UNIVERZITET U BEOGRADU

EKONOMSKI FAKULTET

Zlatko J. Kovačić

MULTIVARIJACIONA ANALIZA

Beograd, 1994.

Predgovor

U knjizi su izloženi osnovi multivarijacionih statističkih metoda. Neposredan povod za njenu pojavu jeste potreba za učilom studenata Ekonomskog fakulteta u Beogradu. Naime, na tom fakultetu studenti redovnih i poslediplomskih studija slušaju Multivarijacionu analizu u okviru istoimenog predmeta. Iz tog razloga su korišćeni primeri iz ekonomije i marketinga za ilustraciju izloženih metoda. Međutim, kao prvi udžbenik iz ove oblasti na našem jeziku, ova knjiga je namenjena i studentima drugih fakulteta kao i onima koji koriste metode multivarijacione analize u praksi. Zbog njih sam pokušao naći kompromis između potrebne rigoroznosti matematičkog izlaganja i praktične upotrebljivosti teorijskih modela. Nagalasak je pri tome na razumevanju i usvajanju osnovnih pojmova izloženih metoda kao i pravilnoj interpretaciji dobijenih rezultata. Na čitaocu je da oceni u kojoj sam meri u tome uspeo.

Struktura svake glave je takva da upućuje čitaoca na proveru razumevanja pročitanog. Tako je nakon izlaganja svakog značajnijeg teorijskog rezultata urađen primer koji ilustruje numeričke aspekte metode. Nivo izlaganja prilagođen je onom obimu znanja iz statističke analize i matrične algebre koju dobijaju studenti Ekonomskog fakulteta u okviru jednogodišnjih kurseva.

Zahvalnost dugujem kolegama profesoru Vladislavu Miloševiću i profesoru Tomislavu Zečeviću na korisnim sugestijama tokom pisanja knjige. Njihovo kritičko čitanje prvobitne verzije doprinelo je poboljšanju knjige u finalnoj verziji. Posebnu zahvalnost izražavam svom studentu Branku Jovanoviću koji je pedantno prošao kroz rukopis ispravljajući prisutne greške, istovremeno sugerišući bolju formulaciju u izvesnim delovima teksta.

Preostale greške ili neusklađenosti, kojih nije imuna ni ova knjiga, treba pripisati isključivo autoru. Zato su sve primedbe i sugestije čitaoca dobrodošle.

U Beogradu, jula 1994. Z.J.K.

- Znaš li sabiranje? - upitala je Bela Kraljica. - Koliko je jedan i jedan i jedan i jedan i jedan i jedan i jedan i jedan i jedan?
- Ne znam - odgovorila je Alisa - izgubila sam se u tom dugačkom nizu.
- Ona ne zna Sabiranje - rekla je Crna Kraljica.
- A znaš li Oduzimanje? Hajde da vidimo. Koliko je osam manje devet.
- Da oduzmem devet od osam? Ne znam - odgovorila je Alisa malo stidljivo - ali...
- Ona ne zna Oduzimanje - zaključila je Bela Kraljica.
- A znaš li Delenje? Podeli veknicu hleba nožem, šta ćeš dobiti?
- Mislim... - počela je Alisa, ali Crna Kraljica je preduhitri odgovorom:
- Sendvič s maslacem, naravno, Hajdemo još jedan minut da pokušamo sa Oduzimanjem. Oduzmi od jednog psa jednu kost. Šta ostaje?
Alisa je razmišljala.
-Kosti više neće biti, ako je uzmem. Onda više neće biti ni psa, jer će dojuriti da me ugrize, a onda ni mene više neće biti.
- I po tvome ništa ne bi trebalo da ostane? - upitala je Crna Kraljica.
- Ja mislim da je to pravi odgovor - uzvratila je Alisa.
- Opet pogrešan odgovor - rekla je Crna Kraljica - jer bi svakako trebalo da ostane odjek psećeg laveža.
- Ali ne shvatam kako...
- Vidiš - rekla je Crna Kraljica - po svoj prilici pas bi trebalo da laje pre no što te ujede, zar ne?
- Pa trebalo bi da bude tako - odgovorila je Alisa oprezno.
- A kada ode pas, ostaće samo njegov lavež! - uzviknula je Crna Kraljica pobedonosno.

Luis Kerol: Alisa u svetu s one strane ogledala

SADRŽAJ

1 UVOD

Poslednjih godina svedoci smo široke primene metoda multivarijacione analize podataka skoro u svim naučnim oblastima. Dva su osnovna razloga za tako nešto. Prvi, razvoj kompjuterske tehnike i softverskih proizvoda koji su omogućili relativno jednostavnu primenu metoda multivarijacione analize, i drugi, sagledavanje potrebe mnogih naučnih istraživanja da se analiziraju simultane međuzavisnosti između tri ili više promenljivih. U okviru ove uvodne glave ukazaćemo na definicije multivarijacione analize kao i na različite klasifikacije njenih metoda. U bliskoj vezi s tim je razmatranje vrste podataka koje koristimo u multivarijacionoj analizi, a koji sa svoje strane uslovljavaju izbor korišćene metode. Na kraju je posvećena pažnja grafičkoj prezentaciji multivarijacionih podataka kao istraživačkoj fazi multivarijacione analize.

1.1 DEFINICIJA I KLASIFIKACIJA METODA MULTIVARIJACIONE ANALIZE

U procesu naučnog objašnjenja prirode nekog fenomena polaznu osnovu analize sačinjavaju podaci koji se odnose na jedan ili više skupova objekata. Ovi objekti u analizi mogu biti: pojedinci, ljudske zajednice, različiti predmeti, a takođe prirodni fenomeni ili one pojave koje su proizvod aktivne delatnosti čoveka. Često nismo u prilici da kompleksnu prirodu objekata sagledamo u potpunosti. Međutim, na raspolaganju nam stoji mogućnost obuhvata različitih karakteristika jedne, po svojoj prirodi, višedimenzione pojave. Te karakteristike, odnosno obeležja predstavljaju predmet našeg merenja. Njih ćemo jednostavno zvati *promenljive*. Pokušavamo, dakle, ispitati prirodu objekata istovremenim merenjem većeg broja promenljivih na svakoj jedinici posmatranja iz jednog ili više skupova objekata.

Mada ne postoji opšte prihvaćena definicija multivarijacione analize u gornjim redovima naveli smo nekoliko elemenata koje bi takva jedna definicija morala da sadrži. To su: veći broj obeležja i posmatranje simultanih međuzavisnosti među promenljivama. Ako bismo se ipak opredelili za jednu definiciju, tada bismo rekli da ***multivarijaciona***

analiza **predstavlja skup statističkih metoda koje simultano analiziraju višedimenziona merenja dobijena za svaku jedinicu posmatranja iz skupa objekata koji ispitujemo**.

Pretpostavimo da smo tokom merenja sakupili podatke za i objekata, $i = 1, 2, ..., n$, o njihovih j svojstava, $j = 1, 2, ..., p$. Dobijeni podaci predstavljaju osnovu multivarijacione analize i prezentiramo ih u vidu matrice podataka (tabela u kojoj se red odnosi na objekat, a kolona na promenljivu). Pretpostavivši da imamo n redova (objekata) i p kolona (obeležja, odnosno promenljivih), tabela podataka ima izgled:

	Promenljiva 1	Promenljiva 2	$\cdots$	Promenljiva j	$\cdots$	Promenljiva p
Objekat 1:	X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1p}
Objekat 2:	X_{21}	X_{22}	$\cdots$	X_{2j}	$\cdots$	X_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Objekat i :	X_{i1}	X_{i2}	$\cdots$	X_{ij}	$\cdots$	X_{ip}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Objekat n :	X_{n1}	X_{n2}	$\cdots$	X_{nj}	$\cdots$	X_{np}

gde (i, j) element matrice predstavlja vrednost $j-$te promenljive merene na $i-$tom objektu. U matričnoj notaciji ovu matricu podataka označavamo sa $\mathbf{X}$, odnosno $[X_{ij}]$, $i = 1, 2, ..., n;\ \ j = 1, 2, ..., p$.

Izbor odgovarajućeg metoda za analizu matrice podataka zavisi od mnogih faktora, a opredeljen je pre svega vrstom problema, tipom podataka, karakteristikama same metode i u krajnjem slučaju ciljem istraživanja. Direktno zaključivanje o međuzavisnosti promenljivih, odnosno o međuodnosu objekata je veoma teško, ako je uopšte moguće, s obzirom na dimenzije matrice podataka. U te svrhe možemo koristiti metode multivarijacione analize za redukciju velike količine podataka i njihovom iskazivanju preko nekoliko veličina. Ovim metodama istovremeno postižemo pojednostavljivanje složene strukture posmatranog fenomena u cilju njegove lakše interpretacije. Pored ovog, pre svega deskriptivnog zadatka, metode multivarijacione analize koristimo u procesu zaključivanja, tako što ocenjujemo, na primer stepen međuzavisnosti promenljivih i/ili testiramo njihovu statističku značajnost. Naposletku, napominjemo da su neke od metoda multivarijacione analize istraživačkog karaktera, što će reći da se koriste ne za testiranje apriori definisanih hipoteza, nego za njihovo generisanje, odnosno konstruisanje.

Klasifikacije metoda multivarijacione analize zasnovane su na različitim klasifikacionim kriterijumima. Prva klasifikacija metoda pravi razliku među njima prema tome da li su orijentisane ka ispitivanju međuzavisnosti promenljivih ili im je osnovni zadatak ispitivanje međuzavisnosti objekata. Kada istražujemo međuzavisnost promenljivih, tada posmatramo kolone matrice podataka. Jedan od načina merenja međuzavisnosti promenljivih baziran je na izračunatom koeficijentu korelacije među

njima. Osnovu ovih metoda multivarijacione analize predstavlja kovarijaciona ili korelaciona matrica. Kod drugog pristupa, u cilju poređenja dva objekta ili osobe, posmatramo odgovarajuće redove u matrici podataka, odnosno definišemo različite mere bliskosti između dva objekta ili osobe. Osnovu ovih metoda multivarijacione analize predstavlja matrica odstojanja između objekata.

Prema drugoj klasifikaciji, metode delimo u dve grupe: metodi zavisnosti i metodi međuzavisnosti. Ukoliko smo u istraživanju zainteresovani za ispitivanje zavisnosti između dva skupa promenljivih, gde jedan skup predstavlja zavisne promenljive, a drugi nezavisne promenljive, tada se odgovarajuća klasa metoda naziva *metodi zavisnosti*. S druge strane, ako nema apriornog, teorijskog osnova za podelu svih promenljivih na dva podskupa promenljivih (zavisnih i nezavisnih), tada koristimo *metode međuzavisnosti*. Treba uočiti da kod metoda zavisnosti težimo da objasnimo ili predvidimo jednu ili više zavisnih promenljivih na osnovu skupa nezavisnih promenljivih. Metodi međuzavisnosti, s druge strane, nisu po svojoj prirodi prediktivni. Njima se pokušava učiniti prodor u kompleksnu unutrašnju strukturu podataka i to njenim pojednostavljenjem, prvenstveno kroz redukciju podataka.

Na osnovu podele metoda multivarijacione analize na metode zavisnosti i međuzavisnosti klasifikujmo konkretne metode u jednu od ovih klasa, istovremeno dajući njihov sažet opis.

Metode zavisnosti

1. *Multivarijaciona regresija*. Ovo je najpoznatija metoda multivarijacione analize. Koristimo u njenom nazivu izraz multivarijaciona da bismo i na taj način razlikovali dva slučaja. Prvi, u okviru koga se bavimo analizom zavisnosti jedne promenljive (zavisna promenljiva) od skupa drugih promenljivih (nezavisne promenljive). Ovaj metod analize poznatiji je pod nazivom *metod višestruke regresije*. Drugi slučaj je kad skup zavisnih promenljivih sadrži više od jednog člana. Za ovaj slučaj kažemo da predstavlja opštiji model multivarijacione regresije. Kod oba modela zadatak nam je ocenjivanje ili predviđanje srednje vrednosti zavisne, odnosno srednjih vrednosti zavisnih promenljivih na bazi poznatih vrednosti nezavisnih promenljivih.
2. *Kanonička korelaciona analiza*. Ova analiza se može smatrati uopštenjem višestruke regresione analize. Naime, njome želimo uspostaviti linearnu zavisnost između skupa nezavisnih i skupa zavisnih promenljivih. Kod izračunavanja kanoničke korelacije formiramo dve linearne kombinacije, jednu za skup nezavisnih, a drugu za skup zavisnih promenljivih. Koeficijente ovih linearnih kombinacija određujemo tako da običan koeficijent korelacije između njih bude maksimalan.
3. *Diskriminaciona analiza*. Bavi se problemom razdvajanja grupa i alokacijom opservacija u ranije definisane grupe. Primena diskriminacione analize omogućava identifikaciju promenljive koja je najviše doprinela razdvajanju grupa kao i predviđanje verovatnoće da će objekat pripasti jednoj od grupa, na osnovu

vrednosti skupa nezavisnih promenljivih.

4. *Multivarijaciona analiza varijanse (MANOVA)*. Multivarijaciona analiza varijanse je odgovarajuća metoda analize kada nam je cilj ispitivanje uticaja različitih nivoa jedne ili više "eksperimentalnih" promenljivih na dve ili više zavisnih promenljivih. U tom smislu ona predstavlja uopštenje jednodimenzione analize varijanse (ANOVA). Od posebne je koristi u situaciji kada je moguće sprovesti kontrolisani eksperiment (manipulišući sa nekoliko tretmana). Osnovni cilj je testiranje hipoteze koja se tiče varijanse efekata grupa dve ili više zavisnih promenljivih.
5. *Logit analiza*. Kada je u regresionom modelu zavisna promenljiva dihotomnog tipa (na primer, promenljiva pola sa modalitetima: muško-žensko), tada takav model nazivamo regresioni model sa kvalitativnom zavisnom promenljivom. Kod njih je zavisna promenljiva, tzv. logit funkcija, logaritam količnika verovatnoća da će dihotomna zavisna promenljiva uzeti jednu ili drugu vrednost. Modele ovog tipa nazivamo i modeli logističke regresione analize.

Metode međusobne zavisnosti

1. *Analiza glavnih komponenti*. Analiza glavnih komponenata je metoda za redukciju većeg broja promenljivih koje razmatramo, na manji broj novih promenljivih (nazivamo ih glavne komponente). Najčešće manjim brojem glavnih komponenata objašnjavamo pretežan deo varijanse originalnih promenljivih, što omogućava lakše razumevanje informacije sadržane u podacima. Osnovni zadatak jeste konstruisanje linearne kombinacije originalnih promenljivih (glavnih komponenata) uz uslov da obuhvate što je moguće veći iznos varijanse originalnog skupa promenljivih. Sukcesivne glavne komponente izdvajaju se uz ograničenje da su međusobom nekorelisane i da obuhvataju u maksimalnom iznosu preostali deo ukupne varijanse koji nije obuhvaćen prethodno izdvojenim komponentama.
2. *Faktorska analiza*. Slična je metodi glavnih komponenti po tome što se koristi za opis varijacija između promenljivih na osnovu manjeg broja promenljivih (nazivamo ih faktori). Međutim, za razliku od nje, pretpostavlja postojanje odgovarajućeg statističkog modela kojim originalnu promenljivu iskazujemo kao linearnu kombinaciju faktora plus greška modela, odnosno veličina koja odražava stepen nezavisnosti posmatrane promenljive od svih ostalih. Na taj način se celokupna kovarijansa ili korelacija objašnjava zajedničkim faktorima, a neobjašnjeni deo se pridružuje grešci (naziva se specifičan faktor). Dakle, kod faktorske analize, za razliku od glavnih komponenata gde smo zainteresovani za objašnjenje varijanse, naš interes je usmeren ka objašnjenju kovarijanse, odnosno onog dela ukupne varijanse koji promenljiva deli sa ostalim promenljivama iz posmatranog skupa promenljivih.
3. *Analiza grupisanja*. Analiza grupisanja je metoda za redukciju podataka, no za razliku od prethodne dve metode koje su orijentisane ka kolonama (promenljivama), ona je orijentisana ka redovima (objektima) matrice podataka.

Ovom analizom kombinujemo objekte u grupe relativno homogenih objekata. Zadatak u mnogim istraživanjima upravo je identifikovanje manjeg broja grupa, tako da su elementi koji pripadaju nekoj grupi u izvesnom smislu sličniji jedan drugom, nego što su to elementi koji pripadaju drugim grupama.

4. *Višedimenziono proporcionalno prikazivanje*. Pripada klasi metoda koji su orijentisani kao objektima, a koristi meru sličnosti, odnosno razlike između njih u cilju njihovog prostornog prikazivanja. Izvedena prostorna reprezentacija sadrži geometrijski raspored tačaka na mapi, gde se svaka tačka odnosi na jedan od objekata. Ukoliko se za ovo proporcionalno prikazivanje koristi mera bliskosti dobijena na osnovu merljivih (kvantitativnih) promenljivih nazivu metode dodajemo pridev *kvantitativno*, a ako smo za računanje mera sličnosti koristili kvalitativne promenljive, tada nazivu metode dodajemo pridev *kvalitativno*.

5. *Loglinearni modeli*. Ovi modeli omogućavaju ispitivanje međusobne zavisnosti kvalitativnih promenljivih koje formiraju višedimenzionu tabelu kontingencije. Ukoliko se jedna od promenljivih u tabeli kontigencije može smatrati zavisnom, tada na osnovu ocenjenih loglinearnih modela možemo izvesti, ranije spomenute logit modele. Međutim, kod tabela kontingencije logit funkcija se izražava preko ćelijskih frekvencija, za razliku od modela logističke regresije gde logit funkciju iskazujemo preko skupa nezavisnih promenljivih koje mogu biti kvantitativne ili kvalitativne.

1.2 VRSTE PODATAKA I MERNE SKALE

Statistička obeležja mogu biti kvantitativna (merljiva) ili kvalitativna (nemerljiva). Kvantitativne promenljive su one kod kojih se vrednosti razlikuju po veličini. Primeri kvantitativnih promenljivih su cena, dohodak, dužina radnog staža i veličina porodice. Kvalitativne promenljive su one kod kojih se vrednosti razlikuju ne po veličini nego po vrsti. Primeri kvalitativnih promenljivih su pol, bračni status i socijalno poreklo. Klasifikaciju metoda multivarijacione analize moguće je izvršiti i prema vrsti podataka koji se koriste. Primer takve klasifikacije uz ranije uveden kriterijum podele metoda na metode zavisnosti i međuzavisnosti prikazan je u Tabeli 1.1.

Merenja kvantitativnog obeležja iskazujemo na različitim vrstama skala i u različitim jedinicama mere. Ukoliko se jedinica mere može beskonačno deliti (na primer: tone, kilogrami, grami), tada kažemo da je promenljiva iskazana u toj jedinici mere *neprekidna*. Kada jedinica mere nije deljiva (na primer: veličina porodice), tada merenu promenljivu nazivamo *prekidnom*. Najčešće korišćena merna skala kod kvantitativnih promenljivih jeste *skala odnosa*. Ona ima sledeće osobine: (a) količnik ma koje dve vrednosti na ovoj skali ima smislenu interpretaciju, (b) rastojanje između dva objekta mereno na ma kom delu ove skale je jednako i (c) opservacijama pozicioniranim na ovoj skale mogu se dodeliti rangovi od višeg ka nižem.

Na primer, dohodak se obično meri na skali odnosa. (a) Odnos (količnik) dohodaka 50 i 100 novčanih jednica ima jasno značenje u smislu da je drugi dohodak dva puta veći od prvog. (b) Vrednost od 20 novčanih jedinica ista je bez obzira da li je dobijena

na osnovu razlike dohodaka, na primer: 180 i 160 ili 970 i 950 novčanih jedinica. (c) Za dohodak od 300 novčanih jedinica kažemo da je veći, odnosno da ima viši rang od dohotka u iznosu 200 novčanih jedinica.

Tabela 1.1 Klasifikacija metoda multivarijacione analize

		Kvantitativne promenljive	Kvalitativne promenljive
Metodi međuzavisnosti		Glavne komponente Faktorska analiza Analiza grupisanja Kvantitativno višedimenziono proporcionalno prikazivanje	Kvalitativno višedimenziono proporcionalno prikazivanje Loglinearni modeli
Metodi zavisnosti	Jedna zavisna promenljiva	Višestruka korelacija Višestruka regresija	Diskriminaciona analiza (samo zavisna mora biti kvalitativna) Logit analiza
	Više zavisnih promenljivih	Višedimenziona regresija Višedimenziona analiza varijanse Kanonička korelaciona analiza	Kanonička korelaciona analiza sa veštačkim promenljivama

Sledeći tip merne skale je *intervalna skala*, koja za razliku od skale odnosa nema fiksni početak. Na primer, kod merenja temperature arbitrarno je izabrana nulta vrednost kako na Celzijusovoj, tako i na Farenhajtovoj skali. Pri tome, vrednost temperature jednaka nuli ne znači odsustvo temperature, a možemo registrovati i temperature niže od nule. Takođe je poznato da količnik dve temperature iskazane na Celzijusovoj skali nije jednak njihovom količniku iskazanom na skali Farenhajta. Dakle, intervalna skala ima samo osobine (b) i (c) navedene kod skale odnosa. Naposletku, kod kvantitativnih obeležja poslednji tip skale je ordinarna skala koja poseduje samo osobinu (c). Podaci iskazani na ordinarnoj skali su rangovi pridruženi svakoj jedinici posmatranja.

Najniži nivo merne skale koriste kvalitativna obeležja i naziva se *nominalna skala*. Ona ne omogućava čak ni rangiranje jedinica posmatranja. Pri analizi kvalitativnog obeležja kao što je na primer bračni status (kategorije: neoženjen, oženjen, razveden, udovac i razdvojen) proizvoljno pridružujemo vrednosti 1, 2, 3, 4 ili 5 svakoj od navedenih kategorija. To ne znači da smo rangirali kategorije nego smo ih samo kodirali radi lakše statističke obrade.

Razumevanje različitih vrsta podataka i tipova mernih skala veoma je bitno, jer oni uslovljavaju izbor odgovarajuće metode multivarijacione analize. Na primer, ako su nezavisne promenljive kvantitativne, a zavisne promenljive kvalitativne, tada je adekvatan metod analize tih podataka diskriminaciona analiza i analiza kanoničke korelacije sa veštačkim promenljivama.

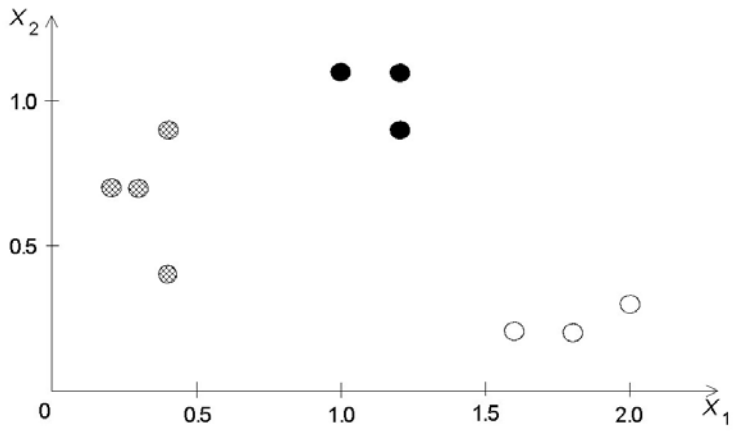
1.3 GRAFIČKI PRIKAZ MULTIVARIJACIONIH PODATAKA

Prvobitno korišćenje grafikona u statistici bilo je u cilju ilustracije tabelarno prezentiranih statističkih podataka. Grafički prikazi su očiglednije, čak i u slučaju relativno malog skupa podataka, ukazivali na određene pravilnosti u podacima nego numerički podaci iskazani u tabeli. Kao ilustraciju ove naše tvrdnje neka posluže hipotetički podaci prezentirani u Tabeli 1.2. Posmatrajmo izdvojeno prve dve promenljive X_1 i X_2. Pregledom opservacija tih promenljivih ne uočavamo neke pravilnosti u podacima. Međutim, ako se poslužimo grafičkim prikazom te dve promenljive, odnosno standardnim dijagramom rasturanja (Slika 1.1), odmah postaje jasno da se parovi opservacija ove dve promenljive grupišu u tri grupe, međusobno jasno razdvojene. Često se već ovakvim rezultatom, odnosno grafikonom, iscrpljuje naš dalji interes za složenijim metodama statističke analize.

Ograničenje klasičnih načina grafičkog prikaza odnosi se na broj promenljivih (dve do tri) koje se mogu istovremeno prikazati. Ta ograničenja se pokušavaju prevazići konstruisanjem novih grafičkih prikaza posebno za multivarijacione statističke podatke. Istovremeno, oni ne predstavljaju više samo alternativan način prezentacije podataka, nego se koriste i za njihovu jednostavnu analizu. U tom smislu, u poslednje vreme razvijani grafički prikazi u vidu tačaka, linija, površina, lica ili drugih geometrijskih formi i figura predstavljaju korisno sredstvo za objašnjenje, interpretaciju i analizu podataka. U ovom odeljku diskutovaćemo i ilustrovati najčešće grafičke prikaze podataka: zvezde, Andrewsove krive i Chernoffova lica.

Tabela 1.2 Hipotetički multivarijacioni podaci

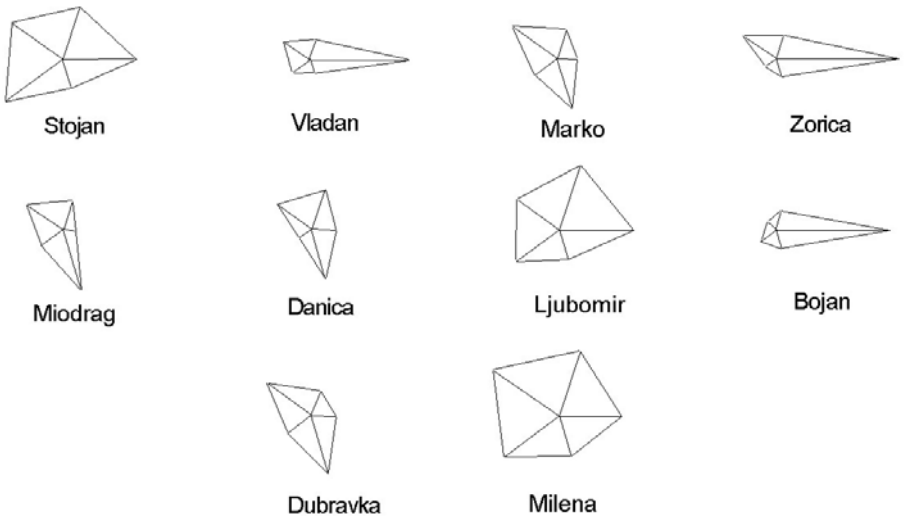
Osoba	X_1	X_2	X_3	X_4	X_5
Stojan	1.2	0.9	1.0	1.1	0.5
Vladan	1.6	0.2	0.4	0.3	0.2
Marko	0.3	0.7	1.1	0.6	1.1
Zorica	2.0	0.3	0.5	0.1	0.3
Miodrag	0.2	0.7	0.9	0.6	1.2
Danica	0.4	0.9	0.9	0.3	1.1
Ljubomir	1.2	1.1	0.9	0.9	0.5
Bojan	1.8	0.2	0.1	0.2	0.2
Dubravka	0.4	0.4	1.1	0.7	1.2
Milena	1.0	1.1	1.3	1.1	0.6



Slika 1.1 Dijagram rasturanja za X_1 i X_2

a) Zvezde

Pretpostavimo da raspolažemo podacima za $p \geq 2$ promenljivih. Za jedinicu posmatranja (na primer, osobu) formiramo zvezdu tako što za svaku promenljivu iz ishodišta povlačimo zrak čija je dužina proporcionalna vrednosti odnosne promenljive. Zrake povlačimo pod jednakim uglom međusobno, počevši sa prvom promenljivim "u tri sata" i krećući se u smeru suprotnom kretanju kazaljke na satu. Spajanjem krajeva zraka konačno formiramo zvezdu. Svaka zvezda dakle, reprezentuje jednu multivarijacionu opservaciju. Pored navedenog načina konstruisanja zvezde postoje brojne modifikacije osnovnog prikaza. Navedimo samo jedan primer takve modifikacije, poznat pod nazivom *sunčevi zraci* (eng. *sun rays*).



Slika 1.2 Zvezde

Za ovaj grafički prikaz potrebno je prethodno izvršiti standardizaciju opservacija. Zatim, povlačimo zrake jednake dužine iz centra u koji lociramo odnosnu najmanju standardizovanu opservaciju unutar posmatranog skupa podataka. Sredina svakog zraka

predstavlja lokaciju sredine odnosne promenljive. Ukoliko je vrh zvezde na tom zraku lociran ispod, odnosno iznad sredine zraka, to znači da je kod posmatrane osobe vrednost te promenljive ispodprosečna, odnosno iznadprosečna. Zvezde nam omogućavaju grupisanje osoba u relativno homogene grupe bez potrebe za korišćenjem formalizovanih postupaka grupisanja koji su inače sastavni elementi analize grupisanja.

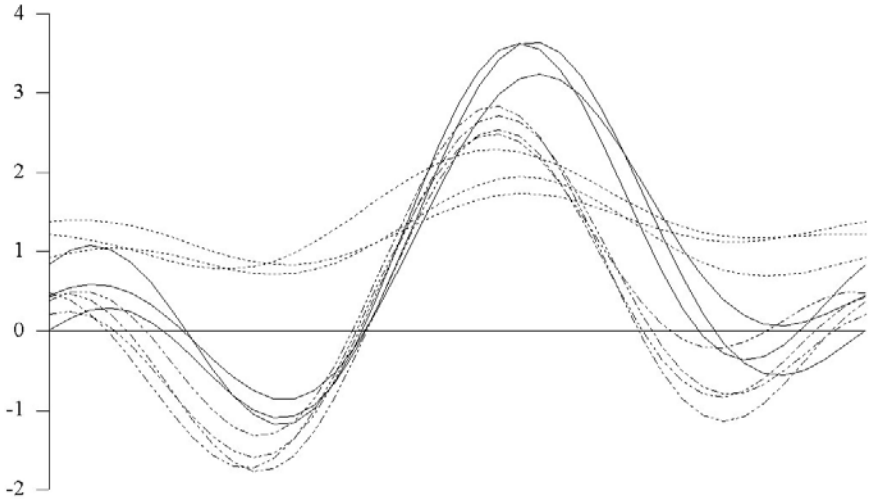
Podaci iz Tabele 1.2 su istovrsne veličine, pa nije potrebno vršiti njihovu standardizaciju. Ovim bi se postupkom izgubila informacija o međusobnoj srazmeri posmatranih promenljivih. Na Slici 1.2 prikazani su podaci iz Tabele 1.2 za deset osoba korišćenjem grafičkog prikaza u vidu zvezde. Na osnovu vizuelne sličnosti njihovih zvezda izdvajaju se tri grupe osoba: {Milena, Stojan, Ljubomir}, {Dubravka, Miodrag, Danica, Marko} i {Zorica, Bojan, Vladan}.

b) Andrewsove krive

Andrews (1972) je sugerisao da se p promenljivih za svaku jedinicu posmatranja prikaže konačnom Fourierovom serijom u kojoj će opservacije biti koeficijenti ove periodične funkcije:

$$f(t) = \frac{X_1}{\sqrt{2}} + X_2 \sin t + X_3 \cos t + X_4 \sin 2t + X_5 \cos 2t + ..., \; -\pi \le t \le \pi \qquad (1.1)$$

Grafički prikaz Fourierovih serija svakog multivarijacionog podatka biće krive koje se mogu, slično zvezdama, vizuelno grupisati. Ukažimo na dva svojstva Andrewsove krive: a) Andrewsova kriva sredine multivarijacionih opservacija jednaka je sredini Andrewsovih krivi pojedinih multivarijacionih opservacija i b) odstojanje između dve Andrewsove krive proporcionalno je Euklidskom odstojanju između odgovarajućih višedimenzionih tačaka. Kako se izmenom koeficijenata u izrazu (1.1) menja i Andrewsova kriva, praktična iskustva sugerišu da se eksperimentiše sa njihovim različitim redosledom. Takođe se sugeriše da se prethodno izvrši standardizacija podataka.



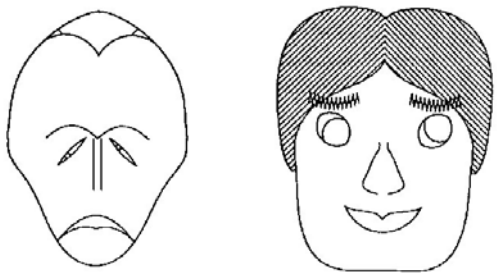
Slika 1.3 Andrewsove krive

Andrewsove krive mogu se, slično zvezdama, koristiti za subjektivno grupisanje osoba u relativno homogene grupe i za otkrivanje nestandardnih opservacija u skupu podataka.

Na osnovu podataka iz Tabele 1.2 (bez standardizacije i u originalnom redosledu promenljivih) izračunate su Andrewsove krive prezentirane na Slici 1.3. Tri krive koje leže iznad ose (uzimaju samo pozitivne vrednosti) odnose se na grupu {Zorica, Bojan, Vladan}. Punom linijom označene su krive koje se odnose na grupu {Milena, Stojan, Ljubomir}, a preostale krive sugerišu formiranje posebno izdvojene grupe {Dubravka, Miodrag, Danica, Marko} koja je ipak bliža drugoj nego prvo naznačenoj grupi.

c) Chernoffova lica

Chernoff (1973) je predložio prikazivanje multivarijacionih opservacija korišćenjem ljudskog lica čije karakteristike (oblik lica, zakrivljenost usta, dužina nosa, veličina očiju, položaj obrva itd.) su određene opservacijama promenljivih. Prvobitno predloženi grafički prikaz pretrpeo je brojne izmene, pa pri njegovom korišćenju, prema Flury i Riedwylu (1988, 4. glava), treba obratiti pažnju na sledeće. Potrebno je izvršiti standardizacija promenljivih na interval od nule do jedinice. Ekstremne vrednosti (nula, odnosno jedinica) svake od osamnaest promenljivih prikazani su na Slici 1.4 u vidu lica kod kojih je svaka od promenljivih, odnosno parametar pridružen karakteristici lica (na levoj i desnoj strani) u skladu sa Tabelom 1.3.

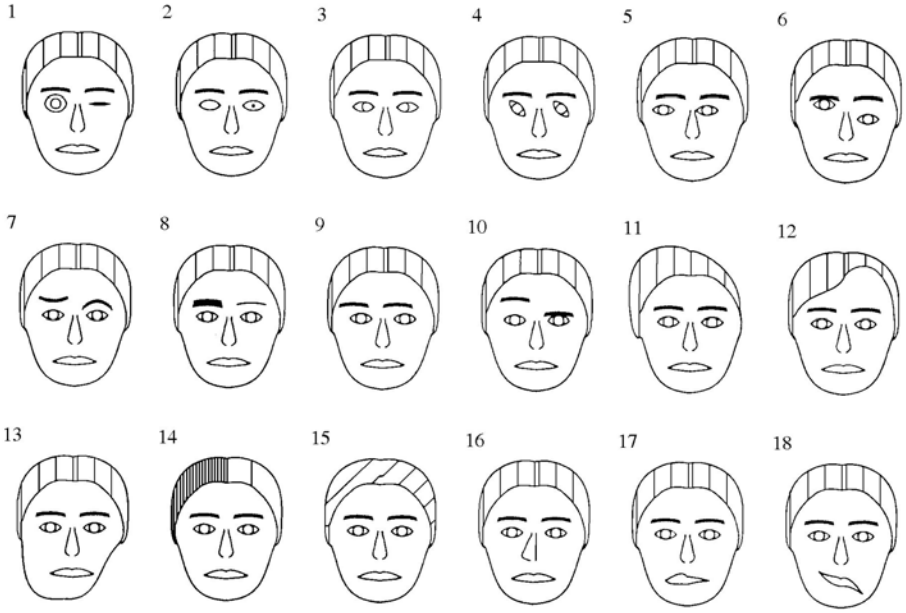


Slika 1.4 Chernoffova lica - ekstremni slučajevi (Izvor: Flury i Reidwyl (1981))

Tabela 1.3 Spisak karakteristika lica

Promenljiva	Karakteristika lica	Promenljiva	Karakteristika lica
1	veličina očiju	10	vertikalna položaj obrva
2	veličina zenice	11	gornja linija kose
3	položaj zenice	12	donja linija kose
4	iskošenost očiju	13	linija lica
5	horizontalna položaj očiju	14	zatamnjenost kose
6	vertikalna položaj očiju	15	iskošenost šrafure kose
7	zakrivljenost obrva	16	dužina nosa
8	debljina obrva	17	otvorenost usta
9	horizontalna položaj obrva	18	zakrivljenost usta

Za vrednost promenljivih između nule i jedinice odgovarajuća karakteristika lica dobijena je interpolacijom između ova dva ekstremna slučaja. Na Slici 1.5 prikazane su karakteristike lica iz Tabele 1.3, tako što je za svaku promenljivu individualno postavljena ekstremna vrednost, pri čemu su ostale karakteristike lica uzele srednju vrednost, odnosno 0.5. Tako je na primer kod prvog lica na njegovoj levoj strani samo prva promenljiva uzela vrednost 1, a ostalih sedamnaest su jednaki 0.5, dok je na desnoj strani vrednost prve promenljive jednaka 0, a ostalih sedamnaest su jednaki 0.5. Pregledom lica sa Slike 1.5 istraživač je u prilici da proceni koja karakteristika lica u najvećoj meri doprinosi opštem utisku i da u skladu sa tim pridruži značajniju promenljivu upravo toj karakteristici lica. Ovo pridruživanje predstavlja svojevrstan iterativan proces, sve dok se ne postigne zadovoljavajuća reprezentacija. Ukoliko smo sasvim siguran da su neke dve ili tri promenljive "odgovorne" za razliku, odnosno formiranje relativno homogenih grupa, tada te promenljive treba pridružiti važnijim karakteristikama lica (recimo dužina nosa, veličina očiju, zakrivljenosti usana i sl.).



Slika 1.5 Ekstremne vrednosti 18 parametara lica (Izvor: Flury i Reidwyl (1981))

Što se korišćenja Chernoffovih lica tiče važi isto ono što je rečeno kod zvezda, odnosno Andrewsovih krivi. Njihova primena, uz dobro pridruživanje promenljivih karakteristikama lica, olakšava formiranje grupa osoba ili objekata međusobno sličnih, objašnjenje eventualnih razlika između njih kao i otkrivanje nestandardnih opservacija. No, za razliku od prethodnih grafičkih prikaza crtanje lica je tehnički veoma zahtevan posao. Za njihovo korišćenje u statističkoj analizi zahteva se korišćenje računara, plotera i odgovarajućeg programa.

Na Slici 1.6 prikazani su podaci iz Tabele 1.2 za deset osoba korišćenjem grafičkog prikaza Chernoffova lica. Promenljivima u Tabeli 1.2 su pridružene sledeće karakteristike lica iz Tabele 1.3:

X_1: Linija lica (Promenljiva 13 u Tabeli 1.3)

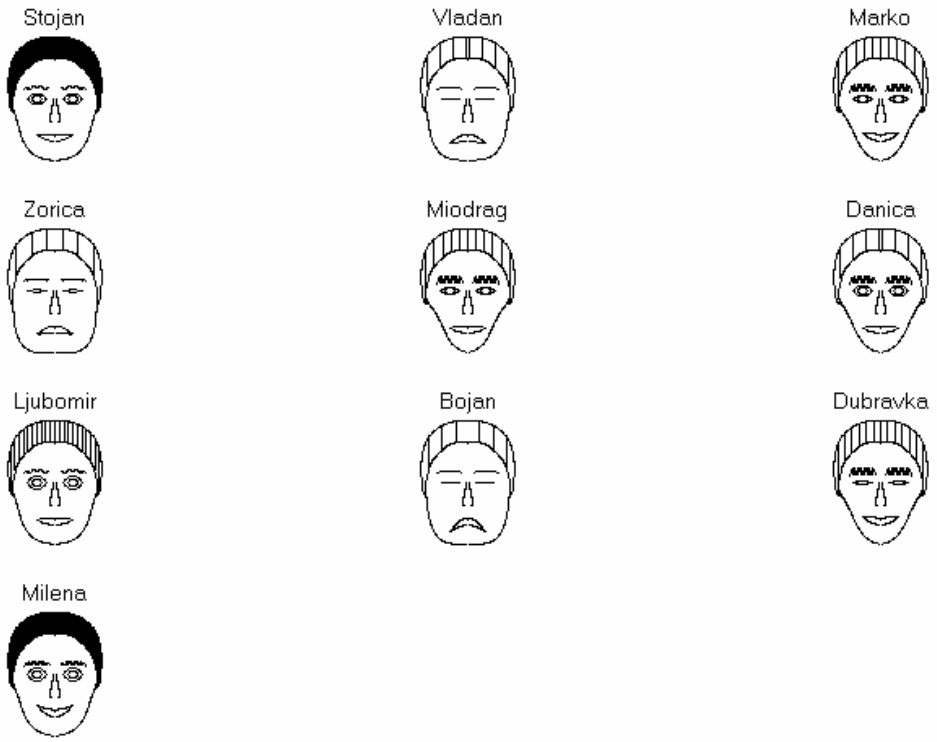
X_2: Veličina očiju (Promenljiva 1 u Tabeli 1.3)

X_3: Zakrivljenost usta (Promenljiva 18 u Tabeli 1.3)

X_4: Zatamnjenost kose (Promenljiva 14 u Tabeli 1.3)

X_5: Debljina obrva (Promenljiva 8 u Tabeli 1.3)

Na osnovu vizuelne sličnosti njihovih lica izdvajaju se tri grupe osoba: {Milena, Stojan, Ljubomir}, {Dubravka, Miodrag, Danica, Marko} i {Zorica, Bojan, Vladan}.



Slika 1.6 Chernoffova lica - Primer

Iscrpan prikaz navedenih i ostalih grafičkih prikaza u analizi multivarijacionih podataka može se naći kod Chambersa i Kleinera (1982), Clevelanda (1985) i Tuftea (1983).

2 VIŠEDIMENZIONI RASPOREDI

Višedimenzionost fenomena koje izučavamo metodama multivarijacione analize zahtevaju definisanje odgovarajućeg statističkog teorijskog modela. Kao što smo u slučaju jednog statističkog obeležja teorijsku osnovu gradili na konceptu slučajne promenljive i njoj pridružene funkcije rasporeda, tako u okviru ove glave uvodimo koncept višedimenzione slučajne promenljive i višedimenzione funkcije rasporeda. Preko njih zatim iskazujemo osnovne pokazatelje kako populacije tako i uzorka: vektor srednjih vrednosti, kovarijacionu i korelacionu matricu. Sledi zatim geometrijska interpretacija uzorka i njegovih osnovnih pokazatelja. Ona omogućava jednostavno objašnjenje i usvajanje osnovnih pojmova multivarijacione analize kao što su: generalizovano odstojanje i generalizovana varijansa.

2.1 VIŠEDIMENZIONE SLUČAJNE PROMENLJIVE

Označimo p jednodimenzionih slučajnih promenljivih sa $X_1, X_2, ..., X_p$. Skup ovih slučajnih promenljivih pišemo kao $(p \times 1)$ *slučajan vektor* $\mathbf{X}$

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

Znači da je slučajan vektor vektor čiji su elementi slučajne promenljive. Zbog kasnijih potreba napominjemo da kolekciju slučajnih vektora možemo predstaviti u vidu matrice. Takvu matricu čiji su elementi slučajne promenljive nazivamo *slučajna matrica*.

Što se korišćenih oznaka tiče, podsećamo da prema uobičajenim oznakama u statistici, slučajne vektore i slučajne matrice označavamo masnim velikim slovima, a malim masnim slovima njihove realizovane vrednosti.

2.1.1 Združeni rasporedi

Funkcija rasporeda $(p\times 1)$ slučajnog vektora $\mathbf{X}$, odnosno združena funkcija rasporeda slučajnih promenljivih $X_1, X_2, ..., X_p$ definiše se na sledeći način

$$F_{\mathbf{X}}(\mathbf{x}) = F_{\mathbf{X}}(x_1, x_2, ..., x_p) = P\left[\mathbf{X} \le \mathbf{x}\right] = P\left[X_1 \le x_1, X_2 \le x_2, ..., X_p \le x_p\right] \tag{2.1}$$

i ona u potpunosti definiše združenu raspodelu svih p slučajnih promenljivih. Napominjemo da izlaganje u ovoj glavi zasnivamo na neprekidnim slučajnim promenljivim. U prekidnom slučaju jedina izmena, pojednostavljeno rečeno, tiče se zamene oznake za integral oznakom za sumiranje kao što na primer u izrazu:

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_p} f_{\mathbf{X}}(\mathbf{x})dx_1\, dx_2 \ldots dx_p \tag{2.2}$$

koja definiše funkciju raspodele neprekidnog slučajnog vektora $\mathbf{X}$, prelaskom na prekidan slučaj integrale na desnoj strani treba zameniti sumama. U izrazu (2.2) $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_p)$ označava združenu, odnosno *višedimenzionu funkciju gustine* od $\mathbf{X}$. Do višedimenzione funkcije gustine slučajnog vektora $\mathbf{X}$ dolazimo diferencirajući višedimenzionu funkciju rasporeda $F_{\mathbf{X}}(\mathbf{x})$.

Višedimenziona funkcija gustine $f_{\mathbf{X}}(\mathbf{x})$ ima sledeće osobine: (1) $f_{\mathbf{X}}(\mathbf{x}) \ge 0$, za svako $\mathbf{x} \in R^p$ i (2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})dx_1\, dx_2 \ldots dx_p = 1$.

2.1.2 Marginalni rasporedi

Funkciju rasporeda ili funkciju gustine jedne slučajne promenljive dobijenu na osnovu višedimenzione funkcije rasporeda ili višedimenzione funkcije gustine nazivamo marginalna jednodimenziona funkcija rasporeda, odnosno marginalna jednodimenziona funkcija gustine. Pretpostavimo da nam je poznata funkcija rasporeda, $F_{\mathbf{X}}(\mathbf{x})$, tada do marginalne funkcije rasporeda slučajne promenljive X_1 dolazimo na osnovu izraza

$$F_{X_1}(x_1) = F_{\mathbf{X}}(x_1, \infty, \infty, ..., \infty) \tag{2.3}$$

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

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, x_2, ..., x_p)\, dx_2 \ldots\, dx_p \tag{2.4}$$

U opštem slučaju združenu funkciju rasporeda ili združenu funkciju gustine ma kog podskupa od $X_1, X_2, ..., X_p$ dobijenu na osnovu $F_{\mathbf{X}}(\mathbf{x})$ ili $f_{\mathbf{X}}(\mathbf{x})$, nazivamo *marginalnom*. Pretpostavimo da je $(p \times 1)$ slučajan vektor $\mathbf{X}$ podeljen na dva podskupa

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \text{ gde je } \mathbf{X}_1 (q \times 1), \text{ a } \mathbf{X}_2 (s \times 1) \text{ slučajan vector i } p = q + s. \tag{2.5}$$

Tada na osnovu funkcije rasporeda slučajnog vektora $\mathbf{X}$, $F_{\mathbf{X}}(\mathbf{x})$, možemo odrediti marginalnu funkciju rasporeda od $\mathbf{X}_1$, tj. $F_{\mathbf{X}_1}(\mathbf{x}_1)$, preko izraza

$$F_{\mathbf{X}_1}(\mathbf{x}_1) = F_{\mathbf{X}_1}(x_1, x_2, ..., x_q) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_q} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2)\, dx_1\, dx_2 \ldots dx_p \tag{2.6}$$

gde smo sa $f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2)$ označili združenu funkciju gustine od $\mathbf{X}$. Na osnovu združene funkcije gustine od $\mathbf{X}$, možemo odrediti združenu marginalnu funkciju gustinu od $\mathbf{X}_1$, tj. $f_{\mathbf{X}_1}(\mathbf{x}_1)$, preko izraza

$$f_{\mathbf{X}_1}(\mathbf{x}_1) = f_{\mathbf{X}_1}(x_1, x_2, ..., x_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2)\, dx_{q+1} \ldots dx_p \tag{2.7}$$

2.1.3 Uslovni rasporedi

Korišćenjem podele (2.5) slučajnog vektora $\mathbf{X}$, možemo odrediti raspored jednog podskupa slučajnih promenljivih za date vrednosti drugog podskupa. Takav raspored nazivamo *uslovni raspored*.

Uslovni raspored od $\mathbf{X}_2$ za dato $\mathbf{X}_1 = \mathbf{x}_1^*$ proporcionalan je $f_{\mathbf{X}}(\mathbf{x}_1^*, \mathbf{x}_2)$, gde se koeficijent proporcionalnosti može odrediti na osnovu činjenice da je integral ove uslovne funkcije gustine verovatnoće jednak jedinici. Na osnovu toga, uslovna funkcija gustine verovatnoće od $\mathbf{X}_2$ za dato $\mathbf{X}_1 = \mathbf{x}_1^*$ je

$$f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 \mid \mathbf{X}_1 = \mathbf{x}_1^*) = \frac{f_{\mathbf{X}}(\mathbf{x}_1^*, \mathbf{x}_2)}{f_{\mathbf{X}_1}(\mathbf{x}_1^*)} \tag{2.8}$$

gde je pretpostavljeno da je $f_{\mathbf{X}_1}(\mathbf{x}_1^*)$ različito od nule. Uslovnu funkciju gustine od $\mathbf{X}_1$ za dato $\mathbf{X}_2 = \mathbf{x}_2^*$ slično definišemo.

Dva slučajna vektora $\mathbf{X}_1$ i $\mathbf{X}_2$ su *nezavisni* ako i samo ako je

$$f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 \mid \mathbf{X}_1 = \mathbf{x}_1^*) = f_{\mathbf{X}_2}(\mathbf{x}_2^*) \text{ za svako dato } \mathbf{X}_1 = \mathbf{x}_1^* \text{ i svako } \mathbf{x}_2 \in R^s \tag{2.9}$$

Uslov nezavisnosti slučajnih vektora možemo izraziti i na osnovu jednakosti združene funkcije gustine i proizvoda marginalnih funkcija gustina

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}_1}(\mathbf{x}_1) f_{\mathbf{X}_2}(\mathbf{x}_2) \text{ za svako } \mathbf{x} \in R^p . \tag{2.10}$$

2.1.4 Sredina i kovarijaciona matrica

Neka je $\mathbf{X}(p\times 1)$ slučajan vektor, čiji svaki element predstavlja jednodimenzionu slučajnu promenljivu sa svojim marginalnim rasporedom. Za svaku jednodimenzionu slučajnu promenljivu možemo odrediti sredinu $\mu_j = E(X_j)$ i varijansu $\sigma_j^2 = E(X_j - \mu_j)^2$, $j = 1, 2, ..., p$. Zbog kasnijih potreba varijansu označavamo σ_{jj} i $Var(X_j)$. *Sredina slučajnog vektora* $\mathbf{X}$ je $(p\times 1)$ vektor čiji su elementi $\mu_j = E(X_j)$, $j = 1, 2, ..., p$ i označavamo ga sa $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}. \tag{2.11}$$

Za ma koji par slučajnih promenljivih X_j i X_k definišemo kovarijansu: $\sigma_{jk} = E[(X_j - \mu_j)(X_k - \mu_k)]$. Nju označavamo i kao $Cov(X_j, X_k)$, pri čemu je na osnovu definicije: $Cov(X_j, X_j) = Var(X_j)$ i $Cov(X_j, X_k) = Cov(X_k, X_j) = \sigma_{kj} = \sigma_{jk}$.

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

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

$$= \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_p) \\ Cov(X_1, X_2) & Var(X_2) & \cdots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_p) & Cov(X_2, X_p) & \cdots & Var(X_p) \end{bmatrix}. \tag{2.12}$$

Kovarijacionu matricu možemo iskazati i kao očekivanu vrednost slučajne matrice. Za slučajan vektor $\mathbf{X}$ sa sredinom $\boldsymbol{\mu}$ definišemo $(p\times p)$ simetričnu slučajnu matricu kvadrata, odnosno uzajamnih proizvoda odstupanja elemenata slučajnog vektora od odgovarajuće sredine

$$\begin{bmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_p-\mu_p) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p-\mu_p)(X_1-\mu_1) & (X_p-\mu_p)(X_2-\mu_2) & \cdots & (X_p-\mu_p)^2 \end{bmatrix}. \qquad (2.13)$$

Slučajna matrica (2.13) proizvod je slučajnih vektora odstupanja od sredine, tj. $(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'$, pa je njena očekivana vrednost

$$E\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\right]=\boldsymbol{\Sigma}\,. \qquad (2.14)$$

Često se tokom istraživanja javlja potreba da p promenljivih razdvojeno posmatramo, odnosno da ih tretiramo kao da pripadaju dvema grupama promenljivih. Metode zavisnosti su bazirane upravo na takvoj podeli promenljivih. Pretpostavimo da ispitujemo uticaj društveno- demografskih karakteristika potrošača na potrošnju izvesne grupe proizvoda. Tada ćemo naš $(p\times 1)$ vector $\mathbf{X}$ podeliti tako da u prvi podskup promenljivih uđu društveno-demografske promenljive (na primer pol, starost, školska sprema i sl.), a u drugi, promenljive potrošnje svakog proizvoda iz posmatrane grupe proizvoda. U skladu sa podelom (2.5) slučajnog vektora $\mathbf{X}$, vršimo podelu sredine $\boldsymbol{\mu}$ i kovarijacione matrice $\boldsymbol{\Sigma}$

$$\boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix} \qquad \text{i} \qquad \boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix} \qquad (2.15)$$

gde su $\boldsymbol{\mu}_1(q\times 1)$ i $\boldsymbol{\mu}_2(s\times 1)$ sredine slučajnih vektora $\mathbf{X}_1$ i $\mathbf{X}_2$ respektivno, a $\boldsymbol{\Sigma}_{11}(q\times q)$ i $\boldsymbol{\Sigma}_{22}(s\times s)$ su njihove kovarijacione matrice. Elementi matrice $\boldsymbol{\Sigma}_{12}(q\times s)$ su kovarijanse između elemenata slučajnih vektora $\mathbf{X}_1$ i $\mathbf{X}_2$ sa osobinom $\boldsymbol{\Sigma}_{12}=\boldsymbol{\Sigma}'_{21}$.

Navedimo nekoliko važnih rezultata vezanih za osobine sredine slučajnog vektora $\mathbf{X}$ i njegove kovarijacione matrice. Prethodno se podsetimo dobro poznatog svojstva kovarijanse. Neka su slučajne promenljive X_j i X_k linearno transformisane. To znači da su definisane nove slučajne promenljive cX_j+a i dX_k+b, gde su a, b, c i d realne konstante. Na osnovu definicije kovarijanse sledi

$$Cov(cX_j+a,\ dX_k+b)=cd\,Cov(X_j,X_k)\,. \qquad (2.16)$$

Jednostavno rečeno, znači da samo promena jedinice mere posmatranih promenljivih utiče u istom iznosu na kovarijansu. Uopštimo ovaj slučaj na linearnu kombinaciju p slučajnih promenljivih iz slučajnog vektora $\mathbf{X}$, sa sredinom $\boldsymbol{\mu}$ i kovarijacionom matricom $\boldsymbol{\Sigma}$. Linearnom kombinacijom $Y = a_1X_1 + a_2X_2 + ... + a_pX_p = \mathbf{a}'\mathbf{X}$, za dati vektor koeficijenata linearne kombinacije: $\mathbf{a}' = [a_1, a_2, ..., a_p]$, definišemo novu slučajnu promenljivu Y čija funkcija gustine $f_Y(y)$ zavisi od $f_{\mathbf{X}}(\mathbf{x})$. Njena očekivana vrednost (sredina) je

$$\mu_Y = E(Y) = E(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\boldsymbol{\mu}, \tag{2.17}$$

i varijansa

$$\sigma_Y^2 = Var(Y) = Var(\mathbf{a}'\mathbf{X}) = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j \sigma_{ij} = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}. \tag{2.18}$$

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

Razmotrimo opštiji slučaj q linearnih kombinacija p slučajnih promenljivih:

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \\ &\vdots \\ Y_q &= a_{q1}X_1 + a_{q2}X_2 + ... + a_{qp}X_p \end{aligned} \tag{2.19}$$

Ove linearne kombinacije su u matričnoj notaciji: $\mathbf{Y} = \mathbf{AX}$, gde je $\mathbf{Y}(q \times 1)$ vektor i $\mathbf{A}(q \times p)$ matrica koeficijenata linearnih kombinacija. Sredina slučajnog vektora $\mathbf{Y}$ je

$$\boldsymbol{\mu}_{\mathbf{Y}} = E(\mathbf{Y}) = E(\mathbf{AX}) = \mathbf{A}\boldsymbol{\mu}_{\mathbf{X}}, \tag{2.20}$$

a kovarijaciona matrica

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = Cov(\mathbf{Y}) = Cov(\mathbf{AX}) = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{A}'. \tag{2.21}$$

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

Naposletku, neka su date $(q \times p)$ matrica $\mathbf{A}$ i $(s \times r)$ matrica $\mathbf{B}$ čiji su elementi realni brojevi. Formiramo q linearnih kombinacija p slučajnih promenljivih, odnosno $\mathbf{Y} = \mathbf{AX}$ i s linearnih kombinacija r slučajnih promenljivih $\mathbf{Z}' = [Z_1, Z_2, ..., Z_r]$, odnosno $\mathbf{W} = \mathbf{BZ}$. Ako $\boldsymbol{\Sigma}_{\mathbf{XZ}}$ označava $(p \times r)$ matricu kovarijansi između $X_j, j = 1, 2, ..., p$ i $Z_k, k = 1, 2, ..., r$, tada je

$$Cov(\mathbf{Y},\mathbf{W}) = Cov(\mathbf{AX},\mathbf{BZ}) = \mathbf{A\Sigma_{XZ}B}'. \tag{2.22}$$

Na osnovu ovog najopštijeg rezultata svi prethodni mogu se tretirati kao njegovi specijalni slučajevi.

2.1.5 Korelaciona matrica

Koeficijent korelacije između dve slučajne promenljive X_j i X_k definišemo kao

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}} \tag{2.23}$$

što predstavlja normalizovanu kovarijansu između X_j i X_k. On uzima vrednost iz intervala od -1 do $+1$. Ukoliko ρ_{jk} uzme donju ili gornju graničnu vrednost, tada kažemo da postoji perfektna linearna veza između X_j i X_k i to sa negativnim $(\rho_{jk} = -1)$, odnosno pozitivnim predznakom $(\rho_{jk} = +1)$.

Korelacionu matricu $\boldsymbol{\rho}$ možemo dakle dobiti na osnovu poznate kovarijacione matrice, a njen $(j,k)-$ti element definisan je gornjim izrazom. U matričnoj notaciji veza između korelacione i kovarijacione matrice data je sa

$$\begin{aligned}
\boldsymbol{\rho} &= \left(\mathbf{D}^{1/2}\right)^{-1}\boldsymbol{\Sigma}\left(\mathbf{D}^{1/2}\right)^{-1} \\
&= \begin{bmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\sigma_{pp}}} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\sigma_{pp}}} \end{bmatrix} \\
&= \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}
\end{aligned} \tag{2.24}$$

gde smo sa $\mathbf{D}$ označili dijagonalnu matricu koja sadrži elemente na glavnoj dijagonali kovarijacione matrice Σ. Na osnovu uspostavljene relacije između korelacione i kovarijacione matrice, možemo ovu potonju pisati kao: $\boldsymbol{\Sigma} = \mathbf{D}^{1/2}\boldsymbol{\rho}\mathbf{D}^{1/2}$.

2.2 UZORAK IZ VIŠEDIMENZIONOG RASPOREDA

Matricu podataka[1] $\mathbf{X}$ dimenzija $(n \times p)$ možemo dvojako tretirati prema tome da li je posmatramo po redovima ili po kolonama. Ako je posmatramo po redovima, tada svaki red predstavlja jednu višedimenzionu opservaciju slučajnih promenljivih $X_1, X_2, ..., X_p$. U tom slučaju matrica podataka $\mathbf{X}$ sadrži n opservacija višedimenzione slučajne promenljive, u oznaci $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$, tako da matricu podataka možemo pisati kao

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \\ \vdots \\ \mathbf{X}'_n \end{bmatrix}.$$

Ako matricu podataka posmatramo po kolonama, tada svaka kolona predstavlja n opservacija jedne od p promenljivih, odnosno svaka kolona od $\mathbf{X}$ je $(n \times 1)$ vektor opservacija jedne od p promenljivih koji označavamo sa $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p$, gde je

$$X_j = \begin{bmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{bmatrix}, \quad j = 1, 2, ..., p \text{ , i } \mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p].$$

U zavisnosti od metode koja se izlaže nekad će nam interes biti da matricu podataka posmatramo po redovima, a nekad po kolonama. U prvom slučaju pod vektorom X_i podrazumevaćemo $(p \times 1)$ slučajni vektor opservacija promenljivih $X_1, X_2, ..., X_p$ za $i-$tu jedinicu posmatranja, a u drugom slučaju podrazumevaćemo $(n \times 1)$ vektor opservacija $i-$te promenljive.

2.2.1 Uzoračka sredina, kovarijaciona i korelaciona matrica

Bez obzira na formu funkcije gustine slučajnog vektora na osnovu matrice podataka, kao oblika prezentacije slučajnog uzorka uzetog iz višedimenzionog rasporeda, u mogućnosti smo da izračunamo pokazatelje uzorka: sredinu, kovarijacionu i korelacionu matrica. Ove pokazatelje koristimo u deskriptivne svrhe, a njihovim uvođenjem

[1] Za slučajan vektor i matricu njegovih realizovanih vrednosti (matrica podataka) koristimo istu oznaku. U toku izlaganja značenje korišćene oznake biće jasno iz konteksta. Isti način označavanja slučajnog vektora i matrice podataka koriste i autori kao što su npr. Karson (1982, s. 67) i Morrison (1976, s. 98).

pripremamo osnovu za statističko zaključivanje o pokazateljima populacije, odnosno funkcije gustine slučajnog vektora $\mathbf{X}$.

Uzoračka sredina, u oznaci $\bar{\mathbf{X}}$, definiše se kao $(p\times 1)$ slučajan vektor čiji su elementi uzoračke sredine odnosnih promenljivih

$$\bar{\mathbf{X}}=\begin{bmatrix}\bar{X}_1\\ \bar{X}_2\\ \vdots\\ \bar{X}_p\end{bmatrix},\ \text{gde je}\ \bar{X}_j=\frac{1}{n}\sum_{i=1}^{n}X_{ij},\ j=1,2,...,p\,. \tag{2.25}$$

Primer 2.1 Realizovane vrednosti slučajnog uzorka od $n=3$ elementa iz dvodimenzionog osnovnog skupa $(p=2)$ date su sledećom matricom podataka

$$\mathbf{X}=\begin{bmatrix}x_{11} & x_{12}\\ x_{21} & x_{22}\\ x_{31} & x_{32}\end{bmatrix}=\begin{bmatrix}-2 & 3\\ 4 & 2\\ 1 & 1\end{bmatrix}.$$

Odrediti realizovanu vrednost sredine $\bar{\mathbf{X}}$.

Prema definiciji sredine (2.25), a na osnovu realizovanih vrednosti uzorka imamo

$$\bar{\mathbf{x}}=\begin{bmatrix}\frac{1}{n}\sum_{j=1}^{n}x_{j1}\\ \frac{1}{n}\sum_{j=1}^{n}x_{j2}\end{bmatrix}=\begin{bmatrix}\frac{-2+4+1}{3}\\ \frac{3+2+1}{3}\end{bmatrix}=\begin{bmatrix}1\\ 2\end{bmatrix}$$

Znači da je sredina prve promenljive $\bar{x}_1=1$, a druge $\bar{x}_2=2$. ■

Uzoračku matricu uzajamnih proizvoda odstupanja od sredine, u oznaci $\mathbf{X}^{*\prime}\mathbf{X}^{*}$, definišemo kao $(p\times p)$ simetričnu slučajnu matricu

$$\mathbf{X}^{*\prime}\mathbf{X}^{*}=\begin{bmatrix}\sum_{i=1}^{n}(X_{i1}-\bar{X}_1)^2 & \sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2) & \cdots & \sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{ip}-\bar{X}_p)\\ \sum_{i=1}^{n}(X_{i2}-\bar{X}_2)(X_{i1}-\bar{X}_1) & \sum_{i=1}^{n}(X_{i2}-\bar{X}_2)^2 & \cdots & \sum_{i=1}^{n}(X_{i2}-\bar{X}_2)(X_{ip}-\bar{X}_p)\\ \vdots & \vdots & \ddots & \vdots\\ \sum_{i=1}^{n}(X_{ip}-\bar{X}_p)(X_{i1}-\bar{X}_1) & \sum_{i=1}^{n}(X_{ip}-\bar{X}_p)(X_{i2}-\bar{X}_2) & \cdots & \sum_{i=1}^{n}(X_{ip}-\bar{X}_P)^2\end{bmatrix}$$

gde je $\mathbf{X}^*$ matrica centriranih podataka

$$\mathbf{X}^* = \begin{bmatrix} X_{11} - \overline{X}_1 & X_{12} - \overline{X}_2 & \cdots & X_{1p} - \overline{X}_p \\ X_{21} - \overline{X}_1 & X_{22} - \overline{X}_2 & \cdots & X_{2p} - \overline{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} - \overline{X}_1 & X_{n2} - \overline{X}_2 & \cdots & X_{np} - \overline{X}_p \end{bmatrix}.$$

Preko matrice $\mathbf{X}^{*\prime}\mathbf{X}^*$ definišemo uzoračku kovarijacionu matricu, u oznaci $\mathbf{S}$, kao $(p \times p)$ simetričnu slučajnu matricu

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^{*\prime}\mathbf{X}^* = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})' \tag{2.26}$$

gde $(p \times 1)$ vector $(\mathbf{X}_i - \overline{\mathbf{X}})$ ima za elemente članove $i-$tog reda matrice $\mathbf{X}^*$.

Primer 2.2 Za podatke date u Primeru 2.1 odrediti matricu centriranih opservacija i matricu uzajamnih proizvoda centriranih opservacija, kao i realizovanu vrednost uzoračke kovarijacione matrice.

Na osnovu matrice originalnih podataka i izračunate sredine izračunavamo matricu centriranih opservacija

$$\mathbf{X}^* = \begin{bmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 \\ x_{31} - \overline{x}_1 & x_{32} - \overline{x}_2 \end{bmatrix} = \begin{bmatrix} -2-1 & 3-2 \\ 4-1 & 2-2 \\ 1-1 & 1-2 \end{bmatrix} = \begin{bmatrix} -3 & 1 \\ 3 & 0 \\ 0 & -1 \end{bmatrix}.$$

Matrica uzajamnih proizvoda centriranih opservacija je

$$\mathbf{X}^{*\prime}\mathbf{X}^* = \begin{bmatrix} -3 & 3 & 0 \\ 1 & 0 & -1 \end{bmatrix}\begin{bmatrix} -3 & 1 \\ 3 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 18 & -3 \\ -3 & 2 \end{bmatrix}$$

Konstatujemo da je matrica $\mathbf{X}^{*\prime}\mathbf{X}^*$ simetrična dimenzija $(p \times p)$, odn. (2×2). Realizovanu vrednost uzoračke kovarijacione matrice određujemo prema izrazu

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$$

Uočiti da $i-$ti sabirak u izrazu za realizovanu vrednost uzoračke kovarijacione matrice ima u razvijenom obliku izgled

$$\begin{bmatrix} x_{i1}-\bar{x}_1 \\ x_{i2}-\bar{x}_2 \end{bmatrix}\begin{bmatrix} x_{i1}-\bar{x}_1 & x_{i2}-\bar{x}_2 \end{bmatrix}$$

što znači da je $(\mathbf{x}_i-\bar{\mathbf{x}})'$ $i-$ti red matrice centriranih opservacija. Realizovana vrednost uzoračke kovarijacione matrice je prema tome

$$\mathbf{S}=\frac{1}{3-1}\left\{\begin{bmatrix}-3\\1\end{bmatrix}\begin{bmatrix}-3 & 1\end{bmatrix}+\begin{bmatrix}3\\0\end{bmatrix}\begin{bmatrix}3 & 0\end{bmatrix}+\begin{bmatrix}0\\-1\end{bmatrix}\begin{bmatrix}0 & -1\end{bmatrix}\right\}$$

$$=\frac{1}{2}\left\{\begin{bmatrix}9 & -3\\-3 & 1\end{bmatrix}+\begin{bmatrix}9 & 0\\0 & 0\end{bmatrix}+\begin{bmatrix}0 & 0\\0 & 1\end{bmatrix}\right\}=\frac{1}{2}\begin{bmatrix}18 & -3\\-3 & 2\end{bmatrix}=\begin{bmatrix}9 & -\frac{3}{2}\\-\frac{3}{2} & 1\end{bmatrix}$$

Pretposlednja matrica dobijena kod realizovane vrednosti uzoračke kovarijacione matrice jeste matrica uzajamnih proizvoda centriranih opservacija. Zato se kovarijaciona matrica može direktno preko nje izračunati. Iz kovarijacione matrice čitamo varijansu prve promenljive: $s_{11}=9$ i varijansu druge promenljive: $s_{22}=1$, dok je njihova kovarijansa $s_{12}=s_{21}=-3/2$. Znači da postoji negativna linearna veza ove dve promenljive. ■

Na osnovu $(j,k)-$tog elementa uzoračke kovarijacione matrice

$$S_{jk}=\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)(X_{ik}-\bar{X}_k),\ j,k=1,2,...,p\,, \tag{2.27}$$

možemo odrediti uzoračku korelacionu matricu čiji je $(j,k)-$ti element

$$r_{jk}=\frac{S_{jk}}{\sqrt{S_{jj}}\sqrt{S_{kk}}},\ j,k=1,2,...,p\,, \tag{2.28}$$

a sama korelaciona matrica je

$$\mathbf{R}=\begin{bmatrix}1 & r_{12} & r_{13} & \cdots & r_{1p}\\ r_{12} & 1 & r_{23} & \cdots & r_{2p}\\ r_{13} & r_{23} & 1 & \cdots & r_{3p}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ r_{1p} & r_{2p} & r_{3p} & \cdots & 1\end{bmatrix}. \tag{2.30}$$

Do uzoračke korelacione matrice mogli smo doći koristeći relaciju koju smo uspostavili između populacione kovarijacione i korelacione matrice. Kod uzoračke kovarijacione i korelacione matrice imamo relaciju: $\mathbf{R} = (\mathbf{D}^{1/2})^{-1}\mathbf{S}(\mathbf{D}^{1/2})^{-1}$.

Slično postupku dolaženja do uzoračke kovarijacione matrice preko matrice centriranih opservacija, $\mathbf{X}^*$, do uzoračke korelacione matrice možemo doći preko $(n \times p)$ matrice standardizovanih opservacija $\widetilde{\mathbf{X}}$, gde je

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \frac{X_{11}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{12}-\bar{X}_2}{\sqrt{S_{22}}} & \cdots & \frac{X_{1p}-\bar{X}_p}{\sqrt{S_{pp}}} \\ \frac{X_{21}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{22}-\bar{X}_2}{\sqrt{S_{22}}} & \cdots & \frac{X_{2p}-\bar{X}_p}{\sqrt{S_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{X_{n1}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{n2}-\bar{X}_2}{\sqrt{S_{22}}} & \cdots & \frac{X_{np}-\bar{X}_p}{\sqrt{S_{pp}}} \end{bmatrix},$$

pa je uzoračka korelaciona matrica data izrazom

$$\mathbf{R} = \frac{1}{n-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}. \tag{2.31}$$

Primer 2.3 Za podatke iz Primera 2.1 odrediti realizovanu vrednost uzoračke korelacione matrice koristeći relaciju uspostavljenu između nje i kovarijacione matrice.

Na osnovu izračunate kovarijacione matrice formiramo dijagonalnu matricu $\mathbf{D}$, a prema relaciji $\mathbf{R} = (\mathbf{D}^{1/2})^{-1}\mathbf{S}(\mathbf{D}^{1/2})^{-1}$ izračunavamo korelacionu matricu. Kako je

$$\mathbf{D} = \begin{bmatrix} 9 & 0 \\ 0 & 1 \end{bmatrix}$$

to je realizovana vrednost uzoračke korelacione matrice

$$\mathbf{R} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 9 & -\frac{3}{2} \\ -\frac{3}{2} & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{3} & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}.$$

Pokazati da se isti rezultat dobija ako se korelaciona matrica računa na osnovu matrice standardizovanih opservacija. ■

Pored matrice uzajamnih proizvoda centriranih opservacija, u oznaci $\mathbf{X}^{*\prime}\mathbf{X}^{*}$, i matrice uzajamnih proizvoda standardizovanih opservacija, u oznaci $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$, u multivarijacionoj analizi od interesa je definisati i matricu uzajamnih proizvoda originalnih opservacija, matricu koju ćemo označiti sa $\mathbf{X}'\mathbf{X}$. Zbir uzajamnog odstupanja dve promenljive od svojih sredina koji figuriše u izrazu za kovarijansu, možemo odrediti bez izračunavanja odstupanja za svaku opservaciju, tj. bez njihovog centriranja, što sledi na osnovu relacije

$$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}\,. \tag{2.32}$$

Višedimenzioni analogon relacije (2.32) sugeriše da je matricu uzajamnih proizvoda centriranih opservacija, $\mathbf{X}^{*\prime}\mathbf{X}^{*}$, moguće izraziti preko matrice uzajamnih proizvoda originalnih opservacija, $\mathbf{X}'\mathbf{X}$. Ovaj rezultat dobijamo polazeći od definicije matrice $\mathbf{X}^{*\prime}\mathbf{X}^{*}$ i na osnovu matričnog računa kao što sledi

$$\mathbf{X}^{*\prime}\mathbf{X}^{*}=\sum_{i=1}^{n}(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})'=\mathbf{X}'\mathbf{X}-n\bar{\mathbf{X}}\bar{\mathbf{X}}'\,. \tag{2.33}$$

Napominjemo da svi rezultati i relacije navedene za populacionu sredinu i kovarijacionu matricu, važe i u slučaju da se populacioni pokazatelji zamene odgovarajućim uzoračkim pokazateljima. Ova napomena odnosi se takođe i na podelu uzoračke sredine i kovarijacione matrice.

