Metoda analize glavnih komponenti i primjene na tržište vrijednosnih papira

Lučić, Paula

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Paula Lučić

METODA ANALIZE GLAVNIH KOMPONENTI I PRIMJENE NA TRŽIŠTE VRIJEDNOSNIH PAPIRA

Diplomski rad

Voditelj rada:
izv. prof. dr. sc.
Nikola Sandrić

Zagreb, 2023.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem mentoru izv. prof. dr. sc. Nikoli Sandriću na pomoći i strpljenju prilikom pisanja ovog diplomskog rada.
Najveće hvala mojoj obitelji na podršci tijekom studiranja.

Sadržaj

Sadržaj **iv**

Uvod **1**

1 Izvod i svojstva glavnih komponenti **2**
1.1 Izvod 2
1.2 Algebarska svojstva 5
1.3 Geometrijska svojstva 8
1.4 Glavne komponente na temelju korelacijske matrice 10

2 Glavne komponente uzorka **12**
2.1 Notacija 12
2.2 Algebarska svojstva 13
2.3 Geometrijska svojstva 15

3 Grafički prikaz glavnih komponenti **20**
3.1 Prikaz podataka uzimajući u obzir dvije (ili tri) glavne komponente 20
3.2 Biplot 23

4 Odabir glavnih komponenti **29**
4.1 Kumulativni postotak ukupne varijacije 29
4.2 Vrijednost varijanci glavnih komponenti 30

5 Primjena - optimizacija dioničkog portfelja **32**

Dodatak **39**

Bibliografija **49**

Uvod

Metoda analize glavnih komponenti široko je korištena metoda koja pronalazi primjenu u raznim područjima kao što su financije i ekonomija, istraživanje hrane, genetika, meteorologija, oceanografija, psihologija i agrikultura. To je metoda redukcije dimenzionalnosti (i algoritam nenadziranog strojnog učenja) koja smanjuje kompleksnost višedimenzionalnih podataka zadržavajući bitne informacije o strukturi i trendovima podataka, na način da ih linearno transformira u novi koordinantni sustav u kojem je većina varijacije opisana manjim brojem varijabli. Time je olakšana vizualizacija podataka, a osobito u slučaju kada su dvije glavne komponente (dimenzije) dovoljne da se obuhvati većina varijacije, jer se na taj način mogu identificirati klasteri.

Moguće je naći najviše onoliko glavnih komponenti koliko ima varijabli u početnom skupu, a prva glavna komponenta odabire se tako da se maksimizira varijanca projektiranih točaka, odnosno da se minimizira udaljenost podataka od njihove projekcije na tu glavnu komponentu. Svaka iduća komponenta odabire se na isti način, uz dodatan uvjet o ortogonalnosti na prethodne komponente. Bit će pokazano kako se korištenje metode svodi na SVD dekompoziciju kovarijacijske ili korelacijske matrice polaznih podataka te da bitno značenje imaju njeni svojstveni vektori i pripadajuće svojstvene vrijednosti.

Rad je podijeljen u pet poglavlja. Prvo i drugo poglavlje bave se teorijskom pozadinom metode te su predstavljeni izvod i definicija glavnih komponenti uz bitna algebarska i geometrijska svojstva u kontekstu populacije i uzorka. Treće poglavlje obrađuje dva načina prikaza glavnih komponenti korištenih u praksi – prikaz koji uzima u obzir dvije ili tri početne glavne komponente te biplot. Četvrto poglavlje daje dva kriterija pomoću kojih možemo odrediti dovoljan broj glavnih komponenti za opisivanje polaznog skupa podataka. Konačno, u petom poglavlju, ranije navedena znanja primijenjuju se na podacima za 12 dionica S&P 500 indeksa.

Poglavlje 1

Izvod i svojstva glavnih komponenti

U prvom poglavlju definiramo analizu glavnih komponenti (skraćeno se piše PCA, kao *principal component analysis*), prikazujemo izvod glavnih komponenti te navodimo njihova bitna matematička i statistička svojstva.

Neka je $\mathbf{x}$ vektor p slučajnih varijabli $x_1, \ldots, x_p$ te promatramo varijance i strukturu kovarijanci (ili korelacija) između tih varijabli. U slučaju kada p nije mali ili struktura kovarijanci nije jednostavna, od interesa će nam biti pronaći nekoliko izvedenih varijabli s (pri čemu je $s \ll p$) koje sadrže većinu informacija o varijanci i kovarijanci početnih varijabli. PCA uzima u obzir kovarijance i korelacije, no usredotočuje se na varijance.
Najprije tražimo linearnu funkciju $\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}$ elemenata od $\mathbf{x}$ koja ima maksimalnu varijancu. Kako vektor $\boldsymbol{\alpha}_1$ može biti proizvoljno velik, postavit ćemo ograničenje u vidu njegove norme (u izvodu ćemo vidjeti da je korisno staviti $\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_1 = 1$) takvu da vrijedi:

$$\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1p}x_p,$$

pri čemu je $\boldsymbol{\alpha}_1$ vektor od p konstanti $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p}$.
Zatim tražimo linearnu funkciju $\boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{x}$ koja nije korelirana s prethodnom $\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}$ te ima maksimalnu varijancu. Nadalje, tražimo k-tu linearnu funkciju $\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{x}$ maksimalne varijance koja nije korelirana s $\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}, \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{x}, \ldots, \boldsymbol{\alpha}_{k-1}^{\mathrm{T}}\mathbf{x}$. Ta k-ta varijabla $\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{x}$ je k-ta glavna komponenta.

Jasno je da je na ovaj način moguće naći p takvih varijabli koje predstavljaju glavne komponente, no cilj ove metode je pronaći s glavnih komponenti ($s \ll p$) koje sadrže većinu varijacije od $\mathbf{x}$ te ćemo tako reducirati kompleksnost problema.

1.1 Izvod

Prikazat ćemo način na koji možemo izračunati glavne komponente. U najčešćem slučaju, kada kovarijacijska matrica $\boldsymbol{\Sigma}$ vektora $\mathbf{x}$ nije poznata, umjesto nje uzimamo kovarijacijsku

matricu uzorka $\mathbf{S}$. Pokazat će se da je k-ta glavna komponenta za $k = 1, 2, \ldots, p$ jednaka $z_k = \boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{x}$, gdje je $\boldsymbol{\alpha}_k$ svojstveni vektor od $\boldsymbol{\Sigma}$ koji odgovara k-toj najvećoj svojstvenoj vrijednosti λ_k. Nadalje, ako je $\boldsymbol{\alpha}_k$ jedinični vektor ($\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\alpha}_k = 1$), onda je $\mathrm{Var}(z_k) = \lambda_k$.

Uzmimo varijablu $\boldsymbol{\alpha}_1\mathbf{x}^{\mathrm{T}}$. Vektor $\boldsymbol{\alpha}_1$ maksimizira $\mathrm{Var}(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}) = \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$ uz ograničenje $\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_1 = 1$ (zbroj kvadrata elemenata od $\boldsymbol{\alpha}_1$ je jednak 1). Umjesto toga, također je moguće uvesti neka druga korisnija ograničenja u određenim situacijama (poput $\max_j \left|a_{1j}\right| = 1$). Gornji problem maksimizacije riješit ćemo metodom Lagrangeovih multiplikatora. Maksimiziramo

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

gdje je λ Lagrangeov multiplikator. Deriviranjem po $\boldsymbol{\alpha}_1$ dobijemo

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

gdje je $\mathbf{I}_p$ jedinična matrica reda p. Iz navedenog slijedi da je λ svojstvena vrijednost od $\boldsymbol{\Sigma}$ te je $\boldsymbol{\alpha}_1$ pripadajući svojstveni vektor. Kako bismo odredili koji od p svojstvenih vektora maksimizira varijancu od $\boldsymbol{\alpha}_1^{\mathrm{T}}x$, moramo maksimizirati

$$\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1^{\mathrm{T}}\lambda\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_1 = \lambda,$$

iz čega možemo zaključiti da λ mora biti najveća moguća svojstvena vrijednost. Tada je $\boldsymbol{\alpha}_1$ svojstveni vektor koji odgovara najvećoj svojstvenoj vrijednosti od $\boldsymbol{\Sigma}$ te je $\mathrm{Var}(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}) = \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda_1$ najveća svojstvena vrijednost.

Generalno će vrijediti da je $\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{x}$ k-ta glavna komponenta od $\mathbf{x}$ te je $\mathrm{Var}(\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{x}) = \lambda_k$, pri čemu je λ_k k-ta najveća svojstvena vrijednost od $\boldsymbol{\Sigma}$, a $\boldsymbol{\alpha}_k$ je pripadajući svojstveni vektor. To ćemo pokazati za slučaj $k = 2$, a za $k \geq 3$ će slijediti na sličan način.
Druga glavna komponenta, $\boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{x}$, maksimizira $\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$ te nije korelirana s $\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}$, odnosno $\mathrm{Cov}(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}, \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{x}) = 0$. No,

$$\mathrm{Cov}(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}, \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{x}) = \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2^{\mathrm{T}}\lambda_1\boldsymbol{\alpha}_1^{\mathrm{T}} = \lambda_1\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_2 \tag{1.1}$$

pa bilo koja od

$$\begin{aligned} \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = 0, \quad \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = 0, \\ \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_2 = 0, \quad \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_2 = 0, \end{aligned}$$

mora biti zadovoljena kako bismo imali uvjet nekoreliranosti između varijabli. Potpuno proizvoljno izabiremo zadnju jednadžbu te, ponovo uzimajući u obzir ograničenje $\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_2 = 1$, trebamo maksimizirati:

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

pri čemu su λ i ϕ Lagrangeovi multiplikatori. Kao i prije, deriviramo izraz po $\boldsymbol{\alpha}_2$ te dobivamo

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

te množenjem slijeva s $\boldsymbol{\alpha}_1^{\mathrm{T}}$ slijedi

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

Kako su prva dva pribrojnika jednaka 0 te je $\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_1 = 1$, proizlazi $\phi = 0$. Tada je $\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 = 0$, odnosno $(\boldsymbol{\Sigma} - \lambda\mathbf{I}_p)\boldsymbol{\alpha}_2 = 0$ pa je λ ponovo svojstvena vrijednost od $\boldsymbol{\Sigma}$, a $\boldsymbol{\alpha}_2$ pripadajući svojstveni vektor.
Jer je $\lambda = \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$, λ je što je moguće veća. Pretpostavimo da $\boldsymbol{\Sigma}$ nema ponavljajuće svojstvene vrijednosti, λ nije jednaka λ_1, jer bi u protivnom vrijedilo $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2$ pa onda ne bi vrijedio uvjet $\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_2 = 0$. Dakle, λ je druga najveća svojstvena vrijednost od $\boldsymbol{\Sigma}$, a $\boldsymbol{\alpha}_2$ pripadajući svojstveni vektor.
Na analogan način bi se pokazalo da su za treću, četvrtu, ..., p-tu glavnu komponentu vektori $\boldsymbol{\alpha}_3, \boldsymbol{\alpha}_4, \ldots, \boldsymbol{\alpha}_p$ svojstveni vektori od $\boldsymbol{\Sigma}$ koji odgovaraju svojstvenim vrijednostima $\lambda_3, \lambda_4, \ldots, \lambda_p$, pri čemu je svaka svojstvena vrijednost manja od prethodne. Dodatno,

$$\mathrm{Var}(\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{x}) = \lambda_k \qquad k = 1, 2, \ldots, p.$$

Označimo sa $\mathbf{z}$ vektor čiji je k-ti element z_k, odnosno k-ta glavna komponenta za $k = 1, 2, \ldots, p$. Podrazumijevamo da je k-ta glavna komponenta ona s najvećom varijancom, tj. ona kojoj pripada k-ta po veličini svojstvena vrijednost. Tada je

$$\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{x}, \tag{1.2}$$

pri čemu je $\mathbf{A}$ ortogonalna matrica čiji je k-ti stupac, $\boldsymbol{\alpha}_k$, k-ti svojstveni vektor od $\boldsymbol{\Sigma}$. Dakle, glavne komponente definirane su kao ortonormirana transformacija od $\mathbf{x}$. Iz prethodnog izvoda imamo

$$\boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}\boldsymbol{\Lambda}, \tag{1.3}$$

gdje je $\boldsymbol{\Lambda}$ dijagonalna matrica čiji je k-ti dijagonalni element λ_k (k-ta svojstvena vrijednost od $\boldsymbol{\Sigma}$), a $\lambda_k = \mathrm{Var}(\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{x}) = \mathrm{Var}(z_k)$. Izraz 1.3 možemo napisati na još dva načina koji će nam kasnije biti korisni, a proizlaze iz ortogonalnosti matrice $\mathbf{A}$:

$$\mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{A} = \boldsymbol{\Lambda} \tag{1.4}$$

i

$$\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^{\mathrm{T}}. \tag{1.5}$$

Sada ćemo formalno napisati definiciju glavne komponente.

Definicija 1.1.1. *Neka je* $\mathbf{x}$ *slučajni vektor dimenzije* p *s očekivanjem* $\boldsymbol{\mu}$ *i matricom kovarijance* $\boldsymbol{\Sigma}$*. Neka je* $\mathbf{A} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p)$ *ortogonalna matrica takva da je*

$$\mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{A} = \boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_p),$$

gdje su $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$ *svojstvene vrijednosti od* $\boldsymbol{\Sigma}$*, i* $\mathbf{y} = [(y_k)] = \mathbf{A}^{\mathbf{T}}(\mathbf{x} - \boldsymbol{\mu})$*. Tada je* $y_k = \boldsymbol{\alpha}_k^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu})$ *(k=1, 2, …, p) k-ta* ***glavna komponenta*** *od* $\mathbf{x}$*, a* $z_k = \lambda_k^{-1/2} y_k$ *je k-ta* ***standardizirana glavna komponenta*** *od* $\mathbf{x}$*.*

Kako je svojstveni vektor $\boldsymbol{\alpha}_k$ jedinični, primjećujemo da je y_k ortogonalna projekcija od $\mathbf{x} - \boldsymbol{\mu}$ u smjeru $\boldsymbol{\alpha}_k$.

U drugom i trećem potpoglavlju predstavit ćemo neka od optimalnih svojstava koje ima ortonormirana linearna transformacija $\mathbf{z}$ od $\mathbf{x}$ iz 1.2.

1.2 Algebarska svojstva

Propozicija 1.2.1. *Za neki cijeli broj* $q, 1 \leq q \leq p$*, neka je*

$$\mathbf{y} = \mathbf{B}^{\mathrm{T}}\mathbf{x}$$

ortonormirana linearna transformacija gdje je $\mathbf{y}$ *vektor duljine q, a* $\mathbf{B}^{\mathrm{T}}$ *je ortogonalna matrica dimenzije* $q \times p$ *te neka je* $\boldsymbol{\Sigma}_y = \mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B}$ *kovarijacijska matrica za* $\mathbf{y}$*. Tada trag od* $\boldsymbol{\Sigma}_y$*, u oznaci* $tr(\boldsymbol{\Sigma}_y)$*, postiže maksimum za* $\mathbf{B} = \mathbf{A}_q$*, gdje se* $\mathbf{A}_q$ *sastoji od prvih q stupaca od* $\mathbf{A}$*.*

Dokaz. Neka je β_k k-ti stupac od $\mathbf{B}$. Kako su stupci od $\mathbf{A}$ baza za p-dimenzionalni prostor, imamo

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

gdje su c_{jk}, $j = 1, 2, \ldots, p, k = 1, \ldots, q$, definirane konstante. Prema tome je $\mathbf{B} = \mathbf{AC}$, gdje je $\mathbf{C}$ matrica dimenzije $p \times q$ koja na mjestu (j, k) ima c_{jk} te je

$$\begin{aligned}\mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B} = \mathbf{C}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{AC} = \mathbf{C}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{C} \quad (koristimo 1.4)\\ = \sum_{j=1}^{p} \lambda_j \mathbf{c}_j \mathbf{c}_j^{\mathrm{T}},\end{aligned}$$

pri čemu je $\mathbf{c}_j^{\mathrm{T}}$ j-ti redak matrice $\mathbf{C}$. Stoga,

$$\begin{aligned}\operatorname{tr}(\mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B}) &= \sum_{j=1}^{p} \lambda_j \operatorname{tr}(\mathbf{c}_j\mathbf{c}_j^{\mathrm{T}}) \\ &= \sum_{j=1}^{p} \lambda_j \operatorname{tr}(\mathbf{c}_j^{\mathrm{T}}\mathbf{c}_j) \\ &= \sum_{j=1}^{p} \lambda_j \mathbf{c}_j^{\mathrm{T}}\mathbf{c}_j \\ &= \sum_{j=1}^{p}\sum_{k=1}^{q} \lambda_j c_{jk}^2. \end{aligned} \tag{1.6}$$

