zajedničku varijancu jednaka sumi kvadrata redaka matrice $\hat{\mathbf{\Lambda}}$:

$$\hat{h}_i^2 = \sum_{j=1}^{m} \hat{\lambda}_{ij}^2,$$

Varijanca i-te varijable, X_i, particionirana je na zajedničku varijancu i na specifičnu varijancu:

$$\begin{aligned} s_{ii} &= \hat{h}_i^2 + \hat{\phi}_i \\ &= \hat{\lambda}_{i1}^2 + \hat{\lambda}_{i2}^2 + \cdots + \hat{\lambda}_{im}^2 + \hat{\phi}_i. \end{aligned}$$

Dakle, j-ti faktor doprinosi varijanci s_{ii} s vrijednošću $\hat{\lambda}_{ij}^2$. Doprinos j-tog faktora ukupnoj uzoračkoj varijanci $tr(\boldsymbol{S}) = s_{11} + s_{22} + \cdots + s_{pp}$ je stoga:

$$\textit{Doprinos j-tog faktora varijanci} = \sum_{i=1}^{p} \hat{\lambda}_{ij} = \hat{\lambda}_{1j}^2 + \hat{\lambda}_{2j}^2 + \cdots + \hat{\lambda}_{pj}^2$$

što je jednako sumi kvadrata težina j-tog stupca matrice $\hat{\Lambda}$.

Suma kvadrata j-tog stupaca matrice $\hat{\mathbf{\Lambda}}$ jednaka je j-toj svojstvenoj vrijednosti matrice $\boldsymbol{S}$:

$$\begin{aligned} \sum_{i=1}^{p} \hat{\lambda}_{ij}^2 &= \sum_{i=1}^{p} (\sqrt{\theta_j} c_{ij})^2 \\ &= \theta_j \sum_{i=1}^{p} c_{ij}^2 \\ &= \theta_j, \end{aligned}$$

jer je vektor c_j normaliziran.

Iz ovoga dobivamo da je odnos ukupne uzoračke varijance i varijance uzrokovane j-tim faktorom jednaka:

$$\frac{\sum_{i=1}^{p} \hat{\lambda}_{ij}}{tr(\boldsymbol{S})} = \frac{\theta_j}{tr(\boldsymbol{S})}.$$

Prikladnost faktorskog modela za dane podatke možemo ocijeniti uspoređujući lijevu i desnu stranu (3.6). Matrica grešaka

$$\boldsymbol{E} = \boldsymbol{S} - (\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Phi}})$$

ima nule na glavnoj dijagonali i ne-nul elemente na ostalim mjestima. Sljedeća nejednakost daje graničnu vrijednost za veličinu elemenata u matrici $\boldsymbol{E}$:

$$\sum_{ij} e_{ij}^2 \leq \theta_{m+1}^2 + \theta_{m+2}^2 + \cdots + \theta_p^2,$$

tj. suma kvadrata elemenata matrice $\boldsymbol{E}$ je manja ili jednaka sumi kvadrata svojstvenih vrijednosti $\boldsymbol{S}$ koje nismo koristili u metodi. Ako su svojstvene vrijednosti male, reziduali matrice grešaka su mali i model je dobar.

Za provođenje metode maksimalne vjerodostojnosti za procjenu matrica $\boldsymbol{\Lambda}$ i $\boldsymbol{\Phi}$ je nužna je pretpostavka da podaci dolaze iz multivarijatne normalne distribucije. U tom slučaju se može pokazati da procjene $\hat{\boldsymbol{\Lambda}}$ i $\hat{\boldsymbol{\Phi}}$ zadovoljavaju sljedeće:

$$\begin{aligned} \boldsymbol{S}\hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Lambda}} &= \hat{\boldsymbol{\Lambda}}(\boldsymbol{I} + \hat{\boldsymbol{\Lambda}}^T\hat{\boldsymbol{\Phi}}^{-1}\hat{\boldsymbol{\Lambda}}) \\ \hat{\boldsymbol{\Phi}} &= diag(\boldsymbol{S} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^T), \\ \hat{\boldsymbol{\Lambda}}^T\hat{\boldsymbol{\Phi}}^{-1}\hat{\boldsymbol{\Lambda}} &\ \textit{je dijagonalna.} \end{aligned}$$

Ove jednadžbe se rješavaju iterativno.

U nekim primjenama faktorske analize cilj je utvrditi odgovara li faktorski model podacima i identificirati te faktore. S druge strane, u nekim primjenama je cilj konkretno dobiti skorove za faktore, tj, $\hat{f}_i = (\hat{f}_{i1}, \hat{f}_{i2}, \ldots, \hat{f}_{im})^T$ za $i = 1, 2, \ldots, n$ koji se definiraju kao procjene vrijednosti faktora za svako opažanje. Primjerice, u interesu nam može biti interpretacija opažanja u terminima faktora. Faktorske skorove procjenjujemo na temelju podataka, a najpopularnija metoda za procjenu faktorskih skorova bazirana je na višestrukoj multivarijatnoj linearnoj regresiji. Vidi [20].

3.4 Usporedba analize glavnih komponenti i faktorske analize

Uz dane slučajne varijable $X_1, X_2, \ldots, X_p$ s pripadnom matricom kovarijanci $\boldsymbol{\Sigma}$, glavni cilj analize glavnih komponenti i faktorske analize je smanjiti dimenzionalnost skupa varijabli s p na m uz minimalan gubitak informacija. Glavna razlika između analize glavnih komponenti i faktorske analize je da se faktorska analiza temelji na konkretnom modelu, a analiza glavnih komponenti ne pretpostavlja postojanje modela.

Obje metode objašnjavaju neki dio strukture matrice kovarijance $\boldsymbol{\Sigma}$. Analiza glavnih komponenti se fokusira na njene dijagonalne elemente. Naime, cilj analize glavnih komponenti je maksimizirati $\sum_{k=1}^{m} Var(Z_k)$ (Vidi Propoziciju 2.1) kako bismo s $m < p$ glavnih komponenti objasnili što je više moguće varijacije u podacima, tj. time želimo obuhvatiti što je više moguće sume dijagonalnih elemenata matrice $\boldsymbol{\Sigma}$.

S druge strane, kod faktorskog modela iz (3.3):

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Phi},$$

faktorski dio modela $\boldsymbol{\Lambda}f$ u potpunosti objašnjava ne-dijagonalne elemente matrice $\boldsymbol{\Sigma}$ jer je $\boldsymbol{\Phi}$ dijagonalna matrica. Dijagonalni elementi matrice $\boldsymbol{\Phi}$, ϕ_j za $j = 1, 2, \ldots, p$, imat će malu vrijednost ukoliko sve varijable imaju značajnu zajedničku varijaciju. U slučaju da postoji varijabla X_j koja je gotovo nezavisna od drugih varijabli tada će $Var(\varepsilon_j) = \phi_j$ biti bliska $Var(X_j)$. Zbog činjenice da je fokus faktorskog modela na nedijagonalnim elementima matrice $\boldsymbol{\Sigma}$, ponekad se koriste prvih m glavnih komponenti za dobivanje početne aproksimacije faktorskih koeficijenata.

Još jedna razlika između analize glavnih komponenti i faktorske analize je da mijenjanje dimenzionalnosti m ima veći utjecaj na faktorsku analizu. Kod analize glavnih komponenti, ako povećamo dimenzionalnost s m_1 na m_2, tj. ako dodamo $m_2 - m_1$ novih komponenti, originalne m_1 komponente se ne mijenjaju. S druge strane, povećavanje dimenzije s m_1 na m_2 u faktorskoj analizi daje novih m_2 faktora koji ne moraju imati nikakvu poveznicu s originalnih m_1 faktora.

Osim dosad navedenih razlika, ove dvije tehnike se razlikuju i kod određivanja broja dimenzije m za adekvatan prikaz p-dimenzionalnog prostora vektora $X = (X_1, X_2, \ldots, X_p)$. Kod analize glavnih komponenti, ako postoje varijable koje su gotovo nezavisne od drugih varijabli, tada će postojati i glavne komponente koje su gotovo ekvivalentne originalnim varijablama. U usporedbi, zajednički faktor u faktorskoj analizi mora davati doprinos za barem dvije varijable. Dakle, nije moguće imati zajednički faktor koji daje doprinos samo jednoj varijabli. Takvi se faktori tretiraju kao specifični faktori (greške) i ne doprinose dimenzionalnosti modela. Stoga, za dani skup podataka, broj faktora za adekvatan faktorski model je manji ili jednak broju glavnih komponenti potrebnih za opisivanje što je više moguće varijacije podataka.

Konačno, treba uočiti da se glavne komponente mogu direktno dobiti iz vektora X za razliku od zajedničkih faktora. Glavne komponente su definirane kao linearne funkcije od X:

$$Z = \boldsymbol{A}^T X,$$

a faktori nisu egzaktna linearna funkcija od X. U faktorskoj analizi, vektor X je definiran kao linearna funkcija od f uz grešku, a inverzna veza ne daje nužno egzaktnu vezu između f i X. Činjenica da je očekivana vrijednost od X linearna funkcija od f ne implicira da je očekivana vrijednost od f linearna funkcija od X (osim ako je u pitanju višedimenzionalna normalna distribucija).

4 Primjena metoda na primjeru

U ovom ćemo dijelu rada provesti analizu glavnih komponenti i faktorsku analizu na primjerima koristeći R. Istražit ćemo postoje li temeljni obrasci ili veze među velikim skupom varijabli i mogu li se te informacije sažeti u manji broj glavnih komponenti ili faktora te interpretirati rezultate.

4.1 Primjer analize glavnih komponenti

Analizu glavnih komponenti provest ćemo na bazi podataka "Abalone" preuzetoj sa stranice [9]. Baza podataka "Abalone" sadrži podatke o puževima puzlatka (Petrovo uho) te ima 9 varijabli i 4177 podatka. Varijable su sljedeće:

- X_1: Spolna zrelost - muško (M), žensko (F), nezreo (I)
- X_2: Duljina kućice puža u milimetrima
- X_3: Širina kućice puža u milimetrima
- X_4: Visina puža u milimetrima
- X_5: Masa puža i kućice u gramima
- X_6: Masa puža u gramima
- X_7: Masa unutrašnjih organa puža u gramima
- X_8: Masa kućice puža u gramima
- Y: Broj prstenova na kućici

Broj prstenova na kućici nam govori koja je starost puža - naime, puž puzlatka, nakon prve godine života, svake godine dodaje jedan prsten svojoj kućici. Starost puža možemo procijeniti dodavanjem broja 1.5 broju prstenova na kućici. Detaljniji opis baze i način prikupljanja podataka se može pronaći u [18].