Primer 2.4 Izračunati sredine, varijanse i kovarijanse sledeće dve linearne kombinacije na osnovu podataka datih u Primeru 2.1

$$Y_1=\mathbf{a}_1'\mathbf{X}=\begin{bmatrix}a_{11} & a_{12}\end{bmatrix}\begin{bmatrix}X_1\\X_2\end{bmatrix}=\begin{bmatrix}-1 & 2\end{bmatrix}\begin{bmatrix}X_1\\X_2\end{bmatrix}=-X_1+2X_2$$

$$Y_2=\mathbf{a}_2'\mathbf{X}=\begin{bmatrix}a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}X_1\\X_2\end{bmatrix}=\begin{bmatrix}3 & -2\end{bmatrix}\begin{bmatrix}X_1\\X_2\end{bmatrix}=3X_1-2X_2$$

na dva načina: direktno na osnovu opservacija za Y_1 i Y_2 i indirektno na osnovu izračunate uzoračke sredine $\bar{\mathbf{x}}$ i izračunate uzoračke kovarijacione matrice $\mathbf{S}$.

Za prvu linearnu kombinaciju imamo realizovane vrednosti opservacija

$$y_{11}=\mathbf{a}_1'\mathbf{x}_1=-1(-2)+2(3)=8\,,$$

$$y_{21}=\mathbf{a}_1'\mathbf{x}_2=-1(4)+2(2)=0\,,$$

$$y_{31} = \mathbf{a}_1' \mathbf{x}_3 = -1(1) + 2(1) = 1,$$

pa je izračunata vrednost sredine $\bar{y}_1 = (8+0+1)/3 = 3$, a izračunata vrednost varijanse

$$s_{Y_1}^2 = \frac{(8-3)^2 + (0-3)^2 + (1-3)^2}{3-1} = 19.$$

Slično izračunavamo realizovane vrednosti opservacija druge linearne kombinacije. Dobijamo da je $y_{12} = -12$, $y_{22} = 8$ i $y_{32} = 1$. Na osnovu ovih vrednosti opservacija druge linearne kombinacije određujemo izračunatu vrednost sredine i varijanse: $\bar{y}_2 = -1$ i $s_{Y_2}^2 = 103$. Proveriti dobijeni rezultat. Izračunata vrednost kovarijanse dve linearne kombinacije je

$$Cov(Y_1, Y_2) = \frac{(8-3)(-12+1) + (0-3)(8+1) + (1-3)(1+1)}{3-1} = -43.$$

Do ovih rezultata mogli smo doći indirektno. Napišimo dve linearne kombinacije u matričnom obliku: $\mathbf{Y} = \mathbf{AX}$, gde je

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}.$$

Kako je $\bar{\mathbf{y}} = \mathbf{A}\bar{\mathbf{x}}$, sledi izračunata vrednost sredine

$$\bar{\mathbf{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 3 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}.$$

Izračunatu vrednost uzoračke kovarijacione matrice dobijamo na osnovu relacije $\mathbf{S_Y} = \mathbf{ASA}'$, odakle sledi

$$\mathbf{S_Y} = \begin{bmatrix} -1 & 2 \\ 3 & -2 \end{bmatrix} \begin{bmatrix} 9 & -\frac{3}{2} \\ -\frac{3}{2} & 1 \end{bmatrix} \begin{bmatrix} -1 & 3 \\ 2 & -2 \end{bmatrix} = \begin{bmatrix} 19 & -43 \\ -43 & 103 \end{bmatrix}.$$

Dakle, na ovaj indirektan način dobili smo rezultate koje smo ranije direktno izračunali na osnovu realizovanih vrednosti opservacija linearnih kombinacija. ■

Primer 2.5 Ilustracije radi razmotrimo sledeći hipotetički primer. Zavod za statistiku u svojoj redovnoj *Anketi o potrošnji domaćinstava* prikuplja podatke o njihovim prihodima i rashodima po različitim kategorijama za poljoprivredna, mešovita i nepoljoprivredna domaćinstva. Izabran je slučajan uzorak od $n = 6$ poljoprivrednih domaćinstava, i $p = 4$ promenljive su merene kod svakog domaćinstva: $X_1 =$ broj članova domaćinstva, $X_2 =$ godišnja raspoloživa sredstva, $X_3 =$ izdaci na ishranu i $X_4 =$ izdaci na obrazovanje i razonodu. Realizovane vrednosti slučajnog uzorka prikazane su u donjoj tabeli.

Poljoprivredno domaćinstvo	X_1	X_2	X_3	X_4
1	6	9.40	6.20	0.20
2	7	12.10	7.20	0.10
3	4	10.70	3.10	0.22
4	5	9.30	5.40	0.12
5	2	11.20	2.10	0.26
6	3	13.30	1.80	0.30

U narednoj tabeli prezentirani su međurezultati potrebni za određivanje izračunate uzoračke kovarijacione i korelacione matrice prema jednoj od navedenih formula. Kako su sve tri prezentirane matrice simetrične, to je prikazan samo njihov donji levi trougao.

Promenljiva	$\mathbf{X}'\mathbf{X}$				$\mathbf{X}^{*}{}'\mathbf{X}^{*}$				$\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$			
	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
X_1	139.0	-	-	-	17.50	-	-	-	5.000	-	-	-
X_2	292.7	738.08	-	-	-4.30	12.08	-	-	-1.479	5.000	-	-
X_3	136.6	276.25	136.70	-	20.50	-7.55	25.760	-	4.828	-2.140	5.00	-
X_4	4.8	13.46	4.38	0.2704	-0.60	0.262	-0.784	0.0304	-4.113	2.162	-4.43	5.000

Izračunata uzoračka sredina, kovarijaciona i korelaciona matrica prikazani su u narednoj tabeli. Koeficijenti korelacije dati su u donjem levom trouglu, a kovarijaciona matrica u gornjem desnom trouglu zajedno sa dijagonalom.

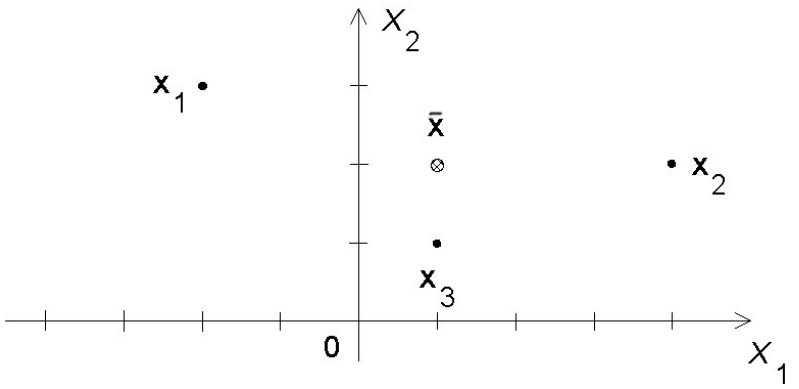
Promenljiva	**Sredina**	**Kovarijaciona i korelaciona matrica**			
		X_1	X_2	X_3	X_4
X_1	4.5	3.5000	-0.8600	4.1000	-0.12000
X_2	11.0	-0.2957	2.4160	-1.5100	0.05240
X_3	4.3	0.9655	-0.4280	5.1520	-0.15680
X_4	0.2	-0.8226	0.4323	-0.8859	0.00608

Na osnovu dobijenih rezultata može se konstatovati da postoji slabo izražena inverzna zavisnost između broja članova domaćinstva i godišnjih raspoloživih sredstava $(r_{12} = -0.2957)$, ali da istovremeno sa porastom broja članova domaćinstva rastu izdaci za ishranu $(r_{13} = -0.9655)$, a smanjuju se izdaci za obrazovanje i razonodu $(r_{14} = -0.8226)$. Interesantno je istaći negativnu vrednost koeficijenta korelacije između godišnjih raspoloživih sredstava i izdataka na ishranu $(r_{23} = -0.428)$. Ovaj fenomen smanjivanja izdataka na ishranu sa porastom dohotka dobro je poznat u teoriji potrošačke tražnje. Domaćinstvo, odnosno potrošač koji raspolaže relativno malim primanjima biće prinuđen da pretežan deo tih skromnih sredstava izdvaja na ishranu, i obratno, sa porastom prihoda biće u mogućnosti povećati izdvajanja za kulturne potrebe na račun relativno manjih izdataka na ishranu. ■

2.2.2 Geometrijska interpretacija uzorka

Matricu podataka možemo dvojako posmatrati sa stanovišta njene geometrijske interpretacije. Ako je posmatramo po redovima tada svaki red predstavlja jednu tačku u $p-$dimenzionom vektorskom prostoru. Stepen raspršenosti n tačaka u ovom prostoru zavisi od sredine i varijanse promenljivih, kao i njihove uzajamne povezanosti.

Primer 2.6 Grafički prikazati podatke iz Primera 2.1 na dijagramu rasturanja i nacrtati tačku koja predstavlja sredinu realizovanih vrednosti dvodimenzione promenljive. Komentarisati nacrtani dijagram rasturanja sa stanovišta stepena varijabiliteta pojedinih promenljivih kao i njihove međusobne povezanosti.



Na osnovu dijagrama rasturanja možemo konstatovati sledeće. Ucrtana tačka $\bar{\mathbf{x}}$ predstavlja težište skupa tačaka ili trougla čija su temena ove tri opservacije. Ukoliko svaku opservaciju posmatramo kao kraj vektora sa ishodištem u koordinatnom početku, tada se do vektora čiji je kraj lociran u tački $\bar{\mathbf{x}}$ dolazi jednostavnim sabiranjem tri vektora i množenjem rezultirajućeg vektora skalarom 1/3 (recipročna vrednost broja

opservacija). Zatim, uočavamo da je varijabilitet prve promenljive znatno veći od varijabiliteta druge promenljive (podsećamo da je računski dobijeno $s_{11}^2 = 9$ i $s_{22}^2 = 1$). Takođe se uočava negativna nagnutost skupa tačaka što sugeriše negativnu vrednost kovarijanse, odn. koeficijenta korelacije (izračunato je $r_{12} = -0.5$). ■

Rečeno je da se redovi matrice podataka odnose na osobu ili objekat. Prirodan način poređenja p realizovanih vrednosti opservacija u dva reda matrice podataka $\mathbf{x}_r$ i $\mathbf{x}_s$, jeste izračunavanje njihovog međusobnog odstojanja. Uobičajena definicija odstojanja jeste tzv. *Euklidsko odstojanje* zasnovano na Pitagorinoj teoremi. Prema njemu kvadrat odstojanja između dve tačke u $p-$dimenzionom prostoru dat je izrazom

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s) = \sum_{j=1}^{p} (x_{rj} - x_{sj})^2 \tag{2.34}$$

gde je $(\mathbf{x}_r - \mathbf{x}_s)$ $(p \times 1)$ vektor razlika koordinata $r-$tog i $s-$tog vektora. Kako se često sredina koristi kao reperna tačka u odnosu na koju posmatramo odstojanje ostalih tačaka uzorka to je kvadrat Euklidskog odstojanja vektora $\mathbf{x}_r$ u odnosu na sredinu, u oznaci d_r^2, dat izrazom

$$d_r^2 = (\mathbf{x}_r - \overline{\mathbf{x}})'(\mathbf{x}_r - \overline{\mathbf{x}}) = \sum_{j=1}^{p} (x_{rj} - \overline{x}_j)^2 . \tag{2.35}$$

Ukoliko smo odredili međusobno odstojanje između svaka dva para objekata, tada njihovu prezentaciju u formi $(n \times n)$ matrice nazivamo *matricom kvadrata Euklidskih odstojanja* i označavamo je sa $\mathbf{D}$

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12}^2 & d_{13}^2 & \cdots & d_{1n}^2 \\ d_{21}^2 & 0 & d_{23}^2 & \cdots & d_{2n}^2 \\ d_{31}^2 & d_{32}^2 & 0 & \cdots & d_{3n}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1}^2 & d_{n2}^2 & d_{n3}^2 & \cdots & 0 \end{bmatrix}. \tag{2.36}$$

Na osnovu definicije Euklidskog odstojanja sledi da je matrica $\mathbf{D}$ simetrična matrica sa nulama na glavnoj dijagonali. U narednom odeljku definisaćemo statističko odstojanje koje uzima u obzir varijansu i kovarijansu posmatranih promenljivih, pa je prema tome primerenije za statističko izučavanje slučajnog uzorka u $p-$dimenzionom prostoru. No, bez obzira na definiciju odstojanja, matrica odstojanja predstavlja osnovu analize grupisanja i drugih metoda multivarijacione analize koje su orijentisane ka redovima matrice podataka.

Primer 2.7 Na osnovu podataka iz Primera 2.1 odrediti Euklidsko odstojanje između tri objekta kao i njihovo Euklidsko odstojanje u odnosu na sredinu.

Odredimo Euklidsko odstojanje između $\mathbf{x}_1$ i $\mathbf{x}_2$. Prema definiciji kvadrat Euklidskog odstojanja između prvog i drugog objekta je

$$d_{12}^2 = (x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 = (-2-4)^2 + (3-2)^2 = 37 .$$

Znači da je Euklidsko odstojanje između $\mathbf{x}_1$ i $\mathbf{x}_2$ jednako $d_{12} = \sqrt{37}$. Proveriti rezultat ostalih Euklidskih odstojanja između objekata koja su prezentirana u matrici $\mathbf{D}$:

$$\mathbf{D} = \begin{bmatrix} 0 & \sqrt{37} & \sqrt{13} \\ \sqrt{37} & 0 & \sqrt{10} \\ \sqrt{13} & \sqrt{10} & 0 \end{bmatrix}.$$

Odstojanje d_1 ili odstojanje $\mathbf{x}_1$ od $\bar{\mathbf{x}}$ je prema definiciji

$$d_1^2 = (x_{11} - \bar{x}_1)^2 + (x_{12} - \bar{x}_2)^2 = (-2-1)^2 + (3-2)^2 = 10 ,$$

pa je $d_1 = \sqrt{10}$. Proveriti da je $d_2 = 3$ i $d_3 = 1$. ■

Ako posmatramo realizovane vrednosti u matrici podataka po kolonama $\mathbf{x}_j' = [x_{1j}, x_{2j}, ..., x_{nj}]$, $j = 1, 2, ..., p$, tada svaku kolonu prikazujemo kao jednu od p tačaka u $n-$dimenzionom prostoru. Svaka promenljiva predstavlja $n-$dimenzion vektor povučen iz koordinatnog početka do tačke čije su koordinate upravo elementi tog vektora. Dužina ovog vektora data je Euklidskim odstojanjem d_{j0}, koordinatnog početka i tačke koja definiše završetak vektora. Prema tome dužine vektora je

$$d_{j0} = \sqrt{(x_{1j} - 0)^2 + (x_{2j} - 0)^2 + ... + (x_{nj} - 0)^2} = \sqrt{\sum_{i=1}^{n} x_{ij}^2} = \sqrt{\mathbf{x}_j' \mathbf{x}_j}, \; j = 1, 2, ..., p . \qquad (2.37)$$

Pokažimo da uzoračka sredina, kovarijaciona matrica i koeficijent korelacije imaju jasnu geometrijsku interpretaciju u $n-$dimenzionom prostoru. U tom cilju definišimo $(n \times 1)$ vektor jedinica, $\mathbf{1}_n$, takav da je $\mathbf{1}_n' = [1, 1, ..., 1]$. Osobina mu je da zaklapa isti ugao sa svakom koordinatnom osom. Njegova dužina je jednaka $\sqrt{n}$, pa će vektor $(1/\sqrt{n})\mathbf{1}_n$ biti vektor jedinične dužine podjednako udaljen od svake koordinatne ose. Posmatrajmo projekciju vektora $\mathbf{x}_j$ na vektor $(1/\sqrt{n})\mathbf{1}_n$:

$$\mathbf{x}_j^{'}\left(\frac{1}{\sqrt{n}}\mathbf{1}_n\right)\left(\frac{1}{\sqrt{n}}\mathbf{1}_n\right)=\frac{x_{1j}+x_{2j}+\ldots+x_{nj}}{n}\mathbf{1}_n=\overline{x}_j\mathbf{1}_n. \tag{2.38}$$

Znači da sredina $j-$te promenljive, $\overline{x}_j=\mathbf{x}_j^{'}\mathbf{1}_n/n$ predstavlja skalar kojim treba pomnožiti vektor $\mathbf{1}_n$ da bi se dobila projekcija vektora $\mathbf{x}_j$ na pravu određenu sa $\mathbf{1}_n$. Kažemo da $\overline{x}_j\mathbf{1}_n$ predstavlja ortogonalnu projekciju vektora $\mathbf{x}_j$ na vector $\mathbf{1}_n$. Razliku vektora $\mathbf{x}_j$ i njegove projekcije $\overline{x}_j\mathbf{1}_n$ označimo sa $\mathbf{e}_j$, gde je

$$\mathbf{e}_j=\mathbf{x}_j-\overline{x}_j\mathbf{1}_n=\begin{bmatrix}x_{1j}-\overline{x}_j\\x_{2j}-\overline{x}_j\\\vdots\\x_{nj}-\overline{x}_j\end{bmatrix}. \tag{2.39}$$

Elementi vektora $\mathbf{e}_j$ su odstupanja pojedinačnih opservacija $j-$te promenljive od svoje sredine.

Da bismo ilustrovali uvedene koncepte vezane za geometrijsku interpretaciju vektora i pokazatelje uzorka poslužićemo se primerom koji je na granici trivijalnog. Naime, posmatraćemo uzorak od $n=2$ elementa sa $p=2$ promenljive. Iz osnovnog kursa statistike poznato je da u ovom slučaju koeficijent korelacije uzima vrednost ±1. Izabrali smo uzorak ove veličine jer se samo u ovom slučaju možemo poslužiti jasnim i preglednim grafičkim prikazom vektora.

Primer 2.8 Sledećom matricom podataka date su realizovane vrednosti slučajnog uzorka

$$\mathbf{X}=\begin{bmatrix}x_{11}&x_{12}\\x_{21}&x_{22}\end{bmatrix}=\begin{bmatrix}1&3\\5&1\end{bmatrix}.$$

Prikazati $p=2$ tačke u $n=2$ prostoru. Odrediti računski sredinu, a zatim grafički, korišćenjem projekcije dva vektora na vektor $\mathbf{1}_2$. Odrediti vektore odstupanja na dva načina (računski i grafički).

Kako je

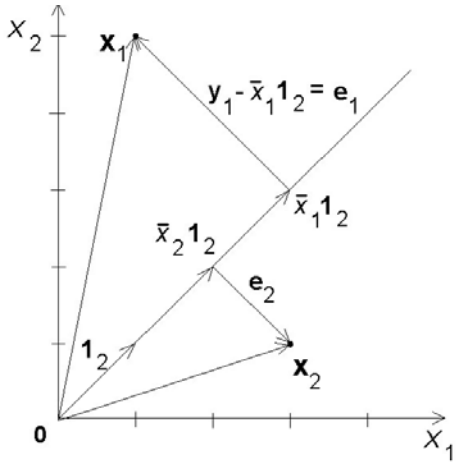
$$\mathbf{x}_1=\begin{bmatrix}1\\2\end{bmatrix}\ \text{i}\ \mathbf{x}_2=\begin{bmatrix}3\\1\end{bmatrix}$$

to je izračunata vrednost uzoračke sredine $\overline{\mathbf{x}}=\begin{bmatrix}3\\2\end{bmatrix}$.

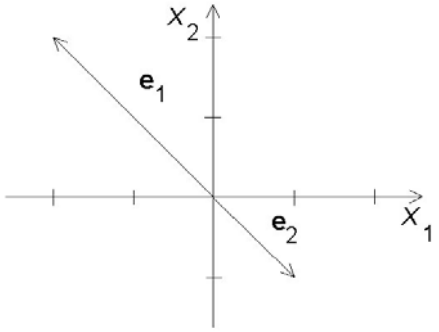
Sa grafikona čitamo da su vektori $\bar{x}_1\mathbf{1}_2 = \begin{bmatrix}3\\3\end{bmatrix}$ i $\bar{x}_2\mathbf{1}_2 = \begin{bmatrix}2\\2\end{bmatrix}$, pa je $\bar{x}_1 = 3$ i $\bar{x}_2 = 2$. Vektori odstupanja su

$$\mathbf{e}_1 = \mathbf{x}_1 - \bar{x}_1\mathbf{1}_2 = \begin{bmatrix}1\\5\end{bmatrix} - \begin{bmatrix}3\\3\end{bmatrix} = \begin{bmatrix}-2\\2\end{bmatrix}$$

$$\mathbf{e}_2 = \mathbf{x}_2 - \bar{x}_2\mathbf{1}_2 = \begin{bmatrix}3\\1\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix} = \begin{bmatrix}1\\-1\end{bmatrix}.$$



Ako grafički prikažemo vektore odstupanja, tako što ih translatorno pomerimo u koordinatni početak, dobićemo sledeću sliku



■

Pogledajmo čemu je jednak kvadrat dužine vektora odstupanja

$$d_j^2 = \mathbf{e}_j'\mathbf{e}_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2,\ j = 1, 2, ..., p\,. \tag{2.40}$$

Znači da je proporcionalan uzoračkoj varijansi $j-$te promenljive, odn. dužina je proporcionalna standardnoj devijaciji. Takođe, za ma koja dva vektora odstupanja $\mathbf{e}_j$ i $\mathbf{e}_k$ imamo unutrašnji ili skalarni proizvod

$$\mathbf{e}_j'\mathbf{e}_j = \sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k) \tag{2.41}$$

koji je proporcionalan kovarijansi između $j-$te i $k-$te promenljive. Odredimo čemu je jednak ugao između dva vektora odstupanja. Kosinus ugla između vektora $\mathbf{e}_j$ i $\mathbf{e}_k$ je po definiciji

$$\cos(\theta_{jk}) = \frac{\mathbf{e}_j'\mathbf{e}_k}{d_j d_k} = \frac{\mathbf{e}_j'\mathbf{e}_k}{\sqrt{\mathbf{e}_j'\mathbf{e}_j}\sqrt{\mathbf{e}_k'\mathbf{e}_k}} = \frac{S_{jk}}{\sqrt{S_{jj}}\sqrt{S_{kk}}} = r_{jk}\,, \tag{2.42}$$

što znači da je jednak uzoračkom koeficijentu korelacije, gde smo sa θ_{jk} označili ugao koji međusobom zaklapaju vektori odstupanja $\mathbf{e}_j$ i $\mathbf{e}_k$. Na osnovu osobine koeficijenta korelacije možemo zaključiti da ukoliko su vektori odstupanja međusobom ortogonalni, tada je koeficijent korelacije jednak 0, jer je njihov skalarni proizvod jednak nuli. U suprotnom, ako dva vektora odstupanja zaklapaju mali ugao, tada je i koeficijent korelacije blizak 1. Naposletku, ako vektori odstupanja leže na istom pravcu, ali su suprotno orijentisani, koeficijent korelacije jednak je -1. Upravo u takvom su međusobnom položaju vektori odstupanja u Primeru 2.8, pa je koeficijent korelacije jednak -1.

Primer 2.9 Na osnovu podataka iz Primera 2.8 izračunati uzoračku kovarijacionu i korelacionu matricu korišćenjem navedenih geometrijskih veličina.

Kako je $\mathbf{e}_1' = \begin{bmatrix}-2 & 2\end{bmatrix}$ i $\mathbf{e}_2' = \begin{bmatrix}1 & -1\end{bmatrix}$ to je

$$\mathbf{e}_1'\mathbf{e}_1 = \begin{bmatrix}-2 & 2\end{bmatrix}\begin{bmatrix}-2\\2\end{bmatrix} = 8$$

$$\mathbf{e}_2'\mathbf{e}_2 = \begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}1\\-1\end{bmatrix} = 2$$

$$\mathbf{e}_1'\mathbf{e}_2 = \begin{bmatrix}-2 & 2\end{bmatrix}\begin{bmatrix}1\\-1\end{bmatrix} = -4$$

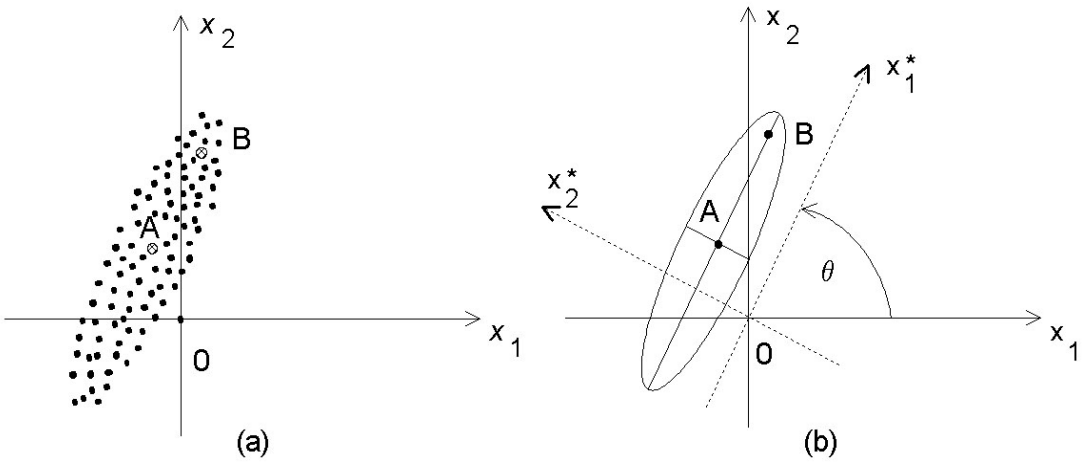
$$\cos(\theta_{12}) = r_{12} = \frac{\mathbf{e}_1'\mathbf{e}_2}{\sqrt{\mathbf{e}_1'\mathbf{e}_1}\sqrt{\mathbf{e}_2'\mathbf{e}_2}} = \frac{-4}{\sqrt{8}\sqrt{2}} = -1$$

Dobijen je, kao što rekosmo, očekivan rezultat. Između X_1 i X_2 postoji perfektna inverzna veza, što se moglo uočiti i na osnovu grafikona vektora odstupanja. Znači da je

$$\mathbf{S} = \begin{bmatrix} 8 & -4 \\ -4 & 2 \end{bmatrix} \quad \text{i} \quad \mathbf{R} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \qquad \blacksquare$$

2.2.3 Generalizovano odstojanje

Kod metoda multivarijacione analize koje su orijentisane ka redovima matrice podataka, odnosno ka ispitivanju međuzavisnosti objekata, osnovna veličina u analizi jeste odstojanje. Da bismo ukazali na neprimerenost korišćenja Euklidskog odstojanja u statističkoj analizi poslužićemo se jednostavnim dijagramom rasturanja za slučaj $p = 2$ promenljive.



Slika 2.1 Dijagram rasturanja i odstojanje
(a) Euklidsko i statističko odstojanje, (b) Rotacija koordinatnog sistema

Na Slici 2.1(a) uočavamo roj tačaka među kojima smo označili tačku njegovog težišta A, tačku B i koordinatni početka 0. Ako bismo odredili Euklidsko odstojanje od tačke A do B i od A do 0 dobili bi da je koordinatni početak bliži tački A nego što je to slučaj sa tačkom B. Međutim, posmatrano sa statističkog stanovišta tačka B pripada roju tačaka, za razliku od koordinatnog početka, i kao takva je sličnija tačkama unutar roja, pa i tački A nego što je to slučaj sa tačkom 0. Ako uzmemo u obzir različit varijabilitet posmatranih promenljivih (druga ima veći varijabilitet od prve) i njihovu korelaciju (očito je prisutna pozitivna korelisanost promenljivih) tada će se pokazati da je tačka B bliža u statističkom smislu tački A, no što je to koordinatni početak. Odstojanje koje uzima u obzir varijanse i kovarijansu promenljivih nazivamo *statističkim odstojanjem*.

Osnovnu ideju konstruisanja statističkog odstojanja izložićemo na osnovu Slike 2.1(b). Na njoj je elipsom predočen roj tačaka. Pozitivna orijentisanost roja tačaka u odnosu na osu x_1, odn. ugao θ, ukazuje na pozitivnu korelisanost promenljivih. Nju ćemo uzeti u obzir prilikom određivanja odstojanja tako što ćemo koordinatni sistem rotirati upravo za ugao θ. Zatim ćemo koordinatni početak tog rotiranog sistema pomeriti tako da se poklopi sa težištem roja tačaka (tačka A). Ovim pomeranjem koordinatnog početka u tačku A ustvari smo centrirali opservacije u novom koordinatnom sistemu. Posmatrajući roj tačaka u odnosu na novi koordinatni sistem, $x_1^* \mathrm{A}\, x_2^*$, možemo konstatovati da on svojim oblikom ukazuje na nekorelisanost promenljivih. Napominjemo još jednom da ova konstatacija važi ako roj tačaka posmatramo u odnosu na nove koordinatne ose. Na ovaj način smo rotacijom, odnosno promenom ugla gledanja na roj tačaka, "eliminisali" korelaciju promenljivih. Ono jedino što sada treba uzeti u obzir jeste različit varijabilitet promenljivih. Sa Slike 2.1(b) uočavamo da u apsolutnom smislu, velike vrednosti promenljive x_1 nisu neočekivane za razliku od velikih vrednosti promenljive x_2. Ovo je razumljivo ako imamo u vidu veći varijabilitet promenljive x_2 u odnosu na varijabilitet promenljive x_1. Zato je razumno pri izračunavanju odstojanja ponderisati većim ponderom koordinatu x_2 u odnosu na ponder koordinate x_1. Za pondere uzimamo recipročne vrednosti standardnih devijacija. Dakle, statističko odstojanje ma koje tačke uzorka od težišta roja tačaka predstavlja Euklidsko odstojanje računato na osnovu standardizovanih koordinata tačaka u novom koordinatnom sistemu

$$d = \sqrt{\left(\frac{x_1^* - \overline{x_1^*}}{\sqrt{S_{11}^*}}\right)^2 + \left(\frac{x_2^* - \overline{x_2^*}}{\sqrt{S_{22}^*}}\right)^2}\,. \tag{2.43}$$

Može se pokazati[2] da kvadrat odstojanja iskazan preko originalnih koordinata ima formu

$$d^2 = a_{11}\left(x_1 - \overline{x}_1\right)^2 + 2a_{12}\left(x_1 - \overline{x}_1\right)\left(x_2 - \overline{x}_2\right) + a_{22}\left(x_2 - \overline{x}_2\right)^2 \tag{2.44}$$

ili u matričnom obliku

$$d^2 = \begin{bmatrix} x_1 - \overline{x}_1 & x_2 - \overline{x}_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 - \overline{x}_1 \\ x_2 - \overline{x}_2 \end{bmatrix}, \tag{2.45}$$

gde su koeficijenti a_{11}, a_{12}, a_{22} takvi da je d nenegativno za ma koji par vrednosti x_1 i x_2. Ovi koeficijenti zavise od ugla θ i elemenata kovarijacione matrice S_{11}, S_{12} i S_{22}.

[2] Videti: Johnson i Wichern (1982, s. 24).

Sve tačke čije je odstojanje od težišta roja tačaka konstantno, zadovoljavaju jednakost (2.45). Ona predstavlja jednačinu elipse sa centrom u tački A. Ako na mesto matrice koeficijenata $\left[a_{ij}\right]$ stavimo inverznu uzoračku kovarijacionu matricu, tada ćemo definisati statističko odstojanje ili drugačije nazvano *Mahalanobisovo odstojanje*. Dakle, uopštenje standardnog Euklidskog odstojanja, odnosno kvadrat Mahalanobisovog odstojanja dat je izrazom

$$d^2 = \begin{bmatrix} x_1 - \overline{x}_1 & x_2 - \overline{x}_2 \end{bmatrix} \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \overline{x}_1 \\ x_2 - \overline{x}_2 \end{bmatrix} \tag{2.46}$$

Izraz za d^2, s jedne strane predstavlja jednačinu elipse, odnosno opisuje tačke dvodimenzionalnog prostora koje su podjednako udaljene od centra $(\overline{x}_1, \overline{x}_2)$ za fiksno d^2, a s druge strane predstavlja kvadrat Mahalanobisovog odstojanja proizvoljne tačke $\left(x_1, x_2\right)$ od tog centra.

Primer 2.10 Na osnovu podataka iz Primera 2.1 odrediti Euklidsko i Mahalanobisovo odstojanje tačaka od sredine. Napisati jednačinu elipse za $d^2 = 4/3$.

U Primeru 2.2 izračunali smo da je

$$\mathbf{X}^* = \begin{bmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 \\ x_{31} - \overline{x}_1 & x_{32} - \overline{x}_2 \end{bmatrix} = \begin{bmatrix} -3 & 1 \\ 3 & 0 \\ 0 & -1 \end{bmatrix} \text{ i}$$

$$\mathbf{S} = \begin{bmatrix} 9 & -\frac{3}{2} \\ -\frac{3}{2} & 1 \end{bmatrix}, \text{ odakle je } \mathbf{S}^{-1} = \begin{bmatrix} 1 & \frac{3}{2} \\ \frac{3}{2} & 9 \end{bmatrix}.$$

Prema definiciji kvadrat Euklidskog odstojanja $i-$te opservacije je

$$d_{ei}^2 = \begin{bmatrix} x_{i1} - \overline{x}_1 & x_{i2} - \overline{x}_2 \end{bmatrix} \begin{bmatrix} x_{i1} - \overline{x}_1 \\ x_{i2} - \overline{x}_2 \end{bmatrix}$$

na osnovu čega izračunavamo $d_{e1}^2 = 10$, $d_{e2}^2 = 9$ i $d_{e3}^2 = 1$.

Na osnovu definicije kvadrata Mahalanobisovog odstojanja imamo za prvu opservaciju $(-2,3)$

$$d_1^2 = \begin{bmatrix} -3 & 1 \end{bmatrix} \frac{4}{27} \begin{bmatrix} 1 & \frac{3}{2} \\ \frac{3}{2} & 9 \end{bmatrix} \begin{bmatrix} -3 \\ 1 \end{bmatrix} = \frac{4}{3}.$$

Slično računamo $d_2^2 = \frac{4}{3}$ i $d_3^2 = \frac{4}{3}$. Ovaj rezultat znači da su u statističkom smislu sve tri tačke na istoj udaljenosti od sredine.

Na osnovu izraza

$$\begin{bmatrix} x_1 - \overline{x}_1 & x_2 - \overline{x}_2 \end{bmatrix} \begin{bmatrix} s^{11} & s^{12} \\ s^{21} & s^{22} \end{bmatrix} \begin{bmatrix} x_1 - \overline{x}_1 \\ x_2 - \overline{x}_2 \end{bmatrix} = \frac{4}{3},$$

gde smo sa s^{ij} označili $(i, j)-$element matrice $\mathbf{S}^{-1}$, i izračunatih elemenata, imamo

$$\begin{bmatrix} x_1 - 1 & x_2 - 2 \end{bmatrix} \begin{bmatrix} \frac{4}{27} & \frac{2}{9} \\ \frac{2}{9} & \frac{4}{3} \end{bmatrix} \begin{bmatrix} x_1 - 1 \\ x_2 - 2 \end{bmatrix} = \frac{4}{3},$$

odakle dobijamo jednačinu elipse

$$\frac{4}{27}(x_1 - 1)^2 \frac{4}{9}(x_1 - 1)(x_2 - 2) + \frac{4}{3}(x_2 - 2)^2 = \frac{4}{3}.$$

Centar elipse je u tački $(1, 2)$. Proveriti da sve tri tačke uzorka: $(-2, 3)$, $(4, 2)$ i $(1,1)$ zadovoljavaju jednačinu elipse, odnosno da leže na njoj.

■

Do sada smo razmatrali dvodimenzioni, ali je uopštenje na $p-$dimenzioni slučaj direktan i očigledan na osnovu definicije kvadrata Mahalanobisovog odstojanja. Naime za tačku $\mathbf{x}' = [x_1, x_2, ..., x_p]$ kažemo da je njeno Mahalanobisovo odstojanje u odnosu na sredinu $\overline{\mathbf{x}}' = [\overline{x}_1, \overline{x}_2, ..., \overline{x}_p]$ dato sa d, gde je

$$d^2 = (\mathbf{x} - \overline{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x} - \overline{\mathbf{x}}) \tag{2.47}$$

Izrazom (2.47) istovremeno je opisan i *elipsoid* u $p-$dimenzionom prostoru za fiksnu vrednost Mahalanobisovog odstojanja.

Primer 2.11 Na osnovu podataka iz Primera 2.5 odrediti za svako domaćinstvo Mahalanobisovo odstojanje.

Kvadrat Mahalanobisovog odstojanje svake višedimenzione opservacije od sredine dat je u narednoj tabeli

Domaćinstvo	1	2	3	4	5	6
Kvadrat Mahalanobisovog odstojanja	4.16603	3.86646	2.86753	2.57068	3.63525	2.89405

Na osnovu dobijenih vrednosti kvadrata Mahalanobisovog odstojanja možemo konstatovati da je opservacija za prvo domaćinstvo najviše udaljena od središta ovog skupa višedimenzionih opservacija dok je opservacija za četvrto domaćinstvo najbliža središtu podataka. ■

2.2.4 Generalizovana varijansa

Jedan od najčešće korišćenih pokazatelja stepena raspršenosti podataka u slučaju jedne promenljive jeste varijansa, ili standardna devijacija. Za $p \geq 2$ promenljive definisana je kovarijaciona matrica koja pruža informaciju o varijansi i kovarijansi promenljivih, ali na osnovu $p(p+1)/2$ elemenata. Interes nam je da u višedimenzionom slučaju definišemo sintetički pokazatelj koji će na osnovu jednog broja iskazati stepen varijabiliteta $p-$dimenzionog skupa podataka. Koriste se dve definicije generalizovane varijanse. Prema prvoj, češće korišćenoj definiciji, ona je u uzorku jednaka determinanti uzoračke kovarijacione matrice, u oznaci $|\mathbf{S}|$, a prema drugoj generalizovana varijansa je jednaka tragu uzoračke kovarijacione matrice (zbir elemenata na glavnoj dijagonali matrice), u oznaci $\text{tr}(\mathbf{S})$. Kao što će se pokazati, kod velikog broja statističkih testova u multivarijacionoj analizi koristi se upravo prva definicija generalizovane varijanse, odnosno $|\mathbf{S}|$.

Primer 2.12 Na osnovu uzoračkih kovarijacionih matrica iz Primera 2.2, Primera 2.5 i Primera 2.9 odrediti generalizovanu varijansu prema prvoj i drugoj definiciji.

(a) Primer 2.2:

, $|\mathbf{S}| = 9(1) - \left(-\frac{3}{2}\right)\left(-\frac{3}{2}\right) = -\frac{27}{4}$, $\text{tr}(\mathbf{S}) = 9+1 = 10$

(b) Primer 2.5:

. $|\mathbf{S}| = 0.002192$, $\text{tr}(\mathbf{S}) = 11.07408$

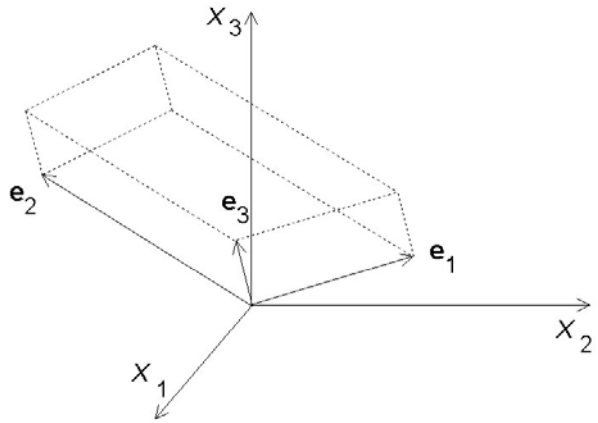
(c) Primer 2.9:

$|\mathbf{S}| = 0$, $\text{tr}(\mathbf{S}) = 10$ ■

Sa stanovišta geometrijske interpretacije generalizovane varijanse ona se može dvojako tumačiti. Ako posmatramo $n-$dimenzioni prostor tada se može pokazati[3] da je zapremina tela generisanog vektorima odstojanja (Slika 2.2) u relaciji sa generalizovanom varijansom. Naime, u slučaju p vektora odstupanja imamo da je

$$|\mathbf{S}| = (n-1)^{-p}(\text{zapremina})^2 . \tag{2.48}$$

Očito je da će se zapremina, a time i generalizovana varijansa povećavati sa porastom dužine vektora odstupanja, s jedne strane i povećanjem ugla koji oni zaklapaju međusobom, s druge strane. Kada su vektori odstupanja međusobno pod pravim uglom tada, uz fiksnu veličinu uzorka, imamo maksimalnu generalizovanu varijansu. Upravo je taj slučaj prikazan na Slici 2.2. U suprotnom, kada postoji visoka kolinearnost promenljivih $|\mathbf{S}|$ biće blisko nuli (videti Primer 2.12(b)), a u ekstremnom slučaju perfektne zavisnosti promenljivih $|\mathbf{S}| = 0$ (Primer 2.12(c)).



Slika 2.2 Zapremina tela za $p = 3$

Alternativna geometrijska interpretacija generalizovane varijanse daje se u $p-$dimenzionom prostoru. Koristeći se Mahalanobisovim odstojanjem, koordinate tačaka $\mathbf{x}' = [x_1, x_2, ..., x_p]$ koje su na konstantnoj udaljenosti c od $\bar{\mathbf{x}}$ zadovoljavaju jednačinu elipsoida

$$(\mathbf{x} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) = c^2 . \tag{2.49}$$

Zapremina ovog elipsoida (za $p = 2$ reč je o površini elipse) u relaciji je sa generalizovanom varijansom $|\mathbf{S}|$. Na osnovu te relacije generalizovana varijansa je[4]

---

[3] Videti: Anderson (1972, ss. 176-8) i Johnson i Wichern (1982, s. 104).

[4] Videti: Anderson (1972, ss. 168-170) i Johnson i Wichern (1982, s. 106).

$$|\mathbf{S}| = c^{-2p} K(p)(\text{zapremina elipsoida})^2 \tag{2.50}$$

gde je $K(p)$ konstanta koja zavisi od p. Povećanjem zapremine elipsoida povećava se i vrednost generalizovane varijanse.

Nedostatak generalizovane varijanse kao sintetičkog pokazatelja varijabiliteta višedimenzionog skupa podataka ogleda se u tome što ona zanemaruje kovarijacionu strukturu podataka. Tako se za različite kovarijacione matrice može dobiti ista vrednost generalizovane varijanse. Zato se preporučuje da se pored njene vrednosti prezentira i sama kovarijaciona matrica.

Generalizovanu varijansu možemo odrediti i na osnovu korelacione matrice i to prema prvoj definiciji kao $|\mathbf{R}|$. Ako su svi elementi van glavne dijagonale korelacione matrice jednaki nuli tada je $|\mathbf{R}| = 1$. Sa povećanjem vrednosti vandijagonalnih elemenata vrednost $|\mathbf{R}|$ se smanjuje, a praktično je jednaka nuli kada postoji perfektna linearna zavisnost između dve promenljive. Znači da generalizovana varijansa bazirana na korelacionoj matrici predstavlja meru odsustva korelacije između promenljivih. U pogledu geometrijske interpretacije veličine $|\mathbf{R}|$ ona je istovetna izloženoj interpretaciji veličine $|\mathbf{S}|$, uz jedinu razliku da se koristimo vektorom standardizovanih promenljivih, a ne vektorom odstupanja (centriranih opservacija).

Veza koja postoji između $|\mathbf{S}|$ i $|\mathbf{R}|$ data je izrazom

$$|\mathbf{S}| = S_{11} S_{22} \cdots S_{pp} |\mathbf{R}| . \tag{2.51}$$

gde je $S_{jj}, j = 1, 2, ..., p$ uzoračka standardna devijacija $j-$te promenljive.

3 VIŠEDIMENZIONI NORMALAN RASPORED

Jednodimenzioni normalan raspored je od posebnog značaja u statističkom zaključivanju. Brojne metode zaključivanja polaze od pretpostavke da je raspored normalan, što omogućava relativno jednostavno određivanje osobina i rasporeda ocena nepoznatih parametara ili test statistika. Ovo stoga što je normalan raspored u potpunosti određen sa dva parametra (prvi i drugi momenat). Pored ove poželjne matematičke osobine normalnog rasporeda, njegov praktičan značaj ogleda se u tome da se on javlja kao prirodan statistički model mnogih fenomena u stvarnosti. Naposletku, on predstavlja dobru aproksimaciju rasporeda uzoračkih statistika bez obzira na raspored populacije odakle je uzorak uzet, što je posledica centralne granične teoreme.

Onu ulogu i značaj koji ima jednodimenzioni normalan raspored u analizi jedne slučajne promenljive, ima višedimenzioni normalan raspored kod slučajnog vektora. Kod većeg broja multivarijacionih metoda pretpostavka koja omogućava definisanje postupka statističkog zaključivanja tiče se višedimenzionog normalnog rasporeda. Njime se detaljno bavimo u ovoj glavi pripremajući tako osnovu za izlaganje metoda multivarijacione analize u narednim glavama. Izlaganje ćemo završiti statističkim postupkom zaključivanja o sredini, kovarijacionoj i korelacionoj matrici.

3.1 FUNKCIJA GUSTINE VEROVATNOĆE

Funkciju gustine verovatnoće normalno raspoređene slučajne promenljive X pišemo kao

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \; -\infty < x < \infty, \tag{3.1}$$

gde je μ sredina, a σ^2 varijansa slučajne promenljive X. Raspored slučajne promenljive X označavamo $N(\mu, \sigma^2)$. Ako drugačije napišemo izraz u eksponentu funkcije gustine verovatnoće

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu) \tag{3.2}$$

tada kvadrat odstupanja $\mathbf{x}$ od $\boldsymbol{\mu}$ u jedinicama standardne devijacije u $p-$dimenzionom slučaju pišemo kao

$$(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{3.3}$$

odnosno, reč je o generalizovanom odstojanju od $\mathbf{x}$ do $\boldsymbol{\mu}$, gde smo sa $\boldsymbol{\mu}$ označili $(p\times 1)$ vektor očekivane vrednosti slučajnog vektora $\mathbf{X}$, a sa Σ njegovu $(p\times p)$ kovarijacionu matricu.

Funkcija gustine verovatnoće višedimenzione normalno raspoređene slučajne promenljive $\mathbf{X}'=[X_1, X_2, ..., X_p]$ data je izrazom

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \text{ gde je } -\infty < x_j < \infty,\ j=1,2,...,p \tag{3.5}$$

ili u oznaci $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Vrednost konstante u izrazu za gustinu verovatnoće $p-$dimenzionog normalnog rasporeda odredili smo na osnovu uslova da višestruki integral funkcije gustine verovatnoće nad celim dimenzionim prostorom bude jednak jedinici.

Primer 3.1 Za slučaj $p=2$ napisati funkciju gustine verovatnoće normalno raspoređenog slučajnog vektora $\mathbf{X}'=\left[X_1, X_2\right]$.

Kako je $\boldsymbol{\mu}=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}$ i $\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$ to je

$$\boldsymbol{\Sigma}^{-1}=\frac{1}{1-\rho^2}\begin{bmatrix}\frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2}\\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2}\end{bmatrix} \text{ i } |\boldsymbol{\Sigma}|=\sigma_1^2\sigma_2^2(1-\rho^2).$$

Generalizovano odstojanje $\left(\mathbf{x}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)$ je

$$\frac{1}{1-\rho^2}\begin{bmatrix}x_1-\mu_1 & x_2-\mu_2\end{bmatrix}\begin{bmatrix}\frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2}\\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2}\end{bmatrix}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}$$

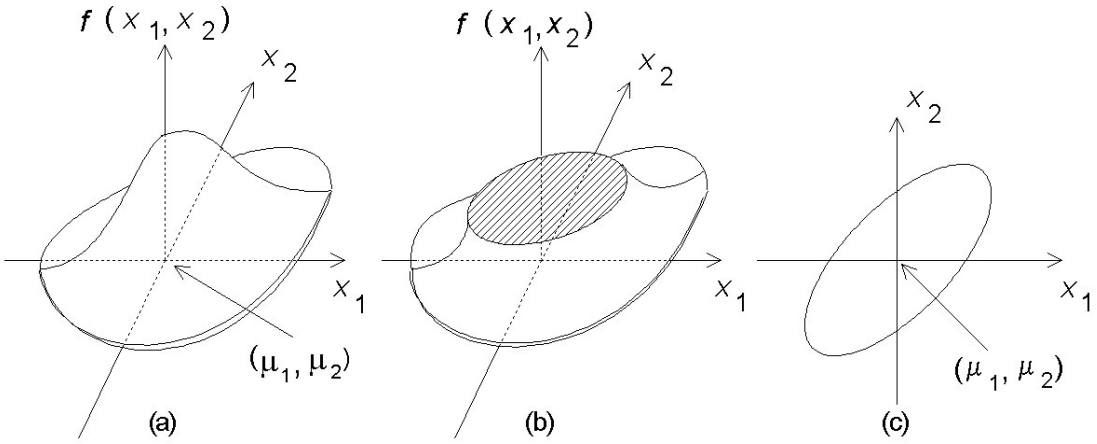
$$= \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]$$

Funkcija gustine verovatnoće je

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

Na Slici 3.1(a) prikazana je funkcija gustine verovatnoće dvodimenzione normalno raspoređene slučajne promenljive za slučaj $\sigma_1^2 = \sigma_2^2$ i $\rho > 0$.



Slika 3.1 Dvodimenzioni normalan raspored
(a) Funkcija gustine verovatnoće, (b) Presek funkcije gustine verovatnoće,
(c) Kontura konstantne gustine verovatnoće

Funkcija gustine verovatnoće biće konstantna ako je $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$, ($c$ je konstanta), gde navedeni izraz predstavlja kvadrat Mahalanobisovog odstojanja. Ovaj izraz znači da smo funkciju gustine verovatnoće presekli na visini c^2 sa ravni koja je paralelna ravni u kojoj leže koordinatne ose x_1 i x_2 (Slika 3.1(b)). Šrafirana površina sugeriše oblik tog preseka, a njegova projekcija u ravan u kojoj su ose x_1 i x_2 prikazana je na Slici 3.1(c). Dobijena je elipsa sa centrom u tački (μ_1, μ_2). Kasnije ćemo ukazati na direktnu vezu kovarijacione matrice i osa elipse, a sada pogledajmo kako se menja izgled elipse sa promenom njenih osa. Uočavamo da glavna, duža osa elipse prolazi kroz

prvi i treći kvadrant pod uglom od 45^0, ukazujući na postojanje pozitivne korelacije izmedu X_1 i X_2. Druga osa je pod uglom od 90^0 u odnosu na prvu. Što je prva osa relativno duža u odnosu na drugu, to je veći, po apsolutnoj vrednosti, koeficijent korelacije. U ekstremnom slučaju, kada je $\rho = \pm 1$, dobijamo pravu liniju jer je dužina sporedne ose jednaka nuli. Povećavajući dužinu sporedne ose, približavajući je time dužini glavne ose, elipsa postaje krug, a koeficijent korelacije $\rho = 0$. ■

U opštem slučaju dobijamo rezultat da je kvadratom Mahalanobisovog odstojanja izmedu $\mathbf{x}$ i $\boldsymbol{\mu}$ definisan omotač elipsoida čiji je centar u $\boldsymbol{\mu}$. Funkcija gustine verovatnoće od $\mathbf{X}$ je konstantna po omotaču elipsoida. Taj omotač nazivamo *kontura konstantne gustine verovatnoće*. Kao i u jednodimenzionom slučaju maksimum funkcije gustine verovatnoće postiže se u tački $\mathbf{x} = \boldsymbol{\mu}$.

Prva glavna osa elipsoida je linija koja prolazi kroz njenu najveću dimenziju. Da bismo odredili njen pravac i dužinu posmatraćemo liniju koja prolazi kroz tačku $\boldsymbol{\mu}$ i seče konturu konstantne gustine verovatnoće u tački $\mathbf{x}$. Prvu osu odredićemo tako što ćemo maksimizirati kvadrat njene poludužine $(\mathbf{x}-\boldsymbol{\mu})'(\mathbf{x}-\boldsymbol{\mu})$ uz ograničenje da tačka $\mathbf{x}$ leži na konturi, tj. da je $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$. Korišćenjem metode Lagranžovih multiplikatora odredićemo vrednost koja maksimizira sledeću funkciju

$$\varphi(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu})'(\mathbf{x}-\boldsymbol{\mu}) - \lambda\left[(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - c^2\right]. \tag{3.6}$$

Vektor parcijalnih izvoda je

$$\frac{\partial\varphi(\mathbf{x})}{\partial\mathbf{x}} = 2(\mathbf{x}-\boldsymbol{\mu})' - \lambda\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}), \tag{3.7}$$

pa koordinate najduže ose moraju zadovoljiti sledeću jednačinu

$$[\mathbf{I} - \lambda\boldsymbol{\Sigma}^{-1}](\mathbf{x}-\boldsymbol{\mu}) = 0 \tag{3.8}$$

ili, što je ekvivalentno jednačini (zbog nesingularnosti kovarijacione matrice)

$$[\boldsymbol{\Sigma} - \lambda\mathbf{I}](\mathbf{x}-\boldsymbol{\mu}) = 0. \tag{3.9}$$

Na osnovu jednačine (3.9) konstatujemo da su koordinate prve ose proporcionalne elementima karakterističnih vektora kovarijacione matrice. Kako matrica $\boldsymbol{\Sigma}$ ima p karakterističnih korena treba izabrati jedan od njih. Množenjem jednačine (3.8) sa $4(\mathbf{x}-\boldsymbol{\mu})'$ dobijamo

$$4(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = 4\lambda(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = 4\lambda c^2 \tag{3.10}$$

Za fiksnu vrednost c, dužina prve ose biće maksimalna ako za λ uzmemo najveći karakteristični koren kovarijacione matrice.

Ponovimo ukratko dobijeni rezultat. Pravac prve ose elipsoida određen je normalizovanim karakterističnim vektorom, $\mathbf{e}_1$, koji je pridružen najvećem karakterističnom korenu kovarijacione matrice λ_1. Njena dužina je $2c\sqrt{\lambda_1}$.

Pravac druge ose elipsoida određen je karakterističnim vektorom koji je pridružen drugom po veličini karakterističnom korenu matrice $\mathbf{\Sigma}$. Ukoliko su svi karakteristični koreni kovarijacione matrice različiti i pozitivni, položaj osa elipsoida jednoznačno je određen i ose su ortogonalne međusobom. Ako su dva karakteristična korena međusobom jednaka, tada je presek elipsoida u ravni koja je generisana odgovarajućim karakterističnim vektorima, kružnica. U tom slučaju postoji beskonačno mnogo položaja dve međusobno ortogonalne ose, odnosno njihov položaj nije jednoznačno određen.

Primer 3.2 Za raspored iz Primera 3.1 odrediti ose konture konstantne gustine verovatnoće.

Pretpostavimo da je $\sigma_1^2 = \sigma_2^2$. Da bismo odredili ose elipse potrebno je izračunati karakteristične korene i vektore kovarijacione matrice. Iz $\left|\mathbf{\Sigma} - \lambda\mathbf{I}\right| = 0$, dobijamo

$$0 = \begin{vmatrix} \sigma_1^2 - \lambda & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 - \lambda \end{vmatrix} = (\sigma_1^2 - \lambda)^2 - \sigma_{12} = (\lambda - \sigma_1^2 - \sigma_{12})(\lambda + \sigma_1^2 - \sigma_{12})$$

pa sledi da su karakteristični koreni $\lambda = \sigma_1^2 + \sigma_{12}$ i $\lambda = \sigma_1^2 - \sigma_{12}$, ili ako ih izrazimo preko koeficijenta korelacije: $\lambda_1 = \sigma_1^2(1+\rho)$ i $\lambda_2 = \sigma_1^2(1-\rho)$. Karakteristične vektore određujemo na osnovu izraza $\mathbf{\Sigma e}_j = \lambda_j \mathbf{e}_j$, $j = 1, 2$. Za elemente prvog karakterističnog vektora imamo

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = (\sigma_1^2 + \sigma_{12}) \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

da su jednaki, odnosno $e_1 = e_2$, a uz normirajući uslov (jedinična dužina vektora) dobijamo za karakteristični koren $\lambda = \sigma_1^2 + \sigma_{12}$, da je karakteristični vektor $\mathbf{e}_1' = \left[\frac{1}{\sqrt{2}} \;\; \frac{1}{\sqrt{2}}\right]$. Slično sledi za $\lambda = \sigma_1^2 - \sigma_{12}$ da je $\mathbf{e}_1' = \left[\frac{1}{\sqrt{2}} \;\; -\frac{1}{\sqrt{2}}\right]$. Ose elipse konstantne gustine verovatnoće su

$$2c\sqrt{\sigma_1^2+\sigma_{12}}\begin{bmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{bmatrix} \quad \text{i} \quad 2c\sqrt{\sigma_1^2-\sigma_{12}}\begin{bmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{bmatrix}.$$

Dakle, u slučaju da je $\sigma_1^2=\sigma_2^2$ i $\sigma_{12}>0$ $(\rho>0)$, prvi karakteristični koren je najveći i njemu pridruženi karakteristični vektor ležaće na pravcu koji pod uglom od 45^0 seče I i III kvadrant prolazeći kroz tačku (μ_1,μ_2). Znači da će glavna osa biti pridružena prvom, najvećem karakterističnom korenu. Za negativne vrednosti kovarijanse drugi karakteristični koren, λ_2, biće najveći, pa će glavna osa elipse ležati na pravcu koji pod uglom od 135^0 seče II i IV kvadrant i prolazi kroz tačku (μ_1,μ_2). U oba slučaja sa porastom ρ po apsolutnoj vrednosti, dužina glavne ose se povećava, a sporedne smanjuje. Kada je $\sigma_{12}=0$, odn. $\rho=0$ dužine obe ose su jednake. ■

Postavljaju se dva pitanje: prvo, kolika je verovatnoća da će elipsoid sadržati neku, na slučaj izabranu tačku i drugo pitanje, da li neka tačka pripada elipsoidu uz unapred izabranu verovatnoću. Da bi se odgovorilo na ova pitanja potrebno je odrediti raspored konture gustine verovatnoće. Pokazano[1] je da ako slučajan vektor $\mathbf{X}$ ima normalan raspored $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, da tada $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ ima χ^2 raspored sa p stepeni slobode. Verovatnoća da tačka $\mathbf{X}$ pripada elipsoidu je

$$P\left(\left\{\mathbf{X}:(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\le\chi^2_{p;\alpha}\right\}\right)=1-\alpha \tag{3.11}$$

gde smo sa $\chi^2_{p;\alpha}$ označili gornjih $(100\alpha)\%$ $\chi^2_{p;\alpha}$ rasporeda. Ovaj rezultat je našao primenu kod provere normalnosti rasporeda kao i kod otkrivanja nestandardnih opservacija.

Primer 3.3 Na osnovu podataka iz Primera 2.11 odrediti proporciju opservacija koje pripadaju 50% konturi konstantne gustine verovatnoće pretpostavljenog četvorodimenzionog normalnog rasporeda.

U Primeru 2.11 određen je kvadrat Mahalanobisovog odstojanja, odnosno $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. Pošto je $\chi^2_{4;0.5}=3.36$, za svaku tačku $\mathbf{x}_i'=[x_{i1},x_{i2},x_{i3},x_{i4}]$, $i=1,2,...,6$, proveravamo da li je zadovoljena nejednakost

[1] Videti: Johnson i Wichern (1982, ss. 138-0) ili Mardia, Kent i Bibby (1979, s. 39)

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq 3.36$$

Kako je nejednakost zadovoljena za opservacije 3., 4. i 6. domaćinstva, a $n=6$, konstatujemo da polovina opservacija pripada 50% konturi konstantne gustine verovatnoće. ■

3.2 MARGINALNI I USLOVNI RASPOREDI VIŠEDIMENZIONE NORMALNE PROMENLJIVE

Ukažimo na nekoliko rezultata vezanih za osobine višedimenzionog normalnog rasporeda koji opravdavaju njegovu popularnost i široku upotrebu. Poslužićemo se ranije definisanom linearnom kombinacijom slučajnih promenljivih iz Odeljka 2.1.4. Na osnovu očekivane vrednosti i kovarijanse linearne kombinacije odredićemo pokazatelje marginalnog rasporeda normalno raspoređene slučajne promenljive, a zatim uspostaviti vezu između parametara uslovnog rasporeda i modela regresione analize.

Posmatrajmo slučajan vektor $\mathbf{X}'=[X_1, X_2, ..., X_p]$ koji ima raspored $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Ako definišemo q linearnih kombinacija p slučajnih promenljivih $(q \leq p)$ i napišemo ih kao $\mathbf{Y}=\mathbf{AX}$, gde je $\mathbf{Y}(q\times 1)$ vektor i $\mathbf{A}(q\times p)$ matrica realnih brojeva, tada slučajan vektor $\mathbf{Y}=\mathbf{AX}$ ima raspored $N_p(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

Pretpostavimo da je $q=1$, što znači da smo formirali jednu linearnu kombinaciju. Tada slučajan vektor $\mathbf{Y}$ postaje jednodimenziona normalno raspoređena slučajna promenljiva. Ako su koeficijenti linearne kombinacije svi jednaki nuli osim $j-$tog, koji je jednak jedinici, tada se linearna kombinacija izjednačava sa X_j, $j=1,2,...,p$, pa sledi da svaki element slučajnog vektora $\mathbf{X}$ ima marginalan raspored koji je jednak $N(\mu_j, \sigma_{jj})$. Štaviše, podskup od q $(q>1)$ elemenata slučajnog vektora $\mathbf{X}$ ima takođe normalan raspored. U skladu sa podelom slučajnog vektora $\mathbf{X}$, sredine $\boldsymbol{\mu}$ i kovarijacione matrice $\boldsymbol{\Sigma}$ (Odeljak 2.1.4) parametri tog normalnog rasporeda su $N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$. Takođe, poslednjih $s=p-q$ elemenata slučajnog vektora ima $N_s(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ raspored.

Elementi submatrice $\boldsymbol{\Sigma}_{12}$ ukazuju na stepen i oblik zavisnosti između dva skupa slučajnih promenljivih $\mathbf{X}_1$ i $\mathbf{X}_2$ Ako je ona jednaka nula matrici, tada su slučajne promenljive $\mathbf{X}_1$ i $\mathbf{X}_2$ međusobom nezavisno raspoređene.

Na osnovu ranije definisane uslovne funkcije gustine verovatnoće možemo odrediti uslovni raspored od $\mathbf{X}_2$ pod uslovom da je $\mathbf{X}_1$ fiksno, odnosno da je $\mathbf{X}_1=\mathbf{x}_1$. Pokazano[2] je da se radi o normalnom rasporedu sa sredinom:

[2] Videti: Anderson (1972, s. 28) ili Morrison (1976, ss. 91-2).

$E(\mathbf{X}_2 \mid \mathbf{x}_1) = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)$ i kovarijacionom matricom: $Cov(\mathbf{X}_2 \mid \mathbf{x}_1) = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. Znači da je sredina slučajne promenljive $\mathbf{X}_2$ pod uslovom da je $\mathbf{X}_1$ fiksno (uslovna sredina), linearna funkcija fiksne vrednosti $\mathbf{x}_1$, dok njena kovarijaciona matrica ne zavisi od fiksne vrednosti $\mathbf{x}_1$.

Primer 3.4 Za dvodimenzionu normalno raspoređenu slučajnu promenljivu odrediti funkciju uslovne gustine verovatnoće.

Kako je $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} = \sigma_{12}$, $\boldsymbol{\Sigma}_{11} = \sigma_1^2$ i $\boldsymbol{\Sigma}_{22} = \sigma_2^2$ to se uslovna sredina od X_2 svodi na skalar

$$E(X_2 \mid x_1) = \mu_2 + \frac{\sigma_{12}}{\sigma_1^2}(x_1 - \mu_1),$$

kao i njena kovarijaciona matrica, pa je uslovna varijansa

$$Var(X_2 \mid x_1) = \sigma_2^2 - \sigma_{12}\left(\sigma_1^2\right)^{-1}\sigma_{12} = \sigma_2^2 - \frac{\sigma_{12}}{\sigma_1^2} = \sigma_2^2(1-\rho).$$

Znači da je raspored od X_2 pod uslovom $X_1 = x_1$ normalan sa sredinom $E(X_2 \mid x_1)$ i varijansom $Var(X_2 \mid x_1)$.

Funkcija uslovne gustine verovatnoće je

$$f(x_2 \mid x_1) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_2^2(1-\rho^2)}} e^{-\frac{1}{2\sigma_2^2(1-\rho^2)}\left[x_2 - \mu_2 - \frac{\sigma_{12}}{\sigma_1^2}(x_1-\mu_1)\right]^2}, -\infty < x_2 < \infty.$$

■

U statistici je uobičajeno da se uslovna sredina $E(\mathbf{X}_2 \mid \mathbf{x}_1)$, naziva *multivarijaciona regresija* $\mathbf{X}_2$ na $\mathbf{X}_1$. Kada je $q = s = 1$, pa je $p = 2$, regresiona funkcija se naziva *jednostavna*, a za $s = 1$ i $q = p - 1$ dobijamo funkciju dobro poznate *višestruke regresije*. Za $q \geq 2$ i $s \geq 2$ dobijamo niz od s funkcija višestruke regresije koje zajedno nazivamo multivarijaciona regresiona funkcija i kod koje smo regresirali X_{q+1} na $X_1, X_2, ..., X_q$; X_{q+2} na $X_1, X_2, ..., X_q$, ..., X_p na $X_1, X_2, ..., X_q$ pri čemu svaka od regresija predstavlja linearnu funkciju navedenih promenljivih, odnosno regresora. Regresioni koeficijenti multivarijacione regresije $\mathbf{X}_2$ na $\mathbf{X}_1$ dati su $(s \times q)$ matricom $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$. Elementi u $i-$tom redu te matrice predstavljaju regresione koeficijente u višestrukoj regresiji svake od s zavisnih promenljivih $X_i, i = q+1, q+2, ..., p$ na

istovetan skup nezavisnih promenljivih $X_1, X_2, ..., X_q$. Ako $E(\mathbf{X}_2 \mid \mathbf{x}_1)$ napišemo u obliku

$$E(\mathbf{X}_2 \mid \mathbf{x}_1) = (\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1) + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1 = \boldsymbol{\beta}_0 + \boldsymbol{\beta}\mathbf{x}_1 \tag{3.12}$$

tada je $\boldsymbol{\beta}_0 = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1$ $(s \times 1)$ vektor slobodnih članova, a $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$ $(s \times q)$ matrica regresionih koeficijenata. Dakle, na osnovu strukture modela multivarijacione regresije konstatujemo da se on sastoji od s višestrukih regresionih funkcija od kojih svaka ima identičnu listu regresora $X_1, X_2, ..., X_q$.

Primer 3.5 Ilustrovati korišćenje modela multivarijacione regresije na osnovu podataka iz Primera 2.5. Pretpostaviti da se radi o populaciji kod koje je slučajan vektor $\mathbf{X} \sim N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i gde su $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ dati u Primeru 2.5. Slučajan vektor $\mathbf{X}$ podelili smo na sledeći način: $\mathbf{X}_1 = [X_1, X_2]$ i $\mathbf{X}_2 = [X_3, X_4]$. Cilj nam je da utvrdimo međusobnu zavisnost izdataka na ishranu (X_3) i izdataka na obrazovanje i razonodu (X_4) od broja članova domaćinstva (X_1) i godišnje raspoloživih sredstava (X_2). Drugim rečima model multivarijacione regresije glasi:

$$X_3 = \beta_{01} + \beta_{11}x_1 + \beta_{12}x_2$$
$$X_4 = \beta_{02} + \beta_{21}x_1 + \beta_{22}x_2 .$$

Elementi sredine podeljenog slučajnog vektora $\mathbf{X}$ i njegove kovarijacione matrice su:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 4.5 \\ 11.0 \end{bmatrix},\ \boldsymbol{\mu}_2 = \begin{bmatrix} 4.3 \\ 0.2 \end{bmatrix},\ \boldsymbol{\Sigma}_{11} = \begin{bmatrix} 3.500 & -0.860 \\ -0.860 & 2.416 \end{bmatrix},\ \boldsymbol{\Sigma}_{22} = \begin{bmatrix} 5.1520 & -0.15680 \\ -0.1568 & 0.00608 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{12} = \begin{bmatrix} 4.1000 & -0.1200 \\ -1.5100 & 0.0524 \end{bmatrix},\ \boldsymbol{\Sigma}_{21} = \begin{bmatrix} 4.1000 & -1.5100 \\ -0.1200 & 0.0524 \end{bmatrix}$$

Kako je $\boldsymbol{\Sigma}_{11}^{-1} = \begin{bmatrix} 0.3131 & 0.1115 \\ 0.1115 & 0.4535 \end{bmatrix}$, to je matrica regresionih koeficijenata:

$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} = \begin{bmatrix} 1.1154 & -0.2280 \\ -0.0317 & 0.0104 \end{bmatrix}$, a vektor slobodnih članova:

$$\boldsymbol{\beta}_0 = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1 = \begin{bmatrix} 4.3 \\ 0.2 \end{bmatrix} - \begin{bmatrix} 1.1154 & -0.2280 \\ -0.0317 & 0.0104 \end{bmatrix}\begin{bmatrix} 4.5 \\ 11.0 \end{bmatrix} = \begin{bmatrix} 1.7881 \\ 0.2285 \end{bmatrix}.$$

Znači da je model multivarijacione regresije:

$$E(\mathbf{X}_2 \mid \mathbf{x}_1) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}\mathbf{x}_1 = \begin{bmatrix} 1.7881 \\ 0.2285 \end{bmatrix} + \begin{bmatrix} 1.1154 & -0.2280 \\ -0.0317 & 0.0104 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \begin{bmatrix} 1.7881 + 1.1154x_1 - 0.2280x_2 \\ 0.2285 - 0.0317x_1 + 0.0104x_2 \end{bmatrix}. \qquad \blacksquare$$

Rečeno je da kovarijaciona matrica uslovnog rasporeda, $Cov(\mathbf{X}_2 \mid \mathbf{x}_1)$, ne zavisi od $\mathbf{x}_1$. Prema tome ni korelaciona matrica, na njoj zasnovana, neće zavisiti od fiksne vrednosti $\mathbf{x}_1$. Ovu matricu nazivamo *matrica parcijalnih koeficijenata korelacije* između promenljivih koje pripadaju podskupu $\mathbf{X}_2$ pod uslovom da je $\mathbf{X}_1 = \mathbf{x}_1$ fiksno. Prema poznatoj relaciji između kovarijacione i korelacione matrice (2.24), parcijalna korelaciona matrica je data izrazom

$$\mathbf{D}_{2\bullet 1}^{-1/2}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})\mathbf{D}_{2\bullet 1}^{-1/2} \tag{3.13}$$

gde smo sa $\mathbf{D}_{2\bullet 1}$ označili dijagonalnu matricu koja sadrži dijagonalne elemente matrice $(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$. Ako elemente kovarijacione matrice uslovnog rasporeda označimo sa $\sigma_{jk\bullet 1,2,\ldots,q}$, gde je $j,k = 1,2,\ldots,s$, tada je element parcijalne korelacione matrice dat izrazom

$$\rho_{jk\bullet 1,2,\ldots,q} = \frac{\sigma_{jk\bullet 1,2,\ldots,q}}{\sqrt{\sigma_{jj\bullet 1,2,\ldots,q}}\sqrt{\sigma_{kk\bullet 1,2,\ldots,q}}}, \; j,k = 1,2,\ldots,s \tag{3.14}$$

i nazivamo ga parcijalni koeficijent korelacije $q-$tog reda. Na osnovu rekurzivne relacije koja je uspostavljena između parcijalnih koeficijenata korelacije višeg i nižeg reda

$$\rho_{jk\bullet 1,2,\ldots,q} = \frac{\rho_{jk\bullet 1,2,\ldots,q-1} - \rho_{jq\bullet 1,2,\ldots,q-1}\,\rho_{kq\bullet 1,2,\ldots,q-1}}{\sqrt{1-\rho^2_{jq\bullet 1,2,\ldots,q-1}}\sqrt{1-\rho^2_{kq\bullet 1,2,\ldots,q-1}}} \tag{3.15}$$

parcijalne koeficijente korelacije možemo odrediti polazeći od matrice običnih koeficijenata korelacije, s obzirom da je za $q = 1$

$$\rho_{jk\bullet 1} = \frac{\rho_{jk} - \rho_{j1}\,\rho_{k1}}{\sqrt{1-\rho^2_{j1}}\sqrt{1-\rho^2_{k1}}} \tag{3.16}$$

Primer 3.6 Na osnovu podataka iz Primera 3.5 odrediti uslovnu kovarijacionu matricu , kao i odgovarajuću parcijalnu korelacionu matricu.

