Univerzitet u NovomSadu
Prirodno-matematički fakultet
Departman za matematiku i informatiku

Metode za smanjenje dimenzionalnosti podataka i njihova primena u prirodnim naukama

- master rad -

Mentor:
prof. dr Zagorka Lozanov-Crvenković

Kandidat:
Vladimir Rančić
181m/11

U Novom Sadu,
2013. godine

Sadržaj

Predgovor

Tema ovog master rada su statističke metode za smanjenje dimenzionalnosti podataka i to analiza glavnih komponenti, (stvarna) faktorska analiza i analiza parcijalnih najmanjih kvadrata. Ove metode spadaju u statističku oblast koja se naziva faktorska analiza. Kreator faktorske analize bio je engleski psiholog Charles Spearman. On je 1904. godine prvi postavio empiriskijsku teoriju inteligencije i pritom razvio novu statističku metodu faktorska analiza, do tada nepoznatu u svetu. U vrlo kratkom periodu, došlo je do nagle ekspanzije same metode, tako da se danas ne može utvrditi tačan broj metoda unutar faktorske analize, ali se zna da je primenljiva u skoro svim prirodnim i društvenim naukama. Glavni cilj faktorske analize jeste rešavanje problema multikolinearnosti, ali i smanjenje broja promenljivih, odnosno ekonomičnost.

Rad se sastoji iz šest glava, gde su prvih pet glava teorijske, a šesta predstavlja rezultat praktičnog rada nad određenim problemom u hemiji. Tekst je koncizno napisan, razumljiv za sve nivoe matematičkog znanja.

U prvoj glavi uvedena je definicija faktorske analize, osnovne karakteristike i problem koji se javlja sa podacima. Zatim je prikazan motivacioni primer faktorske analize u ekonomiji, kako bi čitalac bolje razumeo samo ideju analize. Na kraju je dat kratak istorijat ove statističke oblasti, od Spearman-a, preko Pearson-a i Hotelling-a, do današnjih dana.

U drugoj glavi se nalazi osnovna algebarska i statistička baza koja će biti korišćena u celom radu. U njoj je obrađena: matrica podataka, matrični proizvod veće i manje dimenzije, ortogonalne i ortonormirani vektori i matrice, deskriptivna statistika, rotacija sistema koordinatnih osa, karakteristični koreni i vektori i njihova geometrijska intepretacija, SVD i NIPALS dekompozicija i dr.

U trećoj glavi je obrađena metoda analize glavnih komponenti (principal component analysis). Kreatori ove metode su bili Pearson i Hotelling. Na početku je izveden osnovni (teorijski) populacioni model u okviru kog je definisana glavna komponenta (principal component), kovarijansna matrica i optimizacioni modeli za pronalaženje glavnih komponenti. Takođe, diskutuje se o ideji zamene kovarijansne matrice korelacionom matricom i posledice po tom pitanju. Zatim je dat uzorački model analize glavnih komponenti, analiziran broj komponenti i dat primer, radi lakšeg razumevanja analize.

U četvrtoj glavi je obrađena metoda (prava) faktorska analiza (true factor analysis), kreirana od Spearman. Nakon uvodne priče, dati su modeli faktorske analize: opšti model faktorske analize, model faktorske analize putem kovarijansne matrice i transformacioni model. Zatim je prikazan primer faktorske analize gde je detaljno, korak po korak, prikazan sam postupak analize, rotacija koordinata, grafički prikaz i dr. Na kraju je data metoda za izračunavanje matrice koeficijenata i primer.

U petoj glavi su obrađeni modeli višestruke regresije: regresija običnih najmanjih kvadrata, regresija glavnih komponenti, rubna regresija i kao glavna stavka ove oblasti – regresija parcijalnih najmanjih kvadrata (partial least-squares analysis). Švedski hemometričar Wold je 1970. godine dao prvu ideju regresije parcijalnih najmanjih kvadrata i od tada ona ima široku primenu u hemiji, a i drugim oblastima prirodnih nauka. Za sve

regresije će biti izveden model za izračunavanje regresionih koeficijenata, s tim što je za analizu parcijalnih najmanjih kvadrata dat algoritam, koji koristi NIPALS dekompoziciju.

Šesta glava predstavlja praktičnu primenu sve tri metode nad podacima iz hemije. Tačnije, nakon uvodnog priče o hemometriji, primenjene su analiza glavnih komponenti i faktorska analiza na skupu od 14 molekulskih deskriptora, merena nad 24 derivata hindatoina. Zatim, putem regresije analize glavnih komponenti, analizirana je kvantitativna veza između molekulskih deskriptora i vrednosti retencionog faktora R_M^0. Svi rezultati su dobijeni primenom softverskog paketa STATISTICA.

***

Najveću zahvanolst za odabir i pomoć oko ove teme dala mi je moj mentor prof. dr. Zagorka Lozanov-Crvenković. Bilo je čast sarađivati sa njom, jer kako me je zainteresovala za ovu temu, tako mi je i pomogla u samoj izradi, izboru literature, analiziranju podataka i davanju korsnih sugestija i primedbi. Hvala na ukazanom poverenju!

Takođe, veliku zahvalnost dugujem i prof. dr. Tatjani Đaković-Sekulić, sa Departmana za hemiju, biohemiju i zaštitu životne sredine, na pomoći oko izbora ideje za praktičnu primenu metoda u hemiji, razumevanju hemijske literature i omogućavanju korišćenja neophodnih hemijskih podataka.

Velku pomoć mi je pružila i prof. dr. Ivana Štajner-Papuga, koja je tekst detaljno analizirala, a zatim dala neophodne sugestije i primedbe kako bi ovaj tekst dobio sadašnji oblik.

I na kraju, posebnu zahvalnost dugujem svom bratu, majci i ocu kao i svim priajteljima i profesorima. Hvala im na razumevanju i bezuslovnoj podršci koju su mi pružali tokom celokupnog školovanja.

Novi Sad, jun 2013. godine *Vladimir Rančić*

Glava 1

Uvod u faktorsku analizu

1.1 Povezanost među pojavama

Međusobna povezanost pojava predstavlja jedan od temeljnih načela i postulata koji leži u osnovi svih naučnih aktivnosti. Prema tom načelu, pojave koje su predmet istraživanja na području neke nauke su međusobno povezane. Nauka ima zadatak da utvrdi, što je moguće tačnije i detaljnije, povezanost između pojava jer se time omogućava da se na osnovu jedne pojave predvide promene i događaji kod druge pojave. Paralelno s tim, istraživanjem međusobne povezanosti pojava i događaja postižu se i drugi ciljevi naučnog rada, npr. razumevanje pojave i događaja na nekom području.

Povezanost može mati najmanje dvojaki karakter. Ona može biti uzročno-posledična i korelaciona. Uzročno-posledična povezanost znači da je jedna pojava ili događaj uzrok nastajanja ili javljanja neke druge pojave ili događaja. Time je i druga pojava posledica prve. U ovakvim slučajevima ne očekuje se postojanje promene druge pojave ukoliko ne dođe do promene prve pojave, i obrnuto. Uzročno-posledična povezanost može biti jednostavna ili složena zavisno od toga da li se radi o povezanosti dve ili više pojava ili događaja. Korisnost poznavanaj ove povezanosti je i te kako bitna za istraživače jer što je to poznavanje veće time su mogućnosti predviđanja i kontrole događaja na području nekog istraživanja veća.

Druga vrsta međusobne povezanosti pojava ili događaja jeste korelaciona povezanost. To se dešava kada dođe do paralelnih promena u jednoj i drugoj pojavi, a da nisu međusobno uzrokovane (jedna pojava nije uzrok druge pojave). Ovaka vid povezanosti se javlja u širokom spektru nauka. Poznavanjem ili merenjem neke promene, može se predvideti promene koje će se dogoditi u nekoj drugoj pojavi, koja je u korelacionoj vezi sa posmatranom.

Obe vrste povezanosti imaju veliki značaj. U okviru oređenog područja istraživanja, primarni značaj će imati ona povezanost koja bolje prikazuje vezu između pojava ili događaja. Pojave se u istraživanju registruju, dok se promene mere na objektivan način. Podaci dobijeni merenjem promena na pojavi nazivaju se promenljive. Koeficijenti korelacije ili koeficijenti asocijacije između promenljivih predstavljaju indikatore stepena povezanosti kod međusobno povezanih promenljivih (posmatranih, registrovanih ili merenih pojava). On ređe ukazaje na uzročno-posledičnu povezanost u odnosu na korelacionu, koja je češća.

Prilikom proučavanja pojava teži se obuhvatiti što je moguće veći broj međusobno povezanih pojava. Ovakav pristup istraživanju rezultira registraciju velikog broja promenljivih, a i veliki broj koeficijenata korelacije. Broj koeficijenata korelacije je jednak broju

$$\frac{n(n-1)}{2},$$

gde je n broj promenljivih. Slika 1.1 prikazuje matricu dimenzije 4×4, odnosno međusobnu povezanost između 4 promenljive. Ukupno ima $\frac{4(4-1)}{2} = 6$ korealcija.

$$\begin{bmatrix} 1 & 0.83 & 0.78 & 0.70 \\ 0.83 & 1 & 0.67 & 0.67 \\ 0.78 & 0.67 & 1 & 0.64 \\ 0.70 & 0.67 & 0.64 & 1 \end{bmatrix}$$

Slika 1.1: *Korelaciona matrica*

Međutim, veliki broj koeficijenata korealcije ne omogućava dublji i jasniji uvid u zakonitost i strukturu pojava koje se proučavaju. Zato je u nauci odavno poznat zakon štednje kojim se želi objasniti što veći broj pojava ili događaja na osnovu što manjeg broja promenljivih. Dakle, u istraživanjima se traže takve promenljive koje će omogućiti što veći broj predikcija bez istovremenog povećanja broj samih promenljivih od kojih te predikcije polaze. Generalno gledano, promenljive na osnovu kojih se vrše brojna predviđanja su interesantne ukoliko međusobno nisu povezane (problem multikolinearnosti), odnosno ako im je korelacija jednaka nuli ili su ortogonalne [4].

1.2 Faktorska analiza

Postoje mnoge definicije faktorske analize, zavisno od naučnika i godine kad je nastala, jer se metoda stalno usavršava. Jedna opšteprihvaćena definicija glasi: *''Faktorska analiza predstavlja skup statističko-matematičkih postupaka koji omogućavaju da se u većem broju promenljivih, među kojima postoji povezanost, utvrdi manji broj ''temeljnih'' promenljivih koje objašnjavaju takvu međusobnu povezanost.''* Te temeljne promenljive se nazivaju faktori.

Faktore koji se utvrđuju u postupku faktorske analize objašnjavaju međusobni odnos posmatranih promenljivih. Prema tome, cilj je da se umesto velikog broja međusobno povezanih i zavisnih proemnljivih, koje su dobijene na osnovu nekog istraživanja, utvrdi manji broj međusobno nezavisnih faktora koje mogu objasniti međusobne odnose promenljivih. Takvi faktori se smatraju uzrocima ili izvorima kovarijanse (korelacije) između promenljivih.

Utvrđivanje osnovnih faktora koji leže na jednom području istraživanja ukazuju na uzroke varijanse i kovarijanse pojava koje mogu izgledati potpuno nejasno i neodređeno bez ove analize, iako se zna korelacija između njih. Time, faktorska anliza je ta koja upućuje i usmerava na analizu temeljnih uzroka i izvora različitih pojava, koje su predmet istraživanja.

Postoje dve osnovne strategije u korišćenju analize: *eksploratorna faktorska analiza* i *konfirmatorna faktorska analiza*. U početku se najviše koristila eksploratorna strategija međutim, novijim dostignućima, sve je popularnija i konfirmatorna faktorska analiza. Bitno je napomenuti da se obe strategije faktorske analize mogu međusobno dopunjavati i da zajedno čine nezamenljivi instrument u vrlo velikom broju društvenih i prirodnih nauka.

Cilj eksploatorne faktorske analize je faktorska opisivanje određenog područja istraživanja. Odnosno, ovom analizom se žele utvrditi bazni faktori (na nekom području istraživanja) i time dobit temeljni uvid u uzroke i izvore različitih manifestacija na posmatranom području. Na primer, ova analizu se koristi ukoliko su interesantni bazni faktori koji leže u osnovi kovarijanse velikog broja ekonomskih, socijalnih, bioloških, psiholoških i drugih indikatora.

Suprotno, sasvim drugačija situacija je kod konfirmatorne faktorske analize. Ona predstavlja objektivni test određenog strukturalnog modela ili teorije. U takvoj faktorskoj analizi, istraživač polazi od unapred formulisanog modela, hipoteze ili teorije o strukturi temeljnih izvora varijanse i kovarijanse među posmatranim promenljivima. Svaka teorija ili hipoteza moraju biti podvrgnuti empirijskoj proveri ili testu. Taj test je baš faktorska analiza. Ukoliko su hipoteze i teorije potvrđene faktorkom analizom, time postoji velika verovatnoća prihvatanju takvih modela. U suprotnom, to znači da objektivni podaci ne potvrđuju zadati model.

Velika popularnost je uticala na nagli razvoj faktorske analize tako da danas postoji veliki broj tehnike unutar nje. U ovom radu, detaljno će biti prikazane tehnike:

- analiza glavnih komponenti (principal component analysis),
- factor analysis (true factor analysis)
- i regresija parcijalnih najmanjih kvadrata (partial least square regression).

Prve dve tehnike se koriste za smanjenje dimenzionalnosti matrice podataka, dok se treća odnosi na pronalaženje matrice podataka između zavisne matrice osobina i nezavisne matrice podataka, tako da ona najbolje opisuje njihove vezu.

Postoji široki spektar nauka gde se sve koristi faktorska analiza. Ona se uspešno primenjuje u psihologiji (psihologiji ličnosti, inteligencije, pamćenja, socijalnoj, industriskoj, pedagoškoj, kliničkoj, psihomotorike i dr.), fiziologiji, politici, minerologiji, fizici, hemiji, zoologiji,... Na svim ovim područjima, faktorska analiza je doprinela uočavanju reda i smisla između mnogo povezanih promenljivih i događaja [4].

1.3 Osnovne karakteristike faktorske analize

Kao što je rečeno u prethodnom delu, faktorska analiza ima široku primenu kako u prirodnim tako i u društvenim naukama. U većini tih nauka se javlja pitanja: *Koliki broj faktora utiču na analizu?* i *Koji faktori su prirodno, fizički značajni*? Faktorska analiza omogućava npr. hemičarima da probleme iz prošlosti tipa uticaja nekontrolisanih promenljivih na podatke, budu izbegnuti [2]. Ona rešava i mnoge druge probleme, u drugim naukama.

Osnovnih pet karakteristika faktorske analize su:

- *Omogućava ispitivanje veoma složenih podataka.* Faktorska analiza, kao metod multivarijacione analize, može da radi istovremene sa mnogo faktora. Ova odlika je veoma važna u prirodnim nauka jer intepretacija mnogih podataka zahteva multivarijacioni pristup;

- *Veliki broj podataka može biti analiziran.* Faktorska analiza se može efikasno spovesti korišćenjem standardnih kompjuterskih programa;

- *Razni tipovi problema se mogu proučavati.* Faktorska analiza se može primeniti bez obzira na inicijalni nedostatak uvida u podatke;

- *Podaci se mogu podjednostaviti.* Matrice se mogu modelovati sa minimalni brojem faktora i time se ogormne količine podataka mogu kompresovati u manja ''pakovanja'', bez gubitka tačnosti;

- *Faktori se mogu intepretirati na svrsishodan način.* Vrlo je bitno objasniti i razumeti prirodu dobijenih faktora. Dobijeni modeli se mogu sistematično razvijati i time kasnije koristiti u predikciji novih podataka.

1.4 Motivacioni primer

Kao što je spomenuto, faktorska analiza ima široku primenu kako u društvenim, tako i u prirodnim naukama. Primeri koji najlakše mogu oslikati samu ideju faktorske analize, a time i uvesti čitaoca u gradivo jesu primeri iz primena analize glavnih komponenti u ekonomiji. U nastavku, odnosno u samom primeru neće biti reči o načinu kako se došlo do određenih rezultata već samo komentarisanje istih.

Ideja je anketiranje stanovništva o bankarstvu i pronalaženje glavnih komponenti (faktora) koji utiču na takav stav. Ovaj istraživanje je sproveden od strane velike grupe autora i nalazi se u knjizi Marketing Research [5]. Dakle, prikupljanjem određenih informacija, moguće je uz pomoć analize izračunati komponente, koje sadrže pouzdane informacije o uzrocima ili stavovima koje imaju stanovništvo prilikom odabira banaka.

Prvo je spovedena anketa na kojoj je određena grupa stanovništva popunjavala upitnik. Upitnik je sadržao 5 rečenica (tvrđenja) na koje su ispitanici odgovarali tako što su zaokruživali brojeve od 0 do 9, zavisno da li se slažu sa navedenim tvrđenjem ili ne. Tvrđenja su glasila:

1. Manje banke imaju niže provizije u odnosu na velike banke;

2. Veća je verovatnoća da će velike banke da naprave grešku u poređenju sa malim bankama;

3. Nije neophodno da bankarski službenici budu jako ljubazni.; dovoljno je da budu učtivi;

4. Želim da me u mojoj banci lično poznaju i da se prema meni ophode posebno ljubazno;

5. Ako se finansijka institucija prema meni ponaša neučtivo, nikad više neću poveriti poverenje toj organizaciji.

Svaka rečenica predstavljati jednu promenljivu, dakle, postoji 5 promenljivih x_1, x_2, x_3, x_4, x_5. Anketu je popunjavalo skup od 15 ljudi, tj. taj skup predstavlja uzorak nad kojim je sprovedeno istraživanje. Tako je dobijena matrica podataka $\boldsymbol{X}$ koja je prikazana u tabeli 1.1.

	x_1	x_2	x_3	x_4	x_5
1.	9	6	9	2	2
2.	4	6	2	6	7
3.	0	0	5	0	0
4.	2	2	0	9	9
5.	6	9	8	3	3
6.	3	8	5	4	7
7.	4	5	6	3	6
8.	8	6	8	2	2
9.	4	4	0	8	8
10.	2	8	4	5	7
11.	1	2	6	0	0
12.	6	9	7	3	5
13.	6	7	1	7	8
14.	2	1	7	1	1
15.	9	7	9	2	1

Tabela 1.1: *Matrica podataka* $\boldsymbol{X}$ *(15 ljudi/5 tvrđenja*

Na osnovu dobijene matrice, određuje se koerlaciona matrica $\boldsymbol{R}$ (tabela 1.2). Elementi ove matrice su korelacioni koeficijenti i predstavljaju stepen povezanosti dve promenljive.

	x_1	x_2	x_3	x_4	x_5
x_1	1	0.610	0.469	0.018	0.096
x_2	0.610	1	0.230	0.190	0.319
x_3	0.469	0.230	1	0.832	0.774
x_4	0.018	0.190	0.832	1	0.927
x_5	0.096	0.319	0.774	0.927	1

Tabela 1.2: *Korelaciona matrica* $\boldsymbol{R}$

Dalje, sve se svodi na izračunavanje glavnih komponenti primenom analize glavnih komponenti. Svaka glavna komponenta predstavlja ortonormirani vektor koji je povezan sa karakterističnim korenom. Ovaj model sadrži 5 glavnih komponenti jer je rang od $\boldsymbol{R}$ jednak 5. Vrednosti njihovih karakterističnih korena i postotak varijanse, dat je u tabeli 1.3. Dobija se da prvi faktor objašnjava oko 55% varijanse (rasipanja), drugi oko 36%, treća oko 8%, itd.

Promenljive/Komponente	1. komponenta	2. komponenta
x_1	-0.295	0.853
x_2	0.048	0.920
x_3	0.938	0.278
x_4	0.950	0.228
x_5	0.940	0.268

Tabela 1.3: *Prve dve komponente*

Istraživač namerno bira samo prve dve komponente jer smatra da one objašnjavaju najveći deo varijanse i bez ostale tri, greška je veoma mala. Nakon određivanja broja komponenti, moguće je primeniti rotaciju sve radi lakše intepretacije podataka. Na osnovu podataka iz tabele 1.3, zaključuje se da promenljive x_3, x_4 i x_5 se kombinuju da bi se formirala prva komponenta (faktor) i ona se zove lični faktor. Promenljive x_1 i x_2 se kombinuju da bi se formirao drugi faktor, i on se zove faktor ''male banke''.

1.5 Kratka istorija faktorske analize

Faktorska analiza je nastala u psihologiji gde se i danas razvija. Osnivač faktorske analize je engleski psiholog Charles Spearman, koji je postavio prvu empirijsku teoriju inteligencije. Njegovo delo ''General intelligence'', Objectively Determined and Measured izdato u American Journal of Psychology 1904. godine je prvo delo takve vrste. Prema njegovoj teoriji, postoje opšti i specifični faktor, gde je opšti faktor prisutan u svim inetelektualnim aktivnostima i odgovoran je za postojanje korelacije između rezultata istih ispitanika u tim aktivnostima, a specifični faktori je odgovoran za netpotpunu korelaciju između aktivnosti [6].

U razvoju faktorske analize zaslužni su i drugi psiholozi. Postojale su dve suprostavljene strane između američkih i engleskih psihologa. Tako Thurston i drugi američki psiholozi smatraju da su grupni faktori međusobno nezavisni ili se moraju stavljati u što je moguće manju interkorelaciju. S druge strane, Burt, Vernon i drugi engleski psiholozi kreiraju psihološke modele prema kojima postoji hijerahija pojedinih faktora inteligencije. Na vrhu te hijerarhije su generalni faktori inteligencije, koji participiraju u svim intelektnim aktivnostima dok su ispod njega grupni faktori višeg reda, a zatim grupni faktori nižeg reda.

Veza između faktorske analize i matematičke statistike je godinama bila ignorisana. Ali s obzirom da polazni podaci svake faktorske analize jesu koeficijenti korelacije, njihovo izračunvanje spadaju pod statističke postupke. U ovoj oblasti, najveći doprinos je dao Pearson, odnosno njegov koncept o linijama i ravnima koji najbolje predstavljaju skup tačaka u prostoru. Ta koncepcija ne samo što je bila osnova regresione jednačine već je bila i osnova razvoja metode glavnih komponenti i glavnih osovina. Kasnije je metodu glavnih komponenti i osovina detaljnije razvio Hotelling. Takođe, bitnu vezu između statistike i faktorske analize dao je i Yulea, njegovom studijom o višestrukoj i parcijalnoj korelaciji [8].

Razvoj faktorske analize je dobio još značajniju dimenziju nakon tkz. ''rođenja'' hemometrije 1970. godine.. Razlog je to što se metoda razvijala u prekompjuterskoj eri, pa su problemi morali imati podjednostavljene pretpostavke i pomoćne uslove. Međutim, razvojem kompjutera, metoda je doživela nagli razvoj i od tada naučnik koji se bavi faktorskom analizom, pored teorijskog znanja iz oblasti koju posmatra, mora odlično da poznaje i računarstvo, matematiku i statistiku. Danas, ukupan broj tehnika analize je veliki i svake godine se razvijaju nove [2].

Glava 2

Osnovni matematički i statistički pojmovi

2.1 Matrica podataka

Matrica $\boldsymbol{X}$ predstavlja matricu podataka koja se sastoji od p karakteristika na uzorku obima n, odnsono matrica je dimenzije $n \times p$. Na primer, tabela 2.1 sadrži 8 vrsta (8 kamena) i 4 kolone (4 minerala) i predstavlja količinu minerala za svaki kamen.

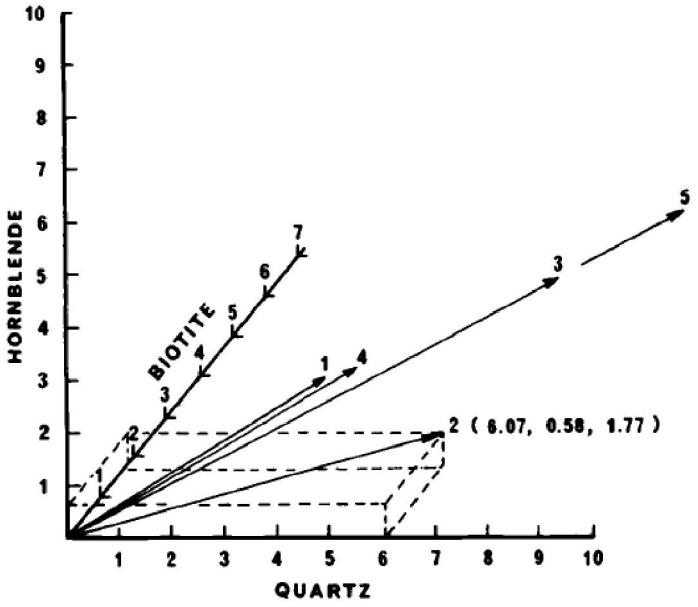
Vrste kamena	Vrste minerala			
	Kvarc	Amfibol	Biotit	Feldspat
1	4.51	2.66	0.42	4.1
2	6.07	0.58	1.77	1.54
3	6.42	1.32	4.65	4.05
4	4.46	2.16	1.41	4.47
5	8.92	2.54	4.66	2.5
6	7.60	2.39	4.14	4.49
7	5.29	1.69	3.66	3.77
8	4.73	2.65	3.29	5.08

Tabela 2.1: *Primer matrice podataka* $\boldsymbol{X}$

Dakle, opšta matrica $\boldsymbol{X}$ se sastoji od n vrsta i p kolona. I vrste i kolone kod matrice $\boldsymbol{X}$ imaju svoj naziv. Vrste matrice $\boldsymbol{X}$ se nazivaju objekti (uzorci) i predstavljaju entitete na kojima su vršena merenja. Kolone matrice $\boldsymbol{X}$ se nazivaju promenljive i one predstavljaju skup registrovanih vrednosti za svaki objekat.

2.2 Geometrijska intepretacija

Kao što se podaci mogu predstaviti algebarski, tako se mogu predstaviti i geometrijski. U geometrijskoj intepretaciji, promenljive grade koordinantnu osu za objekte, pa se tako može reći da su objekti definisani unutar prostora promenljivih. Slika 2.1 prikazuje objekte unutar tri-dimenzionalne ose koristeći tabelu 2.1 ali tako što se koriste sam prve tri promenljive, da bi se predstavilo u trodimenzionalnom prostoru.



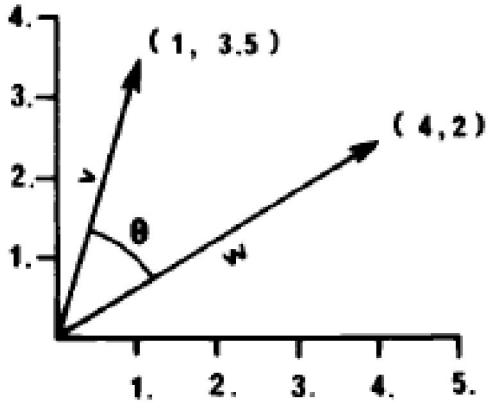
Slika 2.1: *Geometrijska intepretacija matrice podataka* $\boldsymbol{X}$ *(bez 4. kolone)*

Uopšteno, ukoliko je matrica dimenzije $n \times p$, tada se objekti mogu posmatrati kao n vektora unutar p-dimenzionalnog prostora. Ovaj prostor se dalje može definisati kao Euklidski prostor promenljivih.

Dalje, veoma je bitno izračunavanje ugla između vektora podataka unutar prostora. Koristi se jednačina:

$$\cos\theta = \frac{v'w}{(v'v)^{0.5}(w'w)^{0.5}} .$$

Dakle, izračunavanjem kosinusa ugla putem date formule, lako se dobija ugao između vektora $\boldsymbol{v}$ i $\boldsymbol{w}$ (slika 2.2).



Slika 2.2: *Ugao između dva vektora*

2.3 Matrični proizvod veće i manje dimenzije

U multivarijacionoj analizi postoje dva tipa proizvoda matrice $\boldsymbol{X}$ same sa sobom. To su matrični prozivod veće dimenzije (*major product moment*) i matrični proizvod manje dimenzije (minor product moment) i oba termina je definisao Horst (1963). Matrični proizvod veće dimenzije je definisan kao proizvod matrice $\boldsymbol{X}$ sa transponovanom matricom $\boldsymbol{X'}$, tj.

$$\boldsymbol{C} = \boldsymbol{X}\boldsymbol{X}' .$$

Matrica $\boldsymbol{C}$ je kvadratna i simetrična, dimenzije $n \times n$.

Matrični proizvod manje dimenzije je definisan kao proizvod transponovane matrice $\boldsymbol{X'}$ sa matricom $\boldsymbol{X}$, tj.

$$\boldsymbol{E} = \boldsymbol{X}'\boldsymbol{X}.$$

Matrice $\boldsymbol{E}$ je, takođe, kvadratna i simetrična ali dimenzije $p \times p$.

2.4 Ortogonalni i ortonormirani vektori i matrice

Za vektor $\boldsymbol{x}$ se kaže da je normalizovan ako mu je norma (dužina) jednaka 1. Veoma lako se svaki vektor može normalizovati deljenjem vektora $\boldsymbol{x}$ sa njegovom normom 2, $(\boldsymbol{x}'\boldsymbol{x})^{1/2}$, odnosno

$$\frac{\boldsymbol{x}}{(\boldsymbol{x}'\boldsymbol{x})^{1/2}} .$$

Dva vektora $\boldsymbol{u}$ i $\boldsymbol{v}$ su ortogonalna ukoliko je njihov unutrašnji proizvod jednak nula, tj. $\boldsymbol{u}'\boldsymbol{v} = 0$. Takođe, ugao između njih je 90° stepeni.

Ovaj koncept se može preneti i na matrice. Ukoliko matrični proizvod manje dimenzije matrice $\boldsymbol{X}$ daje dijagonalnu matricu $\boldsymbol{D}$, tada je matrica $\boldsymbol{X}$ ortogonalna. Dakle, ukoliko važi

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{D}$$

to implicira da su parovi kolona matrice $\boldsymbol{X}$ međusobno ortogonalni.

Ukoliko važi da je $\boldsymbol{D} = \boldsymbol{I}$, tada je matrica $\boldsymbol{X}$ ortonormirana. To znači da su sve kolone dužine 1 i da su međusobno ortogonalne. Specijalno, od posebnog interesa je kvadratna ortonormirana matrica $\boldsymbol{Q}$ koja ima svojstva:

$$\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I} \quad \text{i} \quad \boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{I} .$$

Postoji beskonačan broj kvadratnih ortonormiranih matrica bilo koje dimenzije.

2.5 Deskriptivne statistike u matričnoj notaciji

U ovom delu su predstavljene neke jednostavne statističke formule u matričnoj formi. Veoma je bitno sve te jednakosti razumeti jer one imaju fundamentalnu ulogu u kasnijem definisanju metoda faktorske analize.

Aritmetička sredina (srednja vrednost) matrice $\boldsymbol{X}$, dimenzije $n \times p$, se računa kao:

$$\hat{\boldsymbol{x}} = \frac{\mathbf{1}'\boldsymbol{X}}{\mathbf{1}'\mathbf{1}},$$

gde je $\hat{\boldsymbol{x}}$ vektor dimenzije $1 \times p$, čije komponente predstavljaju srednje vrednosti, a $\mathbf{1}$ je jedinični vektor dimenzije $n \times 1$.

Razlika između registrovane vrednosti promenljive i srednje vrednosti promenljive se naziva devijacija (centrirana vrednost). Sa y_{ij} se označava devijacija i ona je jednaka:

$$y_{ij} = x_{ij} - \bar{x}_j \, .$$

Podrazumeva se da važi:

$$\sum_{j=1}^{n} y_{ij} = \sum_{j=1}^{n} x_{ij} - \sum_{j=1}^{n} \bar{x}_j = \sum_{j=1}^{n} x_{ij} - n\bar{x}_j = \sum_{j=1}^{n} x_{ij} - n \sum_{j=1}^{n} \frac{x_{ij}}{n} = 0.$$

Pretvaranjem promenljivih u centrirane promenljive dobijaju se promenljive čija je srednja vrednost jednaka nuli. Matrica čiji su elementi centrirane vrednosti se naziva matrica centriranih podataka i označava sa $\boldsymbol{Y}$.