Za provedbu analize glavnih komponenti (a i faktorske analize), prema [14], veličina uzorka generalno ne bi trebala biti manja od 50, a preferiraju se uzorci veličine 100 ili više. U terminima omjera podataka i varijabli, također prema [14], u pravilu bi omjer između broja podataka i broja varijabli trebao biti minimalno 5:1. Korišteni uzorak zadovoljava ove preporuke.

Osnovni zahtjev za provedbu analize glavnih komponenti je mogućnost računanja korelacija među varijablama. Za numeričke varijable postoji više vrsta korelacija koje možemo izračunati. Kategorijalne varijable se najčešće uklanjaju iz razmatranja u analizi jer se za njih ne mogu koristiti iste korelacijske mjere kao i za numeričke varijable. Iz navedenih razloga provest ćemo analizu glavnih komponenti na varijablama $X_2 - X_8$ i pomoću njih ilustrirati uporabu glavnih komponenti u linearnoj regresiji za predviđanje varijable Y. Ideja analize glavnih komponenti u ovom kontekstu je: 1. smanjiti dimenzionalnost podataka jer naslućujemo da su varijable visoko korelirane, 2. predviđati broj prstenova na kućici puža (starost) pomoću glavnih komponenti da bismo izbjegli problem multikolinearnosti.

Prvi korak kod analize glavnih komponenti je analizirati korelacije među varijablama. Ukoliko su korelacije između varijabli vrlo niske nije opravdano koristiti ovu tehniku. Vizualan pristup ovom problemu je promotriti matricu korelacija varijabli i odbaciti ideju korištenja ove metode ukoliko je mali broj korelacija između varijabli veći od 0.3.

	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_2	1.00						
X_3	0.99	1.00					
X_4	0.83	0.83	1.00				
X_5	0.93	0.93	0.82	1.00			
X_6	0.90	0.89	0.77	0.97	1.00		
X_7	0.90	0.90	0.80	0.97	0.93	1.00	
X_8	0.90	0.91	0.82	0.96	0.88	0.91	1.00

Tablica 4.1: Procijenjeni korelacijski koeficijenti za podatke "Abalone"

Iz matrice korelacija varijabli uočavamo da su varijable u $X_2 - X_8$ visoko korelirane (sve su korelacije veće od 0.7), vizualno gledajući ima smisla raditi analizu glavnih komponenti na ovom uzorku.

Objektivniji pristup ovom problemu je korištenje Bartlettovog testa za matrice kovarijanci koji testira ekstreman slučaj - je li korelacijska matrica statistički značajno različita od jedinične matrice. Ako je korelacijska matrica slična jediničnoj matrici onda su varijable nekorelirane i nema smisla provoditi faktorsku analizu i analizu glavnih komponenti. Provođenjem Bartlettovog testa u R-u dobivamo da je p-vrijednost jednaka 0 te zaključujemo da je korelacijska matrica statistički značajno različita od jedinične matrice.

Alternativni način kvantificiranja stupnja korelacije između varijabli je Kaiser-Meyer-Olkinova (KMO) mjera primjerenosti uzorka. KMO mjera predstavlja omjer sume kvadrata korelacija između varijabli i zbroja sume parcijalnih korelacija i sume kvadrata korelacija između varijabli. Formula za računanje KMO mjere j-te varijable je:

$$KMO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} u_{ij}}$$

pri čemu su r_{ij} elementi matrice korelacija i u_{ij} elementi matrice parcijalnih korelacija.

Ova mjera nam govori koliko je uzorak prikladan za analizu glavnih komponenti. Vrijednost KMO mjere nalazi se između 0 i 1. Prema [11], uzorak možemo ocijeniti koristeći sljedeću ljestvicu:

- 0.0-0.49 neprihvatljivo
- 0.50-0.59 loše
- 0.60-0.69 prosječno
- 0.70-0.79 osrednje
- 0.80-0.89 dobro
- 0.90- 1.0 odlično.

Vrijednost KMO mjere za varijable $X_1 - X_8$ prikazane su u sljedećoj tablici:

	X_2	X_3	X_4	X_5	X_6	X_7	X_8
KMO	0.86	0.86	0.99	0.75	0.81	0.89	0.81

Tablica 4.2: Vrijednost Kaiser-Meyer-Olkinove mjere primjerenosti uzorka za varijable

Iz prethodne analize korelacija varijabli zaključujemo da je uzorak prikladan za provedbu analize glavnih komponenti.

Sljedeći korak analize je izračun i odabir glavnih komponenti. Izračun analize glavnih komponenti provest ćemo na matrici korelacija, vidi Poglavlje 2.5. Vektori koeficijenata glavnih komponenti dani su u sljedećoj tablici:

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
X_2	0.3832508	0.03786529	-0.5932799	-0.089331673	0.040512600	0.699651086	0.0238929537
X_3	0.3835732	0.06532324	-0.5853661	-0.008285814	0.008517628	-0.711025627	-0.0158868202
X_4	0.3481438	0.86683603	0.3148764	-0.165564868	-0.027110424	0.009841283	-0.0007077915
X_5	0.3906735	-0.23327117	0.2308252	0.052280164	-0.110183954	-0.021653298	0.8510792809
X_6	0.3781883	-0.34801069	0.2315678	-0.496179039	-0.545339050	-0.011030516	-0.3721942422
X_7	0.3815134	-0.25290295	0.2702527	-0.140972073	0.809328460	-0.023996063	-0.2049136954
X_8	0.3789217	-0.05837478	0.1621047	0.834110000	-0.181668556	0.060561675	-0.3071190581

Tablica 4.3: Vektori koeficijenata glavnih komponenti

Promotrimo i sljedeću tablicu:

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
Standardna devijacija	2.5209	0.52861	0.40908	0.3377	0.25427	0.11282	0.08159
Proporcija varijance	0.9079	0.03992	0.02391	0.0163	0.00924	0.00182	0.00095
Kumulativna proporcija varijance	0.9079	0.94779	0.97170	0.9880	0.99723	0.99905	1.00000

Tablica 4.4: Važnost glavnih komponenti

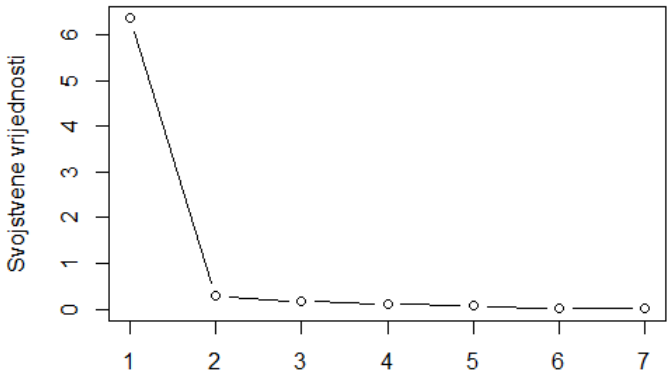
Iz Tablice 4.8 iščitavamo da prva glavna komponenta objašnjava 90.79% varijacije podataka, druga glavna komponenta 3.92% varijacije podataka, a ostale komponente u zbroju objašnjavaju manje od 6% varijacije podataka. Postavlja se pitanje koliko glavnih komponenti treba odabrati za daljnju analizu.

Najpoznatija metoda odabira broja komponenti je metoda svojstvenih vrijednosti poznata kao Kaiserov kriterij. Tehnika je jednostavna za primijeniti - zadržana komponenta treba objašnjavati varijancu barem jedne varijable. Kod analize glavnih komponenti varijanca svake varijable je jedan ukoliko koristimo matricu korelacija. Stoga bismo kod analize glavnih komponenti zadržali samo one komponente čije su varijance veće od 1. Prema [16] je Kaiserov kriterij prestrog i predlaže se zadržavanje komponenti čije su varijance veće od 0.7. Prema oba ova kriterija zadržali bismo samo prvu glavnu komponentu.

Drugi način odabira broja komponenti je odabir granične vrijednosti za ukupnu objašnjenu varijancu, primjerice, 80% ili 90%. U prirodnim znanostima najčešće se uzima granica od 95% dok je kod društvenih znanosti već i postotak veći od 60 zadovoljavajući. Prema ovome kriteriju bismo zadržali dvije ili tri komponente.

U svrhu odabira broja komponenti koristi se i Bartlettov test za varijance koji smo opisali u Poglavlju 2.4.4.

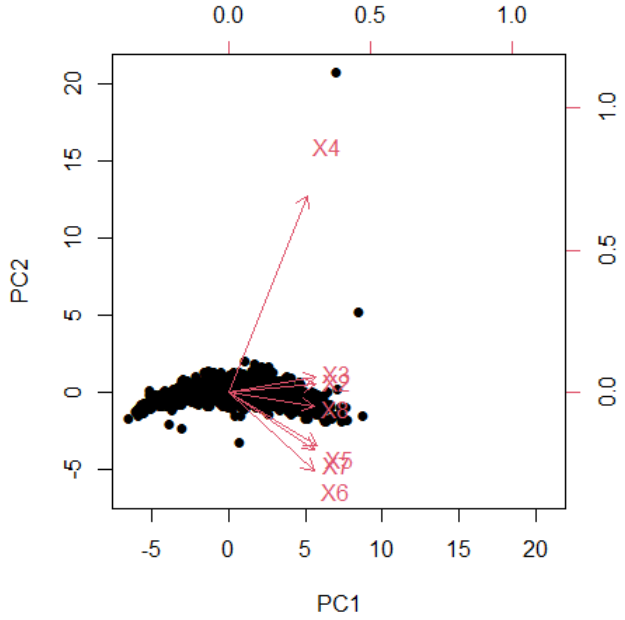
Još jedna metoda za odabir broja komponenti je tzv. scree test. Scree test je grafički prikaz varijanci, l_k, k glavnih komponenti. Ideja je vizualno odrediti za koju vrijednost k dolazi do infleksije i taj broj komponenti zadržati. Pogledajmo scree test za naš uzorak:



Slika 4.1: Svojstvene vrijednosti glavnih komponenti

Uočavamo da dolazi do infleksije kod komponente broj dva, stoga bismo prema ovom kriteriju zadržali do dvije komponente. Iz ovoga i prethodnih razmatranja donosimo odluku da ćemo zadržati dvije glavne komponente.