Kako je na osnovu rezultata iz Primera 3.5

$$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}=\begin{bmatrix}1.1154 & -0.2280\\ -0.0317 & 0.0104\end{bmatrix} \text{ to je } \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}=\begin{bmatrix}4.9174 & -0.1458\\ -0.1458 & 0.0044\end{bmatrix}, \text{ a}$$

$$Cov(\mathbf{X}_2\mid\mathbf{x}_1)=\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}=\begin{bmatrix}0.2346 & -0.0110\\ -0.0110 & 0.0017\end{bmatrix} \text{ i}$$

$$\mathbf{D}_{2\bullet 1}=\begin{bmatrix}0.2346 & 0\\ 0 & 0.0017\end{bmatrix}.$$

Parcijalna korelaciona matrica je $\mathbf{D}_{2\bullet 1}^{-1/2}(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})\mathbf{D}_{2\bullet 1}^{-1/2}=$

$$\begin{bmatrix}2.0647 & 0\\ 0 & 24.0594\end{bmatrix}\begin{bmatrix}0.2346 & -0.0110\\ -0.0110 & 0.0017\end{bmatrix}\begin{bmatrix}2.0647 & 0\\ 0 & 24.0594\end{bmatrix}=\begin{bmatrix}1 & -0.5467\\ -0.5467 & 1\end{bmatrix}.$$

Elemenat ove matrice $\rho_{34\bullet 1,2}=-0.5467$ poredimo sa odgovarajućim običnim koeficijentom korelacije $\rho_{34}=-0.8859$. Konstatujemo da je eliminisanjem linearnog uticaja broja članova domaćinstva i godišnjeg iznosa raspoloživih sredstava došlo do smanjenja međuzavisnosti izdataka na ishranu i izdataka na obrazovanje i razonodu.

Do ovog koeficijenta mogli smo doći na osnovu rekurzivne relacije

$$\rho_{34\bullet 1,2}=\frac{\rho_{34\bullet 1}-\rho_{23\bullet 1}\rho_{24\bullet 1}}{\sqrt{1-\rho_{23\bullet 1}^2}\sqrt{1-\rho_{24\bullet 1}^2}} \text{ gde je}$$

$$\rho_{34\bullet 1}=\frac{\rho_{34}-\rho_{13}\rho_{14}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{14}^2}}=\frac{-0.8859-0.9655(-0.8226)}{\sqrt{1-0.9655^2}\sqrt{1-(-0.8226)^2}}=-0.61916,$$

$$\rho_{23\bullet 1}=\frac{\rho_{23}-\rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}}=-0.57285,\ \rho_{24\bullet 1}=\frac{\rho_{24}-\rho_{12}\rho_{14}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{14}^2}}=0.34805$$

pa je $\rho_{34\bullet 1,2}=\dfrac{-0.61916-(-0.57285)\,0.34805}{\sqrt{1-(-0.57285)^2}\sqrt{1-0.34805^2}}=-0.5467$. ■

Pozabavimo se specijalnim slučajem uslovnog rasporeda, kada je $s=1$ i $q=p-1$. Tada je reč o uslovnom rasporedu od X_2 za fiksne vrednosti $\mathbf{X}_1=\mathbf{x}_1$. Kovarijaciona matrica uslovnog rasporeda svodi se na varijansu (skalar)

$$Var(X_2 \mid \mathbf{x}_1) = \sigma_{22} - \boldsymbol{\sigma}_2' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\sigma}_2 \tag{3.17}$$

gde smo sa $\boldsymbol{\sigma}_2'$ označili $[(p-1)\times 1]$ vektor kovarijansi između X_2 i svih promenljivih iz vektora $\mathbf{X}_1$. Ova uslovna varijansa meri stepen raspršenosti podataka oko regresione linije. Pokazano[3] je da od svih linearnih kombinacija (gde je $\mathbf{a} = [a_1, a_2, ..., a_q]$ vektor koeficijenata linearne kombinacije), kombinacija koja minimizira varijansu od $X_2 - \mathbf{aX}_1$, odnosno $Var(X_2 \mid \mathbf{x}_1)$ i istovremeno maksimizira korelaciju između X_2 i $\mathbf{aX}_1$ je linearna kombinacija $\boldsymbol{\beta}\mathbf{X}_1$, gde je $\boldsymbol{\beta}$ $(1\times q)$ vektor regresionih koeficijenata ($\boldsymbol{\beta} = \boldsymbol{\sigma}_2' \boldsymbol{\Sigma}_{11}^{-1}$). Prema ovom rezultatu definišemo *koeficijent višestruke korelacije*, u oznaci $R_{X_2 \cdot \mathbf{x}_1}$, kao maksimalnu korelaciju između X_2 i linearne kombinacije $\mathbf{aX}_1$. Prema definiciji sledi da je

$$R_{X_2 \cdot \mathbf{x}_1} = \frac{Cov(\boldsymbol{\beta}\mathbf{X}_1, X_2)}{\sqrt{Var(X_2)}\sqrt{Var(\boldsymbol{\beta}\mathbf{X}_1)}} = \frac{\boldsymbol{\sigma}_2' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\sigma}_2}{\sqrt{\sigma_{22}}\sqrt{\boldsymbol{\sigma}_2' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\sigma}_2}} = \frac{\sqrt{\boldsymbol{\sigma}_2' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\sigma}_2}}{\sqrt{\sigma_{22}}} \tag{3.18}$$

Kvadrat koeficijenta višestruke korelacije definiše koeficijent determinacije, u oznaci R^2 gde je

$$R^2 = \frac{\boldsymbol{\sigma}_2' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\sigma}_2}{\sigma_{22}} \tag{3.19}$$

Ako izraz u brojiocu razlomka proširimo tako što dodamo i oduzmemo σ_{22}, dobićemo

$$R^2 = \frac{\sigma_{22} - (\sigma_{22} - \boldsymbol{\sigma}_2' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\sigma}_2)}{\sigma_{22}} = \frac{Var(X_2) - Var(X_2 \mid \mathbf{x}_1)}{Var(X_2)} \tag{3.20}$$

Razlika $Var(X_2)$ i $Var(X_2 \mid \mathbf{x}_1)$ predstavlja deo varijanse zavisne promenljive X_2 koja je objašnjena regresijom. Stoga je koeficijent determinacije proporcija varijanse zavisne promenljive koja je objašnjena regresionom funkcijom.

Primer 3.7 Odrediti koeficijent determinacije u modelu multivarijacione regresije za prvu i drugu regresionu funkciju iz Primera 3.5.

Za prvu regresionu jednačinu imamo: $\mathbf{X}_1 = [X_1, X_2]$ i $\mathbf{X}_2 = [X_3]$, pa je

$$\sigma_{22} = 5.152\,,\ \boldsymbol{\sigma}_2' = [4.1 \quad -1.51] \text{ i } \boldsymbol{\Sigma}_{11}^{-1} = \begin{bmatrix} 0.3131 & 0.1115 \\ 0.1115 & 0.4536 \end{bmatrix}.$$

[3] Videti: Anderson (1972, ss. 31-32).

Njen koeficijent determinacije je $R^2 = \frac{\boldsymbol{\sigma}_2' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\sigma}_2}{\sigma_{22}} = \frac{4.9174}{5.1520} = 0.9545$.

Za drugu regresionu jednačinu imamo: $\mathbf{X}_1 = [X_1, X_2]$ i $\mathbf{X}_2 = [X_4]$, pa je $\sigma_{22} = 0.00608$ i $\boldsymbol{\sigma}_2' = [-0.12 \quad 0.0524]$. Njen koeficijent determinacije je $R^2 = 0.7159$.

Znači da je 95.45% varijabiliteta izdataka na ishranu i 71.59% izdataka na obrazovanje i razonodu objašnjeno brojem članova domaćinstva i godišnje raspoloživim sredstvima. ■

3.3 UZORAK IZ VIŠEDIMENZIONOG NORMALNOG RASPOREDA

Izlaganje višedimenzionog normalnog rasporeda zasnivali smo na pretpostavci da su nam njegovi parametri poznati. U stvarnosti polazimo od višedimenzionog normalnog rasporeda kao statističkog modela kojim opisujemo posmatrani višedimenzioni fenomen, a njegove je parametre potrebno oceniti na osnovu uzorka.

Pretpostavimo da je $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ slučajan uzorak od n elemenata slučajne promenljive $\mathbf{X}$ sa $p-$dimenzionim normalnim rasporedom čiji je sredina $\boldsymbol{\mu}$ i kovarijaciona matrica $\boldsymbol{\Sigma}$ nepoznati. To znači da su i elementi slučajnog uzorka identično raspoređeni, $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i uz to međusobom nezavisni. Na osnovu definicije slučajnog uzorka njegov združeni raspored jednak je proizvodu marginalnih rasporeda. Ova funkcija se za fiksan uzorak naziva *funkcija verodostojnosti uzorka* i predstavlja se kao funkcija nepoznatih parametara, u oznaci $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, gde je

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \tag{3.21}$$

Maksimiziranjem funkcije verodostojnosti s obzirom na $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ dobijamo ocenu najveće verodostojnosti nepoznatih parametara. Tako[4] je ocena najveće verodostojnosti sredina vektora

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{3.22}$$

a ocena najveće verodostojnosti kovarijacione matrice $\boldsymbol{\Sigma}$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' = \frac{n-1}{n} \mathbf{S} . \tag{3.23}$$

[4] Videti: Anderson (1972, ss. 44-49).

Na osnovu ranije iznetog opšteg rezultata o rasporedu linearne kombinacije višedimenzione normalno raspoređene slučajne promenljive, može se pokazati da je sredina slučajnog uzorka od n elemenata uzetog iz višedimenzionog normalnog rasporeda $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ takođe normalno raspoređena. Parametri rasporeda sredine slučajnog uzorka su $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}/n$, odnosno imamo da je $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$. Ovaj rezultat predstavlja generalizaciju poznatog rezultata koji se tiče rasporeda sredine slučajnog uzorka $(\bar{X})$ jedne slučajne promenljive, prema kome je $\bar{X} \sim N_1(\mu, \sigma^2/n)$ i gde su μ i σ^2 populaciona sredina i varijansa respektivno.

Takođe smo imali kod jedne slučajne promenljive da je raspored $(n-1)S^2 \sim \sigma^2 \chi^2_{n-1}$, gde je S^2 uzoračka varijansa. Iz teorije rasporeda poznato je da suma kvadrata nezavisnih, standardizovanih normalno raspoređenih slučajnih promenljivih ima χ^2 raspored. Po analogiji u višedimenzionom slučaju imamo na osnovu $(n \times p)$ matrice podataka $\mathbf{X}$, koja je dobijena kao slučajan uzorak iz $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, da slučajna matrica $\mathbf{X}'\mathbf{X}$ ima $p-$dimenzioni *Wishartov raspored*, u oznaci $W_p(\boldsymbol{\Sigma}, n)$, gde n predstavlja broj stepeni slobode. U tom smislu Wishartov raspored predstavlja generalizaciju χ^2-rasporeda. Za $p=1$ i $\boldsymbol{\Sigma}=1$ Wishartov raspored svodi se na χ^2-raspored. Kako na osnovu uzoračke kovarijacione matrice $\mathbf{S}$ statistika $(n-1)\mathbf{S}$ ima Wishartov raspored sa $n-1$ stepen slobode, što se označava sa $(n-1)\mathbf{S} \sim W_p[\boldsymbol{\Sigma}, (n-1)]$, to se ovaj raspored koristi u statističkom zaključivanju o populacionoj kovarijacionoj matrici $\boldsymbol{\Sigma}$. Navedimo[5] dve osobine Wishartovog rasporeda. Prva, osobina aditivnosti, prema kojoj zbir dve matrice, na primer $\mathbf{A}_1$ i $\mathbf{A}_2$, nezavisne međusobom i raspoređene po Wishartovom rasporedu, $\mathbf{A}_i \sim W_p(\boldsymbol{\Sigma}, n_i)$, $i=1,2$, ima Wishartov raspored: $\mathbf{A}_1 + \mathbf{A}_2 \sim W_p(\boldsymbol{\Sigma}, n_1 + n_2)$. Druga osobina, prema kojoj na osnovu slučajne matrice $\mathbf{A}$ koja je raspoređena po Wishartovom rasporedu, $\mathbf{A} \sim W_p(\boldsymbol{\Sigma}, n)$, sledi da je matrica $\mathbf{BAB}'$ raspoređena po Wishartovom rasporedu, u oznaci $\mathbf{BAB}' \sim W_p(\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}', n)$, gde je $\mathbf{B}(p \times p)$ nesingularna matrica. Naposletku, kao i u jednodimenzionom slučaju, uzoračka sredina $\bar{\mathbf{X}}$ i uzoračka kovarijaciona matrica $\mathbf{S}$ međusobom su nezavisni.

Poslednji raspored koji definišemo na osnovu višedimenzionog normalnog rasporeda predstavlja generalizaciju Studentovog $t-$rasporeda. Ako je raspored jednodimenzione slučajne promenljive $X \sim N(\mu, \sigma^2)$ odakle je uzet uzorak od n elemenata, tada je

$$t^2 = n\frac{(\bar{X}-\mu)^2}{S^2} = n(\bar{X}-\mu)(S^2)^{-1}(\bar{X}-\mu). \tag{3.24}$$

kvadrat Studentove $t-$statistike sa $n-1$ stepen slobode. Po analogiji u višedimenzionom slučaju imaćemo

[5] Za dokaz videti kod Andersona (1972, ss. 161-2).

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \tag{3.25}$$

gde T^2 nazivamo *Hotellingova* T^2 *statistika*. Njen raspored je $[(n-1)p/(n-p)]F_{p,n-p}$, gde smo sa $F_{p,n-p}$ označili raspored sa p i $n-p$ stepeni slobode. Hotellingova T^2 statistika našla je svoju primenu u postupku statističkog zaključivanja o populacionoj sredini $\boldsymbol{\mu}$. Ukažimo[6] na osobinu invarijantnosti T^2 – statistike. Naime, ako smo opservacije linearno transformisali tako što umesto $\mathbf{x}$ posmatramo $\mathbf{Ax} + \mathbf{b}$, gde je $\mathbf{A}(p \times p)$ nesingularna matrica, a $\mathbf{b}(p \times 1)$ vektor čiji su elementi realni brojevi, tada je vrednost T^2 – statistike nepromenjena.

Dosadašnje izlaganje rasporeda uzoračkih pokazatelja bazirano je na pretpostavci da slučajan uzorak potiče iz višedimenzionog normalnog rasporeda. Međutim, na osnovu generalizacije centralne granične teoreme za slučajan uzorak iz ma koje populacije sa sredinom $\boldsymbol{\mu}$ i konačnom kovarijacionom matricom $\boldsymbol{\Sigma}$ pokazano[7] je da $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ ima aproksimativno $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ raspored kada $n \to \infty$. Primena multivarijacione centralne granične teoreme na uzoračku kovarijacionu matricu $\mathbf{S}$ sugeriše da elementi matrice $\sqrt{n}(\mathbf{S} - \boldsymbol{\Sigma})$ konvergiraju u rasporedu ka $N_p(\mathbf{0}, \mathbf{G})$ rasporedu, gde je opšti član kovarijacione matrice $\mathbf{G}$ dat izrazom $(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})$. Na osnovu ovoga možemo za velike uzorke zameniti $\boldsymbol{\Sigma}$ sa $\mathbf{S}$ u rasporedu za $\bar{\mathbf{X}}$, pa da ne bude narušena aproksimativna normalnost rasporeda sredine uzorka. Takođe, raspored koji je naveden za konturu gustine verovatnoće važi asimptotski ako je $\bar{\mathbf{X}}$ aproksimativno normalno raspoređeno, a populaciona kovarijaciona matrica $\boldsymbol{\Sigma}$ zamenjena uzoračkom. Naime, kažemo da $n(\bar{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu})$ ima aproksimativno χ^2 – raspored sa p stepeni slobode. Ova statistika koristi se između ostalog za određivanje oblasti poverenja populacione sredine $\boldsymbol{\mu}$.

3.4 ZAKLJUČIVANJE O SREDINI

Na osnovu slučajnog uzorka uzetog iz normalno raspoređene populacije definisali smo brojnu ocenu nepoznatog parametra, na primer sredinu μ. Međutim, često se postavlja zahtev ocenjivanja prave vrednosti parametra, ne brojem, nego intervalnom vrednošću. U tom smislu definiše se intervalna ocena sredine μ. Razlikovali smo dva slučaja. Prvi, kada nam je varijansa populacije poznata i drugi, kada nam je varijansa populacije nepoznata. Detaljnije o intervalnom ocenjivanju u jednodimenzionom slučaju videti kod Miloševića (1983, ss. 130-147). Zadatak u ovom poglavlju jeste generalizacija upravo navedenih elemenata statističkog zaključivanja o sredini jedne promenljive na višedimenzioni slučaj.

---

[6] Za dokaz videti kod Morrisona (176, s. 133).

[7] Videti: Anderson (1972, ss. 74-6).

3.4.1 Oblast poverenja sredine

Za jednu slučajnu promenljivu sa poznatom varijansom populacije imamo $100(1-\alpha)\%$ interval poverenja sredine:

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}; \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right) \tag{3.26}$$

gde je $\bar{X}$ sredina slučajnog uzorka od n elemenata, σ je standardna devijacija populacije, a predstavlja gornji $100(\alpha/2)$ percentil standardizovanog normalnog rasporeda.

U slučaju da nam varijansa populacije nije poznata $100(1-\alpha)\%$ interval poverenja sredine slučajne promenljive je

$$\left(\bar{X} - \frac{S}{\sqrt{n}} t_{n-1;\alpha/2}; \bar{X} + \frac{S}{\sqrt{n}} t_{n-1;\alpha/2} \right) \tag{3.27}$$

gde je S uzoračka ocena standardne devijacije populacije, a $t_{n-1;\alpha/2}$ označava gornji $100(\alpha/2)$ percentile $t-$rasporeda sa $n-1$ stepen slobode. Interval poverenja za jednu slučajnu promenljivu nazivamo $t-$intervalom. Pre nego što smo uzeli uzorak interval poverenja predstavlja slučajan interval, jer mu granice zavise od slučajnih promenljivih $\bar{X}$ i S. Verovatnoća da taj interval sadrži pravu vrednost sredine iznosi $1-\alpha$, što znači da od velikog broja takvih nezavisnih intervala njih $100(1-\alpha)\%$ sadržaće pravu vrednost sredine.

U višedimenzionom slučaju definišemo slučajnu oblast, *oblast poverenja*, tako da sa unapred izabranom verovatnoćom $1-\alpha$, oblast obuhvata tačku $\boldsymbol{\mu}' = [\mu_1, \mu_2, ..., \mu_p]$. Na osnovu rezultata (3.11) koji se odnosi na raspored konture gustine verovatnoće konstatujemo da $n(\bar{\mathbf{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu})$ ima χ^2-raspored sa p stepena slobode. Ovaj rezultat koristimo kada nam je poznata kovarijaciona matrica da definišemo oblast poverenja sredine populacije.

Ukoliko nam je kovarijaciona matrica $\boldsymbol{\Sigma}$ nepoznata, koristimo njenu uzoračku ocenu $\mathbf{S}$. U ovom slučaju pri definisanju oblasti poverenja sredine polazimo od rezultata da

$$n(\bar{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \tag{3.28}$$

ima Hotellingov T^2-raspored. Kako je T^2 raspoređeno kao $\frac{p(n-1)}{n-p} F_{p,n-p}$, odakle imamo da je $\frac{(n-p)T^2}{p(n-1)} \sim F_{p,n-p}$, to je $100(1-\alpha)\%$ oblast poverenja za $\boldsymbol{\mu}$ data elipsoidom i njegovom unutrašnjošću. Pri tome je elipsoid definisan izrazom

$$\frac{n-p}{p(n-1)} n(\bar{\mathbf{X}}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}) = F_{p,n-p;\alpha} \tag{3.29}$$

gde je $F_{p,n-p;\alpha}$ gornji 100α percentil $F-$rasporeda sa p i $n-p$ stepeni slobode. Za $p>3$ nismo u mogućnosti grafički prikazati oblast poverenja sredine, ali možemo proveriti da li joj neka tačka pripada ili ne. Ako je kvadrat odstojanja te tačke veći od vrednosti $\frac{p(n-1)}{n-p}F_{p,n-p;\alpha}$ tada tačka ne pripada oblasti poverenja. Takođe smo na osnovu veze ose elipsoida i karakterističnih korena i karakterističnih vektora uzoračke kovarijacione matrice u mogućnosti odrediti pravac i dužinu tih osa. Ose elipsoida leže na pravcima koje definišu karakteristični vektori, a dužina im je jednaka

$$2\sqrt{\lambda_j}\sqrt{\frac{p(n-1)}{n-p}F_{p,n-p;\alpha}} \quad j=1,2,...,p \tag{3.30}$$

gde su λ_j karakteristični koreni uzoračke kovarijacione matrice. Za par osa elipsoida možemo, stavljajući u međusobni odnos njihove dužine, odrediti relativnu razvučenost projekcije elipsoida u ravni definisanoj izabranim parom promenljivih (videti sliku u okviru Primera 3.8).

Primer 3.8 Na osnovu podataka iz Primera 2.5 odrediti 90% oblast poverenja za sredinu dvodimenzione promenljive koju sačinjaju: $X_1=$ broj članova domaćinstva i $X_3=$ izdaci na ishranu.

Izračunato je $\bar{\mathbf{x}}=\begin{bmatrix}4.5\\4.3\end{bmatrix}$, $\mathbf{S}=\begin{bmatrix}3.500 & 4.100\\4.100 & 5.152\end{bmatrix}$, $\mathbf{S}^{-1}=\begin{bmatrix}4.2160 & -3.3552\\-3.3552 & 2.8642\end{bmatrix}$, a za $n=6$ i $p=2$, kritična vrednost je $F_{2,4;0.10}=4.32$, pa oblast poverenja sadrži sve tačke koje zadovoljavaju nejednakost

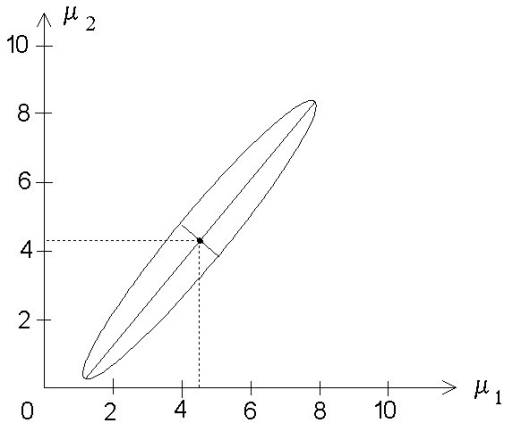
$$\frac{6-2}{2(6-1)}6\begin{bmatrix}4.5-\mu_1 & 4.3-\mu_2\end{bmatrix}\begin{bmatrix}4.2160 & -3.3552\\-3.3552 & 2.8642\end{bmatrix}\begin{bmatrix}4.5-\mu_1\\4.3-\mu_2\end{bmatrix} \le 4.32$$

odakle nakon sređivanja imamo

$$10.1185(\mu_1-4.5)^2-16.1047(\mu_1-4.5)(\mu_2-4.3)+6.8740(\mu_2-4.3)^2 \le 4.32\,.$$

Da bismo nacrtali oblast poverenja (elipsa i njena unutrašnjost) odredimo pravac osa elipse i njihovu dužinu. Karakteristični koreni uzoračke kovarijacione matrice su: $\lambda_1=8.5084$ i $\lambda_2=0.1436$, a njima pridruženi karakteristični vektori su: $\mathbf{e}_1'=\begin{bmatrix}0.6334 & 0.7738\end{bmatrix}$ i $\mathbf{e}_2'=\begin{bmatrix}0.7738 & -0.6334\end{bmatrix}$. Dužina glavne ose je 10.501, a dužina sporedne

ose elipse je 1.364, što znači da je glavna osa 7.697 puta duža od sporedne. Na slici μ_1 predstavlja sredinu promenljive "broj članova domaćinstva", a μ_2 predstavlja sredinu promenljive " izdaci na ishranu".



■

3.4.2 Testiranje hipoteze o vrednosti sredine

Pri testiranju hipoteze o vrednosti sredine normalno raspoređene slučajne promenljive, razlikujemo dva slučaja. Prvi, kada nam je poznata varijansa populacije i drugi, relevantniji slučaj sa praktičnog stanovišta, kada nam je varijansa nepoznata. Pažnju ćemo posvetiti samo drugom slučaju. Kada nam je nepoznata varijansa populacije za test hipoteze

$$H_0 : \mu = \mu_0 \text{, protiv hipoteze } H_1 : \mu \neq \mu_0$$

koristimo statistiku $t-$testa zasnovanu na Studentovom rasporedu. Kada je tačna nulta hipoteza ova statistika raspoređena je po $t-$rasporedu sa $n-1$ stepenom slobode. Na osnovu uzetog uzorka prihvatamo nultu hipotezu ako je apsolutna vrednost statistike testa manja od kritične vrednosti $t_{n-1;\alpha/2}$, gde je α nivo značajnosti testa, a u suprotnom je odbacujemo. Alternativan način provere gornje hipoteze zasnovan je na izračunatom intervalu poverenja sredine. Naime, ukoliko je realizovana vrednost sredine unutar izračunatog intervala poverenja, tada prihvatamo nultu hipotezu na izabranom nivou značajnosti, a u suprotnom slučaju je odbacujemo.

Na osnovu uzorka iz višedimenzionog normalnog rasporeda $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, sa nepoznatom kovarijacionom matricom, generalizacija kvadrata statistike $t-$testa omogućava proveru hipoteze

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \text{, protiv hipoteze } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

ili što u razvijenom obliku pišemo kao

$$H_0: \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{bmatrix}, \text{ protiv hipoteze } H_1: \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \neq \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{bmatrix}$$

Rečeno je u Poglavlju 3.3 da Hotellingova T^2-statistika predstavlja višedimenzioni analogon statistike $t-$rasporeda. Na osnovu nje, korišćenjem realizovanih vrednosti slučajnog uzorka, na nivou značajnosti α odbacujemo H_0, ako je

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0) > \frac{p(n-1)}{n-p} F_{p,n-p;\alpha} \tag{3.31}$$

gde je $\bar{\mathbf{x}}$ i $\mathbf{S}$ realizovana vrednost uzoračke sredine i kovarijacione matrice respektivno. Očigledno je da se, kao i u slučaju jednodimenzione promenljive, i u višedimenzionom slučaju može suditi o tome da li se prihvata nulta hipoteza o vrednosti sredine populacije na osnovu pripadnosti realizovane vrednosti sredine izračunatoj oblasti poverenja.

Primer 3.9 Na osnovu podataka iz Primera 3.8 testirati hipotezu $H_0: \boldsymbol{\mu}' = [1.9 \quad 4.0]$ protiv hipoteze $H_1: \boldsymbol{\mu}' \neq [1.9 \quad 4.0]$ na nivou značajnosti $\alpha = 0.10$.

Izračunata vrednost T^2-statistike je

$$T^2 = 6[4.5-1.9 \quad 4.3-4.0] \begin{bmatrix} 4.2160 & -3.3552 \\ -3.3552 & 2.8642 \end{bmatrix} \begin{bmatrix} 4.5-1.9 \\ 4.3-4.0 \end{bmatrix} = 152.0035$$

Poredeći izračunatu vrednost $T^2 = 152$ sa kritičnom vrednošću

$$\frac{p(n-1)}{n-p} F_{p,n-p;\alpha} = \frac{2(6-1)}{6-2} F_{2,6;0.10} = 2.5 \cdot 4.32 = 10.8$$

zaključujemo da, pošto je $T^2 = 152 > 10.8$, treba odbaciti nultu hipotezu na nivou značajnosti od 10%.

Ako koordinate tačke $[1.9 \quad 4.0]$ uvrstimo umesto μ_1 i μ_2 na levoj strani nejednačine iz Primera 3.8, dobićemo da je vrednost tog izraza jednaka 60.8014. Očigledno je ova vrednost veće od veličine (4.32) na desnoj strani nejednačine, pa konstatujemo da tačka $[1.9 \quad 4.0]$ ne pripada 90% oblasti poverenja. Ovaj rezultat predstavlja ilustraciju ekvivalentnosti dva alternativna pristupa testiranju hipoteze o vrednosti sredine. ■

Ukoliko na osnovu metoda količnika verodostojnosti konstruišemo statistiku testa za testiranje hipoteze $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, protiv hipoteze $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, dobijamo da je količnik verodostojnosti, u oznaci Λ, dat izrazom

$$\Lambda = \frac{\max\limits_{\boldsymbol{\Sigma}} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{\max\limits_{\boldsymbol{\mu},\boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \left(\frac{\left|\hat{\boldsymbol{\Sigma}}\right|}{\left|\hat{\boldsymbol{\Sigma}}_0\right|} \right)^{n/2} \tag{3.32}$$

U izrazu (3.32) je sa $\max\limits_{\boldsymbol{\mu},\boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ označena maksimalna vrednost funkcije verodostojnosti koja se postiže za $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}} = \overline{\mathbf{X}}$ i $\boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\mathbf{S}$, odnosno ukoliko parametre zamenimo njihovim ocenama najveće verodostojnosti; $\max\limits_{\boldsymbol{\Sigma}} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ označava maksimalnu vrednost funkcije verodostojnosti kada je vrednost $\boldsymbol{\mu}$ fiksirana, pa $\hat{\boldsymbol{\Sigma}}_0$ predstavlja ocenu najveće verodostojnosti kovarijacione matrice uz ograničenje da je $\boldsymbol{\mu} = \boldsymbol{\mu}_0$. Ekvivalentna statistika poznata je pod nazivom *Wilksova lambda*. Na osnovu veze[8] koja postoji između T^2 i Λ statistike

$$\Lambda^{2/n} = \left(1 + \frac{T^2}{n-1} \right)^{-1} \tag{3.33}$$

možemo koristiti T^2-statistiku, odnosno njen raspored, prilikom primene Λ statistike u testu hipoteze $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, protiv hipoteze $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. Drugim rečima, kažemo da su T^2-test i test zasnovan na količniku verodostojnosti međusobom ekvivalentni.

Popularnost testa količnika verodostojnosti proizlazi i otuda što u velikim uzorcima statistika $-2\ln\Lambda$ ima aproksimativno χ^2-raspored, pri čemu je broj stepeni slobode jednak razlici broja parametara koji se ocenjuje u modelu bez i sa ograničenjem na parametre rasporeda.

Primer 3.10 Na osnovu podataka iz Primera 2.5 i Primera 3.9 testirati hipotezu $H_0 : \boldsymbol{\mu}' = \begin{bmatrix} 1.9 & 4.0 \end{bmatrix}$ protiv hipoteze $H_1 : \boldsymbol{\mu}' \neq \begin{bmatrix} 1.9 & 4.0 \end{bmatrix}$ na nivou značajnosti $\alpha = 0.10$, korišćenjem asimptotskog rasporeda statistike $-2\ln\Lambda$, bez obzira što aproksimacija rasporeda ove statistike važi samo u velikim uzorcima.

Kako je $\hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\mathbf{S} = \frac{6-1}{6}\begin{bmatrix} 3.500 & 4.100 \\ 4.100 & 5.152 \end{bmatrix} = \begin{bmatrix} 2.9167 & 3.4167 \\ 3.4167 & 4.2708 \end{bmatrix}$ i

[8] Videti: Johnson i Wichern (1982, ss. 185-6).

$$\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_0)(\mathbf{x}_i - \boldsymbol{\mu}_0)' = \frac{1}{n}\mathbf{X}_0'\mathbf{X}_0$$

$$= \frac{1}{6}\begin{bmatrix} 6.0-1.9 & 7.0-1.9 & \cdots & 3.0-1.9 \\ 6.2-4.0 & 3.1-4.0 & \cdots & 1.8-4.0 \end{bmatrix}\begin{bmatrix} 6.0-1.9 & 6.2-4.0 \\ 7.0-1.9 & 3.1-4.0 \\ \vdots & \vdots \\ 3.0-1.9 & 1.8-4.0 \end{bmatrix} = \begin{bmatrix} 9.6767 & 5.5633 \\ 5.5633 & 6.5167 \end{bmatrix}$$

to je $\left|\hat{\boldsymbol{\Sigma}}\right| = 1.275$ i $\left|\hat{\boldsymbol{\Sigma}}_0\right| = 32.1089$. Na osnovu

$$\lambda = \left(\frac{\left|\hat{\boldsymbol{\Sigma}}\right|}{\left|\hat{\boldsymbol{\Sigma}}_0\right|}\right)^{n/2} = \left(\frac{1.1275}{32.1089}\right)^3 = 0.00004324,$$ sledi da je $-2\ln\lambda = 20.0948$.

Kako smo u modelu bez ograničenja ocenili dve srednje vrednosti i tri elementa kovarijacione matrice, ukupno 5 parametara, a u modelu sa ograničenjem na parametre ocenili tri elementa kovarijacione matrice, to je broj stepeni slobode $5-3=2$. Kritična vrednost je $\chi^2_{2;0.10} = 4.61$, pa pošto je $-2\ln\lambda = 20.1 > 4.61$ zaključujemo da treba odbaciti hipotezu $H_0 : \boldsymbol{\mu}' = \begin{bmatrix} 1.9 & 4.0 \end{bmatrix}$. Do izračunate vrednosti statistike Λ mogli smo doći i na osnovu relacije koja postoji između nje i T^2 – statistike (izraz (3.33)). ■

3.4.3 Simultani intervali poverenja

Prilikom testiranja hipoteze o vrednosti sredine može se dobiti rezultat, kao što je uostalom dobijen u Primeru 3.9, da se odbacuje nulta hipoteza. Tada se postavlja pitanje koja od komponenti sredine je svojom razlikom u odnosu na hipotetsku vrednost doprinela odbacivanju nulte hipoteze. Do odgovora na ovo pitanje možemo doći tako što ćemo testirati hipotezu o individualnoj vrednosti svake od komponenti sredine. Međutim, primena baterije individualnih t – testova ne omogućava kontrolisanje nivoa značajnosti testa za sve komponente sredine posmatrane zajedno. Isto važi ako računamo individualne intervale poverenja bazirane na t – rasporedu za svaku od komponenti. Tada nećemo biti u mogućnosti odrediti pravu vrednost koeficijenta poverenja za sve komponente sredine posmatrane zajedno. Određujući, na primer 95% interval poverenja za svaki element sredine populacije, verovatnoća da će interval poverenja uključiti pravu vrednost tog elementa sredine jednaka je 0.95. Ako su promenljive međusobom nezavisne, a formirali smo p individualnih intervala poverenja, tada je verovatnoća da će svi individualni intervali poverenja obuhvatiti pravu vrednost sredine. Ili, posmatrano iz drugog ugla, $1-0.95^p$ predstavlja verovatnoću da bar jedan interval poverenja neće obuhvatiti pravu vrednost komponente sredine. Računajući veći

broj intervala poverenja, na primer $p=4$, verovatnoća da bar jedan interval poverenja ne obuhvati pravu vrednost iznosi $1-0.95^4=0.1855$. Znači da za veće vrednosti p raste i ova verovatnoća, pa se igrom slučaja može dobiti rezultat da prava vrednost komponente sredine ne pripada individualnom intervalu poverenja. Međutim, u višedimenzionom slučaju promenljive su retko međusobom nezavisne, pa neće biti ni individualni intervali poverenja. Znači da će verovatnoća da *svi* individualni intervali poverenja obuhvate pravu vrednost sredine biti veća od 0.95^p.

Jedno rešenje navedenog problema kontrole verovatnoće greške prve vrste, α, odnosno zajedničkog koeficijenta poverenja, $1-\alpha$, je u korišćenju T^2-statistike za formiranje simultanih intervala poverenja. Da bismo definisali simultane intervale poverenja na osnovu sredine slučajnog uzorka $\overline{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$ formiraćemo linearnu kombinaciju

$$\mathbf{a}'\overline{\mathbf{X}} = a_1\overline{X}_1 + a_2\overline{X}_2 + ... + a_p\overline{X}_p \tag{3.34}$$

gde je $\mathbf{a}' = [a_1, a_2, ..., a_p]$ vektor konstanti. Za $\mathbf{a}'\overline{\mathbf{X}}$ je rečeno da ima raspored $N_p(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}/n)$. Za fiksnu vrednost koeficijenata linearne kombinacije i nepoznatu varijansu, definišemo realizovanu vrednost $100(1-\alpha)\%$ intervala poverenja linearne kombinacije $\mathbf{a}'\overline{\mathbf{X}}$, bazirajući se na $t-$rasporedu

$$\mathbf{a}'\overline{\mathbf{x}} - \sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}t_{n-1;\alpha/2} < \mathbf{a}'\boldsymbol{\mu} < \mathbf{a}'\overline{\mathbf{x}} + \sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}t_{n-1;\alpha/2} \tag{3.35}$$

gde je $t_{n-1;\alpha/2}$ gornji $100(\alpha/2)$ percentil $t-$rasporeda sa $n-1$ stepenom slobode. Zamenom $\mathbf{a}' = [1, 0, ..., 0]$ dobili bismo $\mathbf{a}'\boldsymbol{\mu} = \mu_1$, pa bi se interval poverenja sveo na individualni $100(1-\alpha)\%$ interval poverenja prve komponente sredine. Tako za prvu komponentu sredine imamo realizovanu vrednost tog intervala

$$\overline{x}_1 - \sqrt{\frac{s_{11}}{n}}t_{n-1;\alpha/2} < \mu_1 < \overline{x}_1 + \sqrt{\frac{s_{11}}{n}}t_{n-1;\alpha/2} \tag{3.36}$$

Na sličan način odredili bismo individualne intervale poverenja za preostale komponente sredine populacije. Mada su koeficijenti poverenja svakog intervala poverenja ponaosob jednaki $1-\alpha$, združeno posmatrano njihov zajednički koeficijent poverenja nije, kao što rekosmo, $1-\alpha$. Ako želimo da fiksiramo koeficijent poverenja na $1-\alpha$ za svaku, proizvoljnu vrednost koeficijenata linearne kombinacije tada ćemo koristiti rezultat, da za ma koje vrednosti $a_1, a_2, ..., a_p$ interval

$$\mathbf{a}'\overline{\mathbf{X}} - \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p;\alpha}\mathbf{a}'\mathbf{S}\mathbf{a}} : \mathbf{a}'\overline{\mathbf{X}} + \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p;\alpha}\mathbf{a}'\mathbf{S}\mathbf{a}} \tag{3.37}$$

sadrži $\mathbf{a}'\boldsymbol{\mu}$ sa verovatnoćom $1-\alpha$. Ove simultane intervale svake komponente sredina vektora, bazirane na T^2-rasporedu, nazivamo T^2-intervalima. Sledećim izborom vektora koeficijenata linearne kombinacije: $\mathbf{a}'=[1,0,...,0]$, $\mathbf{a}'=[0,1,...,0]$ do $\mathbf{a}'=[0,0,...,1]$, dobijamo realizovane vrednosti intervala svake komponente sredine

$$\begin{aligned}
&\bar{x}_1-\sqrt{\frac{s_{11}}{n}}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p;\alpha}}<\mu_1<\bar{x}_1+\sqrt{\frac{s_{11}}{n}}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p;\alpha}}\\
&\bar{x}_2-\sqrt{\frac{s_{22}}{n}}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p;\alpha}}<\mu_2<\bar{x}_2+\sqrt{\frac{s_{22}}{n}}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p;\alpha}}\\
&\qquad\qquad\cdots\\
&\bar{x}_p-\sqrt{\frac{s_{pp}}{n}}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p;\alpha}}<\mu_p<\bar{x}_p+\sqrt{\frac{s_{pp}}{n}}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p;\alpha}}
\end{aligned}\qquad(3.38)$$

koji važe simultano i to sa koeficijentom poverenja $1-\alpha$. Poredeći individualne $t-$intervale i simultane T^2-intervale možemo konstatovati da za fiksno n, p i α ovi potonji su širi u odnosu na individualne intervale. Dakle, da bismo imali, za ma koji izbor koeficijenata linearne kombinacije, fiksnu vrednost zajedničkog koeficijenta poverenja $1-\alpha$, moramo ići na štetu preciznosti intervalnog ocenjivanja. Drugim rečima, cenu za fiksnu vrednost koeficijenta poverenja platili smo povećanjem dužine intervala poverenja u odnosu na individualne intervale poverenja.

Primer 3.11 U Primeru 3.8 odredili smo 90% oblast poverenja za broj članova domaćinstva i izdataka na ishranu (za podatke i objašnjenje promenljivih videti Primer 2.5). Na osnovu podataka u navedenim primerima odrediti 90% $t-$intervale i T^2-intervale. Odrediti za koliko su relativno širi simultani u odnosu na individualne intervale poverenja.

Za vektor $\mathbf{a}'=\begin{bmatrix}1 & 0\end{bmatrix}$ dobijamo $\mathbf{a}'\boldsymbol{\mu}=\mu_1$, $\mathbf{a}'\bar{\mathbf{x}}=\bar{x}_1=4.5$, $\mathbf{a}'\mathbf{S}\mathbf{a}=$ $=s_{11}=3.5$, pa su uz $t_{5;0.05}=2.015$, 90% $t-$intervali

$$4.5-\sqrt{\frac{3.5}{6}}2.015<\mu_1<4.5+\sqrt{\frac{3.5}{6}}2.015,\ \text{odn.}\ 2.961<\mu_1<6.039,$$

a za vektor $\mathbf{a}'=\begin{bmatrix}0 & 1\end{bmatrix}$ proveriti da se dobija da $\mu_2\in(2.433;6.167)$. Kako je $F_{2,4;0.10}=4.32$ to su 90% T^2-intervali

$$4.5-\sqrt{\frac{3.5}{6}}\sqrt{\frac{2(6-1)}{6-2}4.32}<\mu_1<4.5+\sqrt{\frac{3.5}{6}}\sqrt{\frac{2(6-1)}{6-2}4.32},\ \text{odn.}\ 1.99<\mu_1<7.01,$$

a za drugu komponentu sredine imamo da $\mu_2\in(1.255;7.345)$.

Vrednost množitelja $\sqrt{s_{jj}/n}$ kod $t-$intervala i T^2-intervala iznosi $t_{5;0.05}=2.015$ i $\sqrt{\frac{2(6-1)}{6-2}4.32}=3.2863$ respektivno, pa zaključujemo da su simultani intervali $100(3.2863-2.015)/2.015=63.09\%$ širi od individualnih intervala poverenja. ■

Pored korišćenja T^2-rasporeda za formiranje simultanih intervala poverenja u literaturi se definišu i drugi metodi koji razrešavaju problem pridruživanja zajedničkog koeficijenta poverenja simultanim intervalima poverenja. Takav je na primer, Bonferronijev metod, koji u svojoj operacionalizaciji sugeriše zamenu kritičnih vrednosti $T-$statistike kod simultanih intervala poverenja, odnosno $T_{p,n-p;\alpha}$, vrednostima $t_{p,n-1;\alpha/2}$. Time se postiže da zajednički koeficijent poverenja simultanih intervala bude veći ili jednak $1-\alpha$. O ovom metodu višestrukog poređenja sredina videti kod Johnsona i Wicherna (1982, ss. 197-9).

Zaključivanje o sredini na osnovu slučajnog uzorka iz dve i više populacija razmatramo u 5. Glavi u kojoj se bavimo generalizacijom analize varijanse na višedimenzioni slučaj.

3.5 ZAKLJUČIVANJE O KOVARIJACIONOJ I KORELACIONOJ MATRICI

Pretpostavke za primenu nekih metoda multivarijacione analize bazirane su na osobinama kovarijacione ili korelacione matrice. Tako na primer, dijagonalnost kovarijacione matrice sugeriše kod višedimenzionog normalnog rasporeda nezavisnost promenljivih. Ili, u multivarijacionoj analizi varijanse pretpostavlja se jednakost kovarijacionih matrica populacija odakle su uzeti uzorci.

Razmotrimo problem testa nulte hipoteze $H_0:\Sigma=\sigma^2\mathbf{I}$ protiv hipoteze $H_0:\Sigma\neq\sigma^2\mathbf{I}$. Prema nultoj hipotezi vandijagonalni elementi kovarijacione matrice jednaki su nuli, što znači da su p slučajnih promenljivih normalno raspoređenog slučajnog vektora $\mathbf{X}$, međusobom nezavisni. S druge strane, njihove varijanse jednake su međusobom i jednake su zajedničkoj varijansi σ^2. Sa ovakvom strukturom kovarijacione matrice elipsoid konstantne gustine verovatnoće $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=d^2$ svodi se na $(\mathbf{x}-\boldsymbol{\mu})'(\mathbf{x}-\boldsymbol{\mu})=\sigma^2d^2$, što za $p=2$ predstavlja jednačinu kružnice, za $p=3$ jednačinu kugle, a u opštem slučaju sferu sa centrom u tački $\boldsymbol{\mu}$. Iz tog razloga navedena nulta hipoteza naziva se *hipoteza sferičnosti*. Konstruišući test na principu količnika verodostojnosti može se pokazati da se dobijena statistika testa za testiranje nulte hipoteze $H_0:\Sigma=\sigma^2\mathbf{I}$ svodi na

$$n\,p\ln\frac{\frac{1}{p}\sum_{j=1}^{p}\lambda_j}{\left(\prod_{j=1}^{p}\lambda_j\right)^{\frac{1}{p}}} \tag{3.39}$$

gde je razlomkom dat količnik aritmetičke i geometrijske sredine karakterističnih korena $\lambda_1,\lambda_2,...,\lambda_p$ uzoračke kovarijacione matrice $\mathbf{S}$. Ako je tačna nulta hipoteza ova statistika testa ima χ^2-asimptotski raspored sa $(p-1)(p+2)/2$ stepeni slobode. Nulta hipoteza se prihvata na izabranom nivou značajnosti ako je izračunata vrednost statistike manja od kritične vrednosti.

Primer 3.12 Za podatke iz Primera 2.5 odrediti karakteristične korene uzoračke kovarijacione matrice. Na osnovu njih testirati hipotezu sferičnosti za kovarijacionu matricu populacije.

U Primeru 2.5 dobili smo uzoračku kovarijacionu matricu

$$\mathbf{S}=\begin{bmatrix} 3.5 & -0.860 & 4.100 & -0.12000 \\ & 2.416 & -1.510 & 0.05240 \\ & & 5.152 & -0.15680 \\ & & & 0.00608 \end{bmatrix}.$$

Njeni karakteristični koreni su: $\lambda_1=8.9624$, $\lambda_2=2.0098$, $\lambda_3=0.1008$ i $\lambda_4=0.0012$. Kako je njihova aritmetička sredina jednaka 2.7685, a geometrijska sredina je 0.2164, to je izračunata vrednost statistike testa za testiranje nulte hipoteze $H_0:\boldsymbol{\Sigma}=\sigma^2\mathbf{I}$

$$6\cdot 4\cdot\ln\frac{2.7685}{0.2164}=61.1758\,.$$

Ako je tačna nulta hipoteza statistika testa ima asimptotski χ^2-raspored sa $(4-1)(4+2)/2=18$ stepeni slobode. Na nivou značajnosti $\alpha=0.05$, kritična vrednost iznosi $\chi^2_{18;0.05}=28.9$. Kako je izračunata vrednost veća od kritične, odbacuje se hipoteza o sferičnosti. ■

Korišćenjem principa količnika verodostojnosti konstruišemo test za testiranje hipoteze o strukturi korelacione matrice populacije. Nulta hipoteza $H_0:\boldsymbol{\rho}=\mathbf{I}$ protiv hipoteze $H_1:\boldsymbol{\rho}\neq\mathbf{I}$, slično kao kod zaključivanja o kovarijacionoj matrici, u slučaju višedimenzionog normalnog rasporeda sugeriše međusobnu nezavisnost p promenljivih

slučajnog vektora $\mathbf{X}$. Količnik verodostojnosti svodi se na $|\mathbf{R}|^{n/2}$, gde je $\mathbf{R}$ uzoračka korelaciona matrica. Umesto uobičajene transformacije ovog količnika verodostojnosti $(-2\ln\Lambda)$ koristimo *Bartlettovu statistiku* koja bolje aproksimira χ^2-raspored:

$$-\left(n-1-\frac{2p+5}{6}\right)\ln|\mathbf{R}|=-\left(n-1-\frac{2p+5}{6}\right)\sum_{j=1}^{p}\ln\lambda_j \tag{3.40}$$

gde su $\lambda_j, j=1,2,...,p$ karakteristični koreni uzoračke korelacione matrice. Statistika testa ima asimptotski χ^2-raspored sa $p(p-1)/2$ stepeni slobode. Nulta hipoteza se prihvata na izabranom nivou značajnosti, ako je izračunata vrednost statistike testa manja od kritične vrednosti.

Primer 3.13 Na osnovu podataka iz Primera 2.5 testirati nultu hipotezu $H_0:\boldsymbol{\rho}=\mathbf{I}$ korišćenjem Bartlettove statistike (3.40).

U Primeru 2.5 izračunali smo uzoračku korelacionu matricu

$$\mathbf{R}=\begin{bmatrix}1 & -0.2975 & 0.9655 & -0.8226\\ & 1 & -0.4280 & 0.4323\\ & & 1 & -0.8859\\ & & & 1\end{bmatrix}.$$

Njeni karakteristični koreni su $\lambda_1=3.0067$, $\lambda_2=0.7945$ i $\lambda_3=0.1791$. Izračunata vrednost Bartlettove statistike je 13.5514. Broj stepeni slobode je 6. Na nivou značajnosti $\alpha=0.05$ tablična vrednost χ^2-statistike je $\chi^2_{6;0.05}=12.6$. Kako je izračunata vrednost Bartlettove statistike veća od kritične vrednosti odbacuje se hipoteza o nultoj vrednosti korelacije unutar skupa od četiri posmatrane promenljive. ■

U ovom poglavlju razmatrali smo problem zaključivanja o kovarijacionoj i korelacionoj matrici jedne populacije. Statističkim postupcima zaključivanja u slučaju više populacija, kao i specifičnim strukturama kovarijacione i korelacione matrice vratićemo se u 5. Glavi posvećenoj multivarijacionoj analizi varijanse, odnosno 7. i 8. Glavi u kojima se bavimo metodama analize kovarijacione strukture (metod glavnih komponenata i metod faktorske analize).

4 KANONIČKA KORELACIONA ANALIZA

Metod multivarijacione analize koji se bavi utvrđivanjem postojanja veze i jačine povezanosti dva skupa promenljivih nazivamo *kanonička korelaciona analiza*. Na primer, u ekonomiji interes istraživača usmeren je ka ispitivanju povezanosti skupa upravljačkih promenljivih koje su na raspolaganju nosiocima ekonomske politike i skupa ciljnih makroekonomskih promenljivih kao što su lična, opšta i zajednička potrošnja. Ili, postavlja se pitanje povezanosti izvesnih "demografskih" promenljivih domaćinstva (starost domaćina, broj dece, obrazovni nivo, dohodak i sl.) i "potrošačkih" promenljivih (izdaci na određene robe, učestalost posete restoranu i sl.). Oba navedena primera sugerišu da postoje dva međusobom jasno razdvojena skupa promenljivih, ali između kojih je na osnovu makroekonomske teorije ili na osnovu modela ponašanja potrošača moguće apriori uspostaviti relaciju. Kanonička korelaciona analiza u takvim slučajevima omogućava kvantifikaciju međusobne povezanosti i detaljno ispitivanje te veze.

Izlaganje ovog metoda multivarijacione analize započinjemo ukazivanjem na njene ciljeve, da bismo zatim definisali populacioni model i osnovne veličine kanoničke korelacione analize: kanoničke promenljive, kanoničku korelaciju i kanonička opterećenja. Nakon eksplicitno utvrđene veze između kanoničke korelacione analize i drugih koeficijenata korelacije kao i modela regresione analize, izlaganje posvećujemo uzoračkoj kanoničkoj korelacionoj analizi (ocenjivanje i testiranje kanoničke korelacije), a na kraju, na osnovu pokazatelja kanoničke korelacione analize i mere redundantnosti, sugerišemo način interpretacije dobijenih rezultata.

4.1 UVOD

Uobičajeno sredstvo analize međusobne povezanosti dve promenljive jeste koeficijent korelacije. Ukoliko nam je cilj da ispitamo stepen međusobne veze jedne promenljive (zavisne promenljive) i skupa drugih promenljivih (nezavisnih promenljivih), tada se koristimo modelom regresione analize i koeficijentom višestruke korelacije. Međutim, ukoliko u analizi imamo više od jedne zavisne promenljive, odnosno skup dve i više

zavisnih promenljivih, tada se za utvrđivanje stepena povezanosti dva skupa promenljivih, zavisnih i nezavisnih, koristi kanonička korelaciona analiza. Na osnovu rečenog ona se može posmatrati i kao uopštenje modela regresione analize i koeficijenta obične kao i višestruke korelacije. Nešto kasnije ćemo se vratiti na ovakvo viđenje kanoničke korelacije. Napominjemo da bez obzira što u izlaganju kanoničke korelacione analize koristimo nazive skup zavisnih i skup nezavisnih promenljivih, da bismo naznačili dva posmatrana skupa promenljivih, između njih ne mora nužno postojati uzročna veza, kao što je to slučaj sa skupovima promenljivih u regresionoj analizi.

Originalni teorijski doprinos razvoju analize kanoničke korelacije dao je Hotelling 1936. godine. Jedan od prvih primera korišćenja analize kanoničke korelacije u ekonomiji bio je Tintnerov (1946) pokušaj određivanja zavisnosti između tri indeksa cena proizvođača (indeks cena poljoprivrednih proizvoda, indeks cena prehrambenih proizvoda i indeks ostalih cena) i četiri indeksa proizvodnje (indeks proizvodnje potrošnih dobara, indeks proizvodnje trajnih dobara, indeks proizvodnje sirovina i indeks poljoprivredne proizvodnje). Prvi skup promenljivih okarakterisan je kao skup indeksa cena proizvođača, a drugi kao skup indeksa proizvodnje. Analiza kanoničke korelacije treba po Tintneru da odgovori na pitanje "da li se linearna funkcija indeksa cena proizvođača i linearna funkcija indeksa proizvodnje, kao predstavnici svojih skupova indeksa, mogu izučavati na smislen način". Očigledno je da bi za ova dva skupa promenljivih sa velikim brojem elemenata potvrdan odgovor na postavljeno pitanje bio od velike koristi.

Ukratko i veoma pojednostavljeno rečeno, cilj kanoničke korelacione analize jeste formiranje linearne kombinacije unutar skupa zavisnih i posebno unutar skupa nezavisnih promenljivih, ali tako da između te dve linearne kombinacije postoji maksimalna korelacija. Na osnovu tih i tako određenih linearnih kombinacija možemo proniknuti u međusobnu povezanost dva skupa promenljivih. Dakle, ako ove linearne kombinacije dobro reprezentuju svaka svoj skup promenljivih, tada je problem izučavanja veze između dva skupa promenljivih pojednostavljen u tom smislu što posmatramo samo par izvedenih promenljivih umesto većeg broja originalnih promenljivih iz ova dva skupa. Pored ovog glavnog cilja kanoničke korelacione analize može nas interesovati izvođenje dodatnog para linearnih kombinacija koji će maksimizirati preostalu korelaciju uz uslov nekorelisanosti sa prethodno izdvojenim parom linearnih kombinacija. Ovaj proces izdvajanja parova linearnih kombinacija može se dalje nastaviti. Ili, interesuje nas objašnjenje prirode veze između skupa nezavisnih i zavisnih promenljivih. Do tog objašnjenja dolazimo na osnovu merenja relativnog doprinosa svake promenljive povezanosti izdvojenog para linearnih kombinacija.

Korišćenje kanoničke korelacione analize u deskriptivne svrhe nije zasnovano na pretpostavkama o rasporedu slučajnih vektora. U ovom slučaju mogu se koristiti i promenljive čija su merenja iskazana na nominalnoj ili ordinarnoj skali. Međutim, statistički postupak zaključivanja o značajnosti izvedene veze kao i značajnosti kanoničke korelacije zahteva uvođenje pretpostavke o višedimenzionom normalnom rasporedu.

4.2 KANONIČKE PROMENLJIVE I KANONIČKA KORELACIJA

4.2.1 Definicija i osobine

Posmatrajmo dve slučajna vektora $\mathbf{X}'=[X_1, X_2, ..., X_p]$ i $\mathbf{Y}'=[Y_1, Y_2, ..., Y_q]$ koji reprezentuju dva skupa promenljivih. Interesuje nas da utvrdimo stepen međusobne povezanosti ove dve grupe promenljivih. Pretpostavićemo da je $p \le q$. Za potrebe statističkog zaključivanja o kanoničkoj korelaciji pretpostavićemo takođe da je $\mathbf{X} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ i $\mathbf{Y} \sim N_q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$. Sa $Cov(\mathbf{X}, \mathbf{Y}) = \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^{'}$ označićemo $(p \times q)$ kovarijacionu matricu koja meri stepen zavisnosti između $\mathbf{X}$ i $\mathbf{Y}$. Združeno posmatrajmo slučajne vektore $\mathbf{X}$ i $\mathbf{Y}$, tako što ćemo formirati $((p+q)\times 1)$ slučajan vektor $[\mathbf{X}' \vdots \mathbf{Y}']$. Njegova sredina i kovarijaciona matrica su

$$\underset{((p+q)\times 1)}{\boldsymbol{\mu}} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \cdots \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ i } \underset{(p+q)\times(p+q)}{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \vdots & \boldsymbol{\Sigma}_{12} \\ \cdots & \cdots & \cdots \\ \boldsymbol{\Sigma}_{21} & \vdots & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \tag{4.1}$$

Povezanost promenljivih koje pripadaju različitim skupovima promenljivih iskazujemo preko elemenata matrice $\boldsymbol{\Sigma}_{12}$. Sažimanje tog, najčešće velikog broja kovarijansi postižemo u okviru kanoničke korelacione analize izborom manjeg broja kovarijansi koje su dobijene na osnovu linearne kombinacije promenljivih iz ova dva skupa promenljivih.

Neka su $Z = \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_p X_p = \boldsymbol{\alpha}'\mathbf{X}$ i $W = \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_q Y_q = \boldsymbol{\beta}'\mathbf{Y}$ linearne kombinacije promenljivih koje pripadaju slučajnom vektoru $\mathbf{X}$ i $\mathbf{Y}$ respektivno. Na osnovu rezultata (2.17), (2.18) i (2.22) imamo da je

$$E(Z) = \boldsymbol{\alpha}'\boldsymbol{\mu}_1 \qquad Var(Z) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} \tag{4.2a}$$

$$E(W) = \boldsymbol{\beta}'\boldsymbol{\mu}_2 \qquad Var(W) = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta} \tag{4.2b}$$

$$Cov(Z, W) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\beta} \tag{4.2c}$$

Prema tome, koeficijent korelacije između Z i W, u oznaci ρ_{ZW}, je

$$\rho_{ZW} = \frac{Cov(Z, W)}{\sqrt{Var(Z)}\sqrt{Var(W)}} = \frac{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}}\sqrt{\boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}}}. \tag{4.3}$$

Od svih koeficijenata linearnih kombinacija biramo onaj par $\boldsymbol{\alpha}$ i $\boldsymbol{\beta}$ za koji se postiže maksimalna vrednost koeficijenta korelacije ρ_{ZW}. Kako se vrednost koeficijenta korelacije ne menja linearnom transformacijom $\boldsymbol{\alpha}$ i $\boldsymbol{\beta}$, izabraćemo njihovu vrednost

tako da Z i W imaju jedinične varijanse, tj. da je $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta} = 1$. Sada se naš zadatak svodi na klasičan problem maksimiziranja funkcije $\rho_{ZW} = \rho_{ZW}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ po koeficijentima linearne kombinacije, ili što je ekvivalentno, funkcije $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}$, uz ograničenja $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta} = 1$. Pokazano[1] je da je ovaj problem ekvivalentan problemu rešavanja sledećeg sistema homogenih jednačina

$$\left(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \lambda\mathbf{I}_p\right)\boldsymbol{\alpha} = 0 \tag{4.4a}$$

$$\left(\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \lambda\mathbf{I}_q\right)\boldsymbol{\beta} = 0\,, \tag{4.4b}$$

gde su $\mathbf{I}_p\,(p\times p)$ i $\mathbf{I}_q\,(q\times q)$ jedinične matrice, a λ je najveći karakteristični koren sledećih karakterističnih jednačina

$$\left|\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \lambda\mathbf{I}_p\right| = 0 \tag{4.5a}$$

$$\left|\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \lambda\mathbf{I}_q\right| = 0\,. \tag{4.5b}$$

Ovaj najveći karakteristični koren predstavlja *kvadrat koeficijenta kanoničke korelacije*. Dakle, do rešenja sistema homogenih jednačina dolazimo određivanjem karakterističnih korena matrica $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ i $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. Kako smo pretpostavili da je $p \le q$, na osnovu prve matrice imaćemo manji broj karakterističnih korena i vektora nego na osnovu druge. Međutim, pozitivni karakteristični koreni ove dve matrice su identični i njihov broj je jednak p. Ostalih $q - p$ karakterističnih korena druge matrice imaće vrednost nula. Poređajmo dobijene karakteristične korene po veličini, $\lambda_1, \lambda_2, ..., \lambda_p$, a dobijene jedinstveno određene karakteristične vektore označimo sa $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_p$. Jedinstvenost karakterističnih vektora obezbeđuje uslov prema kome su varijanse linearnih kombinacija jednake jedinici. Do karakterističnih vektora $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_p$ dolazimo na osnovu relacije

$$\boldsymbol{\beta}_i = \frac{\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}_i}{\sqrt{\lambda_i}}, i = 1, 2, ..., p\,. \tag{4.6}$$

Samom oznakom koja se koristi za karakteristične vektore sugeriše se da oni predstavljaju istovremeno i koeficijente linearne kombinacije. Tako imamo za najveći karakteristični koren λ_1 (tj. kvadrat koeficijenta kanoničke korelacije) odgovarajući par karakterističnih vektora $\boldsymbol{\alpha}_1$ i $\boldsymbol{\beta}_1$ (tj. koeficijenata linearnih kombinacija) koje nazivamo *kanonički koeficijenti* ili *ponderi*, a $\boldsymbol{\alpha}_1'\mathbf{X}$ i $\boldsymbol{\beta}_1'\mathbf{Y}$ predstavlja prvi par *kanoničkih promenljivih*. Nakon izdvajanje prvog para kanoničkih promenljivih, drugi par biramo

[1] Videti: Andreson (1972, ss. 289-96).

tako da je od preostalih linearnih kombinacija on sa najvećom međusobnom korelacijom i uz to nekorelisan sa prvim parom kanoničkih promenljivih. Na sličan način izdvajamo ostale parove. Preostale linearne kombinacije za $\mathbf{X}$ označene sa $Z_2, Z_3, ..., Z_p$, međusobno su nekorelisane i svaka je nekorelisana sa Z_1. Slično važi i za linearne kombinacije za $\mathbf{Y}$ označene sa $W_2, W_3, ..., W_p$, koje su međusobno nekorelisane i svaka je nekorelisana sa W_1. Štaviše, različiti elementi ova dva skupa linearnih kombinacija su međusobom nekorelisani, odn. Z_j je nekorelisano sa W_k, $j \neq k$, $j, k = 1, 2, ..., p$.

Sumirajući rečeno konstatujemo da je $j-$ti najveći koeficijent kanoničke korelacije $\sqrt{\lambda_j}$, kvadratni koren karakterističnog korena matrice $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Istovremeno λ_j je kvadrat običnog koeficijenta korelacije izmedu $Z_j = \boldsymbol{\alpha}_j'\mathbf{X}$ i $W_j = \boldsymbol{\beta}_j'\mathbf{Y}$, gde je $\boldsymbol{\alpha}_j$ karakteristični vektor koji odgovara λ_j, a $\boldsymbol{\beta}_j = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}_j / \sqrt{\lambda_i}$. Vektori $\boldsymbol{\alpha}_j$ i $\boldsymbol{\beta}_j$ normirani su tako da je varijansa linearne kombinacije jednaka jedinici. Par promenljivih (Z_j, W_j) predstavljaju kanoničke promenljive sa koeficijentom kanoničke korelacije $\sqrt{\lambda_j}$.

Izlaganje populacionog modela bazirali smo na kovarijacionim matricama. U tom slučaju izvedene kanoničke promenljive iskazane su u originalnim jedinicama mere promenljivih iz $\mathbf{X}$ i $\mathbf{Y}$. Ukoliko se umesto kovarijacionih matrica $(\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}$ i $\boldsymbol{\Sigma}_{12})$ koriste korelacione matrice $(\boldsymbol{\rho}_{11}, \boldsymbol{\rho}_{22}$ i $\boldsymbol{\rho}_{12})$ dobiće se ista vrednost koeficijenata kanoničke korelacije, ali će kanonički koeficijenti biti drugačiji. Istovremeno će kanoničke promenljive biti izražene preko standardizovanih vrednosti originalnih promenljivih. Kanoničke koeficijente označavaćemo u oba slučaja sa $\boldsymbol{\alpha}$, odn. $\boldsymbol{\beta}$, a iz konteksta će biti jasno da li su u pitanju standardizovani koeficijenti ili ne.

Primer 4.1 Neka je $\mathbf{X}' = [X_1, X_2]$ i $\mathbf{Y}' = [Y_1, Y_2]$. Za ove slučajne vektore data je kovarijaciona matrica

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 2 & -1 & 3 \\ 2 & 3 & 1 & 1 \\ -1 & 1 & 6 & 2 \\ 3 & 1 & 2 & 8 \end{bmatrix}.$$

Odrediti korelacionu matricu, a zatim na osnovu:

(a) date kovarijacione matrice i

(b) izračunate korelacione matrice,

odrediti koeficijente kanoničke korelacije i kanoničke koeficijente. Napisati prvi par kanoničkih promenljivih.

Na osnovu poznate relacije (2.24) između kovarijacione i korelacione matrice dobijamo korelacionu matricu

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & 0.6667 & -0.2357 & 0.6124 \\ & 1 & 0.2357 & 0.2041 \\ & & 1 & 0.2887 \\ & & & 1 \end{bmatrix}.$$

U skladu sa podelom matrice $\boldsymbol{\Sigma}$ i $\boldsymbol{\rho}$ sledi da je

$$\boldsymbol{\Sigma}_{11} = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix},\ \boldsymbol{\Sigma}_{22} = \begin{bmatrix} 6 & 2 \\ 2 & 8 \end{bmatrix} \text{ i } \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}' = \begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix},$$

$$\boldsymbol{\rho}_{11} = \begin{bmatrix} 1 & 0.6667 \\ & 1 \end{bmatrix},\ \boldsymbol{\rho}_{22} = \begin{bmatrix} 1 & 0.2887 \\ & 1 \end{bmatrix} \text{ i } \boldsymbol{\rho}_{12} = \boldsymbol{\rho}_{21}' = \begin{bmatrix} -0.2357 & 0.6124 \\ 0.2357 & 0.2041 \end{bmatrix}.$$

(a) Na osnovu date kovarijacione matrice potrebno je odrediti karakteristične korene matrice $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Kako je

$$\boldsymbol{\Sigma}_{11}^{-1} = \begin{bmatrix} 0.6 & -0.4 \\ -0.4 & 0.6 \end{bmatrix} \text{ i } \boldsymbol{\Sigma}_{22}^{-1} = \begin{bmatrix} 0.1818 & -0.0455 \\ -0.0455 & 0.1364 \end{bmatrix} \text{ to je}$$

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \begin{bmatrix} 0.9545 & -0.0091 \\ -0.5909 & 0.0818 \end{bmatrix},$$

a njeni karakteristični koreni su $\lambda_1 = 0.9607$ i $\lambda_2 = 0.0757$. Sledi da je prvi koeficijent kanoničke korelacije $\sqrt{\lambda_1} = 0.9801$, a drugi $\sqrt{\lambda_2} = 0.2751$.

Proveriti da se isti rezultat dobija na osnovu matrice

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \begin{bmatrix} 0.4545 & -0.5273 \\ -0.3636 & 0.5818 \end{bmatrix}.$$

Za prvi karakteristični koren odredimo pridruženi karakteristični vektor $\boldsymbol{\alpha}_1^*$ na osnovu izraza

$$\begin{bmatrix} 0.9545 & -0.0091 \\ -0.5909 & 0.0818 \end{bmatrix} \boldsymbol{\alpha}_1^* = 0.9607\boldsymbol{\alpha}_1^*$$

odakle je $\boldsymbol{\alpha}_1^{*\prime} = [1 \quad -0.6724]$, pa se kanonički koeficijenti dobijaju direktno na osnovu uslova normalizacije $Var(Z_1) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} = 1$:

$$\begin{bmatrix}1 & -0.6724\end{bmatrix}\begin{bmatrix}3 & 2\\ 2 & 3\end{bmatrix}\begin{bmatrix}1\\ -0.6724\end{bmatrix} = 1.6668\,.$$

Korišćenjem $\sqrt{1.6668} = 1.2910$ dobijamo kanoničke koeficijente $\boldsymbol{\alpha}_1$

$$\boldsymbol{\alpha}_1 = \frac{1}{1.2910}\begin{bmatrix}1\\ -0.6724\end{bmatrix} = \begin{bmatrix}0.7746\\ -0.5208\end{bmatrix}.$$

Preko relacije $\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}_1 / \sqrt{\lambda_1}$ dobijamo $\boldsymbol{\beta}_1' = [-0.3239 \quad 0.3109]$. Dakle, prvi par kanoničkih promenljivih je

$$Z_1 = \boldsymbol{\alpha}_1'\mathbf{X} = 0.7746\,X_1 - 0.5208\,X_2$$

$$W_1 = \boldsymbol{\beta}_1'\mathbf{Y} = -0.3239\,Y_1 + 0.3109\,Y_2\,,$$

a njihova kanonička korelacija je $\sqrt{0.9607} = 0.9801$. Pokazati da se na osnovu $\lambda_2 = 0.0757$ dobija drugi par kanoničkih promenljivih

$$Z_2 = \boldsymbol{\alpha}_2'\mathbf{X} = 0.0059\,X_1 + 0.5734\,X_2$$

$$W_2 = \boldsymbol{\beta}_2'\mathbf{Y} = 0.2773\,Y_1 + 0.1992\,Y_2\,,$$

sa kanoničkom korelacijom $\sqrt{0.00757} = 0.2751$.