Varijansa promenljive je mera rasipanja individualnih vrednosti oko srednje vrednosti. Definisana je kao prosečna vrednost zbira kvadrata centriranih vrednosti (devijacionih vrednosti), tj.

$$s_j^2 = \sum_{i=1}^{n} y_{ij}^2 / n \, .$$

Prethodna jednakost je podeljena sa n, iako se češće se koristi vrednost $n - 1$ jer se tako dobija nepristrasna ocena za varijansu. Koristeći matričnu notaciju, varijansa promenljive se dobija na sledeći način

$$s_j^2 = \frac{\boldsymbol{y}_j'\boldsymbol{y}_j}{\mathbf{1}'\mathbf{1}}.$$

Kovarijansa odražava vezu između dve promenljive. Na primer, za promenljive $\boldsymbol{x}_i$ i $\boldsymbol{x}_j$ kovarijansa je definisana kao prosečna vrednost zbira proizvoda centriranih vrednosti za sve objekte, odnosno

$$s_{ij} = \sum_{k=1}^{n} y_{ki} y_{kj} / n \, .$$

Za celokupnu matricu podataka $\boldsymbol{X}$, matrica kovarijanse se dobija iz

$$\boldsymbol{S} = \boldsymbol{Y'Y}/\boldsymbol{1'1} \, .$$

Dakle, kovarijansna matrica je matrični proizvod manje dimenzije centriranih matrica podataka podaljene sa n.

Standardna devijacija je definisana kao kvadratni koren varijanse. Dakle, standardnu devijaciju dobijamo iz sledeće jednačine

$$s_j = \frac{(\boldsymbol{y'y})^{1/2}}{n^{1/2}} .$$

Standardizovana promenljiva za pojedinačni objekat, za promenljivu $\boldsymbol{x}_j$, je dat kao

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} .$$

Standardizovana promenljiva ima srednju vrednost nula i devijaciju jednaku jedan. Formula za standardizaciju svih promenljivih u matrici podataka je

$$\boldsymbol{Z} = \boldsymbol{YD}^{-1/2} \, ,$$

gde je $\boldsymbol{D}$ dijagonalna matrica formirana od dijagonalnih elemenata kovarijansne matrice $\boldsymbol{S}$, a $\boldsymbol{Y}$ matrica centriranih vrednosti.

Mera povezanosti između dve promenljive ima veoma važnu ulogu u faktorskoj analizi. Mera linearne povezanosti se naziva Pirsonov koeficijent linearne korelacije. Definisana je kao odnos između kovarijanse dve promenljive i proizvoda njihovih standardnih devijacija:

$$r_{ij} = \frac{s_{ij}}{s_i s_j} = \frac{\sum_{k=1}^{n} y_{ki} y_{kj}}{\left(\sum_{k=1}^{n} y_{ki}^2 \sum_{k=1}^{n} y_{kj}^2\right)^{1/2}} .$$

Matrcija čiji su elementi r_{ij}, naziva se korelaciona matrica $\boldsymbol{R}$ i izračunava se na sledeći način:

$$\boldsymbol{R} = \boldsymbol{Z'Z}/\boldsymbol{N} \, .$$

Korelaciona matrica je kvadratna i simetrična.

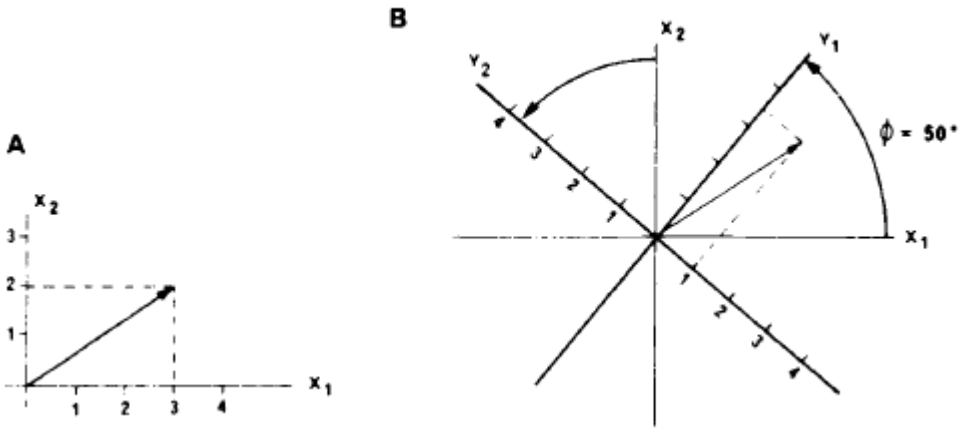
Zanimljivo je uočiti da su elementi r_{ij} matrice $\boldsymbol{R}$ definisani na isti način kao kosinusi ugla između dve vektora, u ovom slučaju između vektora vrsta i i j matrice $\boldsymbol{Y}$:

$$r_{ij} = \frac{\sum_{k=1}^{n} y_{ki} y_{kj}}{\left(\sum_{k=1}^{n} y_{ki}^2 \sum_{k=1}^{n} y_{kj}^2\right)^{1/2}} = \frac{y_i' y_j}{(y_i' y_i)^{1/2} (y_i' y_j)^{1/2}} = \cos(\theta_{ij}).$$

Dakle, matrica $\boldsymbol{R}$ sadrži kosinuse ugla između svaka dva vektora vrste matrice $\boldsymbol{Y}$.

2.6 Rotacija koordinatnog sistema

Rotiranje koordinatni sistemi se može opisati pomoću matričnih operacija. Na slici 2.3 A brojevi x_1, x_2 su koordinate vektora $\boldsymbol{x}$ u kordinatnom sistemu. Neka se iz nekog razloga žele rotirati ose za ugao ϕ° u pravcu suprotnom od kretanje kazaljke na satu, kao što je prikazano na slici 2.3 B. Na ovaj način se ortonormirani koordinatni sistem prevodi u novi ortnormirani koordinatni sistem. Nakon rotiranja, problem je naći nove koordinatne y_1 i y_2 tačke u odnosu na nove ose.



Slika 2.3: *Rotacija koordinatnih osa za* ϕ°

Koristeći elementarne trigonometriske funkcije, nove koordinate se dobijaju na sledeći način:

$$y_1 = \cos\phi\, x_1 + \sin\phi\, x_2\,,$$
$$y_2 = -\sin\phi\, x_1 + \cos\phi\, x_2\,.$$

U matričnoj formi, ove dve jednačine se predstavlju kao

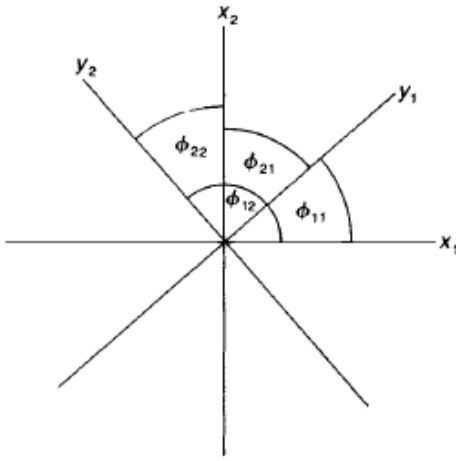
$$\boldsymbol{y}' = \boldsymbol{xT},$$

gde je

$$\boldsymbol{T} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}.$$

Dalje, predstavljen je gornji postupak ali u opštem slučaju. Na slici 2.4, ugao ϕ_{ij} predstavlja ugao između i-te stare koordinatne ose i j-te nove ose. Nakon rotacije, uočava se sledeće veze između uglova:

$$\phi_{12} = \phi_{11} + 90^{\circ}$$
$$\phi_{21} = \phi_{11} - 90^{\circ}$$
$$\phi_{22} = \phi_{11}$$



Slika 2.4: *Grafik rotacije koordinatnih osa*

Primenom trigonometrijskih funkcija, sledi:

$$\sin\phi_{11} = \sin(\phi_{21} + 90^{\circ}) = \cos\phi_{21}$$
$$-\sin\phi_{11} = -\sin(\phi_{12} - 90^{\circ}) = \sin(90^{\circ} - \phi_{12}) = \cos\phi_{12}$$

Konačno, smenom se dobija:

$$y_1 = \cos\phi\, x_1 + \sin\phi\, x_2 = \cos\phi_{11}\, x_1 + \cos\phi_{21}\, x_2$$
$$y_2 = -\sin\phi\, x_1 + \cos\phi\, x_2 = \cos\phi_{12}\, x_1 + \cos\phi_{22}\, x_2\,.$$

Prikazana procedura se može uopštiti za višedimenzionalni koordinatni sistem. Dakle, sistem jednačina za p-osa je

$$y_1 = \cos\phi_{11}\, x_1 + \cos\phi_{21}\, x_2 + \ldots + \cos\phi_{p1}\, x_p$$
$$\vdots$$
$$y_p = \cos\phi_{1p}\, x_1 + \cos\phi_{2p}\, x_2 + \ldots + \cos\phi_{pp}\, x_p\,.$$

Smenom $t_{ij} = \cos\phi_{ij}$, gde se i odnosi na staru koordinatnu osu, a j na novu koordinatnu osu, jednačine se u matričnoj notaciji mogu zapisati kao

$$\boldsymbol{y}' = \boldsymbol{x}'\boldsymbol{T}.$$

Za skup n vektora vrsta u matrici $\boldsymbol{X}$, jednačina

$$\boldsymbol{Y} = \boldsymbol{XT}$$

daće koordinate svih n vektora vrsta u odnosu na p rotiranih osa. Matrica $\boldsymbol{T}$ se naziva transformaciona matrica. Kako se ortonormirani koordinatni sistem prevodi u novi ortnormirani koordinatni sistem, za matricu $\boldsymbol{T}$ mora da važi $\boldsymbol{T'T} = \boldsymbol{I}$, odnosno da je ortonormirana.

Na kraju kratak zaključak. Množenjem matrice podataka $\boldsymbol{X}$ sa ortonormiranom matricom se može smatrati kao jednostavna rotacija. Nove rotirane ose se mogu posmatrati kao nove promenljive. Kolone u matrici $\boldsymbol{Y}$ su nove promenljive koje predstavljaju linearnu kombinaciju kolona (promenljivih) u matrici $\boldsymbol{X}$. Stoga se elementi u matrici $\boldsymbol{T}$ mogu smatrati koeficijentima ovih linearnih kombinacija. Takođe, vektori vrsta matrica $\boldsymbol{X}$ predstavljaju koordinate objekata u odnosu na originalne promenljivie (ose), dok vektori vrste u matrici $\boldsymbol{Y}$ predstavljaju objekte u novodizajniranim promenljivima.

2.7 Rang matričnog proizvoda

Pronalaženje ranga matrice podataka $\boldsymbol{X}$ predstavlja jedan od glavnih ciljeva faktorske analize. Mnoge metode faktorske analize određuju rang matrice koristeći analizu matričnog proizvoda matrica nego samu analizu matrice podataka. Videće se kasnije da je rang proizvoda momenta matrice jednaka broju faktora. Tako, ukoliko je rang matrice $\boldsymbol{X}$ jednak r i ako je

$$\boldsymbol{B} = \boldsymbol{X'X}$$

tada je rang i matrice $\boldsymbol{B}$ jednak r. Slično ukoliko je

$$\boldsymbol{C} = \boldsymbol{XX'}$$

tada je rang matrice $\boldsymbol{C}$ jednak r.

Izračunavanje korelacione matrice je početni korak u svakoj faktorskoj analizi. Ona se dobija matričnim proizvodom manje dimenzije standardizovane matrice podataka. Njeni elementi sadrže kosinuse uglova između svih mogućih elemenata unutar standardizovanog vektora kolone. Moguće je geometriski prikazati odnose između vektora čiji su kosinusi predstavljeni korelaciojom. Na primer, slika 2.5A prikazuje tri vektora koji korespondiraju korelacionoj matrici $\boldsymbol{R}_1$:

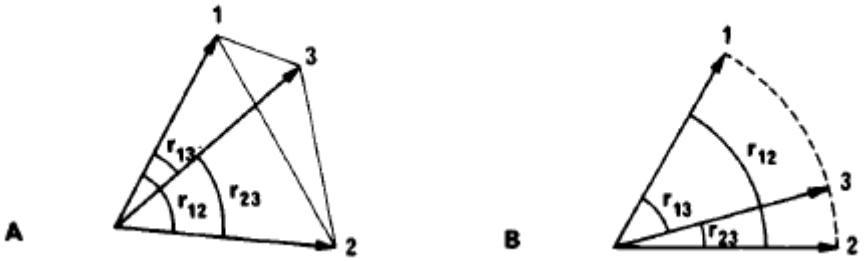
$$\boldsymbol{R}_1 = \begin{bmatrix} 1 & 0.5 & 0.866 \\ 0.5 & 1 & 0.707 \\ 0.866 & 0.707 & 1 \end{bmatrix}$$

Vektori vrsta matrice podataka se smeštaju u trodimenzionalni prostor, i na osnovu slika 2.5 A se zaključuje da su ta tri vektora linearno nezavisna i rang matrice podataka je jednak 3.

Na slika 2.5 B, prikazani su vektori druge korelacione matrice $\boldsymbol{R}_2$:

$$\boldsymbol{R}_2 = \begin{bmatrix} 1 & 0.5 & 0.707 \\ 0.5 & 1 & 0.966 \\ 0.707 & 0.966 & 1 \end{bmatrix}.$$

Geometrijski uslovi primoravaju tri vektora da leže u dvodimenzionalnom prostoru. Tako, bilo koji vektor je linearno zavisan na druga dva vektora i time je rang matrice jednak 2.



Slika 2.5:

Na slici 2.5 A vektori konstuišu tetraedar. Zapremina tetraedra je jednak determinanti korelacione matrice. S obzirom da su vektori na slici 2.5 B koplanarni, zapremina tetraedra koji oni konstruišu je jednak nuli. Tako važi:

$$|\boldsymbol{R}_1| = 0.1124 \text{ i } |\boldsymbol{R}_2| = 0\,.$$

Veza između linearno nezavisnih vektora, ranga, determinante i geometrijske konfiguracije vektora se može uopštiti na bilo koju dimenziji. Opšti zaključci glase:

1. Za bilo koju matricu, čiji je rang (r) manji od njene manje dimenzije (p), mogu se posmatrati dva skupa vektora: prvi koji sadrži $p - r$ linearnih zavisnih vektora i skup od r linearno nezavisnih vektora;
2. Rang određuje dimenziju prostora u kom su sadržani svi vektori, a r linearno nezavisnih vektora čine jednu bazu od beskonačno mnog mogućih baza ovog prostora.

2.8 Karakteristični koreni i vektori

U prethodnom delu je razmatran rang matrice podataka. Karakteristični koreni i vektori ne samo što učestvuju u određivanju vrednosti ranga matrice već se koriste i u mnogim drugim slučajevima. Postoji široki spektar njihovog tumačenja i sigurno imaju fundamentalnu ulogu u mnogim statističkim metodama, posebno u faktorskoj analizi.

Neka je $\boldsymbol{R}$ korelaciona ili kovarijansna matrica. Karakteristični vektor matrice $\boldsymbol{R}$ je vektor $\boldsymbol{u}$, tako da važi

$$\boldsymbol{R}\boldsymbol{u} = \lambda \boldsymbol{u}\,,$$

gde je λ skalar koji se naziva karakteristični koren. Da bi se odredio vektor $\boldsymbol{u}$, izvršiće se niz transformacija jednakosti:

$$\boldsymbol{Ru} = \lambda \boldsymbol{u} \leftrightarrow \boldsymbol{Ru} - \lambda \boldsymbol{u} = \boldsymbol{0} \ \leftrightarrow (\boldsymbol{R} - \lambda \boldsymbol{I})\boldsymbol{u} = \boldsymbol{0}.$$

Ova jednačina implicira da su $\boldsymbol{u}$ i svi vektori vrste matrice $\boldsymbol{R} - \lambda \boldsymbol{I}$ ortogonalni.

Ukoliko je $\boldsymbol{R} - \lambda \boldsymbol{I}$ regularna matrica, množenjem jednačina $(\boldsymbol{R} - \lambda \boldsymbol{I})\boldsymbol{u} = \boldsymbol{0}$ sa $(\boldsymbol{R} - \lambda \boldsymbol{I})^{-1}$ dobija se trivijalno rešenje $\boldsymbol{u} = 0$. Ako je $\boldsymbol{R} - \lambda \boldsymbol{I}$ singularna matrica, dobija se netrivijalno rešenje. Tada je vrednost determinante od $\boldsymbol{R} - \lambda \boldsymbol{I}$ jednaka nuli, odnosno

$$|\boldsymbol{R} - \lambda \boldsymbol{I}| = 0.$$

Ova jednačina se naziva karakteristična jednačina, a njeni koreni su karakteristični koreni. S obzirom da je matrica $\boldsymbol{R}$ realna, kvadratna, simetrična matrica, dimenzije $p \times p$, postoji p relanih karakterističnih korena $\lambda_1, \lambda_2, \dots, \lambda_p$. Svaki karakteristični koren je pridružen jednom karakterističnom vektoru, pa ubacivanjem vrednosti λ_i u jednačinu $\boldsymbol{Ru}_i = \lambda_i \boldsymbol{u}_i$, lako se može izračunati p karakterističnih vektora $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_p$. Važno je napomenuti da su karakteristični vektori ortonormirani.

Dalje, neka su dobijeni karakteristični vektori poređani po dijagonali matrice $\boldsymbol{\Lambda}$, i neka kolone matrice $\boldsymbol{U}$ predstavljaju dobijene karakteristične vektore:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 & & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & & 0 \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_p \end{bmatrix} \text{ i } \boldsymbol{U} = \left[\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \dots, \boldsymbol{u}_p\right].$$

Matrica $\boldsymbol{U}$ je kvadratna i ortnornomirana, pa važi

$$\boldsymbol{U'U} = \boldsymbol{UU'} = \boldsymbol{I} \ ,$$

Uz pomoć matrica $\boldsymbol{U}$ i $\boldsymbol{\Lambda}$, sledi:

$$\boldsymbol{Ru} = \lambda \boldsymbol{u} \ \rightarrow \boldsymbol{RU} = \boldsymbol{U\Lambda},$$

odakle važi

$$\boldsymbol{R} = \boldsymbol{U\Lambda U'} \ .$$

Ova jednačina ima fundamentalnu ulogu u mnogim multivarijacionim metodama. Ona pokazuje da se simetrična matrica $\boldsymbol{R}$ može predstaviti preko karakterističnih korena i vektora, odnosno:

$$\boldsymbol{R} = \lambda_1 \boldsymbol{u}_1 \boldsymbol{u}_1' + \lambda_2 \boldsymbol{u}_2 \boldsymbol{u}_2' + \cdots + \lambda_p \boldsymbol{u}_p \boldsymbol{u}_p'.$$

Matricu $\boldsymbol{R}$ je linearna kombinacija proizvoda koeficijenata λ_i i matrice $\boldsymbol{u}_i \boldsymbol{u}_i'$. Množenjem gornje jednakosti sa $\boldsymbol{u_i}$ (sa desne strane), dobija se

$$\boldsymbol{Ru}_i = \lambda_i \boldsymbol{u}_i.$$

Množenjem iste jednačine sa leve strane i korsiteći da je $\boldsymbol{u}_i\boldsymbol{u}_i'\boldsymbol{u}_j'\boldsymbol{u}_j = 0$, sledi

$$\boldsymbol{\Lambda} = \mathbf{U}'\mathbf{R}\mathbf{U}.$$

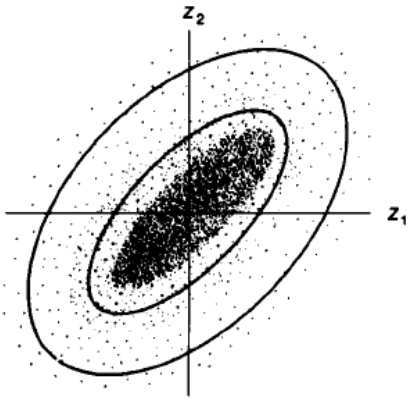
Karakteristični koreni mogu pomoći u određivanju mnogih svojstva matrice. Neke od osobina, koje važe, su:

1. Trag matrice $\boldsymbol{\Lambda}$ je jednaka tragu matrice $\boldsymbol{R}$;
2. Proizvod karakterističnih korena je jednak determinanti matrice $\boldsymbol{R}$;
3. Broj nenula karakterističnih korena je jednak rangu matrice $\boldsymbol{R}$.

Ukoliko je kvadratna matrica $\boldsymbol{R}$ dimenzije $p \times p$ i ukoliko postoji m karakterističnih korena vrednosti nula, tada je rang matrice $\boldsymbol{R}$ jednak $p - m$. Štaviše, skup od $p - m$ karakteristični vektora povezanih sa $p - m$ karakterističnih korena formira skup ortognalnih baza vektora koji generiše prostor od $p - m$ linearno nezavisnih vektora iz $\boldsymbol{R}$.

2.9 Geometrijska intepretacija karakterističnih korena i vektora

Slika 2.6 prikazuje dijagram rasipanja podataka za dve promenljive dobijene na osnovu uzorka. Pretpostavlja se da su podaci u standardizovanoj formi i da imaju normalnu raspodelu. Oko tačaka na grafiku se mogu ocrtati konture tako da se unutar njih nalazi 66%, odnosno 95% podataka (tačaka) respektivno S obzirom da su podaci normalno raspoređeni, važi da su konture elipsastog oblika, kao na slici. Takođe, ukoliko su promenljive nekorelisane, elipsa će dobiti oblik kruga, a s druge strane su potupno korelisane, elipsa će se pretvoriti u pravu. S obzirom da je u prethodnom delu pretpostavljeno da ima p promenljivih, podaci će formirati p-dimenzionalni hiperelipsoid.



Slika 2.6: *Tačke podataka i konture*

Pearson (1901) i Hotelling (1933) su pokazali da se glavna i sporedna osa hiperelipsoida može pronaći pomoću karakterističnih vektora korelacione matrice $\boldsymbol{R}$. Jednačina hiperelipsoida glasi

$$\boldsymbol{w}'\boldsymbol{R}^{-1}\boldsymbol{w} = \boldsymbol{c}\,,$$

gde je $\boldsymbol{x}$ vektor koordinata za tačke na elipsoidu, a $\boldsymbol{c}$ je konstanta. Glavna osa elipsoida se izračunava preko tačaka na elipsoidu koje su najudaljenije od koordinatnog početka. Da bi se pronašli te tačke, potrebno je naći tačke na elipsoidu koje su najudaljenije od koordinatnog početka. Kvadrat rastojanja tačke $\boldsymbol{w}$ na elipsoidu do koordinatnog početka je $\boldsymbol{w'w}$, što daje zaključak da je potrebno naći maksimum funkcije $\boldsymbol{w'w}$ pod uslovom $\boldsymbol{w'R^{-1}w = c}$. Dakle optimizacioni problem glasi:

$$\boldsymbol{w'w} \rightarrow max$$

uslov: $\boldsymbol{w'R^{-1}w = c}$.

Izračunavanjem Lagražijana, pa zatim diferenciranjem po $\boldsymbol{w}$, lako se dobija:

$$\boldsymbol{Rw} = \lambda \boldsymbol{w}\,,$$

gde je $\boldsymbol{w'w} = \lambda$. Sledi da je $\boldsymbol{w}$ karakteristični vektor sa odgovarajućim najvećim karakterističnim korenom λ. Vektor $\boldsymbol{w}$ predstavlja pravac glavne ose elipsoida. Dužina glavne ose tj. rastojanje od koordinatnog početka do tačke na elipsoidu, je jednaka $\sqrt{\lambda}$. Na sličan način, može se pokazati da drugi karakteristični vektor odgovara najvećoj sporednoj osi elipsoida, treći karakteristični vektor odgovara trećoj najvećoj sporednoj osi elispoida, itd.

Na osnovu isloženih činjenica o karakterističnim korenima i vektorima i njihovoj geometrijskoj intepretaciji, dolazi se do sledećih zaključaka:

1. Položaj karakterističnog vektora duž glavne ose na hiperelipsoidu se podudara sa pravcem maksimalne varijanse podataka. Karakteristični vektor čija je karakteristična vrednost najveća određuju pravac maksimalne varijanse podataka; karakteristični vektor povezan sa drugom najvećom karakterističnom vrednošću određuje drugi smer maksimalne varijanse al tako da je ortogonalan na prvi vektor, itd;

2. Elementi vektora vrste matrice $\boldsymbol{W}$ karakterističnih vektora su koeficijenti koji rotiraju ose promenljivih na poziciju duž glavne i sporedne ose. Ovi elementi vektora su kosinusi uglova nastali rotacijom sistema starog ka novom sistemu koordinatnog osa.

3. Karakteristični vektori su linearno nezavisni vektori i njihovom linearnom kobinacijom se mogu dobiti originalne promenljive. Takođe, oni se mogu posmatrati i kao nove promenljive koje imaju lepa svojstva: da su nekorelisane i da objašnjavaju varijansu kod podataka u opadajućoj hijerahiji;

4. Kvadratni koren karakteristične vrednosti predstavlja dužinu karakterističnog vektora;

5. Ukoliko je karakteristična vrednost jednaka nuli to ukazuje da je dužina odgovarjuće sporedne ose jednaka nuli. To ukazuje na to da je dimenzionalnost prostora koji sadrži podatke manja od originalnog prostora, odnosno da je rang matrice podataka manji od dimenzije originalnog prostora.

2.10 SVD dekompozicija

Singularna dekompozicija (*singular value decompozition* - SVD) je veoma važna tehnika za dekomponovanje matrice u matrice karakterističnih korena i vektora. SVD dekompoziciju je kreirao Shrager. Nakon otkrića, sama dekompozicija je doživela ogromu popularnost u svim sferama istraživanja, gde se koriste karakteristični koreni i vektori.

Dakle, neka je data matrica podataka $\boldsymbol{X}$, dimenzije $n \times p$, gde je $n < p$ i neka je r rang matrice. U većini slučajeva će važiti da je dimenzija matrice $\mathbf{X}$ jednaka $r \times p$ $(rang(\boldsymbol{X}) = r = n)$, međutim koristiće se opšte zapis da je dimenzija jednaka $n \times p$. Bazična struktura SVD dekompozicje matrice $\boldsymbol{X}$ je

$$\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Gamma}\boldsymbol{U}',$$

gde je:

- matrica $\boldsymbol{V}$, dimenzije $n \times r$, čije su kolone ortonormirane, važi $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$;
- matrica $\boldsymbol{U}$, dimenzije $p \times r$, čije su kolone takođe ortonormirane, i važi $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$;
- matrica $\boldsymbol{\Gamma}$ je dijgonalna kvadratna matrica, dimenzije $r \times r$, sa pozitivnim singularnim vrednostima $\gamma_1, \gamma_2, \dots, \gamma_r$, gde je, $\gamma_i = \lambda^{1/2},\ i = 1,2, \dots, r$.

Takođe, bazična struktura SVD dekompozicije matrice $\boldsymbol{X}$ se može predstaviti i kao suma:

$$\boldsymbol{X} = \gamma_1 \boldsymbol{u}_1 \boldsymbol{v}_1' + \gamma_2 \boldsymbol{u}_2 \boldsymbol{v}_2' + \cdots + \gamma_r \boldsymbol{u}_r \boldsymbol{v}_r' = \sum_{i=1}^{r} \gamma_i \boldsymbol{u}_i \boldsymbol{v}_i' .$$

Sledi da je matrica $\boldsymbol{X}$, ranga r, predstavljena kao linearna kombinacija r matrica $(\boldsymbol{u}_i \boldsymbol{v}_i')$, ranga 1.

U daljem tekstu, istaknute su neke bitnije veze između matrica $\boldsymbol{U}$, $\boldsymbol{V}$ i $\boldsymbol{\Gamma}$ u bazičnoj stukturi i karakterističnih korena i vektora unutar matričnog proizvoda veće i manje dimenzije matrice $\boldsymbol{X}$. Dakle, fundamentalna svojstva su:

1. Matriči proizvod veće dimenzije $\boldsymbol{X}\boldsymbol{X}'$, dimenzije $n \times n$, ima r pozitivnih i $n - r$ nula karakterističnih korena. Skup od r pozitivnih karakterističnih korena $\gamma_1^2, \gamma_2^2, \dots, \gamma_r^2$ su povezani sa skupom od r odgovarajućih karakterističnih vektora $v_1, v_2, \dots, v_r$.
2. Matriči proizvod manje dimenzije $\boldsymbol{X}'\boldsymbol{X}$, dimenzije $p \times p$, sadrži r pozitivnih i $p - r$ nula karakterističnih korena. Skup od r pozitivnih karakterističnih korena $\gamma_1^2, \gamma_2^2, \dots, \gamma_r^2$ su povezani sa skupom od r odgovarajućih karakterističnih vektora $u_1, u_2, \dots, u_r$.
3. Pozitivni karakteristični koreni $\boldsymbol{X}\boldsymbol{X}'$ i $\boldsymbol{X}'\boldsymbol{X}$ su isti. Štaviše, ukoliko je v_m karakteristični koren matrice $\boldsymbol{X}\boldsymbol{X}'$, u_m karakteristični vektor matrice $\boldsymbol{X}'\boldsymbol{X}$ i γ_m karakteristični koren povezan sa ova dva vektora, tada važi sledća veza:

$$\boldsymbol{v}_m = (1/\gamma_m)\mathbf{X}\boldsymbol{u}_m \text{ i } \boldsymbol{u}_m = (1/\gamma_m)\mathbf{X}\boldsymbol{v}_m ,$$

odnosno u matričnoj normi

$$\boldsymbol{V} = \boldsymbol{X}\boldsymbol{U}\boldsymbol{\Gamma}^{-1} \text{ i } \boldsymbol{U} = \boldsymbol{X}'\boldsymbol{V}\boldsymbol{\Gamma}^{-1}.$$

2.11 NIPALS dekompozicija

NIPALS (*nonlinear iterative partial least squares*) predstavlja algoritam za dobijanje karakterističnih korena i vektora direktno iz centrirane matrice podataka $\boldsymbol{Y}$. Tokom postupka ekstrahuju se parovi karakterističnih korena i vektora, počevši od korena sa najvećom vrednošću pa do najmanje. Prednost ovog postupka jeste da se može prekinuti u bilo kom trenutku (recimo kada je izračunat dovoljan broj karakterističnih korena i vektora), bez javljanja bilo kakve greške.

NIPALS dekompozicija se oslanja na SVD notaciju. Dakle, SVD dekompozicija matrice $\boldsymbol{X}$ glasi:

$$\boldsymbol{X} = \gamma_1 \boldsymbol{u}_1 \boldsymbol{v}_1' + \gamma_2 \boldsymbol{u}_2 \boldsymbol{v}_2' + \cdots + \gamma_r \boldsymbol{u}_r \boldsymbol{v}_r' = \sum_{i=1}^{r} \gamma_i \boldsymbol{u}_i \boldsymbol{v}_i' = \sum_{i=1}^{r} \boldsymbol{X}_j.$$

U prvom koraku, za početnu vrednost vektora $\boldsymbol{v}_1'$ uzima se proizvoljni vektor, koji se zatim normalizuje. S obzirom da su karakteristični vektori međusobno ortonormirani, množenjem vektora $\boldsymbol{v}_1$ sa $\boldsymbol{X}$, sledi:

$$\boldsymbol{X}\boldsymbol{v}_1 = \gamma_1 \boldsymbol{u}_1 \boldsymbol{v}_1' \boldsymbol{v}_1 + \gamma_2 \boldsymbol{u}_2 \boldsymbol{v}_2' \boldsymbol{v}_1 + \cdots + \gamma_r \boldsymbol{u}_r \boldsymbol{v}_r' \boldsymbol{v}_1 = \gamma_1 \boldsymbol{u}_1.$$

Ovaj proizvod, kao što se vidi, se zamenjuje vektorom $\boldsymbol{u}_1$ sa normalizovanom konstantom γ_1. Na isti način, množenjem matrice $\boldsymbol{X}$ sa $\boldsymbol{u}_1'$, sledi:

$$\boldsymbol{u}_1' \boldsymbol{X} = \gamma_1 \boldsymbol{v}_1'.$$

Ova dva postupka se ponavljaju sve dok se ne dobije zadovoljavajuća aproksimacija vektora $\boldsymbol{u}_1$ i $\boldsymbol{v}_1$.