Posljednja jednakost slijedi jer je $\mathbf{c}_j^{\mathrm{T}}\mathbf{c}_j = \sum_{k=1}^{q} c_{jk}^2$ za svaki $j = 1, 2, \ldots, p$.
Sada imamo $\mathbf{C} = \mathbf{A}^{\mathrm{T}}\mathbf{B}$, pa $\mathbf{C}^{\mathrm{T}}\mathbf{C} = \mathbf{B}^{\mathrm{T}}\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{B} = \mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_q$ ($\mathbf{A}$ i $\mathbf{B}$ su ortogonalne). Kako je $\mathbf{C}^{\mathrm{T}}\mathbf{C}$ jedinična matrica reda q, na glavnoj dijagonali imat će 1 (q stupaca), a na ostalim mjestima 0. Kada to raspišemo pomoću suma, dobijemo da je

$$\sum_{j=1}^{p}\sum_{k=1}^{q} c_{jk}^2 = q, \tag{1.7}$$

odnosno $\operatorname{tr}(\mathbf{C}^{\mathrm{T}}\mathbf{C})$ je q. Dodatno, i stupci od $\mathbf{C}$ su ortonormirani. Matricu $\mathbf{C}$ možemo zamisliti kao prvih q stupaca neke ortogonalne matrice $\mathbf{D}$ dimenzije $p \times p$. No kako su retci matrice $\mathbf{D}$ ortonormirani, oni zadovoljavaju $\mathbf{d}_j^{\mathrm{T}}\mathbf{d}_j = 1$, $j = 1, \ldots, p$. Jer se retci od $\mathbf{C}$ sastoje od prvih q elemenata redaka od $\mathbf{D}$, slijedi da je $\mathbf{c}_j^{\mathrm{T}}\mathbf{c}_j \leq 1$, $j = 1, \ldots, p$ te

$$\sum_{k=1}^{q} c_{jk}^2 \leq 1. \quad (\text{zbog } \mathbf{C}^{\mathrm{T}}\mathbf{C} = \mathbf{I}_q) \tag{1.8}$$

$\sum_{k=1}^{q} c_{jk}^2$ je koeficijent od λ_j iz 1.6, suma tih koeficijenata je q (iz 1.7) i niti jedan nije veći od 1 (zbog 1.8). Zato što je $\lambda_1 > \lambda_2 \cdots > \lambda_p$, jasno je da će izraz $\sum_{j=1}^{p}(\sum_{k=1}^{q} c_{jk}^2)\lambda_j$ postići maksimum za skup koeficijenata c_{jk} za koje je

$$\sum_{k=1}^{q} c_{jk}^2 = \begin{cases} 1, & j = 1, 2, \ldots, q \\ 0, & j = q+1, \ldots, p. \end{cases} \tag{1.9}$$

Ali ako je $\mathbf{B}^{\mathrm{T}} = \mathbf{A}_q^{\mathrm{T}}$, onda je

$$c_{jk}^2 = \begin{cases} 1, & 1 \leq j = k \leq q \\ 0, & \text{inače.} \end{cases} \tag{1.10}$$

što zadovoljava 1.9. Stoga $\operatorname{tr}(\boldsymbol{\Sigma}_y)$ postiže maksimum za $\mathbf{B}^{\mathrm{T}} = \mathbf{A}_q^{\mathrm{T}}$. □

Propozicija 1.2.2. *Neka je ponovo*

$$\mathbf{y} = \mathbf{B}^{\mathrm{T}}\mathbf{x}$$

ortonormirana linearna transformacija, gdje su $\mathbf{x}, \mathbf{B}, \mathbf{A}$ *definirani kao i prije. Tada* $\mathrm{tr}(\boldsymbol{\Sigma}_y)$ *postiže minimum za* $\mathbf{B} = \mathbf{A}_q^*$*, gdje se* $\mathbf{A}_q^*$ *sastoji od zadnjih q stupaca od* $\mathbf{A}$*.*

Dokaz. Dokaz ovog svojstva može se dobiti na sličan način kao dokaz prethodnog svojstva uz male izmjene. □

Propozicija 1.2.3. *(Spektralna dekompozicija od* $\boldsymbol{\Sigma}$*) Kovarijacijska matrica* $\boldsymbol{\Sigma}$ *reda p može se zapisati kao*

$$\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^{\mathrm{T}} = \lambda_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^{\mathrm{T}} + \lambda_2\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2^{\mathrm{T}} + \cdots + \lambda_p\boldsymbol{\alpha}_p\boldsymbol{\alpha}_p^{\mathrm{T}}. \tag{1.11}$$

gdje je $\boldsymbol{\Lambda} = diag\{\lambda_1, \ldots, \lambda_p\}$ *matrica reda p koja sadrži svojstvene vrijednosti od* $\boldsymbol{\Sigma}$*, a matrica* $\mathbf{A}$ *je ortogonalna matrica čiji su stupci jedinični svojstveni vektori* $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$ *od* $\boldsymbol{\Sigma}$*.*

Dokaz. Iz 1.4 je $\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^{\mathrm{T}}$. Proširivanjem produkta matrice s desne strane dobivamo

$$\sum_{k=1}^{p} \lambda_k\boldsymbol{\alpha}_k\boldsymbol{\alpha}_k^{\mathrm{T}}, \tag{1.12}$$

što smo i trebali dobiti (vidi izvod od 1.6) □

Gornja propozicija nam nalaže da za dijagonalne elemente kovarijacijske matrice vrijedi

$$\mathrm{Var}(x_j) = \sum_{k=1}^{p} \lambda_k {\alpha_k}_j^2 \qquad j = 1, 2, \ldots, p$$

Dakle, varijancu svakog pojedinog elementa od $\mathbf{x}$ možemo rastaviti na padajuće pribrojnike s obzirom na glavne komponente. No, glavni rezultat ove propozicije je taj što možemo cijelu kovarijacijsku matricu rastaviti na padajuće (ne strogo padajuće) pribrojnike $\lambda_k\boldsymbol{\alpha}_k\boldsymbol{\alpha}_k^{\mathrm{T}}$ s obzirom na glavne komponente. Pribrojnici će padati kako k raste, jer će se λ_k (što je jednako varijanci glavne komponente) smanjivati, a elementi od $\boldsymbol{\alpha}_k$ bit će otprilike "jednaki" zbog ograničenja $\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\alpha}_k = 1$, $k = 1, 2, \ldots, p$. Već smo otprije uvidjeli (1.2.1) da glavne komponente uspješno opisuju $\mathrm{tr}(\boldsymbol{\Sigma})$, ali ova propozicija intuitivno opisuje i elemente od $\boldsymbol{\Sigma}$ izvan glavne dijagonale. Dodatno, rezultatom 1.2.3 moguće je konstruirati kovarijacijsku (ili korelacijsku) matricu ako su nam poznate varijance λ_k i koeficijenti $\boldsymbol{\alpha}_k$ prvih r glavnih komponenti za $k = 1, 2, \ldots, p$, gdje je r rang kovarijacijske matrice. Naime, ako je kovarijacijska matrica reda p te ranga r (simetrična je pa time i dijagonalizabilna), onda ona ima r pozitivnih ne-nul svojstvenih vrijednosti pa će u gornjem izrazu ostati samo pribrojnici s ne-nul svojstvenim vrijednostima.

Propozicija 1.2.4. *Neka je*

$$\mathbf{y} = \mathbf{B}^{\mathrm{T}}\mathbf{x}$$

ortonormirana linearna transformacija, kao u 1.2.1 i 1.2.2. Ako je det $(\mathbf{\Sigma}_y)$ *determinanta kovarijacijske matrice od* $\mathbf{y}$*, onda det* $(\mathbf{\Sigma}_y)$ *postiže maksimum za* $\mathbf{B} = \mathbf{A}_q$*.*

Dokaz. Dokaz ovog teorema može se pronaći u [4]. □

Propozicija 1.2.5. *Neka je* $\mathbf{y} = \mathbf{B}^{\mathrm{T}}\mathbf{x}$*, pri čemu je* $\mathbf{y}$ *linearna funkcija kojom želimo predvidjeti slučajni vektor* $\mathbf{x} = (x_1, x_2, \ldots, x_p)$*. Za navedeno predviđanje* $\mathbf{y}$*, s* σ_j^2 *označimo varijancu njegova reziduala* $r_i = \|x_j - \mathbf{B}^{\mathrm{T}}x_j\|$ *u predviđanju varijable* x_j*. Tada je suma varijanci reziduala* $\sum_{j=1}^{p} \sigma_j^2$ *minimalna za* $\mathbf{B} = \mathbf{A}_q$*.*

Statistički gledano, gornje svojstvo govori da ako želimo pronaći najbolji linearni prediktor od $\mathbf{x}$ u q-dimenzionalnom potprostoru (u smislu minimiziranja sume varijanci reziduala), taj optimalni potprostor bit će razapet s prvih q glavnih komponenti.
Ovo svojstvo može se interpretirati i geometrijski te je zapravo ekvivalent geometrijskom svojstvu za uzorak koje će u idućem poglavlju biti izneseno skupa s dokazom, stoga ovdje nećemo iznositi dokaz gornjeg svojstva.

1.3 Geometrijska svojstva

Propozicija 1.3.1. *Neka je*

$$\mathbf{x}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x} = const \tag{1.13}$$

familija p-dimenzionalnih elipsoida, gdje je $\mathbf{x}$ *slučajni vektor dimenzije p s matricom kovarijance* $\mathbf{\Sigma}$*. Glavne osi tih elipsoida definirane su kao glavne komponente.*

Dokaz. Glavne komponente definirane su kao transformacija $\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{x}$. Kako je $\mathbf{A}$ ortogonalna matrica, inverzna transformacija je $\mathbf{x} = \mathbf{A}\mathbf{z}$. Kada to uvrstimo u izraz iz iskaza, dobivamo

$$(\mathbf{A}\mathbf{z})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{A}\mathbf{z}) = const = \mathbf{z}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{z}.$$

Znamo da su svojstveni vektori od $\mathbf{\Sigma}^{-1}$ isti kao od $\mathbf{\Sigma}$ te da su svojstvene vrijednosti od $\mathbf{\Sigma}^{-1}$ recipročne svojstvenim vrijednostima od $\mathbf{\Sigma}$ (pretpostavljamo da su sve strogo pozitivne). Nadalje slijedi (iz 1.4), da je $\mathbf{A}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{A} = \mathbf{\Lambda}^{-1}$ pa onda i $\mathbf{z}^{\mathrm{T}}\mathbf{\Lambda}^{-1}\mathbf{z} = const$. Zadnja jednakost se može napisati i kao

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

te je to jednadžba za elipsoid s obzirom na njegove glavne osi. Ona također implicira da su poluosi proporcionalne s $\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_p^{1/2}$. □

Ovaj rezultat je statistički bitan kada slučajni vektor $\mathbf{x}$ ima multivarijatnu normalnu razdiobu jer u tom slučaju elipsoidi zadani izrazom 1.13 definiraju konture na kojima se poprima jednaka vjerojatnost, odnosno funkcija distribucije slučajnog vektora $\mathbf{x}$ na tim konturama jednaka je konstanti. Tako prva (i najveća) glavna os tih elipsoida opisuje smjer u kojem je varijanca najveća, što je alternativni način kojim možemo definirati glavne komponente, kao što smo to algebarski učinili u prvom poglavlju. Mala (sekundarna) os elipse maksimizira varijancu te je ortogonalna na glavnu os, tako odgovarajući drugoj glavnoj komponenti (analogno za $i = 1, 2, \ldots, p$).

Propozicija 1.3.2. *Pretpostavimo da su $\mathbf{x}_1$, $\mathbf{x}_2$ nezavisni slučajni vektori s istom vjerojatnosnom distribucijom te su oba podvrgnuta istoj linearnoj transformaciji*

$$\mathbf{y}_i = \mathbf{B}^{\mathrm{T}}\mathbf{x}_i, \quad i = 1, 2.$$

Matrica $\mathbf{B}$ dimenzije $p \times q$ s ortonormiranim stupcima koja maksimizira $E[(\mathbf{y}_1 - \mathbf{y_2})^{\mathrm{T}}(\mathbf{y}_1 - \mathbf{y}_2)]$ je $\mathbf{B} = \mathbf{A}_q$.

Dokaz. Iznad navedeno svojstvo može se protumačiti kao algebarsko svojstvo jer dokaz doista je algebarski. No ono ima geometrijsku interpretaciju: očekivana kvadratna euklidska udaljenost u q-dimenzionalnom potprostoru, između dva vektora p slučajnih varijabli s istom distribucijom, bit će najveća ako je potprostor definiran s prvih q glavnih komponenti.
Uočimo da $\mathbf{x}_1$, $\mathbf{x}_2$ imaju isto očekivanje μ i kovarijacijsku matricu $\mathbf{\Sigma}$. Stoga i $\mathbf{y}_1$, $\mathbf{y}_2$ imaju isto očekivanje i kovarijacijsku matricu, $\mathbf{B}^{\mathrm{T}}\boldsymbol{\mu}$ i $\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{B}$.

$$\begin{aligned} E[(\mathbf{y}_1 - \mathbf{y}_2)^{\mathrm{T}}(\mathbf{y}_1 - \mathbf{y}_2)] &= E\{[(\mathbf{y}_1 - \mathbf{B}^{\mathrm{T}}\boldsymbol{\mu}) - (\mathbf{y}_2 - \mathbf{B}^{\mathrm{T}}\boldsymbol{\mu})]^{\mathrm{T}}[(\mathbf{y}_1 - \mathbf{B}^{\mathrm{T}}\boldsymbol{\mu}) - (\mathbf{y}_2 - \mathbf{B}^{\mathrm{T}}\boldsymbol{\mu})]\} \\ &\overset{\substack{\mathbf{x}_1,\mathbf{x}_2 \text{ nezavisni} \\ \mathbf{y}_1,\mathbf{y}_2 \text{ nezavisni}}}{=} E[(\mathbf{y}_1 - \mathbf{B}^{\mathrm{T}}\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{y}_1 - \mathbf{B}^{\mathrm{T}}\boldsymbol{\mu})] + E[(\mathbf{y}_2 - \mathbf{B}^{\mathrm{T}}\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{y}_2 - \mathbf{B}^{\mathrm{T}}\boldsymbol{\mu})]. \end{aligned}$$

Sada za $i = 1, 2$ imamo

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

Ali iz 1.2.1, $\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{B})$ je maksimalan za $\mathbf{B} = \mathbf{A}_q$, a gornji kriterij je jednak $2 \cdot \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{B})$, stoga je tvrdnja dokazana. □

1.4 Glavne komponente na temelju korelacijske matrice

Izvod i svojstva glavnih komponenti koji su ranije predstavljeni temelje se na svojstvenim vrijednostima i svojstvenim vektorima kovarijacijske matrice. U praksi se glavne komponente češće definiraju kao

$$\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{x}^{*}, \tag{1.14}$$

pri čemu se stupci od $\mathbf{A}$ sastoje od svojstvenih vektora *korelacijske matrice*, a $\mathbf{x}^{*}$ se sastoji od standardiziranih varijabli. Cilj ovog pristupa je naći glavne komponente standardizirane verzije $\mathbf{x}^{*}$ od $\mathbf{x}$, gdje je j-ti element od $\mathbf{x}^{*}$ jednak ${x_j}/{\sigma_{jj}}^{1/2}$, $j = 1, 2, \ldots, p$ (x_j je j-ti element od $\mathbf{x}$, a σ_{jj} je varijanca od x_j). Tada je kovarijacijska matrica od $\mathbf{x}^{*}$ jednaka korelacijskoj matrici od $\mathbf{x}$ te su glavne komponente od $\mathbf{x}^{*}$ dane s 1.14. Sva svojstva iz prethodna dva potpoglavlja i dalje vrijede za korelacijsku matricu, samo sada uzimamo glavne komponente od $\mathbf{x}^{*}$, umjesto od $\mathbf{x}$ (ili od neke druge transformacije od $\mathbf{x}$).

Može se činiti kako se glavne komponente dobivene iz korelacijske matrice mogu lako dobiti iz glavnih komponenti odgovarajuće kovarijacijske matrice jer je $\mathbf{x}^{*}$ jednostavna transformacija od $\mathbf{x}$, no svojstvene vrijednosti i svojstveni vektori korelacijske matrice nisu jednostavno povezani s onima dobivenim iz pripadajuće kovarijacijske matrice. Naime, ako glavne komponente dobivene iz korelacijske matrice izrazimo pomoću $\mathbf{x}$ (koristeći transformaciju $\mathbf{x}^{*}$ od $\mathbf{x}$), te glavne komponente u pravilu neće biti iste onima dobivenima iz korelacijske matrice $\mathbf{\Sigma}$. Razlog tomu je što su glavne komponente invarijantne s obzirom na ortogonalne transformacije od $\mathbf{x}$, ali općenito nisu invarijantne s obzirom na ostale transformacije. Kako transformacija $\mathbf{x}^{*}$ od $\mathbf{x}$ nije ortogonalna, tako ni glavne komponente za korelacijsku i kovarijacijsku matricu neće dati istu informaciju i ne mogu biti izvedene jedna iz druge.