(b) Na osnovu korelacione matrice potrebno je odrediti karakteristične korene matrice $\boldsymbol{\rho}_{11}^{-1}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}$. Kako je

$$\boldsymbol{\rho}_{11}^{-1} = \begin{bmatrix}1.8 & -1.2\\ -1.2 & 1.8\end{bmatrix} \text{ i } \boldsymbol{\rho}_{22}^{-1} = \begin{bmatrix}1.0909 & -0.3149\\ -0.3149 & 1.0909\end{bmatrix}, \text{ to je}$$

$$\boldsymbol{\rho}_{11}^{-1}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21} = \begin{bmatrix}0.9545 & -0.0091\\ -0.5909 & 0.0818\end{bmatrix},$$

a njeni karakteristični koreni su $\lambda_1 = 0.9607$ i $\lambda_2 = 0.0757$. Kao što je rečeno, karakteristični koreni računati na osnovu kovarijacionih i korelacionih matrica su identični. Međutim, karakteristični vektori, odnosno kanonički koeficijenti su različiti. Prvi karakteristični vektor na osnovu $\lambda_1 = 0.9607$ određujemo preko

$$\begin{bmatrix} 0.9545 & -0.0091 \\ -0.5909 & 0.0818 \end{bmatrix} \boldsymbol{\alpha}_1^* = 0.9607 \boldsymbol{\alpha}_1^*,$$

odakle je $\boldsymbol{\alpha}_1^{*\prime} = [1 \quad -0.6724]$, pa na osnovu uslova normalizacije $Var(Z_1) = \boldsymbol{\alpha}' \boldsymbol{\rho}_{11} \boldsymbol{\alpha} = 1$, dobijamo kanoničke koeficijente

$$[1 \quad -0.6724] \begin{bmatrix} 1 & 0.6667 \\ 0.6667 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -0.6724 \end{bmatrix} = 0.5556.$$

Korišćenjem $\sqrt{0.556} = 0.7454$ dobijamo standardizovane kanoničke koeficijente

$$\boldsymbol{\alpha}_1 = \frac{1}{0.7454} \begin{bmatrix} 1 \\ -0.6724 \end{bmatrix} = \begin{bmatrix} 1.3416 \\ -0.9021 \end{bmatrix}.$$

Preko relacije $\boldsymbol{\beta}_1 = \boldsymbol{\rho}_{22}^{-1} \boldsymbol{\rho}_{21} \boldsymbol{\alpha}_1 / \sqrt{\lambda_1}$ dobijamo $\boldsymbol{\beta}_1' = [-0.7934 \quad 0.8794]$. Prvi par standardizovanih kanoničkih promenljivih $\tilde{Z}_1$ i $\tilde{W}_1$ je

$$\tilde{Z}_1 = \boldsymbol{\alpha}_1' \tilde{\mathbf{X}} = 1.3416 \tilde{X}_1 - 0.9021 \tilde{X}_2$$

$$\tilde{W}_1 = \boldsymbol{\beta}_1' \tilde{\mathbf{Y}} = -0.7934 \tilde{Y}_1 + 0.8794 \tilde{Y}_2,$$

a njihova kanonička korelacija je $\sqrt{0.9607} = 0.9801$. Ukoliko računamo vrednosti $\tilde{Z}_1$ i $\tilde{W}_1$, tada na osnovu gornjih relacija koristimo kanoničke koeficijente sa standardizovanim vrednostima $\mathbf{X}$ i $\mathbf{Y}$, u oznaci $\tilde{\mathbf{X}}$ i $\tilde{\mathbf{Y}}$. Pokazati da je na osnovu $\lambda_2 = 0.0757$ drugi par standardizovanih kanoničkih promenljivih

$$\tilde{Z}_2 = 0.0103 \tilde{X}_1 + 0.9931 \tilde{X}_2$$

$$\tilde{W}_2 = 0.6793 \tilde{Y}_1 - 0.5635 \tilde{Y}_2,$$

sa kanoničkom korelacijom $\sqrt{0.0757} = 0.2751$. ■

Karakteristični vektori $\boldsymbol{\alpha}_j$ i $\boldsymbol{\beta}_j$ nazvani kanonički koeficijenti, određuju vrednost kanoničkih promenljivih Z_j i W_j. Kanoničke koeficijente koristimo u intepretaciji samih kanoničkih promenljivih kao i veze između njih. Kako se kanoničke promenljive mogu posmatrati kao regresione funkcije, to se i kanonički koeficijenti mogu interpretirati kao regresioni koeficijenti. Znači da kanonički koeficijenti ukazuju na

parcijalan uticaj odnosne promenljive na kanoničku promenljivu uz pretpostavku da su ostale originalne promenljive u jednačini fiksirane. Nakon što se na ovaj način interpretira svaka kanonička promenljiva Z_j i W_j pristupa se interpretaciji veze između njih. Nešto kasnije iznećemo razloge zašto se ne oslanjamo isključivo na kanoničke koeficijente prilikom interpretacije rezultata.

Interpretacijom rezultata kanoničke korelacione analize bavićemo se detaljnije u poslednjem poglavlju ove glave, a na ovom mestu ukažimo na još jedno svojstvo kanoničkih promenljivih koje nam omogućava njihovu lakšu interpretaciju. Pored korišćenja kanoničkih koeficijenata za vrednovanje relativnog značaja originalnih promenljivih u definisanju kanoničkih promenljivih, izračunavamo i koeficijente korelacije između originalnih i kanoničkih promenljivih. Ove koeficijente korelacije nazivamo *koeficijenti korelacije strukture* ili *kanonička opterećenja* (eng. *canonical loadings*). Ovaj poslednji izraz koristi se u većoj meri u analizi glavnih komponenata i u faktorskoj analizi.

Kako je

$$Cov(Z_j, X_j) = Cov(\boldsymbol{\alpha}_j'\mathbf{X}, X_j) = Cov\left(\sum_{k=1}^{p}\alpha_{ik}X_k, X_j\right) = \sum_{k=1}^{p}\alpha_{ik}Cov(X_k, X_j) \qquad (4.7)$$

to je koeficijent korelacije između Z_j i X_j

$$\rho_{Z_jX_j} = \frac{\sum_{k=1}^{p}\alpha_{ik}Cov(X_k, X_j)}{\sqrt{Var(X_j)}},\ j = 1, 2, ..., p\,. \qquad (4.8)$$

Na sličan način određujemo koeficijente korelacije sledećih parova kanoničkih i originalnih promenljivih: (W_i, Y_j), (W_i, X_j) i (Z_i, Y_j).

Izvedimo u matričnoj notaciji koeficijente korelacije kanoničkih i originalnih promenljivih. U tom cilju označimo sa $\mathbf{A}$ $(p \times p)$ matricu čiji su redovi karakteristični vektori $\boldsymbol{\alpha}_j$, $j = 1, 2, ..., p$, a sa $\mathbf{B}$ $(q \times q)$ matricu čiji su redovi karakteristični vektori $\boldsymbol{\beta}_j$, $j = 1, 2, ..., q$. Sada su vektori kanoničkih promenljivih

$$\underset{(p\times 1)}{\mathbf{Z}} = \mathbf{AX}; \qquad \underset{(q\times 1)}{\mathbf{W}} = \mathbf{BY} \qquad (4.9)$$

a pošto je pretpostavljeno da je $p \le q$, kod vektora kanoničkih promenljivih $\mathbf{W}$ interesuje nas prvih p komponenti. Na osnovu poznate relacije (2.22) imamo

$$Cov(\mathbf{Z}, \mathbf{X}) = Cov(\mathbf{AX}, \mathbf{X}) = \mathbf{A}\boldsymbol{\Sigma}_{11} \qquad (4.10)$$

Kako je na osnovu osobine kanoničkih promenljivih, kovarijaciona matrica vektora kanoničke promenljive $\mathbf{Z}$ jednaka jediničnoj matrici, tj. $Var(\mathbf{Z}) = \mathbf{I}_p$, to je dijagonalna matrica koja sadrži elemente sa glavne dijagonale ove kovarijacione matrice takođe jedinična matrica, tj. $\text{diag}(Var(\mathbf{Z})) = \mathbf{I}_p$. Označimo sa $\mathbf{D}_{11}$ dijagonalnu matricu koja sadrži elemente sa glavne dijagonale kovarijacione matrice $\mathbf{\Sigma}_{11}$ (tj. $\mathbf{D}_{11} = \text{diag}(Var(\mathbf{X}))$). Korelaciona matrica

$$\underset{(q\times q)}{\boldsymbol{\rho}_{\mathbf{ZX}}} = (\text{diag}(Var(\mathbf{Z})))^{-\frac{1}{2}} Cov(\mathbf{Z},\mathbf{X})(\text{diag}(Var(\mathbf{X})))^{-\frac{1}{2}} = \mathbf{A}\mathbf{\Sigma}_{11}\mathbf{D}_{11}^{-\frac{1}{2}} \tag{4.11}$$

ima za elemente koeficijente korelacije između kanoničkih promenljivih $Z_1, Z_2, ..., Z_p$ i originalnih promenljivih $X_1, X_2, ..., X_p$. Na sličan način određujemo korelacione matrice za ostala tri para kanoničkih i originalnih promenljivih

$$\underset{(q\times q)}{\boldsymbol{\rho}_{\mathbf{WY}}} = \mathbf{B}\mathbf{\Sigma}_{22}\mathbf{D}_{22}^{-\frac{1}{2}} \qquad \underset{(q\times p)}{\boldsymbol{\rho}_{\mathbf{WX}}} = \mathbf{B}\mathbf{\Sigma}_{21}\mathbf{D}_{11}^{-\frac{1}{2}} \qquad \underset{(p\times q)}{\boldsymbol{\rho}_{\mathbf{ZY}}} = \mathbf{A}\mathbf{\Sigma}_{12}\mathbf{D}_{22}^{-\frac{1}{2}} \tag{4.12}$$

gde smo sa $\mathbf{D}_{22}$ označili dijagonalnu matricu koja sadrži elemente sa glavne dijagonale kovarijacione matrice $\mathbf{\Sigma}_{22}$. Ukoliko smo pri izračunavanju koristili standardizovane promenljive $\mathbf{X}$ i $\mathbf{Y}$, u oznaci $\tilde{\mathbf{X}}$ i $\tilde{\mathbf{Y}}$, tada se do korelacionih matrica kanoničkih i originalnih promenljivih dolazi na osnovu jednakosti

$$\boldsymbol{\rho}_{\mathbf{ZX}} = \tilde{\mathbf{A}}\boldsymbol{\rho}_{11}; \quad \boldsymbol{\rho}_{\mathbf{ZY}} = \tilde{\mathbf{A}}\boldsymbol{\rho}_{12}; \quad \boldsymbol{\rho}_{\mathbf{WY}} = \tilde{\mathbf{B}}\boldsymbol{\rho}_{22}; \quad \boldsymbol{\rho}_{\mathbf{WX}} = \tilde{\mathbf{B}}\boldsymbol{\rho}_{21}, \tag{4.13}$$

gde smo sa $\tilde{\mathbf{A}}$ i $\tilde{\mathbf{B}}$ označili $(p\times p)$ i $(q\times q)$ matrice čiji su redovi kanonički koeficijenti standardizovanih promenljivih $\tilde{\mathbf{X}}$ i $\tilde{\mathbf{Y}}$ respektivno. Iste oznake koristimo za dobijene korelacione matrice u izrazu (4.13) kao i u izrazu (4.12) jer se standardizacijom promenljivih vrednost koeficijenata korelacije ne menja.

Primer 4.2 Izračunati koeficijente korelacije između kanoničkih i originalnih promenljivih na osnovu podataka iz Primera 4.1.

Na osnovu rezultata iz Primera 4.1 matrice kanoničkih koeficijenata $\mathbf{A}$ i $\mathbf{B}$, kao i dijagonalne matrice $\mathbf{D}_{11}$ i $\mathbf{D}_{22}$ su

$$\mathbf{A} = \begin{bmatrix} 0.7746 & -0.5208 \\ 0.0059 & 0.5734 \end{bmatrix} \text{ i } \mathbf{B} = \begin{bmatrix} -0.3239 & 0.3109 \\ 0.2773 & 0.1992 \end{bmatrix}, \text{ odnosno}$$

$$\mathbf{D}_{11} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \text{ i } \mathbf{D}_{22} = \begin{bmatrix} 6 & 0 \\ 0 & 8 \end{bmatrix}.$$

Na osnovu ovih matrica, korelacione matrice kanoničkih i originalnih promenljivih su

$$\boldsymbol{\rho}_{\mathbf{ZX}} = \mathbf{A}\boldsymbol{\Sigma}_{11}\mathbf{D}_{11}^{-\frac{1}{2}} = \begin{bmatrix} 0.7746 & -0.5208 \\ 0.0059 & 0.5734 \end{bmatrix}\begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}\begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} 0.4272 & -0.0044 \\ 0.3882 & 0.5773 \end{bmatrix},$$

$$\boldsymbol{\rho}_{\mathbf{WX}} = \mathbf{B}\boldsymbol{\Sigma}_{21}\mathbf{D}_{11}^{-\frac{1}{2}} = \begin{bmatrix} -0.3239 & 0.3109 \\ 0.2773 & 0.1992 \end{bmatrix}\begin{bmatrix} 6 & 2 \\ 2 & 8 \end{bmatrix}\begin{bmatrix} \frac{1}{6} & 0 \\ 0 & \frac{1}{8} \end{bmatrix} = \begin{bmatrix} -0.2203 & 0.2299 \\ 0.3437 & 0.2685 \end{bmatrix}.$$

Iz korelacionih matrica čitamo da je $\rho_{Z_1X_1} = 0.4272$ i $\rho_{Z_1X_2} = -0.0044$ što znači da je prva promenljiva iz prvog skupa u većoj meri povezana sa kanoničkom promenljivom Z_1, nego sa Z_2. U drugom skupu obe promenljive su relativno slabo povezane sa kanoničkom promenljivom W_1 pri čemu je ta veza kod Y_1 inverznog tipa ($\rho_{W_1Y_1} = -0.2203$ i $\rho_{W_1Y_2} = 0.2299$). U ovom primeru koeficijenti korelacije pružaju sličnu informaciju kao i standardizovani kanonički koeficijenti.

Dve preostale korelacione matrice su

$$\boldsymbol{\rho}_{\mathbf{WX}} = \mathbf{B}\boldsymbol{\Sigma}_{21}\mathbf{D}_{11}^{-\frac{1}{2}} = \begin{bmatrix} -0.3239 & 0.3109 \\ 0.2773 & 0.1992 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} 0.4189 & -0.0043 \\ 0.1068 & 0.1588 \end{bmatrix}$$

$$\boldsymbol{\rho}_{\mathbf{ZY}} = \mathbf{A}\boldsymbol{\Sigma}_{12}\mathbf{D}_{11}^{-\frac{1}{2}} = \begin{bmatrix} 0.7746 & -0.5208 \\ 0.0059 & 0.5734 \end{bmatrix}\begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{6} & 0 \\ 0 & \frac{1}{8} \end{bmatrix} = \begin{bmatrix} -0.2159 & 0.2254 \\ 0.0946 & 0.0739 \end{bmatrix}$$

Pokazati da se isti rezultat dobija ako se ove korelacione matrice računaju na osnovu ρ_{11}, ρ_{22} i ρ_{12}. ■

4.2.2 Kanonička korelacija, koeficijenti korelacije i regresiona analiza

U Uvodu su ukratko navedeni rezultati koje ćemo u ovom odeljku pokazati. Prema prvom rezultatu kanonička korelacija predstavlja uopštenje koeficijenata korelacije.

Razmotrimo prvo slučaj kada je $p = q = 1$. Tada na osnovu matrica $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ i $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ imamo da je

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \frac{[Cov(X_1, Y_1)]^2}{Var(X_1)Var(Y_1)} = \rho_{X_1Y_1}^2 , \tag{4.14}$$

gde je $\rho_{X_1Y_1}^2$ kvadrat koeficijenta korelacije između X_1 i Y_1. Znači, da kada svaki od skupova promenljivih sadrži po jednu promenljivu, koeficijent kanoničke korelacije jednak je običnom koeficijentu korelacije. Naime, tada se matrice na osnovu kojih određujemo koeficijente kanoničke korelacije svode na skalar. Taj skalar istovremeno predstavlja karakteristični koren, kvadrat koeficijenta korelacije i kvadrat kanoničke korelacije. Pokazano[2] je, da ako je $p, q > 1$, tada je prvi kanonički koeficijent korelacije veći po apsolutnoj vrednosti od svih elemenata korelacione matrice $\boldsymbol{\rho}_{12}$.

Kada prvi skup sadrži p promenljivih, a drugi jednu $(q = 1)$, tada se na osnovu matrice $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ dobija da je

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \frac{\boldsymbol{\sigma}_2'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_2}{\sigma_{22}} = R^2 , \tag{4.15}$$

gde smo, kao i ranije u Poglavlju 3.2, sa σ_{22} označili varijansu od Y_1, sa $\boldsymbol{\sigma}_2$ $(p \times 1)$ vektor kovarijansi između Y_1 i svih promenljivih iz vektora $\mathbf{X}$, a sa R^2 koeficijent determinacije. Znači, da kada skup nezavisnih promenljivih sadrži p promenljivih, a skup zavisnih samo jednu, matrica na osnovu koje određujemo kanoničku korelaciju svodi se na skalar koji je jednak koeficijentu determinacije.

Govoreći o uslovnom rasporedu kod višedimenzione normalne slučajne promenljive rečeno je da uslovna očekivana vrednost definiše multivarijacionu regresionu funkciju. Ako regresiramo $X_1, X_2, ..., X_p$ na $Y_1, Y_2, ..., Y_q$ regresioni koeficijenti dati su $(p \times q)$ matricom $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ (sledi na na osnovu izraza (3.12)). Znači da su regresioni koeficijenti u regresiji svake promenljive (X_i) iz prvog skupa na svih q promenljivih Y_j iz drugog skupa, dati skalarnim proizvodom $i-$reda matrice $\boldsymbol{\Sigma}_{12}$ sa svakom kolonom matrice $\boldsymbol{\Sigma}_{22}^{-1}$.

Kod kanoničke promenljive W_i dobijeno je da je $W_i = \boldsymbol{\beta}_i'\mathbf{Y} = (\boldsymbol{\alpha}_i' / \sqrt{\lambda_i})\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Y}$. Znači da je koeficijent β_{ij} uz Y_j kod kanoničke promenljive W_i, ponderisana suma q regresionih koeficijenata u regresiji svake od promenljivih $X_1, X_2, ..., X_p$, na sve

[2] Videti: Johnson I Wichern (1988, s. 448).

promenljive drugog skupa. Pri tome su ponderi proporcionalni elementima normiranog karakterističnog vektora $\boldsymbol{\alpha}_i$, uz koeficijent proporcionalnosti jednak $(1/\sqrt{\lambda_i})$.

4.3 UZORAČKA KANONIČKA KORELACIONA ANALIZA

Na osnovu slučajnog uzorka od n elemenata svake od $(p+q)$ promenljivih slučajnog vektora $[\mathbf{X}' \vdots \mathbf{Y}']$, formiramo ocenu najveće verodostojnosti kovarijacione matrice $\boldsymbol{\Sigma}$ kao i njenih podmatrica $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22}$ i $\boldsymbol{\Sigma}_{12}$. Zamenom ovih ocena, $\hat{\boldsymbol{\Sigma}}_{11}$, $\hat{\boldsymbol{\Sigma}}_{22}$ i $\hat{\boldsymbol{\Sigma}}_{12}$, dobijamo rešenje $\hat{\lambda}$ uzoračke karakteristične jednačine

$$\left|\hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\Sigma}}_{21} - \hat{\lambda}\mathbf{I}_p\right| = 0 \tag{4.16}$$

To rešenje predstavlja kvadrat uzoračkog koeficijenta kanoničke korelacije. Sve ostale veličine kanoničke korelacione analize definisane u populacionom modelu imaju svoj pandan u uzoračkoj kanoničkoj korelacionoj analizi, a time i identičnu interpretaciju. Umesto ocene kovarijacionih matrica $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22}$ i $\boldsymbol{\Sigma}_{12}$ dobijene metodom najveće verodostojnosti koriste se i njihove nepristrasne ocene, a to su uzoračke kovarijacione matrice $\mathbf{S}$, $\mathbf{S}_{11}$, $\mathbf{S}_{22}$ i $\mathbf{S}_{12}$. Ukoliko se u analizi vrši prethodna standardizacija promenljivih, tada se umesto uzoračkih kovarijacionih matrica koriste uzoračke korelacione matrice $\mathbf{R}$, $\mathbf{R}_{11}$, $\mathbf{R}_{22}$ i $\mathbf{R}_{12}$.

4.3.1 Matrica greške aproksimacije

Kanoničke promenljive reprezentuju dva skupa originalnih promenljivih. Postavlja se pitanje u kojoj meri su one reprezentativne. Ako dobro reprezentuju svaka svoj skup promenljivih, tada na osnovu njih, odnosno koeficijenata kanoničke korelacije, možemo sa većom pouzdanošću izučavati međusobnu povezanost dva skupa promenljivih. Stoga ćemo definisati pokazatelj kojim merimo u kom stepenu kanoničke promenljive obuhvataju varijabilitet u svojim skupovima. U tom cilju izložićemo prethodno postupak konstrukcije *matrice greške aproksimacije*. Ova matrica pokazuje u kojoj meri prvih r $(r \le p \le q)$ uzoračkih kanoničkih promenljivih dobro reprodukuju uzoračke kovarijacione ili korelacione matrice.

Neka su na osnovu slučajnog uzorka formirane matrice $\hat{\mathbf{A}}$ i $\hat{\mathbf{B}}$ koje predstavljaju uzoračke ocene matrica $\mathbf{A}$ i $\mathbf{B}$. Redovi matrica $\hat{\mathbf{A}}$ i $\hat{\mathbf{B}}$ su uzorački kanonički koeficijenti $\hat{\boldsymbol{\alpha}}$ i $\hat{\boldsymbol{\beta}}$. Na osnovu definicije uzoračkog vektora kanoničkih promenljivih: $\hat{\mathbf{Z}} = \hat{\mathbf{A}}\mathbf{X}$ i $\hat{\mathbf{W}} = \hat{\mathbf{B}}\mathbf{Y}$, sledi da je $\mathbf{X} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{Z}}$ i $\mathbf{Y} = \hat{\mathbf{B}}^{-1}\hat{\mathbf{W}}$. Kovarijaciona i korelaciona matrica vektora kanoničkih promenljivih $(\mathbf{Z}, \mathbf{W})$ identične su, pošto im svaki elemenat ima jediničnu varijansu, i date su izrazom

$$\boldsymbol{\Sigma}_{\mathbf{ZW}} = Cov(\mathbf{Z},\mathbf{W}) = \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 & \cdots & 0 & \vdots & \\ 0 & \sqrt{\lambda_2} & 0 & \cdots & 0 & \vdots & \\ 0 & 0 & \sqrt{\lambda_3} & \cdots & 0 & \vdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & \sqrt{\lambda_p} & \vdots & \end{bmatrix}$$

$$= \begin{bmatrix} \rho_1^* & 0 & 0 & \cdots & 0 & \vdots & \\ 0 & \rho_2^* & 0 & \cdots & 0 & \vdots & \\ 0 & 0 & \rho_3^* & \cdots & 0 & \vdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & \rho_p^* & \vdots & \end{bmatrix}, \qquad (4.17)$$

gde smo sa ρ_i^*, $i = 1,2,...,p$, označili koeficijente kanoničke korelacije. Uzoračku kovarijacionu matricu između $\mathbf{X}$ i $\mathbf{Y}$ izražavamo preko uzoračkih kanoničkih koeficijenata i uzoračke kovarijacione matrice kanoničkih promenljivih

$$\mathbf{S}_{12} = \text{Uzoračka } Cov(\mathbf{X},\mathbf{Y}) = Cov(\hat{\mathbf{A}}^{-1}\hat{\mathbf{Z}}, \hat{\mathbf{B}}^{-1}\hat{\mathbf{W}}) = \hat{\mathbf{A}}^{-1}\mathbf{S}_{\mathbf{ZW}}(\hat{\mathbf{B}}^{-1})'. \qquad (4.18)$$

Označimo sa $\hat{\boldsymbol{\alpha}}^{(j)}$ i $\hat{\boldsymbol{\beta}}^{(j)}$ $j-$tu kolonu matrica $\hat{\mathbf{A}}^{-1}$ i $\hat{\mathbf{B}}^{-1}$ respektivno. Izrazimo sada $\mathbf{S}_{12}$ preko uzoračkih koeficijenata kanoničke korelacije $\hat{\rho}_i^*$, $i = 1,2,...,p$

$$\mathbf{S}_{12} = \hat{\mathbf{A}}^{-1}\mathbf{S}_{\mathbf{ZW}}(\hat{\mathbf{B}}^{-1})' = \hat{\rho}_1^*\hat{\boldsymbol{\alpha}}^{(1)}\hat{\boldsymbol{\beta}}^{(1)\prime} + \hat{\rho}_2^*\hat{\boldsymbol{\alpha}}^{(2)}\hat{\boldsymbol{\beta}}^{(2)\prime} + ... + \hat{\rho}_p^*\hat{\boldsymbol{\alpha}}^{(p)}\hat{\boldsymbol{\beta}}^{(p)\prime}. \qquad (4.19)$$

Slično izražavamo $\mathbf{S}_{11}$ i $\mathbf{S}_{22}$ na osnovu uzoračkih kanoničkih koeficijenata

$$\mathbf{S}_{11} = \hat{\mathbf{A}}^{-1}(\hat{\mathbf{A}}^{-1})' = \hat{\boldsymbol{\alpha}}^{(1)}\hat{\boldsymbol{\alpha}}^{(1)\prime} + \hat{\boldsymbol{\alpha}}^{(2)}\hat{\boldsymbol{\alpha}}^{(2)\prime} + ... + \hat{\boldsymbol{\alpha}}^{(p)}\hat{\boldsymbol{\alpha}}^{(p)\prime} \qquad (4.20)$$

$$\mathbf{S}_{22} = \hat{\mathbf{B}}^{-1}(\hat{\mathbf{B}}^{-1})' = \hat{\boldsymbol{\beta}}^{(1)}\hat{\boldsymbol{\beta}}^{(1)\prime} + \hat{\boldsymbol{\beta}}^{(2)}\hat{\boldsymbol{\beta}}^{(2)\prime} + ... + \hat{\boldsymbol{\beta}}^{(p)}\hat{\boldsymbol{\beta}}^{(p)\prime} \qquad (4.21)$$

pošto je $\mathbf{S}_{\mathbf{ZZ}} = \mathbf{I}_p$ i $\mathbf{S}_{\mathbf{WW}} = \mathbf{I}_q$ (prema definiciji - kanonički koeficijenti određeni su uz uslov da kanoničke promenljive imaju jediničnu varijansu).

U analizi najčešće postupamo tako što koristimo prvih r uzoračkih koeficijenata kanoničke korelacije, odnosno kanoničkih promenljivih. Tada su matrice greške aproksimacije

$$\mathbf{S}_{12} - \left[\hat{\rho}_1^*\hat{\boldsymbol{\alpha}}^{(1)}\hat{\boldsymbol{\beta}}^{(1)\prime} + \hat{\rho}_2^*\hat{\boldsymbol{\alpha}}^{(2)}\hat{\boldsymbol{\beta}}^{(2)\prime} + ... + \hat{\rho}_p^*\hat{\boldsymbol{\alpha}}^{(p)}\hat{\boldsymbol{\beta}}^{(p)\prime}\right] = \hat{\rho}_{r+1}^*\hat{\boldsymbol{\alpha}}^{(r+1)}\hat{\boldsymbol{\beta}}^{(r+1)\prime} + ... + \hat{\rho}_p^*\hat{\boldsymbol{\alpha}}^{(p)}\hat{\boldsymbol{\beta}}^{(p)\prime} \qquad (4.22)$$

$$\mathbf{S}_{11} - \left[\hat{\boldsymbol{\alpha}}^{(1)}\hat{\boldsymbol{\alpha}}^{(1)\prime} + \hat{\boldsymbol{\alpha}}^{(2)}\hat{\boldsymbol{\alpha}}^{(2)\prime} + \ldots + \hat{\boldsymbol{\alpha}}^{(r)}\hat{\boldsymbol{\alpha}}^{(r)\prime} \right] = \hat{\boldsymbol{\alpha}}^{(r+1)}\hat{\boldsymbol{\alpha}}^{(r+1)\prime} + \ldots + \hat{\boldsymbol{\alpha}}^{(p)}\hat{\boldsymbol{\alpha}}^{(p)\prime} \tag{4.23}$$

$$\mathbf{S}_{22} - \left[\hat{\boldsymbol{\beta}}^{(1)}\hat{\boldsymbol{\beta}}^{(1)\prime} + \hat{\boldsymbol{\beta}}^{(2)}\hat{\boldsymbol{\beta}}^{(2)\prime} + \ldots + \hat{\boldsymbol{\beta}}^{(r)}\hat{\boldsymbol{\beta}}^{(r)\prime} \right] = \hat{\boldsymbol{\beta}}^{(r+1)}\hat{\boldsymbol{\beta}}^{(r+1)\prime} + \ldots + \hat{\boldsymbol{\beta}}^{(p)}\hat{\boldsymbol{\beta}}^{(p)\prime} \tag{4.25}$$

Elementi ovih matrica pokazuju koliko su uspešne prve r kanoničke promenljive u reprodukovanju kovarijacionih matrica. S obzirom na način i uslove pod kojima su konstruisane kanoničke promenljive, ne treba da čudi što prvih r kanoničkih promenljivih bolje reprodukuju matricu $\mathbf{S}_{12}$ nego $\mathbf{S}_{11}$ i $\mathbf{S}_{22}$. Elementi matrice greške aproksimacije biće relativno mali u slučaju pošto je matrica greške aproksimacije zasnovana na poslednjih $p-r$ uzoračkih koeficijenata kanoničke korelacije, koji su po definiciji najmanji. Matrice greške aproksimacije $\mathbf{S}_{11}$ i $\mathbf{S}_{22}$ zavise samo od kanoničkih koeficijenata, pa se stoga može desiti slučaj da zbog velikih vrednosti elemenata kod poslednjih vektora koeficijenata, njihovi elementi budu relativno veliki. Ukoliko smo u analizi pošli od korelacione matrice, odnosno od standardizovanih opservacija, tada se kovarijacione matrice u gornjem izvođenju zamenjuju odgovarajućim korelacionim matricama. Istovremeno se i matrice kanoničkih koeficijenata zamenjuju odgovarajućim matricama standardizovanih kanoničkih koeficijenata. Napominjemo da kod ovih potonjih koristimo istu oznaku za matrice kanoničkih koeficijenata kao i u slučaju nestandardizovanih kanoničkih promenljivih.

Primer 4.3 Povećajmo uzorak u Primeru 2.5, gde smo na osnovu podataka iz *Ankete o potrošnji domaćinstava* Zavoda za statistiku posmatrali šest poljoprivrednih domaćinstava. Sada ćemo u analizu uključiti i uzorak od 10 nepoljoprivrednih i 8 mešovitih domaćinstava. Kod svakog domaćinstva merili smo sledeće promenljive: $X_1 =$ broj članova domaćinstva, $X_2 =$ godišnja raspoloživa sredstva, $X_3 =$ izdaci na ishranu i $X_4 =$ izdaci na obrazovanje i razonodu. Podaci su prezentirani u narednoj tabeli. Radi celovitosti reprodukovani su podaci iz Primera 2.5 za poljoprivredna domaćinstva.

a) Odrediti uzoračku kovarijacionu i korelacionu matricu.

b) Ako je $\mathbf{X}' = [X_1, X_2]$ i $\mathbf{Y}' = [X_3, X_4]$ odrediti koeficijente kanoničke korelacije između $\mathbf{X}$ i $\mathbf{Y}$ kao i parove standardizovanih kanoničkih promenljivih.

c) Odrediti koeficijente korelacije strukture, odnosno koeficijente korelacije originalnih i kanoničkih promenljivih.

d) Korišćenjem samo prvog para standardizovanih kanoničkih promenljivih odrediti u kom stepenu on efikasno reprodukuje korelacione matrice $\mathbf{R}_{12}$, $\mathbf{R}_{11}$ i $\mathbf{R}_{22}$.

Redni broj domaćinstva	X_1	X_2	X_3	X_4	Tip domaćinstva
1	6	9.40	6.20	0.20	poljoprivredno
2	7	12.10	7.20	0.10	poljoprivredno
3	4	10.70	3.10	0.22	poljoprivredno
4	5	9.30	5.40	0.12	poljoprivredno
5	2	11.20	2.10	0.26	poljoprivredno
6	3	13.30	1.80	0.30	poljoprivredno
7	2	17.0	3.70	0.90	nepoljoprivredno
8	3	16.0	3.40	0.88	nepoljoprivredno
9	5	13.0	6.80	0.41	nepoljoprivredno
10	6	10.0	6.70	0.39	nepoljoprivredno
11	3	12.0	3.80	0.62	nepoljoprivredno
12	2	14.0	4.20	0.71	nepoljoprivredno
13	4	13.5	4.50	0.55	nepoljoprivredno
14	3	15.2	5.20	0.95	nepoljoprivredno
15	3	16.1	3.50	1.20	nepoljoprivredno
16	4	13.0	3.70	0.50	nepoljoprivredno
17	7	9.90	5.50	0.14	mešovito
18	5	12.50	4.30	0.53	mešovito
19	2	16.10	3.70	0.58	mešovito
20	4	10.30	5.30	0.25	mešovito
21	3	16.30	4.30	0.53	mešovito
22	4	13.40	4.40	0.51	mešovito
23	2	12.20	4.60	0.31	mešovito
24	3	11.80	5.20	0.16	mešovito

a) Uzoračka kovarijaciona i korelaciona matrica prikazani su u narednoj tabeli. Koeficijenti korelacije dati su u donjem levom trouglu, a kovarijaciona matrica u gornjem desnom trouglu zajedno sa dijagonalom.

Promenljiva	Kovarijaciona i korelaciona matrica			
	X_1	X_2	X_3	X_4
Broj članova domaćinstva	2.4058	-2.2051	1.5087	-0.23101
Godišnja raspoloživa sredstva	-0.6072	5.4826	-1.2425	0.25605
Izdaci na ishranu	0.7069	-0.3857	1.8933	-0.13130
Izdaci na obrazovanje i razonodu	-0.5075	0.8091	-0.3251	0.08615

b) Standardizovani kanonički koeficijenti i kanonička korelacija dati su tabelom:

Standardizovane promenljive			$\hat{\rho}_j^*$	Standardizovane promenljive		
Vektor	X_1	X_2		Vektor	X_3	X_4
$\hat{\boldsymbol{\alpha}}_1'$	0.3719	-0.7295	0.8376	$\hat{\boldsymbol{\beta}}_1'$	0.3885	-0.8038
$\hat{\boldsymbol{\alpha}}_2'$	1.2023	1.0255	0.5979	$\hat{\boldsymbol{\beta}}_2'$	0.9835	0.6872

Dakle, prvi par standardizovanih uzoračkih kanoničkih promenljivih je

$$\tilde{Z}_1 = 0.3719\tilde{X}_1 - 0.7295\tilde{X}_2$$

$$\tilde{W}_1 = 0.3885\tilde{X}_3 - 0.8038\tilde{X}_4,$$

gde smo sa $\tilde{X}_i$, $i = 1,2,3,4$, označili standardizovane vrednosti originalnih promenljivih. Uzorački koeficijent kanoničke korelacije za prvi par uzoračkih kanoničkih promenljivih iznosi 0.8376. S obzirom na vrednost kanoničkih koeficijenata prve kanoničke promenljive, zaključujemo da ta kanonička promenljiva meri razliku između broja članova domaćinstva i godišnje raspoloživih sredstava. Napominjemo da se radi o standardizovanim promenljivama. Vrednost prve kanoničke promenljive raste ako se dvostruko brže povećava vrednost prve u odnosu na vrednost druge promenljive. Slično je u slučaju druge standardizovane kanoničke promenljive koja meri razliku između izdataka na ishranu i izdataka na obrazovanje i razonodu. Domaćinstva koja pokazuju veću razliku između standardizovanih vrednosti broja članova domaćinstva i raspoloživih sredstava pokazivaće tendenciju ka većim razlikama između standardizovanih vrednosti izdataka na ishranu i izdataka na obrazovanje i razonodu.

c) Koeficijenti korelacije strukture, odnosno koeficijenti korelacije između originalnih promenljivih i kanoničkih dati su narednom tabelom.

Originalne promenljive	Kanoničke promenljive			
	Z_1	W_1	Z_2	W_2
Prva grupa				
X_1	0.8149	0.6825	0.5797	0.3466
X_2	-0.9553	-0.8002	0.2955	0.1767
Druga grupa				
X_3	0.5443	0.6498	0.4544	0.7601
X_4	-0.7790	-0.9301	0.2196	0.3674

Promenljive iz prvog skupa imaju viši koeficijent korelacije sa prvom kanoničkom promenljivom Z_1, a promenljive iz drugog skupa sa drugom kanoničkom promenljivom prvog para, odnosno W_1.

d) Kako je $\hat{\mathbf{A}} = \begin{bmatrix} 0.3719 & -0.7295 \\ 1.2023 & 1.0255 \end{bmatrix}$ i $\hat{\mathbf{B}} = \begin{bmatrix} 0.3885 & -0.8038 \\ 0.9835 & 0.6872 \end{bmatrix}$ to je

$$\hat{\mathbf{A}}^{-1} = \begin{bmatrix} 0.8149 & 0.5797 \\ -0.9554 & 0.2955 \end{bmatrix} \text{ i } \hat{\mathbf{B}}^{-1} = \begin{bmatrix} 0.6498 & 0.7601 \\ -0.9300 & 0.3674 \end{bmatrix}.$$

Na osnovu ovih matrica, matrice greške aproksimacije dobijene korišćenjem samo prvog para kanoničkih promenljivih su

za $\mathbf{R}_{12}$:

$$\hat{\rho}_2^* \hat{\boldsymbol{\alpha}}^{(2)} \hat{\boldsymbol{\beta}}^{(2)\prime} = 0.5979 \begin{bmatrix} 0.5797 \\ 0.2955 \end{bmatrix} \begin{bmatrix} 0.7601 & 0.3674 \end{bmatrix} = \begin{bmatrix} 0.2634 & 0.1273 \\ 0.1343 & 0.0649 \end{bmatrix},$$

za $\mathbf{R}_{11}$

$$\hat{\boldsymbol{\alpha}}^{(2)} \hat{\boldsymbol{\alpha}}^{(2)\prime} = \begin{bmatrix} 0.5797 \\ 0.2955 \end{bmatrix} \begin{bmatrix} 0.5797 & 0.2955 \end{bmatrix} = \begin{bmatrix} 0.3360 & 0.1713 \\ 0.1713 & 0.0873 \end{bmatrix},$$

za $\mathbf{R}_{22}$

$$\hat{\boldsymbol{\beta}}^{(2)} \hat{\boldsymbol{\beta}}^{(2)\prime} = \begin{bmatrix} 0.7601 \\ 0.3674 \end{bmatrix} \begin{bmatrix} 0.7601 & 0.3674 \end{bmatrix} = \begin{bmatrix} 0.5777 & 0.2792 \\ 0.2792 & 0.1350 \end{bmatrix}.$$

Na osnovu matrica greške aproksimacije konstatujemo da prvi par kanoničkih promenljivih relativno slabo reprodukuje koeficijente korelacije iz podmatrice $\mathbf{R}_{12}$. Kada je reč o podmatricama $\mathbf{R}_{11}$ i $\mathbf{R}_{22}$ kod njih su greške aproksimacije još veće nego kod $\mathbf{R}_{12}$, što je očekivan rezultat. ■

4.3.2 Proporcija objašnjene varijanse

Visoka vrednost koeficijenta kanoničke korelacije dobijena u analizi predstavlja, matematičkim rečnikom rečeno, potreban, ali ne i dovoljan uslov za smislenu interpretaciju rezultata. Može se desiti slučaj da su samo jedna ili nekolicina promenljivih iz jednog skupa povezani sa jednom ili nekolicinom promenljivih drugog skupa. Tada se dobija da je koeficijent korelacije strukture te promenljive relativno visok, ali je mali iznos varijanse skupa zavisnih promenljivih objašnjen kanoničkom

promenljivom. Zato je kod interpretacije rezultata analize potrebno koristiti pokazatelj *proporcije objašnjene varijanse* za svaki skup promenljivih. On ukazuje na iznos varijanse originalnih promenljivih objašnjen kanoničkom promenljivom.

Neka su originalne promenljive standardizovane, drugim rečima, analizu zasnivamo na korelacionim matricama. Tada su koeficijenti korelacije kanoničkih i originalnih promenljivih date elementima matrice $\hat{\mathbf{A}}^{-1}$ i $\hat{\mathbf{B}}^{-1}$, pošto su uzoračke kovarijacione matrice standardizovanih promenljivih i kanoničkih promenljivih ili, što je u ovom slučaju isto, korelacione matrice

$$Cov(\tilde{\mathbf{X}}, \hat{\mathbf{Z}}) = Cov(\hat{\mathbf{A}}^{-1}\hat{\mathbf{Z}}, \hat{\mathbf{Z}}) = \hat{\mathbf{A}}^{-1} \tag{4.26}$$

$$Cov(\tilde{\mathbf{Y}}, \hat{\mathbf{W}}) = Cov(\hat{\mathbf{B}}^{-1}\hat{\mathbf{W}}, \hat{\mathbf{W}}) = \hat{\mathbf{B}}^{-1} \tag{4.27}$$

Elementi matrice $\hat{\mathbf{A}}^{-1}$ su uzorački koeficijenti korelacije između elemenata vektora uzoračke kanoničke promenljive $\hat{\mathbf{Z}}$ i promenljivih iz prvog skupa promenljivih, u oznaci $\hat{\rho}_{Z_i X_j}$. Slično imamo za matricu $\hat{\mathbf{B}}^{-1}$; njeni elementi su uzorački koeficijenti korelacije između elemenata vektora uzoračke kanoničke promenljive $\hat{\mathbf{W}}$ i promenljivih iz drugog skupa, u oznaci $\hat{\rho}_{W_i Y_j}$.

Koristimo relacije uspostavljene između uzoračkih kovarijacionih matrica $\mathbf{S}_{11}$ i matrice $\hat{\mathbf{A}}^{-1}$ (izraz (4.20)) i matrice $\hat{\mathbf{B}}^{-1}$ (izraz (4.21)), kao i drugu generalizaciju varijanse (prema kojoj je generalizovana varijansa jednaka tragu kovarijacione ili korelacione matrice). Međutim, sada umesto matrica $\mathbf{S}_{11}$ i $\mathbf{S}_{22}$ imamo $\mathbf{R}_{11}$ i $\mathbf{R}_{22}$ respektivno. Tako dobijamo

$$\mathrm{tr}(\mathbf{R}_{11}) = \mathrm{tr}\left(\hat{\boldsymbol{\alpha}}^{(1)}\hat{\boldsymbol{\alpha}}^{(1)\prime} + \hat{\boldsymbol{\alpha}}^{(2)}\hat{\boldsymbol{\alpha}}^{(2)\prime} + \ldots + \hat{\boldsymbol{\alpha}}^{(p)}\hat{\boldsymbol{\alpha}}^{(p)\prime}\right) = p \tag{4.28}$$

što znači da je ukupna standardizovana uzoračka varijansa prvog skupa promenljivih jednaka veličini tog skupa (broju promenljivih). Slično se dobija

$$\mathrm{tr}(\mathbf{R}_{22}) = \mathrm{tr}\left(\hat{\boldsymbol{\beta}}^{(1)}\hat{\boldsymbol{\beta}}^{(1)\prime} + \hat{\boldsymbol{\beta}}^{(2)}\hat{\boldsymbol{\beta}}^{(2)\prime} + \ldots + \hat{\boldsymbol{\beta}}^{(q)}\hat{\boldsymbol{\beta}}^{(q)\prime}\right) = q \tag{4.29}$$

što znači da je ukupna standardizovana uzoračka varijansa drugog skupa promenljivih jednaka veličini tog skupa (broju promenljivih).

Ukoliko smo u analizi koristili umesto p, svega r kanoničkih promenljivih $(r \le p)$, tada se doprinos prvih r kanoničkih promenljivih $Z_1, Z_2, \ldots, Z_r$ ukupnoj standardizovanoj uzoračkoj varijansi prvog skupa promenljivih definiše kao

$$\mathrm{tr}\left(\hat{\boldsymbol{\alpha}}^{(1)}\hat{\boldsymbol{\alpha}}^{(1)\prime} + \hat{\boldsymbol{\alpha}}^{(2)}\hat{\boldsymbol{\alpha}}^{(2)\prime} + \ldots + \hat{\boldsymbol{\alpha}}^{(r)}\hat{\boldsymbol{\alpha}}^{(r)\prime}\right) = \sum_{i=1}^{r}\sum_{j=1}^{p}\hat{\rho}^2_{Z_i, X_j} \tag{4.30}$$

a doprinos prvih r kanoničkih promenljivih $W_1, W_2, ..., W_r$ ukupnoj standardizovanoj uzoračkoj varijansi drugog skupa promenljivih definiše kao

$$\text{tr}\left(\hat{\boldsymbol{\beta}}^{(1)}\hat{\boldsymbol{\beta}}^{(1)\prime} + \hat{\boldsymbol{\beta}}^{(2)}\hat{\boldsymbol{\beta}}^{(2)\prime} + ... + \hat{\boldsymbol{\beta}}^{(r)}\hat{\boldsymbol{\beta}}^{(r)\prime}\right) = \sum_{i=1}^{r}\sum_{j=1}^{q}\hat{\rho}^2_{W_i,Y_j} \tag{4.31}$$

Proporcija varijanse prvog skupa promenljivih objašnjena sa prvih r kanoničkih promenljivih, u oznaci $R^2_{\mathbf{Z\cdot X}}$, dobija se stavljanjem u odnos veličina iz izraza (4.30) i (4.28):

$$R^2_{\mathbf{Z\cdot X}} = \frac{\text{tr}\left(\hat{\boldsymbol{\alpha}}^{(1)}\hat{\boldsymbol{\alpha}}^{(1)\prime} + \hat{\boldsymbol{\alpha}}^{(2)}\hat{\boldsymbol{\alpha}}^{(2)\prime} + ... + \hat{\boldsymbol{\alpha}}^{(r)}\hat{\boldsymbol{\alpha}}^{(r)\prime}\right)}{\text{tr}(\mathbf{R}_{11})} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{p}\hat{\rho}^2_{Z_i,X_j}}{p} \tag{4.32}$$

Proporcija varijanse prvog skupa promenljivih objašnjena sa prvih r kanoničkih promenljivih, u oznaci $R^2_{\mathbf{W\cdot Y}}$, dobija se stavljanjem u odnos veličina iz izraza (4.31) i (4.29):

$$R^2_{\mathbf{W\cdot Y}} = \frac{\text{tr}\left(\hat{\boldsymbol{\beta}}^{(1)}\hat{\boldsymbol{\beta}}^{(1)\prime} + \hat{\boldsymbol{\beta}}^{(2)}\hat{\boldsymbol{\beta}}^{(2)\prime} + ... + \hat{\boldsymbol{\beta}}^{(r)}\hat{\boldsymbol{\beta}}^{(r)\prime}\right)}{\text{tr}(\mathbf{R}_{22})} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{q}\hat{\rho}^2_{W_i,Y_j}}{q} \tag{4.33}$$

Pokazatelji $R^2_{\mathbf{Z\cdot X}}$ i $R^2_{\mathbf{W\cdot Y}}$ ukazuju na to u kojoj su meri kanoničke promenljive dobri predstavnici svojih skupova. Ne slučajno, njihovom oznakom asociramo na njihovu interpretaciju kao koeficijenata determinacije.

Primer 4.4 Na osnovu podataka iz Primera 4.3 odrediti proporciju varijanse prvog, odnosno drugog skupa promenljivih koji je objašnjen prvim parom kanoničkih promenljivih.

U Primeru 4.3 izračunali smo

$$\hat{\mathbf{A}}^{-1} = \begin{bmatrix} 0.8149 & 0.5797 \\ -0.9554 & 0.2955 \end{bmatrix} \text{ i } \hat{\mathbf{B}}^{-1} = \begin{bmatrix} 0.6498 & 0.7601 \\ -0.9300 & 0.3674 \end{bmatrix},$$

odakle čitamo, na primer $\hat{\rho}^2_{Z_1,X_1} = 0.8149$ i $\hat{\rho}^2_{W_1,Y_2} = -0.93$. Ako je $r = 1$ tada

$$R^2_{\mathbf{Z\cdot X}} = \frac{1}{2}\sum_{j=1}^{2}\hat{\rho}^2_{Z_1,X_j} = \frac{1}{2}[(0.8149)^2 + (-0.9554)^2] = 0.7884,$$

$$R^2_{\mathbf{W}\cdot\mathbf{Y}} = \frac{1}{2}\sum_{j=1}^{2} \hat{\rho}^2_{W_1,Y_j} = \frac{1}{2}[(0.6498)^2 + (-0.9300)^2] = 0.6436$$

Znači da uzoračka kanonička promenljiva $\hat{Z}_1$ "objašnjava" 78.84% ukupne uzoračke varijanse prvog skupa promenljivih, a uzoračka kanonička promenljiva $\hat{W}_1$ "objašnjava" 64.36% ukupne uzoračke varijanse drugog skupa promenljivih. Prema tome, uzoračka kanonička promenljiva $\hat{Z}_1$ bolje reprezentuje svoj (prvi) skup promenljivih, nego što je to slučaj sa $\hat{W}_1$ za njen (drugi) skup promenljivih.

Kako je $r = 1,\ p = q = 2$, to je neobjašnjeni deo varijanse prvog i drugog skupa promenljivih:

$$1 - R^2_{\mathbf{Z}\cdot\mathbf{X}} = 0.2116 \text{ i } 1 - R^2_{\mathbf{W}\cdot\mathbf{Y}} = 0.3564$$

S obzirom da postoje samo dva para kanoničkih promenljivih, to znači da se ovaj neobjašnjeni deo varijanse dva skupa promenljivih pripisuje drugom paru uzoračkih kanoničkih promenljivih $(\hat{Z}_2, \hat{W}_2)$. Kažemo da je 21.16% varijanse prvog i 35.64% varijanse drugog skupa "objašnjeno" uzoračkim kanoničkim promenljivama $\hat{Z}_2$ i $\hat{W}_2$ respektivno. ■

S obzirom da su pokazatelji reprezentativnosti kanoničkih promenljivih zasnovani na koeficijentima korelacije strukture, sugerišemo jednostavniji način njihovog određivanja. Elemente uzoračke matrice korelacije strukture $\mathbf{R_{ZX}}$ i $\mathbf{R_{WY}}$ kvadriramo, pa je $(i, j)-$ti element dobijene matrice proporcija varijanse $i-$te originalne promenljive objašnjene $j-$tom kanoničkom promenljivom. Na primer, kod matrice $\mathbf{R_{ZX}}$ zbir kvadrata elemenata $j-$te kolone podeljen sa p (veličina prvog skupa), daje proporciju ukupne varijanse prvog skupa objašnjene sa $\hat{Z}_j$. Sabiranjem dobijenih proporcija za recimo prvih r kanoničkih promenljivih $Z_1, Z_2, ..., Z_r$, dobićemo pokazatelj $R^2_{\mathbf{Z}\cdot\mathbf{X}}$. Slično bismo došli, na osnovu matrice $\mathbf{R_{WY}}$, do sumarnog pokazatelja reprezentativnosti uzoračkih kanoničkih promenljivih , tj. do pokazatelja $R^2_{\mathbf{W}\cdot\mathbf{Y}}$.

4.3.3 Mera redundantnosti

Rečeno je da postoji mogućnost dobijanja relativno visokog koeficijenta kanoničke korelacije $\sqrt{\lambda_j}$, a da istovremeno bude mala proporcija objašnjene varijanse svakog od dva skupa promenljivih njihovim kanoničkim promenljivama. U tom slučaju nije preporučljivo previše se pouzdati u dobijene rezultate analize. Imamo dakle, s jedne strane, da karakteristični koren λ_j sugeriše visoku vrednost proporcije varijanse Z_j

objašnjene sa W_j, a s druge strane, da $R^2_{Z_j \cdot \mathbf{X}}$ ukazuje na malu vrednost proporcije varijanse u prvom skupu promenljivih objašnjene kanoničkom promenljivom Z_j. Stoga su Stewart i Love (1968) predložili korišćenje *koeficijenta redundantnosti*, koji istovremeno uzima u obzir, kako visinu koeficijenta kanoničke korelacije, tako i proporciju varijanse svakog skupa promenljivih objašnjene njihovim kanoničkim promenljivama. Taj koeficijent meri proporciju varijanse prvog skupa promenljivih objašnjene drugom kanoničkom promenljivom W_j. Definisan je kao proizvod $\lambda_j R^2_{Z_j \cdot \mathbf{X}}$ i ukazuje na kvalitet W_j kao prediktora prvog skupa promenljivih $\mathbf{X}$. Na sličan način definišemo koeficijent redundantnosti $\lambda_j R^2_{W_j \cdot \mathbf{Y}}$ koji daje proporciju varijanse drugog skupa promenljivih $\mathbf{Y}$, objašnjene kanoničkom promenljivom Z_j.

Na osnovu pojedinačnih koeficijenata formiramo *koeficijente ukupne redundantnosti* kao zbir individualnih koeficijenata redundantnosti:

$$\sum_{j=1}^{p} \lambda_j R^2_{Z_j \cdot \mathbf{X}} \qquad \text{i} \qquad \sum_{j=1}^{p} \lambda_j R^2_{W_j \cdot \mathbf{Y}} \tag{4.34}$$

Prvi koeficijent ukupne redundantnosti ukazuje u kom stepenu sve promenljive iz drugog skupa (preko kanoničkih promenljivih W_j) "objašnjavaju" varijansu promenljivih prvog skupa. Slično važi za drugi koeficijent ukupne redundantnosti. On meri u kom stepenu sve promenljive prvog skupa (preko kanoničkih promenljivih Z_j) dobro "objašnjavaju" varijansu promenljivih drugog skupa. Ova dva pokazatelja mogu se međusobom znatno razlikovati. Naime, ako prvi skup promenljivih dobro "objašnjava" varijansu promenljivih drugog skupa, prvi koeficijent ukupne redundantnosti biće visok, ali obrnuto ne mora da važi, pa drugi koeficijent ukupne redundantnosti može biti nizak.

Do koeficijenata ukupne redundantnosti možemo doći i na osnovu koeficijenata korelacije kanoničkih promenljivih jednog skupa i originalnih promenljivih drugog skupa. Ovi koeficijenti korelacije nazivaju se i *unakrsna opterećenja* (eng. cross-loadings). Ako za svaku matricu posebno, saberemo kvadrate elemenata uzoračkih matrica unakrsnih opterećenja ($\hat{\boldsymbol{\rho}}_{\mathbf{WX}}$ i $\hat{\boldsymbol{\rho}}_{\mathbf{ZY}}$), a zatim ih podelimo sa brojem promenljivih, dobićemo gornja dva koeficijenta ukupne redundantnosti. Pored navedene uloge unakrsnih opterećenja ona se koriste i prilikom interpretacije rezultata kanoničke analize.

Koeficijent redundantnosti analogan je koeficijentu determinacije R^2. Stewart i Love su pokazali da se koeficijent ukupne redundantnosti može dobiti tako što će se regresirati svaka promenljiva iz skupa zavisnih promenljivih na sve promenljive iz skupa nezavisnih promenljivih, a zatim uprosečiti p dobijenih koeficijenata determinacije.

Primer 4.5 Na osnovu podataka iz Primera 4.3 i 4.4 sprovesti analizu redundantnosti.

U Primeru 4.3 odredili smo koeficijente kanoničke korelacije: $\sqrt{\lambda_1} = 0.8376$ i $\sqrt{\lambda_2} = 0.5979$, a u Primeru 4.4 dobijene su proporcije objašnjene varijanse:

$$R^2_{Z_1 \cdot \mathbf{X}} = 0.7884,\ R^2_{Z_2 \cdot \mathbf{X}} = 0.2116,\ R^2_{W_1 \cdot \mathbf{Y}} = 0.6436 \text{ i } R^2_{W_2 \cdot \mathbf{Y}} = 0.3564.$$

Za kanoničke promenljive W_1 i W_2 koeficijenti redundantnosti su:

$$\lambda_1 R^2_{Z_1 \cdot \mathbf{X}} = 0.7016 \cdot 0.7884 = 0.5531 \text{ i } \lambda_2 R^2_{Z_2 \cdot \mathbf{X}} = 0.3575 \cdot 0.2116 = 0.0756,$$

a koeficijent ukupne redundantnosti je $0.5531 + 0.0756 = 0.6288$. Znači da je varijansa prvog skupa promenljivih objašnjena u iznosu od 62.88% kanoničkim promenljivama drugog skupa.

Za kanoničke promenljive Z_1 i Z_2 koeficijenti redundantnosti su:

$$\lambda_1 R^2_{W_1 \cdot \mathbf{Y}} = 0.7016 \cdot 0.6436 = 0.4515 \text{ i } \lambda_2 R^2_{W_2 \cdot \mathbf{Y}} = 0.3575 \cdot 0.3564 = 0.1274,$$

a koeficijent ukupne redundantnosti je $0.4515 + 0.1274 = 0.5789$. Znači da je varijansa drugog skupa promenljivih objašnjena u iznosu od 57.89% kanoničkim promenljivama prvog skupa. Analiza redundantnosti je pokazala da kanoničke promenljive drugog skupa W_1 i W_2, objašnjavaju nešto veći iznos varijanse prvog skupa promenljivih no što je to slučaj sa objašnjenjem varijanse drugog skupa promenljivih kanoničkim promenljivama prvog skupa.

Pokazati da se isti rezultat dobija ako se koeficijenti ukupne redundantnosti računaju preko unakrsnih opterećenja ili kao proseci koeficijenata determinacije. ■

Izložene mere redundantnosti predstavljale su osnovu za razvoj alternativne vrste kanoničke analize zasnovane na maksimiziranju koeficijenta redundantnosti. Takav pristup se izdvojio u poseban metod multivarijacione analize pod nazivom *analiza redundantnosti*. U osnovi ona se bavi izdvajanjem nezavisnih promenljivih iz skupa nezavisnih promenljivih tako što će se maksimizirati objašnjena varijansa promenljivih iz skupa zavisnih promenljivih.

4.3.4 Testovi kanoničke korelacije

Pretpostavili smo da su $\mathbf{X}$ i $\mathbf{Y}$ raspoređeni po višedimenzionom normalnom rasporedu. Tada je nezavisnost dva skupa promenljivih ekvivalentna uslovu $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, što

sa svoje strane implicira da su svi koeficijenti kanoničke korelacije jednaki nuli. Za test nulte hipoteze $H_0 : \mathbf{\Sigma}_{12} = \mathbf{0}$, protiv hipoteze $H_1 : \mathbf{\Sigma}_{12} \neq \mathbf{0}$, ili, što je ekvivalentno

$$H_0 : \rho_1^* = \rho_2^* = ... = \rho_p^* = 0, \quad H_1 : \rho_i^* \neq 0, \text{ bar za jedno } i, \tag{4.35}$$

konstruišemo statistiku testa na osnovu principa količnika verodostojnosti. Nultu hipotezu odbacujemo za velike vrednosti statistike testa:

$$-2 \ln \Lambda = n \ln \left(\frac{|\mathbf{S}_{11}||\mathbf{S}_{22}|}{|\mathbf{S}|} \right) = n \ln \prod_{i=1}^{p} (1 - \hat{\rho}_i^{*2}) \tag{4.36}$$

Za velike uzorke ova statistika testa ima aproksimativno $\chi^2 -$raspored sa pq stepeni slobode. Bartlett (1939) je sugerisao njenu modifikaciju u tom smislu što će se koeficijent n zameniti sa $n - 1 - \frac{1}{2}(p+q+1)$. Time se poboljšava navedena aproksimacija $\chi^2 -$rasporeda. Interesantno je istaći da statistika testa bazirana na principu količnika verodostojnosti poredi uzoračku generalizovanu varijansu kada je tačna nulta hipoteza, odnosno uz ograničenje da je $\mathbf{S}_{12} = \mathbf{0}$, sa generalizovanom varijansom bez ovog ograničenja. Na nivou značajnosti α odbacujemo nultu hipotezu ako je na osnovu realizovanih vrednosti slučajnog uzorka

$$-\left(n - 1 - \frac{1}{2}(p+q+1) \right) \ln \prod_{i=1}^{p} (1 - \hat{\rho}_i^{*2}) > \chi^2_{pq;\alpha} \tag{4.37}$$

gde je $\chi^2_{pq;\alpha}$ gornji 100α percentil $\chi^2 -$rasporeda sa pq stepeni slobode. Podsećamo da smo koeficijente kanoničke korelacije poređali po veličini. Ukoliko odbacimo nultu hipotezu da su svi koeficijenti kanoničke korelacije jednaki nuli, tada možemo pretpostaviti da je prvi koeficijent kanoničke korelacije različit od nule, a da su ostali koeficijenti jednaki nuli. Ako se ova hipoteza odbaci tada pretpostavljamo da su prva dva koeficijenta kanoničke korelacije različita od nule, a ostalih $p-2$ koeficijenata jednaki nuli. Ovaj način definisanja niza nultih hipoteza možemo dalje nastaviti. U slučaju $k-$tog koeficijenta kanoničke korelacije nulta i alternativna hipoteza glase:

$$\begin{aligned} &H_0^{(k)} : \rho_1^* \neq 0, \rho_2^* \neq 0, ..., \rho_k^* \neq 0, \rho_{k+1}^* = ... = \rho_p^* = 0, \\ &H_1^{(k)} : \rho_i^* = 0, \quad \text{za neko } i \geq k+1. \end{aligned} \tag{4.38}$$

Statistika testa, za test navedenog niza hipoteza, sugeriše da se na nivou značajnosti odbaci nulta hipoteza ukoliko je na osnovu realizovanih vrednosti slučajnog uzorka

$$-\left(n - 1 - \frac{1}{2}(p+q+1) \right) \ln \prod_{i=k+1}^{p} (1 - \hat{\rho}_i^{*2}) > \chi^2_{(p-k)(q-k);\alpha} \tag{4.39}$$

gde je $\chi^2_{(p-k)(q-k);\alpha}$ gornji 100α percentile χ^2-rasporeda sa $(p-k)(q-k)$ stepeni slobode. Ukoliko se test navedenog niza hipoteza sprovodi tako što se testira jedna po jedna hipoteza sve dok se $H_0^{(k)}$ ne odbaci za neko k, tada opšti nivo značajnosti testa nije α i teško ga je odrediti.

Primer 4.6 Na osnovu podataka iz Primera 4.3 testirati statističku značajnost koeficijenata kanoničke korelacije na nivou značajnosti $\alpha = 0.05$.

Izračunali smo $\hat{\rho}_1^* = 0.8376$ i $\hat{\rho}_2^* = 0.5979$. Kako je $p = q = 2$, $n = 24$, imamo da je $n-1-\frac{1}{2}(p+q+1) = 20.5$. Kritična vrednost χ^2-statistike je $\chi^2_{4;0.05} = 9.49$, a izračunata je

$$\left(n-1-\frac{1}{2}(p+q+1) = 20.5\right)\ln(1-\hat{\rho}_1^{*2})(1-\hat{\rho}_2^{*2}) = 20.5\cdot\ln(0.1918) = 33.8577.$$

Kako je izračunata vrednost statistike testa veća od kritične, na nivou značajnosti od 5% zaključujemo da postoji bar jedan koeficijent kanoničke korelacije statistički značajno različit od nule.

Sada definišemo hipotezu: $H_0^{(1)}: \rho_1^* \neq 0,\ \rho_2^* = 0$, a potrebni elementi za izračunavanje vrednosti statistike testa i test nulte hipoteze su: $(p-1)(q-1) = 1$, $\chi^2_{1;0.05} = 3.84$, $n-1-\frac{1}{2}(p+q+1) = 20.5$.

Izračunata vrednost statistike testa je $-20.5\cdot\ln(1-0.5979^2) = 9.0658$. Kako je izračunata vrednost statistike testa veća od kritične, na nivou značajnosti od 5% odbacujemo nultu hipotezu, pa zaključujemo da je i drugi koeficijent kanoničke korelacije statistički značajno različit od nule.

Proveriti da se do izračunate vrednosti test statistike može doći i na osnovu determinanti kovarijacionih, odnosno korelacionih matrica (generalizovanih varijansi). U tu svrhu prezentiramo sve međurezultate potrebne za proračun: $|\mathbf{S}| = 0.233$, $|\mathbf{S}_{11}| = 8.3277$, $|\mathbf{S}_{22}| = 0.1459$. ■

Pored iznete ograda koja se tiču korišćenja Bartlettovog testa za testiranje značajnosti koeficijenata kanoničke korelacije napominjemo da je potrebno koristiti i druge kriterijume da bismo doneli konačan sud koliko koeficijenata kanoničke korelacije zadržati u analizi. O tome nešto više u narednom poglavlju.

4.4 INTERPRETACIJA KANONIČKIH PROMENLJIVIH I KANONIČKE KORELACIJE

Pre nego što se pristupi interpretaciji dobijenih rezultata kanoničke korelacione analize potrebno je odgovoriti na pitanje koliko parova kanoničkih funkcija, odnosno kanoničkih promenljivih zadržati u analizi. U prethodnom odeljku izloženi su formalni statistički postupci određivanja značajnosti koeficijenata kanoničke korelacije. Kao što je to u statistici uobičajeno u analizi se zadržavaju oni koeficijenti koji su iznad izabranog nivoa značajnosti. Međutim, pored ovog formalnog kriterijuma preporučuje se korišćenje i drugih kriterijuma koji nam olakšavaju donošenje odluke o broju kanoničkih promenljivih u analizi.

Koeficijent kanoničke korelacije može se pokazati statistički značajno različitim od nule, ali istovremeno mora biti ispunjen i uslov da se utvrđena veza dva skupa promenljivih može smisleno da interpretira. U suprotnom, samo uslov statističke značajnosti koeficijenta nije dovoljan da bismo dalje analizirali tu vezu. Slično postupamo ukoliko kanoničke promenljive ne obuhvataju određeni iznos varijanse svoga skupa originalnih promenljivih, ili nam to sugeriše analiza redundantnosti. Znači da se pored rezultata testiranja značajnosti koeficijenata kanoničke korelacije u procesu donošenja odluke oslanjamo i na veličinu tog koeficijenta kao i na mere redundantnosti. Na žalost, ne postoji jasno definisana veličina koeficijenta kanoničke korelacije ispod koje ih ne bi trebalo uzeti u razmatranje. Ta granica je uglavnom uslovljena specifičnom prirodom problema koji se razmatra. Ranije je ukazano da, s obzirom na definiciju koeficijenta kanoničke korelacije (maksimalna korelacija između dva skupa promenljivih), možemo pogrešno zaključiti o postojanju značajne veze između dva skupa čak i kada izdvojene kanoničke promenljive ne izdvajaju značajan deo varijanse iz njihovih respektivnih skupova. Zato se sugeriše korišćenje mera redundantnosti koje ukazuju na deo varijanse zavisnih promenljivih objašnjen nezavisnim promenljivama.

Ukoliko je kanonička korelaciona veza statistički značajna, sa relativno visokom vrednošću koeficijenta kanoničke korelacije i prihvatljivim pokazateljem redundantnosti, dakle ako smo kombinacijom navedenih kriterijuma doneli odluku o broju parova kanoničkih promenljivih u analizi, pristupamo interpretaciji dobijenih rezultata.

U fazi interpretacije rezultata kanoničke korelacione analize oslanjamo se na izračunate kanoničke koeficijente (pondere), koeficijente korelacije strukture (kanonička opterećenja) i unakrsna opterećenja. Predznak i veličina kanoničkih koeficijenata, slično koeficijentima višestrukog regresionog modela, ukazuju na doprinos originalne promenljive svojoj kanoničkoj promenljivoj. Međutim, kao u slučaju višestrukog regresionog modela ovi koeficijenti mogu biti izuzetno nestabilni zbog problema multikolinearnosti. Tako neka varijabla može dobiti mali kanonički koeficijent ili čak sa suprotnim predznakom zato što je varijansa promenljive već obuhvaćena nekom drugom promenljivom ili promenljivama. Zato se preporučuje podela uzorka i sprovođenje odvojenih analiza u cilju ispitivanja stabilnosti izračunatih kanoničkih koeficijenata. Ukoliko su koeficijenti stabilni tada bi koeficijent korelacije između istih kanoničkih

promenljivih iz ova dva poduzorka trebalo da bude relativno visok. Dodatan problem je što, za razliku od promenljivih kod višestrukog regresionog modela, kanoničke promenljive nisu opservabilne. U tom smislu je metod kanoničke analize bliži metodu faktorske analize nego regresionom modelu.

Poslednjih godina koeficijenti korelacije strukture preuzimaju primat kanoničkim koeficijentima u interpretaciji rezultata analize. Koeficijenti korelacije strukture mere linearnu zavisnost između originalnih promenljivih svakog skupa i njihovih kanoničkih promenljivih. Znači da koeficijenti korelacije strukture mere iznos varijanse koju originalne promenljive dele sa kanoničkom promenljivom, pa se na osnovu njih može utvrditi relativan doprinos svake promenljive svojim kanoničkim promenljivama.

Naposletku, interpretaciju međusobne povezanosti dva skupa promenljivih možemo dati i na osnovu unakrsnih opterećenja ili koeficijenata korelacije originalnih promenljivih jednog skupa i kanoničkih promenljivih drugog skupa. Na taj način direktno merimo povezanost zavisnih i nezavisnih promenljivih. Ovaj pristup je u najvećoj meri prikladan prirodi problema koji izučavamo metodom kanoničke korelacione analize.

*

* *

Na kraju ove glave ukažimo na neka ograničenja metode kanoničke korelacione analize, koje treba imati na umu pri izbora ove metode ili prilikom interpretacije njenih rezultata.

1. Koeficijent kanoničke korelacije ukazuje na deo varijanse koji međusobom *dele* kanoničke promenljive dva skupa promenljivih, a ne na varijansu koja je izdvojena iz svakog skupa promenljivih. Da bi se dobili smisleni rezultati potrebno je utvrditi da li je relativno velik iznos varijanse u skupu zavisnih promenljivih koji se objašnjen skupom nezavisnih promenljivih.
2. Potrebno je ispitati visinu koeficijenata korelacije između originalnih promenljivih dva skupa. Visoka vrednost jednog koeficijenta korelacije zavisne i nezavisne promenljive može proizrokovati lažno visoku vrednost koeficijenta kanoničke korelacije.
3. Kanonički koeficijenti mogu biti veoma nestabilni. Stoga je potrebno proveriti njihovu stabilnost na način kako je to sugerisano.
4. Kanonički koeficijenti su izvedeni tako da je maksimizirana korelacija između dve linearne kombinacije, a ne objašnjena varijansa svakog od skupova njihovim kanoničkim promenljivama. Zato je potrebno proveriti visinu proporcije objašnjene varijanse svakog skupa svojim kanoničkim promenljivama.

5. U fazi interpretacije rezultata imamo poteškoća da identifikujemo smislenu vezu između skupova promenljivih usled nedostatka odgovarajućeg statističkog pokazatelja koji bi se koristio u te svrhe. Mada su relativno nepogodni za korišćenje u te svrhe koristimo se koeficijentima korelacije strukture i unakrsnim opterećenjima.

5 MULTIVARIJACIONA ANALIZA VARIJANSE (M A N O V A)

Problem testiranja hipoteze o vrednosti sredine razmatran je u 3. Glavi. Tamo izložen statistički postupak testiranja na osnovu slučajnog uzorka višedimenzione slučajne promenljive iz jedne populacije predstavlja uopštenje jednodimenzionog $t-$testa. Za poređenje sredina nekoliko populacija koristi se jedna od najpoznatijih statističkih metoda poznata pod nazivom analiza varijanse (u oznaci ANOVA). Za testiranje jednakosti sredine višedimenzione slučajne promenljive kod nekoliko populacija koristi se višedimenziono uopštenje ove metode, odnosno *multivarijaciona analiza varijanse* (u oznaci MANOVA).

Pre nego što postupnosti radi, izložimo postupke poređenja sredina vektora za slučaj dve populacije, u uvodnom poglavlju ukazaćemo na situacije i pitanja na koja metodi iz ove glave daju odgovore. Smeštajući problem poređenja sredina populacija u širi kontekst planiranja eksperimenta razlikovaćemo dva slučaja. Prvi, kada se koristi *potpuno slučajan plan* za generisanje uzorka iz više populacija i drugi kada su iste jedinice posmatranja podvrgnute različitim eksperimentalnim situacijama. U ovom drugom slučaju reč je o tzv. *planu ponovljenih merenja*. Kod poređenja sredina više populacija razmatramo dva osnovna modela analize. Prvi, kod koga ispitujemo dejstvo jednog faktora na varijabilitet podatka, tzv. model MANOVA sa jednim faktorom i drugi, kod koga ispitujemo dejstvo dva faktora na varijabilitet podatka, tzv. model MANOVA sa dva faktora. Posebno poglavlje posvećujemo tzv. *analizi profila* u okviru koga postupno, preko niza hipoteza, odgovoramo na pitanje o jednakosti sredina nekoliko populacija.

5.1 UVOD

Metodi poređenja sredina u bliskoj su vezi sa planiranjem eksperimenta kod koga direktno kontrolišemo jednu ili više nezavisnih promenljivih (najčešće su u pitanju

kvalitativne promenljive) da bismo odredili njene efekte na jednu ili više zavisnih promenljivih. Najpoznatiji plan eksperimenta jeste *potpuno slučajan plan* kod koga nezavisna promenljiva (u planiranju eksperimenta koristimo izraz - *faktor*) uzima dve vrednosti (nazivamo ih *tretmani* ili *nivoi* faktora). Kod njega ispitujemo uticaj jednog faktora na zavisnu promenljivu, a na potpuno slučajan način primenjujemo jedan ili drugi tretman (nivo faktora) na eksperimentalne jedinice, što znači da na svaku jedinicu posmatranja možemo primeniti jedan ili drugi tretman sa podjednakom verovatnoćom. Sredina zavisne promenljive kod eksperimentalnih jedinica podvrgnutih prvom ili drugom tretmanu predstavlja polaznu osnovu za ispitivanje uticaja posmatranog faktora na zavisnu promenljivu. Primer istraživanja na koje se odnosi navedeni plan eksperimenta je ispitivanje značajnosti razlike u prodaji nekog proizvoda s obzirom na tip prodavnice (samoposluga i klasična prodavnica). U ovom primeru faktor identifikujemo sa tipom prodavnice, njegov prvi tretman sa prvim tipom prodavnice - samoposlugom, a drugi sa drugim tipom prodavnice - klasičnom prodavnicom. Zavisna promenljiva predstavlja prodaju posmatranog proizvoda. Ako slučajno izaberemo prodavnica oba tipa (ne nužno podjednak broj), odnosno "primenimo tretmane" na eksperimentalne jedinice, dobićemo dva uzorka na osnovu kojih donosimo odluku u pogledu hipoteze da tip prodavnice nema uticaja na prodaju proizvoda. Operacionalizacija ove hipoteze sugeriše jednakost prosečne prodaje u ova dva tipa prodavnica. Na osnovu primene $t-$testa donosimo odluku o prihvatanju ili odbacivanju navedene hipoteze.