Sljedeći grafički prikaz prikazuje podatke iz uzorka u terminima prve i druge glavne komponente, a pokazuje i kako originalne varijable utječu na prve dvije glavne komponente:



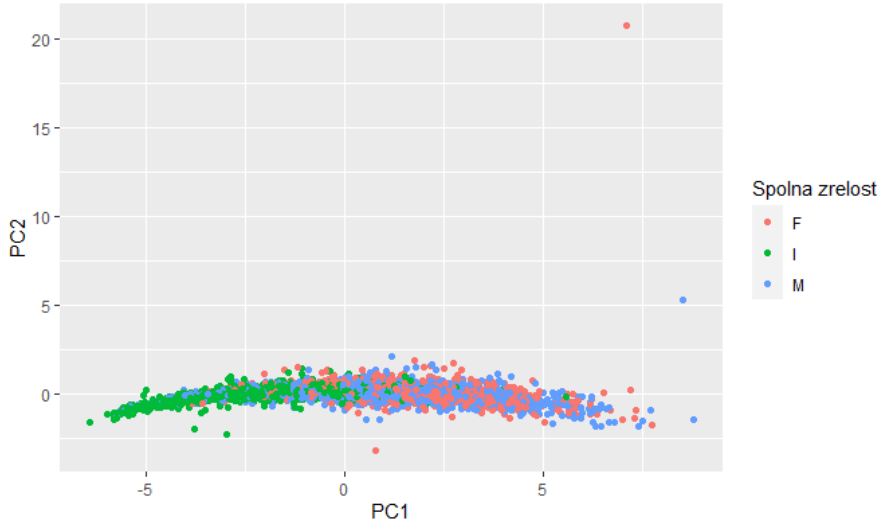
Slika 4.2: Grafički prikaz uzorka u odnosu na PC1 i PC2

Iz Slike 4.2 vidimo da varijable X_2, X_3 i X_8 doprinose gotovo isključivo prvoj glavnoj komponenti što se može iščitati i iz Tablice 4.3. Jedinke koji imaju veće vrijednosti za te varijable bit će smještene desno od ishodišta. Najveći doprinos drugoj glavnoj komponenti daje varijabla X_4, jedinke s velikom vrijednošću X_4 bit će smještene iznad ishodišta. Varijable X_6 doprinosi podjednako objema glavnim komponentama i slično.

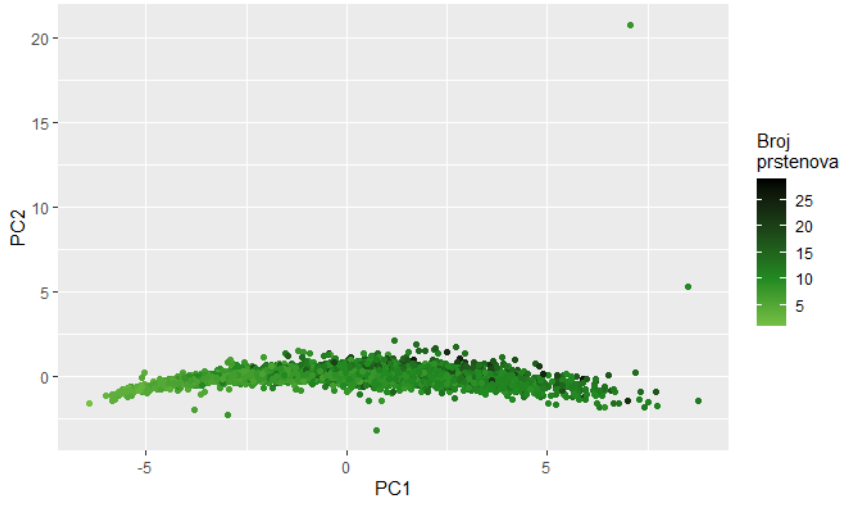
Glavne komponente koje objašnjavaju većinu varijacije interpretirat ćemo pomoću Tablice 4.3. Prva glavna komponenta ima slične (pozitivne) vrijednosti težina za svaku od varijabli $X_2 - X_8$. Interpretacija za prvu glavu komponentu je da ona mjeri ukupnu veličinu puža. Druga glavna komponenta ima pozitivne

vrijednosti težina za varijable $X_2 - X_4$ i negativne vrijednosti za varijable $X_5 - X_8$. Druga komponenta suprotstavlja veličine mjerene u milimetrima (duljina kućice, širina kućica i visina kućice) s veličinama mjerenim u gramima (ukupna masa puža, masa puža bez kućice, masa kućice i masa unutrašnjih organa). Druga komponenta nam govori da je glavni izvor varijacije između puževa, nakon što izuzmemo varijaciju uzrokovanu ukupnom veličinom puža, varijacija uzrokovana puževima čije su mjere u milimetrima relativno velike u usporedbi s njihovom gramažom i obrnuto. U slučaju da originalne glavne komponente nisu jednostavne za interpretirati mogu se koristiti neke tehnike rotacije za pojednostavljenje interpretacije, vidi [16].

Pogledajmo sada koliko dobro prve dvije glavne komponente razlučuju puževe različite spolne zrelosti i različite dobi(različitog broja prstenova na kućici):



Slika 4.3: Grafički prikaz uzorka u odnosu na PC1 i PC2 s obzirom na spolnu zrelost



Slika 4.4: Grafički prikaz uzorka u odnosu na PC1 i PC2 s obzirom na broj prstenova

Iz Slike 4.3 možemo uočiti da su glavne komponente uspjele razdvojiti puževe kategorija M i F od puževa iz kategorije I. Iz Slike 4.5 uočavamo da su glavne komponente djelomično razdvojile i puževe manje dobi od onih veće dobi što nam je jedna od motivacija za predviđanje dobi puževa glavnim komponentama. Naime, do sada smo analizom uzorka zaključili da postoje značajne korelacije između varijabli. Kod predviđanja varijable Y

varijablama $X_2 - X_8$ modelom linearne regresije imali bismo značajan problem s multikolinearnošću. Stoga se okrećemo modelu regresije s glavnim komponentama.

Najprije pogledajmo linearan model za predviđanje Y gdje su prediktori originalne varijable baze:

Koeficijent	Procjena	Standardna devijacija	t-vrijednost	Pr($>|t|$)
Slobodan član	2.9852	0.2691	11.092	$<$ 2e-16
X_2	-1.5719	1.8248	-0.861	0.389
X_3	13.3609	2.2371	5.972	2.53e-09
X_4	11.8261	1.5481	7.639	2.70e-14
X_5	9.2474	0.7326	12.622	$<$ 2e-16
X_6	-20.2139	0.8233	-24.552	$<$ 2e-16
X_7	-9.8297	1.3040	-7.538	5.82e-14
X_8	8.5762	1.1367	7.545	5.54e-14

Tablica 4.5: Linearan model regresije za Y uz originalne varijable

Pripadni VIF-ovi za taj model su:

	X_2	X_3	X_4	X_5	X_6	X_7	X_8
VIF	40.771813	41.845452	3.559939	109.592750	28.353191	17.346276	21.258289

Tablica 4.6: Vrijednosti VIF-a za varijable $X_2 - X_8$

Iz Tablice 4.5 uočavamo da su standardne devijacije procijenjenih regresijskih koeficijenata vrlo visoke, a iz Tablice 4.6 iščitavamo da imamo velik problem s multikolinearnošću što smo i ranije naslutili. Prilagođeni R^2 za ovaj model iznosi 0.5268, tj. 52.68% varijacije dobi puža je objašnjeno ovim modelom.

U usporedbi, promotrimo sada model u kojemu koristimo glavne komponente kao prediktore:

Koeficijent	Procjena	Standardna devijacija	t-vrijednost	Pr($>|t|$)
Slobodan član	9.93368	0.03432	289.481	$<$ 2e-16
$PC1$	0.72456	0.01361	53.222	$<$ 2e-16
$PC2$	1.21464	0.06492	18.709	$<$ 2e-16
$PC3$	-0.59830	0.08390	-7.132	1.16e-12
$PC4$	3.53497	0.10161	34.788	$<$ 2e-16
$PC5$	0.84845	0.13497	6.286	3.59e-10
$PC6$	-1.02053	0.30420	-3.355	0.000801
$PC7$	5.35765	0.42063	12.737	$<$ 2e-16

Tablica 4.7: Linearan model regresije za Y uz glavne komponente

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
Kumulativna proporcija varijance	0.9079	0.94779	0.97170	0.9880	0.99723	0.99905	1.00
Kumulativna proporcija varijance Y	32.09	36.06	36.64	50.35	50.80	50.92	52.68

Tablica 4.8: Važnost glavnih komponenti za uzorak i Y

Kod ovog su modela standardne devijacije procijenjenih regresijskih koeficijenata manje u usporedbi sa standardnim devijacijama regresijskih koeficijenata prethodnog modela, a prilagođeni R^2 iznosi 52.68%. Načine odabira glavnih komponenti u modelu linearne regresije komentirali smo u Poglavlju 2.6.3. U ovom ćemo primjeru napraviti selekciju varijabli s obzirom na vrijednosti t-statistike tj. s obzirom na doprinos glavne komponente regresijskoj jednadžbi. Zadržat ćemo prvu, drugu i četvrtu glavnu komponentu. U

konačnom modelu koeficijenti za odabrane komponente ostaju nepromijenjeni, a prilagođeni R^2 iznosi 0.4974, tj. 49.74% varijacije dobi puža objašnjena je s prvom, drugom i četvrtom glavnom komponentom. Ovim smo postupkom uklonili dio varijacije regresijskih koeficijenata uzrokovan multikolinearnošću varijabli i smanjili broj varijabli za predikciju dobi sa sedam na tri.