Pri korištenju kovarijacijske matrice, postoji osjetljivost glavnih komponenti na mjerne jedinice u kojima su izraženi elementi od $\mathbf{x}$. Stoga, ako postoji velika razlika između varijanci elemenata od $\mathbf{x}$, onda će te varijable s najvećom varijancom dominirati među prvih par glavnih komponenti. To može biti opravdano ako su svi elementi od $\mathbf{x}$ izraženi u istim mjernim jedinicama (primjerice visina u centimetrima). U praksi se češće događa da su elementi od $\mathbf{x}$ izraženi u različitim mjernim jedinicama te će u tom slučaju struktura glavnih komponenti ovisiti o odabiru mjerne jedinice (ukoliko je uopće moguće sve izraziti zajedničkom mjerom). Tu problematiku prikazuje idući primjer.

Pretpostavimo da imamo dvije varijable x_1 i x_2 te da varijabla x_1 predstavlja duljinu koja može biti izražena u centimetrima ili milimetrima. Varijabla x_2 nije duljina već je, primjerice, težina u gramima. Kovarijacijske matrice u tom slučaju su iduće

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix} \qquad \text{i} \qquad \mathbf{\Sigma}_1 = \begin{pmatrix} 8000 & 440 \\ 440 & 8000 \end{pmatrix}$$

Prva glavna komponenta za $\mathbf{\Sigma}_1$ je $0.707x_1+0.707x_2$, a za $\mathbf{\Sigma}_2$ je $0.998x_1+0.055x_2$ te možemo primijetiti da promjena mjerne jedinice jedne varijable bitno utječe na strukturu glavnih komponenti. Naime, za kovarijacijsku matricu $\mathbf{\Sigma}_1$, prva komponenta stavlja jednake težine na obje varijable, dok za matricu $\mathbf{\Sigma}_2$, u prvoj komponenti dominira x_1. Nadalje, za $\mathbf{\Sigma}_1$ prva glavna komponenta sadrži 77.5% ukupne varijacije, dok za $\mathbf{\Sigma}_2$ sadrži 99.3% ukupne varijacije. Stoga se, posebice u ovakvim slučajevima, prednost daje glavnim komponentama iz korelacijske matrice jer one ne ovise o apsolutnim vrijednostima korelacija, već samo o njihovim omjerima. To vrijedi jer, ako pomnožimo sve elemente korelacijske matrice izvan glavne dijagonale istom konstantom, svojstveni vektori neće se promijeniti.

S druge strane, prednost kovarijacijske matrice nad korelacijskom je njeno korištenje u statističkom zaključivanju za populacijske glavne komponente iz glavnih komponenti uzorka. Više o tome u [4] (Potpoglavlje 3.7.).

Poglavlje 2

Glavne komponente uzorka

2.1 Notacija

Ovo poglavlje strukturno je jednako prethodnom, no umjesto razmatranja svojstava glavnih komponenti dobivenih na temelju populacijske kovarijance, sada ćemo uzimati kovarijancu uzorka. Mnoga svojstva ostaju ista, no ona koja su relevantna samo za glavne komponente uzorka bit će detaljnije opisana.
Najprije ćemo uvesti notaciju te izraze koji će kasnije biti korisni.

Promatramo n nezavisnih opservacija $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ p-dimenzionalnog slučajnog vektora $\mathbf{x}$. Neka je $\tilde{z}_{i1} = \mathbf{a}_1^{\mathrm{T}}\mathbf{x}_i$, $i = 1, 2, \ldots, n$ te neka je vektor koeficijenata $\mathbf{a}_1^{\mathrm{T}}$ takav da maksimizira uzoračku varijancu

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

uz uvjet $\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 1$, $\bar{z}_1 = \frac{1}{n}\sum_{i=1}^{n}\tilde{z}_{i1}$. Zatim neka je $\tilde{z}_{i2} = \mathbf{a}_2^{\mathrm{T}}\mathbf{x}_i$, $i = 1, 2, \ldots, n$ te ponovo biramo vektor koeficijenata $\mathbf{a}_2^{\mathrm{T}}$ takav da maksimizira uzoračku varijancu od $\tilde{z}_{i2}$ uz uvjet $\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 = 1$ te takav da $\tilde{z}_{i2}$ nije korelirana s $\tilde{z}_{i1}$ u uzorku. Ponavljajući ovaj postupak za $k = 1, 2, \ldots, p$, dobivamo definiciju glavnih komponenti uzorka kao u Potpoglavlju 1.1 Stoga je $\mathbf{a}_k^{\mathrm{T}}\mathbf{x}$ k-ta glavna komponenta uzorka, $k = 1, 2, \ldots, p$, a $\tilde{z}_{ik}$ je vrijednost za i-to opažanje k-te glavne komponente. Kada bismo proveli izvod iz Potpoglavlja 1.1, zamjenjujući populacijske varijance i kovarijance uzoračkima, dobili bismo da je $\mathrm{Var}(\tilde{z}_{ik}) = l_k$ te je l_k k-ta najveća svojstvena vrijednost uzoračke kovarijacijske matrice $\mathbf{S}$ (za opservacije $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$), a $\mathbf{a}_k$ je pripadajući svojstveni vektor, $k = 1, 2, \ldots, p$.

Definirajmo sada matrice $\tilde{\mathbf{X}}$, $\tilde{\mathbf{Z}}$ dimenzije $n \times p$ kojima je (i, k)-ti element jednak vrijednosti k-tog elementa $\tilde{x}_{ik}$ od $\mathbf{x}_i$, te jednak $\tilde{z}_{ik}$, redom. Matrice $\tilde{\mathbf{X}}$ i $\tilde{\mathbf{Z}}$ zadovoljavaju $\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}\mathbf{A}$,

gdje je $\mathbf{A}$ ortogonalna matrica dimenzije reda p kojoj je k-ti stupac $\mathbf{a}_k$. Ako je srednja vrijednost svakog elementa od $\mathbf{x}$ poznata i iznosi 0, onda kovarijacijska matrica uzorka iznosi $\mathbf{S} = \frac{1}{n}\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}$. U češćem slučaju, kada srednja vrijednost od $\mathbf{x}$ nije poznata, (j,k)-ti element od $\mathbf{S}$ je

$$\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_{ij}-\bar{x}_j)(\tilde{x}_{ik}-\bar{x}_k), \tag{2.1}$$

pri čemu je

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

U izrazu 2.1, zbog nepoznate srednje vrijednosti uzorka te radi nepristrane procjene kovarijacijske matrice, u nazivniku koristimo $n-1$ umjesto n (Besselova korekcija). Stoga se matrica $\mathbf{S}$ može zapisati kao

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}, \tag{2.2}$$

gdje je $\mathbf{X}$ $(n \times p)$ matrica (i, j)-tim elementom jednakim $(\tilde{x}_{ij} - \bar{x}_j)$. Izraz 2.2 će kasnije biti vrlo koristan. S x_{ij} ćemo označavati (i, j)-ti element matrice $\mathbf{X}$, pri čemu je $x_{ij} = \tilde{x}_{ij} - \bar{x}_j$. Za kraj, definirat ćemo matricu vrijednosti (”skorova”) glavnih komponenti (s elementima z_{ij}):

$$\mathbf{Z} = \mathbf{XA}. \tag{2.3}$$

Vrijednosti glavnih komponenti u matrici $\mathbf{Z}$ imat će jednake varijance i kovarijance kao one dane matricom $\tilde{\mathbf{Z}}$ no umjesto $\bar{z}_k$, srednje vrijednosti će biti 0.

Treba napomenuti kako su svojstveni vektori od $\frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$, i $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ jednaki, a svojstvene vrijednosti od $\frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$ su jednake svojstvenim vrijednostima od $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ pomnoženima s $\frac{1}{n-1}$. Zato će nekada biti praktičnije pozivati se na svojstvene vektore i svojstvene vrijednosti od $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, umjesto na one od $\mathbf{S}$.

2.2 Algebarska svojstva

Sada iznosimo svojstva iz prethodnog poglavlja koja, uz manje izmjene, ostaju jednaka i u kontekstu uzorka.

Definirajmo $\mathbf{y}_i = \mathbf{B}^{\mathrm{T}}\mathbf{x}_i$, $i = 1, 2, \ldots, n$, pri čemu je $\mathbf{B}$ matrica dimenzije $p \times q$ s ortonormiranim stupcima, kao u Propozicijama 1.2.1, 1.2.2, 1.2.4, 1.2.5. Svojstva navedena tim

propozicijama i dalje će vrijediti u kontekstu uzorka: zamijenimo $\mathbf{\Sigma}_y$ s uzoračkom matricom kovarijacije opservacija $\mathbf{y}_i$, $i = 1, 2, \ldots, n$ te je sada k-ti stupac matrice $\mathbf{A}$ označen s $\mathbf{a}_k$, a $\mathbf{A}_q$ i $\mathbf{A}_q^*$ predstavljaju, redom, prvih q i zadnjih q stupaca. Možemo primijeniti iste dokaze uz supstituciju populacijskh vrijednosti uzoračkima. Propozicija 1.2.3, odnosno spektralna dekompozicija također vrijedi za uzorke u idućem obliku:

$$\mathbf{S} = l_1\mathbf{a}_1\mathbf{a}_1^{\mathrm{T}} + l_2\mathbf{a}_2\mathbf{a}_2^{\mathrm{T}} + l_2, \ldots, l_p\mathbf{a}_p\mathbf{a}_p^{\mathrm{T}}$$

Iduće algebarsko svojstvo tiče se primjene glavnih komponenti u regresiji. Detaljnije o regresijskoj analizi i korištenoj terminologiji može se pronaći u [6].

Propozicija 2.2.1. *Pretpostavimo da se* $\mathbf{X}$ *sastoji od n opažanja p prediktorskih varijabli x izmjerenih oko njihovih srednjih vrijednosti uzorka te je pripadajuća jednadžba regresije*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

pri čemu je $\mathbf{y}$ *vektor s n opažanja na zavisnoj varijabli, također izmjerenih oko njihovih srednjih vrijednosti uzorka. Uvedimo transformaciju* $\mathbf{Z}$ *od* $\mathbf{X}$*,* $\mathbf{Z} = \mathbf{XB}$*, gdje je* $\mathbf{B}$ *ortogonalna matrica reda p. Jednadžba regresije može se zapisati kao*

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

gdje je $\boldsymbol{\gamma} = \mathbf{B}^{-1}\boldsymbol{\beta}$*. Uobičajeni procjenitelj najmanjih kvadrata za* $\boldsymbol{\gamma}$ *je* $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y}$*. Stoga elementi od* $\hat{\boldsymbol{\gamma}}$ *imaju najmanje moguće varijance ako je* $\mathbf{B}$ *jednako* $\mathbf{A}$*, odnosno matrici čiji je k-ti stupac k-ti svojstveni vektor od* $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ *i k-ti svojstveni vektor od* $\mathbf{S}$*. Zato se* $\mathbf{Z}$ *sastoji od vrijednosti glavnih komponenti uzorka za* $\mathbf{x}$*.*

Dokaz. Poznato je iz modela regresije da je kovarijacijska matrica procjenitelja najmanjih kvadrata $\hat{\gamma}$ proporcionalna sa $(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}$. Naime, iz

$$E[\hat{\boldsymbol{\gamma}}] = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}E[\mathbf{y}]$$

,

$$\begin{aligned} E[\hat{\boldsymbol{\gamma}}]E[\hat{\boldsymbol{\gamma}}]^{\mathrm{T}} &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}E[\mathbf{y}]((\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}E[\mathbf{y}])^{\mathrm{T}} \\ &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}E[\mathbf{y}]E[\mathbf{y}]^{\mathrm{T}}\mathbf{Z}((\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1})^{\mathrm{T}} \\ &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}E[\mathbf{y}]E[\mathbf{y}]^{\mathrm{T}}\mathbf{Z}((\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{\mathrm{T}})^{-1} \\ &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}E[\mathbf{y}]E[\mathbf{y}]^{\mathrm{T}}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1} \end{aligned}$$

i

$$\begin{aligned} E[\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^{\mathrm{T}}] &= E[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y}((\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y})^{\mathrm{T}}] \\ &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}E[\mathbf{y}\mathbf{y}^{\mathrm{T}}]\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}, \end{aligned}$$

oduzimanjem posljednja dva izraza dobivamo varijancu

$$\begin{aligned} E[\hat{\boldsymbol{\gamma}}]E[\hat{\boldsymbol{\gamma}}]^{\mathrm{T}} - E[\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^{\mathrm{T}}] &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}(E[\mathbf{y}\mathbf{y}^{\mathrm{T}}] - E[\mathbf{y}]E[\mathbf{y}]^{\mathrm{T}})\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1} \\ &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\operatorname{Var}(\mathbf{y})\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1} \\ &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\operatorname{Var}(\mathbf{y}) \end{aligned}$$

koja je proporcionalna s

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

jer je $\mathbf{B}$ ortogonalna matrica. Također zahtijevamo da $\operatorname{tr}(\mathbf{B}_q^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{B}_q)$, $q = 1, 2, \ldots, p$ bude minimalan ($\mathbf{B}_q$ se sastoji od prvih q stupaca od $\mathbf{B}$). No, kada zamijenimo $\boldsymbol{\Sigma}_y$ s $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, Propozicija 1.2.2 nalaže da se $\mathbf{B}_q$ mora sastojati od zadnjih q stupaca one matrice čiji je k-ti stupac k-ti svojstveni vektor od $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. Štoviše, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ ima jednake svojstvene vektore kao $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, jedino je njihov redoslijed obrnut, to jest, prvih q stupaca od $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ je jednako matrici $\mathbf{B}_q$. Kako ovo vrijedi za $q = 1, 2, \ldots, p$, gornja propozicija je dokazana. □

2.3 Geometrijska svojstva

Svojstvo iz Propozicije 1.3.1 za populaciju vrijedi i za uzorak kada se zamijeni $\boldsymbol{\Sigma}$ sa $\mathbf{S}$. Elipsoidi $\mathbf{x}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{x} = const$ više nemaju interpretaciju kontura na kojima je funkcija gustoće jednaka konstanti, sada su elipsoidi procjenitelji tih kontura kada $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ dolaze iz multivarijatne normalne razdiobe.

Svojstvo iz Propozicije 1.3.2 za populaciju primjenjuje se na uzorak na idući način. Pretpostavimo da transformiramo opservacije $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ već poznatom transformacijom

$$\mathbf{y}_i = \mathbf{B}^{\mathrm{T}}\mathbf{x}_i, \quad i = 1, 2, \ldots, n,$$

pri čemu je $\mathbf{B}$ matrica dimenzije $p \times q$ s ortonormiranim stupcima tako da su $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ projekcije od $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ na q-dimenzionalni potprostor. Tada će izraz

$$\sum_{h=1}^{n}\sum_{i=1}^{n}(\mathbf{y}_h - \mathbf{y}_i)^{\mathrm{T}}(\mathbf{y}_h - \mathbf{y}_i)$$

postići maksimalnu vrijednost za $\mathbf{y}_i = \mathbf{A}_q^{\mathrm{T}}\mathbf{x}_i$, a minimalan za $\mathbf{y}_i = \mathbf{A}_q^{*\mathrm{T}}\mathbf{x}_i$ (vidi dokaz Propozicije 1.3.2). Naime, ovo svojstvo govori da, ako projiciramo n opservacija na q-dimenzionalni potprostor, onda je zbroj kvadrata Euklidskih udaljenosti među svim parovima opservacija u potprostoru maksimalan kada je potprostor razapet s prvih q glavnih komponenti (minimalan kada je razapet sa zadnjih q glavnih komponenti). Dokaz ovog rezultata vrlo je sličan dokazu istog svojstva za populaciju te je zato izostavljen.

Iduća propozicija opisuje svojstvo koje je ekvivalent populacijskom svojstvu iz Propozicije 1.2.5 u kontekstu uzorka. Oba svojstva dovode se u vezu s metodom najmanjih kvadrata za svaku varijablu x_j na q varijabli sadržanih u $\mathbf{y}$.

Propozicija 2.3.1. *Kao ranije, pretpostavimo da transformiramo opservacije* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ *s* $\mathbf{y}_i = \mathbf{B}^{\mathrm{T}}\mathbf{x}_i$, $i = 1, 2, \ldots, n$, *gdje je* $\mathbf{B}$ *matrica dimenzije* $p \times q$ *s ortonormiranim stupcima, takva da su* $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ *projekcije od* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ *na q-dimenzionalni potprostor. Mjera valjanosti prilagodbe ("goodness-of-fit") tog q-dimenzionalnog potprostora s* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ *definirana je kao zbroj kvadrata okomitih udaljenosti* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ *od potprostora. Ta mjera je minimalna za* $\mathbf{B} = \mathbf{A}_q$.