Pored prodaje proizvoda možemo uzeti u obzir i druge elemente u vezi posmatranog proizvoda. Na primer, opšta zainteresovanost ili pažnja koju kupci poklanjaju tom proizvodu s obzirom da ona može rezultirati u većoj prodaji u narednom periodu. Sada se nalazimo u situaciji da zavisna promenljiva nije jednodimenziona već višedimenziona. U takvom slučaju individualni pristup $t-$testom svakom paru zavisnih promenljivih za dva tretmana zanemaruje korelaciju između zavisnih promenljivih i ne dozvoljava kontrolu greške prve vrste kao što smo to pokazali u Odeljku 3.4.3 o simultanim intervalima poverenja. Za ovaj plan eksperimenta kod koga ispitujemo uticaj faktora sa dva nivoa na višedimenzionu zavisnu promenljivu u Poglavlju 5.2 izlažemo statistički postupak koja omogućava zaključivanje o jednakosti sredine dve populacije.

Proširujući naš primer pretpostavićemo da pored prodaje u samoposlugama i klasičnim prodavnicama posmatrani proizvod kupci mogu naći i u specijalizovanim radnjama. Uz ostale elemente koje smo ranije definisali u ovom primeru, možemo identifikovati plan eksperimenta koji ima karakteristike potpuno slučajnog plana sa nezavisnom promenljivom koja uzima tri vrednosti. Drugim rečima, kažemo da ispitivani faktor (tip prodavnice) ima tri nivoa ili tretmana. Statistički postupak ispitivanja odsustva uticaja faktora na prodaju proizvoda u ovom proširenom primeru zasnovan je na metodu analize varijanse, a ukoliko pored prodaje proizvoda uzimamo u obzir i druge elemente u vezi proizvoda, na primer zainteresovanost kupaca, tada analizu ove višedimenzione zavisne promenljive za tri tretmana vršimo multivarijacionom analizom varijanse. Model koji predstavlja opis izloženog primera

naziva se model multivarijacione analize varijanse sa jednim faktorom i njime se bavimo u Poglavlju 5.4.

Ako u razmatranje uključimo regione u kojima su locirane prodavnice kao dodatni faktor koji utiče na prodaju, pristup sa metodološkog stanovišta ostaje isti (multivarijaciona analiza varijanse) kao i plan eksperimenta (potpuno slučajan plan). Ono što dodatno opterećuje analizu jeste eventualno prisustvo interakcije između faktora. Ovaj primer u kome ispitujemo dejstvo dva faktora, tip prodavnice i region, na višedimenzionu zavisnu promenljivu teorijski modeliramo u Poglavlju 5.6, gde se izlaže model multivarijacione analize varijanse sa dva faktora.

Zajednička nit koja se provlači kroz sve navedene primere jeste identičan plan eksperimenta, reč je o potpuno slučajnom planu eksperimenta. Međutim, modeli multivarijacione analize varijanse koje izlažemo u ovoj glavi primereni su i drugim planovima eksperimenata, ali se o njima može pročitati u literaturi posvećenoj planiranju eksperimenata, npr. Box i dr. (1978). Pored potpuno slučajnog plana eksperimenta pozabavićemo se analizom još jednog plana eksperimenta, planom ponovljenih merenja.

Za razliku od potpuno slučajnog plana eksperimenta kod *plana ponovljenih merenja* nalazimo se u eksperimentalnoj situaciji da opservacije prikupljamo u različito vreme na istim eksperimentalnim jedinicima. Ovaj plan eksperimenta javlja se u situaciji kada, na primer jedna grupa ispitanika rešava određeni problem pre i nakon dodatne obuke i pri tome se registruje vreme rešavanja problema. Tada se razlika u zabeleženim vremenima može pripisati upravo efektima te dodatne obuke. Za poređenje parova odziva koristimo $t-$test, ako je zavisna promenljiva jednodimenziona, odnosno T^2-test, ako je zavisna promenljiva višedimenziona. Statistički postupak poređenja parova izlažemo u Odeljku 5.2.2. Međutim, puni smisao plana ponovljenih merenja sagledava se u eksperimentalnoj situaciji u kojoj beležimo jednu promenljivu odziva na *istoj* eksperimentalnoj jedinici u *različitim* vremenskim trenucima, tj. nakon primene jednog po jednog tretmana u uzastopnim vremenskim periodima.

Primer u kome se javlja ovaj plan eksperimenta odnosi se na istraživanje uticaja tipa pakovanja (porodično i individualno) i vrste ambalaže (brik pakovanje i plastična flaša) na prodaju osvežavajućeg pića. U izabranim prodavnicama u sukcesivnim mesecima beležimo prodaju osvežavajućeg pića, pri čemu na slučaj u jednom mesecu primenjujemo kombinaciju određenog tipa pakovanja i vrste ambalaže. Dakle, u istim prodavnicama (eksperimentalnim jedinicama) beležimo prodaju osvežavajućeg pića (jedna promenljiva odziva) koja varira iz meseca u mesec pod dejstvom uslova prodaje (tretmana). Ovaj primer koristimo kao ilustraciju u Poglavlju 5.3 u kome se izlažu teorijske osnove statističkog postupka analize plana ponovljenih merenja.

5.2 SLUČAJ DVE POPULACIJE

5.2.1 Potpuno slučajan plan

Pretpostavimo da smo na potpuno slučajan način pridružili eksperimentalne jedinice skupu eksperimentalnih uslova (opisanih kao prvi i drugi tretman). Na taj način smo u izvesnoj meri eliminisali uticaj koji na rezultat imaju varijacije od jedne do druge eksperimentalne jedinice. Kao elemente prve populacije posmatraćemo odzive eksperimentalnih jedinica na prvi tretman, a kao elemente druge populacije odzive eksperimentalnih jedinica na drugi tretman. U slučaju da je odziv eksperimentalnih jedinica iz ove dve populacije jednodimenziona promenljiva, primenom $t-$testa možemo testirati hipotezu o jednakosti sredina ove dve populacije. U višedimenzionom slučaju, po analogiji sa $t-$testom, definišemo T^2-test za test hipoteze da su sredine dve višedimenzione populacije jednake.

Neka je $\mathbf{X}_{11},\mathbf{X}_{12},...,\mathbf{X}_{1n_1}$ slučajan uzorak od n_1 elemenata iz $N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$ populacije, a $\mathbf{X}_{21},\mathbf{X}_{22},...,\mathbf{X}_{2n_2}$ slučajan uzorak od n_2 elemenata iz $N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$ populacije i neka su elementi ova dva uzorka nezavisni međusobom. Uzoračke pokazatelji su:

za prvi uzorak: $\bar{\mathbf{X}}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1i}$, $\mathbf{S}_1=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(\mathbf{X}_{1i}-\bar{\mathbf{X}}_1)(\mathbf{X}_{1i}-\bar{\mathbf{X}}_1)'$

za drugi uzorak: $\bar{\mathbf{X}}_2=\frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_{2i}$, $\mathbf{S}_2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(\mathbf{X}_{2i}-\bar{\mathbf{X}}_2)(\mathbf{X}_{2i}-\bar{\mathbf{X}}_2)'$

Kako su uzorci nezavisni, a $\bar{\mathbf{X}}_1\sim N_p(\boldsymbol{\mu}_1,\frac{1}{n_1}\boldsymbol{\Sigma})$ i $\bar{\mathbf{X}}_2\sim N_p(\boldsymbol{\mu}_2,\frac{1}{n_2}\boldsymbol{\Sigma})$, njihova razlika ima takođe normalan raspored, odnosno $\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2\sim N_p(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2,\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\boldsymbol{\Sigma})$, pa se testiranje hipoteze $H_0:\boldsymbol{\mu}_1-\boldsymbol{\mu}_2=\mathbf{0}$ može sprovesti korišćenjem sličnog postupka zasnovanog na T^2-testu kao kod testa hipoteze $H_0:\boldsymbol{\mu}_1-\boldsymbol{\mu}_2=\mathbf{0}$. Može se pokazati[1] da je uz navedene pretpostavke

$$T^2=\left[\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]'\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\bar{\mathbf{S}}\right]^{-1}\left[\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right] \qquad (5.1)$$

raspoređeno kao $\frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)}F_{p,n_1+n_2-p-1}$ gde je F_{p,n_1+n_2-p-1} $F-$raspored sa p i n_1+n_2-p-1 stepeni slobode, a $\bar{\mathbf{S}}=\frac{(n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2}{n_1+n_2-2}$ je ocena uzoračke kovarijacione

[1] Videti na primer Johnson I Wichern (1982, ss. 239-240).

matrice združenih uzoraka. Kada uzorke iz obe populacije smatramo jednim, koristimo naziv združeni uzorci (ili ceo uzorak).

Ako su populacione sredine $\boldsymbol{\mu}_1$ i $\boldsymbol{\mu}_2$ jednake međusobom, tada se T^2-statistika svodi na

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\bar{\mathbf{S}}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \tag{5.2}$$

Na nivou značajnosti α odbacujemo hipotezu $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$, ako je

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\bar{\mathbf{S}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) > \frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1;\alpha} \tag{5.3}$$

gde je $F_{p,n_1+n_2-p-1;\alpha}$ gornji $100(1-\alpha)$ percentil $F-$rasporeda sa p i n_1+n_2-p-1 stepeni slobode.

Oblast poverenja za razliku populacionih sredina $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, konstruišemo na osnovu navedenog T^2-rasporeda. Tako je $100(1-\alpha)\%$ oblast poverenja definisana svim onim vektorima $\boldsymbol{\delta}$ koji zadovoljavaju nejednakost

$$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta})'\bar{\mathbf{S}}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta}) \le \frac{n_1 n_2}{n_1+n_2}\frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1;\alpha} \tag{5.4}$$

Ukoliko se odbaci hipoteza $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ na osnovu $100(1-\alpha)\%$ simultanih intervala poverenja za individualne razlike, možemo utvrditi koja je od njih dovela do odbacivanja nulte hipoteze. Za *sve* linearne kombinacije $\mathbf{a}'\boldsymbol{\delta}$, gde je $\mathbf{a}(p\times 1)$ vektor konstanti, sa nivoom poverenja od $100(1-\alpha)\%$ kažemo da pripadaju intervalu poverenja koji je dat izrazom

$$\mathbf{a}'(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \pm \sqrt{\mathbf{a}'\bar{\mathbf{S}}\mathbf{a}}\sqrt{\frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1;\alpha}}\ , \tag{5.5}$$

odakle dobijamo $100(1-\alpha)\%$ simultani interval poverenja za $j-$ti element vektora $\boldsymbol{\delta}$:

$$(\bar{X}_{1j} - \bar{X}_{2j}) \pm \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\bar{S}_{jj}^2}\sqrt{\frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1;\alpha}}\ , \tag{5.6}$$

gde je $\bar{X}_{1j} - \bar{X}_{2j}$ $j-$ti element vektora $\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2$ i $\bar{S}_{jj}^2$ je $j-$ti dijagonalni element kovarijacione matrice združenih uzoraka.

Primer 5.1 Prodaja gaziranog i negaziranog osvežavajućeg pića beleži se u dva regiona. Na osnovu slučajnog uzorka od $n_1 = 40$ prodavnica iz prvog regiona i $n_2 = 50$ prodavnica iz drugog regiona dobijeni su uzorački pokazatelji

$$\bar{\mathbf{x}}_1' = \begin{bmatrix} 28 & 28 \end{bmatrix},\ \bar{\mathbf{x}}_2' = \begin{bmatrix} 25 & 30 \end{bmatrix},\ \mathbf{S}_1 = \begin{bmatrix} 10 & 5 \\ 5 & 12 \end{bmatrix} \text{ i } \mathbf{S}_2 = \begin{bmatrix} 10 & 4 \\ 4 & 8 \end{bmatrix}.$$

Da li postoji razlika u prodaji osvežavajućih pića u ova dva regiona? Odrediti 95% oblast poverenja za razliku prosečnih prodaja u ova dva regiona. Ako postoji razlika u prodaji osvežavajućih pića između ova dva regiona, kako se ona ispoljava. Objasniti dobijene rezultate.

Uzoračka razlika sredina je $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' = \begin{bmatrix} 3 & -2 \end{bmatrix}$, a kovarijaciona matrica združenih uzoraka je

$$\bar{\mathbf{S}} = \frac{(40-1)\mathbf{S}_1 + (50-1)\mathbf{S}_2}{(40+50-2)} = \begin{bmatrix} 10 & 4.4431 \\ 4.4431 & 9.7727 \end{bmatrix}.$$

Vrednost $T^2 -$ statistike je

$$T^2 = \frac{40 \cdot 50}{40+50} \begin{bmatrix} 3 & -2 \end{bmatrix} \begin{bmatrix} 0.12532 & -0.05697 \\ -0.05697 & 0.12823 \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix} = 51.6543.$$

Kritična vrednost testa na nivou značajnosti $\alpha = 0.05$ iznosi

$$\frac{(40+50-2)2}{(40+50-2-1)} F_{2,87;0.05} = \frac{176}{87} 3.1013 = 6.2739.$$

Kako je $T^2 = 51.6543 > 6.2739$ hipotezu H_0 odbacujemo na nivou značajnosti od 5%. Zaključujemo da postoji razlika u prodaji dva osvežavajuća pića s obzirom na region.

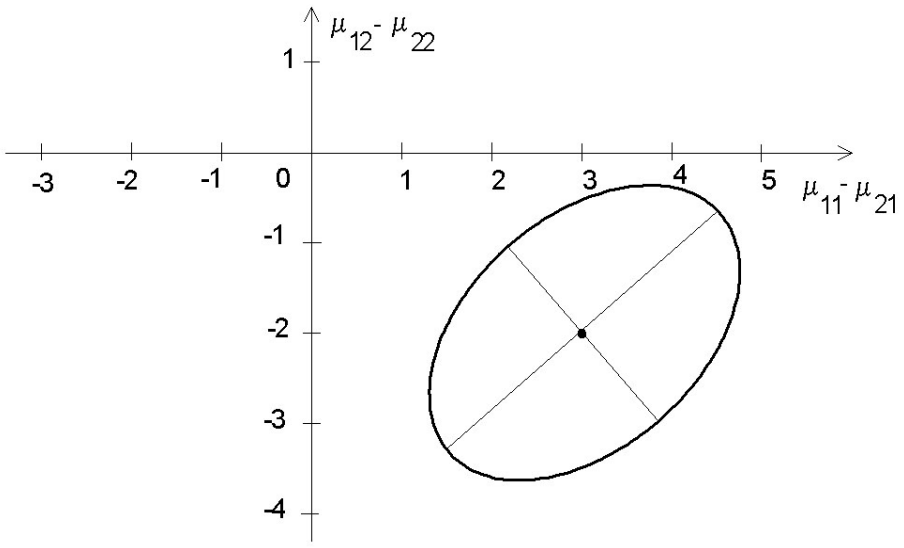
Da bismo formirali 95% oblast poverenja za $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, odredićemo karakteristične korene i vektore kovarijacione matrice, $\bar{\mathbf{S}}$.

$$0 = \left| \bar{\mathbf{S}} - \lambda \mathbf{I} \right| = \begin{vmatrix} 10-\lambda & 4.4431 \\ 4.4431 & 9.7727-\lambda \end{vmatrix} = \lambda^2 - 19.7727\lambda + 77.9854,$$

pa je $\lambda_1 = 14.331$ i $\lambda_2 = 5.442$, a odgovarajući karakteristični vektori su $\mathbf{e}_1' = \begin{bmatrix} -0.716 & -0.698 \end{bmatrix}$ i $\mathbf{e}_2' = \begin{bmatrix} 0.698 & -0.716 \end{bmatrix}$. Ose elipse leže na pravcima određenim karakterističnim vektorima, a dužina su im jednake

$$2\sqrt{\lambda_j}\sqrt{\frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1;\alpha}},\ j = 1,2.$$

Na osnovu ovoga sledi da je dužina prve ose elipse 4.02, a druge 2.48. 95% oblast poverenja prikazana je na donjoj slici.



Očigledno je da tačka $(0,0)$ ne pripada 95% oblasti poverenja, pa je prodaja osvežavajućih pića u ova dva regiona statistički značajno različita. Za izračunavanje 95% simultanih intervala poverenja $(\mu_{11}-\mu_{21})$ i $(\mu_{12}-\mu_{22})$ koristimo ranije određene elemente:

$$\overline{x}_{11}-\overline{x}_{21}=3\,,\ \overline{x}_{12}-\overline{x}_{22}=-2\,,\ \overline{s}_{11}^2=10 \text{ i } \overline{s}_{22}^2=9.7727\,.$$

Na osnovu njih 95% simultani intervali poverenja su:

$$(\mu_{11}-\mu_{21})\in(1.3197,\ 4.6803) \text{ i } (\mu_{12}-\mu_{22})\in(-3.6611,\ -0.3389)\,.$$

Odbacivanju nulte hipoteze doprinele su razlike u prodaji između dva regiona kako za gazirana tako i za negazirana osvežavajuća pića. S tim što je razlika u prodaji gaziranih pića između dva regiona više izražena. ■

5.2.2 Poređenje parova

Drugi, direktan način kontrole varijacija od jedne do druge eksperimentalne jedinice, za razliku od pristupa kome je u osnovi potpuno slučajan plan, je pridruživanje prvog, odnosno drugog tretmana *istoj* ili *sličnoj* eksperimentalnoj jedinici. Posmatranjem razlike dva odziva eksperimentalnih jedinica otkloniće se u većoj meri uticaj varijacija među jedinicama posmatranja.

U društvenim naukama često se suočavamo sa planom eksperimenta koga karakteriše to da na istoj jedinici posmatranja (eksperimentalnoj jedinici) merimo vrednost zavisne promenljive pre i nakon primene tretmana. Na primer, za istu grupu potrošača beležimo prodaju nekog proizvoda pre i posle propagandne kampanje koja se odnosi na taj proizvod. Razlika u prodaji proizvoda pripisuje se efektima propagandne kampanje.

Pored prodaje proizvoda možemo beležiti i druge reakcije kupaca na propagandnu kampanju, pa zavisna promenljiva može biti višedimenziona.

Pretpostavimo da je zavisna promenljiva $p-$dimenziona, tj. beležimo p odziva $i-$te eksperimentalne jedinice na dejstvo prvog ili drugog tretmana. Označimo te odzive po tretmanima na sledeći način:

Prvi tretman		Drugi tretman	Razlika parova
X_{11i}	prva promenljiva	X_{21i}	$D_{1i} = X_{11i} - X_{21i}$
X_{12i}	druga promenljiva	X_{22i}	$D_{2i} = X_{12i} - X_{22i}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{1pi}	$p-$ta promenljiva	X_{2pi}	$D_{pi} = X_{1pi} - X_{2pi}$

U poslednjoj koloni prikazane su razlike parova slučajnih promenljivih. Ako za $\mathbf{D}_i' = [D_{1i}, D_{2i}, ..., D_{pi}]$, $i = 1, 2, ..., n$, pretpostavimo da je

$$E(\mathbf{D}_i) = \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_p \end{bmatrix} \text{ i } Cov(\mathbf{D}_i) = \boldsymbol{\Sigma}_d , \tag{5.7}$$

pri čemu su $\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_n$ elementi slučajnog uzorka iz $N_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_d)$ populacije, tada zaključivanje o sredini $\boldsymbol{\delta}$ možemo bazirati na T^2-statistici. Naime, statistika

$$T^2 = n(\bar{\mathbf{D}} - \boldsymbol{\delta})' \mathbf{S}_d^{-1} (\bar{\mathbf{D}} - \boldsymbol{\delta}) \tag{5.8}$$

raspoređena je kao $[(n-1)p/(n-p)]F_{p,n-p}$ za ma koje $\boldsymbol{\delta}$ i $\boldsymbol{\Sigma}_d$, gde je

$$\bar{\mathbf{D}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{D}_i \text{ i } \mathbf{S}_d = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{D}_i - \bar{\mathbf{D}})(\mathbf{D}_i - \bar{\mathbf{D}})' \tag{5.9}$$

a $F_{p,n-p}$ je $F-$raspored sa p i $n-p$ stepeni slobode. Pored ovog egzaktnog rasporeda T^2-statistike, možemo koristiti i rezultat za veliko n i $n-p$ prema kome T^2-statistika ima aproksimativno χ^2-raspored sa p stepeni slobode, bez obzira na raspored razlika parova.

Raspored statistike (5.8) koristimo da bismo testirali odsustvo razlika u uticaju prvog, odnosno drugog tretmana na višedimenzionu zavisnu promenljivu. Odsustvo razlika iskazujemo preko nulte hipoteze $H_0 : \boldsymbol{\delta} = \mathbf{0}$, protiv hipoteze $H_1 : \boldsymbol{\delta} \neq \mathbf{0}$. Na osnovu realizovanih vrednosti razlika parova $\mathbf{d}_i' = [d_{1i}, d_{2i}, ..., d_{pi}]$, $i = 1, 2, ..., n$ testiramo

navedenu hipotezu na nivou značajnosti α i donosimo odluku da H_0 treba odbaciti ako je

$$T^2 = n\bar{\mathbf{d}}'\mathbf{S}_d^{-1}\bar{\mathbf{d}} > \frac{(n-1)p}{(n-p)}F_{p,n-p;\alpha} \tag{5.10}$$

gde je $F_{p,n-p;\alpha}$ gornji 100α percentil $F-$rasporeda sa p i $n-p$ stepeni slobode. Koristeći navedeni rezultat o rasporedu T^2-statistike definišemo oblast poverenja za vektor razlike parova $\boldsymbol{\delta}$, koji sa $100(1-\alpha)\%$ nivoom poverenja sadrži sve razlike parova promenljivih $\boldsymbol{\delta}$:

$$(\bar{\mathbf{d}}-\boldsymbol{\delta})'\mathbf{S}_d^{-1}(\bar{\mathbf{d}}-\boldsymbol{\delta}) \le \frac{(n-1)p}{(n-p)}F_{p,n-p;\alpha} \tag{5.11}$$

Naposletku, da bismo identifikovali koji je od elemenata vektora razlike parova doprineo odbacivanju nulte hipoteze, formiramo $100(1-\alpha)\%$ simultane intervale poverenja za individualne razlike parova δ_j. Ovi intervali dati su izrazom

$$\bar{d}_j \pm \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p;\alpha}}\sqrt{\frac{s_{d_j}^2}{n}}, \tag{5.12}$$

gde je $\bar{d}_j$, $j-$ti element vektora $\bar{\mathbf{d}}$ i $S_{d_j}^2$ $j-$ti dijagonalni element uzoračke kovarijacione matrice $\mathbf{S}_d$. Pri korišćenju dobijenih simultanih intervala poverenja uzeti u obzir da su oni određeni za specifične vrednosti koeficijenata linearne kombinacije, a da se koeficijent pouzdanosti odnosi na sve moguće vrednosti koeficijenata linearne kombinacije.

Primer 5.2 U želji da proveri efikasnost svoje propagandne kampanje u promociji novog osvežavajućeg pića proizvođač je sproveo istraživanje na uzorku od 12 potrošača.

Ispitanik (i)	**Pre kampanje**		**Posle kampanje**	
	Ukus	**Hranljiva vrednost**	**Ukus**	**Hranljiva vrednost**
1	6	7	4	4
2	2	4	3	2
3	5	6	4	5
4	3	2	5	3
5	5	5	7	4
6	4	6	4	5
7	6	2	7	3
8	7	6	6	4
9	1	3	3	2
10	1	6	4	6
11	6	7	7	6
12	2	6	3	4

Pre početka propagandne kampanje izabrani ispitanici su na skali od 1 do 9 ocenjivali dva elementa osvežavajućeg pića: $X_1 = \text{ukus}$ i $X_2 = \text{hranljivu}$ vrednost. Nakon dva meseca propagandne kampanje isti ispitanici su ponovo ocenjivali ista dva elementa osvežavajućeg pića, a rezultati ocenjivanja prezentirani su u tabeli.

Korišćenjem testa poređenja po parovima proveriti da li je propagandna kampanja imala efekta? Ako postoji razlika u proceni ukusa i hranljive vrednosti osvežavajućeg pića od strane potrošača pre i posle propagandne kampanje, kako se ona ispoljava? Objasniti dobijene rezultate.

Ako je propagandna kampanja imala efekta tada će sredina razlike parova za posmatrane dve promenljive biti statistički značajno različita od nula vektora. Stoga testiramo hipotezu $H_0 : \boldsymbol{\delta}' = [\delta_1, \delta_2] = [0, 0]$ na osnovu realizovanih razlika parova za dve promenljive:

$d_{1i} = x_{11i} - x_{21i}$	2	-1	1	-2	-2	0	-1	1	-2	-3	-1	-1
$d_{2i} = x_{12i} - x_{22i}$	3	2	1	-1	1	1	-1	2	1	0	1	2

Uzoračka sredina i kovarijaciona matrica razlike parova su

$$\bar{\mathbf{d}} = \begin{bmatrix} \bar{d}_1 \\ \bar{d}_2 \end{bmatrix} = \begin{bmatrix} -0.75 \\ 1.00 \end{bmatrix} \text{ i } \mathbf{S}_d = \begin{bmatrix} 2.204545 & 1.090909 \\ 1.090909 & 1.454545 \end{bmatrix},$$

pa je $T^2 = 12 \begin{bmatrix} -0.75 & 1.00 \end{bmatrix} \begin{bmatrix} 0.72131 & -0.54098 \\ -0.54098 & 1.09324 \end{bmatrix} \begin{bmatrix} -0.75 \\ 1.00 \end{bmatrix} = 27.7254.$

Na nivou značajnosti $\alpha = 0.05$ kritična vrednost testa je

$$\frac{p(n-1)}{n-p} F_{p,n-p;\alpha} = \frac{2(12-1)}{12-2} 4.1028 = 9.026.$$

Kako je $T^2 = 27.7254 > 9.026$, odbacujemo hipotezu H_0 i zaključujemo da je propagandna kampanja imala efekte. Pri tome je pre kampanje zabeležena niža ocena kada je u pitanju ukus osvežavajućeg pića, a viša kada se ocenjuje njegova hranljiva vrednost u odnosu na vrednovanje posle propagandne kampanje. Da bismo utvrdili koja je od komponenti dovela do odbacivanja nulte hipoteze, odredićemo 95% simultane intervale poverenja razlika ove dve promenljive. Tako za δ_1 imamo

$$\bar{d}_1 \pm \sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p;\alpha}} \sqrt{\frac{s_{d_1}^2}{n}} = -0.75 \pm \sqrt{9.026} \sqrt{\frac{2.204545}{12}},$$

odnosno $\delta_1 \in (-2.0377,\ 0.53772)$, a za δ_2 imamo da $\delta_2 \in (-0.04598,\ 2.04598)$. Na osnovu rezultata T^2-testa uočavamo da je $H_0: \boldsymbol{\delta} = \mathbf{0}$ odbačena na nivou značajnosti od 5%, a da će 95% simultani intervali poverenja sadržati nulu. Kako objasniti ova dva, na prvi pogled protivurečna rezultata. Može se pokazati da 95% oblast poverenja za vektor $\boldsymbol{\delta}$ ne obuhvata tačku $\boldsymbol{\delta} = \mathbf{0}$ što je u skladu sa rezultatom T^2-testa. S druge strane, podsećamo da se 95% koeficijent poverenja kod simultanih intervala poverenja odnosi na sve moguće linearne kombinacije oblika $a_1\delta_1 + a_2\delta_2$. Intervali koje smo mi odredili i koji obuhvataju nulu, tiču se specifične vrednosti koeficijenata linearne kombinacije $(a_1 = 1,\ a_2 = 0)$ i $(a_1 = 0,\ a_2 = 1)$. Neka druga vrednost ovih koeficijenata generisaće simultane intervale koji neće sadržavati nulu. Ako bi se desilo da ne odbacimo nultu hipotezu $H_0: \boldsymbol{\delta} = \mathbf{0}$, korišćenjem T^2-testa, tada bi svi simultani intrervali sadržali nulu. ■

Dosadašnje izlaganje postupka poređenja parova zasnivali smo na sredini razlika parova i njihovoj kovarijacionoj matrici. Do potrebnih statistika za proveru hipoteze o statističkoj značajnosti razlika parova možemo doći i na osnovu uzoračkih pokazatelja originalnih promenljivih. Pokažimo to formirajući $(2p \times 1)$ vektor sredine $\bar{\mathbf{x}}$, kod koga prvih p elemenata predstavljaju sredine p promenljivih prvog tretmana, a narednih p elemenata (od $p+1$ do $2p$ elementa) su sredine p promenljivih drugog tretmana, odnosno to je vektor

$$\bar{\mathbf{x}}' = [\bar{x}_{11}, \bar{x}_{12}, ..., \bar{x}_{1p}, \bar{x}_{21}, \bar{x}_{22}, ..., \bar{x}_{2p}] \tag{5.13}$$

Uzoračka kovarijaciona matrica $\mathbf{S}$ je $(2p \times 2p)$ simetrična matrica čije su podmatrice uzoračke kovarijacione matrice originalnih promenljivih kod prvog $(\mathbf{S}_{11})$ i drugog tretmana $(\mathbf{S}_{22})$, kao i kovarijanse između promenljivih prvog i drugog tretmana $(\mathbf{S}_{12} = \mathbf{S}_{21}')$:

$$\mathbf{S} = \begin{bmatrix} \underset{(p\times p)}{\mathbf{S}_{11}} & \underset{(p\times p)}{\mathbf{S}_{12}} \\ \underset{(p\times p)}{\mathbf{S}_{21}} & \underset{(p\times p)}{\mathbf{S}_{22}} \end{bmatrix}. \tag{5.14}$$

Da bismo uspostavili vezu između ovih uzoračkih pokazatelja originalnih promenljivih i izraza za T^2-test kod poređenja parova, definišimo matricu

$$\underset{(p\times 2p)}{\mathbf{C}} = \begin{bmatrix} 1 & 0 & \cdots & 0 & \vdots & -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \vdots & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & \vdots & 0 & 0 & \cdots & -1 \end{bmatrix} \tag{5.15}$$

kod koje svaki red predstavlja *vektor kontrasta* (razlika) zato što je zbir njegovih elemenata jednak nuli. Kod poređenja po parovima upravo nas interesuju razlike u vrednosti zavisne promenljive pri utvrđivanju efekata tretmana. Matricu čiji su $(p-1)$ redova linearno nezavisni i svaki predstavlja vektor kontrasta nazivamo *matrica kontrasta*. Može se pokazati da važe sledeće relacije

$$\begin{aligned} &\mathbf{d}_i = \mathbf{C}\mathbf{x}_i, \quad i = 1, 2, ..., n \\ &\bar{\mathbf{d}} = \mathbf{C}\bar{\mathbf{x}} \ \text{ i } \ \mathbf{S}_d = \mathbf{C}\mathbf{S}\mathbf{C}' \end{aligned} \tag{5.16}$$

Na osnovu (5.16) sledi da je $T^2 = n\bar{\mathbf{x}}'\mathbf{C}'[\mathbf{C}\mathbf{S}\mathbf{C}']^{-1}\mathbf{C}\bar{\mathbf{x}}$, što nam omogućava primenu postupka testiranja bez potrebe da se razlike po parovima računaju za svaku promenljivu pre i posle primene tretmana.

Primer 5.3 Na osnovu podataka i zahteva iz Primera 5.2 odrediti uzoračke pokazatelje originalnih promenljivih i konstruisati matricu kontrasta. Pokazati da se njihovim korišćenjem dobijaju identični rezultati za statistiku T^2 – testa kao u Primeru 5.2.

Realizovana vrednost uzoračke sredine i kovarijacione matrice originalnih promenljivih su:

$$\bar{\mathbf{x}}' = \begin{bmatrix} \bar{x}_{11} & \bar{x}_{12} & \bar{x}_{21} & \bar{x}_{22} \end{bmatrix} = \begin{bmatrix} 4 & 5 & 4.75 & 4 \end{bmatrix},$$

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \vdots & \mathbf{S}_{12} \\ \cdots & \vdots & \cdots \\ \mathbf{S}_{21} & \vdots & \mathbf{S}_{22} \end{bmatrix} = \begin{bmatrix} 4.5454 & 1.0909 & 2.4545 & 0.7273 \\ 1.0909 & 3.2727 & -0.1818 & 1.8181 \\ 2.4545 & -0.1818 & 2.5682 & 0.5454 \\ 0.7273 & 1.8181 & 0.5454 & 1.8181 \end{bmatrix}.$$

Prema zahtevima u Primeru 5.2 matrica kontrasta je

$$\underset{(2\times 4)}{\mathbf{C}} = \begin{bmatrix} 1 & 0 & \vdots & -1 & 0 \\ 0 & 1 & \vdots & 0 & -1 \end{bmatrix}.$$

Vektor razlika parova je

$$\begin{bmatrix} d_{1i} \\ d_{2i} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_{11i} \\ x_{12i} \\ x_{21i} \\ x_{22i} \end{bmatrix} = \begin{bmatrix} x_{11i} - x_{21i} \\ x_{12i} - x_{22i} \end{bmatrix}.$$

Realizovana vrednost uzoračke sredine je

$$\begin{bmatrix}\overline{d}_{1i}\\ \overline{d}_{2i}\end{bmatrix}=\begin{bmatrix}1 & 0 & -1 & 0\\ 0 & 1 & 0 & -1\end{bmatrix}\begin{bmatrix}\overline{x}_{11i}\\ \overline{x}_{12i}\\ \overline{x}_{21i}\\ \overline{x}_{22i}\end{bmatrix}=\begin{bmatrix}\overline{x}_{11i}-\overline{x}_{21i}\\ \overline{x}_{12i}-\overline{x}_{22i}\end{bmatrix}=\begin{bmatrix}4-4.75\\ 5-4\end{bmatrix}=\begin{bmatrix}-0.75\\ 1\end{bmatrix},$$

a realizovana vrednost njene kovarijacione matrice $\mathbf{S}_d$ je

$$\mathbf{S}_d=\begin{bmatrix}1 & 0 & -1 & 0\\ 0 & 1 & 0 & -1\end{bmatrix}\begin{bmatrix}4.5454 & 1.0909 & 2.4545 & 0.7273\\ 1.0909 & 3.2727 & -0.1818 & 1.8181\\ 2.4545 & -0.1818 & 2.5682 & 0.5454\\ 0.7273 & 1.8181 & 0.5454 & 1.8181\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\\ -1 & 0\\ 0 & -1\end{bmatrix}=\begin{bmatrix}2.2045 & 1.0909\\ 1.0909 & 1.4545\end{bmatrix}$$

Na osnovu dobijenih rezultata očito da je vrednost statistike T^2-testa ista kao u Primeru 5.2, odn. $T^2=27.7254$, što je trebalo pokazati. ■

5.3 PLAN PONOVLJENIH MERENJA

Plan eksperimenta u kome poredimo p tretmana s obzirom na jednu promenljivu odziva nazivamo plan ponovljenih merenja. Kod njega, istih n eksperimentalnih jedinica posmatramo pod različitim uslovima, odnosno tretmanima (na primer u p različitih vremenskih perioda) registrujući promenljivu odziva.

Neka je $i-$ta opservacija, odn. vektor odziva $i-$te eksperimentalne jedinice za p tretmana

$$\mathbf{X}_i=\begin{bmatrix}X_{i1}\\ X_{i2}\\ \vdots\\ X_{ip}\end{bmatrix},\ i=1,2,...,n \tag{5.17}$$

gde je X_{ij} odziv $i-$te jedinice na dejstvo $j-$tog tretmana. Pretpostavićemo da svaka opservacija X_i ima $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ raspored. Interesuje nas da ispitamo da li je odziv eksperimentalnih jedinica isti pod uslovima primene p tretmana. Dakle, nultom hipotezom pretpostavljamo da je prosečan odziv n jedinica isti za svaki od p tretmana, tj. $H_0:\mu_1=\mu_2=...=\mu_p$, protiv alternativne hipoteze da je pod dejstvom bar jednog od tretmana prosečan odziv eksperimentalnih jedinica različit u odnosu na odzive pod dejstvom preostalih tretmana. Nulta hipoteza o jednakosti prosečnih odziva po tretmanima ekvivalentna je hipotezi izraženoj u obliku $H_0:(\mu_1-\mu_2)=(\mu_2-\mu_3)=...$ $=(\mu_{p-1}-\mu_p)=0$.

Ako formiramo $\mathbf{Y}_i$ $[(p-1)\times 1]$ vektor kao vektor razlika uzastopnih elemenata vektora opservacija čiji su elementi

$$Y_{ij} = X_{ij} - X_{i,j+1},\ i = 1,2,...,n,\ j = 1,2,...,p-1, \tag{5.18}$$

tada možemo na osnovu njegovog uzoračkog sredina $\bar{\mathbf{Y}}$ i uzoračke kovarijacione matrice $\mathbf{S}_Y$, a na bazi Hotellingove T^2-statistike, testirati hipotezu o jednakom odzivu eksperimentalnih jedinica na dejstvo tretmana. Naime, ako je tačna nulta hipoteza, statistika

$$T^2 = n\bar{\mathbf{Y}}'\mathbf{S}_Y^{-1}\bar{\mathbf{Y}} \tag{5.19}$$

je raspoređeno kao $\frac{(n-1)(p-1)}{n-p+1}F_{p-1,n-p+1}$, gde je $F_{p-1,n-p+1}$ $F-$raspored sa $p-1$ i $n-p+1$ stepeni slobode. Ako je izračunata vrednost ove statistike testa veća od kritične vrednosti odbacujemo nultu hipotezu na nivou značajnosti α. Nultu hipotezu o jednakim efektima tretmana mogli smo iskazati i preko razlike prve komponente sredine $\boldsymbol{\mu}$ i ostalih komponenata, tj. $H_0:(\mu_1-\mu_2)=(\mu_1-\mu_3)=...=(\mu_1-\mu_p)=0$, te formirati statistiku testa na isti način kao u izrazu (5.19). Ova jednoznačnost testa u dva navedena primera drugačijeg izražavanja nulte hipoteze, posledica je osobine T^2-statistike prema kojoj se njena vrednost ne menja linearnom transformacijom opservacija. Ako smo na osnovu izraza $\mathbf{Y}_i = \mathbf{C}\mathbf{X}_i$ transformisali $i-$tu originalnu opservaciju, tada za na prvi način iskazanu nultu hipotezu matricu $\mathbf{C}$ pišemo u obliku

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}, \tag{5.20}$$

a samu nultu hipotezu možemo napisati u obliku $H_0:\mathbf{C}\boldsymbol{\mu}=\mathbf{0}$. Matricu $\mathbf{C}$ nazivamo matrica kontrasta jer nam omogućava raznovrsno poređenje efekata tretmana, a ne samo proveru jednakog dejstva tretmana na eksperimentalne jedinice.

Podsetimo se rasporeda linearne transformacije originalnih opservacija. Pretpostavili smo da je $\mathbf{X}_i \sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, pa će tada $\mathbf{Y}_i$ imati $N_{p-1}(\mathbf{C}\boldsymbol{\mu},\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')$ raspored, a statistika $n(\mathbf{C}\bar{\mathbf{X}})'(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')^{-1}(\mathbf{C}\bar{\mathbf{X}})$ će imati χ^2-raspored sa $p-1$ stepeni slobode. Ako nam je pri tome kovarijaciona matrica $\boldsymbol{\Sigma}$ nepoznata, koristimo njenu uzoračku ocenu $\mathbf{S}$, pa statistika $n(\mathbf{C}\bar{\mathbf{X}})'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}(\mathbf{C}\bar{\mathbf{X}})$ ima T^2-raspored kao što navedeno u 3. Glavi. Sada možemo preko uzoračkih pokazatelja originalnih opservacija testirati hipoteze o raznovrsnom međusobnom odnosu sredina odziva eksperimentalnih jedinica na dejstvo različitih tretmana.

Hipotezu $H_0 : \mathbf{C\mu} = \mathbf{0}$ protiv hipoteze $H_1 : \mathbf{C\mu} \neq \mathbf{0}$, odbacujemo na nivou značajnosti α ako je

$$T^2 = n(\mathbf{C}\bar{\mathbf{X}})'(\mathbf{CSC}')^{-1}(\mathbf{C}\bar{\mathbf{X}}) > \frac{(n-1)(p-1)}{n-p+1} F_{p-1,n-p+1;\alpha} \tag{5.21}$$

gde je $F_{p-1,n-p+1;\alpha}$ gornji 100α percentile $F-$rasporeda sa $p-1$ i $n-p+1$ stepeni slobode.

Oblast poverenja za kontrast $\mathbf{C\mu}$ konstruišemo na osnovu T^2-rasporeda. Kažemo da $100(1-\alpha)\%$ oblast poverenja obuhvata sve kontraste $\mathbf{C\mu}$ za koje je

$$n(\mathbf{C}\bar{\mathbf{X}})'(\mathbf{CSC}')^{-1}(\mathbf{C}\bar{\mathbf{X}}) \leq \frac{(n-1)(p-1)}{n-p+1} F_{p-1,n-p+1;\alpha} \tag{5.22}$$

Ukoliko je hipoteza o jednakim efektima tretmana odbačena, na osnovu simultanih intervala poverenja možemo utvrditi koji od tretmana se po svojim efektima razlikuju međusobno. $100(1-\alpha)\%$ simultani intervali poverenja za jedan kontrast $\mathbf{c}'\boldsymbol{\mu}$, za ma koji vektor kontrasta $\mathbf{c}$, dat je sa

$$\mathbf{c}'\bar{\mathbf{x}} \pm \sqrt{\frac{(n-1)(p-1)}{n-p+1} F_{p-1,n-p+1;\alpha}} \sqrt{\frac{\mathbf{c}'\mathbf{S}\mathbf{c}}{n}} \tag{5.23}$$

Primer 5.4 Kao ilustraciju postupka provere jednakih efekata tretmana kod plana ponovljenih merenja poslužićemo se primerom koji smo naveli u Poglavlju 5.1.

U izabranim prodavnicama $(n=15)$ tokom četiri meseca registrovali smo prodaju osvežavajućeg pića. Ono se prodaje u plastičnoj ambalaži (plastične flaše) i kartonskoj ambalaži (tzv. brik pakovanje). Pored vrste ambalaže razlikujemo veličinu pakovanja i to porodično pakovanje od 1 litra i individualno pakovanje od 0.2 litra. Pri tome nas je interesovalo da utvrdimo kakav je uticaj na prodaju osvežavajućeg pića vrste ambalaže i veličine pakovanja. Prodaju smo registrovali za svaku od četiri kombinacije, odnosno u ovom primeru imamo četiri tretmana. U donjoj tabeli date su prodate količine u litrama ovog osvežavajućeg pića za svaki od tretmana.

Tretman 1 = Porodično pakovanje - plastična ambalaža
Tretman 2 = Individualno pakovanje - plastična ambalaža
Tretman 3 = Porodično pakovanje - kartonska ambalaža
Tretman 4 = Individualno pakovanje - kartonska ambalaža

U analizi prodaje možemo testirati jednakost prosečnih prodaja tokom ova četiri meseca (tretmana). Međutim, smatramo da je interesantnije utvrditi da li postoji razlika u prodaji s obzirom na vrstu ambalaže, zatim s

obzirom na veličinu pakovanja i naposletku zbog međusobnog uticaja vrste ambalaže i veličine pakovanja.

Prodavnica	T r e t m a n			
	1	2	3	4
1	125	135	209	187
2	287	271	304	297
3	311	301	399	321
4	198	205	254	308
5	155	183	162	218
6	167	171	178	230
7	244	286	309	300
8	181	193	261	219
9	243	264	234	317
10	213	227	219	302
11	179	169	222	226
12	202	209	285	257
13	195	208	273	262
14	206	201	261	233
15	199	187	255	208

Ako označimo prosečne prodaje osvežavajućeg pića za ova četiri tretmana sa μ_1, μ_2, μ_3 i μ_4 tada se gore, rečima iskazani kontrasti mogu pisati na sledeći način

$(\mu_3+\mu_4)-(\mu_1+\mu_2)=$ Kontrast ambalaže: razlika u prodaji između plastične i kartonske ambalaže

$(\mu_1+\mu_3)-(\mu_2+\mu_4)=$ Kontrast veličine pakovanja: razlika u prodaji između porodičnog i individualnog pakovanja

$(\mu_1+\mu_4)-(\mu_2+\mu_3)=$ Interakcija vrste ambalaže i veličine pakovanja: razlika u prodaji zbog međusobnog uticaja vrste ambalaže i veličine pakovanja

Prema ovako iskazanim kontrastima matrica $\mathbf{C}$ je

$$\mathbf{C}=\begin{bmatrix}-1 & -1 & 1 & 1\\ 1 & -1 & 1 & -1\\ 1 & -1 & -1 & 1\end{bmatrix}.$$

Realizovana vrednost uzoračke sredine i kovarijacione matrice je

$$\bar{\mathbf{x}}=\begin{bmatrix}207\\214\\255\\259\end{bmatrix} \text{ i } \mathbf{S}=\begin{bmatrix}2337.143 & 2155.214 & 2286.786 & 1748.429\\ & 2230.571 & 2039.357 & 1840.571\\ & & 2247.857 & 1445.286\\ & & & 2019.143\end{bmatrix}.$$

Može se pokazati da je

$$\mathbf{C}\bar{\mathbf{x}} = \begin{bmatrix} 93 \\ -11 \\ -3 \end{bmatrix}, \ \mathbf{CSC}' = \begin{bmatrix} 1305.429 & 640.286 & -542.857 \\ & 3412.857 & -2219.143 \\ & & 2054.571 \end{bmatrix}$$

i $T^2 = n(\mathbf{C}\bar{\mathbf{x}})'(\mathbf{CSC}')^{-1}(\mathbf{C}\bar{\mathbf{x}}) = 15 \cdot 7.6597 = 114.8962$. Kritična vrednost na nivou značajnosti $\alpha = 0.05$ iznosi

$$\frac{(n-1)(p-1)}{n-p+1} F_{p-1,n-p+1;\alpha} = \frac{(15-1)(4-1)}{15-4+1} F_{3,12;0.05} = \frac{14 \cdot 3}{12} 3.49 = 12.2161$$

Izračunata vrednost T^2 statistike veća je od kritične vrednosti, pa odbacujemo $H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$ (nema efekata među tretmanima) na nivou značajnosti od 5%. U ovom slučaju simultani intervali poverenja kontrasta pružaju odgovor koji od njih je odgovoran za odbacivanje nulte hipoteze. Za prvi kontrast $\mathbf{c}_1'\boldsymbol{\mu}$ (razlika u prodaji između plastične i kartonske ambalaže) 95% simultani interval poverenja je

$$(\bar{x}_3 + \bar{x}_4) - (\bar{x}_1 + \bar{x}_2) \pm \sqrt{\frac{(n-1)(p-1)}{n-p+1} F_{p-1,n-p+1;\alpha}} \sqrt{\frac{\mathbf{c}_1'\mathbf{S}\mathbf{c}_1}{n}}$$

$$= 93 \pm \sqrt{12.216} \sqrt{\frac{1305.429}{15}}$$

$$= 93 \pm 32.6059 .$$

$$(\mu_3 + \mu_4) - (\mu_1 + \mu_2) \in (60.3941;\ 125.6059) .$$

Preostala dva simultana intervala poverenja za kontraste su:

$$(\mu_1 + \mu_3) - (\mu_2 + \mu_4) \in (-63.7204;\ 41.7204)$$

$$(\mu_1 + \mu_4) - (\mu_2 + \mu_3) \in (-43.9054;\ 37.9054) .$$

Na osnovu ovih simultanih intervala poverenja zaključujemo da postoji efekat ambalaže na prodaju. Prodaja je veća ako se osvežavajuće piće prodaje u kartonskoj, a ne u plastičnoj ambalaži. Ovo se dešava bez obzira na veličinu pakovanja, s obzirom da interakcija vrste ambalaže i veličine pakovanja nije statistički značajna. Kako drugi simultani interval za kontrast veličine pakovanja uključuje nulu znači da ne postoji efekat veličine pakovanja na prodaju. ■

Pri definisanju statistike testa korišćene za testiranje jednakosti tretmana u planu ponovljenih merenja nismo pretpostavljali da kovarijaciona matrica $\boldsymbol{\Sigma}$ ima određenu

strukturu. Ukoliko je međutim, moguće pretpostaviti postojanje određene strukture kod kovarijacione matrice, tada je moguće konstruisati i statistiku kojoj će odgovarati veća jačina testa. Primer kovarijacione matrice sa specijalnom strukturom je matrica koja ima jednake varijanse i jednake kovarijanse. Kovarijaciona matrica sa ovakvom strukturom predstavlja dovoljan, ali ne i potreban uslov za valjanost $F-$statistike u analizi varijanse podataka koji su proistekli iz plana eksperimenta ponovljenih merenja. Potreban i dovoljan uslov za valjanost $F-$statistike nameće određena ograničenja na elemente kovarijacione matrice, a za matricu sa takvom strukturom kažemo da je kovarijaciona matrica *Huynh-Feldtovog tipa*. Ti uslovi su ekvivalentni zahtevu da razlike između svih parova odziva imaju istu varijansu i istu kovarijansu[2].

5.4 M A N O V A SA JEDNIM FAKTOROM

Pretpostavimo da želimo da utvrdimo da li postoji značajna razlika u opredeljenju kupaca prema kupovini neke robe s obzirom na to da se kupovina može obaviti u različitom tipu prodavnice (samoposluga, klasična prodavnica i specijalizovana radnja). Opredeljenje kupaca možemo iskazati korišćenjem niza objektivnih pokazatelja (na primer kupljena količina te robe, broj kupovina u toku meseca i dr.) i niza subjektivnih pokazatelja (na primer na subjektivnoj skali od 0 do 9 iskazan stepen pažnje koju prodavci posvećuju kupcu po njegovoj oceni, osećaj lagodnosti pri kupovini i dr.). Ovi pokazatelji uzimaće različite vrednosti od kupca do kupca, a razumno je pretpostaviti da se razlika koja se javlja u prosečnoj vrednosti svih ovih pokazatelja može pripisati upravo tipu prodavnice u kojoj je obavljena kupovina.

Teorijski opis situacije sa kojom se suočavamo u našem primeru, kao i pretpostavke koje su implicitno sadržane u njemu, mogu se naći u odgovarajućem planu eksperimenta. Da bismo definisali odgovarajući statistički model uspostavićemo korespondenciju između elemenata iz gornjeg primera i teorijskih pojmova koji su u osnovi statističkog modela.

Eksperimentalna jedinica u ovom primeru je kupac, a faktor čiji uticaj na opredeljenost kupaca ispitujemo nazivamo kontrolisani faktor i identifikujemo ga sa tipom prodavnice. Svi ostali uticaji na opredeljenost kupaca svrstani su u tzv. nekontrolisane faktore. Kažemo da ispitujemo uticaj jednog faktora na varijabilitet posmatrane pojave, pa se stoga odgovarajući model analize naziva model sa jednim faktorom. Kontrolisani faktor u našem primeru ima tri nivoa ili tretmana i svaki se tiče posebnog tipa prodavnice. Kako su tretmani fiksirani unapred, model koji predstavlja opis gornje situacije nazivamo model sa fiksiranim efektima. Uslovi u kojima se odvija ovaj statistički eksperiment sugerišu da se tretmani na slučajan način pridružuju eksperimentalnim jedinicama. One su sa svoje strane međusobom relativno slične, čime se smanjuje uticaj nekontrolisanih faktora na dobijene rezultate.

---

[2] Detalji vezani za ovaj model analize varijanse kao i statistiku testa mogu se naći kod Morrisona (1976, ss. 150-3), a kod istog autora izloženi su i testovi za kovarijacione matrice sa ovom specijalnom strukturom (Morison (1976, ss. 250-2)).

Reakcija ili odziv eksperimentalnih jedinica na različite tretmane jeste upravo opredeljenje kupaca prema kupovini neke robe shvaćen kao višedimenziona pojava. Njegove komponente su objektivni i subjektivni pokazatelji kroz koje se ta pojava ispoljava. Odzive eksperimentalnih jedinica dobijene pod uticajem jednog od tretmana smatramo uzorkom iz višedimenzione populacije određenih karakteristika. Pretpostavivši da su populacije (tretmani) identične postavili smo hipotezu da kontrolisani faktor nema uticaja na odziv eksperimentalnih jedinica. Odziv varira od jedne do druge eksperimentalne jedinice i osnovni zadatak u okviru analize jeste razlaganje ukupne varijacije odziva eksperimentalnih jedinica na dve komponente od kojih je jedna deo varijacija koji se pripisuje uticaju kontrolisanog faktora, a druga komponenta obuhvata uticaj nekontrolisanih faktora. Ukoliko je udeo varijacija koji je nastao pod dejstvom kontrolisanog faktora u ukupnoj varijaciji statistički zanemarljiv, tada uočeno variranje odziva eksperimentalnih jedinica pripisujemo isključivo uticaju nekontrolisanih faktora. Na osnovu primera i opisa statističkog eksperimenta u prilici smo da preciznije definišemo statistički model i iskažemo pretpostavke koje su u osnovi MANOVA.

5.4.1 Model MANOVA sa jednim faktorom

Posmatrajmo g populacija, tj. pretpostavimo da kontrolisani faktor ima g tretmana. Sredina $(p\times 1)$ slučajnog vektora $\mathbf{X}_k$, $k=1,2,...,g$, $k-$te populacije označićemo sa $\boldsymbol{\mu}_k$, gde je

$$\boldsymbol{\mu}_k=\begin{bmatrix}\mu_{1k}\\ \mu_{2k}\\ \vdots\\ \mu_{pk}\end{bmatrix},\ k=1,2,...,g \tag{5.24}$$

a opšta sredina svih populacija, u oznaci $\boldsymbol{\mu}$, je

$$\boldsymbol{\mu}=\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_p\end{bmatrix}. \tag{5.25}$$

Pretpostavićemo da je kovarijaciona matrica ista kod svake populacije i označićemo je sa $\boldsymbol{\Sigma}$. Neka je iz svake populacije uzet slučajan uzorak od po n_k, $k=1,2,...,g$, elemenata: $\mathbf{X}_{1k},\mathbf{X}_{2k},...,\mathbf{X}_{n_k k}$, odnosno

$$\begin{bmatrix}X_{11k}\\ X_{12k}\\ \vdots\\ X_{1pk}\end{bmatrix},\begin{bmatrix}X_{21k}\\ X_{22k}\\ \vdots\\ X_{2pk}\end{bmatrix},\ \ldots,\ \begin{bmatrix}X_{n_k 1k}\\ X_{n_k 2k}\\ \vdots\\ X_{n_k pk}\end{bmatrix},\ k=1,2,...,g\ , \tag{5.26}$$

gde smo sa $\mathbf{X}_{ik}$, $i=1,2,...,n_k$, označili $(p\times 1)$ vektor odziva $i-$te eksperimentalne jedinice iz $k-$te populacije. Slučajni uzorci iz različitih populacija međusobom su nezavisni. Pokazatelji slučajnog uzorka za $k-$tu populaciju je uzoračka sredina

$$\bar{\mathbf{X}}_k=\begin{bmatrix}\bar{X}_{\cdot 1k}\\ \bar{X}_{\cdot 2k}\\ \vdots\\ \bar{X}_{\cdot pk}\end{bmatrix},\ \text{gde je}\ \bar{X}_{\cdot jk}=\frac{1}{n_k}\sum_{i=1}^{n_k}X_{ijk},\ j=1,2,...,p,\ k=1,2,...,g \tag{5.27}$$

i uzoračka kovarijaciona matrica

$$\mathbf{S}_k=\frac{1}{n_k-1}\sum_{i=1}^{n_k}(\mathbf{X}_{ik}-\bar{\mathbf{X}}_k)(\mathbf{X}_{ik}-\bar{\mathbf{X}}_k)',\ k=1,2,...,g \tag{5.28}$$

Na osnovu njih formiramo opštu uzoračku sredinu svih g populacija, u oznaci $\bar{\mathbf{X}}$,

$$\bar{\mathbf{X}}=\begin{bmatrix}\bar{X}_{\cdot 1\cdot}\\ \bar{X}_{\cdot 2\cdot}\\ \vdots\\ \bar{X}_{\cdot p\cdot}\end{bmatrix},\ \text{gde je}\ \bar{X}_{\cdot j\cdot}=\frac{1}{\sum_{k=1}^{g}n_k}\sum_{k=1}^{g}\sum_{i=1}^{n_k}X_{ijk}=\frac{1}{n}\sum_{k=1}^{g}n_kX_{\cdot jk},\ n=\sum_{k=1}^{g}n_k \tag{5.29}$$

Model MANOVA sa jednim faktorom formulišemo za proizvoljnu $i-$tu opservaciju iz $k-$te populacije

$$\mathbf{X}_{ik}=\boldsymbol{\mu}+\boldsymbol{\alpha}_k+\boldsymbol{\varepsilon}_{ik},\ i=1,2,...,n_k,\ k=1,2,...,g, \tag{5.30}$$

gde su $\boldsymbol{\varepsilon}_{ik}$ slučajne greške međusobom nezavisne sa $N_p(\mathbf{0},\boldsymbol{\Sigma})$ rasporedom, $\boldsymbol{\mu}$ je opšta sredina svih populacija, a $\boldsymbol{\alpha}_k$ je efekat $k-$tog tretmana. Kako je efekat $k-$tog tretmana $\boldsymbol{\alpha}_k=\boldsymbol{\mu}_k-\boldsymbol{\mu}$ (pokazuje za koliko sredina $k-$te populacije odstupa od opšte sredine), to model pišemo kao

$$\mathbf{X}_{ik}=\boldsymbol{\mu}+(\boldsymbol{\mu}_k-\boldsymbol{\mu})+(\mathbf{X}_{ik}-\boldsymbol{\mu}_k) \tag{5.31}$$

ili u razvijenom obliku

$$\begin{bmatrix}X_{i1k}\\ X_{i2k}\\ \vdots\\ X_{ipk}\end{bmatrix}=\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_p\end{bmatrix}+\begin{bmatrix}\mu_{1k}-\mu_1\\ \mu_{2k}-\mu_2\\ \vdots\\ \mu_{pk}-\mu_p\end{bmatrix}+\begin{bmatrix}X_{i1k}-\mu_{1k}\\ X_{i2k}-\mu_{2k}\\ \vdots\\ X_{ipk}-\mu_{pk}\end{bmatrix},\ i=1,2,...,n_k,\ k=1,2,...,g, \tag{5.32}$$

Na osnovu vrednosti slučajnog uzorka iz $k-$te populacije $\mathbf{X}_{1k}, \mathbf{X}_{2k}, ..., \mathbf{X}_{n_k k}$, model MANOVA sa jednim faktorom možemo oceniti, odnosno razložiti opservacije na tri dela:

$$\underset{\text{opservacija}}{\mathbf{X}_{ik}} = \underset{\substack{\text{ocena zajedničke}\\ \text{sredine}}}{\bar{\mathbf{X}}} + \underset{\substack{\text{ocenjeni efekat}\\ \text{tretmana}}}{(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})} + \underset{\text{ocenjeni rezidual}}{(\bar{\mathbf{X}}_{ik} - \bar{\mathbf{X}}_k)}$$

Pokažimo kako se po analogiji sa jednodimenzionom analizom varijanse može izvršiti razlaganje ukupne varijacije na deo koji se pripisuje efektima tretmana (faktorska varijacija) i deo koji se pripisuje nekontrolisanim faktorima (rezidualna varijacija). Polazaći od razlaganja opservacije, odstupanje $i-$te opservacije kod $k-$te populacije od opšte sredine je $\mathbf{X}_{ik} - \bar{\mathbf{X}} = (\bar{\mathbf{X}}_k - \bar{\mathbf{X}}) + (\mathbf{X}_{ik} - \bar{\mathbf{X}}_k)$. Ako formiramo proizvod vektora $\mathbf{X}_{ik} - \bar{\mathbf{X}}$ sa samim sobom dobićemo matricu, koju pišemo kao

$$\begin{aligned}(\mathbf{X}_{ik} - \bar{\mathbf{X}})(\mathbf{X}_{ik} - \bar{\mathbf{X}})' &= \left[(\mathbf{X}_{ik} - \bar{\mathbf{X}}_k) + (\bar{\mathbf{X}}_k - \bar{\mathbf{X}})\right]\left[(\mathbf{X}_{ik} - \bar{\mathbf{X}}_k) + (\bar{\mathbf{X}}_k - \bar{\mathbf{X}})\right]' \\ &= (\mathbf{X}_{ik} - \bar{\mathbf{X}}_k)(\mathbf{X}_{ik} - \bar{\mathbf{X}}_k)' + (\mathbf{X}_{ik} - \bar{\mathbf{X}}_k)(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})' \\ &\quad + (\bar{\mathbf{X}}_k - \bar{\mathbf{X}})(\mathbf{X}_{ik} - \bar{\mathbf{X}}_k)' + (\bar{\mathbf{X}}_k - \bar{\mathbf{X}})(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})'\end{aligned} \tag{5.33}$$

Sumirajući po indeksima i i k izraze na levoj i desnoj strani dobićemo

$$\underset{\substack{\text{Matrica ukupne}\\ \text{sume kvadrata}}}{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{X}_{ik} - \bar{\mathbf{X}})(\mathbf{X}_{ik} - \bar{\mathbf{X}})'} = \underset{\substack{\text{Matrica sume kvadrata}\\ \text{između tretmana}}}{\sum_{k=1}^{g} n_k(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})'} + \underset{\substack{\text{Matrica sume kvadrata}\\ \text{unutar tretmana}}}{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{X}_{ik} - \bar{\mathbf{X}}_k)(\mathbf{X}_{ik} - \bar{\mathbf{X}}_k)'}$$

jer su dva središnja sabirka nula matrice, što sledi na osnovu jednakosti $\sum_{i=1}^{n_k}(\mathbf{X}_{ik} - \bar{\mathbf{X}}_k) = 0$ i $\sum_{k=1}^{g} n_k(\bar{\mathbf{X}}_k - \bar{\mathbf{X}}) = 0$.

Matricu sume kvadrata unutar tretmana, u oznaci $\mathbf{W}$

$$\mathbf{W} = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{X}_{ik} - \bar{\mathbf{X}}_k)(\mathbf{X}_{ik} - \bar{\mathbf{X}}_k)' \tag{5.34}$$

možemo izraziti kao ponderisanu sumu uzoračkih kovarijacionih matrica, $\mathbf{S}_k$, $k = 1, 2, ..., g$

$$\mathbf{W} = (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + ... + (n_g - 1)\mathbf{S}_g \tag{5.35}$$

Na osnovu nje formiramo ocenu opšte kovarijacione matrice uz pretpostavku o njenoj homogenosti unutar populacija, u oznaci $\bar{\mathbf{S}}$, gde je:

$$\overline{\mathbf{S}} = \tfrac{1}{n-g}\mathbf{W}. \tag{5.36}$$

Matrica sume kvadrata između tretmana, u oznaci $\mathbf{B}$ je:

$$\mathbf{B} = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})' = \sum_{k=1}^{g} n_k(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})' \tag{5.37}$$

Njen $j-$ti element na glavnoj dijagonali je suma kvadrata odstupanja sredine $k-$te populacije od opšte sredine $j-$te promenljive. Vandijagonalni elementi predstavljaju sume uzajamnih proizvoda odstupanja sredine $k-$te populacije od opšte sredine.

Matrica sume kvadrata unutar tretmana $\mathbf{W}$, za $j-$ti element na glavnoj dijagonali ima sumu kvadrata odstupanja opservacija od sredine $k-$te populacije, dok su vandijagonalni elementi sume uzajamnih proizvoda odstupanja opservacije od sredine $k-$te populacije.

Matrica ukupne sume kvadrata, u oznaci $\mathbf{T}$, gde je $\mathbf{T} = \mathbf{B} + \mathbf{W}$, predstavlja uopštenje ukupne sume kvadrata iz jednodimenzionog slučaja. Njeni elementi su sume kvadrata i uzajamnih proizvoda odstupanja opservacija od opšte sredine.

Primer 5.5 Data su tri nezavisna uzorka iz tri nezavisne populacije.

$$\text{Populacija 1: } \begin{bmatrix}7\\4\end{bmatrix}, \begin{bmatrix}9\\6\end{bmatrix},$$

$$\text{Populacija 2: } \begin{bmatrix}6\\8\end{bmatrix}, \begin{bmatrix}4\\5\end{bmatrix}, \begin{bmatrix}5\\8\end{bmatrix},$$

$$\text{Populacija 3: } \begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}6\\2\end{bmatrix}, \begin{bmatrix}1\\4\end{bmatrix}.$$

Korišćenjem izraza za razlaganje opservacija izvršiti razlaganje gornjih opservacija. Formirati matrice sume kvadrata: između i unutar tretmana i ukupne sume kvadrata.

Odredimo prvo sredinu za svaki od tri uzorka, a takođe i opštu sredinu

$$\overline{\mathbf{x}}_1 = \begin{bmatrix}8\\5\end{bmatrix},\ \overline{\mathbf{x}}_2 = \begin{bmatrix}5\\7\end{bmatrix},\ \overline{\mathbf{x}}_3 = \begin{bmatrix}3\\3\end{bmatrix},\ \overline{\mathbf{x}} = \begin{bmatrix}5\\5\end{bmatrix}.$$

Formirajmo matricu opservacija čije kolone predstavljaju realizacije slučajnog uzorka. Tada se zahtevano razlaganje opservacija može kompaktnije prezentirati za prvu populaciju na sledeći način:

$$\begin{bmatrix} x_{111} & x_{211} \\ x_{121} & x_{221} \end{bmatrix} = \begin{bmatrix} \overline{x}_{\cdot 1\cdot} & \overline{x}_{\cdot 1\cdot} \\ \overline{x}_{\cdot 2\cdot} & \overline{x}_{\cdot 2\cdot} \end{bmatrix} + \begin{bmatrix} \overline{x}_{\cdot 11} - \overline{x}_{\cdot 1\cdot} & \overline{x}_{\cdot 11} - \overline{x}_{\cdot 1\cdot} \\ \overline{x}_{\cdot 21} - \overline{x}_{\cdot 2\cdot} & \overline{x}_{\cdot 21} - \overline{x}_{\cdot 2\cdot} \end{bmatrix} + \begin{bmatrix} x_{111} - \overline{x}_{\cdot 11} & x_{211} - \overline{x}_{\cdot 11} \\ x_{121} - \overline{x}_{\cdot 21} & x_{221} - \overline{x}_{\cdot 21} \end{bmatrix}$$

$$\begin{bmatrix} 7 & 9 \\ 4 & 6 \end{bmatrix} = \begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix} + \begin{bmatrix} 3 & 3 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}.$$

Za drugu populaciju:

$$\begin{bmatrix} 6 & 4 & 5 \\ 8 & 5 & 8 \end{bmatrix} = \begin{bmatrix} 5 & 5 & 5 \\ 5 & 5 & 5 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 2 & 2 & 2 \end{bmatrix} + \begin{bmatrix} 1 & -1 & 0 \\ 1 & -2 & 1 \end{bmatrix}.$$

Za treću populaciju:

$$\begin{bmatrix} 2 & 6 & 1 \\ 3 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 5 & 5 & 5 \\ 5 & 5 & 5 \end{bmatrix} + \begin{bmatrix} -2 & -2 & -2 \\ -2 & -2 & -2 \end{bmatrix} + \begin{bmatrix} -1 & 3 & -2 \\ 0 & -1 & 1 \end{bmatrix}.$$

Matrica sume kvadrata između tretmana je

$$\mathbf{B} = \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix}\begin{bmatrix} 3 & 0 \\ 3 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 2 & 2 & 2 \end{bmatrix}\begin{bmatrix} 0 & 2 \\ 0 & 2 \\ 0 & 2 \end{bmatrix} + \begin{bmatrix} -2 & -2 & -2 \\ -2 & -2 & -2 \end{bmatrix}\begin{bmatrix} -2 & -2 \\ -2 & -2 \\ -2 & -2 \end{bmatrix}$$

$$= \begin{bmatrix} 18 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 12 \end{bmatrix} + \begin{bmatrix} 12 & 12 \\ 12 & 12 \end{bmatrix} = \begin{bmatrix} 30 & 12 \\ 12 & 24 \end{bmatrix}.$$

Matrica sume kvadrata unutar tretmana je

$$\mathbf{W} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 & 0 \\ 1 & -2 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -1 & -2 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} -1 & 3 & -2 \\ 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} -1 & 0 \\ 3 & -1 \\ -2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} + \begin{bmatrix} 2 & 3 \\ 3 & 6 \end{bmatrix} + \begin{bmatrix} 14 & -5 \\ -5 & 2 \end{bmatrix} = \begin{bmatrix} 18 & 0 \\ 0 & 10 \end{bmatrix}.$$

Matrica ukupne sume kvadrata je na osnovu izračunatih matrica **B** i **W** jednaka:

$$\mathbf{T} = \mathbf{B} + \mathbf{W} = \begin{bmatrix} 30 & 12 \\ 12 & 24 \end{bmatrix} + \begin{bmatrix} 18 & 0 \\ 0 & 10 \end{bmatrix} = \begin{bmatrix} 48 & 12 \\ 12 & 34 \end{bmatrix}. \qquad \blacksquare$$

5.4.2 Zaključivanje kod MANOVA sa jednim faktorom

Nultu hipotezu koju testiramo metodom multivarijacione analize varijanse iskazujemo preko jednakosti sredina razmatranih populacija i ona glasi

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_g = \boldsymbol{\mu} \tag{5.38}$$

Alternativna hipoteza je da se sredine bar dve populacije razlikuju međusobom. Uobičajeno je sredinu $k-$te populacije posmatrati kao da se sastoji iz dva dela, opšte sredine i komponente koja je specifična za tu populaciju. Tada $\boldsymbol{\mu}_k$ pišemo kao

$$\boldsymbol{\mu}_k = \boldsymbol{\mu} + (\boldsymbol{\mu}_k - \boldsymbol{\mu}) = \boldsymbol{\mu} + \boldsymbol{\alpha}_k . \tag{5.39}$$

Ukoliko eksperimentalni faktor nema uticaja na zavisnu promenljivu, tada nema razlike između specifičnih sredina posmatranih populacija i opšte sredine. U tom slučaju kažemo da je efekat tretmana jednak nuli. Znači da nultu hipotezu, na osnovu vrednosti tretmana, možemo iskazati na sledeći način:

$$H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = ... = \boldsymbol{\alpha}_g = \mathbf{0} \tag{5.40}$$

Alternativna hipoteza u ovom slučaju glasi da je efekat bar jednog tretmana različit od nule. Ako je tačna nulta hipoteza $i-$ta opservacija iz $k-$te populacije $\mathbf{X}_{ik}$, sastoji se samo od opšte sredine $\boldsymbol{\mu}$ i slučajne greške $\boldsymbol{\varepsilon}_{ik}$ (prema modelu MANOVA sa jednim faktorom). To znači da opservacije variraju samo pod uticajem nekontrolisanih faktora.

Podsećamo da smo za testiranje nulte hipoteze o jednakosti sredina više populacija u slučaju jednodimenzione zavisne promenljive (metod ANOVA) koristili statistiku testa konstruisanu na principu količnika verodostojnosti. U krajnjem izrazu te statistike, na bazi $F-$testa, poredimo sumu kvadrata tretmana (S_A) i sumu kvadrata reziduala (S_R). Za velike vrednosti količnika S_A / S_R, odnosno $1 + S_A / S_R$, odbacujemo nultu hipotezu. Alternativno, kažemo da nultu hipotezu odbacujemo za male vrednosti izraza

$$\frac{1}{1+\frac{S_A}{S_R}} = \frac{S_R}{S_R + S_A} . \tag{5.41}$$

Kod MANOVA sa jednim faktorom navedene sume kvadrata zamenjujemo odgovarajućim vrednostima generalizovanih varijansi odnosnih matrica sume kvadrata i uzajamanih proizvoda unutar i između tretmana. Drugim rečima odbacujemo H_0 ako je količnik generalizovanih varijansi (Wilksova lambda), u oznaci Λ:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} \tag{5.42}$$

mali broj. U izvesnim specifičnim slučajevima egzaktni raspored Wilksove lambde je poznat i naveden je u narednoj tabeli. U ostalim slučajevima koristi se jedna od aproksimacija od kojih je najpoznatija Bartlettova.