4.2 Primjer faktorske analize

U ovom ćemo dijelu rada ilustrirati primjenu faktorske analize na podacima, objasniti kada je uzorak prikladan za faktorsku analizu, kako odabrati broj faktora za analizu i kako interpretirati faktorsku strukturu.

Jedna od primjena faktorske analize je izrada upitnika. Naime, ukoliko upitnikom želimo mjeriti neke sposobnosti ili osobine bitno je osigurati da se postavljena pitanja odnose na konstrukciju koju namjeravamo izmjeriti. Podaci na kojima ćemo provesti faktorsku analizu odgovori su hrvatskih učitelja i nastavnika na upitnik SPTKTT (Survey of Preservice Teachers Knowledge of Teaching and Technology) koji su korišteni u radu [7]. Upitnik SPTKTT dizajniran je za mjerenje Tehnološko pedagoškog predmetnog znanja (TPACK) kod budućih učitelja, [22]. Za prvotni razvoj ovog upitnika, čiji je cilj mjerenje TPACK-a, prikupljani su podaci studenata sa sveučilišta Midwestern, SAD. Upitnik (vidi Tablicu 4.14) se sastoji od 47 čestica s pet mogućih odgovora Likertove ljestvice :

1. Uopće se ne slažem
2. Ne slažem se
3. Niti se slažem niti se ne slažem
4. Slažem se
5. U potpunosti se slažem,

a čestice ispituju sljedeće:

- tehnološka znanja (7 čestica)
- sadržajna znanja: matematika (3 čestice), društvene znanosti (3 čestice), znanost (3 čestice), pismenost (3 čestice)
- pedagoška znanja (7 čestica)
- pedagoško sadržajna znanja (4 čestice)
- tehnološka sadržajna znanja (4 čestice)
- tehnološko pedagoška znanja (9 čestica)
- tehnološko pedagoško sadržajna znanja (4 čestice).

U radovima [7] i [8] napravljena je validacija ovog upitnika za hrvatski obrazovni sustav. Ideja faktorske analize u ovom kontekstu je utvrditi postoje li razlike u strukturi faktora upitnika SPTKTT s obzirom na to da se hrvatski i američki obrazovni sustav razlikuju. Faktorsku analizu ćemo provesti na podacima koji su prikupljeni iz populacije učitelja i nastavnika koji su bili zaposleni u Republici Hrvatskoj u osnovnim školama tijekom 2015. godine, a riječ je o 266 ispitanika. Za nekoliko ispitanika nedostaju neki od odgovora

pa izbacivanjem tih ispitanika iz uzorka dolazimo do 254 ispitanika. Veličina uzorka zadovoljava preporuke koje smo naveli u prethodnom poglavlju, tj. uzorak ima više od 100 opservacija i omjer između broja varijabli i broja podataka veći je od 5:1.

Kao i prije, da bismo ocijenili adekvatnost podataka za faktorsku analizu, analizirat ćemo korelacijsku strukturu podataka. Procijenjena korelacijski koeficijenti za podatke mogu se vidjeti u Dodatku, Tablica 4.15, a za izračun korelacija korišten je Pearsonov koeficijent korelacije. Iz matrice uočavamo da postoji značajan broj varijabli između kojih je korelacija veća od 0.4. Također, detaljnijom analizom matrice korelacija možemo uočiti grupiranja nekih od varijabli. Primjerice, varijable P1-P6 (čestice koje ispituju tehnološka znanja) su međusobno visoko korelirane, varijable P8-P10 su visoko korelirane (čestice koje ispituju sadržajna znanja iz matematike) itd.

Bartlettov test za matrice korelacije u ovom slučaju također odbacuje hipotezu da je matrica korelacija jedinična matrica s p-vrijednošću 0.

Vrijednosti Kaiser-Meyer-Olkinove mjere za primjerenost uzorka su sljedeće:

	$P1$	$P2$	$P3$	$P4$	$P5$	$P6$	$P7$	$P8$	$P9$	$P10$	$P11$	$P12$	$P13$	$P14$	$P15$	$P16$
KMO	0.89	0.89	0.92	0.90	0.92	0.90	0.98	0.84	0.83	0.86	0.89	0.85	0.87	0.89	0.85	0.91

	$P17$	$P18$	$P19$	$P20$	$P21$	$P22$	$P23$	$P24$	$P25$	$P26$	$P27$	$P28$	$P28$	$P30$	$P31$	$P32$
KMO	0.84	0.88	0.86	0.93	0.91	0.88	0.92	0.93	0.93	0.93	0.89	0.86	0.88	0.86	0.92	0.88

	$P33$	$P34$	$P35$	$P36$	$P37$	$P38$	$P39$	$P40$	$P41$	$P42$	$P43$	$P44$	$P45$	$P46$	$P47$	
KMO	0.90	0.88	0.85	0.88	0.90	0.92	0.93	0.92	0.96	0.95	0.96	0.91	0.88	0.89	0.90	

Tablica 4.9: KMO mjera za varijable $P1 - P47$

Sve KMO vrijednosti veće su od 0.8 stoga zaključujemo da je ovaj uzorak prikladan za faktorsku analizu. Iz ovog i prethodnih razmatranja zaključujemo da su varijable međusobno visoko korelirane te da na ovim podacima ima smisla provesti faktorsku analizu.

Broj faktora za daljnju analizu možemo odrediti na nekoliko načina:

1. Prethodno znanje o broju potrebnih faktora
2. Kaiserov kriterij svojstvenih vrijednosti matrice korelacija
3. Scree test
4. Odabrati broj faktora m potreban za objašnjavanje nekog postotka varijacije podataka, primjerice 80%
5. Testirati hipotezu da je m korektan broj faktora, $H_0 : \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Phi}$

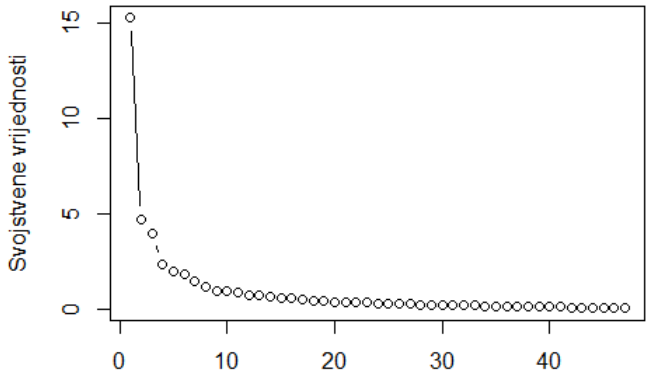
U originalnom je istraživanju, [22], za upitnik SPTKTT utvrđeno 10 faktora pa bismo prema tome zadržali 10 faktora.

Za Kaiserov kriterij, pogledajmo prvih nekoliko najvećih svojstvenih vrijednosti matrice korelacija:

$$\begin{aligned} l_1 &= 15.29, & l_7 &= 1.52, \\ l_2 &= 4.73, & l_8 &= 1.17, \\ l_3 &= 3.98, & l_9 &= 1.01, \\ l_4 &= 2.39, & l_{10} &= 0.96, \\ l_5 &= 2.01, & l_{11} &= 0.93, \\ l_6 &= 1.84, & l_{12} &= 0.78. \end{aligned}$$

Prema Kaiserovom kriteriju odabrali bismo 9 faktora.

Pogledajmo sada grafički prikaz scree testa:



Slika 4.5: Scree test

Prema scree testu bismo ostavili do 4 faktora.

Za odabir faktora prema kriterijima 5 i 6 trebamo odrediti faktorski model za različite vrijednosti faktora. Faktorski model odredit ćemo u R-u pomoću metode maksimalne vjerodostojnosti. Provođenjem faktorske analize za različite brojeve m dobivamo da je za objašnjavanje 60% varijacije podataka potrebno 7 faktora, za objašnjavanje 70% varijacije je potrebno 12 faktora, a za objašnjavanje 80% varijacije podataka je potrebno 22 faktora, pri čemu je količina objašnjene varijacije procijenjena aritmetičkom sredinom zajedničke varijance varijabli P1-P47.

Faktorsku analizu ćemo provesti koristeći 9 faktora, slijedeći Kaiserov kriterij. Procijenjena količina objašnjene varijacije podataka u faktorskom modelu s 9 faktora je 66%.

Za interpretaciju faktora promatramo faktorske koeficijente i donosimo odluku o tome koji faktorski koeficijenti doprinose objašnjenju faktora. Kako je faktorski koeficijent korelacija između nekog faktora i varijable, kvadrat koeficijenta je iznos ukupne varijance varijable objašnjene tim faktorom. Stoga, faktorski koeficijent koji iznosi 0.30 prevodi se u otprilike 10% objašnjene varijance, 0.50 označava da je 25% varijance objašnjeno tim faktorom. Koeficijent mora biti veći od 0.7 da bi faktor objašnjavao 50% varijance varijable. Zaključno, s povećanjem apsolutne vrijednosti faktorskog koeficijenta se povećava i njegova značajnost za interpretaciju faktorske matrice. U [14] se navode sljedeći kriteriji:

- Faktorski koeficijenti manji od $\pm$ 0.10 nisu značajni za interpretaciju jednostavne faktorske strukture

- Faktorski koeficijenti između $\pm$ 0.30 i $\pm$ 0.40 zadovoljavaju minimalnu razinu za interpretaciju faktorske strukture
- Faktorski koeficijenti $\pm$ 0.50 ili veći su značajni za faktorsku strukturu
- Koeficijenti koji su veći od $\pm$ 0.70 pokazatelji su dobro definirane strukture i cilj su svake faktorske analize.