Dokaz. Vektor $\mathbf{y}_i$ je ortogonalna projekcija od $\mathbf{x}_i$ na q-dimenzionalan potprostor definiran matricom $\mathbf{B}$. S $\mathbf{m}_i$ označimo poziciju vektora $\mathbf{y}_i$ s obzirom na koordinate ishodišta te neka je $\mathbf{r}_i = \mathbf{x}_i - \mathbf{m}_i$. Jer je $\mathbf{m}_i$ ortogonalna projekcija od $\mathbf{x}_i$ na q-dimenzionalni potprostor, onda je i vektor $\mathbf{r}_i$ ortogonalan na potprostor pa je $\mathbf{r}_i^{\mathrm{T}}\mathbf{m}_i = 0$. Nadalje, $\mathbf{r}_i^{\mathrm{T}}\mathbf{r}_i$ je kvadrat okomite udaljenosti $\mathbf{x}_i$ od potprostora te je stoga zbroj kvadrata okomitih udaljenosti $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ od potprostora jednaka

$$\sum_{i=1}^{n} \mathbf{r}_i^{\mathrm{T}}\mathbf{r}_i.$$

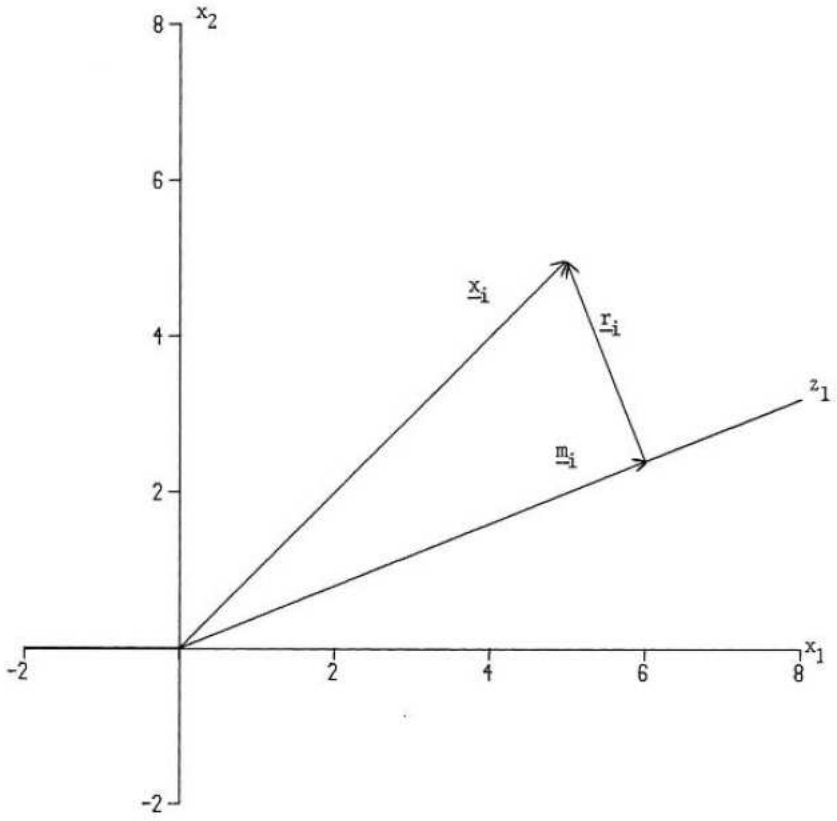
Sada je

$$\begin{aligned}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i &= (\mathbf{m}_i + \mathbf{r}_i)^{\mathrm{T}}(\mathbf{m}_i + \mathbf{r}_i)\\ &= \mathbf{m}_i^{\mathrm{T}}\mathbf{m}_i + \mathbf{r}_i^{\mathrm{T}}\mathbf{r}_i + 2\mathbf{r}_i^{\mathrm{T}}\mathbf{m}_i\\ &= \mathbf{m}_i^{\mathrm{T}}\mathbf{m}_i + \mathbf{r}_i^{\mathrm{T}}\mathbf{r}_i.\end{aligned}$$

Zato je,

$$\sum_{i=1}^{n} \mathbf{r}_i^{\mathrm{T}}\mathbf{r}_i = \sum_{i=1}^{n} \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{i=1}^{n} \mathbf{m}_i^{\mathrm{T}}\mathbf{m}_i,$$

pa za dani skup opservacija, umjesto minimiziranja sume kvadrata okomitih udaljenosti, jednako je maksimizirati $\sum_{i=1}^{n} \mathbf{m}_i^{\mathrm{T}}\mathbf{m}_i$. Kako ortogonalne transformacije ne mijenjaju udaljenosti, kvadrat udaljenosti $\mathbf{y}_i$ od ishodišta ($\mathbf{m}_i^{\mathrm{T}}\mathbf{m}$) ima iste x i y koordinate. Iz toga, trebamo



Slika 2.1: Ortogonalna projekcija dvodimenzionalnog vektora $\mathbf{x}_i$ na jednodimenzionalni potprostor razapet sa $\mathbf{z}_1$ (*slika preuzeta iz* [4])

maksimizirati $\sum_{i=1}^{n} \mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i$. Ali

$$\begin{aligned}
\sum_{i=1}^{n} \mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i &= \sum_{i=1}^{n} \mathbf{x}_i^{\mathrm{T}}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{x}_i \\
&= \mathrm{tr}\Big(\sum_{i=1}^{n} \mathbf{x}_i^{\mathrm{T}}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{x}_i\Big) \\
&= \sum_{i=1}^{n} \mathrm{tr}(\mathbf{x}_i^{\mathrm{T}}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{x}_i) \\
&= \sum_{i=1}^{n} \mathrm{tr}(\mathbf{B}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{B}) \\
&= \mathrm{tr}\left[\mathbf{B}^{\mathrm{T}}\left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\mathbf{B}\right] \\
&= \mathrm{tr}[\mathbf{B}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{B}] \\
&= (n-1)\,\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{S}\mathbf{B}).
\end{aligned}$$

Konačno, iz Propozicije 1.2.1, $\text{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{S}\mathbf{B})$ je maksimalan za $\mathbf{B} = \mathbf{A}_q$. □

Gornje svojstvo također se može smatrati alternativnim izvodom glavnih komponenti jer se one mogu definirati i geometrijski — glavne komponente su linearne funkcije (projekcije) od $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ koje definiraju potprostore dimenzija $1, 2, \ldots, q, \ldots, (p-1)$ te za koje je zbroj kvadrata okomitih udaljenosti između $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ i potprostora minimalan.

Propozicija 2.3.2. *Neka je* $\mathbf{X}$ *matrica dimenzije* $n \times p$ *čiji je* (i, j)*-ti element jednak* $\tilde{x}_{ij} - \bar{x}_j$ *te razmotrimo matricu* $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ *kojoj je* i*-ti element na dijagonali jednak* $\sum_{j=1}^{p}(\tilde{x}_{ij} - \bar{x}_j)$*. Posljednji izraz je Euklidska udaljenost* x_i *od* $\bar{x}$*, srednje vrijednosti točaka* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$*. Također,* (h, i)*-ti element od* $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ *je* $\sum_{j=1}^{p}(\tilde{x}_{hj} - \bar{x}_j)(\tilde{x}_{hj} - \bar{x}_j)$ *te on mjeri kosinus kuta između linija koje spajaju* $\mathbf{x}_h$ *i* $\mathbf{x}_i$ *s* $\bar{\mathbf{x}}$*, pomnožen udaljenostima* $\mathbf{x}_h$ *i* $\mathbf{x}_i$ *od* $\bar{\mathbf{x}}$*. Stoga* $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ *sadrži informacije o konfiguraciji (ili strukturi) točaka* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ *u odnosu na* $\bar{\mathbf{x}}$*.*
Sada pretpostavimo da projiciramo $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ *na* q*-dimenzionalni potprostor ortogonalnom transformacijom* $\mathbf{y}_i = \mathbf{B}^{\mathrm{T}}\mathbf{x}_i$*,* $i = 1, 2, \ldots, n$*. Tada transformacija za koju je* $\mathbf{B} = \mathbf{A}_q$ *minimizira iskrivljenje u konfiguraciji (strukturi) mjereno pomoću* $\|\mathbf{Y}\mathbf{Y}^{\mathrm{T}} - \mathbf{X}\mathbf{X}^{\mathrm{T}}\|$ *(*$\|\cdot\|$ *je oznaka Euklidske norme, a* $\mathbf{Y}$ *je matrica s* (i, j)*-tim elementom jednakim* $\tilde{y}_{ij} - \bar{y}_j$*).*

Dokaz. Dokaz propozicije može se naći u [4]. □

Primijetimo kako zbroj prvih q pribrojnika u spektralnoj dekompoziciji uzoračke kovarijacijske (ili korelacijske) matrice $\mathbf{S}$ daje matricu ${}_q\mathbf{S}$ koja minimizira $\|{}_q\mathbf{S} - \mathbf{S}\|$. Nadalje, $\|{}_q\mathbf{S} - \mathbf{S}\| = \sum_{k=q+1}^{p} l_k$, gdje je l_k k-ta svojstvena vrijednost od $\mathbf{S}$. Rezultat slijedi jer

$$\begin{aligned}
\left\|{}_q\mathbf{S} - \mathbf{S}\right\| &= \left\| \sum_{k=q+1}^{p} l_k \mathbf{a}_k \mathbf{a}_k^{\mathrm{T}} \right\| \\
&= \sum_{k=q+1}^{p} l_k \left\| \mathbf{a}_k \mathbf{a}_k^{\mathrm{T}} \right\| \\
&= \sum_{k=q+1}^{p} l_k \left[\sum_{i=1}^{p} \sum_{j=1}^{p} (a_{ki} a_{kj})^2 \right]^{1/2} \\
&= \sum_{k=q+1}^{p} l_k \left[\sum_{i=1}^{p} a_{ki}^2 \sum_{j=1}^{p} a_{kj}^2 \right]^{1/2} \\
&= \sum_{k=q+1}^{p} l_k
\end{aligned} \tag{2.4}$$

Svojstvo iz gornje propozicije vrlo je slično optimalnom svojstvu glavnih komponenti promatrano u smislu RV-koeficijenta (po Robertu i Escoufieru). RV-koeficijent je multivarijatna generalizacija kvadrata Pearsonovog koeficijenta korelacije, odnosno koeficijenta

determinacije r^2 te poprima vrijednosti između 0 i 1([8]). On je mjera sličnosti između dvije konfiguracije skupa od n točaka, opisanih s $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ i $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$. Udaljenost između te dvije konfiguracije definira se kao

$$\left\| \frac{\mathbf{X}\mathbf{X}^{\mathrm{T}}}{\{\mathrm{tr}\,(\mathbf{X}\mathbf{X}^{\mathrm{T}})^2\}^{1/2}} - \frac{\mathbf{Y}\mathbf{Y}^{\mathrm{T}}}{\{\mathrm{tr}\,(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^2\}^{1/2}} \right\|, \tag{2.5}$$

a nazivnici su uvedeni radi standardizacije, tako da je

$$\left\| \frac{\mathbf{X}\mathbf{X}^{\mathrm{T}}}{\{\mathrm{tr}\,(\mathbf{X}\mathbf{X}^{\mathrm{T}})^2\}^{1/2}} \right\| = \left\| \frac{\mathbf{Y}\mathbf{Y}^{\mathrm{T}}}{\{\mathrm{tr}\,(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^2\}^{1/2}} \right\| = 1,$$

Lako slijedi da je izraz 2.5 jednak $[2(1 - \mathrm{RV}((\mathbf{X}, \mathbf{Y}))]^{1/2}$, gdje je RV-koeficijent definiran kao

$$\mathrm{RV}(\mathbf{X}, \mathbf{Y}) = \frac{\mathrm{tr}(\mathbf{X}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\mathbf{X}^{\mathrm{T}})}{\{\mathrm{tr}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^2\,\mathrm{tr}(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^2\}^{1/2}}. \tag{2.6}$$

Mjera udaljenosti iz izraza 2.5 jednaka je uvjetu iz iskaza Propozicije 2.3.2, ne uzimajući u obzir standardizaciju. Stoga je minimiziranje izraza 2.5 ekvivalentno maksimizaciji RV-koeficijenta iz 2.6. Osim PCA, postoji još nekoliko multivarijatnih tehnika u kojima se optimalnost definira maksimizacijom RV-koeficijenta za odabrane $\mathbf{X}$ i $\mathbf{Y}$ (multivarijatna regresija, kanonička analiza, diskriminacijska analiza...više u [8]). Specijalno, ako je $\mathbf{Y}$ oblika $\mathbf{Y} = \mathbf{X}\mathbf{B}$, gdje je $\mathbf{B}$ $(p \times q)$ matrica takva da su stupci od $\mathbf{Y}$ nekorelirani, onda ćemo maksimizacijom $\mathrm{RV}(\mathbf{X}, \mathbf{Y})$ dobiti da je $\mathbf{B} = \mathbf{A}_q$, odnosno da je $\mathbf{Y}$ matrica vrijednosti za prvih q glavnih komponenti. RV-koeficijent također se koristi kod odabira podskupa varijabli.

Poglavlje 3

Grafički prikaz glavnih komponenti

Postoje mnoge metode prikaza višedimenzionalnih podataka, no za potrebe ovog rada obradit ćemo dvije – prikaz s obzirom na prve dvije/tri glavne komponente i biplot. Ideja za prvu metodu vrlo je jednostavna: prikažemo li na grafu vrijednosti, primjerice, prve dvije glavne komponente za svaku opservaciju, dobit ćemo najbolji dvodimenzionalni prikaz podataka.
Biplot dolazi u mnogim varijantama, no svaka daje prikaz n opservacija i p varijabli na istom dvodimenzionalnom grafu. U nekim varijantama on će čak dati graf kao u prethodno navedenoj metodi, no općenito je specifičnost biplota ta što, osim prikaza opservacija, simultano prikazuje i informaciju o odnosu među varijablama.
Od ostalih metoda valja istaknuti analizu glavnih koordinata i analizu korespodencije. Analiza glavnih koordinata daje dvodimenzionalni prikaz podataka formiranih u matricu sličnosti/različitosti, dok analiza korespodencije također daje dvodimenzionalni graf, no za posebno oblikovane podatke. Naime, ona se provodi na kontigencijskoj tablici gdje su podaci raspoređeni s obzirom na dvije kategoričke varijable, odnosno na matrici $\mathbf{N} = \{n_{ij} : i = 1, 2, \ldots, r; j = 1, 2, \ldots, c\}$, pri čemu je n_{ij} broj opservacija koje poprimaju i-tu vrijednost za prvu varijablu i j-tu vrijednost za drugu varijablu.

3.1 Prikaz podataka uzimajući u obzir dvije (ili tri) glavne komponente

Ako skup podataka $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ ima p varijabli, onda opservacije mogu biti prikazane kao točke u p-dimenzionalnom prostoru. Nadalje, želimo li prikazati podatke na grafu u "najbolje prilagođenom" q-dimenzionalnom prostoru ($q < p$), onda će odgovarajući potprostor biti definiran s prvih q glavnih komponenti. Pri tome je "najbolje prilagođeni" prostor definiran u kontekstu mjere valjanosti prilagodbe iz Propozicije 2.3.1. Dvodimenzionalni

grafovi su uobičajeno vrlo korisni za uočavanje određenih pravilnosti u podacima, dok su trodimenzionalni grafovi teži za interpretirati, ali mogu dati dodatan uvid u podatke.

U slučaju da udaljenost podataka od dvodimenzionalnog ili trodimenzionalnog potprostora nije mala, tada ti prikazi neće dati zadovoljavajuću reprezentaciju. No, ako podaci leže blizu q-dimenzionalnog potprostora, većina varijacije bit će sadržana u prvih q glavnih komponenti te će graf opservacija s obzirom na te glavne komponente dati realističan prikaz podataka (naravno, ako bitan dio o strukturi podataka nije opisan glavnim komponentama manje varijance).