Tabela 5.1 Transformacija Wilksove lambde u statistiku

Broj promenljivih (p)	Broj populacija (g)	Uzorački raspored za višedimenzione normalne podatke
$p=1$	$g\geq 2$	$\frac{1-\Lambda}{\Lambda}\frac{n-g}{g-1}\sim F_{g-1,n-g}$
$p=2$	$g\geq 2$	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{n-g-1}{g-1}\sim F_{2(g-1),n-g-1}$
$p\geq 1$	$g\geq 2$	$\frac{1-\Lambda}{\Lambda}\frac{n-p-1}{p}\sim F_{p,n-p-1}$
$p\geq 1$	$g=3$	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{n-p-2}{p}\sim F_{2p,2(n-p-2)}$

Bartlett (1938) je pokazao da kada je tačna H_0 i kada $n\to\infty$, statistika

$$-\left(n-1-\frac{p+g}{2}\right)\ln\Lambda=-\left(n-1-\frac{p+g}{2}\right)\ln\left(\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}\right) \tag{5.43}$$

ima aproksimativno χ^2-raspored sa $p(g-1)$ stepeni slobode. Na osnovu ovog rasporeda odbacujemo H_0 na nivou značajnosti α ako je

$$-\left(n-1-\frac{p+g}{2}\right)\ln\left(\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}\right)>\chi^2_{p(g-1);\alpha} \tag{5.44}$$

gde je $\chi^2_{p(g-1);\alpha}$ gornji 100α percentile χ^2-rasporeda sa $p(g-1)$ stepeni slobode. Pored ove koriste se i druge statistike. Spomenimo neke: Raova statistika, Pillaijev trag, Royov najveći koren i Hotellingov trag. Prva od njih, Raova statistika, predstavlja transformaciju Wilksove lambde, i neki njeni specijalni slučajevi prikazani su u Tabeli 5.1. Dakle, neke od statistika predstavljaju transformaciju Wilksove lambde, dok su druge bazirane na matricama $\mathbf{B}$ i $\mathbf{W}$, tj. njihovim karakterističnim korenima (na primer, Royov najveći koren zasnovan je na najvećem karakterističnom korenu matrice $\mathbf{BW}^{-1}$) ili tragu (na primer, Hotellingov trag je zasnovan na tragu matrice $\mathbf{BW}^{-1}$). Međutim, i sama Wilksova lambda kao i ostale statistike testa mogu se iskazati preko karakterističnih korena matrice $\mathbf{BW}^{-1}$. Tako se može pokazati[3] da je Wilksova lambda

$$\Lambda=\prod_{j=1}^{p}\frac{1}{1+\theta_j} \tag{5.45}$$

[3] Videti na primer Chatfield (1980, s. 147).

gde smo sa θ_j, $j=1,2,...,p$ označili $j-$ti karakteristični koren matrice $\mathbf{BW}^{-1}$. Ispitivanja jačine različitih testova ukazuju da njihova jačina zavisi od tipa alternativne hipoteze, te da nema testa koji je superioran u odnosu na ostale. Korišćenjem računarskih statističkih programa među izlaznim rezultatima dobijamo vrednost svih navedenih statistika, kao i aproksimativnu vrednost $F-$statistike na osnovu njihove transformacije što nam omogućava izbor testa u zavisnosti od tipa alternativne hipoteze.

Primer 5.6 Na osnovu *Ankete o potrošnji domaćinstava* Zavoda za statistiku posmatrali smo uzorak od ukupno 24 domaćinstva (videti Primer 4.3). Kod svakog domaćinstva merili smo sledeće promenljive: $X_1=$broj članova domaćinstva, $X_2=$godišnja raspoloživa sredstva, $X_3=$izdaci na ishranu i $X_4=$izdaci na obrazovanje i razonodu. Na osnovu prezentiranih podataka ispitati da li između različitih tipova domaćinstava (poljoprivredno, nepoljoprivredno i mešovito domaćinstvo) postoji razlika u sredini.

U donjoj tabeli navedene su uzoračke sredine odnosnih promenljivih za svaki tip domaćinstva

k	Tip domaćinstva	Veličina uzorka	Sredina			
			X_1	X_2	X_3	X_4
1	Poljoprivredno	6	4.5000	11.0000	4.3000	0.2000
2	Nepoljoprivredno	10	3.5000	13.9800	4.5500	0.7111
3	Mešovito	8	3.7500	12.8125	4.6625	0.3763

Uzoračke kovarijacione matrice za svaki tip domaćinstva su

$$\mathbf{S}_1=\begin{bmatrix} 3.5 & -0.860 & 4.100 & -0.1200 \\ & 2.417 & -1.510 & 0.0520 \\ & & 5.152 & -0.1570 \\ & & & 0.00608 \end{bmatrix},\quad \mathbf{S}_2=\begin{bmatrix} 1.611 & -2.044 & 1.250 & -0.2360 \\ & 4.566 & -1.639 & 0.4750 \\ & & 1.629 & -0.2030 \\ & & & 0.07032 \end{bmatrix}.$$

$$\mathbf{S}_3=\begin{bmatrix} 2.786 & -2.411 & 0.575 & -0.110 \\ & 5.658 & -1.288 & 0.357 \\ & & 0.380 & -0.105 \\ & & & 0.0328 \end{bmatrix}.$$

Na osnovu ovih rezultata matrica sume kvadrata unutar tretmana je

$$\mathbf{W}=(n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2+(n_3-1)\mathbf{S}_3=\begin{bmatrix} 51.5 & -39.575 & -35.775 & -3.493 \\ & 92.785 & -31.316 & 7.041 \\ & & 43.084 & -3.342 \\ & & & 0.8929 \end{bmatrix}$$

ocena opšte uzoračke kovarijacione matrice je

$$\overline{\mathbf{S}} = \frac{\mathbf{W}}{24-3} = \begin{bmatrix} 2.452 & -1.885 & 1.704 & -0.166 \\ & 4.418 & -1.1491 & 0.335 \\ & & 2.052 & -0.159 \\ & & & 0.0425 \end{bmatrix}.$$

Kako je uzoračka opšta sredina

$$\overline{\mathbf{x}} = \frac{n_1\overline{\mathbf{x}}_1 + n_2\overline{\mathbf{x}}_2 + n_3\overline{\mathbf{x}}_3}{n_1 + n_2 + n_3} = \begin{bmatrix} 3.833 \\ 12.846 \\ 4.425 \\ 0.472 \end{bmatrix}$$

to na osnovu njega i izračunatih sredina vektora za svaki tip domaćinstva određujemo matricu sume kvadrata između tretmana

$$\mathbf{B} = \sum_{k=1}^{3} n_k (\overline{\mathbf{x}}_k - \overline{\mathbf{x}})(\overline{\mathbf{x}}_k - \overline{\mathbf{x}})' = \begin{bmatrix} 3.833 & -11.142 & -1.075 & -1.821 \\ & 33.315 & 2.739 & 5.749 \\ & & 0.461 & 0.322 \\ & & & 1.088 \end{bmatrix}.$$

Zahtev u zadatku možemo formulisati u vidu nulte hipoteze $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \boldsymbol{\mu}$, prema kojoj ne postoji razlika u sredini od jednog do drugog tipa domaćinstva. Na osnovu matrica sume kvadrata između i unutar tretmana izračunavamo matricu ukupne sume kvadrata, tj.

$$\mathbf{T} = \mathbf{B} + \mathbf{W} = \begin{bmatrix} 55.333 & -50.717 & 34.700 & -5.313 \\ & 126.100 & -28.578 & 12.789 \\ & & 43.545 & -3.020 \\ & & & 1.981 \end{bmatrix}.$$

Vrednost Wilksove lambde zavisi od generalizovanih varijansi matrica $\mathbf{W}$ i $\mathbf{T}$. Kako je $|\mathbf{W}| = 18983.2$ i $|\mathbf{T}| = 65191.5$, vrednost Wilksove lambde je $\Lambda = 0.2912$. Uzimajući u obzir da je $p = 4$ i $g = 3$ na osnovu rezultata datog u četvrtom redu Tabele 5.1, uz pretpostavljenu normalnost rasporeda, možemo odrediti egzaktan test za testiranje nulte hipoteze. Prema tom rezultatu vrednost statistike testa

$$\left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right)\frac{n-p-2}{p} - \frac{1-\sqrt{0.2912}}{\sqrt{0.2912}}\,\frac{24-4-2}{4} = 3.8392$$

poredimo sa gornjim 100α percentilom $F-$rasporeda sa 8 i 36 stepeni slobode. Za $\alpha = 0.05$ je $F_{8,36;0.05} = 2.2085 < 3.8392$, pa odbacujemo H_0 na nivou značajnosti od 5% i zaključujemo da postoje razlike u sredinama između domaćinstava s obzirom na njihov tip.

U cilju poređenja rezultata ovog testa i testa zasnovanog na Bartlettovoj aproksimaciji, izračunaćemo ovaj potonji

$$-\left(n-1-\frac{p+g}{2}\right)\ln\Lambda = -19.5\cdot\ln(0.2912) = 24.0586.$$

Poredeći izračunatu vrednost statistike testa sa kritičnom vrednošću $\chi^2_{8;0.05} = 15.5073$, konstatujemo da je izračunata vrednost veća od kritične što nam dozvoljava da zaključimo, da treba odbaciti H_0 na nivou značajnosti 5%. Dakle, dobili smo rezultat koji je u saglasnosti sa rezultatom prethodnog, egzaktnog testa. ■

Jedna od pretpostavki MANOVA odnosi se na jednakost kovarijacionih matrica populacija odakle su uzeti uzorci. Ova pretpostavka može biti kritična pogotovo ako se veličine uzoraka iz populacija znatno razlikuju. Stoga je od interesa proveriti njenu tačnost pre nego što se primeni MANOVA.

Testiramo hipotezu

$$H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_g, \tag{5.46}$$

gde je $\boldsymbol{\Sigma}_k$, $k = 1, 2, ..., g$ kovarijaciona matrica $k-$te populacije, pri čemu smo pretpostavili normalnost rasporeda svake od g populacija. Alternativna hipoteza je tvrđenje da postoji barem jedna kovarijaciona matrica ovih populacija koja se razlikuje od ostalih. Na osnovu principa količnika verodostojnosti definišemo statistiku testa za testiranje navedene hipoteze. Naime, statistika testa je

$$M = \frac{\prod_{k=1}^{g} |\mathbf{S}_k|^{\frac{n_k-1}{2}}}{|\overline{\mathbf{S}}|^{\frac{n-g}{2}}}, \tag{5.47}$$

gde smo sa $\mathbf{S}_k$ označili uzoračke kovarijacione matrice $k-$te populacije, a sa $\overline{\mathbf{S}}$ uzoračku ocenu opšte kovarijacione matrice, ako je tačna nulta hipoteza. Ova statistika je generalizacija Bartlettove statistike za testiranje homogenosti varijansi. U primeni ovog testa koristimo se jednom od aproksimacija rasporeda M statistike s obzirom na složenost njenog rasporeda.

Transformacijom M statistike prema izrazu $-2(1-c_1)\ln M$ dobićemo statistiku koja je aproksimativno raspoređena po χ^2-rasporedu sa $\frac{1}{2}p(p+1)(g-1)$ stepeni slobode. Pri tome je

$$c_1=\frac{2p^2+3p-1}{6(p+1)(g-1)}\left(\sum_{k=1}^{g}\frac{1}{n_k-1}-\frac{1}{n-g}\right). \tag{5.48}$$

Za veće vrednosti p i g i malo n_k predložena je aproksimacija zasnovana na $F-$rasporedu. Tako $-2b\ln M$ ima aproksimativno $F-$raspored sa ν_1 i ν_2 stepeni slobode, gde su

$$\nu_1=\tfrac{1}{2}p(p+1)(g-1),\ \nu_2=\frac{\nu_1+2}{\left|c_2-c_1^2\right|},\ b=\frac{1}{\nu_1}\left[\frac{1-c_1-\nu_1}{\nu_2}\right]$$

$$c_2=\frac{(p-1)(p+2)}{6(g-1)}\left(\sum_{k=1}^{g}\frac{1}{(n_k-1)^2}-\frac{1}{(n-g)^2}\right). \tag{5.49}$$

U slučaju da je $c_2-c_1^2<0$ koristimo statistiku $\frac{-2b_1\nu_2\ln M}{\nu_1+2b_1\nu_1\ln M}$ koja ima aproksimativno $F-$raspored sa ν_1 i ν_2 stepeni slobode, gde je $b_1=\frac{1-c_1-(2/\nu_2)}{\nu_2}$. Aproksimacija bazirana na $F-$rasporedu bolja je u poređenju sa χ^2 aproksimacijom.

Primer 5.7 Na osnovu podataka iz Primera 5.6 testirati hipotezu da su kovarijacione matrice za tri tipa domaćinstva međusobom jednake.

Generalizovane varijanse uzoračkih kovarijacionih matrica su

$$\left|\mathbf{S}_1\right|=0.00219,\ \left|\mathbf{S}_2\right|=0.04128,\ \left|\mathbf{S}_3\right|=0.00215,\ \left|\overline{\mathbf{S}}\right|=0.09761.$$

Prema tome, vrednost M statistike iznosi $2.49542\cdot10^{-12}$. Odredimo vrednost konstante c_1 za primenu χ^2 aproksimacije. Kako je $p=4$, $g=3,\ n_1=6$, to je . Transformisana vrednost M statistike je $-2(1-c_1)\ln M=29.3081$. Ako je tačna H_0 ova statistika ima χ^2-rasporeda sa 20 stepeni slobode. Kritična vrednost na nivou značajnosti 5% iznosi $\chi^2_{20;0.05}=31.4104$. Kako je izračunata vrednost statistike testa manja od kritične vrednosti prihvatamo hipotezu o jednakosti kovarijacionih matrica ove tri populacije. Ovime je pokazano da smo opravdano koristili pretpostavku o jednakim kovarijacionim matricama u prethodnom primeru.

Skrećemo pažnju da je izračunata vrednost bliska kritičnoj, a realizovani nivo značajnosti iznosi 0.082 (to je najmanji nivo značajnosti sa kojim se nulta hipoteza može odbaciti na osnovu podataka iz uzorka). Znači da bi se odbacila nulta hipoteza na nivou značajnosti od 10% (kritična vrednost je $\chi^2_{20;0.10} = 28.412$). Do sličnog rezultata došli bismo primenom F aproksimacije.

Ilustrujmo korišćenje F aproksimacije u ovom primeru. Pored izračunate vrednosti konstante $c_1 = 0.4515$ odredimo i ostale potrebne elemente: $c_2 = 0.1057$, $\nu_1 = 20$, $\nu_2 = 224.211$ Kako je $c_2 - c_1^2 < 0$ i $b_1 = 0.00241$ izračunata vrednost statistike $\frac{-2b_1\nu_2 \ln M}{\nu_1 + 2b_1\nu_1 \ln M}$ je 1.6543, a statistika ima aproksimativno $F-$raspored sa 20 i 224 stepeni slobode. Kritična vrednost testa na nivou značajnosti $\alpha = 0.05$ iznosi $F_{20,224;0.05} = 1.6176$. Kako je izračunata vrednost veća od kritične vrednosti, na nivo značajnosti od 5% odbacujemo hipotezu o jednakosti kovarijacionih matrica tri populacije. Realizovani nivo značajnosti ovog testa je 0.0425. S obzirom da je F aproksimacija bolja u poređenju sa χ^2 aproksimacijom, to bi se sa više pouzdanja oslonili na rezultate dobijene korišćenjem testa zasnovanog na $F-$rasporedu. ■

Na kraju ovog odeljka ukažimo na interesantnu interpretaciju Wilksove lambde. Kao što je poznato iz regresione analize i jednodimenzione analize varijanse, koeficijent determinacije (R^2) meri jačinu veze između zavisne i nezavisnih promenljivih. Definišemo ga na sledeći način

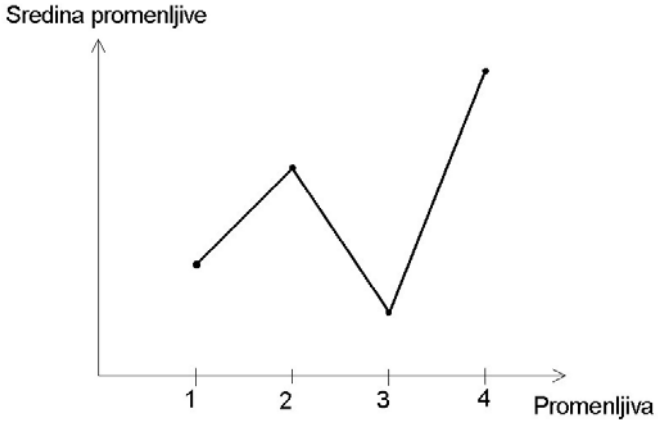
$$R^2 = 1 - \frac{\text{suma kvadrata reziduala}}{\text{ukupna suma kvadrata}}. \tag{5.50}$$

U višedimenzionom slučaju meru analognu R^2 definišemo sa $1-\Lambda$, gde je $\Lambda = |\mathbf{W}| / |\mathbf{B} + \mathbf{W}|$. Znači da Wilksova lambda meri udeo generalizovane varijanse unutar tretmana u ukupnoj generalizovanoj varijansi. Prema tome, što je veće $1-\Lambda = 1 - |\mathbf{W}| / |\mathbf{T}|$ veći je deo ukupne generalizovane varijanse koji se može pripisati varijacijama između grupa, odn. tretmana. U našem Primeru 5.5 veličina $1-\Lambda$ iznosi 0.7088, pa je oko 70% ukupne varijacije između opservacija rezultat varijacija između tri sredine. Pri korišćenju mere $1-\Lambda$, treba imati u vidu da je ona pozitivno pristrasna, odnosno precenjuje pravi stepen zavisnosti.

5.5 ANALIZA PROFILA

Pretpostavimo eksperimentalnu situaciju u kojoj registrujemo opservacije za p promenljivih kod eksperimentalnih jedinica koje pripadaju jednoj od g grupa. Ukoliko

je u pitanju potpuno slučajan plan tada ćemo pretpostaviti i da se p promenljivih mogu izraziti u uporedivim jedinicama mere. Ako su podaci proistekli iz eksperimenta zasnovanog na planu ponovljenih merenja, znači da smo na istim eksperimentalnim jedinicama merili odzive pod p različitih uslova (tretmana). Dakle, u svakoj od g grupa, svaki od n_k, $k = 1, 2, ..., g$ ispitanika podvrgnut je dejstvu p tretmana. Kod plana ponovljenih merenja svi odzivi za različite tretmane odnose se na istu promenljivu, pa su iskazani u istim jedinicama mere. Ovde nas može interesovati kako poređenje sredina p tretmana za ovih g grupa, tako i poređenje sredina g grupa za ovih p tretmana, što znači da analizu profila možemo da vršimo i po grupama i po tretmanima. Sa stanovišta razumevanja suštine analize profila svejedno je koji ćemo eksperimentalni plan razmatrati. Opredelili smo se za potpuno slučajan plan, a jednostavnosti radi zadržaćemo se na slučaju dve grupe.



Slika 5.1 Profil za četiri promenljive

Neka smo kod dve grupe ispitanika $(g = 2)$, formirane na osnovu obeležja pola (muško-žensko), beležili odgovore na p pitanja u vezi određenog proizvoda, pri čemu su odgovori iskazani na ordinarnoj skali od 0 do 9. Ispitanike iz ove dve grupe smo na potpuno slučajan način birali iz muške i ženske populacije. Tada se može postaviti standardno pitanje u analizi varijanse, da li su sredine ove dve populacije jednake, ili jednostavnije rečeno, da li je stav prema određenom proizvodu muškog i ženskog dela populacije jednak. Pre nego što odgovorimo na postavljeno pitanje, ukažimo na grafički prikaz koji koristimo u daljoj analizi.

Ako grafički prikažemo komponente sredine $\boldsymbol{\mu}$ jedne od grupa počevši od prvog do $p-$tog elementa i nacrtane tačke spojimo, tada smo dobili *profil* odgovora (promenljivih) za n eksperimentalnih jedinica. Kod plana ponovljenih merenja imamo profil promenljivih i profil grupa. Na Slici 5.1 ilustrovan je primer profila za sredinu sa četiri komponente.

Unošenjem na grafikon odgovarajućeg profila za svaku grupu nameću nam se, na osnovu međusobnog odnosa profila ove dve grupe, sledeća pitanja preko kojih postepeno formulišemo pitanje o jednakosti sredina. Ta pitanja glase:

1. Da li su profili paralelni?
2. Pretpostavivši da su profili zaista paralelni, da li su istovremeno i podudarni?
3. Pretpostavivši da se profili zaista podudaraju, da li su profili na istom nivou?

Da bismo odgovorili na ova pitanja i definisali statistički postupak za proveru ovih hipoteza, eksplicitno ćemo navesti pretpostavke na kojima je bazirana analiza profila. Odzivi eksperimentalnih jedinica predstavljaju $p-$dimenzioni normalno raspoređen slučajan vektor. Sredinu prve grupe označimo sa $\boldsymbol{\mu}_1' = [\mu_{11}, \mu_{12}, ..., \mu_{1p}]$, a druge sa $\boldsymbol{\mu}_2' = [\mu_{21}, \mu_{22}, ..., \mu_{2p}]$. Za obe grupe pretpostavićemo da imaju identičnu, nepoznatu kovarijacionu matricu $\boldsymbol{\Sigma}$. Neka su nezavisni slučajni uzorci veličine n_1 i n_2 uzeti iz ove dve populacije (grupe). Testiramo nultu hipotezu $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, a u analizi profila nju razlažemo na sledeći niz hipoteza u skladu sa gore postavljenim pitanjima.

Paralelnost profila iskazujemo u prvom koraku preko nulte hipoteze

$$H_{01} : \mu_{1j} - \mu_{1j+1} = \mu_{2j} - \mu_{2j+1}, \quad j = 1, 2, ..., p-1. \tag{5.51}$$

Korišćenjem matrice kontrasta $\mathbf{C}$ nultu hipotezu pišemo u obliku

$$H_{01} : \mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2, \tag{5.52}$$

gde je matrica kontrasta

$$\underset{((p-1)\times p)}{\mathbf{C}} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}. \tag{5.53}$$

Na osnovu promenljivih koje su dobijene linearnom transformacijom originalnih promenljivih $\mathbf{C}\mathbf{X}_{1i}$, $i = 1, 2, ..., n_1$ i $\mathbf{C}\mathbf{X}_{2i}$, $i = 1, 2, ..., n_2$ definišemo statistiku testa za testiranje nulte hipoteze bazirajući se na rasporedu linearne transformacije promenljivih (Poglavlje 5.3). Naime, navedene linearne transformacije imaju normalan raspored $N_{p-1}(\mathbf{C}\boldsymbol{\mu}_k, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')$, $k = 1, 2$, pa na osnovu toga statistika

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' \mathbf{C}' (\mathbf{C}\bar{\mathbf{S}}\mathbf{C}')^{-1} \mathbf{C} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \tag{5.54}$$

ima T^2-raspored, gde je $\overline{\mathbf{S}}$ uzoračka kovarijaciona matrica združenih uzoraka. Nultu hipotezu o paralelnosti profila odbacujemo na nivou značajnosti α ako je

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\mathbf{C}'(\mathbf{C}\overline{\mathbf{S}}\mathbf{C}')^{-1}\mathbf{C}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) > \frac{(n_1+n_2-2)(p-1)}{n_1+n_2-p}F_{p-1,n_1+n_2-2;\alpha}\,. \tag{5.55}$$

U narednom koraku, ako je prethodno prihvaćena hipoteza o paralelnosti profila, možemo odgovoriti na drugo pitanje, odnosno testirati hipotezu o podudarnosti profila. Ako su profili paralelni, tada je moguće da je prvi profil iznad drugog, tj. da je $\mu_{1j} > \mu_{2j}$, za svako j, ili obratno. Podudarnost profila imaćemo u slučaju da je ukupna visina jednog i drugog profila iskazana preko zbira pojedinih komponenti odgovarajućih sredina vektora jednaka, tj. kada je $\mu_{11}+\mu_{12}+...+\mu_{1p} = \mathbf{1}'\boldsymbol{\mu}_1$ i $\mu_{21}+\mu_{22}+...+\mu_{2p} = \mathbf{1}'\boldsymbol{\mu}_2$ jednako međusobom, gde je sa $\mathbf{1}$ označen $p-$dimenzioni vektor jedinica, $\mathbf{1}' = [1,1,...,1]$. Znači da je nulta hipoteza u drugom koraku

$$H_{02}: \mathbf{1}'\boldsymbol{\mu}_1 = \mathbf{1}'\boldsymbol{\mu}_2\,. \tag{5.56}$$

U našem primeru sa dve grupe statistika testa za testiranje navedene nulte hipoteze svodi se na primenu običnog $t-$testa na testiranje jednakosti zbira svih promenljivih odziva u dva uzorka. Nultu hipotezu o podudarnosti dva profila odbacujemo na nivou značajnosti α ako je

$$T^2 = \frac{n_1 n_2}{n_1+n_2}\left(\frac{\mathbf{1}'(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)}{\sqrt{\mathbf{1}'\overline{\mathbf{S}}\mathbf{1}}}\right)^2 > t^2_{n_1+n_2-2;\alpha/2} = F_{1,n_1+n_2-2;\alpha}\,. \tag{5.57}$$

Ako raspolažemo apriornom informacijom ili nam eksperimentalni podaci sugerišu da je jedan od profila viši u odnosu na drugi, tada se koristi jednostrani $t-$test.

Naposletku, pod pretpostavkom da su profili podudarni, možemo odgovoriti na poslednje pitanje, da li su profili na istom nivou. Ovu pretpostavku formulišemo u vidu sledeće hipoteze

$$H_{03}: \mu_{11} = \mu_{12} = ... = \mu_{1p} = \mu_{21} = \mu_{22} = ... = \mu_{2p}\,. \tag{5.58}$$

Korišćenjem matrice kontrasta ovu hipotezu pišemo u obliku

$$H_{03}: \mathbf{C}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \mathbf{0}\,. \tag{5.59}$$

Da bismo testirali navedenu hipotezu odredimo realizovanu vrednost ocene opšte sredine na osnovu izraza . Nultu hipotezu odbacujemo na nivou značajnosti α ako je

$$T^2 = (n_1+n_2)\overline{\mathbf{x}}'\mathbf{C}'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}\mathbf{C}\overline{\mathbf{x}} > F_{p-1,n_1+n_2-p;\alpha}\,. \tag{5.60}$$

Analiza profila za slučaj više od dve grupe izložena je kod Morrisona (1976, ss. 207-16).

Primer 5.8 U nameri da se ispita da li je stav muškog i ženskog dela populacije isti u pogledu određenog proizvoda sprovedena je anketa među muškarcima i ženama. Na slučaj su izabrana po 20 ispitanika iz dve populacije čiji su odgovori izraženi na ordinarnoj skali od 0 (najniža ocena) do 9 (najviša ocena). Ispitanici su odgovarali na sledeća pitanja:

X_1 = Kako ocenjujete hranljivu vrednost ovog proizvoda?

X_2 = Kako ocenjujete ukus ovog proizvoda?

X_3 = Kako ocenjujete pakovanje ovog proizvoda?

X_4 = Kako ocenjujete cenu ovog proizvoda s obzirom na njegov kvalitet?

Rezultati ankete prezentirani su u donjoj tabeli.

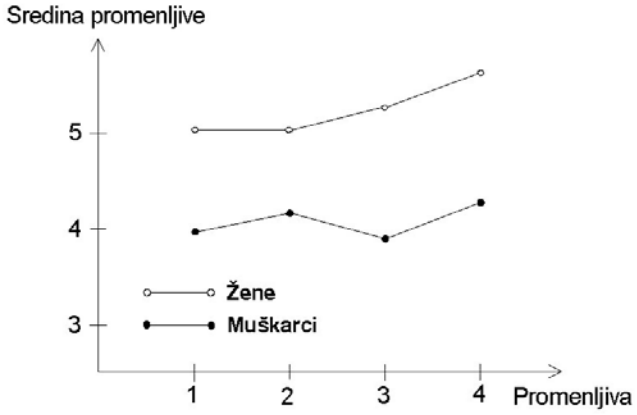
Redni broj	Muškarci				Žene			
	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
1	8	9	7	8	2	1	2	3
2	5	3	4	4	8	9	7	9
3	7	6	9	5	6	4	8	5
4	8	8	5	8	5	6	8	7
5	0	1	3	2	5	5	5	7
6	0	4	2	1	5	7	8	8
7	8	7	5	6	4	4	8	7
8	6	5	4	4	0	1	0	3
9	0	2	2	1	8	7	6	6
10	4	2	1	6	5	3	6	5
11	2	3	5	4	7	6	7	7
12	4	5	3	2	3	1	4	3
13	2	3	4	3	3	5	3	4
14	5	4	4	6	2	3	0	3
15	7	3	8	9	7	8	9	5
16	1	0	2	0	9	5	6	7
17	2	3	1	3	6	7	4	6
18	3	2	4	1	5	7	5	5
19	6	9	4	8	7	6	5	8
20	1	3	0	4	4	6	4	4

Na osnovu odgovora anketiranih ispitanika izračunali smo uzoračke sredine i uzoračke kovarijacione matrice:

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 3.95 \\ 4.10 \\ 3.85 \\ 4.25 \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 8.1553 & 5.5316 & 4.6763 & 6.2763 \\ & 6.5158 & 2.8579 & 4.6579 \\ & & 5.2921 & 3.4079 \\ & & & 7.2500 \end{bmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 5.05 \\ 5.05 \\ 5.25 \\ 5.60 \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 5.3132 & 3.9447 & 4.1974 & 3.1789 \\ & 5.4184 & 3.5658 & 3.0211 \\ & & 6.7237 & 3.3158 \\ & & & 3.5158 \end{bmatrix}$$

Na donjoj slici prikazani su profili odgovora muške i ženske populacije. Na osnovu grafičkog prikaza može se uočiti da pretpostavka o paralelnosti profila ima osnova. Pre nego što to formalno proverimo testirajući hipotezu o paralelnosti profila, izračunajmo opštu sredinu i kovarijacionu matricu.



Opšta uzoračka sredina i uzoračka kovarijaciona matrica su:

$$\bar{\mathbf{x}} = \begin{bmatrix} 4.500 \\ 4.575 \\ 4.550 \\ 4.925 \end{bmatrix}, \quad \bar{\mathbf{S}} = \begin{bmatrix} 6.7342 & 4.7382 & 4.4368 & 4.7276 \\ & 5.9671 & 3.2118 & 3.8395 \\ & & 6.0079 & 3.3618 \\ & & & 5.3829 \end{bmatrix}$$

U prvom koraku testiramo hipotezu da su profili paralelni, pa je nulta hipoteza

$$H_{01} : \mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2 ,$$

gde je matrica kontrasta

$$\mathbf{C} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

Na osnovu nje, a prema izrazu statistike testa za testiranje paralelnosti dva profila, odredimo potrebne elemente za primenu testa:

$$\mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \begin{bmatrix} 0.15 \\ -0.45 \\ 0.05 \end{bmatrix}, \quad \mathbf{C}\bar{\mathbf{S}}\mathbf{C}' = \begin{bmatrix} 3.2250 & -2.4539 & 0.3368 \\ & 5.5513 & -3.2737 \\ & & 4.6671 \end{bmatrix}.$$

Odredimo vrednost statistike za testiranje nulte hipoteze (paralelnost profila):

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{C}'(\mathbf{C}\bar{\mathbf{S}}\mathbf{C}')^{-1}\mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = .$$

$$= \frac{20\cdot 20}{20+20}\begin{bmatrix} 0.15 & -0.45 & 0.05 \end{bmatrix}\begin{bmatrix} 0.5918 & 0.4032 & 0.2401 \\ & 0.5819 & 0.3791 \\ & & 0.4628 \end{bmatrix}\begin{bmatrix} 0.15 \\ -0.45 \\ 0.05 \end{bmatrix}.$$

Izračunata vrednost statistike testa je $T^2 = 0.6442$. Kako je $F_{3,36;0.05} = 2.8663$ to je kritična vrednost $T^2 -$ statistike jednaka

$$T^2 = \frac{(n_1+n_2-2)(p-1)}{n_1+n_2-p} F_{p-1,n_1+n_2-p;\alpha} = \frac{(20+20-2)(4-1)}{20+20-4} F_{3,36;0.05} = 9.0765.$$

Izračunata vrednost statistike testa manja je od kritične vrednosti, pa se prihvata nulta hipoteza o paralelnosti dva profila.

Pošto je, kao što je očekivano, prihvaćena hipoteza o paralelnosti profila prelazimo na naredan korak u proveri jednakosti sredine ove dve populacije. Pretpostavka koju proveravamo jeste da su profili podudarni, što znači da su odgovarajući elementi dve sredine jednaki međusobom. Prema izrazima koji figurišu u statistici testa izračunavamo još zbir elemenata razlike sredina i zbir elemenata opšte uzoračke kovarijacione matrice

$$\mathbf{1}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} -1.10 \\ -0.95 \\ -1.40 \\ -1.35 \end{bmatrix} = -4.8$$

$$\mathbf{1}'\bar{\mathbf{S}}\mathbf{1} = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 6.7342 & 4.7382 & 4.4368 & 4.7276 \\ & 5.9671 & 3.2118 & 3.8395 \\ & & 6.0079 & 3.3618 \\ & & & 5.3829 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Izračunata vrednost T^2-statistike je

$$T^2=\left(\frac{20\cdot 20}{20+20}\right)\left(\frac{-4.8}{\sqrt{72.724}}\right)^2=3.1682.$$

Na nivou značajnosti $\alpha=0.05$ kritična vrednost testa je $F_{1,38;0.05}=4.0982$. Kako je izračunata vrednost statistike testa manja od kritične vrednosti prihvatamo hipotezu o podudarnosti profila.

Naposletku, pošto su profili podudarni, testiramo hipotezu da su nivoi ova dva profila jednaki, što znači da se dva profila poklapaju jer su im sve komponente dve sredine jednake međusobom. Odredimo vrednost sledećeg izraza

$$\mathbf{C}\bar{\mathbf{x}}=\begin{bmatrix}1 & -1 & 0 & 0\\ 0 & 1 & -1 & 0\\ 0 & 0 & 1 & -1\end{bmatrix}\begin{bmatrix}-1.10\\ -0.95\\ -1.40\\ -1.35\end{bmatrix}=\begin{bmatrix}0.075\\ -0.025\\ 0.375\end{bmatrix}.$$

Izračunata vrednost T^2-statistike je

$$T^2=(20+20)\begin{bmatrix}0.075 & -0.025 & 0.375\end{bmatrix}\begin{bmatrix}0.5918 & 0.4032 & 0.2401\\ & 0.5819 & 0.3791\\ & & 0.4628\end{bmatrix}\begin{bmatrix}0.075\\ -0.025\\ 0.375\end{bmatrix}.$$

Na nivou značajnosti $\alpha=0.05$ kritična vrednost testa je $F_{3,36;0.05}=2.8663$. Kako je izračunata vrednost statistike testa veća od kritične, odbacujemo nultu hipotezu da se profili poklapaju na nivou značajnosti od 5%.

Striktno posmatrajući, teoriju normalnog rasporeda ne bismo mogli primeniti u ovom primeru pošto su opservacije celi brojevi, odnosno promenljive su iskazane na ordinarnoj skali, a uzorci su relativno mali. Ovo predstavlja delimično objašnjenje dobijenih, pomalo protivurečnih rezultata. Naime, izračunata vrednost statistike testa za testiranje hipoteze o jednakosti sredine ove dve populacije iznosi $T^2=5.096$, što je manje od odgovarajuće kritične vrednosti (11.472), na nivou značajnosti $\alpha=0.05$. Znači da nam se sugeriše prihvatanje hipoteze o jednakosti dve sredine za muški i ženski deo populacije. ■

U opštem slučaju analiza profila omogućava nam analizu nejednakosti sredina u tom smislu što se razlika među njima može pripisati ili obliku ili nivou profila.

5.6 M A N O V A SA DVA FAKTORA

U primeru koji smo izložili kod MANOVA sa jednim faktorom pretpostavili smo da postoji uticaj tipa prodavnice na opredeljenje kupaca pri kupovini određene robe. Ako proširimo primer posmatrajući pored uticaja tipa prodavnice i uticaj regionalne lokacije prodavnica na opredeljenje kupaca, suočavamo se sa situacijom za koju je primeren model analize varijanse sa dva faktora. Drugi faktor u ovoj eksperimentalnoj situaciji jeste region, a broj njegovih nivoa, odn. tretmana, predstavlja broj region koje smo uključili u analizu. Sve ostalo što je u uvodu kod MANOVA sa jednim faktorom u opisu eksperimentalne situacije rečeno o prvom kontrolisanom faktoru važi i za drugi faktor. Na osnovu odziva eksperimentalnih jedinica za svaku kombinaciju različitih nivoa dva faktora u prilici smo da utvrdimo koji se deo varijacija može pripisati uticaju tipa prodavnice, a koji uticaju njihovog regionalnog razmeštaja. Takođe nas interesuje da li postoji interakcija između tipa prodavnice i regiona. Ona bi se iskazala tako što bi na primer prodaja u jednom regionu bila viša u klasičnoj prodavnici u odnosu na ostale tipove prodavnica, dok bi u ostalim regionima prodaja bila približno jednaka kod sva tri tipa prodavnice. Na ovaj primer vratićemo se kasnije, nakon definisanja odgovarajućeg modela i postupka statističkog zaključivanja.

5.6.1 Model MANOVA sa dva faktora

Pretpostavimo da prvi eksperimentalni faktor ima g, a drugi b nivoa i da smo za svaku od gb kombinacija nivoa dva faktora prikupili n opservacija $p-$dimenzionog slučajnog vektora $\mathbf{X}$. Pretpostavimo da je tih gbn nezavisnih opservacija generisano modelom

$$\mathbf{X}_{ikl} = \boldsymbol{\mu} + \boldsymbol{\alpha}_k + \boldsymbol{\beta}_l + \boldsymbol{\gamma}_{kl} + \boldsymbol{\varepsilon}_{ikl}, \; i = 1,2,...,n, \; k = 1,2,...,g, \; l = 1,2,...,b, \tag{5.61}$$

gde su slučajne greške normalno raspoređene $\boldsymbol{\varepsilon}_{ikl} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ je opšta sredina, $\boldsymbol{\alpha}_k$ je efekat $k-$tog nivoa prvog faktora, $\boldsymbol{\beta}_l$ je efekat $l-$tog nivoa drugog faktora, a $\boldsymbol{\gamma}_{kl}$ je efekat interakcije između $k-$tog i $l-$tog nivoa prvog, odnosno drugog faktora respektivno. Svi efekti predstavljaju $p-$dimenzione vektore koji zadovoljavaju sledeće uslove:

$$\sum_{k=1}^{g} \boldsymbol{\alpha}_k = \sum_{l=1}^{b} \boldsymbol{\beta}_l = \sum_{k=1}^{g} \boldsymbol{\gamma}_{kl} = \sum_{l=1}^{b} \boldsymbol{\gamma}_{kl} = 0. \tag{5.62}$$

Prisustvo komponente $\boldsymbol{\gamma}_{kl}$, interakcije između prvog i drugog faktora, u gornjem modelu, čini da faktori u modelu MANOVA sa dva faktora nisu više aditivni. Ova osobina modela sa svoje strane otežava interpretaciju rezultata analize. Ukoliko je broj opservacija po svakoj od gb kombinacija nivoa dva faktora jednak jedinici $(n = 1)$ model MANOVA sa dva faktora ne omogućava ispitivanje interakcije faktora. Takav

model nazivamo model MANOVA bez ponavljanja. U suprotnom slučaju kada je $n>1$, kažemo da se radi o modelu MANOVA sa ponavljanjem.

Na osnovu vrednosti slučajnog uzorka od po n opservacija za svaku kombinaciju nivoa dva faktora, možemo $i-$tu opservaciju razložiti na pet delova

$\mathbf{X}_{ikl} =$	$\bar{\mathbf{X}} +$	$(\bar{\mathbf{X}}_{k\cdot} - \bar{\mathbf{X}}) +$	$(\bar{\mathbf{X}}_{\cdot l} - \bar{\mathbf{X}}) +$	$(\bar{\mathbf{X}}_{kl} - \bar{\mathbf{X}}_{k\cdot} - \bar{\mathbf{X}}_{\cdot l} - \bar{\mathbf{X}}) +$	$(\mathbf{X}_{ikl} - \bar{\mathbf{X}}_{kl})$
opservacija	ocena opšte sredine	ocenjeni efekat 1. faktora	ocenjeni efekat 2. faktora	ocenjeni efekat interakcije 1. i 2. faktora	ocenjeni rezidual

gde je $\bar{\mathbf{X}}$ ocena opšte sredine, $\bar{\mathbf{X}}_{k\cdot}$ ocena sredine na $k-$tom nivou prvog faktora, $\bar{\mathbf{X}}_{\cdot l}$ ocena sredine $l-$tog nivoa drugog faktora i $\bar{\mathbf{X}}_{kl}$ je ocena sredine na $k-$tom nivou prvog i $l-$tom nivou drugog faktora, odnosno

$$\bar{\mathbf{X}} = \frac{1}{gbn}\sum_{i=1}^{n}\sum_{k=1}^{g}\sum_{l=1}^{b} X_{ikl} \qquad \bar{\mathbf{X}}_{k\cdot} = \frac{1}{bn}\sum_{i=1}^{n}\sum_{l=1}^{b} X_{ikl} \tag{5.63a}$$

$$\bar{\mathbf{X}}_{\cdot l} = \frac{1}{gn}\sum_{i=1}^{n}\sum_{k=1}^{g} X_{ikl} \qquad \bar{\mathbf{X}}_{kl} = \frac{1}{n}\sum_{i=1}^{n} X_{ikl} \tag{5.63b}$$

Prema tome, $\hat{\boldsymbol{\alpha}}_k = \bar{\mathbf{X}}_{k\cdot} - \bar{\mathbf{X}}$ je ocenjeni efekat $k-$tog nivoa prvog faktora, $\hat{\boldsymbol{\beta}}_l = \bar{\mathbf{X}}_{\cdot l} - \bar{\mathbf{X}}$ je ocenjeni efekat $l-$tog nivoa drugog faktora, $\hat{\boldsymbol{\gamma}}_{kl} = \bar{\mathbf{X}}_{kl} - \bar{\mathbf{X}}_{k\cdot} - \bar{\mathbf{X}}_{\cdot l} + \bar{\mathbf{X}}$ je ocenjeni efekat interakcije $k-$tog nivoa prvog i $l-$tog nivoa drugog faktora, a $\hat{\boldsymbol{\varepsilon}}_{ikl} = \mathbf{X}_{ikl} - \bar{\mathbf{X}}_{kl}$ je ocenjeni rezidual (greška) modela MANOVA sa dva faktora.

Na liniji izlaganja modela MANOVA sa jednim faktorom, a u skladu sa gornjim razlaganjem opservacija vršimo razlaganje matrice ukupne sume kvadrata i unakrsnih proizvoda odstupanja opservacija od opšte sredine, u oznaci $\mathbf{T}$, kao i odgovarajućeg broja stepeni slobode:

$\mathbf{T}$	$=$	$\mathbf{G}$	$+$	$\mathbf{B}$	$+$	$\mathbf{W}$	$+$	$\mathbf{E}$
$gbn-1$		$(g-1)$		$(b-1)$		$(g-1)(b-1)$		$gb(n-1)$

Sa $\mathbf{G}$ smo označili matricu sume kvadrata i uzajamnih proizvoda odstupanja između sredina nivoa prvog faktora i opšte sredine

$$\mathbf{G} = \sum_{k=1}^{g} bn(\bar{\mathbf{X}}_{k\cdot} - \bar{\mathbf{X}})(\bar{\mathbf{X}}_{k\cdot} - \bar{\mathbf{X}})'. \tag{5.64}$$

Sa $\mathbf{B}$ smo označili matricu sume kvadrata i uzajamnih proizvoda odstupanja između sredina nivoa drugog faktora i opšte sredine

$$\mathbf{B} = \sum_{l=1}^{b} gn(\bar{\mathbf{X}}_{\cdot l} - \bar{\mathbf{X}})(\bar{\mathbf{X}}_{\cdot l} - \bar{\mathbf{X}})' . \tag{5.65}$$

Sa $\mathbf{W}$ smo označili matricu sume kvadrata i uzajamnih proizvoda interakcija između nivoa prvog i drugog faktora

$$\mathbf{W} = \sum_{k=1}^{g}\sum_{l=1}^{b} n(\bar{\mathbf{X}}_{kl} - \bar{\mathbf{X}}_{k\cdot} - \bar{\mathbf{X}}_{\cdot l} + \bar{\mathbf{X}})(\bar{\mathbf{X}}_{kl} - \bar{\mathbf{X}}_{k\cdot} - \bar{\mathbf{X}}_{\cdot l} + \bar{\mathbf{X}})' . \tag{5.66}$$

Sa $\mathbf{E}$ smo označili matricu sume kvadrata i uzajamnih proizvoda ocenjenih grešaka modela

$$\mathbf{E} = \sum_{i=1}^{n}\sum_{k=1}^{g}\sum_{l=1}^{b} (\mathbf{X}_{ikl} - \bar{\mathbf{X}}_{kl})(\mathbf{X}_{ikl} - \bar{\mathbf{X}}_{kl})' . \tag{5.67}$$

5.6.2 Zaključivanje kod MANOVA sa dva faktora

S obzirom da izlažemo model dvofaktorske MANOVA sa interakcijom, postupak statističkog zaključivanja uslovljen je ovom činjenicom. Naime, pre nego što pristupimo testiranju hipoteze o uticaju prvog ili drugog faktora na zavisnu promenljivu potrebno je proveriti da li je pretpostavka o odsustvu interakcije između faktora opravdana. Testiramo nultu hipotezu

$$H_0 : \boldsymbol{\gamma}_{kl} = \mathbf{0} ,\ k = 1, 2, ..., g ,\ l = 1, 2, ..., b \tag{5.68}$$

protiv alternativne hipoteze da bar za jednu kombinaciju nivoa prvog i drugog faktora postoji interakcija. Statistika testa konstruisana na principu količnika verodostojnosti. Wilksova lambda

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{W} + \mathbf{E}|} \tag{5.69}$$

sugeriše svojim malim vrednostima da treba odbaciti hipotezu o odsustvu interakcije među faktorima. Za velike uzorke koristimo Bartlettovu aproksimaciju rasporeda statistike Λ prema kojoj odbacujemo H_0 na nivou značajnosti α ako je

$$-\left[gb(n-1) - \frac{p + 1 - (g-1)(b-1)}{2} \right] \ln \Lambda > \chi^2_{(g-1)(b-1)\,p;\alpha} . \tag{5.70}$$

Ukoliko smo odbacili hipotezu o odsustvu interakcije između faktora tada se ne preporučuje korišćenje dalje multivarijacionog postupka testiranja uticaja faktora. Umesto toga ANOVA sa dva faktora za svaku od p promenljivih treba da odgovori na pitanja o prisustvu interakcije između pojedinih odziva. Oni odzivi za koje se utvrdi

odsustvo interakcije između faktora mogu se interpretirati u smislu njihovog aditivnog efekta, ako on postoji. Stoga ćemo u daljem izlaganju statističkog postupka pretpostaviti da je hipoteza o odsustvu interakcije prihvaćena.

Sada možemo testirati hipotezu da postoji jednakost efekata različitog nivoa prvog faktora

$$H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = ... = \boldsymbol{\alpha}_g = \mathbf{0} \tag{5.71}$$

protiv alternativne hipoteze da postoji barem jedan nivo prvog faktora čiji je efekat različit od nula vektora. Wilksova lambda, u oznaci Λ_1

$$\Lambda_1 = \frac{|\mathbf{E}|}{|\mathbf{G} + \mathbf{E}|} \tag{5.72}$$

uzima male vrednosti kada H_0 nije tačna. Bartlettova aproksimacija sugeriše da se odbaci nulta hipoteza o odsustvu uticaja prvog faktora na zavisnu promenljivu na nivou značajnosti α ako je

$$-\left[gb(n-1) - \frac{p+1-(g-1)}{2} \right] \ln \Lambda_1 > \chi^2_{(g-1)p;\alpha} . \tag{5.73}$$

Na sličan način donosimo odluku o tome da li treba prihvatili ili odbaciti hipotezu

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = ... = \boldsymbol{\beta}_b = \mathbf{0} \tag{5.74}$$

o odsustvu efekata drugog faktora. Wilksova lambda, u oznaci Λ_2, u ovom slučaju je

$$\Lambda_2 = \frac{|\mathbf{E}|}{|\mathbf{B} + \mathbf{E}|} \tag{5.75}$$

i njene male vrednosti sugerišu da treba odbaciti H_0. Bartlettovu aproksimaciju koristimo u velikim uzorcima da bismo na osnovu nje odbacili nultu hipotezu na nivou značajnosti α ako je

$$-\left[gb(n-1) - \frac{p+1-(b-1)}{2} \right] \ln \Lambda_2 > \chi^2_{(b-1)p;\alpha} . \tag{5.76}$$

Kao i kod prethodnih statističkih modela i kod modela MANOVA sa dva faktora moguće je dalje analizirati prirodu efekata faktora određivanjem simultanih intervala poverenja za kontraste parametara modela. Ti intervali bazirani su na rasporedu najvećeg karakterističnog korena. Detaljnije o simultanim intervalima poverenja u ovom kontekstu može se pročitati kod Morrisona (1976, ss. 197-204).

Primer 5.9 Prodaja gaziranog i negaziranog osvežavajućeg pića obavlja se u tri tipa prodavnica: klasična prodavnica, samoposluga i specijalizovana prodavnica. Ove prodavnice su locirane u dva regiona. Da bi se ispitao uticaj tipa prodavnice na prodaju dve vrste pića kao prvog faktora i regiona kao drugog faktora za svaku kombinaciju tipa prodavnice i regiona registrovana je prodaja dve vrste pića kod pet eksperimentalnih jedinica. Radi se dakle o problemu kome je primeren model MANOVA sa dva faktora. U tabeli su prezentirani podaci uzetog uzorka od elemenata za svaku kombinaciju nivoa prvog i drugog faktora.

Redni broj	Region 1					
	Klasična prodavnica		Samoposluga		Specijalizovana prodavnica	
	Gazirano	Negazirano	Gazirano	Negazirano	Gazirano	Negazirano
1	36	31	13	10	24	23
2	38	27	12	12	32	28
3	30	30	15	22	19	28
4	20	26	25	18	15	21
5	26	26	20	28	25	25
Redni broj	**Region 2**					
	Klasična prodavnica		Samoposluga		Specijalizovana prodavnica	
	Gazirano	Negazirano	Gazirano	Negazirano	Gazirano	Negazirano
1	43	39	32	26	36	35
2	37	38	26	31	40	32
3	38	41	27	29	28	34
4	39	44	25	27	37	33
5	43	43	25	27	39	31

Realizovane vrednosti ocene sredina su:

$$\bar{\mathbf{x}} = \begin{bmatrix} 28.833 \\ 28.833 \end{bmatrix}, \ \bar{\mathbf{x}}_{1\cdot} = \begin{bmatrix} 35.0 \\ 34.5 \end{bmatrix}, \ \bar{\mathbf{x}}_{2\cdot} = \begin{bmatrix} 22 \\ 23 \end{bmatrix}, \ \bar{\mathbf{x}}_{3\cdot} = \begin{bmatrix} 29.5 \\ 29.0 \end{bmatrix}$$

$$\bar{\mathbf{x}}_{\cdot 1} = \begin{bmatrix} 23.33 \\ 23.66 \end{bmatrix}, \ \bar{\mathbf{x}}_{\cdot 2} = \begin{bmatrix} 34.33 \\ 34.00 \end{bmatrix}$$

$$\bar{\mathbf{x}}_{11} = \begin{bmatrix} 30 \\ 28 \end{bmatrix}, \ \bar{\mathbf{x}}_{21} = \begin{bmatrix} 17 \\ 16 \end{bmatrix}, \ \bar{\mathbf{x}}_{31} = \begin{bmatrix} 23 \\ 25 \end{bmatrix}$$

$$\bar{\mathbf{x}}_{12} = \begin{bmatrix} 40 \\ 41 \end{bmatrix}, \ \bar{\mathbf{x}}_{22} = \begin{bmatrix} 27 \\ 28 \end{bmatrix}, \ \bar{\mathbf{x}}_{23} = \begin{bmatrix} 36 \\ 33 \end{bmatrix}.$$

Na osnovu ocene sredina odredili smo matrice sume kvadrata:

$$\mathbf{G} = \begin{bmatrix} 851.667 & 749.167 \\ 749.167 & 661.667 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 907.50 & 852.50 \\ 852.50 & 800.83 \end{bmatrix}, \ \mathbf{W} = \begin{bmatrix} 15.0 & -17.5 \\ -17.5 & 31.6 \end{bmatrix}$$

$$\mathbf{E} = \begin{bmatrix} 656 & 146 \\ 146 & 328 \end{bmatrix}, \ \mathbf{T} = \begin{bmatrix} 2430.17 & 1730.17 \\ 1730.17 & 1822.17 \end{bmatrix}.$$

Determinante matrica, odnosno generalizovane varijanse matrica koje figurišu u izrazu za Wilksovu lambdu su:

$$|\mathbf{E}| = 193852, \ |\mathbf{W} + \mathbf{E}| = 224824, \ |\mathbf{G} + \mathbf{E}| = 690764, \ |\mathbf{B} + \mathbf{E}| = 767929.$$

Za testiranje hipoteze o odsustvu interakcije između tipa prodavnice i regiona izračunavamo vrednost Wilksove lambde

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{W} + \mathbf{E}|} = \frac{193852}{224824} = 0.8622.$$

Vrednost Bartlettove aproksimacije je

$$-\left[3 \cdot 2(5-1) - \frac{2+1-(3-1)(2-1)}{2}\right] \ln(0.8622) = 3.4833.$$

Pošto je kritična vrednost statistike testa , veća od izračunate vrednosti, zaključujemo da ne treba odbaciti hipotezu o odsustvu interakcije između faktora.

Za testiranje efekata prvog faktora izračunavamo vrednost Wilksove lambde

$$\Lambda_1 = \frac{|\mathbf{E}|}{|\mathbf{G} + \mathbf{E}|} = \frac{193852}{690764} = 0.2806.$$

Vrednost Bartlettove aproksimacije je

$$-\left[3 \cdot 2(5-1) - \frac{2+1-(3-1)}{2}\right] \ln(0.2806) = 29.8615.$$

Pošto je kritična vrednost statistike testa $\chi^2_{(3-1)2;0.05} = \chi^2_{4;0.05} = 9.4877$, manja od izračunate vrednosti, zaključujemo da treba odbaciti hipotezu o nultim efektima prvog faktora.

Za testiranje efekata drugog faktora izračunavamo vrednost Wilksove lambde

$$\Lambda_2 = \frac{|\mathbf{E}|}{|\mathbf{B} + \mathbf{E}|} = \frac{193852}{767929} = 0.2524.$$

Vrednost Bartlettove aproksimacije je

$$-\left[3\cdot 2(5-1)-\frac{2+1-(2-1)}{2}\right]\ln(0.2524)=31.6618\,.$$

Pošto je kritična vrednost statistike testa $\chi^2_{(2-1)2;0.05}=\chi^2_{2;0.05}=5.9915$, manja od izračunate vrednosti, zaključujemo da treba odbaciti hipotezu o nultim efektima drugog faktora.

Na osnovu dobijenih rezultata zaključujemo da oba faktora i tip prodavnice i regionalna lokacija utiču na prodaju dve vrste proizvoda. Pri tome oba faktora to čine na aditivan način, pošto nema interakcije među njima. ■

6 DISKRIMINACIONA ANALIZA

Metod multivarijacione analize koji se bavi razdvajanjem različitih grupa i alokacijom opservacija u unapred definisane grupe nazivamo *diskriminaciona analiza*. Kao ilustraciju istraživanja kod koga koristimo metod diskriminacione analize za razdvajanje različitih grupa pretpostavimo da smo sproveli anketu o mišljenju muškog i ženskog dela populacije u pogledu određenog prehrambenog proizvoda (njegovoj hranljivoj vrednosti, ukusu, pakovanju, ceni i dr.). Ukoliko smo multivarijacionom analizom varijanse utvrdili da postoji razlika u mišljenju ova dva dela populacije, pitamo se koja karakteristika proizvoda u najvećoj meri doprinosi njihovom razdvajanju. Odgovor na postavljeno pitanje dobijamo primenom diskriminacione analize. Upotrebna vrednost tog podatka o bitnoj karakteristici, sastoji se u njegovom korišćenju u daljem profiliranju promotivne aktivnosti za svaku od ove dve grupe potrošača.

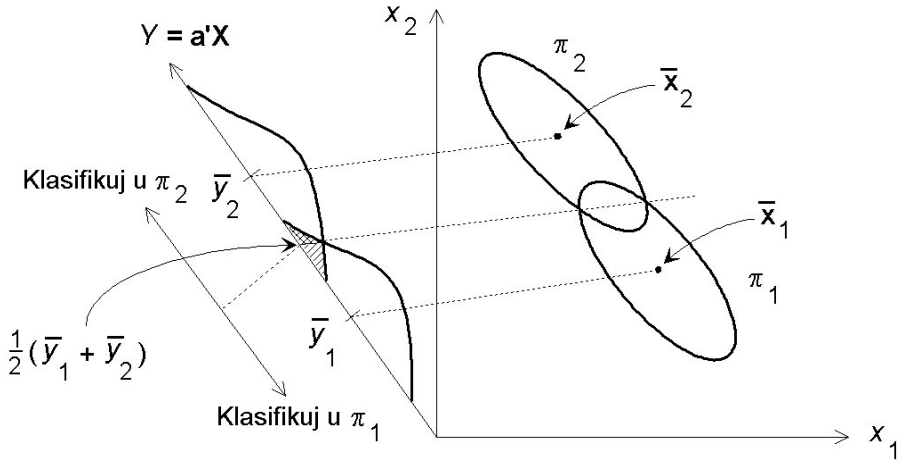
Drugi primer ilustruje istraživanje u kome je moguće koristiti metod diskriminacione analize za svrstavanje opservacija u definisane grupe. Pretpostavimo da se na tržište uvodi novi proizvod. Anketom se traži izjašnjavanje potencijalnih kupaca o karakteristikama tog proizvoda kao i odgovor na pitanje da li bi se opredelili za kupovinu istog. Obradom ankete omogućeno je klasifikovanje ispitanika prema njihovim odgovorima u dve grupe: one koji bi proizvod kupili i one koji proizvod ne bi kupili. Interesuje nas odgovor na pitanje, kolika je verovatnoća da će potencijalni kupac kupiti novi proizvod. Uz pomoć metoda diskriminacione analize, kombinujući odgovore ispitanika iz ove dve grupe, utvrdićemo traženu verovatnoću. Na osnovu nje omogućeno nam je da, u zavisnosti od odgovora nekog ispitanika koji nije obuhvaćen anketom, istoga klasifikujemo u grupu potencijalnih kupaca ili u grupu onih koji proizvod neće kupiti.

Izlaganje u okviru ove glave podelili smo u tri dela. Nakon uvoda u kome ukazujemo na osnovnu ideju i ciljeve diskriminacione analize, u prvom delu izlaganje je posvećeno metodama deskriptivne diskriminacione analizi (razmatra se Fisherov diskriminacioni kriterijum, metode zaključivanja i interpretacije diskriminacionih

funkcija), odnosno metodama koje objašnjavaju razliku između dve i više grupa. Problemom klasifikacije opservacija u unapred definisane grupe korišćenjem metoda diskriminacione analize bavimo se u drugom poglavlju. Ovde se, između ostalog izlažu različiti kriterijumi alokacije opservacija. Naposletku, u poslednjem poglavlju ukazujemo na probleme pri praktičnoj primeni metoda diskriminacione analize, a pre svega izbor promenljivih, anormalnost podataka i korišćenje kvalitativnih promenljivih.

6.1 UVOD

Diskriminaciona analiza ima dva osnovna cilja. Prvi, da utvrdi da li postoji statistički značajna razlika u sredinama dve ili više grupa, a zatim da odredi koja od promenljivih daje najveći doprinos utvrđenoj razlici. Ovaj cilj analize nazivamo *diskriminacija* ili *razdvajanje* između grupa. Drugi cilj odnosi se na utvrđivanje postupka za klasifikaciju opservacija na osnovu vrednosti nekoliko promenljivih u dve ili više razdvojenih, unapred definisanih grupa. Ovaj cilj analize nazivamo *klasifikacija* ili *alokacija* opservacija. U konkretnom istraživanju ova dva cilja često se međusobom preklapaju, pa sredstva analize koja koristimo za razdvajanje između grupa istovremeno služe i za klasifikaciju opservacija u te, unapred definisane grupe. U literaturi, metode diskriminacione analize koje se odnose na prvi cilj - razdvajanje između grupa, izlažu se pod naslovom deskriptivna diskriminaciona analiza, dok se metode primerene drugom cilju - alokaciji opservacija, izlažu pod naslovom metode klasifikacije.



Slika 6.1 Grafička ilustracija diskriminacione analize dve grupe

Sa tehničke strane osnovni cilj diskriminacione analize jeste formiranje linearnih kombinacija nezavisnih promenljivih kojima će se diskriminacija između unapred definisanih grupa tako izvršiti da greška pogrešne klasifikacije opservacija bude minimizirana. Ili, drugačije rečeno, da se maksimizira relativan odnos varijansi između i unutar grupa. Linearnom kombinacijom nezavisnih promenljivih za svakog ispitanika ili objekta odredimo broj, nazvan *diskriminacioni skor*, koji se zatim transformiše u aposteriornu verovatnoću da ispitanik ili objekat potiče iz jedne od grupa. Osnovnu ideju diskriminacione analize izlažemo koristeći grafički prikaz na Slici 6.1 na kome je

ilustrovan problem diskriminacije između dve grupe (π_1 i π_2) na osnovu dve nezavisne promenljive (X_1 i X_2). Elipsama je obuhvaćeno, recimo 95% opservacija prve i druge grupe na odnosnim dijagramima rasturanja. Diskriminacioni skor za svakog ispitanika formira se na osnovu linearne kombinacije vrednosti dve nezavisne promenljive. Znači da u opštem slučaju imamo $Y = \mathbf{a}'\mathbf{X}$, gde je Y diskriminacioni skor, $\mathbf{a}$ je $p-$dimenzioni vektor diskriminacionih koeficijenata (koeficijenti linearne kombinacije), a $\mathbf{X}$ je $p-$dimenzioni vektor nezavisnih promenljivih. Projekcija tačaka sa dijagrama rasturanja na Y osu generiše jednodimenzione rasporede diskriminacionih skorova dveju populacija π_1 i π_2. Istovremeno su pravim linijama težišta elipsi $\bar{\mathbf{x}}_1$ i $\bar{\mathbf{x}}_2$ (realizovana vrednost sredine odnosnih grupa) spojena sa njihovim projekcijama na Y osu (tačke $\bar{y}_1$ i $\bar{y}_2$). Sredine diskriminacionih skorova za ove dve grupe nazivamo prvi, odnosno drugi *centroid*. Njihovim poređenjem možemo utvrditi koliko su grupe međusobom udaljene. Za potrebe klasifikacije opservacija uneta je i sredina centroida, tj. $\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$. Šrafirana površina na krajevima dva jednodimenziona rasporeda dobijena njihovim preklapanjem biće minimalna upravo za projekciju na Y osu one prave linije koja prolazi kroz dve tačke preseka elipsi. Šrafirana površina pod krivom levo od tačke $\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$ predstavlja verovatnoću pogrešne klasifikacije onih ispitanika koji pripadaju grupi π_1, a mi smo ih alocirali u grupu π_2, a šrafirana površina pod krivom desno od tačke $\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$ predstavlja verovatnoću pogrešne klasifikacije ispitanika iz grupe π_2 u grupu π_1.

Na ovaj način problem razdvajanja dve populacije pojednostavljen je u tom smislu što umesto dvodimenzionih rasporeda promenljivih X_1 i X_2 posmatramo jednodimenzione rasporede diskriminacionih skorova Y uz maksimalno razdvajanje sredina grupa. Istovremeno je pokazano kako se linija kojom razdvajamo dve grupe koristi za alokaciju opservacija u jednu od grupa.

Podsećamo da smo u uvodnoj glavi klasifikovali metod diskriminacione analize u grupu metoda zavisnosti kao i metod multivarijacione regresione analize i multivarijacione analize varijanse. Zajednička karakteristika sva tri metoda jeste da na osnovu skupa nezavisnih promenljivih formiraju predviđanje ili opisuju ponašanje zavisne promenljive. Pored ove, nesumljivo bitne osobine koja odlikuje sva tri metoda, postoje i niz drugih osobina po kojima se ove metode suštinski razlikuju. Tako smo kod regresione analize formirali linearnu kombinaciju merljivih (kvantitativnih) promenljivih. Pri tome smo pretpostavili da je zavisna promenljiva merljiva i normalno raspoređena, a da su nezavisne promenljive fiksne vrednosti u ponovljenim uzorcima. U diskriminacionoj analizi formiramo linearnu kombinaciju merljivih promenljivih, ali je zavisna promenljiva nemerljiva (kvalitativna). Za razliku od regresione analize ovde je zavisna promenljiva fiksna (uzima vrednosti 0 i 1 ako razmatramo problem diskriminacije dve grupe), a nezavisne promenljive su slučajne promenljive normalno raspoređene. Cilj regresione analize jeste predviđanje sredine zavisne promenljive na osnovu skupa nezavisnih promenljivih, dok smo kod diskriminacione analize zainteresovani da odredimo linearnu kombinaciju nezavisnih promenljivih tako da se

minimizira verovatnoća pogrešne klasifikacije opservacija u jednu od grupa. Naposletku, strateški pristup kod regresione analize zasnovan je na formalnom modelu za koji uvodimo pretpostavke na osnovu kojih određujemo ocene parametara sa poželjnim statističkim osobinama. Strateški pristup u diskriminacionoj analizi podređen je nalaženju sredstava za razdvajanje grupa, odnosno pouzdanu klasifikaciju opservacija. Sve nabrojane razlike sugerišu da su ove dve metode računski slične, ali da im je bitno drugačija osnova. Pru poređenju analize varijanse i diskriminacione analize uočavamo da je kod prve zavisna promenljiva kvantitativna, a nezavisne su kvalitativne promenljive, dok je kod potonje suprotno, zavisna promenljiva je kvalitativna, a nezavisne promenljive su kvantitativne.

6.2 DESKRIPTIVNA DISKRIMINACIONA ANALIZA

Da bi se postigao prvi cilj diskriminacione analize, što je moguće bolje razdvajanje grupa kao i analiza samog razdvajanja, potrebno je definisati kriterijum diskriminacije. Fisher (1936) je prvi definisao taj kriterijum i izložio postupak diskriminacije između dve i više grupa.

6.2.1 Fisherov pristup - slučaj dve grupe

Osnovnu ideju Fisherovog pristupa izlažemo prvo na primeru populacije koja se sastoji od dve grupe π_1 i π_2. Postupak razdvajanja ove dve grupe zasnivamo na $p-$dimenzionom slučajnom vektoru $\mathbf{X}$. Realizacije slučajnog vektora razlikuju se međusobom već prema tome da li potiču iz populacije π_1 ili populacije π_2. Stoga svaku od ove dve populacije opisujemo odgovarajućom višedimenzionom funkcijom gustine, $f_1(\mathbf{x})$ i $f_2(\mathbf{x})$ respektivno. Da bi se smanjila dimenzija razmatranog problema, Fisher je predložio da se $p-$dimenzione opservacije $\mathbf{X}$ transformišu (linearnom kombinacijom) u jednodimenzionu opservaciju y (diskriminacioni skor), tako da su vrednosti y iz grupe π_1 i π_2 međusobom razdvojene što je moguće više. Ako označimo sredine diskriminacionih skorova prve populacije sa μ_{1Y}, a druge sa μ_{2Y}, tada Fisher ustvari sugeriše da se izaberu koeficijenti linearne kombinacije tako da se maksimizira rastojanje između μ_{1Y} i μ_{2Y} izraženo u jedinicama standardne devijacije diskriminacionih skorova.

Označimo sa $\boldsymbol{\mu}_k = E(\mathbf{X} \mid \pi_k)$, $k = 1, 2$, očekivanu vrednost slučajnog vektora $\mathbf{X}$ ako potiče iz $k-$te populacije. Pretpostavimo da je kovarijaciona matrica $\boldsymbol{\Sigma}$ ista u obe grupe. Na osnovu rezultata iz Odeljka 2.1.4 očekivana vrednost linearne kombinacije $Y = \mathbf{a}'\mathbf{X}$ je $\mu_{1Y} = E(Y \mid \pi_1) = \mathbf{a}'\boldsymbol{\mu}_1$, odn. $\mu_{2Y} = E(Y \mid \pi_2) = \mathbf{a}'\boldsymbol{\mu}_2$, već prema tome da li se diskriminacioni skorovi odnose na prvu ili drugu grupu. Varijansa diskriminacionih skorova je $\sigma_Y^2 = Var(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$, bez obzira o kojoj je grupi reč. Fisher je pošao od količnika

$$\frac{(\mu_{1Y}-\mu_{2Y})^2}{\sigma_Y^2}=\frac{(\mathbf{a}'\boldsymbol{\mu}_1-\mathbf{a}'\boldsymbol{\mu}_2)^2}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}=\frac{(\mathbf{a}'\boldsymbol{\delta})^2}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}, \tag{6.1}$$

gde je $\boldsymbol{\delta}=\boldsymbol{\mu}_1-\boldsymbol{\mu}_2$. Zadatak je da se odredi vektor koeficijenata linearne kombinacije $\mathbf{a}$ tako da se maksimizira količnik $(\mathbf{a}'\boldsymbol{\delta})^2/\mathbf{a}'\mathbf{a}$. Ovaj količnik naziva se *Fisherov diskriminacioni kriterijum*. Pokazano[1] je da se maksimalna vrednost Fisherovog diskriminacionog kriterijuma postiže sledećim izborom koeficijenata linearne kombinacije

$$\mathbf{a}=c\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \tag{6.2}$$

za ma koje $c\neq 0$. Ako je $c=1$ na osnovu vektora $\mathbf{a}$ dobijamo linearnu kombinaciju

$$Y=\mathbf{a}'\mathbf{X}=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{X} \tag{6.3}$$

koju nazivamo *Fisherova linearna diskriminaciona funkcija*. Za ovaj izbor koeficijenata linearne kombinacije maksimalna vrednost Fisherovog diskriminacionog kriterijuma je

$$\max_{\mathbf{a}}\frac{(\mathbf{a}'\boldsymbol{\delta})^2}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}=\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2). \tag{6.4}$$

Koeficijenti linearne kombinacije nisu jednoznačno određeni, jer za ma koje $c\neq 0$ bilo koji vektor $c\mathbf{a}$ maksimiziraće vrednost Fisherovog diskriminacionog kriterijuma. Stoga je uobičajeno da se problem optimizacije rešava uz ograničenje $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}=1$. Time se eliminiše uticaj jedinice mere promenljivih na dobijene rezultate i omogućava njihova lakša interpretacija.

Fisherovu linearnu diskriminacionu funkciju koristimo i za alokaciju opservacija u jednu od dve grupe. Neka je središnja tačka između sredina diskriminacionih skorova iz prve i druge grupe označena sa μ_Y, gde je

$$\mu_Y=\frac{1}{2}(\mu_{1Y}+\mu_{2Y})=\frac{1}{2}(\mathbf{a}'\boldsymbol{\mu}_1+\mathbf{a}'\boldsymbol{\mu}_2)=\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \tag{6.5}$$

i neka za novu opservaciju $\mathbf{x}_0$ imamo diskriminacioni skor $y_0=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0$. Može se pokazati da je

$$E(Y_0\mid\pi_1)-\mu_Y\geq 0 \text{ i } E(Y_0\mid\pi_2)-\mu_Y<0. \tag{6.6}$$

To znači da je pravilo klasifikacije nove opservacije

$$\text{Alociraj } \mathbf{x}_0 \text{ u } \pi_1 \text{ ako je } y_0=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0\geq\mu_Y \tag{6.7a}$$

$$\text{Alociraj } \mathbf{x}_0 \text{ u } \pi_2 \text{ ako je } y_0=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0<\mu_Y \tag{6.7b}$$

[1] Videti npr. Wichern i Johnson (1982, s. 464).

Do sada smo u izlaganju Fisherovog pristupa problemu diskriminacije koristili parametre $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ i $\boldsymbol{\Sigma}$ kao da su nam poznati. U suprotnom ih ocenjujemo na osnovu slučajnog uzorka od n_1 opservacija iz prve i n_2 opservacija iz druge grupe. Za svaki od slučajnih uzoraka odredimo uzoračke pokazatelje $\overline{\mathbf{X}}_1$, $\overline{\mathbf{X}}_2$, $\mathbf{S}_1$ i $\mathbf{S}_2$. Uzoračke kovarijacione matrice koristimo pri formiranju opšte kovarijacione matrice $\overline{\mathbf{S}}$ na osnovu izraza

$$\overline{\mathbf{S}} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{(n_1 + n_2 - 2)} \tag{6.8}$$

Maksimalna vrednost Fisherovog diskriminacionog kriterijuma u uzorku je

$$\max_{\hat{\mathbf{a}}} \frac{(\hat{\mathbf{a}}'\hat{\boldsymbol{\delta}})^2}{\hat{\mathbf{a}}'\overline{\mathbf{S}}\hat{\mathbf{a}}} = \hat{\boldsymbol{\delta}}'\overline{\mathbf{S}}^{-1}\hat{\boldsymbol{\delta}} = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)'\overline{\mathbf{S}}^{-1}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) = d^2 . \tag{6.9}$$

Prema tome, maksimalna vrednost je jednaka kvadratu Mahalanobisovog odstojanja između sredina vektora $\overline{\mathbf{X}}_1$ i $\overline{\mathbf{X}}_2$. Uz pretpostavku o normalnosti rasporeda, testiranje hipoteze o jednakosti sredina dve grupe svodi se na proveru hipoteze da je razdvajanje dve grupe statistički značajno. Prihvatanjem te hipoteze još ne znači da smo korektno izvršili alokaciju opservacija u posmatrane grupe. Nulta hipoteza je $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ protiv alternativne hipoteze $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, a statistika testa za testiranje ove hipoteze

$$\frac{n_1 n_2}{n_1 + n_2} \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} d^2 \tag{6.10}$$

ima $F-$raspored sa p i $(n_1 + n_2 - p - 1)$ stepeni slobode. Odbacivanje nulte hipoteze sugeriše statističku značajnost razdvajanja grupa π_1 i π_2.