Prva rezidualna matrica $\boldsymbol{E}_1$ se izračunava kao

$$\boldsymbol{E}_1 = \boldsymbol{X} - \boldsymbol{X}_1 = \boldsymbol{X} - \gamma_1 \boldsymbol{u}_1 \boldsymbol{v}_1'.$$

Na osnovu dobijene rezidualne matrice, donosi se odluka da li da se nastavi sa dekompozicijom ili ne. Vekore $\boldsymbol{u}_2$ i $\boldsymbol{v}_2$ izračunavamo na isti način kao što je opisano u prethodnom delu s tim što se matrica $\boldsymbol{X}$ zameni sa dobijenom rezidualnom matricom $\boldsymbol{E}_1$.

U opštem slučaju, formule za izračunavanje karakterističnih vektora i korena su:

$$\boldsymbol{E}_{j-1} \boldsymbol{v}_j = \boldsymbol{u}_j \gamma_j \quad \text{i} \quad \boldsymbol{u}_j' \boldsymbol{E}_{j-1} = \gamma_j \boldsymbol{v}_j,$$

gde je

$$\boldsymbol{E}_j = \boldsymbol{X} - \textstyle\sum_{i=1}^{j} \boldsymbol{X}_i.$$

S obzirom da NIPALS uključuje princip najmanjih kvadrata, postupak ekstrahuje karakteristične korene i vektore u hronološkom redosledu, od najvećeg ka najmanjem.

Glava 3

Analiza glavnih komponenti

3.1 Uvod

Analiza glavnih komponenti predstavlja statističku analizu redukcije dimenzionalnosti skupa podataka, koji sadrže veliki broj međusobno povezanih promenljivih, tako da bude obuhvaćena što veća količina varijanse podataka. To se postiže izračunavanjem novog skupa nekorelisanih promenljivih, zvanih glavne komponente (*principal component* - PC) tako da prvih nekoliko glavnih komponenti obuhvati najveći deo varijanse sadržane u originalnim promenljivima.

Neka je $\boldsymbol{x}$ vektor koji sadrži p slučajnih promenljivih i neka je od interesa struktura varijanse ovih promenljivih i njihova kovarijansa ili korelacija. Sem u slučaju kada je p mali broj, komplikovano je posmatrati svih p varijansi i $\frac{p*(p-1)}{2}$ kovarijansi (korelacija). Drugi pristup je pronaći nekoliko ($\ll p$) izvedenih promenljvih koje očuvavaju najveći deo informacije sadržan u tim varijansama i kovarijansama (korelacijama).

U analizi glavnih komponenti, pimarnu ulogu u određivanju glavnih komponenti ima varijansa, ali se ne zanemaruju i kovarijansa i korelacija. Prva glavna komponenta predstavlja linearnu funkciju $\boldsymbol{a}_1'\boldsymbol{x}$ definisanu kao:

$$\boldsymbol{a}_1'\boldsymbol{x} = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p\,,$$

koja ima maksimalnu varijansu, gde su $x_1, x_2, \dots, x_p$ elementi vektora $\boldsymbol{x}$, a $a_{11}, a_{12}, \dots, a_{1p}$ su koeficijenti. Sledeća glavna komponenta je nova linearna funkcija $\boldsymbol{a}_2'\boldsymbol{x}$ koja je nekorelisana sa $\boldsymbol{a}_1'\boldsymbol{x}$ i ima maksimalnu varijansu, itd. Dakle, ideja glavnih komponenti jeste pronalaženje k-te linearne funkcije $\boldsymbol{a}_k'\boldsymbol{x}$, koja je nekorelisana sa prethodnim funkcijama $\boldsymbol{a}_1'\boldsymbol{x}, \boldsymbol{a}_2'\boldsymbol{x}, \dots, \boldsymbol{a}_{k-1}'\boldsymbol{x}$ i koja ima makismalnu varijansu. Iako se teoretski može pronaći p glavnih komponenti, njih će biti mnogo manje jer će putem prvih m ($m \ll p$) komponenti biti opisana ukupna varijansa [1].

3.2 Pronalaženje glavnih komponenti

Nakon definisanja glavnih komponenti, potrebno je pokazati kako se do njih dolazi. Veliku pomoć pruža kovarijansna matrica $\boldsymbol{\Sigma}$, čiji elementi na dijagonali su varijanse (disperzije) elemenata vektora $\boldsymbol{x}$, $\sigma_i^2, i = 1, \dots, p$, a vandijagonalni elementi su kovarijanse $\sigma_{ij},\ i \neq j, i = 1, \dots, p,\ j = 1, \dots, p$ između elemenata vektora. U realnom slučaju, matrica $\boldsymbol{\Sigma}$ je nepoznata pa se ona dobija računanjem uzoračke kovarijansne matrice $\boldsymbol{S}$, o čemu će biti reči u uzoračkom modelu.

Lako se pokazuje da kod glavnih komponenti, definisanih kao $z_k = \boldsymbol{a}_k'\boldsymbol{x}$, $\boldsymbol{a}_k$ predstavja karkateristični vektor matrice $\boldsymbol{\Sigma}$ kojem odgovara najveći karakterističnih korena matrice $\boldsymbol{\Sigma}$. Šta više, ukoliko se izabere da je $\boldsymbol{a}_k$ vektor dužine 1 ($\boldsymbol{a}_k'\boldsymbol{a}_k = 1$), tada je varijansa k-tog glavnog vektora jednaka k-tom karakterističnom korenu, odnosno $var(z_k) = \lambda_k$.

U nastavku je prikazan postupak kako se dobijaju glavne komponente, uz pomoć kovarijansne matrice $\boldsymbol{\Sigma}$. Dakle, prilikom definisanja glavnih komponenti, rečeno je da z_k sadrži maksimalnu varijansu. To bi značilo, za prvu komponentu z_1, da

$$var(z_1) = var(\boldsymbol{a}_1'\boldsymbol{x}) = \boldsymbol{a}_1'\boldsymbol{x}(\boldsymbol{a}_1'\boldsymbol{x})' = \boldsymbol{a}_1'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{a}_1 = \boldsymbol{a}_1'\boldsymbol{\Sigma}\boldsymbol{a}_1 \rightarrow max$$

pod uslovom da je $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$, odnosno da je suma kvadrata elemenata vektora $\boldsymbol{a}_1$ jednaka 1. Ovaj uslov se postavlja jer se gornji maksimum ne postiže za konačni vektor $\boldsymbol{a}_1$. Može se postaviti i uslov $\boldsymbol{a}_1'\boldsymbol{a}_1 = const$, ali ovo bi dovelo do problema u optimizaciji jer bi se optimizacijom dobila vrednost koja je drugačija od prave vrednosti glavne komponente.

Standardna tehnika u izračunavanju optimizacionog problema jeste Langranžov model. Iz optimizacionog problema

$$\boldsymbol{a}_1'\boldsymbol{\Sigma}\boldsymbol{a}_1 \rightarrow max$$

uslov: $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$,

izračunava se Langražijan $\boldsymbol{\mathcal{L}}$ na sledeći način:

$$\boldsymbol{\mathcal{L}}(\boldsymbol{a}_1; \lambda) = \boldsymbol{a}_1'\boldsymbol{\Sigma}\boldsymbol{a}_1 - \lambda\,(\boldsymbol{a}_1'\boldsymbol{a}_1 - 1).$$

Nalaženjem parcijalnog izvoda po $\boldsymbol{a}_1$, sledi

$$\frac{\partial \boldsymbol{\mathcal{L}}(\boldsymbol{a}_1; \lambda)}{\partial \boldsymbol{a}_1} = 2\boldsymbol{\Sigma}\boldsymbol{a}_1 - 2\lambda\boldsymbol{a}_1 = \boldsymbol{0}\,,$$

odnosno

$$\boldsymbol{\Sigma}\boldsymbol{a}_1 = \lambda\boldsymbol{a}_1.$$

Poznato je iz linearne algebre da je $\boldsymbol{a}_1$ karakteristični vektor matrice $\boldsymbol{\Sigma}$, a λ njegov karakteristični koren. Množenjem jednačine, sa desne strane, sa $\boldsymbol{a}_1'$, dobija se

$$\boldsymbol{a}_1'\boldsymbol{\Sigma}\boldsymbol{a}_1 = var(z_1) = \lambda\boldsymbol{a}_1'\boldsymbol{a}_1 = \lambda = \lambda_1,$$

pa sledi da je λ_1 najveći karakteristični koren matrice $\boldsymbol{\Sigma}$. U opšetm slučaju, može se pokazati matematičkom indukcijom da k-toj glavnoj komponenti odgovara karakteristični vektor $\boldsymbol{a}_k$ čija je karakteristični koren λ_k k-ti po veličini.

Da bi se odredila druga glavna komponenta, koristi se drugi uslov koji se postavlja za glavne komponente - da su međusobno nekorelisane, odnosno da važi $cov(z_1, z_2) = 0$. Zamenjivanjem $z_k = \boldsymbol{a}_k'\boldsymbol{x}$ u $cov(z_i, z_j) = 0$, dobija se

$$cov(z_1, z_2) = cov(\boldsymbol{a}_1'\boldsymbol{x}\,, \boldsymbol{a}_2'\boldsymbol{x}\,) = \boldsymbol{a}_1'\boldsymbol{\Sigma}\boldsymbol{a}_2 = \boldsymbol{a}_2'\boldsymbol{\Sigma}\boldsymbol{a}_1 = \boldsymbol{a}_2'\lambda_1\boldsymbol{a}_1 = \lambda_1\boldsymbol{a}_2'\boldsymbol{a}_1 = \lambda_1\boldsymbol{a}_1'\boldsymbol{a}_2 = 0\,.$$

Zbog toga, svaka od jednačina

$$\begin{aligned}\boldsymbol{a}_1'\boldsymbol{\Sigma}\boldsymbol{a}_2 &= 0,\\ \boldsymbol{a}_2'\boldsymbol{\Sigma}\boldsymbol{a}_1 &= 0,\\ \boldsymbol{a}_1'\boldsymbol{a}_2 &= 0,\\ \boldsymbol{a}_2'\boldsymbol{a}_1 &= 0.\end{aligned}$$

se može koristiti da se opiše nekorelisanost glavnih komponenti z_1 i z_2. Koristeći zadnju jednačinu zajedno sa uslovom da je $\boldsymbol{a}_2'\boldsymbol{a}_2 = 1$, dobija se još jedan optimizacioni problem:

$$\boldsymbol{a}_2'\boldsymbol{\Sigma}\boldsymbol{a}_2 \to max$$

$$\begin{aligned}\text{uslov: } \boldsymbol{a}_2'\boldsymbol{a}_1 &= 0.\\ \boldsymbol{a}_2'\boldsymbol{a}_2 &= 1\,.\end{aligned}$$

Lagražijana $\boldsymbol{\mathcal{L}}$ za ovaj problem je

$$\boldsymbol{\mathcal{L}}(\boldsymbol{a}_1, \boldsymbol{a}_2; \lambda) = \boldsymbol{a}_2'\boldsymbol{\Sigma}\boldsymbol{a}_2 - \lambda\,(\boldsymbol{a}_2'\boldsymbol{a}_2 - 1) - \phi\boldsymbol{a}_2'\boldsymbol{a}_1.$$

Diferenciranjem po $\boldsymbol{a}_1'$, dobija se jednačina

$$\frac{\partial\boldsymbol{\mathcal{L}}(\boldsymbol{a}_1, \boldsymbol{a}_2; \lambda)}{\partial\boldsymbol{a}_2} = \boldsymbol{\Sigma}\boldsymbol{a}_2 - \lambda\boldsymbol{a}_2 - \phi\boldsymbol{a}_1 = \boldsymbol{0}.$$

Množenjem jednačine, sa leve strane, sa $\boldsymbol{a}_1'$, dobija se

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

Koristeći gore navedene uslove, prva dva člana jednačine su jednaka nuli, pa sledi da je $\phi = 0$. Odatle važi

$$\boldsymbol{\Sigma}\boldsymbol{a}_2 - \lambda\boldsymbol{a}_2 - \phi\boldsymbol{a}_1'\boldsymbol{a}_1 = \boldsymbol{0} \to \boldsymbol{\Sigma}\boldsymbol{a}_2 - \lambda\boldsymbol{a}_2 = \boldsymbol{0} \to \boldsymbol{\Sigma}\boldsymbol{a}_2 = \lambda\boldsymbol{a}_2\,.$$

Dakle, λ je karakteristični koren matrice $\boldsymbol{\Sigma}$ i $\boldsymbol{a}_2$ je odgovarajući karakteristični vektor.

Kao kod z_1, tako se i ovde dobija da je $\lambda = \lambda_2 = \boldsymbol{a}_2'\boldsymbol{\Sigma}\boldsymbol{a}_2$ (koje je maksimalno), odnosno λ_2 je drugi po veličini karakteristični koren matrice $\boldsymbol{\Sigma}$. Nemoguće je dobiti da je vrednost $\lambda_1 = \lambda_2$, jer u tom slučaju bi važilo da je $\boldsymbol{a}_1 = \boldsymbol{a}_2$, što bi bilo kontradiktorno sa uslovom $\boldsymbol{a}_1'\boldsymbol{a}_2 = 0$, jer bi važilo da je $\boldsymbol{a}_1'\boldsymbol{a}_1 = 0$ ili $\boldsymbol{a}_2'\boldsymbol{a}_2 = 0$.

Ponavljanjem postupka, ekstrahovali bi se svih p karakterističnih korena $\lambda_1, \lambda_2, \dots, \lambda_p$ i vektora $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p$. Svaki novi karakteristični koren bi bio manji od prethodnog, odnosno bili bi poređani u opadajućem nizu [1].

3.3 Kratki istorijat

Najraniji spisi o tehnici, danas poznatoj kao analizi glavnih komponenti, dali su Pearson (1901) i Hotelling (1933) [7][8]. Njihovi spisi imaju različiti pristup. Hotelling-ov rad se zasniva na standardnim algebarskim izvođenjima, dok se Pearson-ov rad odnosi na pronalaženje linija i ravni koje najbolje fituju skup tačaka p-dimenzionalnog prostora.

Pearson-ov komentar, dat 50 godine pre nastanka kompjutera, u vezi sa izračunavanjem komponenti je interesantan i za današnje naučnike. On je tvrdio da se njegove metode mogu lako primeniti u numeričkim problemima i da, iako izračunavanja postaju komplikovana za četiri i više promenljivih, ona ipak mogu da se izvedu. U njegovim spisima geometrijski optimizacioni problemi su spadali pod glavnim komponentama.

Hotelling-ov pristup je bio sličan pristupu faktorske analize, ali se bitno razlikovao u osobinama od same analize. Njegova ideja je bila da postoji manji fundamentalni skup nezavisnih promenljivih koje određuju vrednosti originalnih p promenljivih. Iako je ime tih promenljivih bio ''faktor'' (iz psihološke literature), on je dao alterantivno ime ''komponente'', da ne bi došlo do zabune sa drugim značenjim reči ''faktor'' (iz matematike). Hotelling je svoje komponente definisao tako da maksimizuju određenu varijansu originalnih promenljivih i zato im dodao epitet ''glavne''.

Način dobijanja glavnih komponenti koje je prikazano u prethodnoj delu, putem Langražijana i pronalaženjem karakterističnih korena i vektora, se bitno razlikuje od metode koju je dao Hotelling, i to u tri pogleda:

- prvo, Hotelling je koristio korelacionu matricu,
- drugo, originalne promenljive je posmatrao kao linearne funkcije komponenti umesto obrnuto, da komponente posmatra kao linearnu funkciju promenljivih,
- treće, nije koristio matričnu notaciju.

Nakon izvođenja, Hotelling je pokazao način kako da se izračunaju glavne komponente, koristeći power metodu. Takođe je dao drugačiju geometrijsku intepretaciju u odnosu na Pearson-ovu, u vidu elipsoida sa konstantnom verovatnoćom za promenljive sa višedimenzionalnom normalnom raspodelom.

Postoje još mnoge razlike u izvođenju i notaciji između njihovih i standardnih pristupa analizi glavnih komponenti, ali se smatra da su oni najzaslužniji matematičari za otkriće analize glavnih komponenti.

3.4 Populacioni model analize glavnih komponenti

U ovom delu je diskutovano o matematičkim i statističkim svojstvima glavnih komponenti, koristeći poznatu populacionu kovarijansnu (ili korelacionu) matricu. Zapravo, populacioni model predstavlja teorijski model glavnih komponenti, a nakon ovog dela biće diskutovano o modelu na osnovu uzorka.

Neka je $\boldsymbol{z}$ vektor koji se sastoji od p glavnih komponenti z_k. U matričnoj notaciji, vektor $\boldsymbol{z}$ je jednak

$$\boldsymbol{z} = \boldsymbol{A}'\boldsymbol{x}\,,$$

gde je $\boldsymbol{A}$ ortogonalna matrica čije kolone čine ortonormirani vektori $\boldsymbol{a}_k$ (karakteristični vektori od $\boldsymbol{\Sigma}$). Dakle, glavne komponente su definisane kao ortonormirana linearna transformacija vektora $\boldsymbol{x}$. Treba napomenuti da neki autori vektor glavnih komponenti predstavljaju kao suma $\boldsymbol{z} = \boldsymbol{A}'\boldsymbol{x} + \boldsymbol{e}$, gde je $\boldsymbol{e}$ vektor reziduala, čiji elementi su približno jednaki nuli. Iz praktičnih razloga, u ovom radu neće biti korišćen dati pristup.

Jednakosti $\boldsymbol{\Sigma a}_k = \lambda_k \boldsymbol{a}_k,\ k = 1,2,\dots,p$, u matričnom zapisu glase:

$$\boldsymbol{\Sigma A} = \boldsymbol{A\Lambda},$$

gde je $\boldsymbol{\Lambda}$ dijagonalna matrica na čijoj dijagonali se nalazi p karakterističnih korena $\lambda_k,\ k = 1,2,\dots,p$. S obzirom da je $\boldsymbol{A}$ ortogonalna matrica, važi jednakost

$$\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}\,.$$

Koristeći ovu osobinu lako se mogu izraziti matrice $\boldsymbol{\Sigma}$ i $\boldsymbol{\Lambda}$:

$$\boldsymbol{\Sigma} = \boldsymbol{A\Lambda A}', \qquad \boldsymbol{\Lambda} = \boldsymbol{A}'\boldsymbol{\Sigma A}\,.$$

Teorema 1. Neka je data je ortonormirana linearna transformacija

$$\boldsymbol{y} = \boldsymbol{B}'\boldsymbol{x},$$

gde je $\boldsymbol{y}$ vektor koji sadrži q elemenata i $\boldsymbol{B}'$ matrica dimenzije $q \times p$, za bilo koji ceo broj q, $1 \le q \le p$. Neka je $\boldsymbol{\Sigma}_y = \boldsymbol{B}'\boldsymbol{\Sigma B}$ kovarijansna matrica vektora $\boldsymbol{y}$. Tada važi:

a) $trag(\boldsymbol{\Sigma}_y)$ ima maksimalnu vrednost ukoliko je $\boldsymbol{B} = \boldsymbol{A}_q$, gde matrica $\boldsymbol{A}_q$ sadrži prvih q kolona od matrice $\boldsymbol{A}$;
b) $trag(\boldsymbol{\Sigma}_y)$ ima minimalnu vrednost ukoliko je $\boldsymbol{B} = \boldsymbol{A}_q^*$, gde matrica $\boldsymbol{A}_q^*$ sadrži zadnjih q kolona od matrice $\boldsymbol{A}$.

Dokaz. Pokazuje se pod a). Neka je β_k k-ta kolona matrice $\boldsymbol{B}$ jednaka linearnoj kombinaciji kolona matrice $\boldsymbol{A}$ u p-dimenzionalnom prostoru:

$$\beta_k = \sum_{j=1}^{p} c_{jk}\boldsymbol{a}_j\,, \qquad k = 1,2,\dots,q\,,$$

gde je $c_{jk},\ j = 1,2,\dots p, k = 1,2,\dots,q$ konstanta. Otuda je $\boldsymbol{B} = \boldsymbol{AC}$, gde je $\boldsymbol{C}$ matrica dimenzije $p \times q$. Tada važi:

$$\boldsymbol{B'\Sigma B} = \boldsymbol{C'A'\Sigma AC} = \boldsymbol{C'\Lambda C} = \textstyle\sum_{j=1}^{p} \lambda_j \boldsymbol{c}_j \boldsymbol{c}_j',$$

gde je $\boldsymbol{c}_j'$ j-ta vrsta matrice $\boldsymbol{C}$.

Dalje važi:

$$tr(\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B}) = \sum_{j=1}^{p} \lambda_j tr(\boldsymbol{c}_j \boldsymbol{c}_j') = \sum_{j=1}^{p} \lambda_j tr(\boldsymbol{c}_j' \boldsymbol{c}_j) = \sum_{j=1}^{p} \lambda_j \boldsymbol{c}_j \boldsymbol{c}_j' = \sum_{j=1}^{p} \sum_{k=1}^{q} \lambda_j c_{jk}^2 .$$

Sada je

$$\boldsymbol{C} = \boldsymbol{A}'\boldsymbol{B},$$

pa je

$$\boldsymbol{C}'\boldsymbol{C} = \boldsymbol{B}'\boldsymbol{A}\boldsymbol{A}'\boldsymbol{B} = \boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}_q ,$$

s obzirom da je $\boldsymbol{A}$ ortogonalna matrica, a i kolone matrice $\boldsymbol{B}$ su ortonormirane. Stoga

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

i kolone matrice $\boldsymbol{C}$ su takođe ortonormirane. Matrica $\boldsymbol{C}$ se može posmatrati kao kao prvih q kolona $(p \times p)$ ortogonalne matrice $\boldsymbol{D}$. S obzirom da su vrste matrice $\boldsymbol{D}$ ortonormirane, važi $\boldsymbol{d}_j' \boldsymbol{d}_j = 1,\ j = 1,2,\dots,p$. Kako matrica $\boldsymbol{C}$ sadrže prvih q elemenata vrsta matrice $\boldsymbol{D}$, proizilazi da važi $\boldsymbol{c}_j \boldsymbol{c}_j' < 1,\ j = 1,2,\dots,p$ odnosno:

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

Posmatranjem $\sum_{j=1}^{p} \sum_{k=1}^{q} \lambda_j c_{jk}^2$, sada je $\sum_{k=1}^{q} c_{jk}^2$ koeficijent karakterističnog vektora λ_j. S obziroma da važi $\lambda_1 > \lambda_2 > \cdots > \lambda_p$, vrednost $\sum_{j=1}^{p} (\sum_{k=1}^{q} c_{jk}^2) \lambda_j = q$ se maksimizuje ukoliko se pronađe skup c_{jk} za koji važi:

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

Ali, ukoliko je $\boldsymbol{B}' = \boldsymbol{A}_q'$, tada

$$c_{jk} = \begin{cases} 1, & j \le 1 = k \le q \\ 0, & ostalo' \end{cases}$$

zadovoljava gore navedenu vrednost $\sum_{k=1}^{q} c_{jk}^2$. Otud $tr(\boldsymbol{\Sigma}_y)$ postiže maksimalnu vrednost kad je $\boldsymbol{B}' = \boldsymbol{A}_q'$.

Pod b) se pokazuje na sličan način.

■

Ova teorema daje značajnost prvim q glavnim komponentama jer one opisuju najveće rasipanje (varijansu), dok zadnjih nekoliko glavnih komponenti ima veoma malu varijansu i zato, u određenom značenju, nisu korisne za model. Međutim, one mogu da pomognu da se otkriju linearne povezanosti između elemenata vektora $\boldsymbol{x}$ i mogu biti korišćene u regresiji, u izboru podskupa promenljvih vektora $\boldsymbol{x}$ i u otkrivanju autlajera.

Teorema 2. Kovarijansna matrica $\boldsymbol{\Sigma}$ se može razložiti na sledeći način:

$$\boldsymbol{\Sigma} = \lambda_1 \boldsymbol{a}_1 \boldsymbol{a}_1' + \lambda_2 \boldsymbol{a}_2 \boldsymbol{a}_2' + \cdots + \lambda_p \boldsymbol{a}_p \boldsymbol{a}_p' \, .$$

Ova dekompozicija se naziva spektralna dekompozicija.

Dokaz.

S obzirom da je poznata jednakost $\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Lambda}\boldsymbol{A}'$, ona napisana u vektorskoj notaciji glasi

$$\boldsymbol{\Sigma} = \sum_{k=1}^{p} \lambda_k \boldsymbol{a}_k \boldsymbol{a}_k' \, .$$

■

Jasno je iz formule $\boldsymbol{\Sigma} = \sum_{k=1}^{p} \lambda_k \boldsymbol{a}_k \boldsymbol{a}_k'$ da se kovarijansna (ili korelaciona) matrica može konstruisati koristeći koeficijenie i varijanse prvih r glavnih komponenti, gde r predstavlja rang kovarijansne matrice. Elementi k-tog člana $\lambda_k \boldsymbol{a}_k \boldsymbol{a}_k'$ gornje sume postoju sve manje kako raste vrednost k, jer vektori $\boldsymbol{a}_k$ ''ostaju iste veličine'' ($\boldsymbol{a}_k \boldsymbol{a}_k' = 1$), a λ_k opadaju sa porastom k. Teorema 1. govori da glavne komponente uspešno opisuju $trag(\boldsymbol{\Sigma})$, ali se intuitivno javlja da one dobro opisuju i vandijagonalne elemente matrice $\boldsymbol{\Sigma}$ [1].

3.5 Dobijanje glavnih komponenti pomoću korelacione matrice

U prethodnom delu, glavne komponente su bazirane na osnovu karakterističnih korena i vektora dobijenih iz kovarijansne matrice $\boldsymbol{\Sigma}$. Mnogo češće se glavni faktori definišu kao

$$\boldsymbol{z} = \boldsymbol{A}' \boldsymbol{x}^*,$$

gde je $\boldsymbol{A}$ matrica čije kolone predstavljaju karakteristični vektori korelacione matrice $\boldsymbol{R}$, a $\boldsymbol{x}^*$ je standardizovani vektor, tj.

$$\boldsymbol{x}^* = \left[\frac{x_j}{\sqrt{\sigma_{jj}}} \right]_{1 \times p} .$$

Dakle, korelaciona matrica vektora $\boldsymbol{x}$ je jednak kovarijasnoj matrici vektora $\boldsymbol{x}^*$.

Umesto kovarijansne i korelacione matrice, treći način kako se mogu izračunati glavne komponente jeste definisanjem vektora $\boldsymbol{x}^*$ kao

$$\boldsymbol{x}^* = \left[\frac{x_j}{w_j}\right]_{1\times p}.$$

Za slučaj da je $w_j = \sqrt{\sigma_{jj}}$ ili $w_j = 1$, dobija se korelaciona, odnosno kovarijansna matrica vektora $\boldsymbol{x}^*$. Međutim, mnogi autori tvrde da izbor drugih vrednosti za w_j je bolje primenljiv u modelu.

Sva svojstva vezana za kovarijansnu matricu, koja su pokazana u prethodnom radu, važe i za korelacionu matricu. Čini se da se glavne komponente za korelacionu matricu mogu dobiti veoma lako iz njene kovarijansne matrice, s obzirom da se $\boldsymbol{x}^*$ može lako dobiti primenom jednostavne transformacije nad $\boldsymbol{x}$. Međutim, to nije tačno. Karakteristični koreni i vektori korelacione matrice nemaju jednostavnu vezu sa karakterističnim korenima i vektorima kovarijansne matrice. Specijalno, ukoliko su glavne komponente dobijene iz korelacione matrice izraze preko vektora $\boldsymbol{x}$ (putem transformacije iz $\boldsymbol{x}^*$ u $\boldsymbol{x}$), tada one nisu iste kao komponente dobijeni iz kovarijansne matrice $\boldsymbol{\Sigma}$, sem u vrlo specijalnim slučajevima. Ovo se može objanisiti činjenicom da su glavne komponente invarijantne u odnosu na ortogonalne transformacije vektora $\boldsymbol{x}$, dok nisu invarijantne u odnosu na neke druge transformaicje. Transformacija iz $\boldsymbol{x}$ u $\boldsymbol{x}^*$ nije ortogonalna. Dakle, zaključak je da glavne komponente dobijene iz kovarijansne i korelacione matrice ne daju ekvivaletne informacije, niti se mogu izraziti direktno jedna iz druge.

Jedan od argumenata za korišćenje korelacione umesto kovarijansne jeste to da se podaci kod korelacione matrice mogu međusobno upoređivati, što nije moguće kod kovarijasne matrice. Problem analize glavnih komponenti zasnovane na kovarijasnoj matrici je da su glavne komponente osetljive na jedinice merenja, koje se koriste za elemente vektora $\boldsymbol{x}$. Ukoliko postoji velika razlika između varijansi elemenata vektora $\boldsymbol{x}$, tada promenljive čije su varijanse najveće imaju tedenciju da dominiraju prvim glavnim komponentama. Problem neće nastati ukoliko svi elementi vektora $\boldsymbol{x}$ imaju iste jedinice mere. npr. u cm. Problem nastaje u praksa, kada se dogodi da sve promenljive imaju različite jedinice mere, npr. jedna promenljiva predstavlja temperaturu, druga težinu, treća dužinu,... U nastavku, prikazan je jedan primer koji govori o uticaju jedinice mere na glavne komponente.

Neka su date dve promenljive: x_1 i x_2. Neka promenljvia x_1 predstavlja dužinu, koja može biti izražena u centimetrima ili milimetrima, dok promenljiva x_2 predstavlja težinu, izraženu u gramima. Nakon sprovedenog istraživanja, dobijaju se dve kovarijansne matrice $\boldsymbol{\Sigma}_1$ i $\boldsymbol{\Sigma}_2$:

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 80 & 44 \\ 44 & 80 \end{bmatrix} \text{ i } \boldsymbol{\Sigma}_2 = \begin{bmatrix} 8000 & 440 \\ 440 & 80 \end{bmatrix}.$$

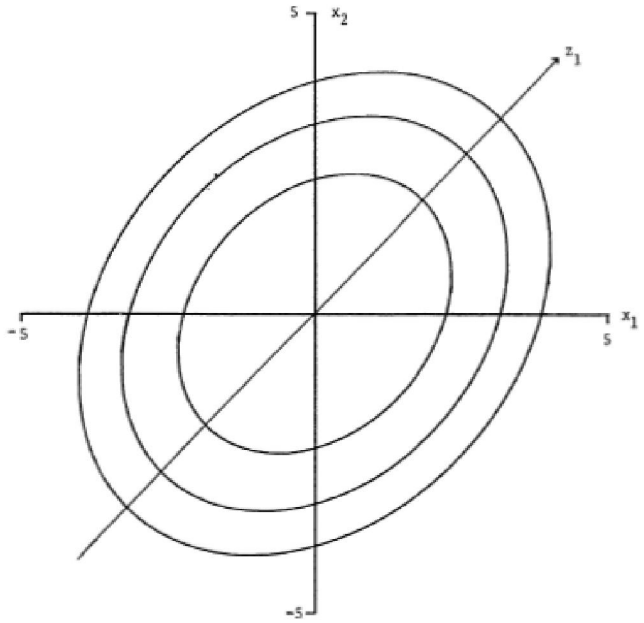
Izračunavanjem, dobijaju se prve glavne komponente date matrice:

$$z_1 = 0.707x_1 + 0.707x_2 \text{ za } \boldsymbol{\Sigma}_1 \text{ i } z_2 = 0.998x_1 + 0.055x_2 \text{ za } \boldsymbol{\Sigma}_2.$$

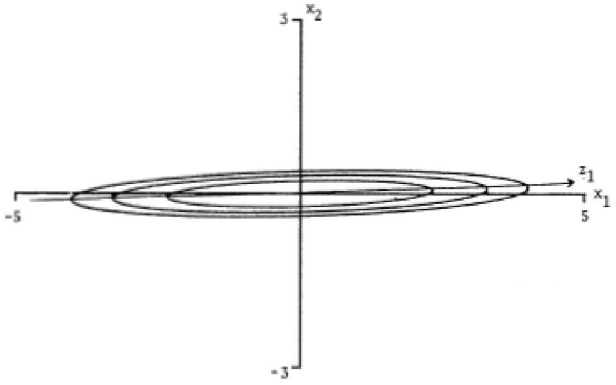
Mala promena (iz cm u mm) u prvoj promenljivoj dovodi do toga da se od glavne komponente (z_1) koja daje iste težine promenljivima (x_1 i x_2), dolazi do glavne komponente

(z_2) u kojoj skoro potpuno dominira promenljiva x_1. Šta više, prva glavna komponenta z_1 opisuje 77.5% totalne varijanse za matricu $\mathbf{\Sigma}_1$, a z_2 opisuje 99.3% totalne varijanse za matricu $\mathbf{\Sigma}_2$.