Gornji prikaz podataka predstavit ćemo na kratkom primjeru dobivenom u programskom jeziku R. Korišteni podaci uzeti su iz [3]. Skup podataka sastoji se od 505 dionica burzovnog indeksa S&P 500 kojima se aktivno trguje u SAD-u. Uz dionice popisano je 10 financijskih pokazatelja, primjerice, cijena dionice, dividendni prinos (*dividend yield*), najveća i najmanja cijena u zadnjih godinu dana, tržišna kapitalizacija (*market cap*; ukupna vrijednost svih izdanih dionica), EBITDA (dobit prije kamata, poreza i amortizacije) itd. Na taj skup podataka, korištenjem funkcije `prcomp`, primijenili smo analizu glavnih kom-

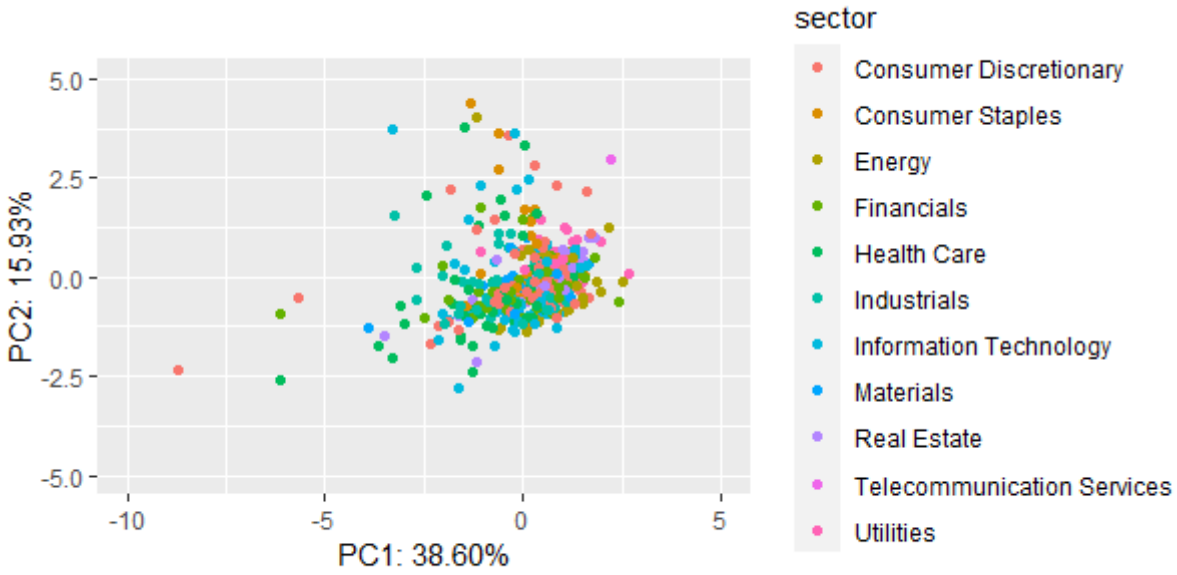
		1. komponenta	**2. komponenta**
Price	koeficijenti	-0.49	-0.12
Price Earnings		-0.12	-0.16
Dividend Yield		0.14	0.36
Earnings Share		-0.35	-0.04
52w low		-0.49	-0.13
52w High		-0.49	-0.12
Market Cap		-0.29	0.55
EBITDA		-0.18	0.67
Price Sales		-0.10	-0.17
Price Book		-0.02	0.11
Svojstvena vrijednost:		3.86	1.593
Kumulativni postotak od ukupne varijance:		0.386	0.5454

Tablica 3.1: Svojstvene vrijednosti i svojstveni vektori za prve dvije glavne komponente - dobiveni analizom glavnih komponenti.

ponenti za korelacijsku matricu podataka. Kumulativni postotak od ukupne varijance u

zadnjem retku Tablice 3.1 predstavlja zbroj postotaka objašnjene varijance za svaku glavnu komponentu (odnosno zbroj udjela svojstvenih vrijednosti glavnih komponenti).

U Tablici 3.1 prikazujemo dobivene vrijednosti za prve dvije glavne komponente te varijable, odnosno financijske pokazatelje. Vidimo da je prva komponenta u gotovo svim koeficijentima negativna. Razlog tomu je taj što je prva komponenta tržišna komponenta, ona reproducira ponašanje cijelog tržišta, tj. indeksa. Uz podatke o financijskim pokazateljima za svaku dionicu, imamo i podatak o sektoru kojem ona pripada.



Slika 3.1: Prikaz S&P 500 indeksa s financijskim pokazateljima - uzimajući u obzir prve dvije glavne komponente.

Graf 3.1 daje prikaz navedenih podataka s obzirom na prve dvije glavne komponente, pri čemu su različitim bojama označeni sektori.
Ako istaknemo, na primjer, sektore *Health Care* i *Utilities*, na grafu 3.2 vidimo da su podaci podijeljeni u dva klastera. To znači da nam metoda potencijalno može dati korisne zaključke o sličnom ponašanju dionica po sektorima te također indicira kako podaci pokazuju razlike među sektorima. No, kako smo spomenuli ranije, pojednostavljeni dvodimenzionalni prikaz pokazat će približno realno stanje jedino ako skup podataka nije previše udaljen od dvodimenzionalnog potprostora. Stoga, valja uzeti u obzir kako u našem slučaju

prve dvije glavne komponente sadrže otprilike 54% od ukupne varijacije, a Poglavlje 4 dat će nam odgovor je li taj postotak dovoljno dobar.

Slika 3.2: Prikaz sektora *Health Care* i *Utilities* S&P 500 indeksa s financijskim pokazateljima - uzimajući u obzir prve dvije glavne komponente.

3.2 Biplot

Biplot je postupak korišten u statistici kojim se grafički opisuju veze između p-dimenzionalnih opservacija $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ te veze između p varijabli. Temelji se na činjenici da se bilo koja matrica $\mathbf{B}$ dimenzije $n \times m$ i ranga r može na nejedinstven način faktorizirati kao

$$\mathbf{B} = \mathbf{G}\mathbf{H}^{\mathrm{T}}, \tag{3.1}$$

gdje je $\mathbf{G}$ matrica tipa $n \times r$, a $\mathbf{H}$ matrica tipa $m \times r$ te su obje ranga r. Stoga je

$$b_{ij} = \mathbf{g}_i^{\mathrm{T}} \mathbf{h}_j,$$

gdje su $\mathbf{g}_i^{\mathrm{T}}$ i $\mathbf{h}_j^{\mathrm{T}}$, redom, retci matrica $\mathbf{G}$ i $\mathbf{H}$ te smo tako dobili reprezentaciju od b_{ij} pomoću r-dimenzionalnih vektora. Kada je $r = 2$, $m + n$ vektora možemo prikazati grafom u

dvodimenzionalnom koordinatnom sustavu te tako dobiti biplot. U slučaju kada je $r > 2$, moguće je da postoji matrica $\mathbf{B}_{(2)}$ ranga 2 koja je dovoljno dobra aproksimacija matrice $\mathbf{B}$ te će odgovarajući biplot pobliže opisati samu matricu $\mathbf{B}$.

Izraz 3.1 možemo napisati kao $\mathbf{B} = (\mathbf{G}\mathbf{R}^{\mathrm{T}})(\mathbf{H}\mathbf{R}^{-\mathbf{1}})^{\mathrm{T}}$, za proizvoljnu regularnu matricu $\mathbf{R}$, te upravo zbog nejedinstvenosti odabira matrice $\mathbf{R}$ dobit ćemo različite biplotove. Osim ortogonalnom transformacijom, koja "čuva" udaljenosti i kutove između vektora, faktorizaciju možemo učiniti jedinstvenom ako uvedemo metriku za stupce matrica $\mathbf{G}$ i $\mathbf{H}$. Dodatno, želimo da matrice imaju određena svojstva, stoga ćemo koristiti singularnu dekompoziciju (**SVD**):

$$\mathbf{B} = \mathbf{L}_r\mathbf{\Delta}_r\mathbf{M}_r^{\mathrm{T}} = \sum_{i=1}^{r} \delta_i \mathbf{l}_i \mathbf{m}_i^{\mathrm{T}}, \tag{3.2}$$

gdje je $\mathbf{L}_r$ $(n \times r)$ matrica ranga r s ortogonalnim stupcima $\mathbf{l}_i$ $(\mathbf{L}_r^{\mathrm{T}}\mathbf{L}_r = \mathbf{I}_r)$, $\mathbf{M}_r$ je $(m \times r)$ matrica ranga r s ortogonalnim stupcima $\mathbf{m}_i$ $(\mathbf{M}_r^{\mathrm{T}}\mathbf{M}_r = \mathbf{I}_r)$, a $\mathbf{\Delta}_r = \mathrm{diag}(\delta_1, \delta_2, ..., \delta_r)$ te su $\delta_1 \geq \delta_2 \geq \cdots \delta_r > 0$ singularne vrijednosti od $\mathbf{B}$, odnosno pozitivni korijeni ne-nul svojstvenih vrijednosti od $\mathbf{B}^{\mathrm{T}}\mathbf{B}$. Iz 3.2, zanemarujući r kao oznaku ranga, dobivamo

$$\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{M} = \mathbf{M}\mathbf{\Delta}\mathbf{L}^{\mathrm{T}}\mathbf{L}\mathbf{\Delta}\mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{M}\mathbf{\Delta}^{\mathbf{2}} \tag{3.3}$$

ili

$$\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{m}_i = \delta_i^2 \mathbf{m}_i, \tag{3.4}$$

tako da su stupci matrice $\mathbf{M}$ svojstveni vektori od $\mathbf{B}^{\mathrm{T}}\mathbf{B}$. Dodatno, definirajmo dijagonalnu matricu $\mathbf{\Delta}^{\alpha}$, za $0 \leq \alpha \leq, 1$, s elementima $\delta_1^{\alpha/2}, \delta_2^{\alpha/2}, \ldots, \delta_r^{\alpha/2}$ te analogno matricu $\mathbf{\Delta}^{1-\alpha}$. Sada vidimo da $\mathbf{B}$ možemo zapisati u formi od 3.1:

$$\mathbf{B} = (\mathbf{L}\mathbf{\Delta}^{\alpha})(\mathbf{\Delta}^{\mathbf{1}-\alpha}\mathbf{M}^{\mathrm{T}}) = \mathbf{G}\mathbf{H}^{\mathrm{T}},$$

odnosno

$$\mathbf{G} = \mathbf{L}\mathbf{\Delta}^{\alpha}, \qquad \mathbf{H} = \mathbf{M}(\mathbf{\Delta}^{\mathbf{1}-\alpha})^{\mathrm{T}}, \qquad b_{ij} = \mathbf{g_i}^{\mathrm{T}}\mathbf{h_j} \tag{3.5}$$

te smo time dokazali egzistenciju te faktorizacije. U faktorizaciji 3.5 uočavamo kako je α varijabilna te će faktorizacija biti jedinstvena ako postavimo fiksnu vrijednost za α. Razmotrit ćemo slučajeve $\alpha = 0$ i $\alpha = 1$ jer će takvi α imati vrlo korisne interpretacije biplotova. Uz $\alpha = 0$, za naš odabir matrica $\mathbf{G}$ i $\mathbf{H}$ vrijedi iduće:

$$\mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{M}\mathbf{\Delta}^{\mathbf{2}}\mathbf{M}^{\mathrm{T}} = \sum_{i=1}^{r} \delta_i^2 \mathbf{m}_i\mathbf{m}_i^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{B}, \tag{3.6}$$

$$\begin{aligned}\mathbf{G}\mathbf{G}^{\mathrm{T}} &= \mathbf{L}\mathbf{L}^{\mathrm{T}} && (3.7)\\ &= (\mathbf{B}\mathbf{M}\boldsymbol{\Delta}^{-1})(\mathbf{B}\mathbf{M}\boldsymbol{\Delta}^{-1})^{\mathrm{T}} \qquad (\text{iz } 3.2\text{je } \mathbf{L} = \mathbf{B}\mathbf{M}\boldsymbol{\Delta}^{-1})\\ &= \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{\mathbf{-1}}\mathbf{B}^{\mathrm{T}} \qquad (\text{iz } 3.2 \text{ je } \mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{M}\boldsymbol{\Delta}^{\mathbf{2}}\mathbf{M}^{\mathrm{T}}) && (3.8)\end{aligned}$$

Ako želimo pronaći aproksimaciju od $\mathbf{B}$ ranga s ($s < r$), to možemo koristeći Frobeniusovu normu. Naime, neka je kao u našem slučaju, $\mathbf{B}$ ($n \times m$) matrica ranga r sa singularnom dekompozicijom $\mathbf{B} = \sum_{i=1}^{r} \delta_i \mathbf{l}_i \mathbf{m}_i^{\mathrm{T}}$ te neka je $\mathbf{C}$ ($n \times m$) matrica ranga $s < r$. Tada je

$$\|\mathbf{B} - \mathbf{C}\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(b_{ij} - c_{ij})^2 \tag{3.9}$$

minimalna za

$$\mathbf{C} = \mathbf{B}_{(s)} = \sum_{i=1}^{s} \delta_i \mathbf{l}_i \mathbf{m}_i^{\mathrm{T}}. \tag{3.10}$$

Izraz

$$\|\mathbf{B} - \mathbf{B}_{(s)}\|^2 = (n-1)(\delta_{s+1}^2 + \delta_{s+2}^2 + \cdots + \delta_r^2) \tag{3.11}$$

daje grešku aproksimacije, a valjanost prilagodbe modela ("*goodness of fit*") dana je izrazom

$$\begin{aligned}\rho_s^{(2)} &= 1 - (\|\mathbf{B} - \mathbf{B}_{(s)}\|^2 / \|\mathbf{B}\|^2) && (3.12)\\ &= \sum_{i=1}^{s} \delta_i^2 \Big/ \sum_{i=1}^{r} \delta_i^2. && (3.13)\end{aligned}$$

Gornji problem minimizacije naziva se *Eckart–Young–Mirsky* teorem te se njegov dokaz može pronaći u [9].

Sada gornju faktorizaciju primijenimo na matricu podataka $\mathbf{B} = \tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_n)^{\mathrm{T}}$, gdje je $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ i-ti redak matrice $\tilde{\mathbf{X}}$ i $m = p = r$. Tada je $\mathbf{B}^{\mathrm{T}}\mathbf{1}_n = \mathbf{0}$ i

$$\mathbf{B}^{\mathrm{T}}\mathbf{B} = \sum_{i=1}^{n} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^{\mathrm{T}} = \mathbf{Q} = (n-1)\mathbf{S}, \tag{3.14}$$

gdje je $\mathbf{S}$ nepristrana procjena kovarijacijske matrice od $\mathbf{x}_i$. Ako su $\delta_1^2 \geq \delta_2^2 \geq \cdots \geq \delta_p^2$ padajuće svojstvene vrijednosti od $\mathbf{Q}$, onda iz 3.6 i 3.7, $\mathbf{G} = (\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_p)$ i $\mathbf{H} = (\delta_1\mathbf{m}_1, \delta_2\mathbf{m}_2, \ldots, \delta_p\mathbf{m}_p)$, gdje je $\mathbf{G}$ dimenzije $n \times p$, a $\mathbf{H}$ dimenzije $p \times p$. Također, iz

3.6 i 3.8 je $\mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{Q}$ i $\mathbf{G}\mathbf{G}^{\mathrm{T}} = \tilde{\mathbf{X}}\mathbf{S}^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}/(n-1)$. Skaliranjem se možemo riješiti faktora $(n-1)^{-1}$ iz $\mathbf{G}\mathbf{G}^{\mathrm{T}}$:

$$\mathbf{G} = (n-1)^{1/2}(\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_p) \qquad \mathbf{H} = (n-1)^{-1/2}(\delta_1\mathbf{m}_1, \delta_2\mathbf{m}_2, \ldots, \delta_p\mathbf{m}_p). \tag{3.15}$$

Stoga, faktorizacija iz 3.1 i dalje vrijedi uz

$$\mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{S} \quad \text{i} \quad \mathbf{G}\mathbf{G}^{\mathrm{T}} = \tilde{\mathbf{X}}\mathbf{S}^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}, \tag{3.16}$$

a retci $\mathbf{g}_i^{\mathrm{T}}$ i $\mathbf{h}_j^{\mathrm{T}}$ matrica $\mathbf{G}$ i $\mathbf{H}$, redom, imaju iduća svojstva [7]:

(1) Iz 3.16, $\mathbf{g}_\alpha^{\mathrm{T}}\mathbf{g}_\beta = \tilde{\mathbf{x}}_\alpha^{\mathrm{T}}\mathbf{S}^{-1}\tilde{\mathbf{x}}_\beta$ te je Euklidska udaljenost između točaka reprezentiranih pomoću $\mathbf{g}_\alpha$ i $\mathbf{g}_\beta$ dana s

$$\begin{aligned} \|\mathbf{g}_\alpha - \mathbf{g}_\beta\| &= \sqrt{\mathbf{g}_\alpha^{\mathrm{T}}\mathbf{g}_\alpha + \mathbf{g}_\beta^{\mathrm{T}}\mathbf{g}_\beta - 2\mathbf{g}_\alpha^{\mathrm{T}}\mathbf{g}_\beta} \\ &= \sqrt{(\tilde{\mathbf{x}}_\alpha - \tilde{\mathbf{x}}_\beta)^{\mathrm{T}}\mathbf{S}^{-1}(\tilde{\mathbf{x}}_\alpha - \tilde{\mathbf{x}}_\beta)} \\ &= \sqrt{(\mathbf{x}_\alpha - \mathbf{x}_\beta)^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}_\alpha - \mathbf{x}_\beta)}, \end{aligned}$$

što je jednako Mahalanobisovoj udaljenosti između opservacija $\mathbf{x}_\alpha$ i $\mathbf{x}_\beta$. Mahalanobisovu udaljenost između dvije opservacije $\mathbf{x}_h$ i $\mathbf{x}_i$, uz pretpostavku da je $\mathbf{X}$ ranga r pa $\mathbf{S}^{-1}$ postoji, definiramo kao

$$\delta_{h_i}^2 = (\mathbf{x}_h - \mathbf{x}_i)\mathbf{S}^{-1}(\mathbf{x}_h - \mathbf{x}_i) \tag{3.17}$$

Ona se često koristi kao alternativa Euklidskoj udaljenosti $d_{h_i}^2 = (\mathbf{x}_h - \mathbf{x}_i)(\mathbf{x}_h - \mathbf{x}_i)$. Osnovna razlika je u tome što Euklidska udaljenost tretira sve varijable jednako, u smislu da pretpostavlja kako su sve varijable nekorelirane s jednakim varijancama, dok Mahalanobisova udaljenost pridodaje relativno manju težinu varijablama s velikom varijancom te skupovima vrlo koreliranih varijabli.