Uzoračka Fisherova linearna diskriminaciona funkcija na osnovu realizovanih vrednosti slučajnog uzorka je

$$y = \hat{\mathbf{a}}'\mathbf{x} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\overline{\mathbf{S}}^{-1}\mathbf{x} , \tag{6.11}$$

a ocena tačke koja se nalazi na sredini između sredina diskriminacionih skorova dve grupe, $\overline{y}_1 = \hat{\mathbf{a}}'\overline{\mathbf{x}}_1$ i $\overline{y}_2 = \hat{\mathbf{a}}'\overline{\mathbf{x}}_2$, u oznaci $\hat{\mu}_Y$ data je izrazom

$$\hat{\mu}_Y = \frac{1}{2}(\overline{y}_1 + \overline{y}_2) = \frac{1}{2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\overline{\mathbf{S}}^{-1}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2) . \tag{6.12}$$

Pravilo klasifikacije nove opservacije na osnovu uzoračkih pokazatelja je

Alociraj $\mathbf{x}_0$ u π_1 ako je $y_0 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\overline{\mathbf{S}}^{-1}\mathbf{x}_0 \geq \hat{\mu}_Y$ (6.13a)

Alociraj $\mathbf{x}_0$ u π_2 ako je $y_0 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\overline{\mathbf{S}}^{-1}\mathbf{x}_0 < \hat{\mu}_Y$ (6.13b)

Kako radimo sa konačnim uzorcima ograničenje koje se nameće na primenu uzoračke Fisherove diskriminacione funkcije jeste $(n_1 + n_2 - 2) > p$. U suprotnom uzoračka opšta kovarijaciona matrica je singularna, pa njena inverzna matrica ne postoji.

Primer 6.1 Na osnovu podataka datih u donjoj tabeli odrediti Fisherovu linearnu diskriminacionu funkciju, a zatim na osnovu nje alocirati opservacije u dve grupe. Ako je raspoloživa nova opservacija $\mathbf{x}_0' = [5 \quad 3]$ korišćenjem izračunate Fisherove linearne diskriminacione funkcije alocirati je u prvu ili drugu grupu.

Uzoračke sredine i kovarijacione matrice po grupama, odnosno opšta kovarijaciona matrica su

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 3.5 \\ 4.5 \end{bmatrix}, \ \bar{\mathbf{x}}_2 = \begin{bmatrix} 6.5 \\ 4.5 \end{bmatrix}, \ \mathbf{S}_1 = \frac{1}{3}\begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}, \ \mathbf{S}_2 = \frac{1}{3}\begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix}, \ \bar{\mathbf{S}} = \frac{1}{6}\begin{bmatrix} 10 & 5 \\ 5 & 10 \end{bmatrix}.$$

Redni broj	X_1	X_2	Grupa
1	2	4	1
2	3	3	1
3	4	6	1
4	5	5	1
5	5	4	2
6	6	5	2
7	7	3	2
8	8	6	2

Prema (6.11) uzoračka Fisherova linearna diskriminaciona funkcija je

$$y = \hat{\mathbf{a}}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\bar{\mathbf{S}}^{-1}\mathbf{x} = [-3 \quad 0]\begin{bmatrix} 0.8 & -0.4 \\ -0.4 & 0.8 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -2.4x_1 + 1.2x_2 .$$

Sredine diskriminacionih skorova po grupama su

$$\bar{y}_1 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_1 = \begin{bmatrix} -2.4 & 1.2 \end{bmatrix}\begin{bmatrix} 3.5 \\ 4.5 \end{bmatrix} = -3.$$

$$\bar{y}_2 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_2 = \begin{bmatrix} -2.4 & 1.2 \end{bmatrix}\begin{bmatrix} 6.5 \\ 4.5 \end{bmatrix} = -10.2,$$

a središnja tačka između njih je

$$\hat{\mu}_Y = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(-3 - 10.2) = -6.6 .$$

Diskriminacione skorove za svaku opservaciju prezentiramo u donjoj tabeli zajedno sa indikatorom pripadnosti skora prvoj ili drugoj grupi:

Redni broj	1	2	3	4	5	6	7	8
Diskriminacioni skor	0.0	-3.6	-2.4	-6.0	-7.2	-8.4	-13.2	-12.0
Grupa	1	1	1	1	2	2	2	2

Za novu opservaciju računamo diskriminacioni skor

$$y_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\bar{\mathbf{S}}^{-1}\mathbf{x}_0 = [-3 \quad 0]\begin{bmatrix} 0.8 & -0.4 \\ -0.4 & 0.8 \end{bmatrix}\begin{bmatrix} 5 \\ 3 \end{bmatrix} = -8.4 .$$

Kako je $y_0 = -8.4 < \hat{\mu}_Y = -6.6$, prema pravilu klasifikacije (6.13) novu opservaciju alociramo u drugu grupu.

Napominjemo da smo korišćenjem Fisherove linearne diskriminacione funkcije ispravno alocirali sve opservacije iz uzorka. U opštem slučaju greške klasifikacije se javljaju, ali ih tokom analize težimo minimizirati. ■

6.2.2 Fisherov pristup - slučaj više grupa

Primenom Fisherovog pristupa u slučaju dve grupe dobili smo jednu linearnu kombinaciju nezavisnih promenljivih. Ona je određena tako da se maksimizira odnos varijacija između i unutar grupa. Dobijene diskriminacione skorove prikazali smo na Y osi. U slučaju više grupa naš zadatak je isti, uz minimalne greške klasifikacije razdvojiti posmatrane grupe. Međutim, teško se može očekivati da će jedna diskriminaciona funkcija uspešno razdvojiti te grupe. Kako raspolažemo $p-$dimenzionim vektorom opservacija koje potiču iz jedne od g grupa, moguće je formirati $\min(p, g-1)$ linearnih kombinacija nezavisnih promenljivih, odnosno diskriminacionih funkcija. Svaka od njih, kao što ćemo videti, ne mora biti statistički značajna. Najčešće se zadovoljavajuće razdvajanje postiže sa dve do tri diskriminacione funkcije. Ako smo odredili dve diskriminacione funkcije, diskriminacione skorove računate na osnovu njih prikazujemo na dijagramu rasturanja, pa kažemo da smo opservacije prikazali u diskriminacionom prostoru. Time smo redukovali problem razdvajanja g grupa na osnovu $p-$dimenzionih opservacija na slučaj koji se vizuelno može analizirati. Na osnovu prikaza opservacija u diskriminacionom prostoru uočavamo vezu između grupa, njihovu bliskost ili eventualno grupisanje, postojanje nestandardnih opservacija ili nekih devijacija u podacima.

Postupak određivanja diskriminacionih funkcija u slučaju više grupa započinje određivanjem prve diskriminacione funkcije tako što će se maksimizirati relativan odnos varijacija između i unutar grupa. Zatim se određuje druga diskriminaciona funkcija tako da se maksimizira odnos preostalog dela varijacija između i unutar grupa (preostao

nakon izdvajanja prve diskriminacione funkcije) uz dodatni uslov da diskriminacioni skorovi određeni na osnovu prve i druge funkcije budu nekorelisani. Svaka naredna diskriminaciona funkcija određena je uz uslov maksimiziranja odnosa preostalih varijacija između i unutar grupa (nakon izdvajanja prethodnih diskriminacionih funkcija) uz ograničenje da su njeni diskriminacionih skorovi nekorelisani sa diskriminacionim skorovima svih prethodno određenih diskriminacionih funkcija. Napominjemo pri tome da Y ose u diskriminacionom prostoru ne moraju biti međusobom ortogonalne, nego se samo zahteva da diskriminacioni skorovi budu nekorelisani.

Pretpostavimo da se populacija sastoji od g grupa, π_k, $k = 1, 2, ..., g$. Zadatak je da se na osnovu $p-$dimenzionog slučajnog vektora $\mathbf{X}$ izvrši optimalno razdvajanje ovih g grupa. Sredinu slučajnog vektora $\mathbf{X}$, ako potiče iz $k-$te populacije, označimo sa $\boldsymbol{\mu}_k = E(\mathbf{X} \mid \pi_k)$, $k = 1, 2, ..., g$, a pretpostavimo da su kovarijacione matrice g grupa jednake međusobom i jednake $\boldsymbol{\Sigma}$. Na osnovu sredina $\boldsymbol{\mu}_k$ odredimo opštu sredinu svih grupa, u oznaci $\boldsymbol{\mu}$, gde je

$$\boldsymbol{\mu} = \frac{1}{g}\sum_{k=1}^{g}\boldsymbol{\mu}_k \tag{6.14}$$

Da bismo definisali Fisherov diskriminacioni kriterijum za slučaj više grupa uvodimo matricu suma uzajamnih proizvoda odstupanja sredina grupa od opšte sredine. Ova matrica sadrži informaciju o varijacijama između grupa. Označimo je sa $\mathbf{B}_0$, gde je

$$\mathbf{B}_0 = \sum_{k=1}^{g}(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})'. \tag{6.15}$$

Zainteresovani smo za takvu linearnu kombinaciju elemenata slučajnog vektora $\mathbf{X}$, tj. $Y = \mathbf{a}'\mathbf{X}$, koja će istaći razlike između grupa s obzirom na elemente sredine populacije. Njena očekivana vrednost i varijansa su

$$\mu_{kY} = E(Y) = \mathbf{a}'E(\mathbf{X} \mid \pi_k) = \mathbf{a}'\boldsymbol{\mu}_k, \; k = 1, 2, ..., g, \tag{6.16}$$

$$\sigma_Y^2 = Var(Y) = \mathbf{a}'Cov(\mathbf{X})\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}, \text{ za sve grupe} \tag{6.17}$$

Naposletku, opšta sredina linearnih kombinacija u oznaci μ_Y je

$$\mu_Y = \frac{1}{g}\sum_{k=1}^{g}\mu_{kY} = \mathbf{a}'\boldsymbol{\mu} \tag{6.18}$$

Sada smo u prilici da definišemo Fisherov diskriminacioni kriterijum

$$\frac{\sum_{k=1}^{g}(\mu_{kY} - \mu_Y)^2}{\sigma_Y^2} = \frac{\sum_{k=1}^{g}(\mathbf{a}'\boldsymbol{\mu}_k - \mathbf{a}'\boldsymbol{\mu})^2}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}} = \frac{\mathbf{a}'\left(\sum_{k=1}^{g}(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})'\right)\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}} \tag{6.19}$$

odnosno

$$\frac{\sum_{k=1}^{g}(\mu_{kY}-\mu_Y)^2}{\sigma_Y^2}=\frac{\mathbf{a}'\mathbf{B}_0\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}. \tag{6.20}$$

Time je izvršena generalizacija Fisherovog diskriminacionog kriterijuma izloženog u slučaju dve grupe. On predstavlja relativan odnos varijacija između i unutar grupa. Koeficijente linearne kombinacije biramo tako da se postigne maksimalna vrednost ovog kriterijuma uz ograničenje da je $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}=1$. Pokazano[2] je da se za određivanje vektora $\mathbf{a}$ koji maksimizira vrednost Fisherovog kriterijuma zahteva rešavanje sledećeg sistema jednačina

$$(\mathbf{B}_0-\lambda\boldsymbol{\Sigma})\mathbf{a}=\mathbf{0} \tag{6.21}$$

ili ekvivalentno

$$(\boldsymbol{\Sigma}^{-1}\mathbf{B}_0-\lambda\mathbf{I})\mathbf{a}=\mathbf{0} \tag{6.22}$$

Potrebno je, dakle, odrediti karakteristične korene i karakteristične vektore matrice $\boldsymbol{\Sigma}^{-1}\mathbf{B}_0$ Ako označimo sa $\lambda_1,\lambda_2,...,\lambda_s>0$ prvih s $\min(g-1,p)$ pozitivnih karakterističnih korena matrice $\boldsymbol{\Sigma}^{-1}\mathbf{B}_0$, a pridružene karakteristične vektore sa $\mathbf{e}_1,\mathbf{e}_2,...,\mathbf{e}_s$ (pri čemu je $\mathbf{e}'\boldsymbol{\Sigma}\mathbf{e}=1$), tada upravo ovi karakteristični vektori jesu koeficijenti linearne kombinacije koji maksimiziraju vrednost Fisherovog diskriminacionog kriterijuma. Za vektor $\mathbf{a}_1=\mathbf{e}_1$ imamo linearnu kombinaciju $Y_1=\mathbf{a}_1'\mathbf{X}$ koju nazivamo prva *diskriminanta*. Vektor koeficijenata $\mathbf{a}_2=\mathbf{e}_2$ kod druge diskriminante $Y_2=\mathbf{a}_2'\mathbf{X}$, maksimizira Fisherov kriterijum uz uslov da su prva i druga diskriminanta međusobom nekorelisane, tj. da je $Cov(\mathbf{a}_1'\mathbf{X},\mathbf{a}_2'\mathbf{X})=0$. Nastavljajući na taj način, vektor koeficijenata $\mathbf{a}_r=\mathbf{e}_r$ kod $r-$te diskriminante $Y_r=\mathbf{a}_r'\mathbf{X}$, maksimizira Fisherov kriterijum uz ograničenje da je $Cov(\mathbf{a}_r'\mathbf{X},\mathbf{a}_i'\mathbf{X})=0$, $i<r$. Pri tome smo izvršili standardizaciju linearne kombinacije uzimajući da je $Var(\mathbf{a}_i'\mathbf{X})$, $i=1,2,...,s$.

Dobijene promenljive ili diskriminante $Y_1,Y_2,...,Y_s$ nazivamo i *kanoničke diskriminacione funkcije*. Kanonička u nazivu diskriminacione funkcije potiče od veze koja postoji između diskriminacione analize na bazi Fisherovog kriterijuma i kanoničke korelacione analize. Naime, ako definišemo veštačke promenljive D_k, $k=1,2,...,g-1$, tako da uzimaju vrednost 1 ako opservacija pripada $k-$toj grupi, 0 ako joj ne pripada, tada je kanonička korelaciona analiza između promenljivih $X_1,X_2,...,X_p$ i veštačkih promenljivih $D_1,D_2,...,D_{g-1}$ ekvivalentna diskriminacionoj analizi zasnovanoj na Fisherovom kriterijumu. Koeficijenti kanoničke korelacije dati su sa $\lambda_i/(1+\lambda_i)$, gde su λ_i karakteristični koreni iz diskriminacione analize.

[2] Videti npr. Wichern i Johnson (1982, ss. 506-7).

Karakteristični koreni matrice $\boldsymbol{\Sigma}^{-1}\mathbf{B}_0$, tj. λ_i, s obzirom na način njihovog definisanja, predstavljaju sumu kvadrata među grupama promenljive Y_i, $i=1,2,...,s$, i mere stepen varijacija između grupa koji se pripisuje $i-$toj kanoničkoj diskriminacionoj funkciji. Ako za svaki karakteristični koren odredimo njegov udeo u ukupnoj varijansi, tj.

$$\frac{\lambda_i}{\sum_{i=1}^{s}\lambda_i},\ i=1,2,...,s, \tag{6.23}$$

tada ćemo dobiti relativan pokazatelj važnosti svake od kanoničkih diskriminacionih funkcija za razdvajanje grupa.

Pretpostavimo da raspolažemo slučajnim uzorkom od n_k opservacija slučajnog vektora $\mathbf{X}$ iz $k-$te grupe, $k=1,2,...,g$ i označimo ga sa $\mathbf{X}_{ik}$, $i=1,2,...,n_k$. Na osnovu uzoračkih pokazatelja grupa, sredina i kovarijacionih matrica

$$\bar{\mathbf{X}}_k=\frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{X}_{ik},\ \mathbf{S}_k=\frac{1}{n_k-1}\sum_{i=1}^{n_k}(\mathbf{X}_{ik}-\bar{\mathbf{X}}_k)(\mathbf{X}_{ik}-\bar{\mathbf{X}}_k)',\ k=1,2,...,g \tag{6.24}$$

ocenjujemo opštu sredinu svih grupa

$$\bar{\mathbf{X}}=\frac{\sum_{k=1}^{g}n_k\bar{\mathbf{X}}_k}{\sum_{k=1}^{g}n_k}=\frac{\sum_{k=1}^{g}\sum_{i=1}^{n_k}\mathbf{X}_{ik}}{\sum_{k=1}^{g}n_k}. \tag{6.25}$$

Određujemo zatim matricu suma kvadrata između grupa u oznaci $\mathbf{B}$

$$\mathbf{B}=\sum_{k=1}^{g}n_k(\bar{\mathbf{X}}_k-\bar{\mathbf{X}})(\bar{\mathbf{X}}_k-\bar{\mathbf{X}})', \tag{6.26}$$

a za ocenu kovarijacione matrice $\boldsymbol{\Sigma}$ formiramo matricu suma kvadrata unutar grupa

$$\mathbf{W}=\sum_{k=1}^{g}(n_k-1)\mathbf{S}_k=\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{X}_{ik}-\bar{\mathbf{X}}_k)(\mathbf{X}_{ik}-\bar{\mathbf{X}}_k)' \tag{6.27}$$

tako da je

$$\bar{\mathbf{S}}=\frac{\mathbf{W}}{\sum_{k=1}^{g}n_k-g} \tag{6.28}$$

tražena ocena opšte kovarijacione matrice. S obzirom na međusobnu vezu matrica $\bar{\mathbf{S}}$ i $\mathbf{W}$ koeficijenti $\hat{\mathbf{a}}$ koji maksimiziraju Fisherov diskriminacioni kriterijum $\hat{\mathbf{a}}'\mathbf{B}\hat{\mathbf{a}}/\hat{\mathbf{a}}'\bar{\mathbf{S}}\hat{\mathbf{a}}$ istovremeno će maksimizirati $\hat{\mathbf{a}}'\mathbf{B}\hat{\mathbf{a}}/\hat{\mathbf{a}}'\mathbf{W}\hat{\mathbf{a}}$. Imajući ovo u vidu opredeljujemo se za uobičajeni pristup prema kome se rešenje problema određivanja kanoničkih

diskriminacionih funkcija zasniva na određivanju koeficijenata $\hat{\mathbf{a}}$ kao karakterističnih vektora matrice $\mathbf{W}^{-1}\mathbf{B}$.

Ako označimo sa $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_s > 0$, $s \le \min(g-1, p)$ pozitivne karakteristične korene matrice $\mathbf{W}^{-1}\mathbf{B}$, a pridružene karakteristične vektore sa $\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, ..., \hat{\mathbf{c}}_s$ (pri čemu je $\hat{\mathbf{c}}'\bar{\mathbf{S}}\hat{\mathbf{c}} = 1$) tada se maksimalna vrednost Fisherovog diskriminacionog kriterijuma

$$\frac{\hat{\mathbf{a}}'\mathbf{B}\hat{\mathbf{a}}}{\hat{\mathbf{a}}'\mathbf{W}\hat{\mathbf{a}}} = \frac{\hat{\mathbf{a}}'\left(\sum_{k=1}^{g}(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})'\right)\hat{\mathbf{a}}}{\hat{\mathbf{a}}'\left(\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{X}_{ik} - \bar{\mathbf{X}}_k)(\mathbf{X}_{ik} - \bar{\mathbf{X}}_k)'\right)\hat{\mathbf{a}}} \tag{6.29}$$

dostiže koeficijentima linearne kombinacije koji su jednaki karakterističnim vektorima, tj. $\hat{\mathbf{a}}_i = \hat{\mathbf{e}}_i$, $i = 1, 2, ..., s$. Sada je $\hat{y}_1 = \mathbf{a}_1'\mathbf{x}$, uzoračka prva kanonička diskriminaciona funkcija (naziva se još i prva Fisherova diskriminaciona funkcija ili prva diskriminanta). U opštem slučaju $\hat{y}_r = \mathbf{a}_r'\mathbf{x}$ predstavlja $r-$tu kanoničku diskriminacionu funkciju za $r \le s$. Diskriminacioni skorovi računati na osnovu različitih kanoničkih diskriminacionih funkcija su međusobom nekorelisani, ali zato u opštem slučaju diskriminacione ose (drugi naziv vektora koeficijenata kanoničkih diskriminacionih funkcija, tj. karakterističnih vektora) neće biti međusobom ortogonalne.

Primer 6.2 Neka je dat slučajan uzorak iz tri populacije $(g = 3)$ za koje pretpostavljamo da imaju istu kovarijacionu matricu. Opservacije dve promenljive $(p = 2)$ za svaku populaciju date su sledećim matricama podataka

$$\mathbf{X}_1 = \begin{bmatrix} 2 & 2 & 2 \\ 1 & 0 & 2 \end{bmatrix}, \ \mathbf{X}_2 = \begin{bmatrix} 0 & -1 & 1 & 0 \\ 3 & 1 & 3 & 1 \end{bmatrix}, \ \mathbf{X}_3 = \begin{bmatrix} -1 & 0 & -2 \\ -1 & -3 & -2 \end{bmatrix}.$$

Odrediti kanoničke diskriminacione funkcije.

Na osnovu uzoračkih pokazatelja

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \ \bar{\mathbf{x}}_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \ \bar{\mathbf{x}}_3 = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

odredimo opštu sredinu, matrice suma kvadrata između i unutar grupa, kao i opštu kovarijacionu matricu:

$$\bar{\mathbf{x}} = \begin{bmatrix} 0.3 \\ 0.5 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 14.1 & 10.5 \\ 10.5 & 28.5 \end{bmatrix}, \ \mathbf{W} = \begin{bmatrix} 4 & 1 \\ 1 & 8 \end{bmatrix}, \ \bar{\mathbf{S}} = \frac{1}{7}\begin{bmatrix} 4 & 1 \\ 1 & 8 \end{bmatrix}.$$

Potrebni elementi za određivanje diskriminacionih funkcija su:

$$\mathbf{W}^{-1} = \frac{1}{31}\begin{bmatrix} 8 & -1 \\ -1 & 4 \end{bmatrix},\ \mathbf{W}^{-1}\mathbf{B} = \frac{1}{31}\begin{bmatrix} 102.3 & 55.5 \\ 27.9 & 103.5 \end{bmatrix}.$$

Kako je $\min(g-1, p) = \min(2,2) = 2$, to znači da se mogu odrediti dve kanoničke diskriminacione funkcije na osnovu determinantne jednačine

$$\left|\mathbf{W}^{-1}\mathbf{B} - \hat{\lambda}\mathbf{I}\right| = \left\|\begin{bmatrix} 3.3 - \hat{\lambda} & 1.7903 \\ 0.9 & 3.3387 - \hat{\lambda} \end{bmatrix}\right\| = 0.$$

Karakteristični polinom je

$$\hat{\lambda}^2 - 6.6387\hat{\lambda} + 9.4065 = 0,$$

a karakteristični koreni su $\hat{\lambda}_1 = 4.5889$ i $\hat{\lambda}_2 = 2.0498$. Rešavanjem sledećeg homogenog sistema jednačina određujemo prvi pridruženi karakteristični vektor

$$\left(\mathbf{W}^{-1}\mathbf{B} - \hat{\lambda}\mathbf{I}\right)\hat{\mathbf{a}}_1 = 0$$

$$\begin{bmatrix} -1.2889 & 1.7903 \\ 0.9000 & -1.2502 \end{bmatrix}\begin{bmatrix} \hat{a}_{11} \\ \hat{a}_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

odakle dobijamo $\hat{\mathbf{a}}_1' = \begin{bmatrix} 1 & 0.7199 \end{bmatrix}$, a uz njegovu normalizaciju $(\hat{\mathbf{a}}_1'\overline{\mathbf{S}}\hat{\mathbf{a}}_1 = 1)$ imamo $\hat{\mathbf{a}}_1' = \begin{bmatrix} 0.8545 & 0.6152 \end{bmatrix}$. Uz isti postupak određujemo drugi karakteristični vektor $\hat{\mathbf{a}}_2' = \begin{bmatrix} 1.0374 & -0.7244 \end{bmatrix}$. Znači da su dve kanoničke diskriminacione funkcije date izrazima

$$\hat{y}_1 = \hat{\mathbf{a}}_1'\mathbf{x} = \begin{bmatrix} 0.8545 & 0.6152 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.8545x_1 + 0.6152x_2$$

$$\hat{y}_2 = \hat{\mathbf{a}}_2'\mathbf{x} = \begin{bmatrix} 1.0374 & -0.7244 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1.0374x_1 - 0.7244x_2.$$

Na osnovu izračunatih karakterističnih korena konstatujemo da je prvom kanoničkom diskriminacionom funkcijom obuhvaćeno 69.12% varijacija između grupa, a drugom svega 30.88%.

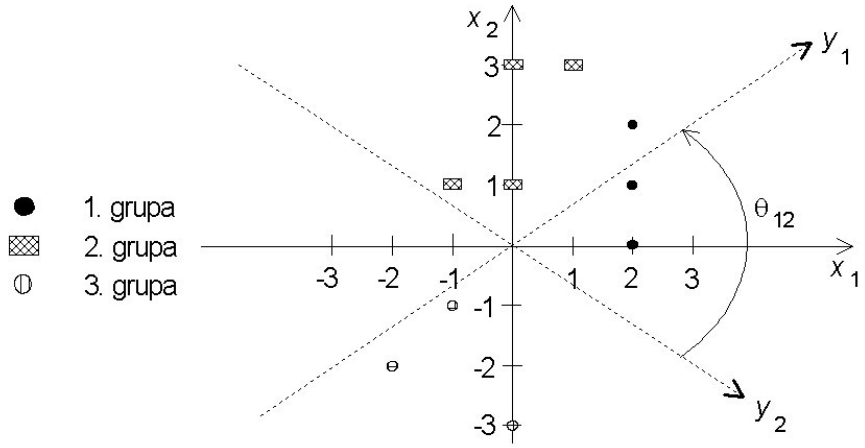
U narednoj tabeli prezentirani su diskriminacioni skorovi računati na osnovu dve kanoničke diskriminacione funkcije zajedno sa originalnim opservacijama.

Redni broj	Originalne promenljive		Grupa	Diskriminacioni skorovi	
	Prva	Druga		1. diskriminanta	2. diskriminanta
1	2	1	1	2.32	1.35
2	2	0	1	1.71	2.07
3	2	2	1	2.94	0.63
4	0	3	2	1.85	-2.17
5	-1	1	2	-0.24	-1.76
6	1	3	2	2.71	-1.14
7	0	1	2	0.62	-0.72
8	-1	-1	3	-1.47	-0.31
9	0	-3	3	-1.85	2.17
10	-2	-2	3	-2.94	-0.63

Karakteristični vektori $\hat{\mathbf{a}}_1$ i $\hat{\mathbf{a}}_2$, koje definišu diskriminacione ose, nisu ortogonalni pošto je $\hat{\mathbf{a}}_1'\hat{\mathbf{a}}_2 = 0.4409 \neq 0$ Takođe ose nisu ortogonalne u odnosu na koordinatni sistem (x_1, x_2). Kako je kosinus ugla između dva karakteristična vektora

$$\cos\theta_{12} = \frac{\hat{\mathbf{a}}_1'\hat{\mathbf{a}}_2}{\sqrt{\hat{\mathbf{a}}_1'\hat{\mathbf{a}}_1}\sqrt{\hat{\mathbf{a}}_2'\hat{\mathbf{a}}_2}} = \frac{0.4409}{\sqrt{1.1087}\sqrt{1.6010}} = 0.3309$$

to je ugao između diskriminacionih osa $\theta_{12} = 71^0$. Sada možemo grafički prikazati originalne opservacije i diskriminacione ose.



Sa grafikona se jasno uočava ono što nam pokazatelj relativnog značaja prve kanoničke diskriminacione funkcije u iznosu od 69.12% kazuje. Naime, projekcije originalnih opservacija na osu y_1 tako su raspoređeni duž nje da su tri grupe dobrim delom potpuno razdvojene međusobom. Ostatak varijacija između grupa u iznosu od 30.88% "objašnjen" je drugom kanoničkom diskriminacionom funkcijom, odnosno u tom iznosu druga diskriminanta pomaže u razdvajanju grupa.

Napominjemo da je ovaj primer korišćen pre svega kao ilustracija funkcionisanja diskriminacione analize u slučaju tri grupe jer se na dijagramu rasturanja originalnih opservacija mogu jasno razdvojiti tri grupe. U kompleksnijem slučaju kada je $p > 2$ u nemogućnosti grafičkog prikaza originalnih opservacija koristimo diskriminacione ose i diskriminacione skorove u cilju razdvajanja grupa. ■

Pokazano je u slučaju dve grupe da se Fisherova diskriminaciona funkcija može koristiti za alokaciju opservacija u jednu od grupa. Mada će u narednom poglavlju biti više reči o tom aspektu diskriminacione analize na ovom mestu ukažimo ukratko na postupak korišćenja kanoničkih diskriminacionih funkcija za klasifikaciju opservacija u slučaju tri grupe. Podsećamo da je klasifikacija opservacija u slučaju dve grupe zahtevala izračunavanje središnje tačke između prvog i drugog centroida (sredine diskriminacionih skorova po grupama). U zavisnosti od položaja diskriminacionog skora nove opservacije u odnosu na tu središnju tačku, alocirali smo opservaciju u prvu ili drugu grupu. Ukoliko je rastojanje diskriminacionog skora nove opservacije manje od prvog centroida u odnosu na njegovo rastojanje od drugog centroida, tada se opservacija alocira u prvu grupu i obratno. Slično postupamo i u opštem slučaju. Koristimo se Mahalanobisovim odstojanjem opservacije od centroida svake grupe, pa novu opservaciju alociramo u grupu od čijeg je centroida ova najmanje udaljena.

Označimo sa $\mathbf{Y}$ $(s\times 1)$ vektor čiji su elementi kanoničke diskriminacione funkcije, tj. $\mathbf{Y}' = [Y_1,...,Y_s]$. Za $k-$tu populaciju , vektor $\mathbf{Y}$ ima sredinu $\boldsymbol{\mu}'_{kY} = [\mu_{kY_1},...,\mu_{kY_s}]$, a kovarijaciona matrica je za sve grupe jednaka jediničnoj matrici (podsećamo da su varijanse jednake jedinici zbog uslova $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = 1$, a kovarijanse između različitih diskriminacionih funkcija su jednake nuli). Tada se Mahalanobisovo odstojanje vektora $\mathbf{Y} = \mathbf{y}$ od sredina vektora $\boldsymbol{\mu}_{kY}$ svodi na Euklidsko odstojanje

$$(\mathbf{y}-\boldsymbol{\mu}_{kY})'(\mathbf{y}-\boldsymbol{\mu}_{kY}) = \sum_{i=1}^{s}(y_i-\mu_{kY_i})^2 \tag{6.30}$$

Novu opservaciju $\mathbf{x}_0$ alociramo u populaciju π_k koristeći pravilo klasifikacije prema kome je

$$\sum_{i=1}^{s}(y_i-\mu_{kY_i})^2 = \sum_{i=1}^{s}\left[\mathbf{a}_i'(\mathbf{x}_0-\boldsymbol{\mu}_k)\right]^2 \le \sum_{i=1}^{s}\left[\mathbf{a}_i'(\mathbf{x}_0-\boldsymbol{\mu}_r)\right]^2 \text{ , za svako } r\neq k \tag{6.31}$$

Odgovarajuće pravilo alokacije u slučaju korišćenja podataka iz uzorka je da se nova opservacija $\mathbf{x}_0$ alocira u π_k ako je

$$\sum_{i=1}^{s}(\hat{y}_i-\overline{y}_{ki})^2 = \sum_{i=1}^{s}\left[\hat{\mathbf{a}}_i'(\mathbf{x}_0-\overline{\mathbf{x}}_k)\right]^2 \le \sum_{i=1}^{s}\left[\hat{\mathbf{a}}_i'(\mathbf{x}_0-\overline{\mathbf{x}}_r)\right]^2 \text{ , za svako } r\neq k \tag{6.32}$$

Napominjemo pri tome da je za klasifikaciju moguće koristiti svih s raspoloživih kanoničkih diskriminacionih funkcija ili manji broj njih.

Primer 6.3 Na osnovu podataka iz Primera 6.2 i ocenjenih kanoničkih diskriminacionih funkcija

$$\hat{y}_1 = \hat{\mathbf{a}}_1'\mathbf{x} = 0.8545x_1 + 0.6152x_2$$

$$\hat{y}_2 = \hat{\mathbf{a}}_2'\mathbf{x} = 1.0374x_1 - 0.7244x_2$$

klasifikovati novu opservaciju $\mathbf{x}_0' = [-1 \quad 2]$ u jednu od tri grupe.

Diskriminacioni skor nove opservacije je $\hat{y}_1 = 0.3758$, $\hat{y}_2 = -2.4862$. Odredimo potom centroide tri grupa $\overline{y}_{ki} = \hat{\mathbf{a}}_i'\overline{\mathbf{x}}_k$, $k = 1,2,3$; $i = 1,2$

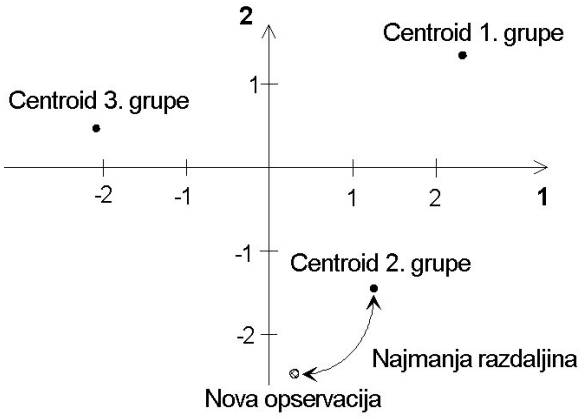
$$\overline{y}_{11} = \hat{\mathbf{a}}_1'\overline{\mathbf{x}}_1 = \begin{bmatrix} 0.8545 & 0.6152 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \end{bmatrix} = 2.3243$$

$$\overline{y}_{12} = \hat{\mathbf{a}}_2'\overline{\mathbf{x}}_1 = \begin{bmatrix} 1.0374 & -0.7244 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \end{bmatrix} = 1.3504$$

Koordinate ostalih centroida su: $\overline{y}_{21} = 1.2304$, $\overline{y}_{22} = -1.4488$, $\overline{y}_{31} = -2.0849$, $\overline{y}_{32} = 0.4114$. Naposletku, odredimo vrednost klasifikacionog kriterijuma

$$\sum_{i=1}^{2} (\hat{y}_i - \overline{y}_{ki})^2, \quad k = 1,2,3$$

Vrednost kriterijuma za prvu grupu je 18.5161, za drugu 1.8065 i treću 14.4516. Kako se minimalna vrednost kriterijuma dobija za $k = 2$, to se opservacija klasifikuje u drugu grupu. Na narednoj slici prikazani su centroidi tri grupe kao i vrednost diskriminacionog skora nove opservacije u diskriminacionom prostoru.



■

6.2.3 Zaključivanje u kanoničkoj diskriminacionoj analizi

Pri radu sa ocenjenim kanoničkim diskriminacionim funkcijama postavlja se pitanje koliko ih je statistički značajno, odnosno koliko funkcija zadržati u daljoj analizi. Odgovor na postavljeno pitanje zasnivamo na primeni Bartlettovog testa. Njime testiramo značajnost karakterističnih korena matrice $\mathbf{W}^{-1}\mathbf{B}$.

Testiramo hipotezi da su prvih s karakterističnih korena matrice $\mathbf{W}^{-1}\mathbf{B}$ jednaki nuli, odnosno $H_0 : \lambda_1 = ... = \lambda_s = 0$. Testiranje zasnivamo na Bartlettovoj χ^2 statistici

$$\left[n - 1 - \frac{1}{2}(p+g) \right] \sum_{i=1}^{s} \ln(1+\hat{\lambda}_j) \tag{6.33}$$

koja ima aproksimativno χ^2 – raspored sa $p(g-1)$ stepeni slobode, ako je tačna nulta hipoteza. Ukoliko je realizovana vrednost ove statistike veća od kritične vrednosti $\chi^2_{p(g-1);\alpha}$, tada zaključujemo da je barem jedan karakteristični koren (λ_1 kao najveći), odnosno diskriminaciona funkcija, statistički značajno različita od nule. Ako smo odbacili H_0, izostavljajući $\hat{\lambda}_1$ iz statistike testa formiramo statistiku

$$\left[n - 1 - \frac{1}{2}(p+g) \right] \sum_{i=2}^{s} \ln(1+\hat{\lambda}_j) \tag{6.34}$$

koja ima aproksimativno χ^2 – raspored sa $(p-1)(g-1)$ stepeni slobode. Testiramo hipotezu $H_{01} : \lambda_2 = ... = \lambda_s = 0$. Ako smo odbacili H_{01}, u narednom koraku izostavljamo $\hat{\lambda}_2$ i ponovo formiramo statistiku testa za test hipoteze $H_{02} : \lambda_3 = ... = \lambda_s = 0$. Pretpostavimo da smo, nastavljajući na isti način, odbacili hipoteze da su prvih r karakterističnih korena jednaki nuli, sada nulta hipoteza postaje $H_{0r} : \lambda_{r+1} = ... = \lambda_s = 0$, a testiramo je korišćenjem statistike testa

$$\left[n - 1 - \frac{1}{2}(p+g) \right] \sum_{i=(r+1)}^{s} \ln(1+\hat{\lambda}_j) \tag{6.35}$$

koja za tačnu nultu hipotezu ima aproksimativno χ^2 – raspored sa $(p-r)(g-r-1)$ stepeni slobode.

Primer 6.4 Na osnovu podataka iz Primera 6.2 utvrditi statističku značajnost diskriminacionih funkcija.

Testiramo hipotezu $H_0 : \lambda_1 = \lambda_2 = 0$. Kako je $n = 10$, $p = 2$, $g = 3$, $\hat{\lambda}_1 = 4.5889$ i $\hat{\lambda}_2 = 2.0498$, to je vrednost statistike testa

$$\left[n-1-\frac{1}{2}(p+g)\right]\sum_{i=1}^{s}\ln(1+\hat{\lambda}_j)=\left[10-1-\frac{1}{2}(2+3)\right][\ln(5.5889)+\ln(3.0498)]$$

$$=6.5\cdot(1.7208+1.1151)=18.43.$$

Na nivou značajnosti $\alpha=0.05$ kritična vrednost testa iznosi $\chi^2_{4;0.05}=9.49$. Pošto je izračunata vrednost veća od kritične vrednosti odbacujemo nultu hipotezu na nivou značajnosti od 5% da je barem jedan karakteristični koren različit od nule.

Pošto smo odbacili nultu hipotezu nastavljamo dalji postupak testiranja. Kao što smo utvrdili prvi, veći, karakteristični koren je različit od nule, pa definišemo nultu hipotezu $H_{01}:\lambda_2=0$, a test statistika ima vrednost

$$6.5\cdot\ln(3.0498)=6.5\cdot1.1151=7.25$$

Na nivou značajnosti $\alpha=0.05$ kritična vrednost testa iznosi $\chi^2_{2;0.05}=5.99$. Pošto je izračunata vrednost veća od kritične odbacili bismo nultu hipotezu da je drugi karakteristični koren jednak nuli na nivou značajnosti od 5%. Konstatujemo da su obe kanoničke diskriminacione funkcije statistički značajne.

Napominjemo da se radi o testu primerenom velikim uzorcima, pa njegovo korišćenje u ovom primeru treba posmatrati samo kao ilustraciju. ■

6.2.4 Interpretacija kanoničke diskriminacione funkcije

Nakon ocene kanoničkih diskriminacionih funkcija postavlja se pitanje da li je potrebno u daljoj analizi zadržati sve kanoničke diskriminacione funkcije ili je analizu razdvajanja grupa moguće obaviti sa njihovim manjim brojem. Kako je broj populacija najčešće znatno manji od broja korišćenih promenljivih, to je broj diskriminacionih funkcija u takvim slučajevima jednak broju populacija manje jedan. Formalnim postupkom statističkog testiranja značajnosti diskriminacionih funkcija ovaj broj možemo dalje smanjiti. Drugi subjektivan pristup zasnovan je na korišćenju relativnog pokazatelja važnosti svake od kanoničkih diskriminacionih funkcija za razdvajanje grupa (udeo njenog karakterističnog korena u zbiru svih karakterističnih korena). Ukoliko ovaj relativan pokazatelj, kumulativno posmatrano za nekoliko prvih diskriminacionih funkcija prelazi neki unapred izabran procenat, recimo 90% varijacija između grupa, tada možemo sa velikom pouzdanošću dalju analizu zasnovati samo na tom broju diskriminacionih funkcija, a ne svih ocenjenih.

Držeći se principa ekonomičnosti: relativno uspešno objasniti glavni izvor varijacija između grupa sa malim brojem funkcija, najčešće se dimenzija diskriminacionog

prostora svodi na dva ili tri. Tada se, u cilju interpretacije dobijenih rezultata diskriminacione analize, preporučuje korišćenje grafičkog prikaza.

Prvi grafički prikaz je već korišćen u Primeru 6.3. Naime, centroidi grupa prikazani su u dvodimenzionom diskriminacionom prostoru. Na slici se jasno uočava razdvajanje tri grupe i njihova pozicija u diskriminacionom prostoru kao i njihov međusobni položaj. Takođe se međusobna sličanost pojedinih grupa bolje ističe u diskriminacionom prostoru nego na parcijalnim dijagramima rasturanja originalnih promenljivih. Ako su u analizi korišćene tri i više diskriminacionih funkcija onda se pribegava grafičkom prikazu diskriminacionih skorova u tzv. redukovanom diskriminacionom prostoru. Diskriminacioni skorovi prikazuju se u dvodimenzionim diskriminacionim prostorima koje generišu parovi diskriminacionih funkcija. Ovaj grafički prikaz se može dopuniti tako što će se oko svakog centroida grupa nacrtati krug koji će obuhvatiti unapred izrabran procenat opservacija te grupe. Ovde se radi o korišćenju ranije izloženog pojma oblasti poverenja. Sada je centar elipse poverenja lociran u centroidu grupe, a oblik mu je određen kovarijacionom matricom $\overline{\mathbf{S}}$. Pošto svaka grupa ima istu kovarijacionu matricu oblik elipse je isti kod svih grupa. Ako smo u analizi koristili standardizovane opservacije, tada je kovarijaciona matrica $\overline{\mathbf{S}}$ jednaka jediničnoj matrici, a elipsa poverenja postaje krug. Eventualno preklapanje nekih od ovih krugova sugeriše veću sličnost odgovarajućih grupa.

Grafički prikaz centroida u diskriminacionom prostoru pruža sumarno objašnjenje uočene razlike među njima, s obzirom na dve diskriminacione funkcije. Da bi se dalo detaljnije objašnjenje utvrđenih razlika potrebno je interpretirati same kanoničke diskriminacione funkcije.

Na osnovu veze koja postoji između regresione i diskriminacione analize, koeficijente kod kanoničkih diskriminacionih funkcija tretiramo kao regresione koeficijente u modelu regresione analize. Koeficijenti $\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, ..., \hat{\mathbf{a}}_s$ su nestandardizovani koeficijenti kanoničkih diskriminacionih funkcija. Ako originalne promenljive nisu prevedene u standardizovan oblik, tada se dobijeni koeficijenti $\hat{\mathbf{a}}$ mogu standardizovati tako što će se pomnožiti dijagonalnom matricom $\mathbf{D}^{1/2}$ čiji su elementi kvadratni koreni dijagonalnih elemenata opšte kovarijacione matrice $\overline{\mathbf{S}}$. Dakle, *standardizovani diskriminacioni koeficijenti* (ili ponderi) su $\hat{\mathbf{a}}^* = \mathbf{D}^{1/2}\hat{\mathbf{a}}$. Oni se interpretiraju kao i regresioni koeficijenti, što znači da ukazuju na parcijalni doprinos odnosne promenljive razdvajanju između grupa. Veličina relativnog doprinosa sagledava se na osnovu apsolutne vrednosti diskriminacionog koeficijenta, a njegov predznak ukazuje na smer tog uticaja. Problemi u korišćenju diskriminacionih koeficijenata u interpretaciji rezultata diskriminacione analize slični su onima u regresionoj analizi. Mala vrednost koeficijenta sugeriše ili irelevantnost te promenljive za diskriminacionu funkciju ili da je njen uticaj umanjen zbog problema multikolinearnosti. Nestabilnost diskriminacionih koeficijenata je takođe problem sa kojim treba računati prilikom njihovog korišćenja u interpretaciji rezultata.

Stoga se sugeriše, slično kao i u kanoničkoj korelacionoj analizi, da se interpretacija diskriminacionih funkcija bazira na koeficijentima korelacije originalnih promenljivih i diskriminacionih funkcija. Ove koeficijente nazivamo i *diskriminaciona opterećenja*, ili

strukturni koeficijenti korelacije. Oni ukazuju na zajedničku varijansu koju originalne promenljive dele sa diskriminacionom funkcijom i u tom smislu su manje podložne nestabilnosti uzrokovane korelacijom originalnih promenljivih. Za razliku od diskriminacionih koeficijenata, opterećenja mere kako su promenljive združeno povezane, a ne njihov marginalni uticaj. Pri njihovom korišćenju u interpretaciji rezultata potrebno je voditi računa da i oni mogu pokazati nestabilnost kao i diskriminacioni koeficijenti, pa se zahtevaju veći uzorci da bi se obezbedila njihova stabilnost.

Inače diskriminaciona opterećenja, u oznaci $\mathbf{r}_i$, računaju se na osnovu standardizovanih diskriminacionih koeficijenata, $\hat{\mathbf{a}}_i^*$, tako što se množe korelacionom matricom originalnih promenljivih $\mathbf{R}$

$$\mathbf{r}_i = \mathbf{R}\hat{\mathbf{a}}_i^*, \ i = 1, 2, ..., s. \tag{6.36}$$

Primer 6.5 Korišćenjem ocenjenih diskriminacionih funkcija i podataka iz Primera 6.2 izračunati diskriminaciona opterećenja. Komentarisati dobijene rezultate.

Nestandardizovani diskriminacioni koeficijenti su $\hat{\mathbf{a}}_1' = \begin{bmatrix} 0.8545 & 0.6152 \end{bmatrix}$ i $\hat{\mathbf{a}}_2' = \begin{bmatrix} 1.0374 & -0.7244 \end{bmatrix}$, a opšta kovarijaciona matrica je

$$\overline{\mathbf{S}} = \begin{bmatrix} 0.5714 & 0.1429 \\ 0.1429 & 1.1429 \end{bmatrix}, \text{ pa je matrica } \mathbf{D}^{1/2} = \begin{bmatrix} 0.7559 & 0 \\ 0 & 1.0690 \end{bmatrix}.$$

Na osnovu ovih elemenata standardizovani diskriminacioni koeficijenti su:

$$\hat{\mathbf{a}}_1^* = \mathbf{D}^{1/2}\hat{\mathbf{a}}_1 = \begin{bmatrix} 0.7559 & 0 \\ 0 & 1.0690 \end{bmatrix}\begin{bmatrix} 0.8545 \\ 0.6152 \end{bmatrix} = \begin{bmatrix} 0.6460 \\ 0.6577 \end{bmatrix}$$

$$\hat{\mathbf{a}}_2^* = \mathbf{D}^{1/2}\hat{\mathbf{a}}_2 = \begin{bmatrix} 0.7559 & 0 \\ 0 & 1.0690 \end{bmatrix}\begin{bmatrix} 1.0374 \\ -0.7244 \end{bmatrix} = \begin{bmatrix} 0.7842 \\ -0.7744 \end{bmatrix}.$$

Korelaciona matrica originalnih promenljivih je

$$\mathbf{R} = \begin{bmatrix} 1 & 0.4474 \\ 0.4474 & 1 \end{bmatrix},$$

pa su diskriminaciona opterećenja

$$\mathbf{r}_1 = \mathbf{R}\hat{\mathbf{a}}_1^* = \begin{bmatrix} 1 & 0.4474 \\ 0.4474 & 1 \end{bmatrix}\begin{bmatrix} 0.6460 \\ 0.6577 \end{bmatrix} = \begin{bmatrix} 0.9402 \\ 0.9467 \end{bmatrix}$$

$$\mathbf{r}_2 = \mathbf{R}\hat{\mathbf{a}}_2^* = \begin{bmatrix} 1 & 0.4474 \\ 0.4474 & 1 \end{bmatrix} \begin{bmatrix} 0.7842 \\ -0.7744 \end{bmatrix} = \begin{bmatrix} 0.4377 \\ -0.4236 \end{bmatrix}$$

Tako na primer, koeficijent korelacije prve originalne promenljive i druge kanoničke diskriminacione funkcije iznosi 0.4377, a druge originalne promenljive . Ovi koeficijenti korelacije su elementi vektora $\mathbf{r}_2$, a elementi vektora $\mathbf{r}_1$ sugerišu da su obe originalne promenljive u većoj meri povezane sa prvom, no sa drugom kanoničkom diskriminacionom funkcijom. Pošto su oba koeficijenta pozitivna, uostalom, kao i diskriminacioni koeficijenti, kažemo da sa porastom obe promenljive dobijamo višu vrednost prve diskriminacione funkcije. U slučaju druge diskriminacione funkcije, koja predstavlja kontrast prve i druge originalne promenljive, samo ako je porast prve relativno veći od porasta druge promenljive, porašće i vrednost diskriminacione funkcije. ■

Grafički prikaz centroida u redukovanom diskriminacionom prostoru može se dopuniti ucrtavanjem diskriminacionih opterećenja. U geometrijskom smislu oni, kao koeficijenti korelacije, predstavljaju kosinus ugla između dva vektora. U diskriminacioni prostor ucrtavamo vektor za svaku originalnu promenljivu. Dužina vektora treba da predoči relativan značaj odnosne promenljive u diskriminaciji između grupa. Njegov položaj, odnosno koordinate definišu diskriminaciona opterećenja uz prvu i drugu diskriminacionu funkciju. Kako su diskriminaciona opterećenja relativni pokazatelji (uzimaju vrednosti iz intervala od -1 do $+1$) to se koristi postupak tzv. *istezanja* vektora (eng. *stretching*) da bi se istakao njihov različit doprinos razdvajanju grupa. Jedan od postupaka istezanja vektora koristi pokazatelj doprinosa promenljive definisan kao $\hat{\mathbf{a}}_{ji}^2 / \hat{\lambda}_i$, odn. kvadrat opterećenja $j-$te promenljive na $i-$toj diskriminacionoj funkciji podeljen sa karakterističnim korenom pridruženim $i-$toj diskriminacionoj funkciji. Ovaj pokazatelj nazivamo *koeficijent istezanja*. Dobijenom vrednošću koeficijenta istezanja množe se koordinate vektora odnosne promenljive. Lokacija centroida na ovom grafikonu takođe se menja u odnosu na prethodni grafički prikaz. Svaka koordinata centroida množi se doprinosom odgovarajuće diskriminacione funkcije razdvajanju grupa, tj. koristimo pokazatelj $\hat{\lambda}_i$ da bismo njime množili $i-$tu vrednost koordinate centroida. Crtajući vektore originalnih promenljivih prema predloženoj proceduri mi ih usmeravamo što više u pravcu one grupe koja ima najveću srednju vrednost odnosne promenljive, a što dalje od grupe koja ima najmanju srednju vrednost te promenljive.

Primer 6.6 Korišćenjem ocenjenih diskriminacionih funkcija i opterećenja iz Primera 6.2 i Primera 6.5, grafički prikazati oba vektora originalnih promenljivih i centroida. Koristiti postupak istezanja vektora. Dati interpretaciju dobijenih rezultata.

Iz Primera 6.2 reprodukujemo sredine dve promenljive po grupama

Grupa	Promenljiva	
	Prva	Druga
Prva	2	1
Druga	0	2
Treća	-1	-2
Ukupno	0.3	0.5

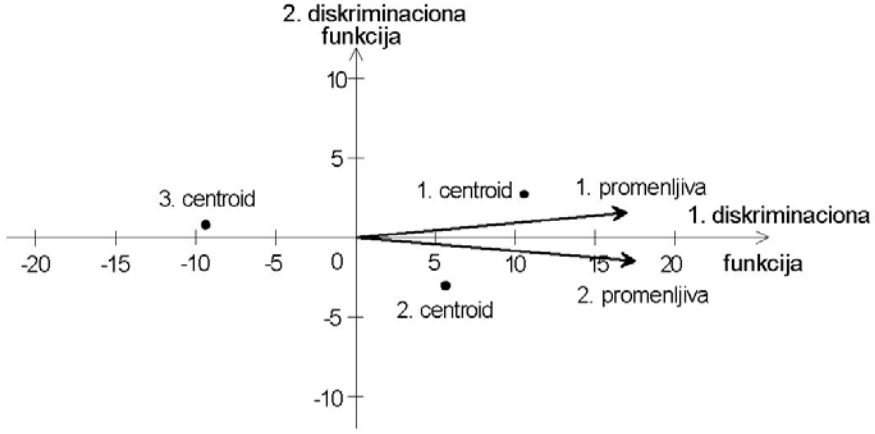
U prvoj tabeli dati su pokazatelji doprinosa svake diskriminacione funkcije razdvajanju grupa, a zatim vrednost centroida pre i nakon modifikacije.

Diskriminaciona funkcija	Karakteristični koren	Centroidi			Modifikovani centroidi		
		Prvi	Drugi	Treći	Prvi	Drugi	Treći
Prva	4.5889	2.3243	1.2304	-2.0849	10.67	5.65	-9.57
Druga	2.0498	1.3504	-1.4488	0.4114	2.77	-2.97	0.84

U drugoj tabeli data su diskriminaciona opterećenja pre i nakon istezanja.

Diskriminaciona funkcija	Opterećenja		Karakteristični koren	Koeficijenti istezanja		Modifikovana opterećenja	
	1. prom.	2. prom.		1. prom.	2. prom.	1. prom.	2. prom.
Prva	0.9402	0.9467	4.5889	19.26	19.53	17.03	17.50
Druga	0.4377	-0.4236	2.0498	9.35	8.75	1.79	1.57

Na osnovu dobijenih podataka prikazujemo u diskriminacionom prostoru kako centroide grupa tako i originalne promenljive.



Na osnovu grafičkog prikaza u diskriminacionom prostoru konstatujemo da je prva diskriminaciona funkcija osnovni izvor razlika između s jedne strane, prve i druge grupe i s druge strane, treće grupe. Položaj vektora sugeriše da su obe promenljive podjednako blizu prvoj diskriminacionoj funkciji, pa je upravo i takav značaj promenljivih u objašnjenju, odnosno opisu prve diskriminacione funkcije, s obzirom da su i dužine vektora

približno jednake. Uočiti da je vektor prve promenljive orijentisan ka centroidu prve, a vektor druge promenljive ka centroidu druge grupe. U oba slučaja vektori su najviše udaljeni od centroida treće grupe što je u skladu sa srednjim vrednostima prezentiranim u tabeli na početku ovog zadatka. Najveća srednja vrednost prve promenljive je 2 kod prve grupe, a najveća srednja vrednost druge promenljive je 2 kod druge grupe. Treća grupa ima najmanju srednju vrednost kod prve i druge promenljive, -1 i -2 respektivno.

Druga diskriminaciona funkcija objašnjava preostale razlike između grupa jasno izdvajajući prvu i treću grupu od druge. Ona predstavlja kontrast između prve i druge promenljive sa kojima je inače slabo povezana. ■

6.3 DISKRIMINACIONA FUNKCIJA I KLASIFIKACIJA

Drugi cilj diskriminacione analize jeste klasifikovanje ili alokacija opservacija nepoznatog porekla u jednu od grupa, odn. populacija. Videli smo da Fisherova linearna diskriminaciona funkcija može poslužiti i za alokaciju opservacija. U ovom poglavlju izložićemo opšti teorijski okvir klasifikacije opservacija i pokazaćemo da pod izvesnim uslovima dobijena Fisherova diskriminaciona funkcija ima optimalna svojstva. Prvo ćemo se pozabaviti slučajem dve populacije.

6.3.1 Klasifikacija za slučaj dve populacije

Navedimo situaciju u kojoj diskriminaciona analiza može biti od pomoći u donošenju odluke o alokaciji opservacija. Banka je suočena sa problemom da neki od klijenata nisu u stanju da vrate kredit na vreme. Zato je bančin interes da definiše objektivni kriterijum na osnovu koga će proceniti, u slučaju novog zahteva za zajmom, da li će podnosilac zahteva biti u stanju vratiti kredit na vreme ili ne. Za klijente koji su do tada podnosili zahteve postoje podaci o ukupnom porodičnom prihodu, veličini porodice, vrednosti nepokretne imovine, starosti glave porodice i dr. kao i podatak da li su ranije uzeti krediti vraćeni na vreme ili ne. Na osnovu ovih podataka određuje se kriterijum pomoću koga će se svi budući tražioci svrstati u grupu kreditno pouzdanih ili u grupu kojima je rizično dati kredit. Pri tom alociranju novog zahteva za kreditom banka može pogrešiti tako što će kreditno pouzdanog tražioca svrstati u grupu onih kojima ne bi trebalo odobriti kredit i obratno, odobriti kredit tražiocu koji neće biti u stanju vratiti kredit. U oba slučaja pogrešna alokacija zahteva za kreditom povlači po banku određene troškove u vidu neostvarenog prihoda u prvom slučaju, odn. direktnih gubitaka u drugom. Težnja banke je da se broj pogrešno klasifikovanih zahteva po oba navedena osnova smanji na najmanju moguću meru jer će time minimizirati svoj gubitak. Prema dosadašnjem iskustvu u banci je uočeno da je broj zahteva kreditno pouzdanih klijenata veći od broja zahteva klijenata kojima je rizično odobriti kredit.

Navedeni primer ilustruje koje sve elemente prilikom analize problema alokacije treba imati u vidu, odnosno kroz koje sve etape u analizi prolazimo da bismo doneli

ispravnu odluku. U primeru se razlikuju dve međusobom jasno razdvojene grupe: kreditno pouzdani klijenti (π_1) i oni kojima je rizično dati kredit (π_2). Označimo sa $\mathbf{X}$ slučajan vektor sa p promenljivih koje su merene kod svakog klijenta, a sa $f_1(\mathbf{x})$ i $f_2(\mathbf{x})$ funkciju gustine verovatnoće od $\mathbf{X}$, za prvu i drugu grupu respektivno. Naš je zadatak da prostor uzorka, odnosno sve moguće realizacije slučajnog vektora $\mathbf{X}$, podelimo u dve oblasti R_1 i R_2 tako da ako nova opservacija $\mathbf{X}_0$ pripada oblasti R_1, tada odgovarajućeg klijenta alociramo u grupu π_1, a ako pripada oblasti R_2, tada klijenta alociramo u grupu π_2. Očigledno je da klijent može biti klasifikovan samo u jednu od grupa, pa kažemo da su oblasti R_1 i R_2 međusobom isključive, a njihova unija pokriva ceo prostor uzorka. Označimo sa p_1 i p_2 apriorne verovatnoće da slučajno izabrani klijent potiče iz populacije π_1 i π_2 respektivno. Uz novi kreditni zahtev koji je podneo potencijalni klijent raspolažemo opservacijom $\mathbf{X}_0$, odnosno vrednostima p promenljivih na osnovu kojih donosimo odluku da li klijenta treba klasifikovati u oblast R_1 ili oblast R_2.

Iz teorijske statistike poznato je da se testiranje hipoteza može posmatrati u okviru statističke teorije odlučivanja. Neodbacivanje ili odbacivanje nulte hipoteze može biti korektna odluka, ali su u procesu odlučivanja moguće i greške, tzv. greške prve i druge vrste. Verovatnoće tih grešaka su uslovne, jer doneta odluka je uslovljena istinitošću nulte hipoteze. Na sličan način pristupamo problemu alokacije. Znači da u postupku klasifikacije klijenata možemo doneti ispravnu ili pogrešnu odluku. Klijenta ispravno klasifikujemo ako ga svrstamo ili u grupu π_1 ili u grupu π_2, a on zaista potiče iz te grupe. U druga dva slučaja kažemo da smo klijenta pogrešno klasifikovali. Prvi slučaj, klijent je potekao iz grupe π_1, a mi smo ga pogrešno svrstali u grupu π_2. Ovu uslovnu verovatnoću označavamo sa $P(2|1)$

$$P(2|1) = P(\mathbf{X} \in R_2 \mid \pi_1) = \int_{R_2} f_1(\mathbf{x}) d\mathbf{x} \tag{6.37}$$

Drugi slučaj, klijent je potekao iz grupe π_2, a mi smo ga pogrešno svrstali u grupu π_1. Ovu uslovnu verovatnoću označavamo sa $P(1|2)$

$$P(1|2) = P(\mathbf{X} \in R_1 \mid \pi_2) = \int_{R_1} f_2(\mathbf{x}) d\mathbf{x} \tag{6.38}$$

Sada možemo odrediti verovatnoću da klijent potiče iz $j-$te populacije i da smo ga alocirali u $i-$tu populaciju

$$P(\mathbf{X} \in R_i, \pi_j) = P(\mathbf{X} \in R_i \mid \pi_j) P(\pi_j) = P(i \mid j) \cdot p_j, \; i, j = 1, 2 . \tag{6.39}$$

Uvedimo poslednji element u analizu, troškove pogrešne alokacije. Ako smo ispravno klasifikovali klijenta ovi troškovi su jednaki nuli, a u suprotnom javljaju se u visini

$c(i \mid j)$ kada smo klijenta iz grupe π_j pogrešno alocirali u grupu π_i, pri čemu je $i \neq j$. Uzimajući u obzir moguće ishode klasifikacije klijenta, formiramo narednu tabelu na kojoj su prikazane sve odluke u pogledu klasifikacije klijenta, odgovarajuće verovatnoće i troškovi pogrešne alokacije.

Tabela 6.1 Ishodi klasifikacije, njihove verovatnoće i troškovi pogrešne alokacije

		Poreklo klijenta	
		π_1	π_2
Odluka	Klasifikuj u π_1	Ispravna klasifikacija $P(1\mid 1)\,p_1 \quad c(1\mid 1)=0$	Pogrešna klasifikacija $P(1\mid 2)\,p_2 \quad c(1\mid 2)$
	Klasifikuj u π_2	Pogrešna klasifikacija $P(2\mid 1)\,p_1 \quad c(2\mid 1)$	Ispravna klasifikacija $P(2\mid 2)\,p_2 \quad c(2\mid 2)=0$

Ukupna verovatnoća pogrešne klasifikacije, u oznaci E, jednaka je zbiru verovatnoća pogrešne klasifikacije iz Tabele 6.1

$$E = P(2\mid 1)\,p_1 + P(1\mid 2)\,p_2 \tag{6.40}$$

Ova verovatnoća naziva se i *stopa greške*. Sada se problem alokacije formuliše na sledeći način: Izvršiti takvu podelu prostora uzorka na oblasti R_1 i R_2 tako da se minimizira stopa greške E ili, ekvivalentno, da se maksimizira verovatnoća ispravne alokacije

$$1-E = P(1\mid 1)\,p_1 + P(2\mid 2)\,p_2 \tag{6.41}$$

Ako u analizu alokacije uključimo i troškove, tada je *očekivani trošak pogrešne klasifikacije*, u oznaci C

$$C = c(2\mid 1)P(2\mid 1)\,p_1 + c(1\mid 2)P(1\mid 2)\,p_2 \tag{6.42}$$

a problem alokacije formulišemo na sledeći način: Izvršiti takvu podelu prostora uzorka na oblasti R_1 i R_2 da se minimizira očekivani trošak pogrešne klasifikacije.

Pretpostavimo da nam je višedimenziona funkcija gustine verovatnoće dve grupe poznata, a takođe i apriorne verovatnoće p_1 i p_2. Zadatak je odrediti oblasti R_1 i R_2 tako da se

$$C = c(2\mid 1)\,p_1 \int_{R_2} f_1(\mathbf{x})d\mathbf{x} + c(1\mid 2)\,p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x} \tag{6.43}$$

minimizira. Uzimajući u obzir da je prostor uzorka jednak uniji oblasti R_1 i R_2 to je

$$1 = \int_{R_1 \cup R_2} f_1(\mathbf{x})d\mathbf{x} = \int_{R_1} f_1(\mathbf{x})d\mathbf{x} + \int_{R_2} f_1(\mathbf{x})d\mathbf{x}\,. \tag{6.44}$$

Sada je očekivani trošak pogrešne alokacije

$$C = c(2|1)p_1\left[1-\int_{R_1} f_1(\mathbf{x})d\mathbf{x}\right]+c(1|2)p_2\int_{R_1} f_2(\mathbf{x})d\mathbf{x}$$

$$=\int_{R_1}[c(1|2)p_2 f_2(\mathbf{x})-c(2|1)p_1 f_1(\mathbf{x})]d\mathbf{x}+c(2|1)p_1 \tag{6.45}$$

Po definiciji su apriorne verovatnoće p_1 i p_2 nenegativni brojevi kao i troškovi pogrešne alokacije. Funkcije gustine verovatnoće su nenegativne funkcije koje zavise od $\mathbf{x}$. To znači da se minimum očekivanih troškova pogrešne alokacije postiže za one vrednosti $\mathbf{x}$ koje pripadaju oblasti R_1 i za koje je podintegralna funkcija manja ili jednaka nuli

$$[c(1|2)p_2 f_2(\mathbf{x})-c(2|1)p_1 f_1(\mathbf{x})]\le 0 \tag{6.46}$$

Odavde sledi da oblast R_1 sadrži one tačke $\mathbf{x}$ za koje važi nejednakost

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\ge\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right], \tag{6.47}$$

a oblast R_2 sadrži one tačke $\mathbf{x}$ za koje važi nejednakost

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}<\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]. \tag{6.48}$$

Oblasti R_1 i R_2 definisane gornjim nejednakostima minimiziraju očekivane troškove pogrešne klasifikacije. Kada su troškovi pogrešne alokacije jednaki međusobom, tada su oblasti date sledećim nejednakostima

$$R_1:\ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\ge\frac{p_2}{p_1};\qquad R_2:\ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}<\frac{p_2}{p_1} \tag{6.49}$$

Do istog rezultata dolazimo i minimiziranjem stope greške E, odnosno maksimiziranjem verovatnoće ispravne alokacije. Ako su i apriorne verovatnoće jednake međusobom tada su oblasti definisane nejednakostima

$$R_1:\ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\ge 1;\qquad R_2:\ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}<1 \tag{6.50}$$

Poslednji slučaj definicije oblasti se najčešće javlja u praktičnoj primeni. Za novu opservaciju $\mathbf{x}_0$ računa se vrednost funkcije gustine verovatnoće, pa ako je $f_1(\mathbf{x}_0)\ge f_2(\mathbf{x}_0)$, tada se alocira u π_1, a ako je $f_1(\mathbf{x}_0)< f_2(\mathbf{x}_0)$ tada se $\mathbf{x}_0$ alocira u π_2. Ovo pravilo alokacije naziva se i *pravilo najveće verodostojnosti*.

Primer 6.7 Pretpostavimo da nam je poznata funkcija gustine verovatnoće kod dve grupe. Na slučaj izabrana opservacija ima apriornu verovatnoću $p_1 = 0.6$ da je izvučena iz grupe π_1 i verovatnoću $p_2 = 0.4$ da je izvučena iz grupe π_2. Ako su troškovi pogrešne klasifikacije $c(2|1) = 4$ i $c(1|2) = 9$ izvesti oblasti R_1 i R_2. Novu opservaciju $\mathbf{x}_0$ alocirati u jednu od grupa ako je dato $f_1(\mathbf{x}_0) = 0.2$ i $f_2(\mathbf{x}_0) = 0.5$.

Definicija oblasti je

$$R_1: \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right] = \left[\frac{9}{4}\right]\left[\frac{0.4}{0.6}\right] = 1.5$$

$$R_2: \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right] = \left[\frac{9}{4}\right]\left[\frac{0.4}{0.6}\right] = 1.5$$

Kako je količnik verodostojnosti

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = \frac{0.2}{0.5} = 0.4$$

manji od vrednosti 1.5 konstatujemo da nova opservacija pripada oblasti R_2 i da je treba alocirati u grupu π_2.

Zanemarivanjem troškova dobijamo novu definiciju oblasti

$$R_1: \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} = \frac{0.4}{0.6} = 0.67; \qquad R_2: \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1} = \frac{0.4}{0.6} = 0.67$$

I u ovom slučaju nova opservacija pripada oblasti R_2 pošto je količnik verodostojnosti manji od 0.67, pa opservaciju alociramo u grupu π_2.

Naposletku, ako zanemarimo apriorne verovatnoće dobijamo novu definiciju oblasti

$$R_1: \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1; \qquad R_2: \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

Kao i u prethodna dva slučaja i u ovom nova opservacija pripada oblasti R_2, pošto je količnik verodostojnosti manji od 1. Znači da se opservaciju alocira u grupu π_2. ■

Navedeni postupci određivanja oblasti R_1 i R_2, odnosno kriterijumi alokacije u literaturi se izlažu pod različitim nazivima. Ako pored funkcije gustine verovatnoća

raspolažemo apriornim verovatnoćama i troškovima pogrešne klasifikacije, koristi se Bayesov rizik koji predstavlja aposteriorni očekivani gubitak. Na njemu zasnovan *Bayesov kriterijum* alokacije zahteva klasifikaciju opservacije u onu grupu koja ima najmanji Bayesov rizik. Osnovna varijanta Bayesovog pravila diskriminacije računa samo sa apriornim verovatnoćama, a ne i troškovima, što je ekvivalentno (6.49). Ova varijanta Bayesovog pravila diskriminacije uz jednake apriorne verovatnoće ekvivalentna je pravilu diskriminacije baziranom na najvećoj verodostojnosti (oblast definisana izrazom (6.50)). U situacijama kada jedna od grupa ima malu apriornu verovatnoću koristi se tzv. *minimax pristup* prema kome se minimizira maksimalna verovatnoća pogrešne klasifikacije. Detaljnije o alternativnim kriterijumima klasifikacije kao i ograničenjima uz koja se uspostavlja relacija ekvivalencije među njima, videti kod Mardia, Kent i Bibby (1979, ss. 301-308).

Specijalan slučaj alokacije opservacija je kada su populacije π_1 i π_2 višedimenzione normalno raspoređene sa poznatim sredinama $\boldsymbol{\mu}_1$ i $\boldsymbol{\mu}_2$ i poznatim kovarijacionim matricama $\boldsymbol{\Sigma}_1$ i $\boldsymbol{\Sigma}_2$. Osnovni rezultat u ovom slučaju ukazuje na to da svi kriterijumi daju identičan rezultat kao i Fisherov diskriminacioni kriterijum, dakle i kriterijum najveće verodostojnosti, minimax kriterijum, kriterijum minimalne ukupne verovatnoće pogrešne klasifikacije i Bayesov kriterijum sa jednakim apriornim verovatnoćama i troškovima.

Količnik verodostojnosti je

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\left[(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_1\right|^{1/2}\right]^{-1}\exp\left[(-1/2)(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right]}{\left[(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_2\right|^{1/2}\right]^{-1}\exp\left[(-1/2)(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right]}$$

$$= \frac{\left|\boldsymbol{\Sigma}_2\right|^{1/2}}{\left|\boldsymbol{\Sigma}_1\right|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] \tag{6.51}$$

odakle se nakon sređivanja dobijaju oblasti R_1 i R_2 koji daju minimalne očekivane troškove pogrešne klasifikacije

$$R_1 : -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\mathbf{x}+(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x} \geq$$

$$\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]+\frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)+\frac{1}{2}\left(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right) \tag{6.52}$$

$$R_2 : -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\mathbf{x}+(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x} <$$

$$\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]+\frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)+\frac{1}{2}\left(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right) \tag{6.53}$$

Ako nova opservacija $\mathbf{x}_0$ pripada oblasti R_1 tada je alociramo u π_1, u suprotnom je alociramo u π_2. Skrećemo pažnju na prvi član s leve strane nejednakosti (6.52) i (6.53), tj. na $-\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\mathbf{x}$. Zbog njega su oblasti definisane kao kvadratne funkcije od $\mathbf{x}$. Zato se ova funkcija za slučaj višedimenzionog normalnog rasporeda sa različitim kovarijacionim matricama $\boldsymbol{\Sigma}_1$ i $\boldsymbol{\Sigma}_2$ naziva *kvadratna diskriminaciona funkcija*.

Kada je $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2$ definicije oblasti klasifikacije pojednostavljuju se i sada glase

$$R_1:(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\geq\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{6.54}$$

$$R_1:(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)<\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{6.55}$$

a pravilo alokacije sugeriše da novu opservaciju $\mathbf{x}_0$ koja pripada oblasti R_1 klasifikujemo u π_1, a ako pripada oblasti R_2 klasifikujemo u π_2.