Značajnost faktorskih koeficijenata ovisi i o veličini promatranog uzorka. Primjerice, za uzorak veličine 50 smatrali bismo da je faktorski koeficijent značajan ako je veći od 0.75, za uzorak veličine 100 smatrali bismo da je faktorski koeficijent značajan ako je veći od 0.55. Tablica 4.10, preuzeta iz [14], prikazuje smjernice za značajnost faktorskih koeficijenata ovisno o veličini uzroka:

Faktorski koeficijent	Veličina uzorka
0.30	350
0.35	250
0.40	200
0.45	150
0.50	120
0.55	100
0.60	85
0.65	70
0.70	60
0.75	50

Tablica 4.10: Smjernice za identificiranje značajnih faktorskih koeficijenata ovisno o veličini uzorka

Također, treba naglasiti da je interpretacija strukture otežana ako neka od varijabli nema nijedan značajan faktorski koeficijent ili ako ima više značajnih faktorskih koeficijenata. U slučaju da neke varijable nemaju nijedan značajan faktorski koeficijent može se poduzeti sljedeće:

- Zanemariti problematične varijable i nastaviti s interpretacijom modela imajući na umu da te varijable nisu dobro predstavljene faktorskim modelom
- Svaku varijablu pojedinačno analizirati i ocijeniti njenu važnost za uzorak. Ukoliko varijabla nije od velikog značaja može se ukloniti i izraditi novi faktorski model
- Rotirati faktorske koeficijente nekom drugom tehnikom
- Smanjiti ili povećati broj zadržanih faktora te provesti ponovno faktorsku analizu i vidjeti jesu li problematične varijable bolje predstavljene u novom modelu
- Modificirati tip faktorskog modela, npr. napraviti analizu glavnih komponenti umjesto faktorske analize.

U slučaju da neke varijable imaju više značajnih faktorskih koeficijenata (engl. cross-loading) smjernice se temelje na sljedećim načelima:

1. Treba uspoređivati varijance, a ne faktorske koeficijente. Varijancu svake varijable u faktorskom modelu možemo rastaviti na specifičnu varijancu i zajedničku varijancu koja je predstavljena kvadratima pripadnih faktorskih koeficijenata, vidi (3.4). Kod provođenja faktorske analize cilj je objasniti što

je više moguće varijance varijable kroz zajedničku varijancu koja je prisutna zbog faktora. Varijanca predstavljena razlikom dvaju faktorskih koeficijenta od 0.1 razlikuje se ovisno o veličini faktorskih koeficijenta. Primjerice, ako uspoređujemo faktorske koeficijente 0.7 i 0.6 (razlika od 0.1) razlika u varijanci je 0.13 (tj. $0.7^2 - 0.6^2 = 0.13$). S druge strane, za faktorske koeficijente 0.4 i 0.3 je razlika u varijanci 0.7 (tj. $0.4^2 - 0.3^2 = 0.7$). Stoga ako želimo shvatiti razliku utjecaja faktorskih koeficijenata, treba uspoređivati razliku u varijanci, a ne razliku između faktorskih koeficijenta.

2. Treba uspoređivati omjere varijanci. Primjerice, neka je razlika u varijanci 0.1. Ova razlika u varijanci se čini značajnija ako je riječ o razlici između 0.5 i 0.4 (npr. faktorski koeficijenti 0.71 i 0.63) nego kada je riječ o razlici između 0.2 i 0.1 (npr. faktorski koeficijenti 0.44 i 0.31). Ovu razliku zornije prikazuje omjer veće i manje varijance. U prvom slučaju je omjer 1.25, a u drugom slučaju 2.0.

Na temelju ovih načela dolazimo do sljedećih smjernica:

1. Pronaći varijablu koja ima dva značajna faktorska koeficijenta
2. Kvadrirati njene značajne faktorske koeficijente i izračunati omjer između veće i manje varijance
3. Napraviti kategorizaciju prema sljedećim uputama:
   - Između 1.0 i 1.5 - problematičan cross-loading. Uklanjanje varijable koja ima ovaj cross-loading može dati jednostavniju faktorsku strukturu.
   - Između 1.5 i 2.0 - upitan cross-loading. Uklanjanje varijable koja ima ovaj cross-loading ovisi o interpretabilnosti novih faktora.
   - Veće od 2.0 - zanemariv cross-loading. Možemo ga zanemariti za potrebe interpretacije.

Faktorsku analizu na podacima upitnika SPTKTT provest ćemo uz pretpostavku da imamo 9 faktora, slijedeći Kaiserov kriterij, a provest ćemo i rotaciju faktorskih koeficijenata radi pojednostavljenja interpretacije faktorskog modela (vidi Poglavlje 3.2). Najpopularnija tehnika rotacije je varimax rotacija. Varimax rotacija rotira faktorske koeficijente s ciljem maksimizacije varijance kvadrata koeficijenata svakog stupca matrice faktorskih koeficijenata $\hat{\mathbf{\Lambda}}$. Rezultat ove rotacije su visoki faktorski koeficijenti uz neke faktore (blizu 1 i -1) i niski faktorski koeficijenti uz ostale faktore (blizu 0). Faktorske koeficijente ćemo smatrati značajnima ukoliko su veći od 0.4 jer imamo 254 podatka (vidi Tablicu 4.10).

Nakon provođenja faktorske analize na podacima upitnika SPTTKT uz pretpostavku da imamo 9 faktora metodom maksimalne vjerodostojnosti uz varimax rotaciju, dobivamo sljedeće vrijednosti za procijenjene specifične varijance:

	$P1$	$P2$	$P3$	$P4$	$P5$	$P6$	$P7$	$P8$	$P9$	$P10$	$P11$	$P12$	$P13$	$P14$	$P15$	$P16$
ϕ_i	0.383	0.290	0.294	0.367	0.235	0.347	0.286	0.231	0.149	0.230	0.503	0.337	0.244	0.299	0.160	0.227

	$P17$	$P18$	$P19$	$P20$	$P21$	$P22$	$P23$	$P24$	$P25$	$P26$	$P27$	$P28$	$P28$	$P30$	$P31$	$P32$
ϕ_i	0.390	0.124	0.179	0.313	0.349	0.359	0.200	0.407	0.605	0.435	0.304	0.541	0.340	0.549	0.259	0.438

	$P33$	$P34$	$P35$	$P36$	$P37$	$P38$	$P39$	$P40$	$P41$	$P42$	$P43$	$P44$	$P45$	$P46$	$P47$	
ϕ_i	0.329	0.516	0.461	0.459	0.715	0.556	0.309	0.264	0.412	0.354	0.322	0.267	0.190	0.166	0.270	

Tablica 4.11: Procijenjene specifične varijance za varijable $P1 - P48$

Iz Tablice 4.11 iščitavamo koji dio varijacije varijable nije objašnjen faktorskim modelom. Ukoliko varijabla ima visoku specifičnu varijancu onda se ona ne uklapa dobro u faktorski model, primjerice za varijablu P37 procijenjena specifična varijanca iznosi 0.715 tj. manje od 30% varijacije varijable P37 je objašnjena ovim faktorskim modelom. Oduzimanjem specifičnih varijanci od broja 1 dobivamo zajedničku varijancu varijabli tj. varijancu koja je objašnjena faktorima.

Vrijednosti faktorskih koeficijenata dane su u sljedećoj tablici:

	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5	Faktor 6	Faktor 7	Faktor 8	Faktor 9
P1	0.15	0.15	0.74	0.14	0.02	0.06	0.02	-0.02	-0.04
P2	0.27	0.12	0.78	0.08	-0.00	0.06	0.08	0.06	0.02
P3	0.22	0.06	0.78	0.04	0.10	0.10	0.11	0.06	0.08
P4	0.19	-0.09	0.74	-0.01	0.11	0.11	0.12	0.06	0.03
P5	0.27	-0.04	0.81	0.07	0.11	0.14	0.05	-0.01	0.06
P6	0.31	0.01	0.73	0.08	0.06	-0.03	-0.05	0.07	0.02
P7	0.45	0.29	0.36	0.27	0.18	0.25	0.26	0.17	0.17
P8	0.07	0.12	0.08	0.84	0.17	0.07	0.11	-0.00	-0.00
P9	0.10	0.11	0.13	0.87	0.19	0.07	0.09	0.04	0.05
P10	0.05	0.16	0.08	0.79	0.16	0.12	0.23	-0.03	0.08
P11	0.12	0.25	0.07	0.14	0.09	0.09	-0.02	0.32	0.53
P12	0.14	0.18	0.02	0.03	0.19	0.08	-0.03	0.18	0.73
P13	0.16	0.25	0.02	0.02	0.28	0.05	0.02	0.19	0.74
P14	0.21	0.14	0.12	0.21	0.73	0.07	0.12	0.06	0.18
P15	0.19	0.04	0.08	0.22	0.84	0.11	0.05	0.04	0.15
P16	0.13	0.11	0.11	0.17	0.79	0.09	0.20	0.07	0.18
P17	0.16	0.33	0.10	0.06	0.02	0.05	-0.08	0.64	0.21
P18	0.14	0.26	0.07	-0.07	0.08	0.12	-0.04	0.85	0.17
P19	0.15	0.34	0.04	-0.04	0.03	0.15	0.02	0.78	0.24
P20	0.22	0.77	0.03	0.06	0.07	0.07	0.07	0.14	0.07
P21	0.13	0.76	0.00	0.14	0.11	0.06	0.04	0.12	0.07
P22	0.13	0.76	0.07	0.06	0.06	0.06	0.05	0.14	0.07
P23	0.13	0.86	0.09	0.11	0.10	0.04	0.07	0.05	0.09
P24	0.13	0.72	0.04	0.05	0.08	0.07	0.09	0.14	0.11
P25	0.26	0.50	0.03	0.11	0.07	0.13	0.06	0.16	0.11
P26	0.23	0.65	-0.00	0.09	-0.04	0.04	0.17	0.11	0.20
P27	0.04	0.31	0.02	0.41	0.15	0.04	0.64	-0.05	0.01
P28	0.05	0.41	0.06	-0.03	0.02	0.25	0.22	0.05	0.41
P29	0.06	0.25	0.10	0.21	0.44	0.17	0.56	-0.05	0.04
P30	0.03	0.40	0.03	0.01	0.01	0.36	0.19	0.26	0.25
P31	0.26	0.19	0.17	0.43	0.08	0.14	0.63	-0.03	0.01
P32	0.34	0.11	0.23	-0.03	0.04	0.39	0.31	0.12	0.35
P33	0.30	0.08	0.25	0.14	0.31	0.30	0.55	0.00	0.03
P34	0.29	0.18	0.11	-0.02	-0.02	0.44	0.22	0.25	0.23
P35	0.66	0.26	0.13	0.06	0.00	0.09	-0.02	0.10	0.02
P36	0.63	0.21	0.19	0.04	0.01	0.19	0.04	0.12	0.05
P37	0.42	0.11	0.13	0.05	0.19	0.09	0.13	0.11	-0.01
P38	0.55	0.10	0.19	-0.00	0.17	0.19	0.08	0.04	0.16
P39	0.71	0.09	0.39	0.03	0.08	0.05	0.08	0.05	0.09
P40	0.80	0.15	0.16	0.14	0.15	0.01	0.02	0.05	0.10
P41	0.66	0.23	0.19	0.08	0.11	0.15	0.08	0.03	0.07
P42	0.65	0.11	0.38	0.06	0.10	0.08	0.17	0.07	0.09
P43	0.68	0.13	0.34	0.07	0.07	0.20	0.10	0.06	0.16
P44	0.29	0.12	0.15	0.49	0.21	0.37	0.43	-0.07	-0.03
P45	0.29	0.13	0.11	0.21	0.10	0.78	0.04	0.17	0.07
P46	0.26	0.14	0.18	0.15	0.20	0.80	0.10	0.05	0.10
P47	0.27	0.12	0.17	0.25	0.42	0.49	0.37	-0.02	-0.02