(2) Kovarijanca j-te i k-te varijable je

$$s_{j_k} = \mathbf{h}_j^{\mathrm{T}}\mathbf{h}_k,$$

a uzoračke varijance su $\|\mathbf{h}_j^2\|$.

(3) Korelacija između j-te i k-te varijable je kosinus kuta između $\mathbf{h}_j$ i $\mathbf{h}_k$ $(= \mathbf{h}_j^{\mathrm{T}}\mathbf{h}_k / \|\mathbf{h}_j^{\mathrm{T}}\mathbf{h}_k\| \cdot \|\mathbf{h}_j^{\mathrm{T}}\mathbf{h}_k\|)$

(4) Izraz

$$\begin{aligned}\|\mathbf{h}_j - \mathbf{h}_j\|^2 &= \mathbf{h}_j^{\mathrm{T}}\mathbf{h}_j + \mathbf{h}_k^{\mathrm{T}}\mathbf{h}_k - 2\mathbf{h}_j^{\mathrm{T}}\mathbf{h}_j\mathbf{h}_j^{\mathrm{T}}\mathbf{h}_k \\ &= s_{jj} + s_{kk} - 2s_{jk}\end{aligned}$$

je uzoračka varijanca razlike između j-te i k-te varijable.

Kao što smo ranije naveli, matricu $\mathbf{B}$ ranga r možemo aproksimirati matricom nižeg ranga $s < r$ (vidi 3.10). Taj izraz možemo također napisati kao

$$\begin{aligned}b_{i_j}^{(s)} &= \sum_{k=1}^{s} g_{ik}h_{j_k} \\ &= \mathbf{g}_i^{*\mathrm{T}}\mathbf{h}_j^*,\end{aligned}$$

gdje se $\mathbf{g}_i^*$ i $\mathbf{h}_j^*$ sastoje od prvih s elemenata vektora $\mathbf{g}_i$ i $\mathbf{h}_j$, respektivno. U slučaju da matrica $\mathbf{B}_{(2)}$ dobro aproksimira matricu $\mathbf{B}$ ($s = 2$), tada će $\mathbf{g}_i^*$ $i = 1, 2, \ldots, n$ i $\mathbf{h}_j^*$ $j = 1, 2, \ldots, p$, zajedno dati dobar dvodimenzionalni prikaz n opservacija i p varijabli. Primjenjujući gore navedena svojstva vektora $\mathbf{g}_i$ i $\mathbf{h}_j$ za te aproksimacije $\mathbf{g}_i^*$ i $\mathbf{h}_j^*$, možemo zaključiti da $\mathbf{h}_j^{*\mathbf{T}}\mathbf{h}_k^*$ daje aproksimaciju za s_{j_k}, odnosno kovarijance j-te i k-te varijable te će graf od $\mathbf{h}_j$ dati dvodimenzionalni prikaz (aproksimaciju) elemenata kovarijacijske matrice $\mathbf{S}$. Nadalje, Mahalanobisova udaljenost između $\mathbf{x}_h$ i $\mathbf{x}_i$ može se aproksimirati (i lako vizualizirati) Euklidskom udaljenosti između $\mathbf{g}_h^*$ i $\mathbf{g}_i^*$.

U gornjem slučaju, promatrali smo svojstva zasebno za $\mathbf{g}_i{}^*$ i $\mathbf{h}_j{}^*$, no postoji svojstvo koje vrijedi za svaki odabir vrijednosti α, te koje interpretira $\mathbf{g}_i{}^*$ i $\mathbf{h}_j{}^*$ zajedno. Iz relacije $x_{ij} = \mathbf{g}_i{}^{\mathrm{T}}\mathbf{h}_j$ slijedi da je x_{ij} projekcija od $\mathbf{g}_i$ na $\mathbf{h}_j$. Kako je x_{ij} vrijednost za i-tu opservaciju j-te varijable izmjerene oko svoje srednje vrijednosti, x_{ij} će biti blizu 0 za one opservacije koje su blizu srednjoj vrijednosti j-te varijable. To će se postići ako su $\mathbf{g}_i$ i $\mathbf{h}_j$ gotovo ortogonalni. Na isti način, za opservacije za koje je x_{ij} daleko od 0, $\mathbf{g}_i$ i $\mathbf{h}_j$ bit će približno istog smjera. Dakle, relativni položaji točaka definiranih s $\mathbf{g}_i$ i $\mathbf{h}_j$ (ili položaji njihovih aproksimacija $\mathbf{g}_i{}^*$ i $\mathbf{h}_j{}^*$ u dvodimenzionalnom prostoru), dat će informaciju o tome koje opservacije poprimaju velike, prosječne ili male vrijednosti za svaku varijablu.

U slučaju kada je $\alpha = 1$, svojstva navedena ranije za $\mathbf{g}_i$ i $\mathbf{h}_j$ neće vrijediti. Sada je

$$\mathbf{G} = \mathbf{L}\boldsymbol{\Delta}, \quad \mathbf{H}^{\mathrm{T}} = \mathbf{M}^{\mathrm{T}},$$

te $\|\mathbf{g}_\alpha - \mathbf{g}_\beta\|$ nije jednako Mahalanonbisovoj udaljenosti između $\mathbf{x}_h$ i $\mathbf{x}_i$, već je jednako

Euklidskoj udaljenosti. To slijedi jer

$$\begin{aligned}\|\mathbf{x}_\alpha - \mathbf{x}_\beta\| &= (\mathbf{x}_h - \mathbf{x}_i)^{\mathrm{T}}(\mathbf{x}_h - \mathbf{x}_i) \\ &= (\mathbf{g}_h - \mathbf{g}_i)^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{H}(\mathbf{g}_h - \mathbf{g}_i) \\ &= (\mathbf{g}_h - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}(\mathbf{g}_h - \mathbf{g}_i) \\ &= (\mathbf{g}_h - \mathbf{g}_i)^{\mathrm{T}}(\mathbf{g}_h - \mathbf{g}_i)\end{aligned}$$

Stoga, ako preferiramo graf na kojem će udaljenost između $\mathbf{g}_h^*$ i $\mathbf{g}_i^*$ biti dobra aproksimacija za Euklidsku (a ne Mahalanobisovu) udaljenost između $\mathbf{x}_h$ i $\mathbf{x}_i$, izabrat ćemo biplot za kojeg je $\alpha = 1$.
Navest ćemo još jedno zanimljivo svojstvo biplota kada je $\alpha = 1$. Naime, položaj vrijednosti od $\mathbf{g}_i^*$, $i = 1, 2, \ldots, n$ na grafu za $s = 2$, identičan je onima kod prikaza podataka s obzirom na prve dvije glavne komponente, kako je opisano u prethodnom potpoglavlju. Stupci matrice $\mathbf{M}$ iz faktorizacije 3.2 su svojstveni vektori matrice $\mathbf{B}^{\mathrm{T}}\mathbf{B}$. Umjesto $\mathbf{B}$ možemo uvrstiti matricu podataka $\mathbf{X}$ pa, kako su stupci matrice $\mathbf{A}$ iz 2.3 također svojstveni vektori od $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\mathbf{M}$ i $\mathbf{A}$ su ekvivalentne. Faktorizaciju 3.2 možemo zapisati kao

$$\mathbf{XA} = \mathbf{L\Delta}, \tag{3.18}$$

uz $\mathbf{B} = \mathbf{X}$ i $\mathbf{M} = \mathbf{A}$. Iz 3.18 i 2.3 zato slijedi:

$$x_{ij} = \sum_{k=1}^{r} z_{ik} a_{jk}, \tag{3.19}$$

gdje je $z_{ik} = l_{ik}{\delta_k}^{1/2}$ vrijednost k-te glavne komponente za i-tu opservaciju. No $\alpha = 1$ implicira da je $\mathbf{G} = \mathbf{L\Delta}$ pa je k-ti element od $\mathbf{g}_i$ jednak $l_{ik}{\delta_k}^{1/2} = z_{ik}$. Vektor ${\mathbf{g}_i}^*$ se sastoji od prva dva elementa vektora $\mathbf{g}_i$, a to su vrijednosti prve dvije glavne komponente i-te opservacije.
Što se tiče interpretacije za $\mathbf{h}_j$, koordinate od $\mathbf{h}_j$ bit će koeficijenti j-te varijable za prve dvije glavne komponente. Stoga, za $\alpha = 1$, prednost je ta što ${\mathbf{g}_i}^*$ i ${\mathbf{h}_j}^*$ možemo interpretirati skupa jer je x_{ij} projekcija od $\mathbf{g}_i$ na $\mathbf{h}_j$. Ipak, taj biplot ne predstavlja ništa novo jer su ${\mathbf{g}_i}^*$ vrijednosti glavnih komponenti, a ${\mathbf{h}_j}^*$ koeficijenti glavnih komponenti.

Poglavlje 4

Odabir glavnih komponenti

Kako je cilj analize glavnih komponenti u praksi odrediti s glavnih komponenti ($s \ll p$) koje zamjenjuju p varijabli vektora $\mathbf{x}$, ovo poglavlje dat će neke metode pomoću kojih možemo odrediti broj s, a da pritom ne dođe do većeg gubitka informacija početnog skupa $\mathbf{x}$, odnosno da zadržimo većinu varijacije od $\mathbf{x}$.

4.1 Kumulativni postotak varijance

U ovoj metodi, kao kriterij za odabir broja s, potrebno je odrediti kumulativni postotak ukupne varijacije (primjerice 80% ili 90%), tako da zbroj pojedinih postotaka ukupne varijacije za odabrane glavne komponente bude najmanje jednak tom kumulativnom postotku. Tada će traženi broj glavnih komponenti s biti najmanji broj glavnih komponenti takav da njihovi postotci ukupne varijacije u zbroju prelaze kumulativni postotak. Definirajmo što je "postotak varijacije sadržan u prvih s glavnih komponenti". Naime, glavne komponente odabrane su tako da imaju najveću moguću varijancu, u oznaci l_k. Nadalje, $\sum_{k=1}^{p} l_k = \sum_{k=1}^{p} s_{jj}$, odnosno zbroj varijanci glavnih komponenti jednak je zbroju varijanci elemenata od $\mathbf{x}$. Stoga se "postotak varijacije sadržan u prvih s glavnih komponenti" definira kao:

$$t_s = 100 \sum_{k=1}^{s} l_k \Big/ \sum_{k=1}^{p} s_{jj} = 100 \sum_{k=1}^{s} l_k \Big/ \sum_{k=1}^{p} l_k,$$

što se svodi na

$$t_s = \frac{100}{p} \sum_{k=1}^{s} l_k$$

za korelacijsku matricu. Odabirom granične vrijednosti t^* otprilike između 70% i 90% te uzimanjem prvih s glavnih komponenti (pri čemu je s najmanji cijeli broj za koji je $t_s > t^*$),

utvrđeno je pravilo po kojem u praksi dobivamo prvih s glavnih komponenti koje sadrže većinu informacija iz $\mathbf{x}$. Najbolja vrijednost za t^* općenito će biti manja kako raste p ili kako raste broj opservacija n. Iako je razumna granična vrijednost t^* između 70% i 90%, ponekad ona može biti viša ili niža ovisno o skupu podataka. Na primjer, bit će prikladno uzeti granicu iznad 90% ako su jedna ili dvije glavne komponente vrlo dominantne te sadrže velik dio varijacije od $\mathbf{x}$. Ostale glavne komponente bi nam mogle biti od interesa te kako bismo ih ubrojali, potrebno je uzeti granicu veću od 90%. Obrnuto, kada je p jako velik, odabir vrijednosti od s sukladno graničnoj vrijednosti 70% može dati vrlo velik s koji nije praktičan za daljnju analizu pa je potrebno korigirati granicu na niže.

Ovo pravilo zapravo je ekvivalentno promatranju spektralne dekompozicije kovarijacijske (ili korelacijske) matrice $\mathbf{S}$ (vidi 1.2.3) ili SVD dekompozicije matrice podataka $\mathbf{X}$. U oba slučaja potrebno je odlučiti koliko varijabli (tj. pribrojnika dekompozicije) uzeti u obzir kako bismo dobili dobru aproksimaciju za matricu $\mathbf{S}$ ili $\mathbf{X}$. To je povezano s odabirom t_s, jer je odgovarajuća mjera za grešku aproksimacije u slučaju odabira prvih s pribrojnika u obje dekompozicije jednaka $\sum_{k=s+1}^{p} l_k$. Za SVD dekompoziciju to slijedi iz 3.11 za $\mathbf{B} = \mathbf{X}$, a za spektralnu dekompoziciju slijedi iz razmatranja nakon Propozicije 2.4.

4.2 Vrijednost varijanci glavnih komponenti - Kaiserovo pravilo

Prethodno pravilo vrijedi u slučaju kada se koristi kovarijacijska ili korelacijska matrica pri računanju glavnih komponenti, no ovo pravilo konstruirano je posebno za korištenje s korelacijskom matricom. Naime, ukoliko su svi elementi od $\mathbf{x}$ nezavisni, onda su glavne komponente iste kao originalne varijable te sve imaju varijance jednake 1 u slučaju korelacijske matrice. Stoga svaka glavna komponenta s varijancom manjom od 1 sadrži manje informacija od neke originalne varijable pa ju nećemo zadržati. Ponekad se pravilo naziva *Kaiserovo pravilo* - zadržavamo samo one glavne komponente čije su varijance l_k veće od 1.

Diskutabilno je kako granična vrijednost $l_k = 1$ zadržava premalo glavnih komponenti u nekim situacijama. Primjerice, uzmimo varijablu koja je u populaciji pretežno nezavisna od drugih varijabli. U uzorku, takva varijabla imat će mali koeficijent u $(p-1)$ glavnih komponenti, no dominirat će u jednoj glavnoj komponenti čija će varijanca l_k biti blizu 1 koristeći korelacijsku matricu. Kako varijabla daje informaciju nezavisnu od drugih varijabli, nije pametno odbaciti ju. Međutim, ako koristimo Kaiserovo pravilo te zbog uzorkovanja bude $l_k < 1$, tu varijablu ćemo odbaciti. Zato je dobro, zbog varijacija u uzorkovanju, odrediti graničnu vrijednost $l^* < 1$ te je na osnovi simulacijskih studija preporučeno da ta vrijednost bude 0.7 [4].

Iako je konstruirano prvotno za korelacijsku matricu, ovo pravilo može se prilagoditi i

za kovarijacijsku matricu tako da graničnu vrijednost l^* računamo kao srednju vrijednost ($\bar{l}$) svojstvenih vrijednosti ili da također uzmemo nešto nižu granicu, primjerice $l^* = 0.7\bar{l}$.

Alternativni način interpretiranja ovog pravila je tzv. *broken stick* model. Ako podijelimo štap jedinične duljine na p nasumičnih segmenata, može se pokazati da će očekivana duljina k-tog najduljeg segmenta biti

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

Kada to primijenimo na naš slučaj – način na koji možemo odlučiti je li udio varijance sadržan u k-toj glavnoj komponenti dovoljan da se ona zadrži je usporedbom tog udjela s $l_k{}^*$. One glavne komponente za koje postotak od ukupne varijance prelazi $l_k{}^*$ ćemo zadržati, a ostale odbacujemo.

Poglavlje 5

Primjena - optimizacija dioničkog portfelja

U ovom poglavlju obrađujemo primjer koji prikazuje primjenu analize glavnih komponenti u diversifikaciji portfelja. S web stranice *Yahoo Finance* preuzeti su povijesni podaci za razdoblje od 1.8.2021. do 1.8.2023. za idućih 12 dionica: *Apple, American Express Company, Cisco System, General Electric, Goldman Sachs, IBM, Intel, JP Morgan and Chase, 3M, Microsoft, The Travelers Companies* i *Verizon Communications*. Za provedbu metode koristit ćemo korelacijsku matricu jer je u našem primjeru cilj proučavanje smjera rizika. Naime, kovarijacijska matrica sadrži informacije o riziku imovine te smjeru tog rizika, dok korelacijska matrica sadrži isključivo informacije o smjeru rizika. To nam je korisno u optimizaciji portfelja jer će korelacijska matrica biti sasvim nezavisna, s obzirom da se neće promijeniti djelovanjem upravitelja portfelja, koji ima mogućnost skaliranja rizika imovine uvođenjem gotovine ili poluge. Pritom će korelacija između imovina biti konstantna.
Iz dnevnih cijena gore navedenih dionica dobiveni su dnevni povrati te smo iz korelacijske matrice povrata dobili svojstvene vektore i pripadajuće svojstvene vrijednosti. U Tablici 5.1 prikazane su dobivene vrijednosti za svojstvene vrijednosti i kumulativne varijance svih glavnih komponenti, a u Tablici 5.2 prikazani su dodatno i svojstveni vektori za prve 3 glavne komponente.