Geometrijska prikaz kontura elipsoida i dobijenih glavnih komponenti, dat je je na slici 3.1 i 3.2. Na slici 3.1 prikazane su konture konstantne verovatnoće, i ono što se može zaključiti jeste da se podaci rasipaju u dva pravca, i zato prva glavna komponenta (z_1) objašnjava 77.5% rasipanja. Na slici 3.2 podaci su drugačije raspoređeni, tj. po jednoj promenljivoj se rasipaju, dok po drugoj skoro i da nema rasipanja. Zato, za ovaj elipsoid konstante verovatnoće, dobijena prva galvna komponenta objašnjava skoro 100% rasipanja podataka (tačnije 99.3%). Zaključak kojii proističe jeste da kod kovarijansne matrice, jedinica mere ima veliki uticaj na rezultate dobijenih glavnih komponenti.



Slika 3.1: Glavna komponenta z_1 na osnovu matrice $\mathbf{\Sigma}_1$



Slika 3.2: Glavna komponenta z_1 na osnovu matrice $\mathbf{\Sigma}_2$

Ovim primerom je pokazano da nije valjano korišćenje kovarijansne matrice za dobijanje glavnih komponenti, kada su promenljive različitog tipa (jedinice). Takođe, javlja se i problem intepretacija glavnih komponenti. Zato korišćenje korelacione matrice daje mnogo bolje rezultate [1].

3.6 Uzorački model analize glavnih komponenti

Kao što je rečeno u uvodnom delu, populacioni model predstavlja teorijsku osnovu analize glavnih komponenti. Uzorački model predstavlja analizu na osnovu reprezentativnog uzorka.

Neka je, kao i ranije, $\boldsymbol{z}$ vektor glavnih komponenti. Element z_k ovkog vektora je definisana kao

$$z_k = \boldsymbol{a}_k'\boldsymbol{x} = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p,$$

gde su $(x_1, x_2, \ldots, x_p)$ slučajne promenljive.

Za svaku slučajnu promenljivu, x_k, $k = 1,2,\ldots p$ posmatra se uzorak obima n posmatranja $(x_{k1}, x_{k2}, \ldots, x_{kn})$, $k = 1,2,\ldots p$, i dobija se matrica podataka $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

Tada je

$$z_{k1} = a_{11}x_{11} + a_{12}x_{21} + \cdots + a_{1p}x_{p1}\ ,$$

odnosno u matričnom zapisu

$$\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1q} \\ z_{21} & z_{22} & & z_{2q} \\ \vdots & & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nq} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & & a_{2q} \\ \vdots & & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{bmatrix},$$

tj.

$$\boldsymbol{Z} = \boldsymbol{XA}.$$

Vrlo je bitno napomenuti da podaci koji se dobiju u matrici $\boldsymbol{X}$ nisu centrirani. Zato, pre bilo kakve analize, bitno je centrirati matricu na sledeći način:

$$\tilde{x}_{ij} = x_{ij} - \sum_{i=1}^{n} \frac{x_{ij}}{n}.$$

Time se dobija centrirana matrica $\widetilde{\boldsymbol{X}}$. Međutim, u nastavku, radi lakše notacije, koristiće se oznaka $\boldsymbol{X}$ za matricu centriranih podatka, umesto oznake $\widetilde{\boldsymbol{X}}$. Dalje, deljenjem matričnog proizvoda manje dimenzije $\boldsymbol{X'X}$ matrice $\boldsymbol{X}$ sa n ($\mathbf{1'1}$ u matričnom obliku) dobija se uzoračka kovarijasna matrica $\boldsymbol{S}$, tj.

$$\boldsymbol{S} = \boldsymbol{X'X}/\mathbf{1'1}\,.$$

3.7 Broj komponenti

Neka je q broj glavnih komponenti koje su dobijene u prethodnom delu. Po pravili, q je nepoznata vrednost koju istraživač određuje tako da bude zadovoljan postotkom objašnjenjenog rasipanja (varijacije). Naravno, podrazumeva se da svi istraživači žele da što manji broj komponenti objasni što veći iznos totalne varijanse. Dakle, zaključuje se da je broj komponenti q direktno povezano sa totalnom varijansom i o tome se u nastavku diskutuje.

Totalna varijansa komponenti je data kao $trag(\boldsymbol{\Lambda}_q)$, odnosno kao suma q karakterističnih korena:

$$trag(\boldsymbol{\Lambda}_q) = \lambda_1 + \lambda_2 + \cdots + \lambda_q.$$

Ukoliko bi važilo da je $q = p$, tada bi ukupna varijansa svih promenljivih bila jednaka ukupnoj varijansi komponenti, jer važi:

$$\boldsymbol{S} = \boldsymbol{U\Lambda U'} \rightarrow trag(\boldsymbol{S}) = trag(\boldsymbol{\Lambda}).$$

Međutim, ukoliko je $q < p$, u tom slučaju suma $p - q$ karakterističnih korena bi činila rezidualnu varijansu, odnosno varijansu neobjašnjenu od strane komponenti. Zaključak je da se totalna varijansa promenljivih može predstaviti kao suma objašnjene varijanse (putem q komponenti) i rezidualne varijanse, odnosno:

$$trag(\boldsymbol{S}) = trag(\boldsymbol{\Lambda}_q) + \lambda_{q+1} + \lambda_{q+2} + \cdots + \lambda_p\,.$$

Najbolji način u određivanju broja glavnih komponenti jeste izračunavanje kumulativne varijanse. Dakle, izračunavanjem novih glavnih komponenti, ukupna kumulativna varijansa se povećava i bitno je zaustaviti se kada je taj broj dovoljno veliki, npr. 90% ili 95%. Koliko daleko se može ići zavisi od prirode podataka, odnosno od stepena kolinearnosti i redundantnosti (ponavljanje istih podataka) u njima. Procenat kumulativnog varijanse se dobija na sledeći način:

$$trag(\boldsymbol{S}) = trag(\boldsymbol{\Lambda}_q) + \lambda_{q+1} + \lambda_{q+2} + \cdots + \lambda_p \;\; / \, trag(\boldsymbol{S})^{-1} * 100$$

$$100\%\,(totalna\ var.) = \%\ var.glav.komp. + \%\ rezidualna\ var.$$

Karakteristični koreni u okviru varijanse glavnih komponenti treba da budu značajno različiti od nule. Rezidualni karakteristični koreni treba da budu mali i približno jednaki [3][11].

3.8 Korelacija između faktora i glavnih komponenti

U nastavku će biti određeni koeficijenti korelacije između originalnih promenljivih i glavnih komponenti. Za matricu podataka $\boldsymbol{X}$ i matricu glavnih komponenti $\boldsymbol{Z}$ važi:

$$var(\boldsymbol{X}) = \boldsymbol{\Sigma} \quad \text{i} \quad var(\boldsymbol{Z}) = \boldsymbol{\Lambda}.$$

Dalje, dobija se vrednost kovarijanse između matrice $\boldsymbol{X}$ i $\boldsymbol{Z}$:

$$cov(\boldsymbol{X}, \boldsymbol{Z}) = cov(\boldsymbol{X}, \boldsymbol{AX}) = \boldsymbol{\Sigma A'} = (\mathbf{A'\Lambda A})\mathbf{A'} = \mathbf{A'\Lambda} = [\boldsymbol{a}_1\lambda_1 \quad \boldsymbol{a}_2\lambda_2 \quad \dots \quad \boldsymbol{a}_p\lambda_p].$$

S obzirom da je cilj izračunavanje koeficijenta korelacije između k-te originalne promenljive i j-te glavne komponente, dat je sledeći izraz

$$\rho_{x_k,z_j} = \frac{cov(x_k, z_j)}{\sqrt{var(x_k) * var(z_j)}} = \frac{a_{jk}\lambda_j}{\sqrt{\lambda_j \sigma_{kk}}} = a_{jk}\frac{\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}, \qquad j, k = 1,2, \dots p,$$

odnosno u matčnom obliku

$$\boldsymbol{\rho}_{x,z} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{AD}^{-1/2},$$

gde je matrica $\boldsymbol{D}$ dijagonalna matrica čiji su elementi varijanse originalnih promenljvih [11].

3.9 Primer

U sledećem primeru će se na osnovu kovarijansne matrice podataka $\boldsymbol{S}$ izračunati glavne komponente matrice $\boldsymbol{X}$, pokazati da je varijansa dobijenih glavnih komponenti jednaka vrednostima odgovarjućih karakterističnih korena i da je kovarijansa između para glavnih komponenti jednaka nuli [11]. Zatim će se izračunati korelaciona matrica $\boldsymbol{R}$ na osnovu koje će se izračunati nove glavne komponente, a zatim izračunati korelaciona matrica između glavnih komponenti (i iz $\boldsymbol{S}$ i iz $\boldsymbol{R}$) i originalnih promenljivih.

Dakle, neka je data realizovana kovarijansna matrica podataka $\boldsymbol{S}$:

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

Na osnovu jednačine

$$|\boldsymbol{S} - \lambda\boldsymbol{I}| = \begin{vmatrix} 4-\lambda & 0 & 0 \\ 0 & 2-\lambda & -2 \\ 0 & -2 & 4-\lambda \end{vmatrix} = (4-\lambda)^2(2-\lambda) - 4(4-\lambda) = 0$$

mogu se izračunati karkteristični koreni matrice $\boldsymbol{S}$. Rešavanjem gore navedene jednačine dobijaju se sledeće vrednosti karakterističnih korena:

$$\lambda_1 = 5.2361,\ \lambda_2 = 4,\ \lambda_3 = 0.7639.$$

Dalje, na osnovu jednakosti

$$\boldsymbol{S}\boldsymbol{a}_i = \lambda_i \boldsymbol{a}_i, \qquad i = 1,2,3.$$

lako se mogu izrčunati karakteristilni vektori. Rešavanjem, dobijaju se sledeće vrednosti:

$$\boldsymbol{a}_1 = \begin{bmatrix} 0 \\ -0.5257 \\ 0.8507 \end{bmatrix}, \qquad \boldsymbol{a}_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{a}_3 = \begin{bmatrix} 0 \\ -0.8507 \\ -0.5257 \end{bmatrix}.$$

Na osnovu izračunatih karakterističnih korena i vektora dobijaju se vrednosti glavnih komponenti matrice $\boldsymbol{X}$:

$$z_1 = \boldsymbol{a}_1'\boldsymbol{x} = -0.5257x_2 + 0.8507x_3,$$

$$z_2 = \boldsymbol{a}_2'\boldsymbol{x} = x_1,$$

$$z_3 = \boldsymbol{a}_3'\boldsymbol{x} = -0.8507x_2 - 0.5257x_3.$$

Varijansa glavnih komponenti je jednaka odgovarajućim karakterističnim korenima kovarijansne matrice. U nastavku, prikazano je da to važi za prvu komponentu:

$$var(z_1) = var(-0.5257x_2 + 0.8507x_3) =$$

$$= (-0.5257)^2 * var(x_2) + (0.8507)^2 * var(x_3) + 2 * (-0.5257) * (0.8507) * cov(x_2, x_3)$$

$$= 0.2764 * 2 + 0.7236 * 4 - 0.8944 * (-2) =$$

$$= 5.2361 = \lambda_1 .$$

Na isti način se pokazuje i za ostale dve glavne komponente.

Nekorelisanost između glavnih komponenti se lako pokazuje:

$$cov(z_1, z_2) = cov(-0.5257x_2 + 0.8507x_3, x_1) =$$

$$= -0.5257 * cov(x_2, x_1) + 0.8507 * cov(x_3, x_1) =$$

$$= -0.5257 * 0 + 0.8507 * 0 = 0.$$

Takođe, na isti način se pokazuje za preostale parove glavnih komponenti.

Korelaciona matrica $\boldsymbol{R}$ matrice $\boldsymbol{X}$ je jednaka:

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -0.7071 \\ 0 & -0.7071 & 1 \end{bmatrix}.$$

Na isti način kao kod kovarijansne matrice, određuju se karakteristični koreni i vektori. Za datu korelacionu matricu $\boldsymbol{R}$, vrednosti karakterističnih korena su:

$$\lambda_1 = 1.7071,\ \lambda_2 = 1,\ \lambda_3 = 0.2929.$$

Karakteristični vektori su jednaki:

$$\boldsymbol{a}_1 = \begin{bmatrix} 0 \\ -0.7071 \\ 0.7071 \end{bmatrix}, \qquad \boldsymbol{a}_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{a}_3 = \begin{bmatrix} 0 \\ -0.7071 \\ -0.7071 \end{bmatrix},$$

pa samim tim se dobijaju nove vrednosti glavnih komponenti matrice $\boldsymbol{X}$:

$$z_1 = \boldsymbol{a}_1'\boldsymbol{x} = -0.7071x_2 + 0.7071x_3,$$

$$z_2 = \boldsymbol{a}_2'\boldsymbol{x} = x_1,$$

$$z_3 = \boldsymbol{a}_3'\boldsymbol{x} = -0.7071x_2 - 0.7071x_3.$$

Na onovu formule za izračunavanje korelacije između i-te originalne promenljive i j-te glavne komponente, lako se može izračunati i korelaciona matrica između originalnih promenljivih i glavnih komponenti. Dakle, korelaciona matrica $\boldsymbol{\rho_{XZ_S}}$ za glavne komponente dobijene na osnovu kovarijansne matrice $\boldsymbol{S}$ je jednaka:

$$\boldsymbol{\rho_{XZ_S}} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2} = \begin{bmatrix} 0 & 1 & 0 \\ -0.8506 & 0 & -0.5258 \\ 0.9733 & 0 & -0.2297 \end{bmatrix}.$$

Na isti način se lako dobija korelaciona matrica $\boldsymbol{\rho_{XZ_R}}$ za glavne komponente ali dobijene na osnovu korelacione matrice $\boldsymbol{R}$:

$$\boldsymbol{\rho_{XZ_R}} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2} = \begin{bmatrix} 0 & 1 & 0 \\ -0.9239 & 0 & -0.3827 \\ 0.9239 & 0 & -0.3827 \end{bmatrix}$$

Glavne komponente		**Kovarijansna matrica**			**Korelaciona matrica**		
		prva	druga	treća	prva	druga	treća
		Koeficijent linearne kombinacije					
Originalne promenljvie	prva	0	1	0	0	1	0
	druga	-0.5257	0	-0.8507	-0.7071	0	-0.7071
	treća	0.8507	0	-0.5257	0.7071	0	-0.7071
Karakteristični koreni		5.2361	4	0.7639	1.7071	1	0.2929
Objašnjena varijansa		52.36%	40%	7.64%	56.90%	33.33%	9.76%
		Korelacija glavnih komponenti i originalnih promenljivih					
Originalne promenljive	prva	0	1	0	0	1	0
	druga	-0.8506	0	-0.5258	-0.9239	0	-0.3827
	treća	0.9733	0	-0.2297	0.9239	0	-0.3827

Tabela 3.1: *Prikaz dobijenih rešenja*

Glava 4

Faktorska analiza

4.1 Uvod

Metoda multivarijacione analize koja se koristi za opisivanje međusobne zavisnosti velikog broja promenljivih uz pomoć manjeg broja osnovnih promenljih jeste (prava) faktorska anliza (true factor analysis). Na samom početku je bitno napomenuti da postoje dva značenja faktorske analize. Ona može da označava oblast statistike (o tome je bilo reči u uvodnom delu) i metode unutar te oblasti (o tome će biti reči u ovom delu). Dakle, u ovom delu će biti detaljno opisana faktorska analiza, metoda kojom se takođe smanjuje broj promenljivih, ali na bitno drugačiji način nego što je kod analize glavnih komponenti. Razlika je u tome što se analiza glavnih komponenti bazira na analizi dijagonalnih elemenata kovarijansne matrice-varijansi, dok su u faktorskoj analizi interesantni vandijagonalni elementi-kovarijanse.

U uvodnom delu master rada spomenuto je da je faktorsku analizu utemeljio Charles Spearmanu (1904). Baveći se korelacijama između različitih testova inteligencije, pokazao je da se sve one mogu iskazati jednostavnijim modelom. Na primer, sprovedeno je istraživanje vezano za pripremljenost dece za školu i dobijena je sledeća korelaciona matrica:

$$\begin{bmatrix} 1 & 0.83 & 0.78 & 0.70 \\ 0.83 & 1 & 0.67 & 0.67 \\ 0.78 & 0.67 & 1 & 0.64 \\ 0.70 & 0.67 & 0.64 & 1 \end{bmatrix},$$

gde su promeljive testovi iz klasike (x_1), francuskog (x_2), engleskog (x_3) i matematike (x_4). Spearman je uočio proporcionalnost između ma koje dve vrste ili kolone matrice, ukoliko se zanemare elementi sa glavne dijagonale. Tako na primer elementi prve i poslednje vrste daju količnik:

$$\frac{0.83}{0.67} = \frac{0.78}{0.64} = 1.2\ .$$

Na osnovu ovakvih dobijenih rezultata, Spearman je predložio smanjivanje dimenzije problema sa $p = 4$ na $p = 1$, tako što će svi testovi (x_i) biti iskazani u sledećem obliku

$$x_i = a_i f + e_i, \qquad i = 1,2,3,4\ ,$$

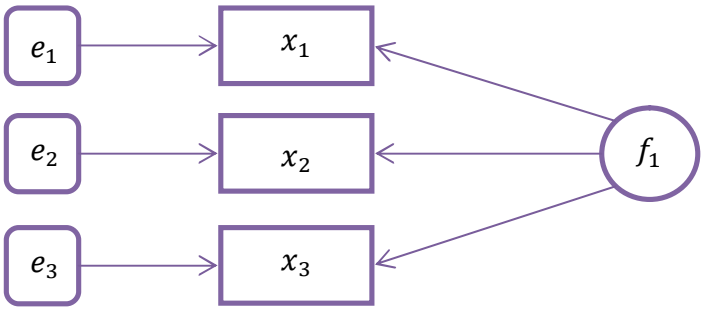
gde f predstavlja zajednički faktor, a_i koeficijent faktorskog opterećenja, a e_i slučajnu grešku (specifični faktor). Uz izvesne pretpostavke, sledi da su elementi u dve vrste korelacione matrice međusobno proporcionalni. Dakle, dobijeni model dobro opisuje korelacionu strukturu podataka. Daljim svojim radom, Spearman razvija svoju dvofaktorsku teoriju testova inteligencije. Zaključio je da se rezultati svakog testa mogu dekomponovati na dva dela. Prvi je zajednički deo za sve testove (f), a drugi specifičan za svaki test (e_i). Zajednički faktor se može definisati kao ''opšta sposobnost'' ili ''inteligencija'' [6].

4.2 Modeli faktorske analize

4.2.1 Razvoj modela

Pretpostavimo da je dovoljan jedan faktor f_1 da objasni sve promenljve unutar vektora $\boldsymbol{x} = (x_1, x_2, x_3)$ (pretpostavlja se da su promenljive centrirane). Dakle, neka koeficijenti a_1, a_2, a_3
zajedno sa faktorom f_1 daju približne vrednosti za x_1, x_2, x_3 respektivno:

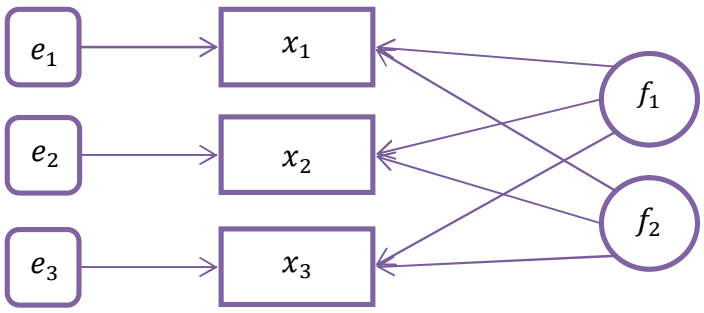
$$x_1 = a_1 f_1 + e_1 ,$$
$$x_2 = a_2 f_1 + e_2 ,$$
$$x_3 = a_3 f_1 + e_3 ,$$



Slika 4.1: *Model sa jednim faktorom*

gde su e_1, e_2, e_3 razlika između aproksimacije i registrovane vrednosti. Ukoliko se proceni da aproksimacija nije dobro objasnila registrovane vrednosti, koristi se dodatni faktor f_2. U ovom slučaju, svaka promenljiva je dobijena uz pomoć dva faktora f_1 i f_2 i odgovarajuće koeficijenta:

$$x_1 = a_{11} f_1 + a_{12} f_2 + e_1 ,$$
$$x_2 = a_{21} f_1 + a_{12} f_2 + e_2 ,$$
$$x_3 = a_{31} f_1 + a_{12} f_2 + e_3 .$$



Slika 4.2: *Model sa dva faktora*

Novi faktori f_i se dodaju sve dok razlika između aproksimacije i registrovane vrednosti ne bude zadovoljavajuće dovoljno mala. U opštem slučaju, ako postoji p promenljivih i m faktora, jedakost glasi

$$x_i = a_{i1}f_1 + a_{i2}f_2 + \cdots + a_{im}f_m + e_i\,,$$

gde je x_i i-ta promenljiva unutar vektora $\boldsymbol{x}$ [3].

4.2.2 Opšti model faktorske analize

Neka je $\boldsymbol{x}$ vektor koji se sastoji od p slučajnih promenljivih $x_1, x_2, \dots, x_p$ sa sredinama $E(x_1) = \mu_1,\ E(x_2) = \mu_2,\ \dots\ ,\ E(x_p) = \mu_p$. Neka se promenljive mogu predstaviti kao linearna funkcija m ($m < p$) hipotetičkih slučajnih promenljivih $f_1, f_2, \dots, f_m$ koje se nazivaju zajednički faktori. Tako, opšti model faktorske analize glasi:

$$\begin{aligned} x_1 &= \mu_1 + a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + e_1 \\ x_2 &= \mu_2 + a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + e_2 \\ &\vdots \\ x_p &= \mu_p + a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + e_p \end{aligned}$$

gde je a_{jk} konstanta koja se naziva faktorski koeficijent, a $e_j,\ j = 1,2,\dots p$ predstavlja grešku. Greška e_j se često naziva i specifični faktor jer je e_j specifično za promenljivu x_j, dok su f_j zajednički za sve promenljive x_j. U matričnoj zapisu, faktorski model glasi

$$\boldsymbol{x} - \boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{e}\,.$$

Na prvi pogled, model faktorske analize liči na višestruku regresiju. Međutim, razlika postoji u broju promenljivih koje su registrovane, jer kod faktorskog modela p odstupanja $x_1 - \mu_1, x_2 - \mu_2, \dots, x_p - \mu_p$ se izražava preko $m + p$ slučajnih promenljivih $f_1, f_2, \dots, f_m$ i $e_1, e_2, \dots, e_p$ koje nisu registrovane, za razliku od višestuke regresije gde su nezavisne promenljive registrovane [11].

Za svaku slučajnu promenljivu, $x_k,\ k = 1,2,\dots p$ posmatra se uzorak obima n centriranih posmatranja $(x_{k1}, x_{k2}, \dots, x_{kn}),\ k = 1,2,\dots p$, i dobija se matrica podataka $\boldsymbol{X}$, za koju važi:

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{A}' + \boldsymbol{E},$$

gde je $\boldsymbol{F}$ matrica faktora, $\boldsymbol{A}$ matrica koeficijenata a $\boldsymbol{E}$ matrica reziduala. Tako se bilo koji elemetn x_{ij} matrice $\boldsymbol{X}$ može izraččunati kao:

$$x_{ij} = \sum_{k=1}^{m} f_{ik}a_{ik} + e_{ij}.$$

4.2.3 Model faktorske analize putem kovarijansne matrice

Gore definisana opšta forma faktorskog modela nije sasvim standardna, pa mnogi autori uvode pretpostavke da bi se dobila drugačija forma. Postoje mnoge pretpostavke za faktorski model, a neke od njih su:

i. $E(\boldsymbol{f}) = \mathbf{0},\ cov(\boldsymbol{f}) = E(\boldsymbol{f}\boldsymbol{f}') = \boldsymbol{\Phi}$

Ovo ograničenje se odnosi na zajedničke faktore. Specijalan slučaj, na osnovu kojeg je baziran model je da važi $\boldsymbol{\Phi} = \mathbf{I}$, odnosno dobija se ortogonalni model kod koga su faktori međusobno nezavisni.

ii. $E(\boldsymbol{e}) = \mathbf{0}, cov(\boldsymbol{e}) = E(\boldsymbol{e}\boldsymbol{e}') = \boldsymbol{\Psi} = diag(\psi_1, \psi_2, \dots, \psi_p)$

Ova ograničenja predstavljaju jedan od uobičajnih pretpostavki u statistici, a to je da je očekivana vrednost greške jednaka nuli. Takođe, pretpostavlja se da je kovarijasna matrica greške dijagonalna matrica, tj. greške nisu međusobno korelisane.

iii. $cov(\boldsymbol{e}\boldsymbol{f}') = cov(\boldsymbol{f}\boldsymbol{e}') = \mathbf{0}$

Pretpostavka govori o nepovezanosti zajedničkih i specifičnih faktora.

Na osnovu pretpostavki, model dobija novu formu:

$$\boldsymbol{\Sigma} = \mathrm{E}((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})') = E((\boldsymbol{A}\boldsymbol{f}+\boldsymbol{e})(\boldsymbol{A}\boldsymbol{f}+\boldsymbol{e})') = E((\boldsymbol{A}\boldsymbol{f}+\boldsymbol{e})(\boldsymbol{f}'\boldsymbol{A}'+\boldsymbol{e}')) =$$

$$= E(\boldsymbol{A}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{A}' + \boldsymbol{A}\boldsymbol{f}\boldsymbol{e}' + \boldsymbol{f}'\boldsymbol{A}' + \boldsymbol{e}\boldsymbol{e}') = E(\boldsymbol{A}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{A}') + E(\boldsymbol{A}\boldsymbol{f}\boldsymbol{e}') + E(\boldsymbol{e}\boldsymbol{f}'\boldsymbol{A}') + E(\boldsymbol{e}\boldsymbol{e}') =$$

$$= \boldsymbol{A}E(\boldsymbol{f}\boldsymbol{f}')\boldsymbol{A}' + \boldsymbol{A}E(\boldsymbol{f}\boldsymbol{e}') + E(\boldsymbol{e}\boldsymbol{f}')\boldsymbol{A} + E(\boldsymbol{e}\boldsymbol{e}') = \boldsymbol{A}\boldsymbol{\Phi}\boldsymbol{A}' + \boldsymbol{\Psi}\,.$$

S obzirom da se posmatra ortogonalni model ($\boldsymbol{\Phi} = \mathbf{I}$), novi model faktorske analize glasi:

$$\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}' + \boldsymbol{\Psi}\,.$$

4.2.4 Transformacija modela

Pri intepretaciji rezultata dobijenih faktorskom analizom, često se primenjuje rotacija zajedničkih faktora. Neka je $\boldsymbol{T}$ ortogonalna matrica, s kojom će biti izvršena ortogonalna rotacija zajedničkih faktora. Osnovni model glasi:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{e} = \boldsymbol{A}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{f} + \boldsymbol{e} = \boldsymbol{A}^*\boldsymbol{f}^* + \boldsymbol{e}.$$

Dalje, ispituju se pretpostavke pomoću kojih se dobija model faktorske analize, ali koristeći nova rotirana rešenja:

i. $E(\boldsymbol{f}^*) = E(\boldsymbol{T}'\boldsymbol{f}) = \boldsymbol{T}'E(\boldsymbol{f}) = \mathbf{0}$,
$cov(\boldsymbol{f}^*) = E(\boldsymbol{f}^*\boldsymbol{f}^{*\prime}) = E(\boldsymbol{T}'\boldsymbol{f}\boldsymbol{f}'\boldsymbol{T}) = \boldsymbol{T}'E(\boldsymbol{f}\boldsymbol{f}')\boldsymbol{T} = \boldsymbol{T}'\boldsymbol{\Phi}\mathbf{T} = \boldsymbol{T}'\mathbf{I}\mathbf{T} = \boldsymbol{T}'\mathbf{T} = \mathbf{I}$;

ii. $E(\boldsymbol{e}) = \mathbf{0}$,
$cov(\boldsymbol{e}) = E(\boldsymbol{e}\boldsymbol{e}') = \boldsymbol{\Psi} = diag(\psi_1, \psi_2, \dots, \psi_p)$;

iii. $cov(\boldsymbol{e}\boldsymbol{f}^{*\prime}) = cov(\boldsymbol{e}\boldsymbol{f}'\boldsymbol{T}) = cov(\boldsymbol{e}\boldsymbol{f}')\boldsymbol{T} = \mathbf{0}$.

Pod ovim pretpostavkama, novi transformisani model glasi:

$$\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}' + \boldsymbol{\Psi} = \mathbf{A}\boldsymbol{T}\boldsymbol{T}'\mathbf{A}' + \boldsymbol{\Psi} = \mathbf{A}^*\mathbf{A}^{*\prime} + \boldsymbol{\Psi}\,.$$

Ova jednačina pokazuje da ne postoji jedinstveno rešenje za matricu $\boldsymbol{A}$, a time i $\boldsymbol{F}$, jer se pomoću ortogonalne transformacione matrice $\boldsymbol{T}$ može promeniti (rotirati) u novo rešenje [11].