Tablica 4.12: Faktorski koeficijenti za upitnik SPTKTT

Prije nego interpretiramo faktorsku strukturu iz Tablice 4.12 uočimo da ne postoje varijable koje nemaju nijedan značajan faktorski koeficijent. Nadalje, neke od varijabli imaju dva značajna faktorska koeficijenta,

točnije, to su varijable P27, P28, P29, P31, P44 i P47. Analizirajmo te cross-loadinge:

Varijabla	Koeficijent_1	Koeficijent_2	Koeficijent_1^2	Koeficijent_2^2	Omjer	Klasifikacija
P27	0.41	0.64	0.17	0.41	2.41	zanemarivo
P28	0.41	0.41	0.17	0.17	1	problematično
P29	0.44	0.56	0.19	0.31	1.61	upitno
P31	0.43	0.63	0.18	0.40	2.22	zanemarivo
P44	0.49	0.43	0.24	0.18	1.33	problematično
P47	0.42	0.49	0.18	0.24	1.33	problematično

Tablica 4.13: Klasifikacija cross-loadinga faktorskih koeficijenata upitnika SPTKTT

Iz Tablice 4.13 iščitavamo da imamo 3 problematična cross-loadinga, trebalo bi razmotriti ideju uklanjanja tih varijabli (P28, P44 i P47). Radi jednostavnosti zanemarit ćemo problematične cross-loadinge i nastaviti s interpretacijom faktora.

Svaki se faktor može imenovati na temelju varijabli koje imaju značajne koeficijente za faktore, navodimo faktore i pripadne čestice značajnih koeficijenata:

- Faktor 1: P7 (Tehnološko znanje) i P35-P43 (Tehnološko pedagoško znanje)
- Faktor 2: P20-P26 (Pedagoško znanje), P28 i P30 (Pedagoško sadržajno znanje)
- Faktor 3: P1-P6 (Tehnološko znanje)
- Faktor 4: P8-P10 (Sadržajno znanje matematike)
- Faktor 5: P14-P16 (Sadržajno znanje znanosti)
- Faktor 6: P32 i P34 (Tehnološko sadržajno znanje) i P45-P47 (Tehnološko pedagoško sadržajno znanje)
- Faktor 7: P27 i P29 (Pedagoško sadržajno znanje), P31, P33 (Tehnološko sadržajno znanje), P44 (Tehnološko pedagoško sadržajno znanje)
- Faktor 8: P18-P19 (Sadržajno znanje pismenosti)
- Faktor 9: P11-P13 (Sadržajno znanje društvenih znanosti).

Faktorizacija upitnika SPTKTT razlikuje se od one u radovima [8] i [22]. Na uzorku učitelja su se pojavili nešto drugačiji faktori nego u navedenim istraživanjima koja su provedena na uzorku studenata. Faktore možemo imenovati na sljedeći način:

- Faktor 1: Tehnološko pedagoško znanje
- Faktor 2: Pedagoško znanje i pedagoško sadržajno znanje pismenosti i društvenih znanosti
- Faktor 3: Tehnološko znanje
- Faktor 4: Sadržajno znanje matematike
- Faktor 5: Sadržajno znanje znanosti
- Faktor 6: Tehnološko i tehnološko pedagoško sadržajno znanje pismenosti i društvenih znanosti
- Faktor 7: Tehnološko i pedagoško sadržajno znanje matematike i znanosti

- Faktor 8: Sadržajno znanje pismenosti
- Faktor 9: Sadržajno znanje društvenih znanosti.

Treba napomenuti da se proces imenovanja faktora temelji na subjektivnom mišljenju istraživača.

Faktorska analiza je na temelju uzorka upitnik podijelila na čestice koje ispituju sadržajna znanja (matematika, znanost, društvene znanosti, pismenost), tehnološka znanja i tehnološko pedagoško znanja slično kao i u [8]. Faktori koji su drugačiji su faktori koji razlikuju radi li se o tehnološkim i pedagoškim sadržajnim znanjima o pismenosti i društvenim znanostima ili o matematici i znanosti (faktor 2, faktor 6, faktor 7). Ovo implicira da kod upitnika SPTKTT za ovaj uzorak za pedagoško sadržajno, tehnološko sadržajno i tehnološko pedagoško sadržajno znanje treba razlikovati koje se sadržajno znanje ispituje, je li riječ o matematici i znanosti ili o društvenim znanostima i pismenosti.

Dodatak

Uporaba tehnologije
1. Znam riješiti tehničke probleme s kojima se suočavam.
2. Lako svladavam tehnologiju.
3. Držim korak s važnim novim tehnologijama.
4. Često se poigravam tehnologijom.
5. Znam mnogo o različitim tehnologijama.
6. Posjedujem tehničke vještine potrebne za upotrebu tehnologije.
7. Imao/imala sam dovoljno prilika za rad s različitim tehnologijama.
Poznavanje sadržaja
Matematika
8. Posjedujem dovoljno znanja iz matematike.
9. Mogu upotrebljavati matematički način razmišljanja.
10. Vlastito razumijevanje matematike razvijam različitim strategijama i na različite načine.
Društvene znanosti
11. Posjedujem dovoljno znanja iz društvenih znanosti.
12. Mogu koristiti povijesni način razmišljanja.
13. Vlastito razumijevanje društvenih znanosti razvijam različitim strategijama i na različite načine.
Znanost
14. Posjedujem dovoljno znanja iz znanosti.
15. Mogu koristiti znanstveni način razmišljanja.
16. Vlastito razumijevanje znanosti razvijam različitim strategijama i na različite načine.
Pismenost
17. Posjedujem dovoljno znanja iz pismenosti.
18. Mogu koristiti književni način razmišljanja.
19. Vlastito razumijevanje pismenosti razvijam različitim strategijama i na različite načine.
Pedagoško znanje
20. Znam kako vrjednovati postignuća učenika u razredu.
21. Prilagođavam poučavanje onome što učenici trenutno razumiju ili ne razumiju.
22. Prilagođavam stil poučavanja različitim učenicima.
23. Znam vrjednovati znanje učenika na različite načine.
24. Koristim široki raspon pristupa poučavanju ovisno o razrednim uvjetima.
25. Upoznat/upoznata sam s uobičajenim učeničkim razumijevanjem i zabludama.
26. Znam kako organizirati i upravljati razredom.
Sadržajno znanje u pedagoškom kontekstu
27. Odabirem efektivne načine poučavanja u svrhu usmjeravanja razmišljanja i učenja učenika u matematici.
28. Odabirem efektivne načine poučavanja u svrhu usmjeravanja razmišljanja i učenja učenika u društvenim znanostima.
29. Odabirem efektivne načine poučavanja u svrhu usmjeravanja razmišljanja i učenja učenika u znanosti.
30. Odabirem efektivne načine poučavanja u svrhu usmjeravanja razmišljanja i učenja učenika u pismenosti.
Tehnološka znanja u kontekstu sadržaja
31. Znam koje tehnologije mogu upotrebljavati kako bih osigurao razumijevanje i primjenu matematike.
32. Znam koje tehnologije mogu upotrebljavati kako bih osigurao razumijevanje i primjenu društvenih znanosti.
33. Znam koje tehnologije mogu upotrebljavati kako bih osigurao razumijevanje i primjenu znanosti.
34. Znam koje tehnologije mogu upotrebljavati kako bih osigurao razumijevanje i primjenu pismenosti.
Tehnološka znanja u pedagoškom kontekstu
35. Mogu izabrati tehnologiju koja će unaprijediti poučavanje tijekom nastavnog sata.
36. Mogu izabrati tehnologiju koja će unaprijediti učenikovo učenje tijekom nastavnog sata.
37. Moj program učiteljskog obrazovanja potaknuo me na dublje razmišljanje o utjecaju tehnologije na pristup poučavanju koji koristim u razredu.
38. Kritički razmišljam o načinima upotrebe tehnologije u razredu.
39. Mogu prilagoditi svoje znanje o tehnologiji različitim aktivnostima tijekom poučavanja.
40. Mogu izabrati tehnologiju za rad u razredu koja unaprjeđuje sadržaj i način poučavanja kao i ono što učenici nauče.

41. Mogu upotrebljavati strategije kombiniranja sadržaja, tehnologije i pristupa poučavanju u razredu o kojima sam učio/učila tijekom obrazovanja.
42. Mogu pružiti podršku drugima u koordinaciji upotrebe sadržaja, tehnologija i pristupa poučavanju u mojoj školi i/ili području.
43. Mogu izabrati tehnologije koje unaprjeđuju sadržaj nastavnog sata.
Tehnološko pedagoško sadržajno znanje (TPACK)
44. Mogu poučavati nastavne sadržaje koji na odgovarajući način kombiniraju matematiku, tehnologiju i pristupe poučavanju.
45. Mogu poučavati nastavne sadržaje koji na odgovarajući način kombiniraju pismenost, tehnologiju i pristupe poučavanju.
46. Mogu poučavati nastavne sadržaje koji na odgovarajući način kombiniraju društvene znanosti, tehnologiju i pristupe poučavanju.
47. Mogu poučavati nastavne sadržaje koji na odgovarajući način kombiniraju znanost, tehnologiju i pristupe poučavanju.