Kako smo ranije već vidjeli, korelacijska matrica može se spektralnom dekompozicijom rastaviti na više pribrojnika koji predstavljaju glavne komponente te stoga korelacijsku matricu možemo podijeliti na tri dijela koja opisuju tri vrste kretanja cijene dionica [10]:

1. Prva glavna komponenta s najvećom svojstvenom vrijednosti predstavlja tržišni učinak koji utječe na sve dionice.

2. Nekoliko glavnih komponenti koje slijede tržišnu komponentu predstavljaju sinkro-

nizirane fluktuacije koje se događaju samo određenoj grupi dionica.

3. Preostale glavne komponente ukazuju na slučajnost kretanja cijene dionica.

PC	**Svojstvena vrijednost**	**Kumulativna varijanca (%)**
1	5.713	47.5
2	1.238	57.8
3	0.942	65.6
4	0.710	71.5
5	0.638	76.9
6	0.583	81.7
7	0.469	85.6
8	0.449	89.3
9	0.446	93.0
10	0.366	96.0
11	0.248	98.2
12	0.222	100

Tablica 5.1: Svojstvene vrijednosti i kumulativne varijance svih glavnih komponenti

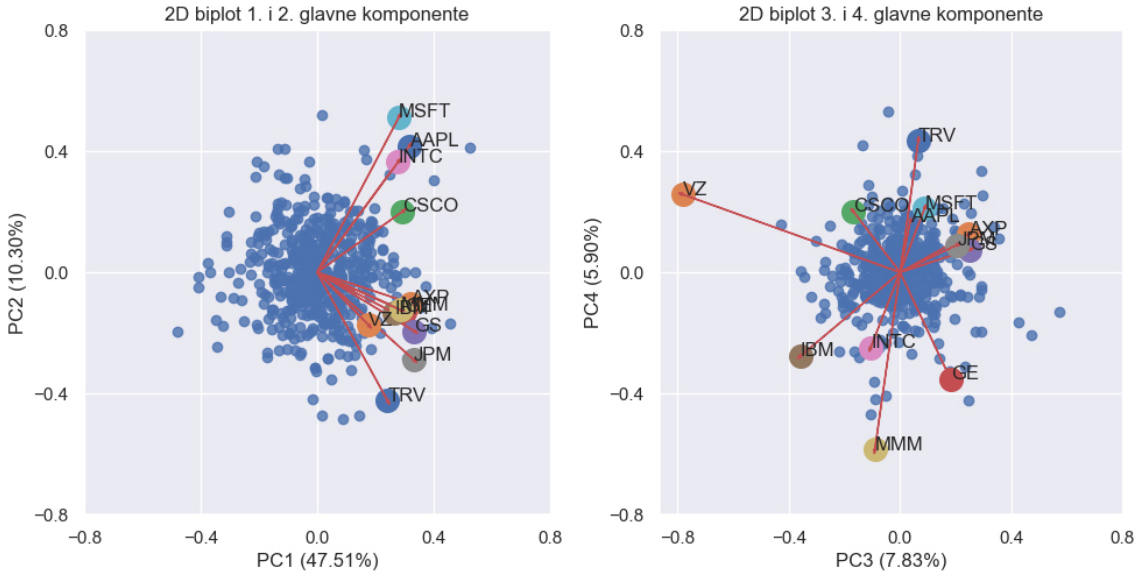
Dakle, prva glavna komponenta uobičajeno se interpretira kao tržišna komponenta, odnosno kao komponenta koja opisuje kretanje cijelog tržišta u kojoj bi sve dionice trebale imati otprilike jednaku težinu. Iz tog razloga očekujemo da će koeficijenti prve komponente imati isti predznak, što je kod nas slučaj. Već kod druge komponente pojavljuju se i pozitivni i negativni koeficijenti te je razlika u težinama dionica u drugoj glavnoj komponenti vidljivija. Osim iz vrijednosti u tablici, to najlakše možemo vidjeti na grafu 5.1 koji prikazuje biplot prve i druge glavne komponente, tako da očitamo projiciranu vrijednost svojstvenog vektora na os prve i druge glavne komponente te tako dobijemo težinu dionice u toj komponenti. Štoviše, međusobni položaj svojstvenih vektora na biplotu daje nam informaciju o koreliranosti pojedinih dionica. Primjerice, ako dva vektora zatvaraju mali kut, dionice su pozitivno korelirane, ako su pod kutom od približno 90° stupnjeva, dionice

nisu korelirane, te za kut od 180° bit će negativno korelirane. Vidimo kako u drugoj glavnoj komponenti koeficijenti dionica MSFT (Microsoft) i TRV (The Travelers Companies) imaju najveće apsolutne vrijednosti. Iz te komponente možemo na neki način naslutiti grupiranje dionica po sektoru. Naime, sve dionice financijskog sektora (AXP, GE, GS, JPM, MMM, TRV) imaju negativne koeficijente, a većina tehnoloških ima pozitivne pa bi se komponenta mogla donekle protumačiti kao mjera pripadnosti pojedinom sektoru dionica. Treća i četvrta komponenta teže su za interpretirati, no iz biplota možemo vidjeti kako velik utjecaj u trećoj komponenti ima dionica VZ (Verizon Communications), a u četvrtoj MMM i TRV.

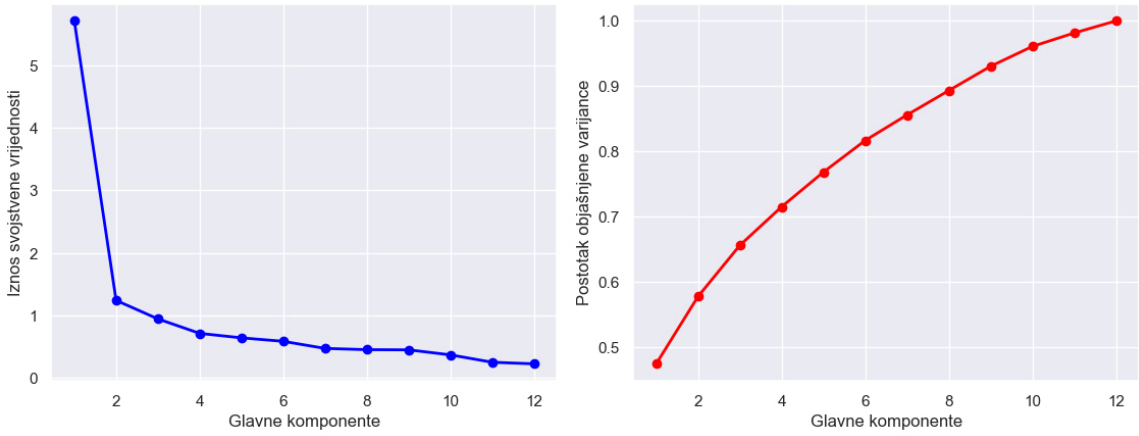
		1. komponenta	**2. komponenta**	**3. komponenta**
AAPL	koeficijenti	0.317	0.417	0.036
AXP		0.321	-0.102	0.247
CSCO		0.293	0.200	0.167
GE		0.302	-0.128	0.182
GS		0.333	-0.194	0.250
IBM		0.265	-0.135	-0.358
INTC		0.278	0.364	-0.106
JPM		0.331	-0.288	0.204
MMM		0.287	-0.124	-0.090
MSFT		0.281	0.513	0.088
TRV		0.242	-0.423	0.065
VZ		0.176	-0.174	-0.783
Svojstvena vrijednost:		5.713	1.238	0.942
Kumulativni postotak od ukupne varijance:		0.475	0.578	0.656

Tablica 5.2: Svojstvene vrijednosti i svojstveni vektori za prve tri glavne komponente - dobiveni analizom glavnih komponenti.

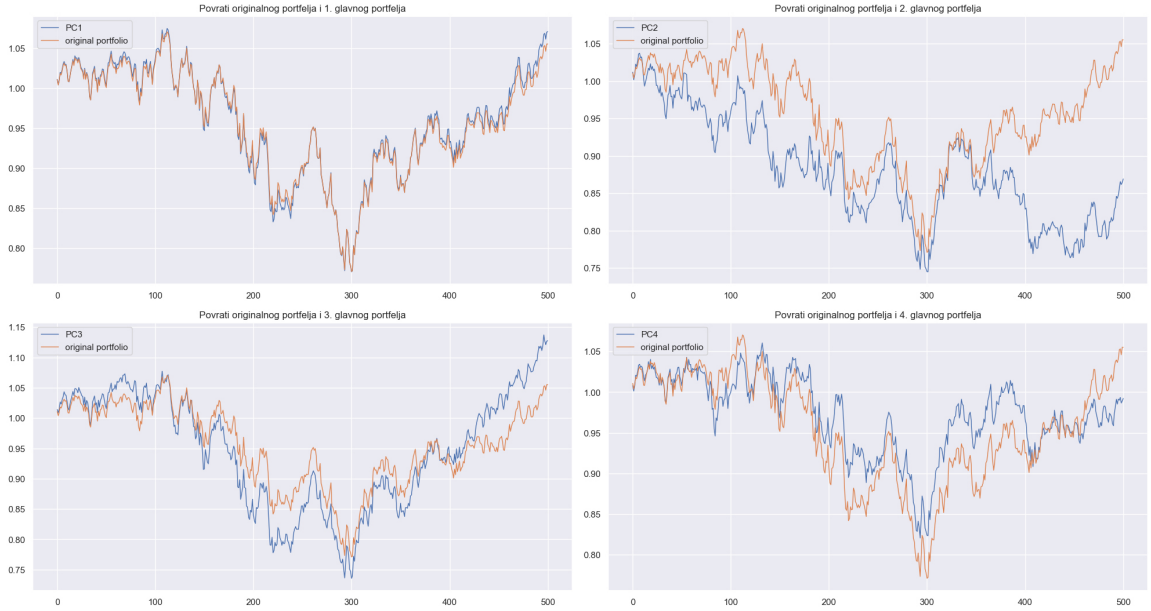
Sada ćemo prema kriterijima opisanima u Poglavlju 4 odrediti broj glavnih komponenti koje ćemo uzeti u obzir. Za kriterij o kumulativnom postotku varijance, možemo se poslužiti prikazom 5.2 (desno). Ako odredimo da nam je kumulativni postotak varijance koji želimo postići jednak 75%, bit će dovoljno uzeti u obzir prvih 5 glavnih komponenti. Nadalje, koristeći Kaiserovo pravilo te promatrajući prikaz 5.2 (lijevo), zaključujemo da je dovoljno zadržati prve četiri glavne komponente jer one imaju svojstvene vrijednosti iznad preporučene granice 0.7. Stoga možemo zaključiti da će nam prve četiri komponente dati dovoljno informacija o podacima te će kumulativna varijanca i dalje biti iznad 70%.



Slika 5.1: Biplot 1. i 2. glavne komponente (lijevo) te 3. i 4. glavne komponente (desno).



Slika 5.2: Graf svojstvenih vrijednosti (lijevo) i graf kumulativnih postotaka od ukupne varijance za glavne komponente (desno).



Slika 5.3: Povrati originalnog portfelja $(1/N)$ skupa s povratima prva četiri glavna portfelja.

No, koja je zapravo interpretacija metode analize glavnih komponenti na korelacijsku matricu povrata te kako može poslužiti kao smjernica u diversifikaciji portfelja? Svojstveni vektori glavnih komponenti (koeficijenti u tablici 5.2) predstavljaju težine za svaku dionicu *glavnog portfelja*. Naime, primjenom metode analize glavnih komponenti, svaka glavna komponenta definira jedan glavni portfelj u kojem pozitivne težine svojstvenog vektora ukazuju na duge pozicije, a negativne težine na kratke pozicije. Nadalje, pripadajuća svojstvena vrijednost jednaka je udjelu standardizirane varijance od originalnog portfelja. Originalni portfelj konstruiran je koristeći strategiju $1/N$, u kojem svaka dionica ima jednaku težinu. Te glavne komponente/glavni portfelji interpretiraju se kao nekorelirani izvori rizika u originalnom skupu podataka (tržištu dionica). Također, iz našeg primjera vidljivo je kako svojstvene vrijednosti, odnosno postoci objašnjene varijance glavnih komponenti brzo opadaju s rastom broja komponente. Stoga, iako je tehnički moguće konstruirati portfelj koristeći sve glavne komponente izlažući se svim izvorima rizika, pomalo je nerazumno raspodijeliti sredstva u više glavne komponente koje nisu izvor nekog većeg rizika. Portfelji konstruirani na ovaj način tretiraju se kao individualne investicije koje nisu korelirane.

Koraci za konstrukciju glavnih portfelja su idući[10]:

1. Primijeniti metodu analize glavnih komponenti te dobiti koeficijente za glavne komponente.

2. Težine ulaganja u svaku dionicu računaju se kao:

   - koeficijent uz dionicu podijeljen zbrojem svih pozitivnih koeficijenata unutar te glavne komponente (ako je koeficijent pozitivan)
   - koeficijent uz dionicu podijeljen apsolutnom vrijednošću zbroja svih negativnih koeficijenata unutar te glavne komponente (ako je koeficijent negativan)

   Ovo će dati skup težina u kojem će duge pozicije u zbroju iznositi 1, a kratke pozicije će u zbroju iznositi -1. Udio kratkih pozicija je omjer zbroja svih negativnih koeficijenata i zbroja svih pozitivnih koeficijenata. Pretpostavlja se da se sredstva dobivena iz kratkih pozicija ulažu po prosječnoj bezrizičnoj stopi tijekom posljednjih 14 godina te je godišnja bezrizična stopa pretvorena u dnevnu jer koristimo dnevne povrate.

3. Povrat glavnog portfelja jednak je: (*zbroj ponderiranih povrata za svaku dionicu*) + (*umnožak bezrizične stope i udjela kratkih pozicija*).

U našem primjeru, originalni portfelj konstruiran je koristeći strategiju $1/N$. Povrati tog originalnog portfelja prikazani su skupa s prva četiri glavna portfelja na grafovima 5.3. Vidimo kako prvi glavni portfelj, koji ima najveću varijancu, najbolje opisuje kretanje originalnog portfelja jer su u prvoj glavnoj komponenti doprinosi dionica približno jednaki što oponaša originalni portfelj jednakih težina. To bi značilo da je ulaganje u originalni portfelj najlakši način kako dobiti portfelj koji ima tržišni rizik, a ne idiosinkratski rizik (inherentan rizik koji pogađa pojedinu dionicu ili tvrtku, također se naziva nesustavni rizik). Već se drugi glavni portfelj bitno razlikuje od originalnog donoseći vidljivo niži povrat te odražavajući veće promjene s obzirom na kretnje povrata u originalnom portfelju. Sve prikazane komponente dugoročno prate kretanje originalnog portfelja s povremenim odstupanjima te uspješno detektiraju veće promjene u povratima, poput pada povrata oko 300-tog dana.
Tablica 5.3 prikazuje Sharpeov omjer koji se koristi kao mjera performanse portfelja, a računa se kao godišnji povrat podijeljen s godišnjom volatilnosti. On predstavlja dodatni povrat koji investitor dobije po jedinici povećanja rizika. Veći Sharpeov omjer znači da je portfelj bolje prilagođen riziku. Gledajući Sharpeov omjer, investitor može odlučiti više novčanih sredstava dodijeliti za treći i četvrti glavni portfelj.

Zaključno, na ovaj način izloženost pojedinačnom riziku postaje izvediva. Investitori mogu odlučiti držati bilo koji glavni portfelj kako bi se izložili jednom izvoru rizika koji nije u korelaciji s drugim rizicima na tržištu. Odluka o tome hoće li se neka imovina

Glavni portfelj	**1**	**2**	**3**	**4**
Povrat	3.48%	-1.75%	14.00%	4.42%
Volatilnost	19.64%	20.03%	21.28%	17.58%
Sharpeov omjer	0.18	-0.09	0.66	0.25

Tablica 5.3: Sharepov omjer za prva četiri glavna portfelja.

uključiti donosi se isključivo na temelju njezine varijance i povrata neovisno o njenom zajedničkom kretanju s ostalima u portfelju.