4.3 Komunalitet i specifična varijansa

Neka je $x_i = c_i + e_i$, gde je $c_i = a_{i1}f_{j1} + a_{i2}f_{j2} + \cdots + a_{im}f_{jm}$, a e_i rezidualna matrica. Varijansa promenljive x_i je tada jednaka:

$$\sigma_{x_i}^2 = \sigma_{c_i}^2 + \psi_i,$$

gde $\sigma_{c_i}^2$ označava zajedničku varijansu (komunalitet) i predstavlja deo varijanse od x_i koja je zajednička za sve promenljive. Rezidualna varijana ψ_i (specifična varijansa) predstavlja deo varijanse x_i koja nije objašnjena delovanjem faktora i nije zajednička sa drugim promenljivima [3][11]. Daljim izvođenjem vrednosti $\sigma_{x_i}^2$ dobija se:

$$\sigma_{x_i}^2 = \sigma_{c_i}^2 + \sigma_{e_i}^2 = \sum_{j=1}^{n} \frac{\left(a_{i1}f_{j1} + a_{i2}f_{j2} + \cdots + a_{im}f_{jm}\right)^2}{n} + \psi_i =$$

$$a_{i1}^2 \frac{\sum f_{j1}^2}{n} + a_{i2}^2 \frac{\sum f_{j2}^2}{n} + \cdots + a_{im}^2 \frac{\sum f_{jm}^2}{n} + a_{i1}a_{i2}\frac{\sum f_{j1}f_{j2}}{n} + + a_{i(m-1)}a_{ik}\frac{\sum f_{j(m-1)}f_{jm}}{n} + \psi_i\,.$$

Varijansa faktora $\frac{\sum f_{jl}^2}{n}$, $l = 1,2,..m$ je jednaka 1, jer se pretpostavlja da su faktori standardizovani. Proizvod $\frac{\sum f_{jl}f_{jp}}{n}$, $l,p = 1,2,..m$, $l \neq p$ predstavlja korelaciju između faktora. Na osnovu rečenog, dobija se:

$$\sigma_{x_i}^2 = {a_{i1}}^2 + {a_{i2}}^2 + \cdots + {a_{im}}^2 + a_{i1}a_{i2}\phi_{12} + a_{i1}a_{i3}\phi_{13} + \cdots + a_{i(m-1)}a_{ik}\phi_{(m-1)m} + \psi_i\,,$$

gde je ϕ_{ij} korelacija između faktora. Ukoliko su faktori nekorelisani ($\phi_{ij} = 0$), sledi

$$\sigma_{x_i}^2 = a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 + \psi_i, \qquad i = 1,2,\dots,p.$$

Dakle, varijansa promenljive x_i je jednaka komunalitetu plus specifičnoj varijansi, odnosno

$$\sigma_{x_i}^2 = a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 + \psi_i = h_i^2 + \psi_i, \qquad i = 1,2,\dots,p.$$

4.4 Primer faktorske analize

U nastavku sledi jednostavan primer faktorske analize [3]. Neka su $x_1, x_2, \dots, x_6$ promenljive koje se dobijaju merenjem osobina velike populacije organizma i nejne okoline, tako da su x_1, x_2 i x_6 različite mere morfoloških osobina organizama, a x_2, x_3 i x_5 mere određenih vrsta ekoloških uslova za koje se sumnja da su uticali na morfologiju ovih organizma. Zbog jednostavnosti, promenljive su date u standardizovanoj formi i korelaciona matrica glasi:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.72 & 0.378 & 0.324 & 0.27 & 0.27 \\ 0.72 & 1 & 0.336 & 0.288 & 0.24 & 0.24 \\ 0.378 & 0.336 & 1 & 0.42 & 0.35 & 0.126 \\ 0.324 & 0.288 & 0.42 & 1 & 0.3 & 0.108 \\ 0.27 & 0.24 & 0.35 & 0.3 & 1 & 0.09 \\ 0.27 & 0.24 & 0.126 & 0.108 & 0.09 & 1 \end{bmatrix}.$$

Elementi matrice $\mathbf{\Sigma}$ predstavljaju populacione vrednosti korelacija ρ_{ij}.

Faktorskom analizom se pokušavaju objasniti korelacije uvođenjem faktora $f_1, f_2, \dots, f_m$. Jedan način da se objasni ovaj postupak jeste da se postavi sledeće pitanje: Da li postoji jedan faktor f_1, takav da kada se on izdvoji, među promenljivama više ne postoji povezanost. U tom slučaju, korelacija između bilo kog para promenljivih x_i i x_j, bi morala biti jednaka nuli.

Dakle, za neke dve različite promenljive, važilo bi

$$x_i = a_i f_1 + e_i \quad \text{i} \quad x_j = a_j f_1 + e_j \ ,$$

gde su e_i i e_j nekorelisane za $i \neq j$. Koelacioni koeficijent ρ_{ij} bi tada bio jednak

$$\rho_{ij} = \rho(x_i, x_j) = a_i a_j, \qquad i \neq j \ .$$

pa bi korelacije u vrstama i i j bile proporcionalne. Kako to ne važi za gornju matricu $\mathbf{\Sigma}$, sledi da nije dovoljan samo jedan faktor da objasni sve korelacije u $\mathbf{\Sigma}$, pa je potrebno uvesti još faktora.

Pod pretpostavkom da u faktori nekorelisani ($\mathbf{\Phi} = \boldsymbol{I}$), osnovni model faktorske analize glasi:

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{A}' + \mathbf{\Psi} \ .$$

Proizvod $\boldsymbol{AA'}$ bi trebao da reprodukuje vandijagonalne elemente korelacione matrice $\mathbf{\Sigma}$, uz postojanje minimalne greške. U proširenom obliku:

$$\rho_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2} + \cdots + a_{ik}a_{jk} \ .$$

Izračunavanjem matrica $\boldsymbol{A}$, zajednička varijansa h_i^2 za promenljivu x_i glasi:

$$h_i^2 = {a_{i1}}^2 + {a_{i2}}^2 + \cdots + {a_{ik}}^2.$$

Korišćenjem određene metode i transformacija, koje će biti naknadno prikazane, izračunava se matrica $\boldsymbol{A}$, koja je u ovom slučaju jednaka

$$\boldsymbol{A} = \begin{bmatrix} 0.889 & -0.138 \\ 0.791 & -0.122 \\ 0.501 & 0.489 \\ 0.429 & 0.419 \\ 0.358 & 0.349 \\ 0.296 & -0.046 \end{bmatrix}.$$

Proverom, lako se pokazuje da važi:

$$\rho_{12} = a_{11}a_{21} + a_{12}a_{22} = 0.889 * 0.791 + (-0.138) * (-0.122) = \\ = 0.7032 + 0.0168 = 0.7200\,,$$

što je jednako odgovarujćem elementu u matrici $\boldsymbol{\Sigma}$.

Vrednost komunaliteta promenljive x_i je jednaka

$$h_1^2 = a_{11} + a_{12} = 0.889^2 + (-0.138)^2 = 0.81\,.$$

Tako se dolazi do činjenice da postoje dva nekorelisana faktora f_1 i f_2, koji objašnjavaju šest promenljivih

$$x_1 = 0.889f_1 - 0.138f_2 + e_1$$
$$x_2 = 0.791f_1 - 0.122f_2 + e_2$$
$$x_3 = 0.501f_1 + 0.489f_2 + e_3$$
$$x_4 = 0.429f_1 + 0.419f_2 + e_4$$
$$x_5 = 0.358f_1 + 0.349f_2 + e_5$$
$$x_6 = 0.296f_1 - 0.046f_2 + e_6$$

i važi $\rho(e_i, e_j) = 0, \forall\, i, j$.

Faktori f_1 i f_2 nisu jedina dva faktora koji zadovoljavaju jednakost $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}' + \boldsymbol{\Psi}$. Na primer, sledeća dva faktora takođe zadovoljavaju jednakost:

$$f_1^* = 0.988f_1 - 0.153f_2 \quad \text{i} \quad f_2^* = -0.153f_1 + 0.988f_2.$$

U suštini, faktori f_1^* i f_2^* su dobijena korišćenjem ortonormirane transformacione matrice $\boldsymbol{T}$ na faktore f_1 i f_2, koja je jednaka:

$$\boldsymbol{T} = \begin{bmatrix} 0.988 & -0.153 \\ -0.153 & 0.988 \end{bmatrix}.$$

Množenjem matrice $\boldsymbol{A}$ sa matricom $\boldsymbol{T}$, dobija se nova matrica faktorskih koeficijenata $\boldsymbol{A}^*$, koja je jednaka

$$\boldsymbol{A}^* = \begin{bmatrix} 0.90 & 0 \\ 0.80 & 0 \\ 0.42 & 0.56 \\ 0.36 & 0.48 \\ 0.30 & 0.40 \\ 0.30 & 0 \end{bmatrix}.$$

Dakle, $\boldsymbol{A}^*$ predstavlja novo rešenje i sledi niz novih jednačina:

$$\begin{aligned} x_1 &= 0.889 f_1^* + e_1 \\ x_2 &= 0.791 f_1^* + e_2 \\ x_3 &= 0.501 f_1^* + 0.489 f_2^* + e_3 \\ x_4 &= 0.429 f_1^* + 0.419 f_2^* + e_4 \\ x_5 &= 0.358 f_1^* + 0.349 f_2^* + e_5 \\ x_6 &= 0.296 f_1^* + e_6 \,. \end{aligned}$$

Proverom, ponov važi

$$\rho_{12} = 0.720,$$

$$h_1^2 = 0.81.$$

Matrica $\boldsymbol{A}^*$ za razliku od $\boldsymbol{A}$ ima mnogo više nula u svojoj strukturi, pa se često takve matrice nazivaju matrice jednostavne strukture.

Generalno, najjednostavnija struktura se postiže tako što se postavi uslov da su faktori korelisani, ali time se gubi svojstvo ortogonalnosti faktora. Na primer, neka su dva nova korelisana faktora definisana kao

$$f_1^{**} = f_1^* \quad \text{i} \quad f_2^{**} = 0.6 f_1^* + 0.8 f_2^* \,,$$

tako da je korelacija jednaka $\rho(f_1^{**}, f_2^{**}) = 1 * 0.6 + 0 * 0.8 = 0.6$. Odatle je kovarijasna matrica faktora jednaka

$$\boldsymbol{\Phi} = \begin{bmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{bmatrix}$$

Transformaciona matrica $\boldsymbol{T}$ koja preslikava $f^* \to f^{**}$ je jednaka

$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 \\ 0.6 & 0.8 \end{bmatrix},$$

odnosno inverzna je jednaka

$$\boldsymbol{T}^{-1} = \begin{bmatrix} 1 & 0 \\ -0.75 & 1.25 \end{bmatrix}.$$

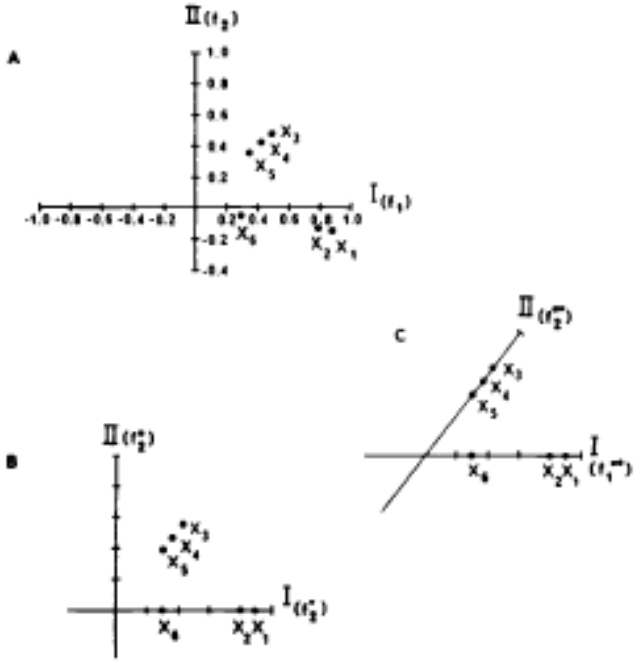
Pomoću matrice $\boldsymbol{T}^{-1}$ lako se izračunava $\boldsymbol{A}^{**}$ koja odgovara faktoru $\boldsymbol{f}^{**}$. Dakle, matrica $\boldsymbol{A}^{**}$ je jednaka

$$\boldsymbol{A}^{**} = \boldsymbol{A}\boldsymbol{T}^{-1} = \begin{matrix} \\ 0.9 \\ 0.8 \\ 0 \\ 0 \\ 0 \\ 0.3 \end{matrix} \begin{matrix} f_1^{**} & f_2^{**} \\ \end{matrix} \begin{bmatrix} 0 \\ 0 \\ 0.7 \\ 0.6 \\ 0.5 \\ 0 \end{bmatrix}.$$

Zaključak je da prvi faktor kombinuje morfološke mere, a drugi faktor ekološke osobine.

Koristći geometrijski prikaz lako se ilustruju rezultati dobijeni u primeru, a time i lakše analizuju. U koordinatnom sistemu čije su ose faktori, prikazane su promenljive $x_1, x_2, \ldots, x_6$. Na slici 4.1, prikazana su tri rešenja dobijena u primeru:

- prikaz nerotiranih rezultata (A),
- efekat rotacije osa (u pravcu kretanaj kazaljke na satu) (B),
- efekairotacije samo druge ose (C).



Slika 4.1 *Prikaz rešenja i rotacija koordinata*

U okviru rešenja $\boldsymbol{A}$ i $\boldsymbol{A}^*$, faktorske ose su predstavljene kao dva ortogonalna vektora dužine 1. Tako, u okviru grafika A, ose su predstavljene sistemom (f_1, f_2). Projekcijom promenljivih na faktore zaključuje se da je jedna grupa promenljivih pođednako udaljena od faktora, do kod druge grupe to ne važi.

U sledećem grafiku, izvršena je rotacija koordinatnog sistema u pravcu kazaljke na satu i dobijen je novi sistem osa (f_1^*, f_2^*) koje su i dalje ortogonalne i dužine 1. Ovom rotacijom je dobijeno da promenljive x_1, x_2, x_6 leže na prvoj osi (faktoru) i time im je

koeficijent (koordinata) u odnosu na drugu osu (faktor) jednak nuli. Sa druge strane promenljive x_3, x_4, x_5 imaju nenula koeficijente u odnosu na oba faktora.

Daljom rotacijom, ali samo jednog faktora, dobija se novi sistem osa (f_1^{**}, f_2^{**}). Prvi faktor je ostao na istom mestu (s obzirom da na njemu leže promenljive, što smo i želeli postići), dok se drugi faktor rotira da bi sadržao promenljive x_3, x_4, x_5. Ugao između novih osa je jednak 0.6, odnosno korelaciji između faktora.

4.5 Metoda za izračunavanje matrice A

4.5.1 Glavna faktorska metoda

Faktorski model putem kovarijansne matrice glasi:

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{A}' + \mathbf{\Psi}\,.$$

Dijagonalni elementi matrice $\mathbf{\Psi}$ su specifične varijanse i procenjuju se na osnovu podataka matrice $\boldsymbol{X}$. Ovi dijagonalni elementi jedinstvene varijanse se procenjuju na osnovu podataka, zajedno sa matricom $\boldsymbol{A}$. Uobičajan kriterijum za određivanje modela sa podacima jeste minimiziranje sume kvadrata svih elemenata $\boldsymbol{S} - \boldsymbol{A}\boldsymbol{A}' - \mathbf{\Psi}$, to jest minimiziranje

$$trag(\boldsymbol{S} - \boldsymbol{A}\boldsymbol{A}' - \mathbf{\Psi})^2.$$

Ukoliko je vrednost matrice $\mathbf{\Psi}$ poznata, vrednost matrice $\boldsymbol{A}\boldsymbol{A}'$ se računa putem analize glavnih komponenti, s tim da se u kolone matrice $\boldsymbol{A}$ stavljaju normalizovani karakteristični vektori matrice $\boldsymbol{S} - \mathbf{\Psi}$ koji odgovaraju prvih k najvećih karakterističnih vrednosti. Međutim, u praksi matrica $\mathbf{\Psi}$ je nepoznata i procenjuje se iz podataka.

Matrica reziduala $\mathbf{\Psi}$ se ocenjuje putem ocenjivanja komunaliteta. Postoji nekoliko metoda u određivanju komunaliteta ali tradicionalno je najpoznatija metoda kod koje se koristi celokupna kovarijansna matrica u cilju ocene komunaliteta (ULS metoda). Prema toj metodi, h_i^2 je predstavljen kvadratom višestrukog koeficijenata kovarijanse promenljive x_i i svih preostalih $p-1$ promenljivih.Ova metoda daje donju granicu ocene komunaliteta. Tako je ocena i-te specifične varijanse $\hat{\psi}_i = 1/s^{ii}$, gde je s^{ii} i-ti dijagonalni element matrice $\boldsymbol{S}^{-1}$. Dakle, procenu matrice reziduala $\widehat{\mathbf{\Psi}}$ računa se kao

$$\widehat{\mathbf{\Psi}} = (diag\ \boldsymbol{S}^{-1})^{-1}.$$

Dalja procedura je istao kao i u slučaju kada je poznata matrica $\mathbf{\Psi}$. Ova metoda se naziva *glavna faktorska metoda*. Matrice $\widehat{\boldsymbol{A}}$ i $\widehat{\boldsymbol{F}}$ su sistemski pristrasne.

Sama metoda se može mnogo poboljšati korišćenjem interativnog postupka. Ukoliko bi se procenila vrednost $\widehat{\mathbf{\Psi}}$, a time i $\widehat{\boldsymbol{A}}$, model se može poboljšati novom vrednošću $\widehat{\mathbf{\Psi}}$ na sledeći način:

$$\widehat{\mathbf{\Psi}} = diag\big(\boldsymbol{S} - \widehat{\boldsymbol{A}}\widehat{\boldsymbol{A}}'\big).$$

Ovaj proces se ponavlja sve dok se ne dobije stabilna vrednost matrice $\widehat{\boldsymbol{A}}$. Ova metoda je poznata kao *interativna principalna faktorska metoda*.

Mnogo efektivnija metoda faktorske analize se postiže minimiziranjem funkcije $trag(\boldsymbol{S}-\boldsymbol{AA'}-\boldsymbol{\Psi})^2$. Jedna od takvih metoda jeste MINRES, koju je kreirao Harman 1967. godine, koja minimizuje sumu kvadrata svih vandijagonalnih elemenata matrice $\boldsymbol{S}-\boldsymbol{AA'}$ i time se izbegavaju komunaliteti. Druga metoda jeste ULS metoda, razvijena od strane Joreskog 1976. godine. Ona putem eliminacije matrice $\boldsymbol{A}$, smanjuje funkciju na funkciju matrice Ψ i onda se minimizira putem Njutn-Rafsonove procedure [3].

4.5.2 Neskalirana ocena matrice A i Ψ

Jedan od problema koji se javlja kod ULS metode jeste da ne daje skalirano rešenje. Problem je u tome što jedinice mere kod promenljive su često proizvoljne. Rešenje predstavlja korelaciona matrica i zato ULS metoad mnogo češće koristi korelacionu matricu umesto kovarijansne.

U nastavku, prikazana je neskalirano metoda ocene matrice $\boldsymbol{A}$ i $\boldsymbol{\Psi}$. Neka je $\boldsymbol{\Psi}$ poznata matrica. U tom slučaju, aproksimativno važi

$$\boldsymbol{S}-\boldsymbol{\Psi}\approx \mathbf{AA}'.$$

Množenjem jednakosti sleve i desne strane, sa matricom $\boldsymbol{\Psi}^{-1/2}$, dobija se

$$\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}-\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Psi}\boldsymbol{\Psi}^{-1/2}\approx \boldsymbol{\Psi}^{-1/2}\mathbf{AA}'\boldsymbol{\Psi}^{-1/2}$$

odnosno

$$\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}-\mathbf{I}\approx \boldsymbol{\Psi}^{-1/2}\mathbf{AA}'\boldsymbol{\Psi}^{-1/2}.$$

Pošto je $\boldsymbol{\Psi}$ poznato, lako se izračunava $\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}-\mathbf{I}$ i time se dobija vrednost $\boldsymbol{A}^*\boldsymbol{A}^{*\prime}$, gde je $\boldsymbol{A}^*=\boldsymbol{\Psi}^{-1/2}\mathbf{A}$. Izračunavanjem karakterističnih korena i vektora iz $\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}-\mathbf{I}$, izračunava se ocena $\widehat{\boldsymbol{A}}^*$, odakle je $\mathbf{A}=\boldsymbol{\Psi}^{1/2}\widehat{\boldsymbol{A}}^*$. Ukoliko je $\boldsymbol{\Psi}$ nepoznato, korsiteći se ocena

$$(diag\ \boldsymbol{S}^{-1})^{1/2}\boldsymbol{S}(diag\ \boldsymbol{S}^{-1})^{1/2}-\mathbf{I}.$$

S obziorm da ocena matrice $\boldsymbol{\Psi}$ je sistemski previše velika, Joreskog (1963) je predložio da ocena matrice $\boldsymbol{\Psi}$ bude

$$\widehat{\boldsymbol{\Psi}}=\boldsymbol{\theta}(diag\ \boldsymbol{S}^{-1})^{-1}$$

gde je θ nepoznati skalara, vrednosti manje od 1 ($\theta<1$), koji se ocenjuje na osnovu podataka. Neka su $\lambda_1,\lambda_2,\dots,\lambda_p$ karakteristični koreni matrice $\boldsymbol{S}^*$:

$$\boldsymbol{S}^*=(diag\ \boldsymbol{S}^{-1})^{1/2}\boldsymbol{S}(diag\ \boldsymbol{S}^{-1})^{1/2}.$$

Metodom najmanjih kvadrata, ocena parametra θ je jednaka:

$$\hat{\theta} = \frac{1}{p-k} \sum_{m=k+1}^{p} \lambda_m \,,$$

odnosno, ocena je jednaka srednjoj vrednosti $p-k$ najmanjih karakterističnih korena matrice $\boldsymbol{S}^*$. Štaviše, neka je $\boldsymbol{U}_k$ matrica dimenzije $p \times k$, gde su njene kolone ortnonormirani karakteristični vektori matrice $\boldsymbol{S}^*$ povezane sa k najvećih karakterističnih korena koji su smešteni na dijagonali matrice $\boldsymbol{\Lambda}_k$. Tada je ocena matrice $\boldsymbol{A}$ metodom najmanjih kvadrata jednaka

$$\widehat{\boldsymbol{A}} = (diag\, \boldsymbol{S}^{-1})^{-1/2} \boldsymbol{U}_k (\boldsymbol{\Lambda}_k - \hat{\theta}\boldsymbol{I})^{1/2}.$$

Dakle, prvo se kolone matrice $\boldsymbol{U}_k$ skaliraju sa sumom korena j-te kolone iznosom $\lambda_j - \hat{\theta}$ pa zatim skalira i-tu vrstu sa $1/\sqrt{s^{ii}}$, gde je s^{ii} i-ti dijagonalni element matrice $\boldsymbol{S}^{-1}$ [3].

4.5.3 Primer

U nastavku, biće prikazan primer ocene matrice $\boldsymbol{A}$ korišćenjem korelacione matrice $\boldsymbol{S}$, ULS metodom koju je razvio Joreskog [3]. Neka matrica $\boldsymbol{S}$ (u ovom slučaju je to korelaciona matrica) ima elementi:

$$\boldsymbol{S} = \begin{bmatrix} \mathbf{1.000} & 0.466 & 0.456 & 0.441 & 0.375 & 0.312 & 0.247 & 0.207 \\ 0.466 & \mathbf{1.000} & 0.311 & 0.296 & 0.521 & 0.286 & 0.483 & 0.314 \\ 0.456 & 0.311 & \mathbf{1.000} & 0.185 & 0.184 & 0.300 & 0.378 & 0.378 \\ 0.441 & 0.296 & 0.185 & \mathbf{1.000} & 0.176 & 0.244 & 0.121 & 0.341 \\ 0.375 & 0.521 & 0.184 & 0.176 & \mathbf{1.000} & 0.389 & 0.211 & 0.153 \\ 0.312 & 0.286 & 0.300 & 0.244 & 0.389 & \mathbf{1.000} & 0.210 & 0.289 \\ 0.247 & 0.483 & 0.378 & 0.121 & 0.211 & 0.210 & \mathbf{1.000} & 0.504 \\ 0.207 & 0.314 & 0.378 & 0.341 & 0.153 & 0.289 & 0.504 & \mathbf{1.000} \end{bmatrix}.$$

Dalje, u tabeli 4.1 su prikazani dijagonalni elementi matrice $\boldsymbol{S}^{-1}$ i određene vrednosti koje će se koristiti u nastavku analize.

	s^{ii}	$\sqrt{s^{ii}}$	$\frac{1}{\sqrt{s^{ii}}}$	$\frac{1}{s^{ii}}$	$1-\frac{1}{s^{ii}}$	$\sqrt{1-\frac{1}{s^{ii}}}$
1	1.785	1.336	0.748	0.560	0.440	0.663
2	1.911	1.382	0.723	0.523	0.477	0.690
3	1.506	1.227	0.815	0.664	0.336	0.580
4	1.422	1.193	0.838	0.703	0.297	0.545
5	1.553	1.246	0.802	0.644	0.356	0.597
6	1.323	1.150	0.869	0.756	0.244	0.494
7	1.687	1.299	0.670	0.593	0.407	0.638
8	1.623	1.274	0.785	0.616	0.384	0.620

Tabela 4.1 *Prikaz potrebnih podataka*

Matrica $\boldsymbol{S}^*$ se lako izračunava, koristeći podatke iz tabele 4.1 i formulu:

$$\boldsymbol{S}^* = (diag\ \boldsymbol{S}^{-1})^{1/2}\boldsymbol{S}(diag\ \boldsymbol{S}^{-1})^{1/2},$$
$$(\ s_{ij}^* = \sqrt{s^{ii}}s_{ij}\sqrt{s^{jj}}\ ),$$

Bitno je napomenuti da bi matrica $\boldsymbol{S}^*$ bila ista ukoliko bi početna matrica $\boldsymbol{S}$ bila kovarijansna, a ne korelaciona matrica. Matrica $\boldsymbol{S}^*$ je jednaka:

$$\boldsymbol{S}^* = \begin{bmatrix} \mathbf{1.785} & 0.861 & 0.748 & 0.703 & 0.624 & 0.479 & 0.429 & 0.352 \\ 0.861 & \mathbf{1.911} & 0.528 & 0.488 & 0.898 & 0.455 & 0.867 & 0.553 \\ 0.748 & 0.528 & \mathbf{1.506} & 0.271 & 0.281 & 0.423 & 0.603 & 0.591 \\ 0.703 & 0.488 & 0.271 & \mathbf{1.422} & 0.262 & 0.335 & 0.187 & 0.518 \\ 0.624 & 0.898 & 0.281 & 0.262 & \mathbf{1.553} & 0.558 & 0.342 & 0.243 \\ 0.479 & 0.455 & 0.423 & 0.335 & 0.558 & \mathbf{1.323} & 0.314 & 0.423 \\ 0.429 & 0.867 & 0.603 & 0.187 & 0.342 & 0.314 & \mathbf{1.687} & 0.834 \\ 0.352 & 0.553 & 0.591 & 0.518 & 0.243 & 0.423 & 0.834 & \mathbf{1.623} \end{bmatrix}.$$

Vrednosti karakterističnih korena i trag matrice $\boldsymbol{S}^*$ su prikazani u tabeli 4.2. U cilju određivanja broja faktora, koristiće se pravilo poznato kao rough rule (gruba pravilo). Najmanja vrednost karakterističnog korena je 0.534 i on je udaljen od 1 za vrednost od 0.466. Sada, posmatraće se vrednosti karakterističnih korena čija je vrednost veća od 1+0.466 odnosno od 1.466. Njih ima ukupno tri, dakle $k = 3$.

λ_1	5.281	λ_5	1.152
λ_2	1.809	λ_6	0.703
λ_3	1.507	λ_7	0.625
λ_4	1.199	λ_8	0.534
trag($\boldsymbol{S}^*$)		12.810	

Tabela 4.2 *Vrednost karakterističnih korena matrice* $\boldsymbol{S}^*$

Da bi se izračunala matrica $\boldsymbol{A}$, prvo mora da se oceni parametar θ. Po formuli, ona je jednaka srednjoj vrednosti $p - k$ karakterističnih korena, odnosno:

$$\hat{\theta} = \frac{1}{5}\sum_{i=4}^{8}\hat{\lambda}_i = 0.843\ .$$

Zatim se računa vrednost matrice $\boldsymbol{\Lambda}_k - \hat{\theta}\boldsymbol{I}$ i dobija se:

$$\boldsymbol{\Lambda}_k - \hat{\theta}\boldsymbol{I} = \begin{bmatrix} 5.281 & 0 & 0 \\ 0 & 1.809 & 0 \\ 0 & 0 & 1.507 \end{bmatrix} - \begin{bmatrix} 0.843 & 0 & 0 \\ 0 & 0.843 & 0 \\ 0 & 0 & 0.843 \end{bmatrix} = \begin{bmatrix} 4.438 & 0 & 0 \\ 0 & 0.966 & 0 \\ 0 & 0 & 0.664 \end{bmatrix},$$

odnosno korenovanjem:

$$\sqrt{\boldsymbol{\Lambda}_k - \hat{\theta}\boldsymbol{I}} = \begin{bmatrix} \sqrt{4.438} & 0 & 0 \\ 0 & \sqrt{0.966} & 0 \\ 0 & 0 & \sqrt{0.664} \end{bmatrix} = \begin{bmatrix} 2.107 & 0 & 0 \\ 0 & 0.983 & 0 \\ 0 & 0 & 0.815 \end{bmatrix}.$$

Dalje se računa matrica karakterističnih korena $\boldsymbol{U}_3$, na osnovi 3 najveća karakteristična korena:

$$\boldsymbol{U}_3 = \begin{bmatrix} 0.41 & -0.37 & 0.36 \\ 0.47 & -0.15 & -0.42 \\ 0.33 & 0.20 & 0.26 \\ 0.27 & -0.15 & 0.56 \\ 0.32 & -0.44 & -0.42 \\ 0.27 & -0.13 & 0.05 \\ 0.37 & 0.52 & -0.33 \\ 0.34 & 0.54 & 0.17 \end{bmatrix}.$$

Konačno, svi potrebni parametri i matrice su izračunate tako da može da se izračuna matrica $\widehat{\boldsymbol{A}}$, i ona je jednaka

$$\widehat{\boldsymbol{A}} = (diag\ \boldsymbol{S}^{-1})^{-1/2}\boldsymbol{U}_3(\boldsymbol{\Lambda}_3 - \hat{\theta}\boldsymbol{I})^{1/2} = \begin{bmatrix} 0.65 & 0.27 & 0.22 \\ 0.71 & -0.11 & -0.25 \\ 0.57 & 0.16 & 0.17 \\ 0.47 & -0.12 & 0.38 \\ 0.55 & -0.35 & -0.28 \\ 0.50 & -0.11 & 0.03 \\ 0.59 & 0.40 & -0.20 \\ 0.56 & 0.42 & 0.11 \end{bmatrix}.$$

4.6 Razlika između analize glavnih komponenti i faktorske analize

Mnogi autori tretiraju analizu glavnih komponenti kao specifičan slučaj ili prvu fazu u faktorskoj analizi. Opravdanje za ovakav tretman dve metode multivarijacione analize nalazi se u sličnosti rezultata koji se dobijaju njihovom primenom, ali kad se uđe u suštinu, one su zaista distancirane tehnike i traže posebna matematička definisanja. Delom je zabuna nastala zbog originalnog rada Hotelling-a (1933) u kojima je on glavne komponente uveo kao manji broj fundamentalnih promenljivih koje određuju vrednost p originalnih promenljivih. Iako je ovaj vid definisanja glavnih komponenti u duhu sa faktorskim modelom, Girschick (1936) je smatrao da su Hotelling-ove glavne komponente neprikladne za faktorsku analizu. U suštini, cilj i faktorske analize i analize glavnih komponenti jeste smanjenje dimenzionalnosti skupa podataka, ali im se metode mnogo razlikuju.

Fundamentalna razlika između analize glavnih komponenti i faktorske analize jeste način na koji su faktori (komponente) definisane i pretpostavke vezanih za reziduale. U faktorskoj analizi postoji određeni model, koji je osnova faktorske analize, dok se u analizi glavnih komponenti ne pretpostavlja model. Dalje, u analizi glavnih komponenti, komponente su određene tako da objašnjavaju maksimalnu varijansu svih registrovanih promenljivih, dok u faktorskoj analizi, faktori su definisani tako da maksimalno objašnjavaju korelaciju između

promenljivih. Sledi zaključak da je analiza glavnih komponenti orijentisana ka varijansi, a faktorska analiza orijentisana ka korelaciji.

Dalje, faktorska anliza polazi od ideje razlaganja varijanse promenljivih na dva dela: zajednički i specifični deo. Zajednički deo je onaj deo varijanse promenljive koji ona deli sa ostalim promenljivama, dok je specifični deo onaj deo varijanse promenljive koji je poseban za tu promenljivu. Dakle, faktorska analiza izučava deo varijanse koji je zajednički za sve promenljive, a anliza glavnih komponenti izučava ukupnu varijansu skupa podataka.

Što se tiče reziduala, pretpostavlja se da imaju veoma malu vrednost u analizi glavnih komponenti, za razliku od faktorske analize gde oni nisu zanemarljivi. Takođe, u obe metode je pretpostavljeno da reziduali nisu korelisani sa faktorima (komponentama). Međutim, u analizi glavnih komponenti ne postoje pretpostavke o korelaciji između reziduala, dok u faktorskoj analizi se pretpostavlja da su reziduali međusobno nekorelisani.