Tablica 4.14: Upitnik SPTKTT

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23
P1	1.00																						
P2	0.71	1.00																					
P3	0.62	0.71	1.00																				
P4	0.53	0.61	0.66	1.00																			
P5	0.61	0.67	0.72	0.75	1.00																		
P6	0.63	0.68	0.64	0.57	0.68	1.00																	
P7	0.42	0.51	0.48	0.40	0.51	0.44	1.00																
P8	0.22	0.18	0.12	0.08	0.16	0.18	0.42	1.00															
P9	0.24	0.24	0.21	0.13	0.23	0.19	0.44	0.82	1.00														
P10	0.21	0.16	0.21	0.12	0.20	0.12	0.43	0.72	0.81	1.00													
P11	0.12	0.14	0.17	0.07	0.15	0.11	0.38	0.22	0.22	0.18	1.00												
P12	0.05	0.11	0.16	0.08	0.11	0.10	0.35	0.09	0.15	0.12	0.56	1.00											
P13	0.05	0.12	0.17	0.11	0.13	0.11	0.36	0.09	0.15	0.20	0.55	0.73	1.00										
P14	0.18	0.17	0.28	0.21	0.29	0.23	0.44	0.39	0.38	0.35	0.35	0.32	0.43	1.00									
P15	0.15	0.17	0.21	0.20	0.24	0.19	0.43	0.35	0.41	0.37	0.21	0.34	0.40	0.76	1.00								
P16	0.15	0.17	0.26	0.22	0.25	0.17	0.41	0.32	0.37	0.38	0.24	0.31	0.43	0.73	0.80	1.00							
P17	0.19	0.19	0.19	0.09	0.14	0.16	0.36	0.12	0.17	0.05	0.53	0.37	0.37	0.20	0.14	0.17	1.00						
P18	0.08	0.18	0.19	0.13	0.10	0.16	0.36	-0.00	0.06	0.00	0.46	0.38	0.40	0.20	0.17	0.19	0.69	1.00					
P19	0.07	0.17	0.15	0.12	0.09	0.14	0.40	0.02	0.08	0.08	0.47	0.39	0.48	0.18	0.14	0.20	0.69	0.84	1.00				
P20	0.18	0.16	0.17	0.03	0.09	0.13	0.45	0.16	0.21	0.21	0.31	0.28	0.32	0.24	0.18	0.22	0.44	0.36	0.45	1.00			
P21	0.14	0.16	0.15	-0.01	0.01	0.08	0.40	0.25	0.26	0.27	0.33	0.25	0.31	0.24	0.23	0.24	0.38	0.32	0.39	0.70	1.00		
P22	0.17	0.18	0.15	0.08	0.09	0.11	0.39	0.18	0.19	0.19	0.33	0.28	0.31	0.26	0.12	0.23	0.36	0.38	0.40	0.62	0.67	1.00	
P23	0.21	0.24	0.18	0.02	0.13	0.12	0.44	0.23	0.24	0.29	0.31	0.27	0.35	0.27	0.20	0.25	0.37	0.31	0.39	0.74	0.71	0.73	1.00
P24	0.20	0.20	0.13	0.06	0.07	0.06	0.36	0.14	0.19	0.24	0.29	0.27	0.37	0.24	0.16	0.22	0.33	0.37	0.42	0.61	0.58	0.65	0.67
P25	0.17	0.21	0.13	0.08	0.12	0.09	0.45	0.21	0.23	0.19	0.31	0.27	0.31	0.27	0.21	0.20	0.32	0.33	0.39	0.50	0.48	0.43	0.48
P26	0.17	0.18	0.10	0.04	0.05	0.09	0.41	0.18	0.18	0.25	0.33	0.33	0.37	0.21	0.11	0.15	0.39	0.32	0.36	0.66	0.55	0.53	0.61
P27	0.15	0.13	0.16	0.10	0.11	0.01	0.42	0.48	0.47	0.56	0.09	0.07	0.17	0.32	0.25	0.37	0.04	0.01	0.10	0.32	0.32	0.30	0.36
P28	0.14	0.21	0.15	0.08	0.10	0.03	0.36	0.09	0.11	0.15	0.39	0.37	0.44	0.20	0.12	0.23	0.24	0.26	0.32	0.34	0.35	0.39	0.43
P29	0.16	0.19	0.23	0.19	0.21	0.09	0.48	0.39	0.39	0.43	0.19	0.14	0.19	0.49	0.49	0.59	0.10	0.06	0.09	0.28	0.27	0.26	0.34
P30	0.10	0.16	0.11	0.03	0.10	0.09	0.41	0.14	0.13	0.17	0.27	0.30	0.35	0.15	0.15	0.22	0.43	0.40	0.44	0.38	0.37	0.36	0.42
P31	0.29	0.32	0.31	0.25	0.28	0.23	0.56	0.48	0.51	0.60	0.15	0.11	0.14	0.33	0.30	0.34	0.10	0.04	0.11	0.30	0.30	0.25	0.32
P32	0.23	0.38	0.39	0.27	0.34	0.27	0.50	0.07	0.16	0.18	0.39	0.38	0.35	0.28	0.24	0.28	0.25	0.29	0.33	0.28	0.23	0.24	0.24
P33	0.26	0.38	0.40	0.35	0.37	0.29	0.57	0.31	0.35	0.34	0.14	0.18	0.19	0.44	0.46	0.47	0.11	0.12	0.12	0.25	0.23	0.21	0.23
P34	0.14	0.27	0.23	0.20	0.23	0.21	0.46	0.08	0.14	0.13	0.28	0.27	0.33	0.18	0.20	0.21	0.34	0.38	0.40	0.29	0.31	0.26	0.26
P35	0.22	0.39	0.29	0.18	0.25	0.34	0.44	0.14	0.17	0.15	0.21	0.16	0.22	0.25	0.16	0.18	0.29	0.26	0.28	0.36	0.30	0.31	0.34
P36	0.27	0.42	0.38	0.25	0.35	0.31	0.50	0.15	0.17	0.16	0.21	0.22	0.22	0.25	0.17	0.21	0.25	0.29	0.31	0.34	0.26	0.31	0.30
P37	0.24	0.34	0.25	0.19	0.21	0.21	0.37	0.17	0.17	0.13	0.17	0.11	0.21	0.33	0.26	0.25	0.15	0.22	0.22	0.18	0.21	0.19	0.16
P38	0.23	0.35	0.35	0.35	0.34	0.25	0.44	0.12	0.16	0.14	0.21	0.25	0.32	0.31	0.31	0.32	0.13	0.23	0.23	0.23	0.19	0.19	0.21
P39	0.39	0.51	0.50	0.45	0.54	0.51	0.58	0.15	0.18	0.15	0.20	0.18	0.26	0.28	0.27	0.25	0.27	0.21	0.23	0.25	0.19	0.23	0.23
P40	0.26	0.36	0.33	0.29	0.38	0.40	0.56	0.23	0.27	0.24	0.21	0.28	0.30	0.33	0.35	0.32	0.27	0.23	0.24	0.33	0.27	0.24	0.30
P41	0.31	0.37	0.34	0.28	0.38	0.35	0.56	0.19	0.23	0.21	0.28	0.22	0.26	0.33	0.29	0.27	0.24	0.24	0.24	0.35	0.28	0.34	0.36
P42	0.43	0.48	0.49	0.43	0.53	0.50	0.61	0.14	0.24	0.19	0.26	0.24	0.24	0.34	0.30	0.28	0.24	0.23	0.24	0.30	0.23	0.24	0.25
P43	0.40	0.46	0.47	0.43	0.52	0.45	0.65	0.20	0.22	0.22	0.30	0.27	0.30	0.32	0.28	0.28	0.29	0.24	0.29	0.34	0.22	0.27	0.28
P44	0.29	0.26	0.32	0.27	0.33	0.21	0.57	0.57	0.56	0.62	0.14	0.11	0.18	0.46	0.40	0.40	0.05	0.02	0.09	0.25	0.24	0.25	0.26
P45	0.23	0.26	0.28	0.25	0.31	0.19	0.53	0.29	0.32	0.34	0.23	0.24	0.25	0.29	0.32	0.29	0.27	0.33	0.34	0.27	0.24	0.25	0.24
P46	0.27	0.27	0.33	0.29	0.37	0.20	0.56	0.26	0.31	0.33	0.28	0.24	0.25	0.35	0.38	0.36	0.20	0.23	0.28	0.27	0.25	0.23	0.24
P47	0.28	0.32	0.34	0.30	0.34	0.22	0.58	0.41	0.41	0.44	0.18	0.19	0.22	0.54	0.54	0.50	0.11	0.12	0.17	0.26	0.22	0.21	0.26