Dodatak

R kod – implementacija primjera iz Poglavlja 3

```
#install.packages("jsonlite", repos="https://cran.rstudio.com/")
library("jsonlite")

#install.packages("tidyverse")
library("scales")
library(tidyverse)
library(scatterplot3d)

json_file <-
'https://datahub.io/core/s-and-p-500-companies-financials/datapackage.json'
json_data <- fromJSON(paste(readLines(json_file), collapse=""))

# get list of all resources:
print(json_data$resources$name)

# print all tabular data(if exists any)
for(i in 1:length(json_data$resources$datahub$type)){
  if(json_data$resources$datahub$type[i]=='derived/csv'){
    path_to_file = json_data$resources$path[i]
    data <- read.csv(url(path_to_file))
    print(data)
  }
}
path_to_file=json_data$resources$path[3]
data <- read.csv(url(path_to_file))
View(data)

data=na.omit(data)
```

```r
pca=prcomp(data[, 4:13], scale = T, center=T)
summary(pca)

pov <- pca$sdevˆ2/sum(pca$sdevˆ2)
pov=label_percent()(pov)

eigenvectors=pca$rotation
eigenvalues=pca$sdevˆ2
sum(eigenvalues)
#scores
PC1=pca$x[, 1]
PC2=pca$x[, 2]

data=cbind(data, pca$x)
sector=data$Sector

#svi sektori
ggplot(data=data, aes(x=PC1, y=PC2, color=sector))+geom_point()+
coord_fixed(xlim=c(-10, 5), ylim=c(-5, 5))+
labs(x=paste0("PC1: ", pov[1]),
     y=paste0("PC2: ", pov[2]))

#istaknuti sektori
data %>%
ggplot(aes(x=PC1, y=PC2))+
geom_point(color="grey")+
geom_point(data=data %>% filter(Sector=="Utilities"),
aes(PC1, PC2, color=Sector))+
geom_point(data=data %>% filter(Sector=="Health Care"),
aes(PC1, PC2,  color=Sector))+
coord_fixed(xlim=c(-5, 2.7), ylim=c(-5, 5))+
labs(x=paste0("PC1: ", pov[1]),
     y=paste0("PC2: ", pov[2]))
```

Python kod – implementacija primjera iz Poglavlja 5

```python
import numpy as np
import pandas as pd
from sklearn.preprocessing import StandardScaler
```

```
from bioinfokit.visuz import cluster
import sklearn.decomposition
from sklearn.decomposition import PCA
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

df = pd.read_csv("dionice2022.csv", index_col = 0)
df.head()

povrati = pd.DataFrame(data=np.zeros(shape=(len(df.index), df.shape[1])),
                               columns=df.columns.values,
                               index=df.index)
povrati = df.pct_change().dropna()
povrati.head()

tickers = povrati.columns.values
povrati_std = StandardScaler().fit_transform(povrati)
povrati_std = pd.DataFrame(povrati_std, columns=df.columns)

pca_out = PCA().fit(povrati_std)
eigenvalues = pca_out.explained_variance_

PC_values = np.arange(pca_out.n_components_) + 1
plt.plot(PC_values, eigenvalues, 'o-', linewidth=2, color='blue')
plt.xlabel('Glavne komponente')
plt.ylabel('Iznos svojstvene vrijednosti')
plt.show()

# Udio varijance (from PC1 to PC6)
pca_out.explained_variance_ratio_
# Kumulativna varijanca (from PC1 to PC6)
cum=np.cumsum(pca_out.explained_variance_ratio_)

loadings = pca_out.components_
num_pc = pca_out.n_features_
pc_list = ["PC"+str(i) for i in list(range(1, num_pc+1))]
load_df = pd.DataFrame.from_dict(dict(zip(pc_list, loadings)))
load_df['variable'] = df.columns.values
```

```
load_df = load_df.set_index('variable')
load_df

PC_values = np.arange(pca_out.n_components_) + 1
plt.plot(PC_values, cum, 'o-', linewidth=2, color='red')
plt.xlabel('Glavne komponente')
plt.ylabel('Postotak objašnjene varijance')
plt.show()

# PC scores
pca_scores = PCA().fit_transform(povrati_std)

#biplot za 1. i 2. gk
pca_df = pd.DataFrame(
    data=pca_scores, columns=['PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'PC6',
                              'PC7', 'PC8', 'PC9', 'PC10', 'PC11', 'PC12'])

pca_df_scaled = pca_df.copy()
scaler_df = pca_df[['PC1', 'PC2']]
scaler = 1 / (scaler_df.max() - scaler_df.min())

for index in scaler.index:
    pca_df_scaled[index] *= scaler[index]

# 2D
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()

xs = loadings[0]
ys = loadings[1]

sns.lmplot(
    x='PC1',
    y='PC2',
    data=pca_df_scaled,
    fit_reg=False,
    )

```

```
for i, varnames in enumerate(tickers):
    plt.scatter(xs[i], ys[i], s=200)
    plt.arrow(
        0, 0, # coordinates of arrow base
        xs[i], # length of the arrow along x
        ys[i], # length of the arrow along y
        color='r',
        head_width=0.01
        )
    plt.text(xs[i], ys[i], varnames)


xticks = np.linspace(-0.8, 0.8, num=5)
yticks = np.linspace(-0.8, 0.8, num=5)
plt.xticks(xticks)
plt.yticks(yticks)
plt.xlabel('PC1 (%.2f%%)' %(100*pca_out.explained_variance_ratio_[0]))
plt.ylabel('PC2 (%.2f%%)' %(100*pca_out.explained_variance_ratio_[1]))

plt.title('2D biplot 1. i 2. glavne komponente')
plt.show()

#biplot za 3. i 4. gk
scaler_df = pca_df[['PC3', 'PC4']]
scaler = 1 / (scaler_df.max() - scaler_df.min())

for index in scaler.index:
    pca_df_scaled[index] *= scaler[index]

# 2D
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()

xs = loadings[2]
ys = loadings[3]

sns.lmplot(
    x='PC3',
```

```
    y='PC4',
    data=pca_df_scaled,
    fit_reg=False,
    )

for i, varnames in enumerate(tickers):
    plt.scatter(xs[i], ys[i], s=200)
    plt.arrow(
        0, 0, # coordinates of arrow base
        xs[i], # length of the arrow along x
        ys[i], # length of the arrow along y
        color='r',
        head_width=0.01
        )
    plt.text(xs[i], ys[i], varnames)


xticks = np.linspace(-0.8, 0.8, num=5)
yticks = np.linspace(-0.8, 0.8, num=5)
plt.xticks(xticks)
plt.yticks(yticks)
plt.xlabel('PC3 (%.2f%%)' %(100*pca_out.explained_variance_ratio_[2]))
plt.ylabel('PC4 (%.2f%%)' %(100*pca_out.explained_variance_ratio_[3]))

plt.title('2D biplot 3. i 4. glavne komponente')
plt.show()

# normaliziranje za prvu komponentu
pc_w = loadings[0, :] / sum(loadings[0, :])

#jednaki portfelj
pc_w_1n=np.repeat(1/12, 12)#1/n portfelj

initial_p = pd.DataFrame(data ={'weights': pc_w_1n.squeeze()*100},
                                              index = tickers)
eigen_p1 = pd.DataFrame(data ={'weights': pc_w.squeeze()*100},
                                             index = tickers)

eigen_p1_povrati = np.dot(povrati.loc[:, eigen_p1.index],
```

```
                                            eigen_p1 / 100)
eigen_p1_povrati = pd.Series(eigen_p1_povrati.squeeze(),
                             index=povrati_std.index)

initial_povrati = np.dot(povrati.loc[:, initial_p.index],
                                            initial_p / 100)
initial_povrati = pd.Series(initial_povrati.squeeze(),
                             index=povrati_std.index)

df_plot = pd.DataFrame({'PC1': eigen_p1_povrati,
                'original portfolio': initial_povrati},
                index=povrati_std.index)
np.cumprod(df_plot+1).plot(title='Povrati originalnog portfelja
                i 1. glavnog portfelja', figsize=(12,6), linewidth=1)

# 2.komponenta ima negativne vrijednosti

#za negativne tezine + risk free rate
l=loadings[1, :]
rat = sum(l[l < 0])/sum(l[l > 0])*0.0002
print(rat)

#omjeri samo za pozitivne tezine
l[l < 0] = 0
l[l > 0] = l[l > 0] / sum(l[l > 0])
print(sum(l))

eigen_p2 = pd.DataFrame(data ={'weights': pc_w.squeeze()*100},
            index = tickers)
eigen_p2.sort_values(by=['weights'], ascending=False, inplace=True)

eigen_p2_povrati = np.dot(povrati.loc[:, eigen_p2.index], l)
eigen_p2_povrati = pd.Series(eigen_p2_povrati.squeeze(),
index=povrati_std.index)

df_plot = pd.DataFrame({'PC2': eigen_p2_povrati + rat,
'original portfolio': initial_povrati}, index=povrati_std.index)
np.cumprod(df_plot+1).plot(title='Povrati originalnog portfelja
                i 2. glavnog portfelja', figsize=(12,6), linewidth=1)
```

```
199
200 # 3.komponenta ima negativne vrijednosti
201
202 #za negativne tezine + risk free rate
203 l=loadings[2, :]
204 rat = sum(l[l < 0])/sum(l[l > 0])*0.0002
205 print(rat)
206
207 #omjeri samo za pozitivne tezine
208 l[l < 0] = 0
209 l[l > 0] = l[l > 0] / sum(l[l > 0])
210 print(sum(l))
211
212 eigen_p3 = pd.DataFrame(data ={'weights': pc_w.squeeze()*100},
213             index = tickers)
214 eigen_p3.sort_values(by=['weights'], ascending=False, inplace=True)
215
216 eigen_p3_povrati = np.dot(povrati.loc[:, eigen_p3.index], l)
217 eigen_p3_povrati = pd.Series(eigen_p3_povrati.squeeze(),
218             index=povrati_std.index)
219
220 df_plot = pd.DataFrame({'PC3': eigen_p3_povrati + rat,
221 'original portfolio': initial_povrati}, index=povrati_std.index)
222 np.cumprod(df_plot + 1).plot(title='Povrati originalnog portfelja
223                 i 3. glavnog portfelja', figsize=(12,6), linewidth=1)
224
225 #za negativne tezine + risk free rate
226 l=loadings[3, :]
227 rat = sum(l[l < 0])/sum(l[l > 0])*0.0002
228 print(rat)
229
230 #omjeri samo za pozitivne tezine
231 l[l < 0] = 0
232 l[l > 0] = l[l > 0] / sum(l[l > 0])
233 print(sum(l))
234
235 eigen_p4 = pd.DataFrame(data ={'weights': pc_w.squeeze()*100},
236 index = tickers)
237 eigen_p4.sort_values(by=['weights'], ascending=False, inplace=True)
```

```
238
239 eigen_p4_povrati = np.dot(povrati.loc[:, eigen_p4.index], l)
240 eigen_p4_povrati = pd.Series(eigen_p4_povrati.squeeze(),
241 index=povrati_std.index)
242
243 df_plot = pd.DataFrame({'PC4': eigen_p4_povrati + rat,
244 'original portfolio': initial_povrati}, index=povrati_std.index)
245 np.cumprod(df_plot + 1).plot(title='Povrati originalnog portfelja
246                 i 4. glavnog portfelja', figsize=(12,6), linewidth=1)
247
248 #za negativne tezine + risk free rate
249 l=loadings[4, :]
250 rat = sum(l[l < 0])/sum(l[l > 0])*0.0002
251 print(rat)
252
253 #omjeri samo za pozitivne tezine
254 l[l < 0] = 0
255 l[l > 0] = l[l > 0] / sum(l[l > 0])
256 print(sum(l))
257
258 def sharpe_ratio(ts_povrati, periods_per_year=252):
259
260     annualized_return = 0.
261     annualized_vol = 0.
262     annualized_sharpe = 0.
263
264     n_years = ts_povrati.shape[0] / periods_per_year
265     annualized_return = np.power(np.prod(1 + ts_povrati),(1 / n_years)) - 1
266     annualized_vol = ts_povrati.std() * np.sqrt(periods_per_year)
267     annualized_sharpe = annualized_return / annualized_vol
268
269
270     return annualized_return, annualized_vol, annualized_sharpe
271
272 ret, vol, sharpe = sharpe_ratio(eigen_p1_povrati)
273 print('First eigen-portfolio:\nReturn = %.2f%%
274 \nVolatility = %.2f%%\nSharpe = %.2f' % (ret*100, vol*100, sharpe))
275
276 ret, vol, sharpe = sharpe_ratio(eigen_p2_povrati)
```

```
277 print('Second eigen-portfolio:\nReturn = %.2f%%
278 \nVolatility = %.2f%%\nSharpe = %.2f' % (ret*100, vol*100, sharpe))
279
280 ret, vol, sharpe = sharpe_ratio(eigen_p3_povrati)
281 print('Third eigen-portfolio:\nReturn = %.2f%%
282 \nVolatility = %.2f%%\nSharpe = %.2f' % (ret*100, vol*100, sharpe))
283
284 ret, vol, sharpe = sharpe_ratio(eigen_p4_povrati)
285 print('Fourth eigen-portfolio:\nReturn = %.2f%%
286 \nVolatility = %.2f%%\nSharpe = %.2f' % (ret*100, vol*100, sharpe))
287
288 ret, vol, sharpe = sharpe_ratio(initial_povrati)
289 print('Original portfolio:\nReturn = %.2f%%
290 \nVolatility = %.2f%%\nSharpe = %.2f' % (ret*100, vol*100, sharpe))
```

Bibliografija

[1] *Biplots in Python*, https://www.jcchouinard.com/python-pca-biplots-machine-learning/, [accessed: 8-August-2023].

[2] *Portfolio Construction Using PCA*, https://github.com/Doj-i/NYU_Machine_Learning_in_Finance/blob/master/Portfolio%20Construction%20using%20PCA/Eigen-portfolio%20construction%20using%20Principal%20Component%20Analysis%20(PCA)_ML2_ex3.ipynb, [accessed: 8-August-2023].

[3] *S&P 500 Financial Information*, https://datahub.io/core/s-and-p-500-companies-financials#r, [accessed: 8-August-2023].

[4] Ian Jolliffe, *Principal component analysis*, Springer Verlag, New York, 2002.

[5] Giorgia Pasini, *Principal component analysis for stock portfolio management*, International Journal of Pure and Applied Mathematics **115** (2017), br. 1, 153–167.

[6] John O. Rawlings, Sastry G. Pantula i David A. Dickey, *Applied regression analysis: a research tool*, Springer, 1998.

[7] G. A. F. Seber, *Multivariate Observations*, Wiley, 1984.

[8] Wikipedia contributors, *RV coefficient — Wikipedia, The Free Encyclopedia*, 2021, https://en.wikipedia.org/w/index.phptitle=RV_coefficient&oldid=1059405688, [Online; accessed 6-July-2023].

[9] ———, *Low-rank approximation — Wikipedia, The Free Encyclopedia*, 2023, https://en.wikipedia.org/w/index.php?title=Low-rank_approximation&oldid=1163165586, [Online; accessed 28-July-2023].

[10] Libin Yang, *An application of principal component analysis to stock portfolio management*, (2015).

Sažetak

Cilj ovog rada bio je obraditi metodu analize glavnih komponenti te prikazati njenu primjenu na tržište vrijednosnih papira. Analiza glavnih komponenti je metoda redukcije dimenzionalnosti koja pojednostavljuje interpretaciju višedimenzionalnih podataka uz očuvanje maksimalne količine informacija. Najprije smo prikazali izvod i definiciju glavnih komponenti dobivenih na temelju populacijske matrice kovarijanci te smo predstavili njihova algebarska i geometrijska svojstva, a ona bitnija i dokazali. Također, razmotrili smo situaciju u kojoj je pogodnije koristiti korelacijsku matricu za primjenu metode umjesto kovarijacijske matrice. Nadalje smo definirali glavne komponente uzorka te su ranije prikazana svojstva predstavljena i u kontekstu uzorka. Obrađena su dva grafička prikaza glavnih komponenti te je primjerom pokazano kako metoda može klasterirati podatke u grupe. Opisali smo kriterije pomoću kojih možemo u praksi odrediti potreban broj glavnih komponenti koje zamjenjuju početne varijable, no da pritom ne dođe do većeg gubitka informacija. U posljednjem poglavlju metoda je primijenjena na povijesne podatke o povratima za 12 dionica te je opisano kako investitor, razmatrajući neovisne rizike, može diversificirati portfelj koristeći metodu.

Summary

The goal of this thesis was to present the principal component analysis method and show its usage in the capital markets. PCA is a dimensionality reduction method that simplifies the interpretation of high-dimensional data while preserving the largest amount of information. At first, we describe the method's mathematical background and its algebraic and geometrical properties. We consider when to use the correlation matrix instead of covariance matrix. Furthermore, we define the sample principal components and show their properties. The method's ability to cluster data into groups is shown using an example. We describe the criteria using which we can set the necessary number of principal components to replace the initial variables in practice. In the last chapter, we apply the method on historical price data for 12 stocks and show how we can better diversify the portfolio by considering independent risks.

Životopis

Rođena sam 7. veljače 1999. godine u Zagrebu. Osnovnu školu pohađam u Sesvetama. 2013. godine upisujem Gimnaziju Sesvete koju završavam 2017. godine. U srednjoj školi razvijam interes za matematiku. Iste godine upisujem preddiplomski studij na matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu, smjer Matematika, te 2020. godine stječem titulu prvostupnika. Po završetku preddiplomskog studija nastavljam obrazovanje upisom diplomskog studija Financijska i poslovna matematika na istom fakultetu. U slobodno vrijeme bavim se pjevanjem.