U obe metode javljaju se i opažljive (originalne) i neopažljive ili latentne (faktori, glavne komponente) promenljive. U opštem slučaju, nije isto izraziti latentne promenljive uz pomoć originalnih promenljivih i obrnuto. Tu leži razlika između metoda. Ideja faktorske analize je

$$x_i = a_{i1}f_1 + a_{i2}f_2 + \cdots + a_{im}f_m + e_i\,,$$

odnosno da se pomoću faktora izraze originalne promenljive. S druge strane, kod analize glavnih komponenti, glavne komponente su izražene pomoću originalnih promenljivih, odnosno:

$$z_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{im}x_m\,.$$

Takođe, ograničenja koja postoje za kovarijasnu matricu $\mathbf{\Sigma}$ se razlikuju u ovim metodama. U analizi glavnih komponenti, s obzirom da je pretpostavljeno da greška ima malu vrednost, važi:

$$\mathbf{\Sigma} \approx \mathbf{A\Phi A}'.$$

Ovo implicira da je matrica $\mathbf{\Sigma}$, dimenzije $p \times p$, ranga m. U faktorskoj analizi, rezidualna matrica $\mathbf{\Psi}$ je dijagonalna matrica jer se pretpostavlja da su reziduali međusobno nekorelisani, odnosno $\psi_{ij} = 0,\ i \neq j$. Sledi da su vandijagonalni elementi matrice $\mathbf{\Sigma}$ jednaki vandijagonalni elementima matrice $\mathbf{A\Phi A}'$. Dijagonala matrice $\mathbf{A\Phi A}'$ sadrži komunalitete (zajedničke varijanse promenljivih). Tako, u analizi glavnih komponenti je cilj da se reprodukuju i dijagonalni i vandijagonalni elementi matrice $\mathbf{\Sigma}$, dok je u faktorskoj analizi cilj da se reprodukuju samo vandijagonalni elementi, jer ti elementi objašnjavaju kovarijansu, a ne varijansu [3],[1].

Glava 5

Modeli višestrukih regresija i regresija parcijalnih najmanjih kvadrata

5.1 Regresija običnih najmanjih kvadrata

Višestruka regresija (*multiple regression*) predstavlja statističku metodu kod koje se matrice zavisnih promenljivih $\boldsymbol{Y}$ izražava kao linearna suma nezavisnih promenljivih $\boldsymbol{X}$:

$$\underset{n\times q}{\boldsymbol{Y}} = \underset{n\times p}{\boldsymbol{X}}\ \underset{p\times q}{\boldsymbol{B}} + \underset{n\times q}{\boldsymbol{E}}\,.$$

Matrica $\boldsymbol{B}$ predstavlja matricu parametara (regresionih koeficijenata), a matrica $\boldsymbol{E}$ predstavlja grešku.

Regresija običnih najmanjih kvadrata (*ordinary least squares regression*) predstavlja specijalan slučaj višestruke regresije, gde je zavisna promenljiva vektor $\boldsymbol{y}$. U tom slučaju, matrica parametara je vektor i regresija običnih najmanjih kvadrata dobija sledeću formu:

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

Dalje, neka je matrica $\boldsymbol{X}$ punog ranga, odnosno $rang(\boldsymbol{X}) = p$. Tada se ocena matrice parametara $\widehat{\boldsymbol{\beta}}_{OLS}$ može dobiti minimizacijom sume najmanjih kvadrata (**SSE**), odnosno diferenciranjem **SSE**-a po $\boldsymbol{\beta}$ i izjednačavanjem sa nulom:

$$\boldsymbol{SSE} = \boldsymbol{e'e} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{y'y} - \boldsymbol{2\beta'X'y} + \boldsymbol{\beta'X'X\beta}$$

$$\frac{d\boldsymbol{SSE}}{d\boldsymbol{\beta}} = -\boldsymbol{2X'y} + \boldsymbol{2X'X\beta} = 0$$

$$\boldsymbol{\beta} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} = \boldsymbol{X^{+}y}\,.$$

Za zadate vrednosti matrice $\boldsymbol{X}$ i $\boldsymbol{y}$, dobijena je ocena vrednosti parametra $\widehat{\boldsymbol{\beta}}_{OLS}$, dimenzije $p \times 1$. $\widehat{\boldsymbol{\beta}}_{OLS}$ daje nepristrasnu ocenu elementa $\boldsymbol{\beta}$, i ta ocena ima minimalnu varijansu.

Situacija kada postoji q zavisnih promenljivih, ocena dobijena regresijom najmanjih kvadrata se može uopštiti kao:

$$\widehat{\boldsymbol{B}}_{OLS} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'ZY},$$

gde je $\widehat{\boldsymbol{B}}_{OLS}$ ocena matrice parametara $\boldsymbol{B}$ dobijena metodom najmanjih kvadrata. Ukoliko je nezavisna promenljiva jako korelisana, tada je matrica $\boldsymbol{X'X}$ loše uslovljena (*ill-conditioned matrix*, matrica teži da bude singularna). Time ocena $\widehat{\boldsymbol{B}}_{OLS}$ ima veliku varijansu. Ukoliko

postoji multikolinearnost, ocena $\widehat{\boldsymbol{B}}_{OLS}$ koeficijenta može biti statistički neznačajna (ima preveliku vrednost, premalu vrednost) čak iako je koeficijent determinacije visok.

Pod nazivom metode pristrasnih ocena spadaju brojne alternativne metode za ocenu parametara tako dizajnirane da utiču na smanjenje multikolinearnosti. Ukoliko se odustane od nepristrasnih metoda, tada se pristrasne metode, kao što su regresija glavnih komponenti, rubna regresija i regresija najmanjih kvadrata, koriste u rešavanju problema nepreciznih predikcija [14].

5.2 Regresija glavnih komponenti

Regresija glavnih komponenti predstavlja regresionu metodu, kojom se rešava problem ill-conditioned matrica. Ideja regresije glavnih komponenti je da se izaberu one glavne komponente koje objašnjava najveći deo varijanse matrice $\boldsymbol{X}$ i time optimizuju prediktivne mogućnosti modela. Zapravo, regresija glavnih komponenti je linearna regresija gde se matrice $\boldsymbol{Y}$ dobija kao linearna kombinacija glavnih komponenti.

Neka je $\boldsymbol{X}$ centrirana i normalizovana matrica. Na osnovu analize glavnih komponenti, važi

$$\boldsymbol{X}'\boldsymbol{X}\gamma_i = \lambda_i\gamma_i, i = 1, 2, \dots, p\ ,$$

gde su λ_i karakteristični koreni matrice $\boldsymbol{X}'\boldsymbol{X}$, $\boldsymbol{X}'\boldsymbol{X}$ korelaciona matrica, a γ_i su karakteristični vektori matrice $\boldsymbol{X}'\boldsymbol{X}$. Glavne komponente z_i matrice podataka $\boldsymbol{X}$ su definisane kao:

$$z_i = \boldsymbol{a}_i'\boldsymbol{X} = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p\ ,$$

i one su međusobno ortogonalne. Razvojem osnovnog modela, dobija se da prvih m glavnih komponenti z_i optimizuje prediktivne mogućnosti modela, odnosno:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = (\boldsymbol{X}\boldsymbol{A}_m)(\boldsymbol{A}'_m\boldsymbol{\beta}) + e = \boldsymbol{Z}_m\boldsymbol{\alpha}_m + e.$$

Odavde važi da je $\boldsymbol{\alpha}_m = (\boldsymbol{Z}'_m\boldsymbol{Z}_m)\boldsymbol{Z}'_m\boldsymbol{y}$, a broj m predstavlja broj glavnih komponenti koje su sadržane u modelu. Koristeći ocenu vrednosti $\boldsymbol{\alpha}$, lako se dobija ocenjena vrednost $\boldsymbol{\beta}$, odnosno:

$$\widehat{\boldsymbol{\beta}}_{PCR} = \boldsymbol{A}_m\boldsymbol{\alpha}_m\ ,$$

gde je $\boldsymbol{A}_m$ matrica koja se sastoji od prvih m karakterističnih vektora [14].

Regresija glavnih komponenti daje pristrasnu ocenu parametara. Kada bi se umesto prvih m glavnih komponenti koristili sve glavne komponente, tada bi važilo da je

$$\widehat{\boldsymbol{\beta}}_{PCR} = \widehat{\boldsymbol{\beta}}_{OLS}\ .$$

5.3 Rubna regresija

Druga metoda za rešavanje problema multikolinearnosti među regresijama jeste rubna regresija, koja je kreirana od strane Hoerl-a. Kada postoji multikolinearnost, matrica $\boldsymbol{X'X}$ teži ka singularitetu. Pošto je $Var(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X'X})^{-1}$, a dijagonalni elementi matrice $(\boldsymbol{X'X})^{-1}$ imaju veliku vrednost, sledi da je vrednost varijanse $\widehat{\boldsymbol{\beta}}$ veoma velika. To dovodi do nestabilne ocene $\boldsymbol{\beta}$, ukoliko se koristi regresija običnih najmanjih kvadrata.

Ideja rubne regresije je da se uvede pozitivna konstanta θ i da se ona doda dijagonalnim elementima matrice $\boldsymbol{X'X}$ (pod pretpostavkom da je $\boldsymbol{X'X}$ standardizovana matrica). Ovom transformacijom se omogućuje da matrica $\boldsymbol{X'X}$ postane nesingularna matrica. Prema tome, ocena je jednaka

$$\widehat{\boldsymbol{\beta}}_{RIDGE} = (\boldsymbol{X'X} + \theta\boldsymbol{I})^{-1}\boldsymbol{X'y}\,,$$

gde je $\boldsymbol{I}$ jedinična matrica, dimenzije $p \times p$, a $\boldsymbol{X'X}$ je korelaciona matrica nezavisnih promenljivih. Vrednost konstante θ ima vrednost u intervalu $(0,1)$. Ukoliko je $\theta = 0$, tada važi da je

$$\widehat{\boldsymbol{\beta}}_{RIDGE} = \widehat{\boldsymbol{\beta}}_{OLS}\,.$$

Cilj rubne regresije je određivanje najbolje vrednosti konstante θ kojom će se maksimizovati predikciona moć modela. U literaturama postoje mnogi algoritmi za određivanje vrednosti θ. Jedan od načina jeste da se grafički prikaže $\widehat{\boldsymbol{\beta}}_{RIDGE}$ u odnosu na promenljivu θ. Za onu vrednost θ, za koju $\widehat{\boldsymbol{\beta}}_{RIDGE}$ bude pokazivao najveću stabilnost, biće izabrana [14].

5.4 Regresija parcijalnih najmanjih kvadrata

5.4.1 Uvod

Regresija parcijalnih najmanjih kvadrata (partial least squares regression-PLS) je metoda zasnovana na višestrukoj regresiji i analizi glavnih komponenata, koju je tokom šezdesetih godina prošlog veka ustanovio Herman Wold, za upotrebu u hemometriji. Za razliku od regresije glavnih komponenti, regresija parcijalnih najmanjih kvadrata pronalazi skup glavnih komponenti (latentni vektori) iz $\boldsymbol{X}$, koje su relevatne i za matricu $\boldsymbol{Y}$. Konkretno, skup glavnih komponenti (latentni vektori) istovremeno razlaže matrice $\boldsymbol{X}$ i $\boldsymbol{Y}$, pod uslovom da komponente objašnjavaju što je moguće veću kovarijansu između tih matrica. Zato, rešenja dobijena regresijom parcijalnih najmanjih kvadrata imaju bolju predikcionu moć. Uopšteno, osnovne prednosti ove metode su:

- može se koristiti na multikolinearnim podacima,
- može uključivati veliki skup nezavisnih promenljivih i
- može se istovremeno modelovati nekoliko zavisnih promenljivih.

Prvu ideju metode je dao švedski istraživač Herman Wold 1970. godine, da bi je u potpunosti razvio 1977. godine [9]. Nakon dve godine, Gerlach, Kowalski i Herman Wold su napisali prvu studiji o regresiji parcijalnih najmanjih kvadrata i dali prvu aplikaciju njene primene u hemiji [10].

Pre samog razvoja modela bitno je napomenuti da se notacija i mnogi nazivi matrica bitno razlikuju od prethodno definisanih. Regresija parcijalnih najmanjih kvadrata razlaže matrice $\boldsymbol{X}$ i $\boldsymbol{Y}$ na proizvod ortogonalnih i skupa specifičnih koeficijenata (loadings). Tako, nezavisna promenljiva $\boldsymbol{X}$ se razlaže na proizvod

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T, \qquad \boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I},$$

gde matrica $\boldsymbol{T}$ predstavlja skor matricu, a matrica $\boldsymbol{P}$ matricu koeficijenata. Ocena matrice $\boldsymbol{Y}$ je jednaka

$$\widehat{\boldsymbol{Y}} = \boldsymbol{TBC},$$

gde je $\boldsymbol{B}$ dijagonalna matrica sa regresionima koeficijentima na dijagonali. Kolone matrice $\boldsymbol{T}$ se nazivaju latentni vektori. Kada je broj latentnih promenljivih jednak rangu matrice $\boldsymbol{X}$, dobija se tačna dekompozicije matrice $\boldsymbol{X}$.

Latentni vektori se mogu birati na nekoliko načina. U suštini, u prethodnom izlaganju bilo koji skup ortogonalnih vektora koji generišu prostor vektora kolona matrice $\boldsymbol{X}$ može se koristiti kao skup latentnih promenljivih (matrica $\boldsymbol{T}$). Za određivanje matrice $\boldsymbol{T}$, potrebni su dodatni uslovi. Ideja je pronalaženje dva skupa težinskih koeficijenata $\boldsymbol{w}$ i $\boldsymbol{c}$ sa ciljem da se formiraju linearne kombinacije vektora kolona matrica $\boldsymbol{X}$ i $\boldsymbol{Y}$ respektivno, tako da njihova kovarijansa bude maksimalna. Konkretno, cilj je pronalaženje prvog para vektora $\boldsymbol{t} = \boldsymbol{Xw}$ i $\boldsymbol{u} = \boldsymbol{Yc}$ (pod uslovima $\boldsymbol{w}^T\boldsymbol{w} = 1$ i $\boldsymbol{t}^T\boldsymbol{t} = 1$) tako da $\boldsymbol{t}^T\boldsymbol{u}$ bude maksimalno. Nakon izračunavanja prvog latentnog vektora, on se oduzima od obe matrice (i $\boldsymbol{X}$ i $\boldsymbol{Y}$), i procedura se ponavlja sve dok matrica $\boldsymbol{X}$ ne postane nula matrica [10],[14].

5.4.2 Algoritam PLS regresije

Pre nego što se definiše sam algoritam, potrebno je uvesti određene notacije. Neka su matrice $\boldsymbol{X}$ i $\boldsymbol{Y}$ zamenjene matricama $\boldsymbol{E}$ i $\boldsymbol{F}$ na sledeći način:

$$\boldsymbol{E} = \boldsymbol{X} \text{ i } \boldsymbol{F} = \boldsymbol{Y}.$$

Podrazumeva se da su matrice centrirane i normalizovane (po kolonama). Suma kvadrata tih matrica se označava sa SS_X i SS_Y. Algoritam se sastoji iz dva dela.

1. Pre početka interativnog postupka, potrebno je dodeliti početnu vrednost vektoru $\boldsymbol{u}$, koja se dodeljuje proizvoljno. Dakle, za početni vektor $\boldsymbol{u}$, interativni postupak glasi (simbol $\propto$ označava dobijanje normalizovanog rezultata):

i. $\boldsymbol{w} \propto \boldsymbol{E}^T\boldsymbol{u}$ (ocena $\boldsymbol{X}$ težinskih koeficijenata),
ii. $\boldsymbol{t} \propto \boldsymbol{E}\boldsymbol{w}$ (ocena $\boldsymbol{X}$ faktorski skorova),
iii. $\boldsymbol{c} \propto \boldsymbol{F}^T\boldsymbol{t}$ (ocena $\boldsymbol{Y}$ težinskih koeficijenata),
iv. $\boldsymbol{u} = \boldsymbol{F}\boldsymbol{c}$ (ocena $\boldsymbol{Y}$ skoreva).

Ukoliko $\boldsymbol{t}$ nije konvergirao, postupak se vraća na korak i., a ukoliko $\boldsymbol{t}$ je kovergirao, tada se izračunava vrednost b, koja se koristi za predikciju $\boldsymbol{Y}$ iz $\boldsymbol{t}$ kao $b = \boldsymbol{t}^T\boldsymbol{u}$ i onda se izračunavju faktorski koeficijenti za $\boldsymbol{X}$ kao $\boldsymbol{p} = \boldsymbol{E}^T\boldsymbol{t}$.

2. Sledeći korak jeste revidiranje matrica $\boldsymbol{E}$ i $\boldsymbol{F}$ na sledeći način:

$$\boldsymbol{E} = \boldsymbol{E} - \boldsymbol{t}\boldsymbol{p}^T \quad \text{i} \quad \boldsymbol{F} = \boldsymbol{F} - b\boldsymbol{t}\boldsymbol{c}^T .$$

Dobijene vrednosti vektora $\boldsymbol{t}$, $\boldsymbol{u}$, $\boldsymbol{w}$, $\boldsymbol{c}$ i $\boldsymbol{p}$ se ubacuju u odgovarjuće matrice, a vrednost skalara b se upisuje se kao dijagonalni element matrice $\boldsymbol{B}$. Suma kvadrata matrice $\boldsymbol{X}$ (repektivno i $\boldsymbol{Y}$) objašnjena latentnim vektorom se računa kao $\boldsymbol{p}^T\boldsymbol{p}$ (repektivno b^2), a procenat objašnjene varijanse se dobija deljenjem obajšnjene sume kvadrata sa odgovarajućom ukupnom sumom kvadrata (tj. SS_X i SS_Y). Interativni postupak je završen, kada je matrica $\boldsymbol{E}$ jednaka nula matrici.

Predstavljeni interativni algoritam je sličan metodi poznatoj kao power metoda, koja služi za pronalaženje karakterističnih vektora. Zato regresija parcijalnih najmanjih kvadrata koristi poznate tehnike kao što je SVD dekompozicija. Na primer, korak 1. računa $\boldsymbol{w} \propto \boldsymbol{E}^T\boldsymbol{u}$ i interativnom zamenom desnog člana $\boldsymbol{u}$, moguće je dobiti sledeći niz jednačina

$$\boldsymbol{w} \propto \boldsymbol{E}^T\boldsymbol{u} \propto \boldsymbol{E}^T\boldsymbol{F}\boldsymbol{c} \propto \boldsymbol{E}^T\boldsymbol{F}\boldsymbol{F}^T\boldsymbol{t} \propto \boldsymbol{E}^T\boldsymbol{F}\boldsymbol{F}^T\boldsymbol{E}\boldsymbol{w}.$$

Ovo pokazuje da prvi težinski koeficijent $\boldsymbol{w}$ je prva desni singularni vektor matrice $\boldsymbol{X}^T\boldsymbol{Y}$. Isti argument pokazuje da prvi vektori $\boldsymbol{t}$ i $\boldsymbol{u}$ su prvi karakteristični vektori matrica $\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T$ i $\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{X}^T$ [10].

5.4.3 Predikcija zavisnih promenljivih

Vrednost zavisne promenljive se određuje korišćenjem regresione formule $\widehat{\boldsymbol{Y}} = \boldsymbol{T}\boldsymbol{B}\boldsymbol{C}^T = \boldsymbol{X}\boldsymbol{B}_{PLS}$, gde se $\boldsymbol{B}_{PLS}$ dobija kao:

$$\boldsymbol{B}_{PLS} = (\boldsymbol{P}^{T+})\boldsymbol{B}\boldsymbol{C}^T,$$

gde je $\boldsymbol{P}^{T+}$ psudoinverzna matrica matrice $\boldsymbol{P}^T$. Ukoliko su uključene sve latentne promenljjive (od $\boldsymbol{X}$), tada je ova regresija jednaka regresiji glavnih komponenti. Suprotno, ukoliko se uključi samo određeni podskup od svih latentnih promenljivih, tada predikcija $\boldsymbol{Y}$ daje optimalnu vrednost za zadati broj predikcija.

Predikciona suma kvadrata (*prediction sum of squares*-PRESS) predstavlja formu kros validacije (*cross-validation*) koja se koristi kod regresione analize da obezbedi meru fitovanja modela na osnovu uzorka. Računa se putem formule:

$$PRESS(m) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij}^{(m)} - y_{ij})^2,$$

gde je m broj komponenti, $y_{ij}^{(m)}$ vrednost (i, j) elementa u ocenjenoj matrici $\boldsymbol{Y}$, a y_{ij} vrednost (i, j) elementa u originalnoj matrici $\boldsymbol{Y}$.

5.4.4 Primer

Ideja primera jeste predikcija subjektivnih ocena pet vrsta vina na osnovnih njihovih određenih svojstava [10]. Zavisna promenljiva koja se predviđa jeste karakteristika vina, npr. kakav ukus ima vino kada se pije uz meso, uz dezert, itd. (tabela 5.1). Nezavisnu promenljivu predstavlja matrica podataka koja meri cenu, sadržaj šećera, alkohol i kiselost vina u uzorku od 5 vrste vina (tabela 5.2).

Korišćenjem kompjuterskog softvera MATLAB, lako se izračunavaju matrice $\boldsymbol{T}$, $\boldsymbol{U}$, $\boldsymbol{P}$, $\boldsymbol{W}$, $\boldsymbol{C}$, $\boldsymbol{B}$, $\boldsymbol{B}_{PLS}$ za dva ili tri latentna vektora (tabela 5.3-5.10). Tabela 5.11 prikazuje da dva latentna vektora objašnjavaju 98% varijanse matrice $\boldsymbol{X}$ i 85% matrice $\boldsymbol{Y}$. To ukazuje da je mnogo bolja dvo-dimenzionalno rešenje (dve latentne promenljive). Ispitivanjem dvodimenzionalnih regresionih koeficijenata (tabela 5.8) pokazuje da je šećer utiče na izbor vina za dezert, a cena je negativna korelisana sa kvalitetom vina, dok je alkohol pozitivno korelisan sa kvalitetom. Tako, latentne promenljiva $\boldsymbol{t}_1$ izražava cenu, a $\boldsymbol{t}_2$ izražava sadržaj šećera.

Vino	Hedonističko	Slaže se sa mesom	Slaže se sa dezertom
1.	14	7	8
2.	10	7	6
3.	8	5	5
4.	2	4	7
5.	6	2	4

Tabela 5.1 Matrica $\boldsymbol{Y}$

Vino	Cena	Šećer	Alkohol	Kiselost
1.	7	7	13	7
2.	4	3	14	7
3.	10	5	12	5
4.	16	7	11	3
5.	13	3	10	3

Tabela 5.2 Matrica $\boldsymbol{X}$

Vino	$\boldsymbol{t}_1$	$\boldsymbol{t}_2$	$\boldsymbol{t}_3$
1.	0.4538	-0.4662	0.5716
2.	0.5399	0.4940	-0.4631
3.	0	0	0
4.	-0.4304	-0.5327	-0.5301
5.	-0.5633	0.5049	0.4217

Tabela 5.3 Matrica $\boldsymbol{T}$

Vino	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	$\boldsymbol{u}_3$
1.	1.9451	-0.7611	0.6191
2.	0.9347	0.5305	-0.5388
3.	-0.2327	0.6084	0.0823
4.	-0.9158	-1.1575	-0.6139
5.	-1.7313	0.7797	0.4513

Tabela 5.4 Matrica $\boldsymbol{U}$

	$\boldsymbol{p}_1$	$\boldsymbol{p}_2$	$\boldsymbol{p}_3$
Cena	-1.8706	-0.6845	-0.1796
Šećer	0.0468	-1.9977	0.0829
Alkohol	1.9547	0.0283	-0.4224
Kiselost	1.9874	0.0556	0.2170

Tabela 5.5 Matrica $\boldsymbol{P}$

	$\boldsymbol{w}_1$	$\boldsymbol{w}_2$	$\boldsymbol{w}_3$
Cena	-0.5137	-0.3379	-0.3492
Šećer	0.2010	-0.9400	0.1612
Alkohol	0.5705	-0.0188	-0.8211
Kiselost	0.6085	0.0429	0.4218

Tabela 5.6 Matrica $\boldsymbol{W}$

	Hedonističko	Ide sa mesom	Ide sa dezertom
Cena	-1.0607	-0.0745	0.1250
Šećer	0.3354	0.2593	0.7510
Alkohol	-1.4142	0.7454	0.5000
Kiselost	1.2298	0.1650	0.1186

Tabela 5.7 Matrica $\boldsymbol{B}_{PLS}$ na osnovu tri latentna vektora

	Hedonističko	Ide sa mesom	Ide sa dezertom
Cena	-0.2662	-0.2498	0.0121
Šećer	0.0616	0.3197	0.7900
Alkohol	0.2969	0.3679	0.2568
Kiselost	0.3011	0.3699	0.2506

Tabela 5.8 Matrica $\boldsymbol{B}_{PLS}$ na osnovu dva latentna vektora

	$\boldsymbol{c_1}$	$\boldsymbol{c_2}$	$\boldsymbol{c_3}$
Hedonističko	0.6093	0.0518	0.9672
Ide sa mesom	0.7024	-0.2684	-0.2181
Ide sa dezertom	0.3680	-0.9619	-0.1301

Tabela 5.9 Matrica $\boldsymbol{C}$

b_1	b_2	b_3
2.7568	1.6272	1.1191

Tabela 5.10 Dijagonalni elementi matrice $\boldsymbol{B}$

Latentni vektor	Procenat objašnjene varijanse za $\boldsymbol{X}$	Kumulativni procenat objašnjene varijanse za $\boldsymbol{X}$	Procenat objašnjene varijanse za $\boldsymbol{Y}$	Kumulativni procenat objašnjene varijanse za $\boldsymbol{Y}$
1	70	70	63	63
2	28	98	22	85
3	2	100	10	95

Tabela 5.11 Prikaz objašnjene varijanse matrice $\boldsymbol{X}$ i $\boldsymbol{Y}$ od strane latentnih vektora

Glava 6

Primena metoda faktorske analize u hemiji

6.1 Hemometrija

6.1.1 Uvod

Hemometrija je multidisciplinarna naučna oblast koja na hemijske podatke primenjuje matematičke, informatičke i statističke metode kako bi se omogućilo efikasno i jednostavno određivanje fizičko-hemijskih osobina, predviđanje ponašanja ili klasifikovanje jedinjenja u neku od kategorija na osnovu podataka o njihovoj molekulskoj strukturi. Naziv hemometrija je prvi put bio upotrebljen 1974. godine [12]. Izveden je prema analogiji sa nazivom naučne discipline ekonometrija, koja ukazuje na obradu ekonomskih podataka. Za uspostavljanje matematičke korelacije između strukture i osobina nekog jedinjenja potrebno je sve informacije pretvoriti u numeričke vrednosti. Matematički obrazac se modeluje pomoću osnovnog skupa ulaznih podataka dobijenih eksperimentalnim putem ili kompjuterskim algoritmom na osnovu molekulske strukture. Dobijeni matematički model omogućava proučavanje odnosa između strukture i osobina molekula, a time i predviđanje ponašanja novosintetisanih jedinjenja i klasifikovanje uzoraka u jednu od kategorija. Kroz matematički model takođe je moguće posmatranje biolških sistema i reakcija koje se u njima odvijaju. Modeli koji daju kvantitativnu vezu između strukture molekula i njihove biološke aktivnosti se skraćeno nazivaju QSAR (od *Quantitative Structure Activity Relationships)*. Prve veze između strukture i osobina potiče iz 1863. godine [13]. Tada je uočena veza između rastvorljivosti primarnih alkohola i njihovog toksičnog dejstva, i postavljen je prvi aksiom modelovanja strukture i toksičnosti. Danas se QSAR najčešće primenjuju u sledećim oblastima hemije: medicinska hemija, toksikologija, agrohemija, zaštita životne sredine. Vremenom QSAR modeli su se proširili i na druge oblasti hemije, i postali su deo jedne šire oblasti SPR (od *Structure Property Relationship*), koja obuhvata kako biološke tako i fizičko-hemijske osobine molekula. Široka primena računara krajem 80-ih godina prošlog veka omogućila je hemometrijski pristup u rešavanju problema iz oblasti primenjene hemije [15].

QSAR modelovanje je poteklo iz oblasti toksikologije, a bazira se na pretpostavci da struktura molekula, odnosno njegova geometrija utiču na fizičke, hemijske i biološke osobine supstanci. Veza između matematike i hemije formulisana je u funkcionalnoj zavisnosti ($\boldsymbol{f}$) biološke aktivnosti različitih supstanci ($\boldsymbol{A}$) od strukture njihovih molekula ($\boldsymbol{D}$), odnosno:

$$\mathbf{A} = \boldsymbol{f}(\boldsymbol{D}).$$

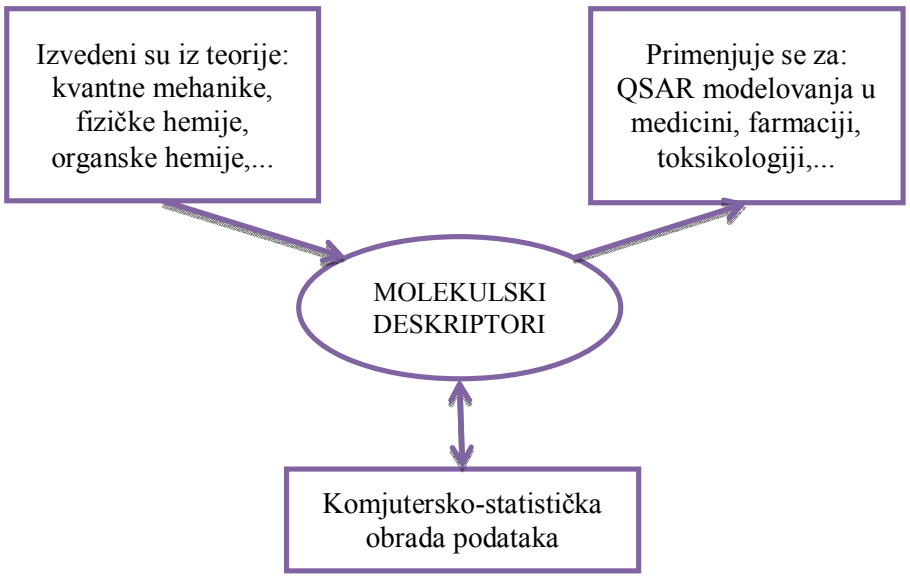
Obično se razmatra jedan osnovni moelkul i njegovi derivati. Promena u strukturi običnog molekula ($\Delta\boldsymbol{D}$) utiče na promene u biološkoj aktivnosti ($\Delta\mathbf{A}$). Ova matematička relacija se smatra opštom matematičkom formulacijom QSAR pristupa [15].

Za oko 30 godina, hemometrija je razvila različite metode koje omogućavaju dobijanje pouzdanih modela za predviđanje nepoznatih vrednosti, za reprodukciju poznatih eksperimentalnih podataka i klasifikaciju jedinjenja. Poslednjih nekoliko godina raste interesovanje za modele koji mogu da omoguće efikasnu i pouzdanu procenu osobina supstance. Hemometrija podrazumeva spajanje različitih izvora informacija, da bi se podaci pretvorili u informacije, a informacije u znanje neophodno za donošenje brzih i boljih odluka u oblasti indtifikaciji supstanci. Oblast hemometrije obuhvata dizajn, organizaciju, upravljanje, pronalaženje, analizu, širenje, vizuelizaciju, kao i upotrebu hemijskih informacija. Molekulski deskriptori imaju važnu ulogu u svim navedenim procesima kao osnova za pretvaranje hemijskih informacija u numeričke vrednosti pogodne za informatičku obradu [15].

6.1.2 Molekulski deskriptori

Prema Livingstonu, molekulski deskriptori su brojevi kojima se numerički kvantifikuju neka osobina molekula [13]. Molekulski deskriptori predstavljaju konačni rezultat logičkog i matematičkog postupka koji pretvara hemijske informacije, kodirane u simboličkom prikazu molekula, u numeričke vrednosti. Do danas je definisano više od 5000 molekulskih deskriptora, što potvrđuje njihov naučni značaj i veliku primenu.