	P24	P25	P26	P27	P28	P29	P30	P31	P32	P33	P34	P35	P36	P37	P38	P39	P40	P41	P42	P43	P44	P45	P46	P47
P1																								
P2																								
P3																								
P4																								
P5																								
P6																								
P7																								
P8																								
P9																								
P10																								
P11																								
P12																								
P13																								
P14																								
P15																								
P16																								
P17																								
P18																								
P19																								
P20																								
P21																								
P22																								
P23																								
P24	1.00																							
P25	0.55	1.00																						
P26	0.59	0.55	1.00																					
P27	0.31	0.22	0.33	1.00																				
P28	0.42	0.27	0.36	0.34	1.00																			
P29	0.29	0.19	0.25	0.67	0.45	1.00																		
P30	0.36	0.30	0.41	0.29	0.55	0.37	1.00																	
P31	0.29	0.29	0.35	0.67	0.18	0.51	0.24	1.00																
P32	0.22	0.26	0.26	0.18	0.45	0.29	0.31	0.43	1.00															
P33	0.17	0.23	0.26	0.42	0.22	0.62	0.25	0.64	0.60	1.00														
P34	0.21	0.33	0.34	0.15	0.29	0.20	0.58	0.36	0.63	0.48	1.00													
P35	0.27	0.31	0.34	0.13	0.22	0.13	0.19	0.25	0.36	0.29	0.35	1.00												
P36	0.27	0.34	0.27	0.16	0.26	0.20	0.20	0.30	0.42	0.34	0.37	0.79	1.00											
P37	0.28	0.28	0.18	0.18	0.19	0.25	0.14	0.27	0.23	0.33	0.16	0.31	0.37	1.00										
P38	0.25	0.31	0.23	0.16	0.29	0.26	0.22	0.28	0.36	0.33	0.34	0.39	0.52	0.48	1.00									
P39	0.23	0.28	0.25	0.16	0.19	0.21	0.19	0.32	0.40	0.39	0.35	0.52	0.51	0.42	0.58	1.00								
P40	0.26	0.26	0.34	0.19	0.19	0.23	0.16	0.34	0.34	0.35	0.27	0.61	0.55	0.37	0.53	0.71	1.00							
P41	0.29	0.35	0.32	0.21	0.28	0.28	0.23	0.37	0.40	0.39	0.35	0.52	0.54	0.47	0.47	0.58	0.65	1.00						
P42	0.26	0.32	0.29	0.21	0.16	0.25	0.12	0.44	0.45	0.46	0.33	0.44	0.48	0.39	0.45	0.66	0.66	0.64	1.00					
P43	0.24	0.34	0.28	0.16	0.22	0.24	0.25	0.37	0.52	0.46	0.41	0.53	0.59	0.38	0.56	0.68	0.64	0.60	0.66	1.00				
P44	0.24	0.30	0.27	0.60	0.16	0.50	0.17	0.72	0.35	0.59	0.28	0.25	0.32	0.35	0.30	0.36	0.35	0.39	0.44	0.41	1.00			
P45	0.26	0.28	0.24	0.21	0.32	0.26	0.45	0.36	0.48	0.44	0.56	0.34	0.41	0.28	0.38	0.32	0.36	0.42	0.37	0.44	0.56	1.00		
P46	0.24	0.31	0.22	0.24	0.35	0.40	0.40	0.36	0.53	0.49	0.48	0.31	0.41	0.28	0.41	0.36	0.32	0.41	0.37	0.46	0.55	0.80	1.00	
P47	0.25	0.26	0.24	0.47	0.23	0.60	0.25	0.54	0.39	0.69	0.31	0.27	0.32	0.34	0.36	0.36	0.38	0.39	0.40	0.44	0.74	0.59	0.69	1.00

Tablica 4.15: Korelacijski koeficijenti za upitnik SPTKTT

Literatura

[1] T. W. Anderson, An Introduction to Multivariate Statistical Analysis, John Wiley & Sons Ltd., New Jersey, 2003.

[2] T. W. Anderson, Asymptotic theory for principal component analysis, Ann. Math. Statist., Volume 34, Number 1, 122-148, 1963.

[3] D. Bakić, Linearna Algebra, Školska knjiga, Zagreb, 2008.

[4] M. Benšić, N. Šuvak, Uvod u vjerojatnost i statistiku, Sveučilište u Osijeku, Odjel za matematiku, Osijek, 2014.

[5] C. Chatfield, A. J. Collins, Introduction to Multivariate Analysis, Chapman and Hall, London, 1989.

[6] M. J. Crawley, The R book, John Wiley & Sons Ltd, London, 2013.

[7] K. Dobi Barišić, Utjecaj vršnjačke procjene i samoprocjene na pristup učenju i primjenu informacijske i komunikacijske tehnologije kod budućih učitelja, Varaždin, 2018.

[8] K. Dobi Barišić, B. Divjak, V. Kirinić, Education Systems as Contextual Factors in the Technological Pedagogical Content Knowledge Framework, Journal of Information and Organizational Sciences, Vol. 43. No. 2 pp. 163-183, 2019.

[9] D. Dua, C. Graff, UCI Machine Learning Repository `http://archive.ics.uci.edu/ml`. Irvine, CA: University of California, School of Information and Computer Science, 2020.

[10] G. H. Dunteman, Principal Components Analysis, Sage Publications, Newbury Park, 1989.

[11] C. D. Dziuban, E. C. Shirkey, When is a correlation matrix appropriate for factor analysis? Some decision rules, Psychological Bulletin, Vol. 81, No. 6, 358-361, 1974.

[12] A. Field, J. Miles, Z. Field, Discovering Statistics using R, Sage Publications, London, 2012.

[13] R. F. Gunst, R. L. Mason, Regression analysis and its application, Marcel Dekker inc, New York, 1980.

[14] J. F. Hair Jr., W. C. Black, B. J. Babin, R. E. Anderson, Multivariate Data Analysis, Cengage Learning EMEA, 2019.

[15] B. E. Hansen, Econometrics, Wisconsin, 2020. `https://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf`

[16] I. T. Jolliffe, Principal Component Analysis (second edition), Springer, New York, 2002.

[17] K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate analysis, Academic Press, London, 1995.

[18] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, W. B. Ford, The Population Biology of Abalone (Haliotis species) in Tasmania. I. BlacklipAbalone (H. rubra) from the North Coast and the Islands of Bass Strait, Sea Fisheries Division. Marine Research Laboratories, 1994.

[19] S. J. Press, Applied multivariate analysis, Holt, Rinehart and Winston, New York, 1972.

[20] A. C. Rechner, Methods of Multivariate Analysis, John Wiley & Sons Ltd., New York, 2002.

[21] N. Sarapa, Teorija vjerojatnosti, Školska knjiga, Zagreb, 2002.

[22] D. A. Schmidt, E. Baran, A. D. Thompson, P. Mishra, M. J. Koehler, T. S. Shin (2009) Technological Pedagogical Content Knowledge (TPACK), Journal of Research on Technology in Education, 42:2, 123-149, DOI

[23] N. Truhar, Numerička linearna algebra, Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za matematiku, Osijek, 2010.

Sažetak

U ovom su radu opisane statističke metode analiza glavnih komponenti i faktorska analiza. U prvom dijelu rada navedeni su osnovni pojmovi i rezultati koji su potrebni u nastavku.

U drugom poglavlju se bavimo analizom glavnih komponenti. Najprije definiramo i izvodimo glavne komponente, a zatim navodimo neka njihova algebarska svojstva i statističke implikacije tih svojstava. Nakon toga definiramo glavne komponente uzorka te se bavimo distribucijskim rezultatima za glavne komponente i procjenom parametara. Nadalje, osvrćemo se na nejedinstvenost glavnih komponenti za slučajan uzorak i njegove realizacije. Naime, glavne komponente nisu jedinstvena karakteristika podataka nego ovise o korištenim mjernim jedinicama i tipu matrice koja se koristi za izračun glavnih komponenti slučajnog uzorka. Zaključno, opisujemo jednu primjenu glavnih komponenti, primjenu glavnih komponenti u linearnoj regresiji.

U trećem poglavlju se bavimo faktorskom analizom. U tom poglavlju prvo definiramo faktorski model, a nakon toga opisujemo procjenu parametara faktorskog modela. U tom poglavlju navodimo i usporedbu metoda, tj. navodimo sličnosti i razlike, analize glavnih komponenti i faktorske analize.

U četvrtom poglavlju provodimo analizu glavnih komponenti i faktorsku analizu na primjerima. Kod primjera analize glavnih komponenti objašnjavamo kada ima smisla koristiti ovu metodu, kako odabrati broj glavnih komponenti za daljnju analizu, kako interpretirati zadržane glavne komponente te ilustriramo primjenu glavnih komponenti u linearnoj regresiji. Slično, kod primjera faktorske analize objašnjavamo kada su podaci prikladni za faktorsku analizu, kako odabrati broj faktora i kako interpretirati faktorsku strukturu.

Ključne riječi: analiza glavnih komponenti, kovarijacijska matrica, korelacijska matrica, spektralna dekompozicija, linearna regresija, faktorska analiza

Principal component analysis and factor analysis

Summary

In this thesis we describe two statistical methods, namely, principal component analysis and factor analysis. In the first chapter of this thesis we list basic concepts and results that are used throughout the thesis.

In the second chapter we deal with principal component analysis. We first define and derive the principal components. In addition, we list some of their algebraic properties and the statistical implications of those properties. After that, we define the sample principal components and deal with distributional results and parameter estimation for the sample principal components. Furthermore, we look at the non-uniqueness of the principal components for random samples and their realizations. The principal components are not a unique characteristic of the data but depend on the units of measurement used and the type of matrix used to construct the sample principal components. To conclude this chapter, we describe one application of principal components and that is how to use prinicpal components in linear regression.

In the third chapter, we first define the factor model and then describe the estimation of the factor model. In this chapter we also list the similarities and differences between principal component analysis and factor analysis.

In the fourth chapter, we perform principal component analysis and factor analysis on examples. In this chapter, we explain when to use these methods, how to select the number of principal components or factors, how to interpret the principal components or factors and we illustrate the application of principal components in linear regression.

Keywords: principal component analysis, covariance matrix, correlation matrix, spectral decomposition, linear regression, principal component regression, factor analysis

Životopis

Ana Vilić rođena je 05. travnja 1996. godine u mjestu Essen, Njemačka. Upisala je Osnovnu školu Tenja u Tenji 2003. godine. Nakon završene osnovne škole pohađa prirodoslovno-matematičku gimnaziju u Osijeku, tj. III. Gimnaziju Osijek. U srednjoj školi sudjeluje na županijskim natjecanjima iz njemačkog i engleskog jezika te iz povijesti. Sveučilišni preddiplomski studij matematike Odjela za matematiku Sveučilišta J. J. Strossmayera u Osijeku upisuje 2015. godine. Završetkom preddiplomskog studija 2018. godine stječe akademski naziv sveučilišne prvostupnice matematike izradom završnog rada na temu "Unitarni operatori" pod mentorstvom doc. dr. sc. Suzane Miodragović. Iste godine upisuje diplomski studij matematike, smjer financijska matematika i statistika na Odjelu za matematiku u Osijeku. Na diplomskom studiju radi kao demonstrator iz kolegija Statistički praktikum kod doc. dr. sc. Danijela Grahovca.