Molekulski deskriptori su izvedeni na osnovu nekoliko različitih teorija, kao što su kvantna hemija, organska hemija, teorija informacija i teorija grafova. Koriste se za modelovanje različitih osobina supstance u oblasti: analitičke hemije, farmacije, toksikologije, fizičke hemije, medicine i zaštite životne sredine. Obrađuju se brojnim statističkim, hemometrijskim i hemoinformacionim metodama u cilju dobijanja pouzdanih procena osobina molekula i indetifikovanja strukturnih karakteristika.



Grafik 6.1: *Veza deskriptora i matematike i hemije*

Molekulski deskriptori se mogu klasifikovati u nekoliko grupa:

1. **u zavisnosti od uptrebljene vrste prikaza molekula i definisanog algoritma za izračunavanje**, teorijski deskriptori mogu biti 1D, 2D i 3D molekulski deskriptori (tabela 6.1);

2. **prema vrsti osobina koju opisuju**, molekusku deskriptorim mogu biti topološki, fizičko-hemijski i kvantno-mehanički;

3. **prema delu molekula na koji se odnosi**, molekulski deskriptori mogu biti globalni i lokalni;

4. **prema načinu dobijanja**, molekulski deskriptori mogu biti eksperimentalni i teorijski.

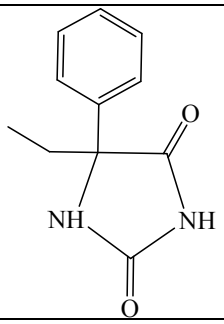
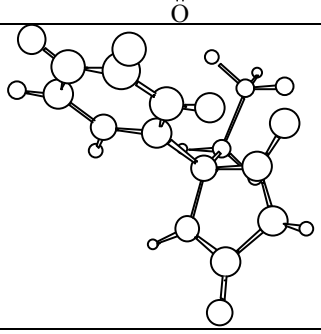
	Prikaz molekula	Deskriptori
1D	$C_{11}H_{12}N_2O_2$	Molekulska masa Atomski deskriptori
2D		Fragmentacioni doprinosi Topološki indeks
3D		Površina molekula Zapremina molekula Interakcione energije Kvantno-mehanički deskriptori

Tabela 6.1: *Primer deskriptora koji se mogu izračunati na osnovu različitog prikaza molekula nirvanola*

6.1.3 Osnovni principi postavljanja matematičkog QSAR modela

Da bi se došlo do adekvatnog matematičkog QSAR modela, prethodno je neophodno definisati:

- podatke dobijene eksperimentalnim merenjem biološke aktivnosti ili neke druge fizičko-hemijske eksperimentalno merljive osobine molekula (zavisne promenljive),
- podatke o strukturi ili osobinama ovih jedinjenja - molekulski deskriptori (nezavisne promenljive),
- i statistički model za pronalaženje matematičke relacije između ove dve grupe podataka.

1. Ulazni podaci

Ograničavajući faktor QSAR modelovanja je dostupnost eksperimentalnih podataka visokog kvaliteta. Da bi se razvio model, važno je da ulazni podaci budu tačni i precizni, jer model je onoliko validan koliko i podaci iz kojih je dobijen. Podaci za model koriste se iz literature ili se posebno generišu za QSAR analizu. Oni mogu da se odnose na seriju strukturno sličnih ili potpuno strukturno različitih molekula, koji čak mogu da pripadaju i različitim grupama hemijskih jedinjenja. QSAR modeli su definisani i ograničeni prirodom i kvalitetom ulaznih podataka, koji su korišćeni za njegovo razvijanje [15].

2. Izbor relevatnih deskriptora za QSAR

Poseban problem koji se javlja u hemometriji jeste kako od velikog broja raspoloživih deskriptora odabrati određeni broj njih i uključiti ih u model. Kod izbora deskriptora prvo se vrši njihovo preliminarna selekcija. Od matematičkih metoda za izbor deskriptora nejčešće se koristi: višestruka linearna regresija (stepwise), metoda glavnih komponenti (PCA), faktorska analiza (FA), metoda grupisanja (HA) i dr. Većina deskriptora su, po prirodi, međusobno korelisani, što znači da sadrže suviše informacija. Na primer, molekulska masa, površina molekula i molarna refraktivnost su uglavno jako korelisani pa je u tom slučaju opravdano uvrstiti samo jednu od njih u QSAR model. Još jedna od procedura verifikovanja podataka, koji bi trebao izvršiti na svakom deskriptoru, je ispitivanje respodele vrednosti skupa deskriptora. U mnogim slučajeviima je poželjno da vrednost deskriptora prati neku posebnu raspodelu, npr. noramalnu raspodelu [15].

3. Izbor statističkog modela

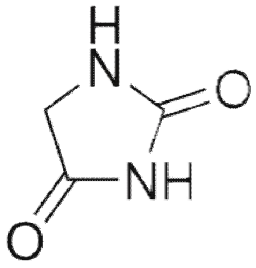
Postoje mnogi statistički modeli koji se mogu koristiti u predikciji vrednosti određene osobine jedinjenja. Na osnovu tipa ulaznih podataka, deskriptora i rezultata koji želi da se dobije, bira se statistička metoda [15]. Neki od najznačajnijih su:

- Višestruka linearna regresija,
- Metoda regresije najmanjih kvadrata,
- Neuralne mreže,
- Algoritam protpornih vektora,
- Stabla odluke,
- Genetski algoritmi.

6.2 Eksperimentalni deo, rezultati i diskusije

6.2.1 Hidantoin

Hidantoin (slika 6.1) je petočlani ciklični molekul sa dva atoma azota u prstenu i dva atoma kiseonika povezanih dvostrukom vezom za prsten. Prvi put je sintetisan 1861. godine.



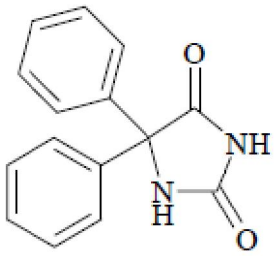
Slika 6.1: *Hidantoin*

Nacionalna medicinska bibiloteka SAD (National Library of Medicine) u svom registru lekova sadrži pet grupa derivata hindatoina, a to su:

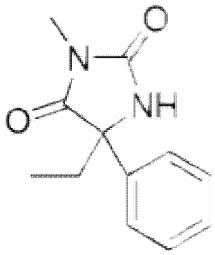
- Alantoin,
- Dantrolen,
- Mefenitoin,
- Fenitoin i
- Tiohidantoini.

Lekovi iz grupe hindatoina imaju anestetičko, antiaritmijsko i antikonvulzivno delovanje [15].

Dva najpoznatija hidantoina su Fenitoin (5,5-difenilhidatoin) (slika 6.2) i Mefenitoin (5-etil-3-metil-5-fenilhidantoin) (Slika 6.3). Fenitoin je jedan od najstarijih lekova za lečenje epilepsije i veoma je efikasan za kontrolisanje niza privremenih poremećaja. Mefenitoin je otkriven 10 godina nakon Fenitoina, krajem 1940. godine. Značajan metabolit Mefenitoina je Nirvanol (5-etil-5-fenilhidantoin) koji je prvi hidantoin korišćen kao hipnotik. Mefenitoin se više ne koristi u terapijama za lečenje, zbog svoje toksičnosti, ali se ipak još proučava [15].



Slika 6.2. *Struktura Fenitoina*



Slika 6.3 *Struktura Mefenitoina*

U radu se koriste eksperimentalni podaci 24 derivata hidantoina, tačnije derivati dva leka: Fenitoina (14 jedinjenja) i Mefenitoina (10 jedinjenja).

6.2.2 Primena faktorske analize na molekulske deskriptore

Za svaki od 24 ispitivanih molekula, izračunati su sledeći moelkulski deskriptori: Uc, Ui, Hy, AMR, TPSA(NO), TPSA(Tot), MLOGP, MLOGP2, ALOGP, ALOGP2, Satot, Saacc, Sadon, Vx (oznake deskriptora objašnjene su u tabeli 6.2, a numeričke vrednosti su date u tabeli 6.3). Zapravo, tabela 6.3 predstavlja matricu podataka nad kojom se primenjuju metode faktorske analize. Cilj je da se data matrica smanji na matricu manje dimenzije koju će sačinjavati faktori, koji su međusobno nekorelisani. Dakle, s obzirom da podaci sadrže deskriptore koji imaju istu ili približno istu brojnu vrednost za određenu grupu deskriptora, oni će putem metoda faktorske analize biti smanjene za broj promenljivih (deskriptora), a time će se rešiti i problem singularnosti matrice molekulskih deskriptora (problem multikolinearnosti). Za izračunavanje vrednosti matrice faktora, koeficijenata, kreiranje grafika i dr. korišćen je statistički softverski paket STATISTICA.

Oznaka	Engleski naziv	Srpski naziv
Uc	Unsaturation count	Nezasićenost
Ui	Unsaturation index	Indeks nezasićenosti
Hy	Hydrophilic factor	Faktor hidrofilnosti
AMR	Ghose-Crippen molar refractivity	Molarna refraktivnost (izračunata po Ghose-Crippen modelu)
TPSA(NO)	Topological polar surface area using N,O polar contributions	Topološka polarna površina izračunata na osnovu polarnih doprinosa atoma N iO
TPSA(Tot)	Topological polar surface area using N,O,S,P polar contributions	Topološka polarna površina izračunata na osnovu polarnih doprinosa atoma N,O,S i P
MLOGP	Moriguchi octanol-water partition coeff. (logP)	Podeoni koeficijent oktanol-voda izračunat po Moriguchi logP modelu
MLOGP2	Squared Moriguchi octanol-water partition coeff. (logP^2)	Kvadrat Moriguchijevog logP podeonog koeficijenta
ALOGP	Ghose-Crippen octanol-water partition coeff. (logP)	Podeoni koeficijen oktanol-voda izračunat po Ghose-Crippen-u
ALOGP2	Squared Ghose-Crippen octanol-water partition coeff. (logP^2)	Kvadrat Ghose-Crippen-novog podeonog koeficijenta
SAtot	Total surface area from P_VSA-like descriptors	Ukupna površina (izračunata na osnovu P_VSA descriptora)
SAacc	Surface area of acceptor atoms from P_VSA-like descriptors	Površina akceprtorskih atoma (izračunata na osnovu P_VSA descriptora)
SAdon	Surface area of donor atoms from P_VSA-like descriptors	Površina donorskih atoma (izračunata na osnovu P_VSA descriptora)
Vx	McGowan volume	McGowan zapremina (molekula)

Table 6.2: *Nazivi molekulskih deskriptora*

	Uc	Ui	Hy	AMR	TPSA(NO)	TPSA(Tot)	MLOGP	MLOGP2	ALOGP	ALOGP2	SAtot	SAacc	SAdon	Vx
1	3,17	2,585	0,496	54,641	58,2	58,2	0,962	0,926	1,394	1,944	274,929	86,311	36,022	256,362
2	3,17	2,585	-0,229	59,538	49,41	49,41	1,243	1,545	1,6	2,56	303,753	71,425	18,011	279,767
3	3,17	2,585	-0,256	64,286	49,41	49,41	1,515	2,295	1,949	3,798	329,829	71,425	18,011	303,173
4	3,17	2,585	-0,28	68,81	49,41	49,41	1,778	3,163	2,473	6,113	355,906	71,425	18,011	326,578
5	3,17	2,585	-0,28	68,704	49,41	49,41	1,778	3,163	2,326	5,412	360,091	71,425	18,011	326,578
6	3,322	2,807	-0,28	68,7	49,41	49,41	1,696	2,876	2,217	4,914	346,247	71,425	18,011	319,435
7	3,17	2,585	-0,301	73,411	49,41	49,41	2,035	4,14	2,929	8,577	381,983	71,425	18,011	349,983
8	3,17	2,585	-0,301	73,282	49,41	49,41	2,035	4,14	2,792	7,794	385,739	71,425	18,011	349,983
9	3,17	2,585	-0,301	73,228	49,41	49,41	2,035	4,14	2,85	8,123	386,168	71,425	18,011	349,983
10	3,907	3,17	-0,354	84,15	49,41	49,41	2,525	6,375	3,184	10,135	395,113	71,425	18,011	380,731
11	3,907	3,17	0,341	69,981	58,2	58,2	1,795	3,221	2,105	4,431	314,136	86,311	36,022	310,515
12	3,907	3,17	-0,32	74,878	49,41	49,41	2,044	4,179	2,311	5,341	342,96	71,425	18,011	333,92
13	3,907	3,17	-0,338	79,626	49,41	49,41	2,287	5,232	2,66	7,075	369,036	71,425	18,011	357,326
14	3,907	3,17	-0,354	84,15	49,41	49,41	2,525	6,375	3,184	10,135	395,113	71,425	18,011	380,731
15	3,907	3,17	-0,354	84,044	49,41	49,41	2,525	6,375	3,037	9,225	399,298	71,425	18,011	380,731
16	4	3,322	-0,354	84,041	49,41	49,41	2,449	5,999	2,928	8,572	385,454	71,425	18,011	373,588
17	3,907	3,17	-0,369	88,751	49,41	49,41	2,757	7,603	3,64	13,247	421,19	71,425	18,011	404,136
18	3,907	3,17	-0,369	88,622	49,41	49,41	2,757	7,603	3,503	12,27	424,946	71,425	18,011	404,136
19	3,907	3,17	-0,369	88,568	49,41	49,41	2,757	7,603	3,561	12,681	425,375	71,425	18,011	404,136
20	4,392	3,585	-0,408	99,491	49,41	49,41	3,205	10,269	3,895	15,167	434,32	71,425	18,011	434,884
21	3,907	3,17	-0,326	85,479	58,64	58,64	2,143	4,594	2,511	6,306	410,229	82,425	18,011	390,482
22	4,392	3,585	-0,394	105,344	58,64	58,64	3,038	9,228	3,746	14,032	475,513	82,425	18,011	468,04
23	4	3,322	-0,301	85,131	75,71	75,71	2,108	4,442	2,133	4,55	414,43	107,57	18,011	393,09
24	4,459	3,7	-0,382	99,72	66,48	66,48	3,02	9,118	3,759	14,133	438,522	96,57	18,011	437,492

Table 6.3: *Vrednost molekulskih deskriptora za 24 derivata hindatoina*

1. Eliminacija suvišnih deskriptora putem faktorske analize

Value	Eigenvalues (faktorska analiza) Extraction: Principal factors (comm.=multiple R-square)		
	Eigenvalue	% Total variance	Cumulative %
1	8,221719	58,72656	58,72656
2	3,426410	24,47436	83,20092
3	1,084476	7,74626	90,94718

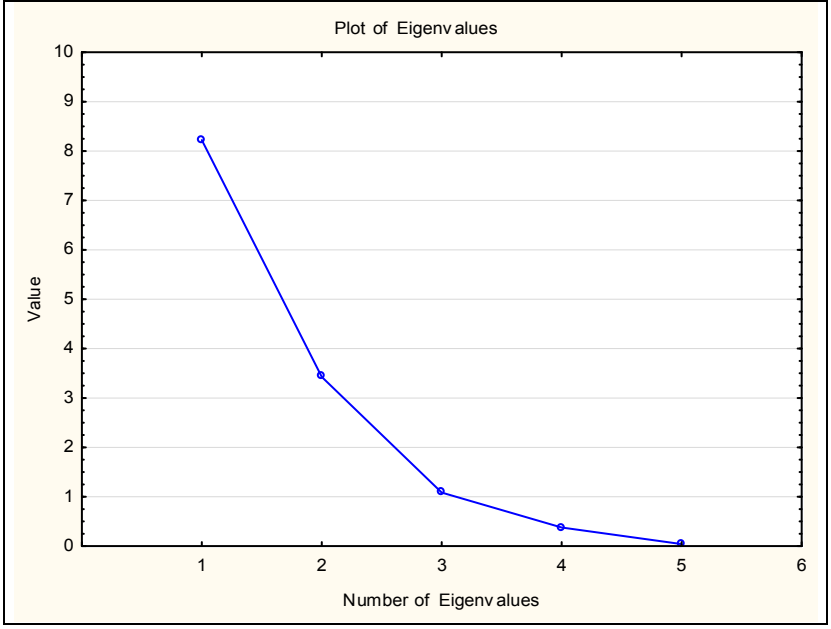
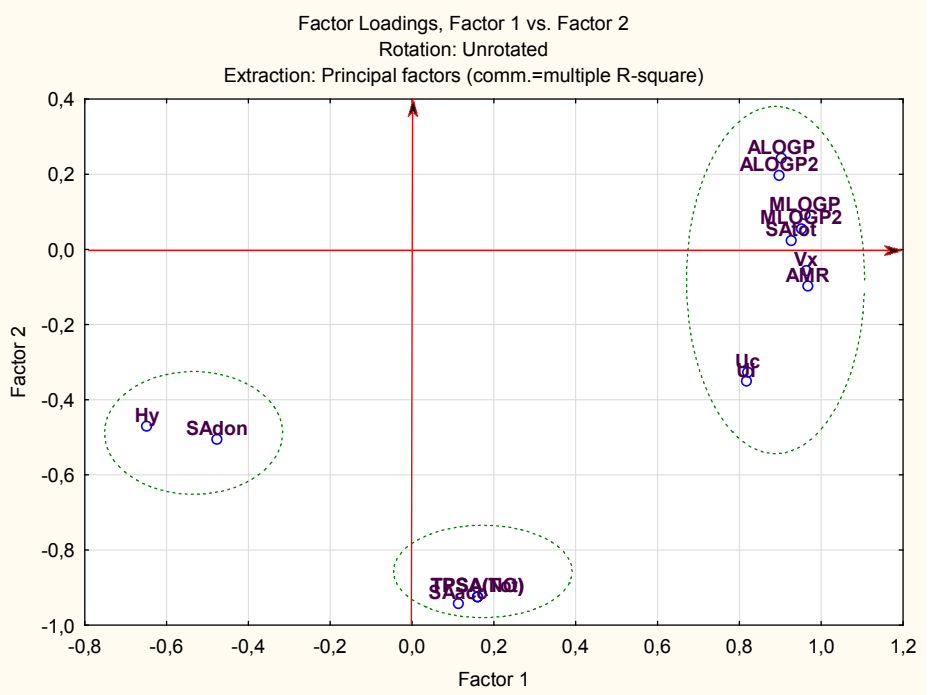


Table 6.4: *Vrednost karakterističnih korena i varijanse*



Grafik 6.2: *Grafički prikaz molekulskih deskriptora unutar sistema faktorskih osa (Factor 1 i Factor 2)*

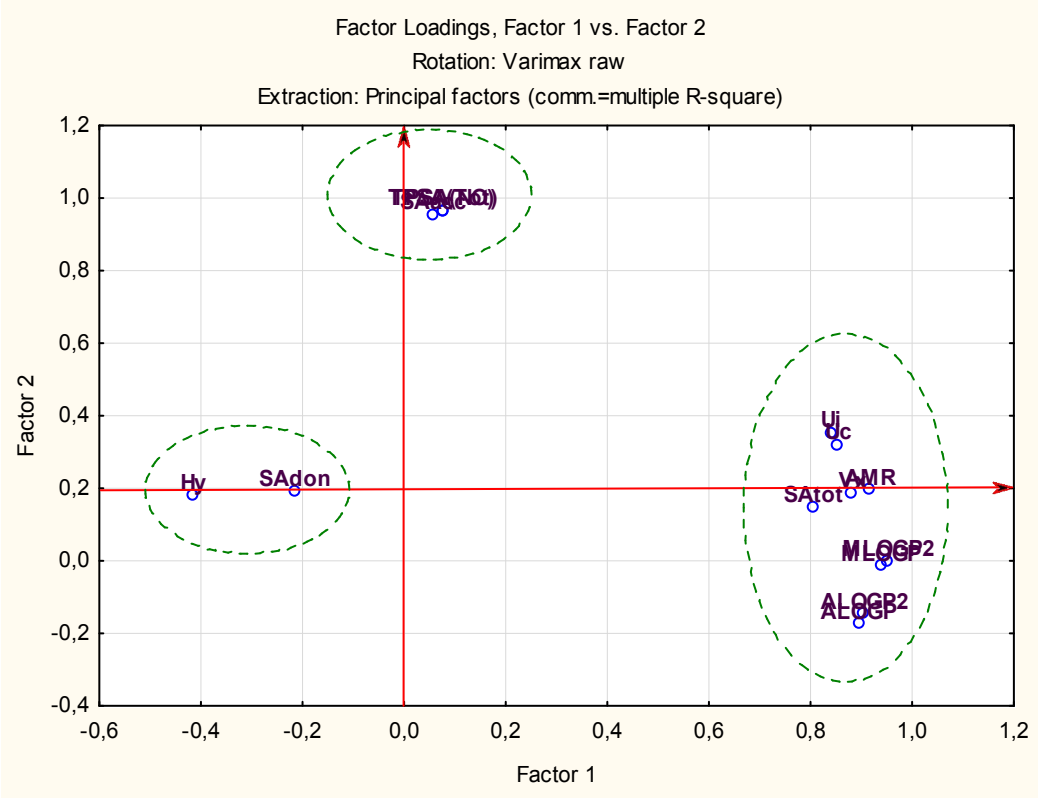
Variable	Factor Loadings (Unrotated) (faktorska analiza) Extraction: Principal factors (comm.=multiple R-square) (Marked loadings are >,700000		
	Factor 1	Factor 2	Factor 3
Uc	0,820627	-0,326547	0,229366
Ui	0,817197	-0,350779	0,196683
Hy	-0,648718	-0,470773	0,529966
AMR	0,967180	-0,097540	0,024136
TPSA(NO)	0,159835	-0,924485	-0,246464
TPSA(Tot)	0,159835	-0,924485	-0,246464
MLOGP	0,960177	0,092019	0,118595
MLOGP2	0,951668	0,055174	0,178007
ALOGP	0,902405	0,242527	0,153331
ALOGP2	0,896946	0,197292	0,196775
SAtot	0,926209	0,023858	-0,174816
SAacc	0,113094	-0,942193	-0,180351
SAdon	-0,477174	-0,505481	0,644642
Vx	0,963486	-0,056385	-0,060125
Expl.Var	8,221719	3,426410	1,084476
Prp.Totl	0,587266	0,244744	0,077463

Table 6.5: *Matrica korelacija između faktora i molekulskih deskriptora*

Case	Factor Scores (faktorska analiza) Rotation: Unrotated Extraction: Principal factors (comm.=multiple R-square)		
	Factor 1	Factor 2	Factor 3
1	-2,21703	-1,66152	1,82719
2	-1,53262	0,42951	-0,78849
3	-1,17687	0,50642	-0,73888
4	-0,77110	0,61284	-0,59659
5	-0,80221	0,59152	-0,66885
6	-0,82877	0,47494	-0,52159
7	-0,36811	0,71242	-0,44752
8	-0,40060	0,69094	-0,51960
9	-0,38457	0,70112	-0,49623
10	0,45362	0,48669	0,33103
11	-1,02178	-1,74829	2,42979
12	-0,34596	0,29922	0,01445
13	0,02609	0,37624	0,12335
14	0,45362	0,48669	0,33103
15	0,41784	0,46272	0,25045
16	0,35790	0,37488	0,32753
17	0,87574	0,59041	0,53454
18	0,83892	0,56647	0,45474
19	0,85677	0,57768	0,48137
20	1,62725	0,45815	1,19397
21	0,21307	-0,71937	-0,91595
22	1,77905	-0,64709	0,11124
23	0,28890	-2,91650	-2,38738
24	1,66084	-1,70608	-0,32958

Table 6.6: *Matrica faktora*

Korišćenjem faktorske analize nad podacima iz tabele 6.3, dobijene su neophodni podaci za analizu (tabele 6.4-6.6 i grafik 6.2). Prilikom određivanaj broja faktora koji opisuju dati model, posmatraju se karakteristični koreni koji imaju vrednost veću od 1. Broj takvih karakterističnih korena je tri, tako da su definisana tri faktora, i oni objašnjavaju 90,94% varijanse (tabela 6.4). Takođe, grafik 6.2 daje isti zaključak - da postoje tri grupe molekulski deskriptora koji su međusobno jako korelisane. U tabeli 6.5, prikazane su korelacija između molekulskih deskriptora i faktora i zaključuje se da su molekulski deskriptori Uc, Ui, AMR, MLOGP, MLOGP2, ALOGP, ALOGP2, Satot i Vx korelisani sa prvim faktorom, TPSA(NO), TPSA(Tot) i Saacc korelisani sa drugim faktorom, a za preostala dva, Hy i Sadon, ne može se dati relevantan zaključak. U ovakvim slučajevim se obično primenjuje rotacija kako bi se došlo do intepretativnijeg rezultata. Primeniće se rotacija Varimax raw i dobijaju se novi rezultati, u nastavku.



Grafik 6.3: *Grafički prikaz molekulskih deskriptora unutar sistema faktorskih osa (Factor 1 i Factor 2) nakon rotacije*

Variable	Factor Loadings (Varimax raw) (faktorska analiza) Extraction: Principal factors (comm.=multiple R-square) (Marked loadings are >,700000)		
	Factor 1	Factor 2	Factor 3
Uc	0,853753	0,318555	-0,047951
Ui	0,839465	0,352087	-0,029552
Hy	-0,415283	0,178359	-0,847971
AMR	0,915842	0,193964	0,262956
TPSA(NO)	0,077186	0,964280	-0,071917
TPSA(Tot)	0,077186	0,964280	-0,071917
MLOGP	0,939519	-0,016085	0,248025
MLOGP2	0,952879	-0,003497	0,180034
ALOGP	0,895328	-0,174859	0,253924
ALOGP2	0,906082	-0,148550	0,197749
SAtot	0,806067	0,145636	0,466947
SAacc	0,056795	0,952319	-0,151354
SAdon	-0,213967	0,190487	-0,903716
Vx	0,882295	0,184617	0,350103
Expl.Var	7,347203	3,205951	2,179451
Prp.Totl	0,524800	0,228996	0,155675

Table 6.7: *Matrica korelacija između faktora i molekulskih deskriptora nakon rotacije*

Case	Factor Scores (faktorska analiza) Rotation: Varimax raw Extraction: Principal factors (comm.=multiple R-square)		
	Factor 1	Factor 2	Factor 3
1	-1,41288	0,652928	-2,93120
2	-1,71780	-0,301684	0,33657
3	-1,36835	-0,349449	0,43924
4	-0,94006	-0,451185	0,48867
5	-0,99417	-0,409768	0,53340
6	-0,96580	-0,355705	0,35380
7	-0,51191	-0,549241	0,52878
8	-0,56724	-0,507893	0,57284
9	-0,54421	-0,523654	0,56154
10	0,53350	-0,515696	0,04059
11	-0,08119	0,661637	-3,09194
12	-0,32344	-0,323303	-0,01718
13	0,06202	-0,389948	0,03936
14	0,53350	-0,515696	0,04059
15	0,47215	-0,469456	0,09005
16	0,44427	-0,421524	-0,02901
17	0,99856	-0,634365	0,04121
18	0,93651	-0,588540	0,08965
19	0,96238	-0,606193	0,07650
20	1,93514	-0,654368	-0,33232
21	-0,11019	1,012908	0,60309
22	1,71492	0,768513	0,25403
23	-0,52138	3,578143	1,10187
24	1,46567	1,893539	0,20988

Table 6.8: *Matrica faktora nakon rotacije*

Varimax rotacijom se došlo do novih rezultata (tabele 6.7-6.8 i grafik 6.3). Iako matrica korelacija između deskriptora i faktora ima nove vrednosti, nije promenjena slika o povezanosti između molekulskih deskriptora i faktora. Šta više, sada se lepo da zaključiti da su:

- molekulski deskriptori Uc, Ui, AMR, MLOGP, MLOGP2, ALOGP, ALOGP2, Satot i Vx korelisani sa Factor 1,
- molekulski deskriptori TPSA(NO), TPSA(Tot) i Saacc su korelisani sa Factor 2 i
- molekulski deskriptori Hy i Sadon su korelisani sa Factor 3.

Dakle, definisana matrica molekulskih deskriptora (tabela 6.3) se može zameniti sa matricom faktora (tabela 6.8) i ova nova matrica objašnjava oko 91% varijanse prethodne matrice. Novom matricom je smanjena dimenzionalnost originalne matrice podataka, a i faktori su međusobno nekorelisani, matric više ne teži ka singularitetu.

2. Eliminacija suvišnih deskriptora putem analize glavnih komponenti

Value	Eigenvalues (faktorska analiza) Extraction: Principal factors (comm.=multiple R-square)		
	Eigenvalue	% Total variance	Cumulative %
1	8,659030	61,85021	61,8502
2	3,626654	25,90467	87,7549
3	1,182058	8,44327	96,1982

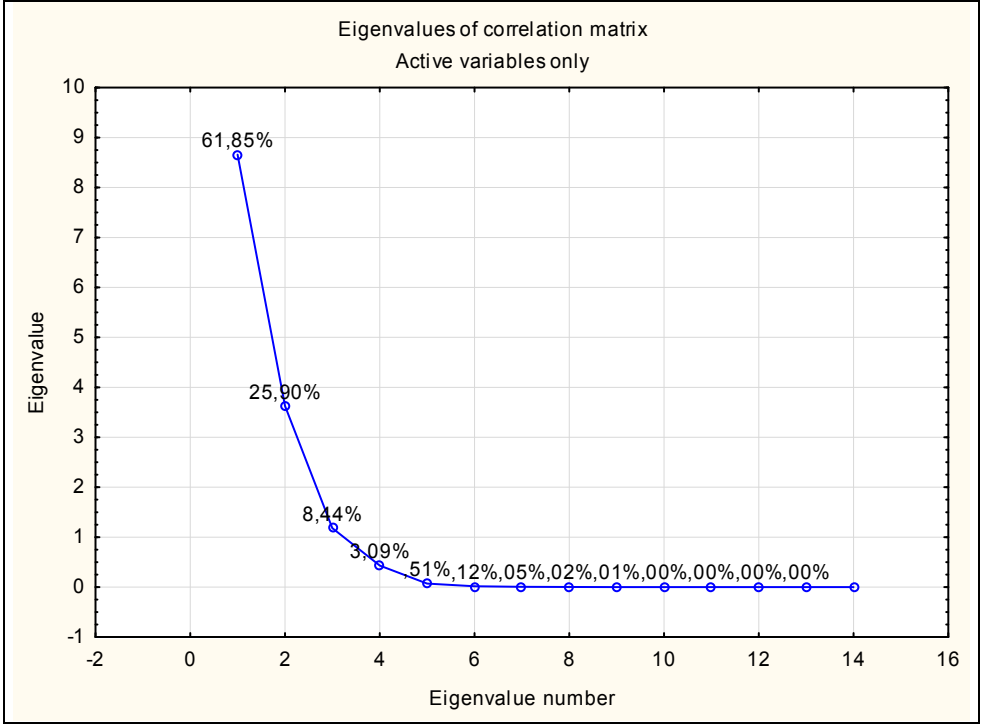
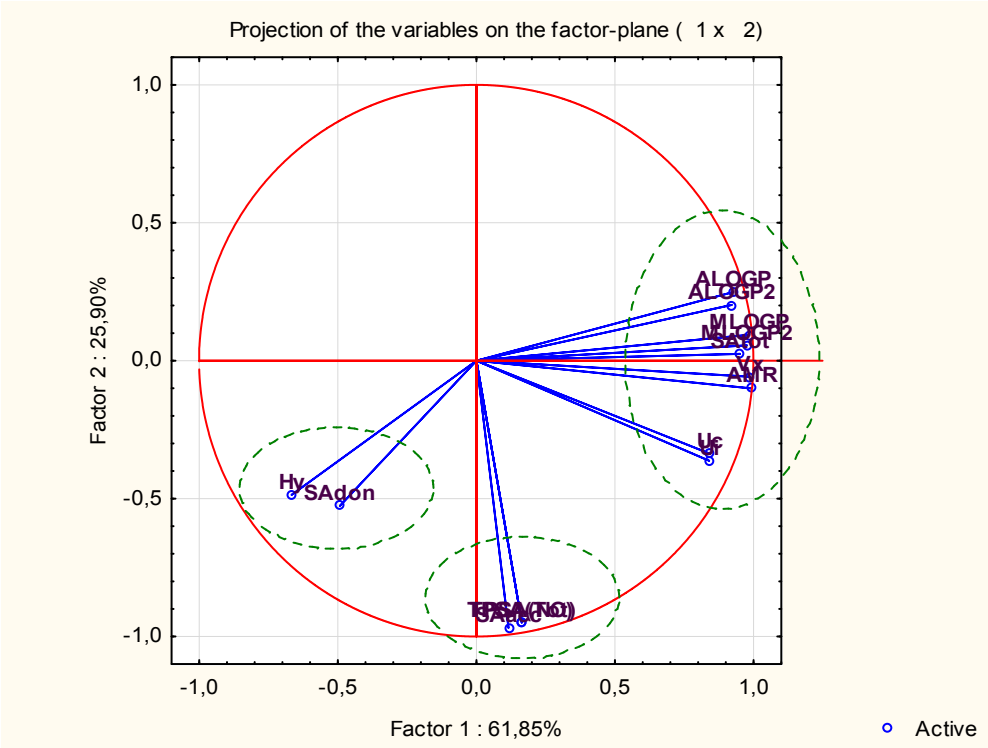


Table 6.9: *Vrednost karakterističnih korena i varijanse*

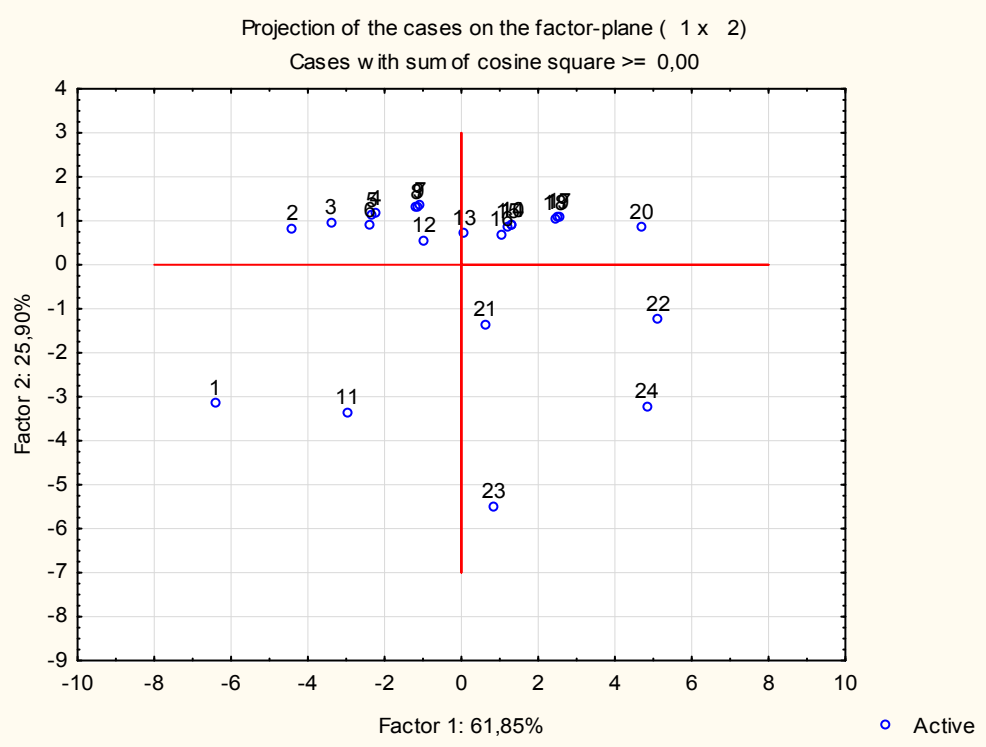
Variable	Factor-variable correlations (factor loadings), based on correlations (faktorska analiza)		
	Factor 1	Factor 2	Factor 3
Uc	0,843166	-0,337648	0,236583
Ui	0,839731	-0,362649	0,202590
Hy	-0,666785	-0,486198	0,546445
AMR	0,990903	-0,100414	0,025441
TPSA(NO)	0,163827	-0,949135	-0,258582
TPSA(Tot)	0,163827	-0,949135	-0,258582
MLOGP	0,984184	0,093597	0,123061
MLOGP2	0,975745	0,055759	0,183775
ALOGP	0,926017	0,248518	0,161148
ALOGP2	0,921257	0,202392	0,206888
SAtot	0,951488	0,024499	-0,178971
SAacc	0,115907	-0,968385	-0,191519
SAdon	-0,491973	-0,525935	0,679771
Vx	0,987453	-0,058135	-0,059811

	Factor coordinates of cases, based on correlations (faktorska analiza)		
	Factor 1	Factor 2	Factor 3
1	-6,42068	-3,15060	2,03541
2	-4,40710	0,83545	-0,90973
3	-3,40446	0,97206	-0,83964
4	-2,23435	1,16899	-0,66660
5	-2,33157	1,12384	-0,74706
6	-2,37717	0,90335	-0,58679
7	-1,07283	1,35424	-0,48731
8	-1,17575	1,30868	-0,56824
9	-1,12147	1,32890	-0,54132
10	1,31798	0,91660	0,37051
11	-2,98003	-3,35323	2,76376
12	-0,98307	0,56677	-0,00179
13	0,07379	0,70750	0,12846
14	1,31798	0,91660	0,37051
15	1,20381	0,86570	0,27974
16	1,04451	0,70050	0,36286
17	2,54460	1,11344	0,60776
18	2,42595	1,06253	0,51726
19	2,48688	1,08501	0,54822
20	4,68745	0,86008	1,32652
21	0,61485	-1,35981	-1,00771
22	5,12955	-1,22639	0,14502
23	0,83672	-5,49329	-2,69755
24	4,82441	-3,20692	-0,40226

Table 6.10: *Matrica korelacija između faktora i molekulskih deskriptora i matrica faktora*



Grafik 6.4: *Grafički prikaz molekulskih deskriptora unutar sistema faktorskih osa (Factor 1 i Factor 2)*



Grafik 6.5: *Grafički prikaz realizovanih vrednosti (vrste matrice) molekulskih deskriptora unutar sistema faktorskih osa (Factor 1 i Factor 2)*

U ovom delu, primenjena je analiza glavnih komponenti. Podaci o karakterističnim korenima i dr. je dobijeno iz korelacione matrice (s obzirom da se drugačije vrednosti dobijaju iz kovarijansne). Ova tri faktora objašnjavaju nešto veću količinu ukupne varijanse (96,19%). Drastičnih razlika u odnosu na metodu faktorske analize nema, što potvrđuje i grafik 6.4, jer ponovo postoje tri grupe međusobno jako korelisanih deskriptora. Tako, na osnovu tabele 6.10, se zaključuje da su isti deskriptori korelisani sa istim faktorima, kao u faktorskoj analizi, s tim što su drugačije vrednosti korelacije.

Dakle, zamenom matrice molekulskih deskriptora sa matricom faktora dobija se nova matrica podataka, koja objašnjava oko 96% varijanse prethodne matrice. Takođe, i ovde su faktori međusobno nekorelisani što rešava problem singularnih matrica.

6.3 Kvantitativna veza između strukture supstance i hromatografskih retencionih parametara

6.3.1 Hromatografska analiza

Eksperimentalni merenjem i primenom hromatografije na tankom sloju određeni su retencioni faktori R_f vrednosti koje su karakteristične za svaku ispitivanu supstancu. R_f vrednosti određivane su u različitim eksperimentalnim uslovima, odnosno u smešama metanola i vode različitog sastava. U intepretaciji rezultata korišćena je vrednost R_M, koja se dobija na sledeći način:

$$R_M = \log\left(\frac{1}{R_f} - 1\right).$$

R_f odnosno R_M vrednosti ispitivanih jedinjenja menjale su se u zavisnosti od zapreminskog udela metanola u smeši (φ) metanloa-vode prema sledećoj jednačini:

$$R_M = R_M^0 + S\varphi,$$

gde je φ zapreminski udeo metanola, R_M^0 je odsečak, a S nagib jednačine. Dobijene vrednosti odsečka R_M^0 korišćene su u intepretaciji rezultata, i one su prikazane u tabeli 5.6.

R_M^0	Metanol
I.1	1,616
I.2	2,122
I.3	2,397
I.4	2,879
I.5	2,948
I.6	2,498
I.7	3,308
I.8	3,32
I.9	3,362
I.10	3,455

R_M^0	Metanol
II.1	2,234
II.2	2,5
II.3	2,986
II.4	3,442
II.5	3,447
II.6	3,25
II.7	4,001
II.8	3,951
II.9	4,123
II.10	4,214
II.11	3,246
II.12	4,529
II.13	2,955
II.14	3,785

Table 6.11: *Vrednosti odsečka R_M^0 za derivate hindatoina u rastvoru metanola*

6.3.2 Metoda regresije parcijalnih najmanjih kvadrata

Da bi se opisala kvantitativna veza između R_M^0 i računski dobijenih molekulskih deskriptora, korišćena je regresija parcijalnih najmanjih kvadata, kao zamena za višestruku regresiju. Zavisna matrica su vrednost R_M^0 (tabela 6.11), a nezavisna matrica je tabela molekulskih deskriptora (tabela 6.3). Za dobijanje vrednosti karakterističnih korena, matrice regresionih koeficijenata, matrice težinskih koeficijenata i dr. korišćen je NIPALS algoritam, unutar PLS regresije softverskog paket STATISTICA

Component	Partial Least Squares Analysis Summary (faktorska analiza) Number of components is 4 98,9783% of sum of squares of the dependent variables has been explained by all the extracted components.									
	R2X	R2X(Cumul.)	Eigenvalues	R2Y	R2Y(Cumul.)	Q2	Limit	Q2(Cumul.)	Significance	Iterations
1	0,614512	0,614512	8,497736	0,898837	0,898837	0,86940	0,00	0,869399	S	1
2	0,251771	0,866283	2,688778	0,035244	0,934082	0,02730	0,00	0,872964	S	1
3	0,042059	0,908342	0,577015	0,051064	0,985146	0,47856	0,00	0,933759	S	1
4	0,043147	0,951489	0,130138	0,004637	0,989783	-8,15861	0,00	0,393322	NS	1

Component	Partial Least Squares Analysis Summary (faktorska analiza) Number of components is 4 98,9783% of sum of squares of the dependent variables has been explained by all the extracted components.									
	R2X	R2X(Cumul.)	Eigenvalues	R2Y	R2Y(Cumul.)	Q2	Limit	Q2(Cumul.)	Significance	Iterations

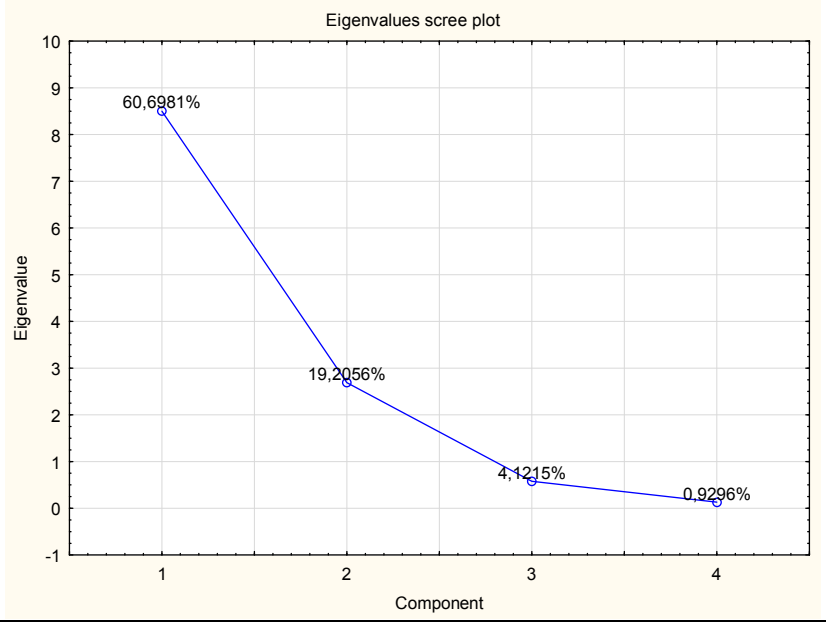


Table 6.12: *Vrednost karakterističnih korena komponenti, objašnjene varijanse kod zavisne i nezavisne matrice i vrednost predikcione moći modela*

	X score spreadsheet (faktorska analiza) Number of components is 4			Y score spreadsheet (faktorska analiza) Number of components is 4		
	Component 1	Component 2	Component 3	Component 1	Component 2	Component 3
1	-6,78526	-1,90313	1,00011	-2,18524	0,021519	0,239408
2	-4,25595	0,54759	-0,92572	-1,48289	-0,098735	-0,161429
3	-3,21394	0,88492	-0,35628	-1,10118	-0,055913	-0,157228
4	-1,98800	1,34309	0,36797	-0,43214	0,214412	0,060642
5	-2,09547	1,27270	0,29730	-0,33637	0,345139	0,199429
6	-2,22213	0,68757	-0,50845	-0,96099	-0,238286	-0,317005
7	-0,77311	1,78578	1,08096	0,16333	0,414764	0,210310
8	-0,88662	1,71220	1,00378	0,17998	0,468338	0,272308
9	-0,82680	1,75279	1,05580	0,23828	0,507180	0,306503
10	1,42761	0,54728	-0,36478	0,36737	-0,096930	-0,159588
11	-3,48416	-2,75765	0,04598	-1,32743	-0,194283	0,121440
12	-0,97627	-0,31392	-1,77656	-0,95821	-0,640701	-0,604761
13	0,12275	0,04891	-1,15682	-0,28362	-0,323544	-0,329144
14	1,42761	0,54728	-0,36478	0,34932	-0,114974	-0,177632
15	1,30156	0,46487	-0,45130	0,35626	-0,067042	-0,120265
16	1,08203	0,02201	-1,07954	0,08282	-0,269088	-0,271608
17	2,71190	1,02524	0,40608	1,12524	0,243251	0,125871
18	2,58115	0,94051	0,31418	1,05584	0,216372	0,108693
19	2,64826	0,98581	0,37242	1,29458	0,433290	0,320425
20	4,73533	0,28047	-0,39469	1,42089	-0,119172	-0,151282
21	0,37630	-1,28671	-0,10744	0,07727	-0,045116	0,102200
22	4,88668	-1,07971	0,62811	1,85813	0,268837	0,392453
23	-0,04692	-4,62991	0,47797	-0,32665	-0,311393	0,218686
24	4,25346	-2,87800	0,43569	0,82542	-0,557926	-0,228423

Table 6.13: *Matrica skora zavisne i nezavisne matrice*

Variable	X weight spreadsheet (faktorska analiza) Number of components is 4			
	Variable number	Component 1	Component 2	Component 3
Uc	38	0,221114	-0,473792	-0,310135
Ui	39	0,215869	-0,502201	-0,333977
Hy	40	-0,250332	-0,054681	0,314450
AMR	41	0,324367	-0,090227	0,060748
TPSA(NO)	42	-0,017534	-0,359668	0,360315
TPSA(Tot)	43	-0,017534	-0,359668	0,360315
MLOGP	44	0,337175	-0,017269	-0,018377
MLOGP2	45	0,332160	-0,025030	0,003156
ALOGP	46	0,347655	0,196637	0,146100
ALOGP2	47	0,342979	0,183592	0,168985
SAtot	48	0,346055	0,175177	0,344762
SAacc	49	-0,036382	-0,376147	0,344613
SAdon	50	-0,195732	-0,086427	0,323733
Vx	51	0,337802	0,022857	0,193342

Table 6.14: *Matrica težinskih koeficijenata nezavisne matrice*

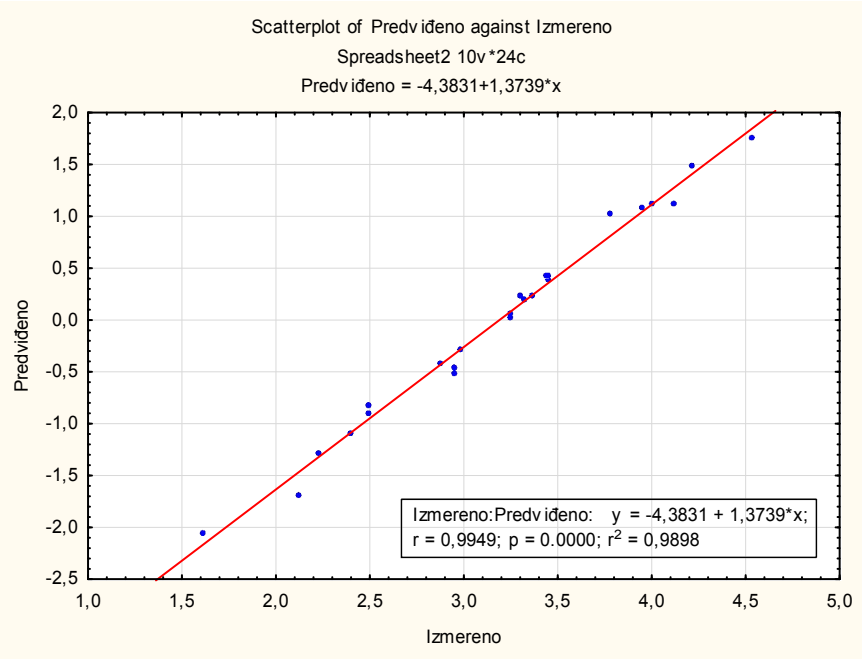
Variable	X loading spreadsheet (faktorska analiza) Number of components is 4			
	Variable number	Component 1	Component 2	Component 3
Uc	38	0,273888	-0,300858	-0,339866
Ui	39	0,271807	-0,315973	-0,338612
Hy	40	-0,244241	-0,230021	0,285230
AMR	41	0,334417	-0,124101	0,026277
TPSA(NO)	42	0,022528	-0,560583	0,380131
TPSA(Tot)	43	0,022528	-0,560583	0,380131
MLOGP	44	0,339098	-0,007022	-0,011580
MLOGP2	45	0,334948	-0,026790	0,019449
ALOGP	46	0,325753	0,115171	0,198038
ALOGP2	47	0,322529	0,089364	0,222762
SAtot	48	0,326543	-0,017066	0,278632
SAacc	49	0,005516	-0,568306	0,391400
SAdon	50	-0,186105	-0,266943	0,279189
Vx	51	0,335256	-0,084952	0,145384

Tabela 6.15: *Matrica koeficijenata nezavisne matrice*

	PLS scaled regression coefficients (faktorska analiza) Responses: Metanol Options: NO-INTERCEPT AUTOSCALE							
	Uc	**Ui**	**Hy**	**AMR**	**TPSA(NO)**	**TPSA(Tot)**	**MLOGP**	**MLOGP2**
Metanol	-0,146280	-0,163207	-0,011019	0,108398	0,000098	0,000098	0,109880	0,112322

	PLS scaled regression coefficients (faktorska analiza) Responses: Metanol Options: NO-INTERCEPT AUTOSCALE					
	ALOGP	**ALOGP2**	**SAtot**	**SAacc**	**SAdon**	**Vx**
Metanol	0,222518	0,224002	0,275030	-0,015912	0,002304	0,184337

Tabela 6.16: *Matrica regresionih koeficijenata* $\boldsymbol{B}_{PLS}$



Grafik 6.6: *Korelacija metodom višestuke regresije predviđenih i izmerenih R_M^0 vrednosti*

Analizom parcijalnih najmanjih kvadrata dobijen je trokomponentni model. Četvrta komponenta nije fizički značajna što se vidi iz tabele 6.12. Njeno uključivanje bi uticalo na smanjenje ukupne predikcione vrednost modela Q2 (u tabeli 6.12 se vidi da se kumulativna Q2 vrednost smanjuje sa ubacivanjem 4. komponente). Takođe, rezultati iz tabele 4.12 pokazuju da je sa tri komponente objašnjavaju 90,83% varijanse nezavisne matrice, a 98,51% varijanse zavisne matrice. Ukupna prediktivna vrednost modela (Q2 kumulativno) je 93,37%.

Kad je reč o pridikcionim koeficeijentim molekulskih deskriptora u donosu na vrednost R_M^0 (tabela 6.16), deskriptor Satot, pa zatim ALOGP i ALOGP2 su najviše pozitivno korelisani. S druge strane, UC i Ui su negativno korelisani sa R_M^0, dok deskriptori Hy, TPSA(NO), TPSA(Tot), Saacc i SAdon skoro da nemaju uticaj na vrednost R_M^0. Na osnovu grafika 6.6, primećuje se da ni jedan derivat hindatoina ne odskače od regresione linije, čime se zaključuje da predikcioni model dobro predviđa vrednosti.

Zaključak

U ovom master radu je prikazan pregled osnovnih pojmova faktorske analize, njenih metoda (analiza glavnih komponenti, (prava) faktorska analiza i regresija parcijalnih najmanjih kvadrata), kao i primena u hemiji.

Osnovna ideja je bila da se promenljive unutar matrice podataka zamene faktorima (komponentama), kako bi ''nove promenljive'' bile međusobno nekorelisane, a čime se rešava problem multikolinearnosti. Takođe, smanjuje se i broj podataka što je mnogo ekonomičnije. Krećući od te ideje, pokazano je da postoje dve metode, koje na sasvim jednostavan, ali idejno drugačiji način, daju rešenja novih faktora. Analiza glavnih komponenti kreće od pretpostavke da nove komponente objašnjavaju (opisuju) maksimalnu varijansu (rasipanje), i to hronološkim redom. S druge strane, faktorska analiza pretpostavlja da novi faktori treba da održe jednaku korelaciju (kovarijansu) između promenljivih kao kod korelacione (kovarijansne) matrice. Različiti pristupi zato daju i različita rešenja, zavisno od ideje kojom se rukovodi.

Opšti zaključak teorijske priče prve dve metode jeste da one mogu da se primene u praksi, i pored brojnih uslova i pretpostavki. Takođe, ''lepota'' metoda jeste da one nemaju jedinstveno rešenje, već se ono može stalno menjati korišćenje ortonormirane transformacione matrice (rotacija sistema osa) što nam pomaže kod intepretacije rezultata.

Regresija parcijalnih najmanjih kvadrata predstavlja analizu koja se primenjuje nad regresionim modelom, kako bi model imao što bolju predikcionu moć. Pokazano je, putem algoritma, da se za što bolju predikcionu moć modela ne treba analizirati samo nezavisna matrica, već i zavisna matrica, korišćenjem NIPALS dekompozicije.

Primena metoda u hemiji (hemometrija) prikazana je praktična priča metoda. Uz pomoć statističkog softvera STATISTICA, lako se može manipulisati sa rešenjim sve u cilju dobijanja boljeg, razumljivijeg rešenja. Tako, na primer, primenom parcijalnih najmanjih kvadrata nad matricama molekulskih deskriptora i matrici R_M^0 vrednosti, dobijena je matrica regresionih koeficijenata iz kojih se lako može doneti zaključak o korelisanosti deskriptora sa R_M^0 vrednosti.

Iako je od nastanka prve metode faktorske analize prošlo više od jednog veka, faktorska analiza je i dalje veoma interesantna i inspirativna za mnoge naučnike. Zato, danas nije poznat tačan broj metoda faktorskih analiza jer se njihov broj stalno povećava. Uzastopno se povećava i broj naučnih oblasti u kojima se ona primenjuje, tako da od psihologije, hemije i medicine koje su bile početne baze ideja nastanka analize, danas ne postoji naučna grana gde se ona ne primenjuje.

Sama tema rada je za mene bila veoma zanimljiva budući da se njome nismo bavili tokom mojih akademskih studija. Težinu u pisanju rada mi je predstavljala literatura. Mnogi isti modeli u literaturi su se razlikovali kako u notaciji tako i u intepretaciji. Upravo zbog te činjenice je bilo potrebno pročitati dosta knjiga kako bi se došlo do modela lako razumljivog kako za mene tako i za čitaoce rada.

Stoga, prezentovani master rad može da predstavlja uvodnu priču za dalju, dublju analizu faktorske analize i njenih metoda.

Literatura

[1] Jolliffe I.T., Principal Component Analysis Second Edition, Springer, 2002.

[2] Malinowski R. Edmundi, Factor Analysis in Chemistry, 3rd Edition, Wiley-Interscience, 2002

[3] Richard A. Reyment, Joreskog K. G., Applied Factor Analysis in the Natural Sciences, Cambridge University Press, 1996.

[4] Fulgosi Ante, Faktorska analiza, Treće dopunjeno izdanje, Školska knjiga Zagreb, 1988.

[5] David Aaker , V. Kumar, Robert Leone, George S. Day, Marketing Research, 11th Edition October 2012

[6] Spearman Charles, "General Intelligence," Objectively Determined and Measured". The American Journal of Psychology, 1904.

[7] Pearson Karl, On lines and planes of closest fit to systems of points in space, Philosophical Magazine, Series 6, vol. 2, no. 11, 1901.

[8] Hotteling, Harold, 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 1933.

[9] Wold H. "Nonlinear Iterative Partial Least Squares (NIPALS) Modeling: Some Current Developments", in P.R. Krishnaiah [ed.]. Multivariate Analysis II, Proceedings of an International Symposium on Multivariate Analysis held at Wright State University, Dayton, Ohio, June 19-24, 1972, New York: Academic Press, pp. 383-407

[10] Robert W. Geralch, Kowalski Bruce R. , Wold Herman O. A., Partial Least Squares Path Modelling with Latent Variables. Washington univ seattle lab for chemometrics, Jun-Sep 79

[11] Kovačić J. Zlatko, Multivarijaciona analiza, Ekonomski fakultet, Univerzitet u Beogradu, 1994.

[12] Wold S., Chemometrics and Intelligent Laboratory Systems 30, 1995., 109

[13] Camile W., The Practice of Medicinal Chemistry, Academic Press, Oxford, 2003.

[14] Ozgur Yeniay, Atilla Goktas, A comparison of partial least squares regression with other prediction methods, Hacettepe Journal of Mathematics and Statistics, Volume 31 (2002), 99/111

[15] Kaleman Svetlana, Analiza hromatografskih podataka novosintetisanih derivata hidantoina primenom hemometrijskih metoda-doktorska disertacija, Univerzitet u Novom Sadu, Prirodno-matematički fakultet, Departman za hemiju, biohemiju i zaštitu životne sredine, 2012.

Kratka biografija

Vladimir Rančić je rođen 23.06.1988. godine u Bačkoj Topoli. U rodnom gradu je završio Osnovnu školu ''Nikola Tesla'', a potom Gimnaziju ''Dositej Obradović'' - opšti smer. Godine 2007. upisuje osnovne studije Finansijske matematike na Departmanu za matematiku i informatiku, Prirodno-matematičkog fakulteta u Novom Sadu. Završava ih za četiri godine sa prosekom 9,06. Odmah nakon osnovnih studija upisuje master studije Finansijske matematike i zaključno sa septembrom 2012. godine, polaže sve predviđene predmete nastavnim planom i time stiče uslov za odbranu master rada. S obzirom na veliko interesovanje u finansije i ekonomiju, tokom studija je odslušao sve ponuđene predmete iz tih oblasti. Tokom studija je bio član parlamenta PMF, studentske organizacije AIESEC, učestvovao je na mnogim projektima, nosilac je prve nagrade Agencije za borbu protiv korupcije i dr.

Univerzitet u Novom Sadu
Prirodno-matematički fakultet
Ključna dokumentacijska informacija

Redni broj:

RBR

Identifikacioni broj:

IBR

Tip dokumentacije: Monografska dokumentacija

TD

Tip zapisa: Tekstualni štampani materijal

TZ

Vrsta rada: Master rad

VR

Autor: Vladimir Rančić

AU

Mentor: Dr Zagorka Lozanov-Crvenković

MN

Naslov rada: Metode za smanjenje dimenzionalnosti podataka i njihova primena u prirodnim naukama

NR

Jezik publikacije: Sprski (latinica)

JP

Jezik izvoda: s/en

JI

Zemlja publikovanja: Republika Srbija

ZP

Uže geografsko područje: Vojvodina

UGP

Godina: 2013

GO

Izdavač: Autorski reprint

IZ

Mesto i adresa: Trg Dositeja Obradovića 4, 21000 Novi Sad

84

MA

Fizički opis rada: (5/ 65/ 1 / 23 / 4 / 0 / 0)

FO

Naučna oblast: Matematika

NO

Naučna disciplina: Statistika

ND

Predmetna odrednica, ključne reči: Faktorska analiza, Analiza glavnih komponenti, Regresija parcijalnih najmanjih kvadrata

PO

UDK

Čuva se: Biblioteka departmana za matematiku i informatiku, Prirodno- matematički fakultet, Trg Dositeja Obradovića 4, Novi Sad

ČU

Važna napomena: Nema

VN

Izvod: U master radu su predstavljene statističke metode za smanjenje dimenzionalnosti podataka i to analiza glavnih komponenti, (stvarna) faktorska analiza i analiza parcijalnih najmanjih kvadrata. Sve tri metode spadaju u statističku oblast koja se naziva faktorska analiza. Osnovna ideja faktorske analize jeste da se broj promenljivih unutar matrice podataka smanji na određeni broj faktora (komponenti), kako bi novi model imao bolju predikcionu moć. Time se rešava i problem multikolinearnosti (singularne matrice), a i smanjuje se broj promenljivih (ekonomičnost). Sve tri metode su primenjene u hemiji, odnosno analizirana je matrica molekulskih deskriptora i kvantitativna veza između njih i hidantoina.

IZ

Datum prihvatanja teme od strane NN veća: 19.3.2013.

DP

Datum odbrane: Jun 2013.

DO

Članovi komisije: (naučni stepen/ime i prezime/zvanje/fakultet)

KO

Predsednik: dr Tatjana Đaković-Sekulić, redovni profesor Prirodno-matematičkog fakulteta u Novom Sadu

Član: dr Zagorka Lozanov-Crvenković, redovni profesor Prirodno-matematičkog fakuteta u Novom Sadu

Član: dr Ivana Štajner-Papuga, vanredni profesor Prirodno-matematičkog fakulteta u Novom Sadu

University of Novi Sad

Faculty of Natural Sciences And Mathematics

Key word documentation

Accession number:

ANO

Identification number:

INO

Document type: Monograph documentation

DT

Type of record: Textual printed material

TR

Contents code: Master's thesis

CC

Author: Vladimir Rančić

AU

Mentor: Zagorka Lozanov-Crvenković Ph.D.

MN

Title: Methods for reducing the dimensionality of data with apllication in natural sciences

TI

Language of text: Serbian (latin)

LT

Language of abstract: en/s

LA

Country of publication: Republic of Serbia

CP

Locality of publication: Vojvodina

LP

Publication year: 2013

PY

Publisher: Author's reprint

PU

Publ. place: Trg Dositea Obradovic 4, Novi Sad

PP

Physical description: (5/ 65/ 1/ 23/ 4/ 0/ 0)

PD

Scientific field: Mathematics

SF

Scientific discipline: Statistics

SD

Subject Key words: Factor analysis, Principal component analysis, Partial least-squares regression

SKW

UC

Holding data: Library of the Department of Mathematics and Computer Sciences, Faculty of Natural Sciences, Trg Dositeja Obradovića 4

HD

Abstract: This master's thesis includes three statistical methods for reducing the dimensionality of date: principal component analysis, true factor analysis and partial least-squares regression. These methods belong to statistical field which that name is factor analysis. The main idea of factor analysis is reducing the number of variables in a date matrix to smaller number of factors (components), because the new model has better prediction power. Futhermore, the new model solves the problem of multicolinearity (singular matrix) and reduces the number of varibles (economical). All methods are apllied in chemistry, in analysing the matrix of molecular descriptors and quantitative relations between them and characteristics of hydantoins.

AB

Accepted on Scientific Board on: 19.3.2013.

AS

Defended:

DE

Thesis Defend board: Jun 2013.

DB

president: Tatajana Đakovic-Sekulić Ph. D., Full professor, Faculty of Natural Sciences and Mathematics, Novi Sad

member: Zagorka Lozanov-Crvenković Ph. D., Assistant Professor, Faculty of Natural Sciences and Mathematics, Novi Sad, mentor

member: Ivana Štajner-Papuga Ph. D., Associate professor, Faculty of Natural Sciences and Mathematics, Novi Sad