Ako na ovom mestu uvedemo pretpostavku o jednakim troškovima i apriornim verovatnoćama, tada će vrednost logaritma na desnoj strani nejednakosti biti jednaka nuli, a definicija regiona klasifikacije biti istovetna Fisherovoj linearnoj diskriminacionoj funkciji. To znači da je Fisherova diskriminaciona funkcija u slučaju normalnog rasporeda sa jednakim kovarijacionim matricama optimalno sredstvo za alokaciju novih opservacija po kriterijumu minimalnih očekivanih troškova pogrešne klasifikacije uz jednake troškove i apriorne verovatnoće.

Populacioni parametri, sredine i kovarijacione matrice, u praktičnoj primeni diskriminacione analize su nepoznati parametri, pa ih zamenjujemo odgovarajućim uzoračkim pokazateljima. Tako na primer, pravilo alokacije nove opservacije $\mathbf{x}_0$ u slučaju normalnog rasporeda sa jednakim kovarijacionim matricama uz kriterijum minimalnih očekivanih troškova pogrešne klasifikacije glasi:

Alociraj $\mathbf{x}_0$ u π_1 ako je

$$(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'\bar{\mathbf{S}}^{-1}\mathbf{x}_0-\frac{1}{2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'\bar{\mathbf{S}}^{-1}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2)\geq\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

u suprotnom alociraj $\mathbf{x}_0$ u π_2. (6.56)

Na sličan način izražavamo pravilo alokacije kada ne možemo usvojiti pretpostavku da su kovarijacione matrice jednake međusobom, tj. u slučaju kvadratne diskriminacione funkcije.

6.3.2 Klasifikacija za slučaj više populacija

Uopštenje postupka alokacije opservacija u slučaju više grupa direktno proizlazi iz pristupa izloženog za dve populacije. Sada je zadatak da se izvrši takva podela prostora uzorka na oblasti $R_1, R_1, ..., R_g$, da se ukupna verovatnoća ispravne klasifikacije maksimizira ili ekvivalentno, da se minimizira ukupna verovatnoća pogrešne klasifikacije. Ako u razmatranje uključimo troškove pogrešne alokacije tada oblasti određujemo tako da se minimiziraju očekivani troškovi pogrešne alokacije.

Označimo sa $f_k(\mathbf{x})$, $k = 1, 2, ..., g$ funkciju gustine verovatnoće slučajnog vektora $\mathbf{X}$ ako ovaj potiče iz populacije π_k. Apriornu verovatnoću da će na slučaj izvučen element iz svih združeno posmatranih populacija pripadati populaciji π_k označimo sa p_k, a trošak pogrešne alokacije elementa u $k-$tu populaciju kada on ustvari potiče iz $i-$te populacije sa $c(k \mid i)$, $k, i = 1, 2, ..., g$.

Sada je očekivani trošak pogrešne alokacije

$$C = \sum_{i=1}^{g} p_i \left[\sum_{\substack{k=1 \\ k \neq i}}^{g} P(k \mid i) c(k \mid i) \right] \tag{6.57}$$

gde smo sa $P(k \mid i)$ označili uslovnu verovatnoću pogrešne klasifikacije elementa koji potiče iz $i-$te populacije, a svrstan je u $k-$tu populaciju.

Pokazano[3] je da se minimalna vrednost očekivanog troška pogrešne alokacije postiže takvom podelom prostora uzorka da se opservacija $\mathbf{x}$ alocira u populaciju π_k, $k = 1, 2, ..., g$ za koju je

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(\mathbf{x}) c(k \mid i) < \sum_{\substack{i=1 \\ i \neq j}}^{g} p_i f_i(\mathbf{x}) c(j \mid i), \quad j = 1, 2, ..., g, \quad j \neq k \tag{6.58}$$

Ako za neki indeks j nejednakost (6.58) ne važi, odnosno može se zameniti jednakošću, tada se opservacija može alocirati u ma koju populaciju, kod koje se to desilo.

Specijalan slučaj gornjeg pravila alokacije javlja se kada su jednaki troškovi, na primer kada je $c(k \mid i) = 1$, za svako k, i $(k \neq i)$. Sada se opservacija $\mathbf{x}$ alocira u π_k, ako za tu populaciju važi

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(\mathbf{x}) < \sum_{\substack{i=1 \\ i \neq j}}^{g} p_i f_i(\mathbf{x}), \quad j \neq k \tag{6.59}$$

[3] Videti: Anderson (1972, ss. 143-4).

Oduzimajući $\sum_{\substack{i=1 \\ i\neq k,j}}^{g} p_i f_i(\mathbf{x})$ od izraza na levoj i desnoj strani nejednakosti, dobijamo

$$p_j f_j(\mathbf{x}) < p_k f_k(\mathbf{x}) \tag{6.60}$$

što znači da je opservacija $\mathbf{x}$ u oblasti R_k, ako se za tu vrednost indeksa postiže maksimalna vrednost od $p_i f_i(\mathbf{x})$, odnosno kada je π_k najverovatnija populacija. Primena ovog pravila alokacije zahteva poznavanje stvarnih ili ocenjenih vrednosti funkcije gustine verovatnoća, apriornih verovatnoća i troškova alokacije.

Primer 6.8 Primeniti pravilo alokacije zasnovano na minimalnoj vrednosti očekivanih troškova pogrešne alokacije (posebno uz pretpostavku jednakih troškova) na opservaciju $\mathbf{x}_0$ za koju znamo da potiče iz jedne od tri populacije i za koje su date apriorne verovatnoće: $p_1 = 0.2$, $p_2 = 0.3$, $p_3 = 0.5$, vrednost funkcije gustine za $\mathbf{x}_0$: $f_1(\mathbf{x}_0) = 0.3$, $f_2(\mathbf{x}_0) = 0.6$, $f_3(\mathbf{x}_0) = 1.2$ i troškovi pogrešne alokacije

		Potiče iz		
		π_1	π_2	π_3
Klasifikuj u	π_1	$c(1\|1) = 0$	$c(1\|2) = 100$	$c(1\|3) = 100$
	π_2	$c(2\|1) = 400$	$c(2\|2) = 0$	$c(2\|3) = 50$
	π_3	$c(3\|1) = 600$	$c(3\|2) = 200$	$c(3\|3) = 0$

Izraz $\sum_{\substack{i=1 \\ i\neq k}}^{g} p_i f_i(\mathbf{x}) c(k\,|\,i)$ za tri populacije u razvijenom obliku glasi:

$$k=1:\ p_2 f_2(\mathbf{x}_0)c(1|2) + p_3 f_3(\mathbf{x}_0)c(1|3) = 0.3\cdot 0.6\cdot 100 + 0.5\cdot 1.2\cdot 100 = 78$$

$$k=2:\ p_1 f_1(\mathbf{x}_0)c(2|1) + p_3 f_3(\mathbf{x}_0)c(2|3) = 0.2\cdot 0.3\cdot 400 + 0.5\cdot 1.2\cdot 50 = 54$$

$$k=3:\ p_1 f_1(\mathbf{x}_0)c(3|1) + p_2 f_2(\mathbf{x}_0)c(3|2) = 0.2\cdot 0.3\cdot 600 + 0.3\cdot 0.6\cdot 200 = 72$$

Na osnovu dobijenih vrednosti zaključujemo da prema kriterijumu najmanjih očekivanih troškova pogrešne alokacije opservaciju $\mathbf{x}_0$ treba svrstati u π_2.

Uz pretpostavku jednakih troškova potrebno je izračunati vrednost $p_i f_i(\mathbf{x}_0)$

$$p_1 f_1(\mathbf{x}_0) = 0.2\cdot 0.3 = 0.06$$

$$p_2 f_2(\mathbf{x}_0) = 0.3 \cdot 0.6 = 0.18$$

$$p_3 f_3(\mathbf{x}_0) = 0.5 \cdot 1.2 = 0.6$$

Kako se najveća vrednost ovog izraza postiže za $k = 3$ to znači da novu opservaciju treba svrstati u populaciju π_3. ■

Specijalan slučaj klasifikacije opservacija dobijamo uvođenjem pretpostavke o normalnom rasporedu populacija $\pi_1, \pi_2, ..., \pi_g$, tj. kada je $\mathbf{X} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 1, 2, ..., g$. Uz dodatnu pretpostavku jednakih troškova pogrešne alokacije, problem se svodi na određivanje oblasti tako da se minimizira ukupna verovatnoća pogrešne klasifikacije. Na osnovu izraza funkcije gustine verovatnoće normalno raspoređenog slučajnog vektora $\mathbf{X}$, zanemarujući konstantu $(2\pi)^{p/2}$ koja je jednaka kod svih populacija, pravilo alokacije glasi

Alociraj $\mathbf{x}$ u π_k ako je

$$y_k^Q(\mathbf{x}) = \ln p_k - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) = \max_i y_i^Q(\mathbf{x}),\ i = 1, 2, ..., g \qquad (6.61)$$

gde $y_k^Q(\mathbf{x})$ nazivamo kvadratni diskriminacioni skor, jer u njegovom izrazu figuriše, slično kao i u slučaju dve populacije, kvadratna funkcija od $\mathbf{x}$ (kvadrat odstojanja $\mathbf{x}$ od sredine $\boldsymbol{\mu}_k$).

Kada je opravdano pretpostaviti da su kovarijacione matrice populacija međusobom jednake ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$) pravilo alokacije glasi

Alociraj $\mathbf{x}$ u π_k ako je

$$y_k(\mathbf{x}) = \boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p_k = \max_i y_i(\mathbf{x}),\ i = 1, 2, ..., g \qquad (6.62)$$

gde smo sa $y_k(\mathbf{x})$ označili linearni diskriminacioni skor.

U empirijskom istraživanju populacioni parametri su nam nepoznati, pa ih u definiciji pravila alokacije zamenjujemo odgovarajućim uzoračkim ocenama. To znači da ćemo u slučaju da se ne može usvojiti pretpostavka o jednakim kovarijacionim matricama za alokaciju opservacija koristiti ocenjene kvadratne diskriminacione skorove $\hat{y}_i^Q(\mathbf{x})$. Među njima biramo najveći, da bi tu opservaciju svrstali u odgovarajuću populaciju. Uz pretpostavku jednakih kovarijacionih matrica umesto kvadratnih, koristimo ocenjene linearne diskriminacione skorove $\hat{y}_i(\mathbf{x})$. Najveći ocenjeni linearni diskriminacioni skor sugeriše da opservaciju treba svstati u odnosnu populaciju za koju je postignuta ta maksimalna vrednost skora.

Primer 6.9 Korišćenjem podataka iz Primera 6.2 odrediti linearne diskriminacione skorove za novu opservaciju $\mathbf{x}_0' = \begin{bmatrix} x_{01} & x_{02} \end{bmatrix} = \begin{bmatrix} -1 & 2 \end{bmatrix}$ ako je $p_1 = 0.2$, $p_2 = 0.3$ i $p_3 = 0.5$.

Kako su uzoračke sredine i opšta kovarijaciona matrica

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix},\ \bar{\mathbf{x}}_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix},\ \bar{\mathbf{x}}_3 = \begin{bmatrix} -1 \\ -2 \end{bmatrix},\ \bar{\mathbf{S}} = \frac{1}{7}\begin{bmatrix} 4 & 1 \\ 1 & 8 \end{bmatrix}$$

to je

$$\bar{\mathbf{x}}_1'\bar{\mathbf{S}}^{-1} = \begin{bmatrix} 2 & 1 \end{bmatrix}\frac{7}{31}\begin{bmatrix} 8 & -1 \\ -1 & 4 \end{bmatrix} = \frac{7}{31}\begin{bmatrix} 15 & 2 \end{bmatrix}$$

$$\bar{\mathbf{x}}_1'\bar{\mathbf{S}}^{-1}\bar{\mathbf{x}}_1 = \frac{7}{31}\begin{bmatrix} 15 & 2 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \frac{224}{31}$$

Sada je prvi ocenjeni linearni diskriminacioni skor

$$\hat{y}_1(\mathbf{x}_0) = \ln p_1 - \frac{1}{2}\bar{\mathbf{x}}_1'\bar{\mathbf{S}}^{-1}\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_1'\bar{\mathbf{S}}^{-1}\mathbf{x}_0$$

$$= \ln(0.2) - \frac{1}{2}\left(\frac{224}{31}\right) + \frac{105}{31}x_{01} + \frac{14}{31}x_{02}$$

$$= 5.2223 + \frac{105}{31}x_{01} + \frac{14}{31}x_{02}$$

Na sličan način određujemo preostala dva linearna diskriminaciona skora

$$\hat{y}_2(\mathbf{x}_0) = -3.0104 - \frac{14}{31}x_{01} + \frac{56}{31}x_{02}$$

$$\hat{y}_3(\mathbf{x}_0) = -2.9512 - \frac{42}{31}x_{01} - \frac{49}{31}x_{02}$$

Za konkretnu vrednost nove opservacije imamo sledeće vrednosti linearnih diskriminacionih skorova: $\hat{y}_1(\mathbf{x}_0) = -7.706$, $\hat{y}_2(\mathbf{x}_0) = 1.054$ i $\hat{y}_3(\mathbf{x}_0) = -4.758$. Kako je $\hat{y}_2(\mathbf{x}_0) = 1.054$ najveći linearni diskriminacioni skor, novu opservaciju $\mathbf{x}_0$ alociramo u π_2. Podsećamo da smo istu odluku u pogledu alokacije ove opservacije doneli i na osnovu Fisherovih diskriminacionih funkcija. ■

Alternativan, ali ekvivalentan pristup alokaciji opservacija u slučaju više grupa zasnovan je na korišćenju količnika verodostojnosti $f_k(\mathbf{x})/f_i(\mathbf{x})$ kao diskriminacione funkcije. Na osnovu količnika verodostojnosti za normalan raspored sa jednakim kovarijacionim matricama uz jednake troškove pravilo alokacije glasi

Alociraj $\mathbf{x}$ u π_k ako je

$$y_{ki}(\mathbf{x}) = (\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_i) \geq \ln\left(\frac{p_i}{p_k}\right), \tag{6.63}$$

za svako $i = 1, 2, ..., g$, $i \neq k$

Dakle, $y_{ki}(\mathbf{x})$ je linearna diskriminaciona funkcija od $X_1, X_2, ..., X_p$ i služi za alokaciju opservacija ili u populaciju π_k ili u π_i. Sada je za definiciju svake oblasti R_k potrebno izračunati $g-1$ diskriminacionih funkcija, a za specifikaciju svih oblasti $g(g-1)/2$ takvih funkcija. Potrebno je odrediti samo $g(g-1)/2$ funkcija pošto jedna diskriminaciona funkcija istovremeno reprezentuje količnik $f_k(\mathbf{x})/f_i(\mathbf{x})$ i njegovu recipročnu vrednost, odn. $f_k(\mathbf{x})/f_i(\mathbf{x})$. Praktično se problem alokacije u više grupa tretira kao problem alokacije u dve grupe, pri čemu se posmatraju svi mogući parovi grupa. Tako na primer, za slučaj tri grupe $(g=3)$ postoje tri oblasti, a svaka je određena dvema diskriminacionim funkcijama $(g-1=2)$. Kako je $g(g-1)/2 = 3$, znači da ima ukupno tri linearne diskriminacione funkcije u specifikaciji ove tri oblasti. Ove oblasti su definisani trima funkcijama na sledeći način:

$$R_1: \; y_{12}(\mathbf{x}) \geq \ln\left(\frac{p_2}{p_1}\right) \text{ i } y_{13}(\mathbf{x}) \geq \ln\left(\frac{p_3}{p_1}\right) \tag{6.64}$$

$$R_2: \; y_{21}(\mathbf{x}) = -y_{12}(\mathbf{x}) \geq \ln\left(\frac{p_1}{p_2}\right) \text{ i } y_{23}(\mathbf{x}) \geq \ln\left(\frac{p_3}{p_2}\right) \tag{6.65}$$

$$R_3: \; y_{31}(\mathbf{x}) = -y_{13}(\mathbf{x}) \geq \ln\left(\frac{p_1}{p_3}\right) \text{ i } y_{32}(\mathbf{x}) = -y_{23}(\mathbf{x}) \geq \ln\left(\frac{p_2}{p_3}\right) \tag{6.66}$$

Kako je $y_{23}(\mathbf{x}) = y_{13}(\mathbf{x}) - y_{12}(\mathbf{x})$ dovoljno je koristiti diskriminacione skorove $y_{12}(\mathbf{x})$ i $y_{13}(\mathbf{x})$ za klasifikaciju opservacija. Uzimajući ovo u obzir, kao i to da je $y_{21}(\mathbf{x}) = -y_{12}(\mathbf{x})$, $y_{31}(\mathbf{x}) = -y_{13}(\mathbf{x})$ i $y_{32}(\mathbf{x}) = -y_{23}(\mathbf{x})$ pravilo alokacije u slučaju tri populacije glasi:

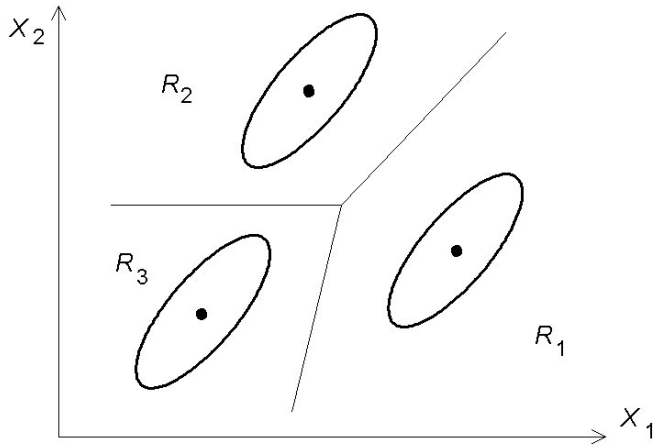
Alociraj $\mathbf{x}$ u

$$\pi_1 \text{ ako je } y_{12}(\mathbf{x}) \geq \ln\left(\frac{p_2}{p_1}\right) \text{ i } y_{13}(\mathbf{x}) \geq \ln\left(\frac{p_3}{p_1}\right)$$

$$\pi_2 \text{ ako je } y_{12}(\mathbf{x}) < \ln\left(\frac{p_1}{p_2}\right) \text{ i } y_{13}(\mathbf{x}) \geq y_{12}(\mathbf{x}) + \ln\left(\frac{p_3}{p_2}\right) \tag{6.67}$$

$$\pi_3 \text{ ako je } y_{13}(\mathbf{x}) < \ln\left(\frac{p_1}{p_3}\right) \text{ i } y_{12}(\mathbf{x}) \geq y_{13}(\mathbf{x}) + \ln\left(\frac{p_2}{p_3}\right)$$

Odgovarajuće uzoračke pokazatelje koristimo u linearnoj diskriminacionoj funkciji $y_{ki}(\mathbf{x})$ da bismo formirali njenu ocenu u uzorku. Kao što je to pokazano u Primeru 6.10. Grafički prikaz problema alokacije u tri populacije sa dve promenljive dat je na Slici 6.2.



Slika 6.2 Oblasti klasifikacije u slučaju tri populacije

Primer 6.10 Korišćenjem podataka iz Primera 6.2 odrediti sve diskriminacione funkcije za parove grupa i primeniti izloženo pravilo alokacije na novu opservaciju $\mathbf{x}_0' = \begin{bmatrix} -1 & 2 \end{bmatrix}$, ako su apriorne verovatnoće $p_1 = 0.2$, $p_2 = 0.3$ i $p_3 = 0.5$.

Ocenjena linearna diskriminaciona funkcija koja razdvaja $k-$tu i $i-$tu populaciju je

$$\hat{y}_{ki}(\mathbf{x}) = (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_i)'\bar{\mathbf{S}}^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_i)'\bar{\mathbf{S}}^{-1}(\bar{\mathbf{x}}_k + \bar{\mathbf{x}}_i)$$

Na osnovu ranije određenih uzoračkih pokazatelja (videti Primer 6.2) računamo

$$(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_i)'\bar{\mathbf{S}}^{-1} = \begin{bmatrix} 2-0 & 1-2 \end{bmatrix}' \frac{7}{31}\begin{bmatrix} 8 & -1 \\ -1 & 4 \end{bmatrix} = \frac{7}{31}\begin{bmatrix} 17 & -6 \end{bmatrix}$$

$$(\overline{\mathbf{x}}_k - \overline{\mathbf{x}}_i)'\overline{\mathbf{S}}^{-1}(\overline{\mathbf{x}}_k + \overline{\mathbf{x}}_i) = \frac{7}{31}\begin{bmatrix}17 & -6\end{bmatrix}\begin{bmatrix}2+0\\1+2\end{bmatrix} = \frac{112}{31}$$

pa je

$$\hat{y}_{12}(\mathbf{x}) = -\frac{1}{2}\cdot\frac{112}{31} + \frac{7}{31}\begin{bmatrix}17 & -6\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} = -\frac{56}{31} + \frac{119}{31}x_1 - \frac{42}{31}x_2$$

Na sličan način određujemo preostale dve diskriminacione funkcije parova

$$\hat{y}_{13}(\mathbf{x}) = -\frac{42}{31} + \frac{147}{31}x_1 + \frac{63}{31}x_2$$

$$\hat{y}_{23}(\mathbf{x}) = \frac{14}{31} + \frac{28}{31}x_1 + \frac{105}{31}x_2$$

Pokazati da važi relacija $\hat{y}_{23}(\mathbf{x}) = \hat{y}_{13}(\mathbf{x}) - \hat{y}_{12}(\mathbf{x})$ i u slučaju ocenjenih diskriminacionih funkcija parova.

Za novu opservaciju $\mathbf{x}_0' = \begin{bmatrix}-1 & 2\end{bmatrix}$ imamo

$$\hat{y}_{12}(\mathbf{x}_0) = -\frac{56}{31} + \frac{119}{31}x_{01} - \frac{42}{31}x_{02} = -\frac{56}{31} + \frac{119}{31}\cdot(-1) - \frac{42}{31}\cdot 2 = -8.3$$

Slično određujemo $\hat{y}_{13}(\mathbf{x}_0) = -2.03$ i $\hat{y}_{23}(\mathbf{x}_0) = 6.32$. Odredimo još i

$$\ln\left(\frac{p_2}{p_1}\right) = 0.41,\ \ln\left(\frac{p_1}{p_2}\right) = -0.41,\ \ln\left(\frac{p_3}{p_1}\right) = 0.92$$

$$\ln\left(\frac{p_1}{p_3}\right) = -0.92,\ \ln\left(\frac{p_3}{p_2}\right) = 0.51,\ \ln\left(\frac{p_2}{p_3}\right) = -0.51$$

Za novu opservaciju imamo da su ispunjene nejednakosti

$$\hat{y}_{12}(\mathbf{x}_0) < \ln\left(\frac{p_1}{p_2}\right) \text{ i } \hat{y}_{13}(\mathbf{x}_0) \geq \hat{y}_{12}(\mathbf{x}_0) + \ln\left(\frac{p_3}{p_2}\right)$$

pa ovu opservaciju alociramo u populaciju π_2. Dobijen je isti rezultat u pogledu alokacije nove opservacije kao i u slučaju korišćenja Fisherove diskriminacione funkcije (Primer 6.2) i diskriminacionih skorova (Primer 6.9) uz bitnu razliku što kod Fisherovih diskriminacionih funkcija nismo koristili apriorne verovatnoće. ■

6.3.3 Vrednovanje funkcije klasifikacije

Definicija diskriminacionih funkcija kao sredstva za alokaciju opservacija bazirana je na minimiziranju ukupne verovatnoće pogrešne alokacije. U dobijenim izrazima za oblasti, odnosno pravilima za alokaciju, figurišu nam kao poznate veličine funkcije gustine verovatnoće. U tom slučaju relativno je jednostavno izračunati minimalnu vrednost ukupne verovatnoće pogrešne alokacije. Problem je grafički prikazan na Slici 6.1, gde diskriminacioni skorovi u slučaju dve normalno raspoređene populacije imaju normalan raspored. Određivanje verovatnoća pogrešne klasifikacije $P(2|1)$ i $P(1|2)$ (šrafirane površine pod krivom) kao elementa ukupne verovatnoće pogrešne klasifikacije predstavlja problem u teoriji verovatnoće koji se rutinski rešava. Na osnovu izračunate vrednosti te verovatnoće možemo proceniti kakav je kvalitet diskriminacionih funkcija pri alokaciji opservacija. Za nas je interesantniji slučaj kada funkcije gustine verovatnoće nisu poznate. Tada na osnovu slučajnog uzorka treba proceniti u kojoj meri diskriminaciona funkcija uspešno svrstava opservacije u odgovarajuće populacije, a u koliko slučajeva to radi pogrešno.

Da bi se definisala mera uspešnosti klasifikacije koja se ne zasniva na poznatom ili pretpostavljenom obliku funkcije gustine verovatnoće, definišemo *matricu konfuzije* koja prikazuje broj ispravno i pogrešno klasifikovanih opservacija po grupama. Neka se problem alokacije odnosi na slučaj dve grupe π_1 i π_2 odakle smo uzeli uzorak od po n_1 i n_2 opservacije respektivno. Matrica konfuzije data je Tabelom 6.2 za slučaj dve populacije.

Tabela 6.2 Matrica konfuzije - slučaj dve populacije

		Stvarna pripadnost	
		π_1	π_2
Predviđena pripadnost	π_1	n_{11}	n_{12}
	π_2	n_{21}	n_{22}
Veličina uzorka		$n_1 = n_{11} + n_{21}$	$n_2 = n_{12} + n_{22}$

U tabeli smo sa n_{ij} označili broj opservacija koje potiču iz $j-$te populacije, a svrstane su u $i-$tu populaciju. Elementi na glavnoj dijagonali, n_{ii} (n_{11} i n_{22} u slučaju dve populacije) predstavljaju broj opservacija koje su ispravno alocirane, dok vandijagonalni elementi (n_{12} i n_{21} slučaju dve populacije) predstavljaju broj opservacija koje su pogrešno alocirane. Ako ne bi bilo grešaka u alokaciji opservacija tada bi bilo $n_{11} + n_{22} = n$, odn. matrica konfuzije bila bi dijagonalna matrica (upravo takav slučaj matrice konfuzije imamo u Primeru 6.2 gde smo sve opservacije iz uzorka ispravno klasifikovali).

Intuitivno se nameće korišćenje n_{ij}/n_j kao ocene uslovne verovatnoće $P(i|j)$ da će opservacija koja potiče iz $j-$te biti svrstana u $i-$tu populaciju. Znači da je u slučaju

dve grupe n_{12}/n_2 ocena od $P(1|2)$, a n_{21}/n_1 ocena od $P(2|1)$. Ove ocene uslovnih verovatnoća dobijene su po tzv. metodu matrice konfuzije. Korišćenjem gore navedenih ocena uslovnih verovatnoća formiramo ocenu stope greške E kao

$$\hat{E} = \frac{n_{12} + n_{21}}{n_1 + n_2} \tag{6.68}$$

Skrećemo pažnju na to da se za vrednovanje diskriminacionih funkcija kao klasifikacionog sredstva koriste iste opservacije iz uzorka koje su korišćene za ocenu diskriminacionih funkcija. Stoga se ocena od E dobijena na opisani način javlja pristrasnom naniže, tj. njome podcenjujemo stvarnu stopu greške. Zato se sugeriše, za dovoljno velike uzorke, podela uzorka na dva dela. Prvi deo se koristi za ocenu diskriminacione funkcije, a drugi za njeno vrednovanje. Stopa greške je ocenjena metodom matrice konfuzije na osnovu alokacije opservacija iz drugog dela uzorka. Na ovaj način prevazilazi se spomenuti problem pristrasnosti u oceni stope greške. Međutim, to je postignuto po cenu gubljenja dela informacije sadržane u drugom delu uzorka koji nije korišćen za ocenu diskriminacione funkcije.

Korišćenjem tzv. *jackknife* metoda[4] autori Lachenbruch i Mickey (1968) predlažu sledeći postupak ocene stope greške. U slučaju dve grupe na osnovu $n_1 - 1$ i n_2 elementa (izostavljajući jednu opservaciju iz prve grupe) ocenjujemo diskriminacionu funkciju, a ovu zatim koristimo u alokaciji opservacije koju smo izostavili. Postupak ponavljamo dok svaku od opservacija iz prve grupe ne izostavimo iz ocene diskriminacione funkcije, ali je potom klasifikujemo u jednu od grupa koristeći ocenjenu diskriminacionu funkciju. Ukupan broj opservacija iz prve grupe koje smo pogrešno klasifikovali u drugu grupu označimo sa $n_{21}^{(J)}$. Navedeni postupak ponovimo i za svaku opservaciju iz druge grupe. Naime, sukcesivno izostavljamo po jednu opservaciju iz druge grupe, a na osnovu preostalih n_1 i $n_2 - 1$ opservacija ocenjujemo diskriminacionu funkciju. U narednom koraku izostavljenu opservaciju klasifikujemo koristeći ocenjenu diskriminacionu funkciju. Broj opservacija iz druge grupe koje smo pogrešno klasifikovali u prvu grupu označavamo sa $n_{12}^{(J)}$. Sada je ocena od $P(1|2)$ data sa $n_{12}^{(J)}/n_2$, a ocena od $P(2|1)$ data sa $n_{21}^{(J)}/n_1$. U prilici smo da formiramo *jackknife* ocenu stope greške, u oznaci $\hat{E}^{(J)}$, gde je

$$\hat{E}^{(J)} = \frac{n_{12}^{(J)} + n_{21}^{(J)}}{n_1 + n_2} \tag{6.69}$$

---

[4] Videti: Efron (1982). *Jackknife* metod koristi se za ocenu pristrasnosti, varijanse i opštijih mera grešaka koje se čine pri ocenjivanju. Najjednostavniji primer korišćenja *jackknife* metoda je u oceni varijanse sredine slučajnog uzorka od *n* opservacija. Sukcesivno izostavljajući po jednu opservaciju uzorka od preostalih elemenata formiramo ocenu sredine uzorka, dobijajući tako niz ocena sredine. Varijansa ovog niza predstavlja tzv. *jackknife* ocenu varijanse sredine uzorka. Ono što je značajnije od primene *jackknife* metoda u navedenom primeru jeste to da se za ma koju ocenu na osnovu slučajnog uzorka, prema navedenom postupku, može formirati *jackknife* ocena varijanse te ocene.

Primer 6.11 Korišćenjem Fisherove linearne diskriminacione funkcije klasifikovati opservacije uzorka u jednu od grupa na osnovu datih matrica podataka $(n_1 = n_2 = 3)$. Formirati matricu konfuzije i oceniti stopu greške. Primenom *jackknife* postupka oceniti stopu greške. Pretpostaviti da su apriorne verovatnoće jednake kao i da su troškovi pogrešne klasifikacije jednaki međusobom.

Za prvu populaciju matrica podataka sa uzoračkim pokazateljima je

$$\mathbf{X}_1 = \begin{bmatrix} 2 & 3 & 4 \\ 4 & 8 & 6 \end{bmatrix},\ \bar{\mathbf{x}}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix},\ \mathbf{S}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

Za drugu populaciju matrica podataka sa uzoračkim pokazateljima je

$$\mathbf{X}_2 = \begin{bmatrix} 3 & 4 & 5 \\ 4 & 2 & 6 \end{bmatrix},\ \bar{\mathbf{x}}_2 = \begin{bmatrix} 4 \\ 4 \end{bmatrix},\ \mathbf{S}_2 = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}.$$

Opšta kovarijaciona matrica je

$$\bar{\mathbf{S}} = \frac{1}{4}(2\mathbf{S}_1 + 2\mathbf{S}_2) = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

Fisherova linearna diskriminaciona funkcija je

$$y = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\bar{\mathbf{S}}^{-1}\mathbf{x} = \begin{bmatrix} -1 & 2 \end{bmatrix}\frac{1}{3}\begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -2x_1 + x_2$$

Centroidi su $\bar{y}_1 = 0$ i $\bar{y}_2 = -4$, $\hat{\mu}_Y = -2$. Proveriti da se alokacijom elemenata uzorka dobija sledeća matrica konfuzije

		Stvarna pripadnost	
		π_1	π_2
Predviđena pripadnost	π_1	$n_{11} = 2$	$n_{12} = 0$
	π_2	$n_{21} = 1$	$n_{22} = 3$
Veličina uzorka		$n_1 = 3$	$n_2 = 3$

odakle je ocena stope greške $\hat{E} = \dfrac{n_{12} + n_{21}}{n_1 + n_2} = \dfrac{1+0}{3+3} = \dfrac{1}{6} = 0.1667$.

Postupak primene *jackknife* metoda započinjemo izostavljanjem prve opservacije $\mathbf{x}_I' = \begin{bmatrix} 2 & 4 \end{bmatrix}$ iz prve grupe, pa su matrica opservacija i uzoračke pokazatelji

$$\mathbf{X}_{1I} = \begin{bmatrix} 3 & 4 \\ 8 & 6 \end{bmatrix},\ \bar{\mathbf{x}}_{1I} = \begin{bmatrix} 3.5 \\ 7 \end{bmatrix},\ \mathbf{S}_{1I} = \begin{bmatrix} 0.5 & -1 \\ -1 & 2 \end{bmatrix}$$

Nova opšta kovarijaciona matrica i njena inverzna matrica su

$$\bar{\mathbf{S}}_I = \frac{1}{3}(1\mathbf{S}_{1I} + 2\mathbf{S}_2) = \frac{1}{3}\begin{bmatrix} 2.5 & 1 \\ 1 & 10 \end{bmatrix},\ \bar{\mathbf{S}}_I^{-1} = \frac{1}{8}\begin{bmatrix} 10 & -1 \\ -1 & 2.5 \end{bmatrix}$$

Fisherova linearna diskriminaciona funkcija je sada

$$y_I = (\bar{\mathbf{x}}_{1I} - \bar{\mathbf{x}}_2)'\bar{\mathbf{S}}_I\mathbf{x} = \begin{bmatrix} -0.5 & 2 \end{bmatrix}\frac{1}{8}\begin{bmatrix} 10 & -1 \\ -1 & 2.5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -x_1 + x_2$$

Centroidi su $\bar{y}_{1I} = 3.5$ i $\bar{y}_2 = 0$, a središnja tačka između njih je 1.75. To znači da izostavljenu opservaciju treba klasifikovati u populaciju π_1 (ispravna odluka). U narednoj etapi izostavljamo drugu opservaciju iz prve grupe, tj. $\mathbf{x}_I' = \begin{bmatrix} 3 & 8 \end{bmatrix}$ i ponavljamo izloženi postupak. Potrebni međurezultati su

$$\mathbf{X}_{1I} = \begin{bmatrix} 2 & 4 \\ 4 & 6 \end{bmatrix},\ \bar{\mathbf{x}}_{1I} = \begin{bmatrix} 3 \\ 5 \end{bmatrix},\ \mathbf{S}_{1I} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

$$\bar{\mathbf{S}}_I = \frac{1}{3}\begin{bmatrix} 4 & 4 \\ 4 & 10 \end{bmatrix},\ \bar{\mathbf{S}}_I^{-1} = \frac{1}{8}\begin{bmatrix} 10 & -4 \\ -4 & 4 \end{bmatrix}$$

Fisherova linearna diskriminaciona funkcija sada glasi $y_I = -1.75x_1 + x_2$, centroidi su $\bar{y}_{1I} = -0.25$ i $\bar{y}_2 = -3$, a središnja tačka između njih je 1.625. Znači da izostavljenu opservaciju treba klasifikovati u populaciju π_1 (ispravna odluka). Nastavljajući prema izloženom postupku poslednju opservaciju iz prve grupe alocirali bi pogrešno u populaciju π_2.

Pokazati da bi se sukcesivno izostavljene opservacije iz druge grupe redom klasifikovale u populaciju: π_1 (pogrešna odluka), π_2 (ispravna odluka) i π_2 (ispravna odluka). To znači da je $n_{12}^{(J)} = 1$ i $n_{21}^{(J)} = 1$, pa je *jackknife* ocena stope greške

$$\hat{E}^{(J)} = \frac{n_{12}^{(J)} + n_{21}^{(J)}}{n_1 + n_2} = \frac{1+1}{3+3} = \frac{2}{6} = 0.3333$$

Kao što je rečeno, ocena $\hat{E}$ je pristrasna, jer podcenjuje pravu stopu greške. Ocena $\hat{E}^{(J)}$ dobijena *jackknife* metodom je pouzdanija. ■

Kako se u primeru dve populacije sa jednakim brojem opservacija na slučajan način opservacije mogu alocirati u jednu od grupa sa podjednakom verovatnoćom, to znači da je kod klasifikacije na slučaj, stopa greške 50%. Ukoliko je korišćenjem jednog od postupaka klasifikacije dobijena stopa greške blizu 50%, konstatujemo da je korišćeni postupak klasifikacije opservacija slabo sredstvo za njihovu pouzdanu alokaciju po grupama.

O drugim metodama ocene stope greške kao što su metod najveće verodostojnosti, metod asimptotskog razvoja i kombinacija *jackknife* metoda sa normalnim rasporedom videti kod Karsona (1982, ss. 180-4).

6.4 PROBLEMI U PRIMENI DISKRIMINACIONE ANALIZE

U dosadašnjem izlaganju modela diskriminacione analize polazili smo od toga da je vektor opservacija $\mathbf{X}$ sa svojih p komponenti dat. Pri tome nismo razmatrali kakva je priroda tih promenljivih sa stanovišta njihove podele na kvantitativne i kvalitativne. Takođe, sama dimenzija vektora $\mathbf{X}$ nije se dovodila u pitanje. Drugim rečima, sve raspoložive promenljive uključili smo u vektor $\mathbf{X}$ i razdvajanje grupa i. alokaciju opservacija vršili na osnovu tako komponovanog vektora. U ovom poglavlju bavićemo se upravo problemom značajnosti izvesnog podskupa promenljivih iz vektora $\mathbf{X}$, te problemima i posledicama korišćenja kvalitativnih promenljivih u diskriminacionoj analizi.

U izlaganju modela diskriminacione analize za slučajan vektor $\mathbf{X}$ vezivali smo izvesne pretpostavke koje su nam omogućile dolaženje do relativno jednostavnih pravila alokacije opservacija ili procedure statističkog testiranja značajnosti diskriminacionih funkcija. Pre svega, pretpostavljali smo da slučajan vektor $\mathbf{X}$ ima višedimenzioni normalan raspored. Ukazaćemo na posledice napuštanja ove pretpostavke, tj. na alternativne modele diskriminacione analize.

6.4.1 Izbor promenljivih

U odeljku o zaključivanju kod kanoničke diskriminacione funkcije razmatran je postupak statističkog testiranja značajnosti diskriminacione funkcije u celini. Navedenim postupkom u prilici smo da utvrdimo koliko diskriminacionih funkcija treba zadržati u daljoj analizi. Pored utvrđivanja broja statistički značajnih diskriminacionih funkcija interes nam je da utvrdimo i broj promenljivih koje ćemo koristiti u analizi. Naime, postavlja se pitanje da li su sve promenljive potrebne za dobro razdvajanje grupa ili klasifikaciju opservacija. Zato se javlja potreba definisanja statističkog kriterijuma provere diskriminacione moći pojedinačnih promenljivih ili grupe promenljivih.

Govoreći o interpretaciji kanoničkih diskriminacionih funkcija rekli smo da na osnovu veličine apsolutne vrednosti standardizovanih diskriminacionih koeficijenata možemo suditi o visini doprinosa odnosne promenljive razdvajanju grupa. Promenljivu čiji je koeficijent blizak nuli mogli smo izostaviti iz diskriminacione funkcije. Za

razliku od ovog subjektivnog pristupa problemu izbora promenljivih izložićemo formalan statistički postupak utvrđivanja značajnosti doprinosa promenljive diskriminaciji između grupa. Podsećamo da je u Odeljku 6.2.1 kod razdvajanja dve grupe na osnovu Fisherovog pristupa dobijen rezultat da je maksimalna vrednost Fisherovog diskriminacionog kriterijuma jednaka kvadratu Mahalanobisovog odstojanja između dve grupe. Ako je izostavljanjem nekih promenljivih novodobijena vrednost Mahalanobisovog odstojanja neznatno različita od predašnje, tada ima osnova zaključku da izostavljene promenljive ne doprinose razdvajanju grupa. Statistički postupak zasnovan je upravo na utvrđivanju statističke značajnosti razlike ova dva Mahalanobisova odstojanja.

Razmotrimo problem razdvajanja dve normalno raspoređene populacije sa sredinama $\boldsymbol{\mu}_1$ i $\boldsymbol{\mu}_2$ i zajedničkom kovarijacionom matricom $\boldsymbol{\Sigma}$. Dobijen je rezultat da su populacioni diskriminacioni koeficijenti u diskriminacionoj funkciji $\mathbf{a}'\mathbf{X}$ dati izrazom

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}, \text{ gde je } \boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \tag{6.70}$$

Izvršimo podelu vektora $\mathbf{X}$ tako da prvi podskup sadrži q, a drugi podskup $s = p - q$ promenljivih, tj. $\mathbf{X}' = \left[\mathbf{X}_1' \vdots \mathbf{X}_2'\right]$, gde je $\mathbf{X}_1$ $(q \times 1)$ vektor, a $\mathbf{X}_2$ $(s \times 1)$ vektor. U skladu sa podelom vektora $\mathbf{X}$ izvršićemo podelu vektora $\mathbf{a}$ i $\boldsymbol{\delta}$, pa prema tome imamo da je $\mathbf{a}' = [\mathbf{a}_1' \vdots \mathbf{a}_2']$ i $\boldsymbol{\delta}' = [\boldsymbol{\delta}_1' \vdots \boldsymbol{\delta}_2']$, gde su dimenzije vektora $\mathbf{a}_1$ i $\boldsymbol{\delta}_1$, kao i $\mathbf{a}_2$ i $\boldsymbol{\delta}_2$ jednake dimenzijama vektora $\mathbf{X}_1$ i $\mathbf{X}_2$ respektivno. Takođe vršimo uobičajenu podelu kovarijacione matrice $\boldsymbol{\Sigma}$. Ako promenljive $X_{q+1}, ..., X_p$ ne doprinose razdvajanju grupa, odnosno njihova diskriminaciona moć je mala, tada će diskriminacioni koeficijenti uz te promenljive biti jednaki nuli ($\mathbf{a}_2 = \mathbf{0}$).

Kao što smo napomenuli, maksimalna vrednost Fisherovog diskriminacionog kriterijuma je kvadrat Mahalanobisovog odstojanja dve populacije. Označimo ga sa Δ_p^2 ako je računat na osnovu svih p promenljivih, a sa Δ_q^2 ako je računat na osnovu prvih q promenljivih. Sada se hipoteza da je $\mathbf{a}_2 = \mathbf{0}$ može izraziti na ekvivalentan način kao $\Delta_p^2 = \Delta_q^2$, gde je

$$\Delta_p^2 = \boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \text{ i } \Delta_q^2 = \boldsymbol{\delta}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1 \tag{6.71}$$

Znači da ako su u statističkom smislu ova dva odstojanja međusobom jednaka, tada promenljive $X_{q+1}, ..., X_p$ ne doprinose razdvajanju grupa, pa ih ne treba uključiti u analizu.

Statistika testa za test hipoteze $H_0 : \mathbf{a}_2 = \mathbf{0}$ zasnovana je na uzoračkim Mahalanobisovim odstojanjima

$$D_p^2 = \hat{\boldsymbol{\delta}}'\bar{\mathbf{S}}^{-1}\hat{\boldsymbol{\delta}} \text{ i } D_q^2 = \hat{\boldsymbol{\delta}}_1'\bar{\mathbf{S}}_{11}^{-1}\hat{\boldsymbol{\delta}}_1 \tag{6.72}$$

i glasi

$$\frac{n_1+n_2-p-1}{p-q}\frac{D_p^2-D_q^2}{c+D_q^2} \tag{6.73}$$

gde je $c=(n_1+n_2)(n_1+n_2-2)/(n_1n_2)$. Ako je H_0 tačna, statistika testa ima $F-$raspored sa $(p-q)$ i (n_1+n_2-p-1) stepeni slobode. Specijalan slučaj od posebnog interesa jeste za $q=p-1$, kada testiramo značajnost samo jedne promenljive u razdvajanju grupa. Ova statistika se može, slično kao u regresionoj analizi, koristiti u tzv. stepenastoj diskriminacionoj analizi za uključivanje i isključivanje promenljivih iz diskriminacione funkcije. Stepenasta diskriminaciona analiza povećava broj uključenih promenljivih u diskriminacionoj funkciji uzimajući sa spiska razpoloživih promenljivih onu koja u najvećoj meri doprinosi razdvajanju grupa, a sa druge strane isključuje onu varijablu čiji je doprinos razdvajanju obuhvaćen nekom kombinacijom ranije uključenih promenljivih.

Primer 6.12 Na osnovu podataka iz Primera 6.1 testirati hipotezu o statistički zanemarljivoj diskriminacionoj moći prve i druge promenljive.

Kako je u našem primeru $p=2$, to je $\mathbf{a}'=[a_1\ a_2]$, tj. vektori $\mathbf{a}_1$ i $\mathbf{a}_2$ svode se na skalare a_1 i a_2. Uzoračka vrednost kvadrata Mahalanobisovog odstojanja računata na osnovu obe promenljive glasi

$$D_p^2=\hat{\boldsymbol{\delta}}'\bar{\mathbf{S}}^{-1}\hat{\boldsymbol{\delta}}=\begin{bmatrix}3.5-6.5 & 4.5-4.5\end{bmatrix}\begin{bmatrix}0.8 & -0.4\\ -0.4 & 0.8\end{bmatrix}\begin{bmatrix}3.5-6.5\\ 4.5-4.5\end{bmatrix}=7.2,$$

a na osnovu samo prve promenljive, odnosno kada želimo testirati hipotezu o diskriminacionoj moći druge promenljive, $H_0:\mathbf{a}_2=\mathbf{0}$,

$$D_q^2=\hat{\boldsymbol{\delta}}_1'\bar{\mathbf{S}}_{11}^{-1}\hat{\boldsymbol{\delta}}_1=(3.5-6.5)^2(0.6)=5.4.$$

Kako je $c=(n_1+n_2)(n_1+n_2-2)/(n_1n_2)=(3+3)(3+3-2)/(3\cdot 3)=2.67$, to je vrednost statistike testa

$$\frac{n_1+n_2-p-1}{p-q}\frac{D_p^2-D_q^2}{c+D_q^2}=\frac{3+3-2-1}{2-1}\frac{7.2-5.4}{2.67+5.4}=0.67.$$

Kritična vrednost statistike testa na nivou značajnosti $\alpha=0.05$ iznosi $F_{1,3;0.05}=10.1$. Pošto je izračunata vrednost manja od kritične, prihvatamo hipotezu da druga promenljiva ne doprinosi statistički značajno razdvajanju grupa.

Na sličan način proveravamo diskriminacionu moć prve promenljive, tj. testiramo hipotezu $H_0 : \mathbf{a}_1 = \mathbf{0}$. Odgovarajuća uzoračka vrednost kvadrata Mahalanobisovog odstojanja iznosi

$$D_q^2 = \hat{\boldsymbol{\delta}}_1' \overline{\mathbf{S}}_{11}^{-1} \hat{\boldsymbol{\delta}}_1 = (4.5 - 4.5)^2 (0.6) = 0 .$$

Vrednost statistike testa sada iznosi 8.09. Ona nam sugeriše da ni prva promenljiva na nivou značajnosti od 5% ne doprinosi značajno razdvajanju grupa, ali da se pokazuje značajna na nivou značajnosti od 10%. Ovo je rezultat ekstremno malog uzorka od po tri opservacije po grupi, pa ovaj primer treba posmatrati samo kao ilustraciju postupka testiranja. ■

Opisani postupak testiranja opravdanosti uvođenja dodatne promenljive, pored ranije uključenih, može se proširiti na slučaj više populacija $(g > 2)$ korišćenjem Wilksove lambde. Ako sa Λ_j označimo vrednost Wilksove lambde u MANOVA baziranoj na prvih j promenljivih, statistika testa za vrednost $j+1$ promenljive iznosi

$$\frac{n-g-j}{g-1}\left(\frac{\Lambda_j}{\Lambda_{j+1}} - 1\right), \tag{6.74}$$

koja u velikim uzorcima ima F – raspored sa $(g-1)$ i $(n-g-j)$ stepeni slobode ako $j+1$ promenljiva ne doprinosi statistički značajno diskriminaciji između grupa. Nedostatak ove F statistike ogleda se u tome što pokazuje tendenciju povećanja razdvajanja dobro razdvojenih grupa, umesto da poboljšava razdvajanje slabije razdvojenih grupa. Ukoliko p nije veliko, predlaže se poređenje rezultata razdvajanja dobijenih preko funkcija koje su računate na osnovu svih podskupova promenljivih. Detaljnije o izboru promenljivih u diskriminacionoj analizi videti kod Krishnaiah (1982).

6.4.2 Uključivanje kvalitativnih promenljivih

U uvodnoj glavi, dajući klasifikaciju metoda multivarijacione analize, diskriminacionu analizu svrstali smo među metode zavisnosti, pri čemu je zavisna promenljiva kvalitativnog tipa. Na ovom mestu interesuje nas mogućnost primene diskriminacione analize kada se u skup nezavisnih promenljivih uključi kvalitativna promenljiva ili u ekstremnom slučaju kada su sve nezavisne promenljive kvalitativnog tipa.

U načelu nema prepreka da se izloženi pristup diskriminaciji između grupa primeni i u slučaju kada je među nezavisnim promenljivama jedna ili više kvalitativnih promenljivih. Međutim, u takvom slučaju ne možemo očekivati da će primenjen metod imati navedena optimalna svojstva, štaviše, ne možemo se nadati ni dobrim rezultatima razdvajanja. U mešovitom slučaju (nezavisne promenljive su kvantitativnog i kvalitativnog tipa) sugeriše se umesto primene linearne diskriminacione analize

korišćenje tzv. logističke diskriminacione analize u čijoj osnovi je logit model. Opširnije o primeni logističke diskriminacione analize videti: Dillon i Goldstein (1984, ss. 386-392).

Ukoliko su svih p promenljivih slučajnog vektora $\mathbf{X}$ prekidnog tipa, odnosno svaka od promenljivih $X_1, X_2, ..., X_p$ uzima konačan broj različitih vrednosti $m_1, m_2, ..., m_p$, tada je skup svih mogućih vrednosti dat proizvodom $\prod_{j=1}^{p} m_j$. Za slučaj dve grupe, sa p_1 i p_2 označimo apriornu verovatnoću da će na slučaj izvučen element iz združeno posmatranih grupa pripadati prvoj, odnosno drugoj grupi respektivno. Sa $f_1(\mathbf{x})$ i $f_2(\mathbf{x})$ označimo verovatnoće da slučajan vektor $\mathbf{X}$ uzme vrednost $\mathbf{x}$ u slučaju prve i druge grupe respektivno. Novu opservaciju $\mathbf{x}_0$ alociramo u grupu π_1 ako je

$$p_1 f_1(\mathbf{x}_0) > p_2 f_2(\mathbf{x}_0) \tag{6.75}$$

U suprotnom, opservaciju alociramo u grupu π_2. Ukoliko je $p_1 f_1(\mathbf{x}_0) = p_2 f_2(\mathbf{x}_0)$ opservaciju $\mathbf{x}_0$ alociramo na slučaj u grupu π_1 ili π_2. Proizvod $p_k f_k(\mathbf{x})$, $k = 1, 2$, nazivamo diskriminacioni skor. Možemo zaključiti da pri alokaciji opservacija u diskretnoj diskriminacionoj analizi postupamo slično kao u ranije izloženom postupku u slučaju dve grupe.

Neka je iz združenih grupa uzet uzorak od n opservacija ($n = n_1 + n_2$). Označimo sa $n_k(\mathbf{x})$ broj opservacija uzetih iz grupe π_k kod kojih je $\mathbf{X} = \mathbf{x}$. Ocene apriornih verovatnoća p_k i uslovnih verovatnoća da će slučajan vektor $\mathbf{X}$ uzeti vrednost $\mathbf{x}$ pod uslovom da opservacija potiče iz te grupe su

$$p_k = \frac{n_k}{n}, \text{ odn. } \hat{f}_k(\mathbf{x}) = \frac{n_k(\mathbf{x})}{n_k}, \; k = 1, 2 \tag{6.76}$$

Znači da je diskriminacioni skor jednak $n_k(\mathbf{x})/n$. Uzoračko pravilo alokacije glasi: alociraj $\mathbf{x}_0$ u π_1 ako je

$$p_1 \hat{f}_1(\mathbf{x}_0) > p_2 \hat{f}_2(\mathbf{x}_0) \tag{6.77}$$

u suprotnom alociraj u π_2. Ako je $p_1 \hat{f}_1(\mathbf{x}_0) = p_2 \hat{f}_2(\mathbf{x}_0)$ novu opservaciju možemo alocirati na slučaj u grupu π_1 ili π_2.

Primer 6.13 U ispitivanju potrošnje osvežavajućeg pića sprovedena je anketa među potrošačima. Dobijen je odgovor 2700 ispitanika koji su prema visini potrošnje svrstani u dve grupe: 1700 ispitanika visoke potrošnje i 1000 ispitanika niske potrošnje. Ispitanici su odgovarali sa "slažem se - ne slažem se" na postavljena pitanja, pa su odgovori kodirani sa 1 i 0. Dve promenljive su:

X_1: Hranljiva vrednost je bitna karakteristika osvežavajućeg pića.

X_2: Vrsta pakovanja utiče na visinu potrošnje osvežavajućeg pića.

Kako svaka promenljiva ima dva modaliteta $(m_1 = m_2 = 2)$ to je ukupan broj različitih kombinacija odgovora ispitanika jednak 4. Distribucija frekvencija ove četiri kombinacije za svaku grupu ispitanika data je u narednoj tabeli.

(x_1, x_2)	1. grupa: Visoka potrošnja			2. grupa: Niska potrošnja		
	Frekvencija	Verovatnoća	Skor	Frekvencija	Verovatnoća	Skor
1,1	450	0.27	0.17	150	0.15	0.06
1,0	910	0.54	0.34	350	0.35	0.13
0,1	300	0.18	0.11	460	0.46	0.17
0,0	40	0.02	0.02	40	0.04	0.02
Zbir	1700	1.00	-	1000	1.00	-

Na primer, ispitanika koji su potvrdno odgovorili na oba pitanja ($\mathbf{x} = (1,1)$) ima u prvoj grupi 450, a u drugoj 150. Uslovne verovatnoće su

$$\hat{f}_1(1,1) = \frac{n_1(1,1)}{n_1} = \frac{450}{1700} = 0.2$$

$$\hat{f}_2(1,1) = \frac{n_2(1,1)}{n_2} = \frac{150}{1000} = 0.1$$

a diskriminacioni skorovi su

$$p_1 \hat{f}_1(1,1) = \frac{n_1(1,1)}{n} = 0.17$$

$$p_2 \hat{f}_2(1,1) = \frac{n_2(1,1)}{n} = 0.06$$

Kako je $p_1 \hat{f}_1(1,1) > p_2 \hat{f}_2(1,1)$, zaključujemo da svakog novog ispitanika koji potvrdno odgovori na oba pitanja treba svrstati u prvu grupu potrošača sa visokom potrošnjom osvežavajućeg pića.

Da bismo proverili uspešnost klasifikacije ispitanika na osnovu njihovih diskriminacionih skorova posmatraćemo vrednosti skorova kod ove dve grupe potrošača. Već smo konstatovali da ispitanika sa odgovorom $\mathbf{x} = (1,1)$ treba svrstati u prvu grupu. To znači da smo 150 ispitanika iz druge grupe pogrešno klasifikovali. Slično konstatujemo za odgovore $\mathbf{x} = (1,0)$, odnosno 350 ispitanika smo pogrešno alocirali u drugu grupu. Ispitanike koji daju odgovore $\mathbf{x} = (0,1)$ svrstavamo u drugu grupu, što znači da smo 300 ispitanika pogrešno svrstali u prvu grupu. Sumaran

rezultat ukazuje da smo 17.65% ispitanika pogrešno svrstali u prvu grupu (visoke potrošnje), a 50% ispitanika pogrešno svrstali u drugu grupu (niske potrošnje). Ispitanici čiji su odgovori $\mathbf{x} = (0, 0)$ imaju jednake diskriminacione skorove u ove dve grupe, pa je sasvim svejedno u koju ćemo ih grupu svrstati. Zato se u alokaciji na slučaj ovih ispitanika mogu javiti 40 dodatnih ispitanika koji će biti pogrešno klasifikovani. ■

Izloženo pravilo alokacije u diskretnoj diskriminacionoj analizi relativno uspešno funkcioniše u velikim uzorcima. Problemi se javljaju kod manjih uzoraka, tj. u slučaju da neka od kombinacija vrednosti prekidnih promenljivih sadrži mali broj opservacija (niska frekvencija date kombinacije).

6.4.3 Anormalnost podataka

Izlaganje diskriminacione analize bazirali smo na pretpostavci da je slučajan vektor $\mathbf{X}$ raspoređen po višedimenzionom normalnom rasporedu. Ako je ova pretpostavka ispunjena tada Fisherova linearna diskriminaciona funkcija predstavlja optimalno sredstvo za diskriminaciju po kriterijumu minimalnih očekivanih troškova pogrešne klasifikacije. Međutim, ukoliko su sve nezavisne promenljive kvalitativne, kao u Primeru 6.13, ili imamo mešovit slučaj da su neke od nezavisnih promenljivih kvalitativne, tada nije razumno pretpostaviti normalnost rasporeda slučajnog vektora $\mathbf{X}$. Korišćenje Fisherove diskriminacione funkcije uprkos tome može dati relativno slabe rezultate klasifikacije uz pristrasne ocene greške klasifikacije.

U slučajevima kada se ne može pretpostaviti normalnost rasporeda preporučuje se korišćenje logističke diskriminacione analize i probit modela. Takođe se predlaže korišćenje neparametarskih metoda diskriminacione analize zasnovanih na rangovima. Detaljnije o metodama diskriminacione analize primerenim slučajevima kada pretpostavka o normalnosti nije ispunjena videti kod Lachenbruch (1975), a o posledicama primene diskriminacione analize u slučaju anormalnosti podataka videti kod Lachenbruch, Sneeringer i Revo (1973).

7 GLAVNE KOMPONENTE

Metod multivarijacione analize koji se koristi za smanjivanje dimenzije skupa podataka (sačinjava ga veliki broj uzajamno korelisanih promenljivih) uz istovremeno zadržavanje maksimalno mogućeg varijabiliteta koji je prisutan u tim podacima, naziva se *metod glavnih komponenata*. Pretpostavimo da ispitujemo stepen socioekonomske razvijenosti regiona neke zemlje ili grupe zemalja. Stepen dostignutog socioekonomskog razvoja je višedimenziona pojava. Njega možemo meriti korišćenjem različitih indikatora koji ističu specifičnu komponentu dostignutog stepena razvijenosti: ekonomski indikatori (društveni proizvod po glavi stanovnika, obim osnovnih sredstava, broj radnika u tercijarnim delatnostima, stopa nezaposlenosti i dr.), socijalni indikatori (stopa nepismenosti, broj telefonskih pretplatnika, broj putničkih automobila i dr.), kulturni indikatori (broj pozorišta, broj TV i radio pretplatnika i dr.) i zdravstveni indikatori (broj stanovnika na jednog lekara, broj bolničkih kreveta, i dr.). Uključivanje veliki broja navedenih indikatora u analizu nameće probleme praktične prirode sagledavanja njihove povezanosti. U slučaju na primer, 20 promenljivih međusobna povezanost indikatora iskazana je sa 190 koeficijenata korelacije.

U našem primeru metod glavnih komponenata pojednostavljuje analizu stepena razvijenosti u tom smislu što veliki broj pokazatelja socioekonomskog razvoja svodi na manji broj koji kao takvi, reprezentuju navedene grupe indikatora. Pri tome oni zadržavaju skoro svu informaciju sadržanu u prvobitnom skupu podataka. Time smo, s jedne strane, redukovali broj promenljivih u analizi, a sa druge strane učinili korak ka razumevanju strukture izučavanog fenomena, odnosno relativnog značaja faktora koji su uslovili dostignuti stepen razvijenosti. Stoga kažemo da metod glavnih komponenata pored toga što redukuje dimenziju skupa podataka predstavlja i istraživačko sredstvo analize pomoću koga se generišu hipoteze o proučavanom fenomenu. Hipotetsku strukturu "otkrivenu" metodom glavnih komponenata izučavamo zatim drugim metodama multivarijacione analize kao što su metod multivarijacione regresije i kanoničke korelacione analize.

Nakon uvodnog poglavlja u kome iznosimo osnovne zadatke i ciljeve metode glavnih komponenata kao i bazičnu ideju njihovog konstruisanja, u okviru drugog poglavlja dajemo formalnu definiciju i glavne osobine populacionih glavnih komponenata kao i način njihove interpretacije. Uzoračkim glavnim komponentama, ocenom i testiranjem njihove značajnosti bavimo se u trećem poglavlju. Naposletku u poslednjem poglavlju izlažemo neke aspekte primene glavnih komponenata. Tako razmatramo njihovo korišćenje u drugim metodama multivarijacione analize, zatim u identifikaciji nestandardnih opservacija kao i robustnu analizu glavnih komponenata u slučaju prisustva nestandardnih opservacija.

7.1 UVOD

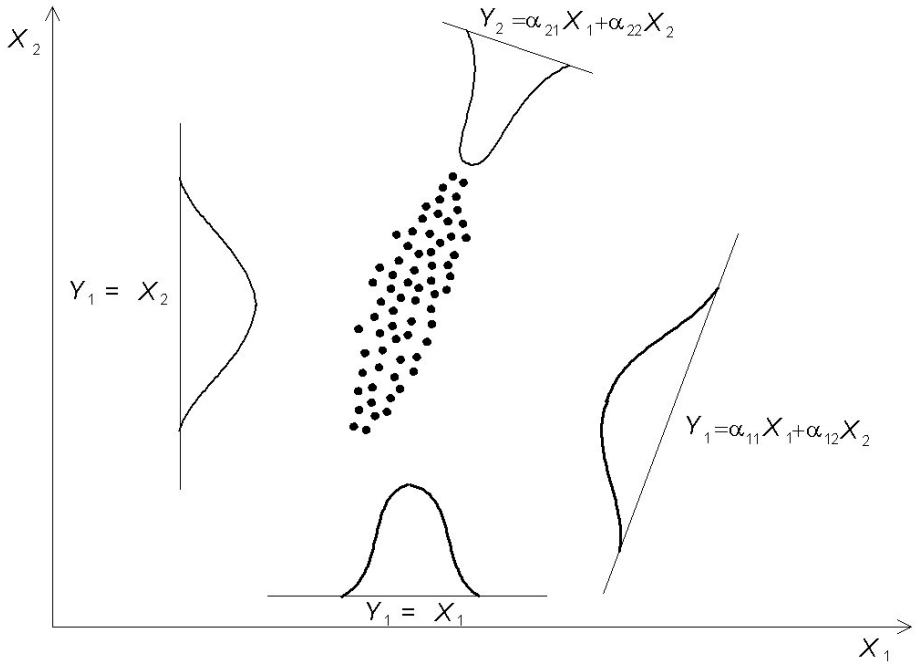
Osnovni zadatak metode glavnih komponenata jeste određivanje linearne kombinacije originalnih promenljivih koja će imati maksimalnu varijansu. U primeru ispitivanja stepena razvijenosti regiona velika vrednost varijanse linearne kombinacije implicira dobro razdvajanje regiona olakšavajući na taj način utvrđivanje razlika između njih. Sada se regioni mogu rangirati prema vrednostima te linearne kombinacije koju bismo mogli nazvati opštim indeksom razvijenosti.

Drugi, opštiji zadatak metode glavnih komponenata jeste određivanje *nekoliko* linearnih kombinacija originalnih promenljivih koje će, pored toga što imaju maksimalnu varijansu, biti međusobom nekorelisane gubeći u što je moguće manjoj meri informaciju sadržanu u skupu originalnih promenljivih. U postupku primene ove metode originalne promenljive transformišu se u nove promenljive (linearne kombinacije) koje nazivamo *glavne komponente*. Prva glavna komponenta konstruisana je tako da "obuhvata" najveći deo varijanse originalnog skupa podataka, a naredne onaj deo varijanse originalnog skupa podataka koji nije obuhvaćen prethodno izdvojenim glavnim komponentama. Analiza je bazirana na pretpostavci da će nekoliko glavnih komponenata (znatno manji broj od broja originalnih promenljivih) biti dobra aproksimacija kovarijacione strukture skupa originalnih promenljivih.

Izvršavajući ove zadatke metoda glavnih komponenata postiže dva cilja: 1) vrši redukciju originalnog skupa podataka i 2) olakšava njegovu interpretaciju. Prvi cilj bismo opisali kao "štedljivo sumiranje" podataka. Naime, ako je moguće važne osobine višedimenzionog skupa podataka sumirati sa manjim brojem linearnih kombinacija, tada ćemo na primer, umesto 20 promenljivih u daljoj analizi koristiti recimo, 4 linearne kombinacije. Istovremeno, ovom redukcijom podataka stižemo do drugog cilja, odnosno olakšavamo interpretaciju kovarijacione strukture originalnog skupa promenljivih na bazi manjeg broja međusobom nekorelisanih glavnih komponenata.

Poslužimo se jednostavnim primerom u kome na dvodimenzionom skupu promenljivih ilustrujemo osnovnu ideju metode i način konstruisanja glavnih komponenata. Na Slici 7.1 prikazan je dijagram rasturanja promenljivih X_1 i X_2. Linearna kombinacija ove dve promenljive glasi: $Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2$, gde smo sa α_{1j} označili koeficijent linearne kombinacije uz $j-$tu promenljivu u prvoj linearnoj

kombinaciji. Izborom koeficijenata $\alpha_{11}=1$ i $\alpha_{12}=0$, odn. $\alpha_{11}=0$ i $\alpha_{12}=1$ dobićemo da je prva linearna kombinacija jednaka prvoj, odnosno drugoj promenljivoj respektivno. U geometrijskom smislu ovakvim izborom koeficijenata linearne kombinacije dobijamo promenljivu Y_1 čija distribucija frekvencije je formirana na osnovu projekcije roja tačaka na X_1 i X_2. osu (na Slici 7.1 ove distribucije frekvencija su jednostavnosti radi prikazane u obliku zvona). Distribucija frekvencije Y_1 u prvom slučaju predstavlja marginalni raspored od X_1, a u drugom slučaju marginalni raspored od X_2. Ako se zahteva reprezentovanje dvodimenzionog skupa samo jednom promenljivom onda bismo izabrali onu koja ima veći varijabilitet. Ovakav izbor se objašnjava time da na osnovu promenljive sa većim varijabilitetom možemo u većoj meri razlikovati pojedinačne opservacije dvodimenzionog skupa. U ekstremnom slučaju kada sve tačke roja leže na pravoj upravnoj na X_1 osu (vrednost promenljive X_1 je jednaka kod svake jedinice posmatranja) tada je dovoljno u analizi zadržati samo promenljivu X_2 jer ona nosi svu informaciju o varijabilitetu dvodimenzionog skupa podataka.



Slika 7.1 Projekcije roja tačaka

Sa Slike 7.1 vidimo da je promenljiva X_2 kandidat za reprezenta dvodimenzionog skupa podataka jer ima veću varijansu od promenljive X_1. Postavlja se pitanje, da li postoji takav izbor koeficijenata linearne kombinacije koji će rezultirati u većoj varijansi promenljive Y_1 nego što je to u slučaju prethodno navedenog izbora koeficijenata. Formalno, naš izbor koeficijenata može se opisati kao zadatak maksimiziranja varijanse linearne kombinacije uz uslov da je zbir kvadrata koeficijenata linearne kombinacije jednak jedinici. Geometrijski ovaj uslov znači da je vektor koeficijenata linearne kombinacije $[\alpha_{11}, \alpha_{12}]'$ jedinične dužine. Normirajući uslov uvodi se u cilju postizanja

jednoznačne definisanosti linearne kombinacije. U geometrijskom smislu izborom koeficijenata menjamo ugao pod kojim projektujemo tačke iz roja tačaka na pravu liniju. Izabraćemo one koeficijente koji će dati projekciju tačaka sa najvećom varijansom. Na Slici 7.1 to je projekcija označena sa $Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2$, odnosno ta linearna kombinacija ima najveću varijansu od svih linearnih kombinacija koje se mogu dobiti promenom ugla projekcije. Ovu linearnu kombinaciju, koja predstavlja projekciju roja tačaka na pravu liniju sa najvećim varijansom, nazivamo *prva glavna komponenta*.

Ako je za potrebe analize dovoljno izdvojiti jednu linearnu kombinaciju na ovom mestu se zaustavljamo, u suprotnom formiramo narednu linearnu kombinaciju $Y_2 = \alpha_{21}X_1 + \alpha_{22}X_2$. Njene koeficijente određujemo tako da joj se maksimizira varijansa uz normirajući uslov kao i kod prve glavne komponente i dodatni uslov da su Y_1 i Y_2 međusobom nekorelisani. Ovaj poslednji uslov u geometrijskom prikazu problema zahteva da prave linije na koje se projektuje roj tačaka kod prve i druge linearne kombinacije budu međusobom upravne.

Pošto je određen položaj prve prave linije na koju se projektuje roj tačaka, tj. prva glavna komponenta, druga linija leži pod uglom od 90^0 u odnosu na nju. Na Slici 7.1 prikazana je ova druga linija kao i odgovarajuća distribucija frekvencije vrednosti te linearne kombinacije (druga glavna komponenta) u oznaci $Y_2 = \alpha_{21}X_1 + \alpha_{22}X_2$. Sa slike se uočava da su ove dve prave linije paralelne sa glavnom i sporednom osom elipse koja se može nacrtati oko roja tačaka. Tu elipsu smo u drugoj glavi kod izlaganja dvodimenzionog normalnog rasporeda upoznali pod nazivom, elipsa konstantne gustine verovatnoće. To znači da alternativan način dolaska do glavnih komponenata zahteva rotaciju osa koordinatnog sistema za neki ugao sve dok se ne poklope sa osama elipse konstantne gustine verovatnoće. Podsećamo da smo ovaj način rotacije osa koordinatnog sistema koristili prilikom izlaganja generalizovanog odstojanja u okviru 2. Glave.

Slikovito bismo u prvom slučaju opisali postupak određivanja glavnih komponenata kao kretanje osobe oko roja tačaka pri čemu posmatrač traži takvo stajno mesto sa koga će imati najširi vidik na taj roj. U drugom slučaju posmatrač stoji u mestu, a ose koordinatnog sistema se rotiraju sve dok se ne poklope sa glavnim osama roja tačaka.

7.2 POPULACIONE GLAVNE KOMPONENTE

7.2.1 Definicija glavnih komponenata

Pretpostavimo da je $\mathbf{X}$ $p-$dimenzioni slučajan vektor sa kovarijacionom matricom $\mathbf{\Sigma}$. Neka je $Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + ... + \alpha_{1p}X_p = \boldsymbol{\alpha}_1'\mathbf{X}$ linearna kombinacija elemenata slučajnog vektora $\mathbf{X}$, gde su $\alpha_{11}, \alpha_{12}, ..., \alpha_{1p}$ koeficijenti linearne kombinacije. Iz 2. Glave poznato je da je $Var(Y_1) = Var(\boldsymbol{\alpha}_1'\mathbf{X}) = \boldsymbol{\alpha}_1'\mathbf{\Sigma}\boldsymbol{\alpha}_1$. Naš zadatak je da odredimo vektor koeficijenata $\boldsymbol{\alpha}_1$ tako da se maksimizira varijansa od Y_1. Kako se $Var(Y_1) = \boldsymbol{\alpha}_1'\mathbf{\Sigma}\boldsymbol{\alpha}_1$ može

proizvoljno povećavati množenjem vektora $\boldsymbol{\alpha}_1$ proizvoljnim skalarem, to se uvodi ograničenje da je vektor koeficijenata jedinične dužine, tj. da je $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$.

Problem maksimiziranja $\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$ uz ograničenje $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$ rešavamo korišćenjem Lagrangeovih množitelja tako što ćemo maksimizirati Lagrangeovu funkciju

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 - 1) \tag{7.1}$$

gde je λ Lagrangeov množitelj. Diferenciranjem Lagrangeove funkcije po koeficijentima $\boldsymbol{\alpha}_1$, a zatim izjednačavanjem dobijenog izraza sa nulom dobijamo

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1 = 0 \tag{7.2}$$

ili

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I})\boldsymbol{\alpha}_1 = 0 \tag{7.3}$$

gde je $\mathbf{I}$ $(p \times p)$ jedinična matrica. Da bi se dobilo netrivijalno rešenje za $\boldsymbol{\alpha}_1$ determinanta $|\boldsymbol{\Sigma} - \lambda\mathbf{I}|$ mora biti jednaka nuli. To znači da λ mora biti jedan od karakterističnih korena kovarijacione matrice $\boldsymbol{\Sigma}$. Odluka o izboru jednog od karakterističnih korena bazirana je na sledećem razmatranju. Ako pomnožimo s leve strane izraz (7.2) sa $\boldsymbol{\alpha}_1'$ dobićemo

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 0 \tag{7.4}$$

Pošto je $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 - 1 = 0$, sledi da je . Kako težimo da maksimiziramo varijansu za λ ćemo uzeti najveći karakteristični koren, recimo λ_1. Na osnovu uslova $(\boldsymbol{\Sigma} - \lambda\mathbf{I})\boldsymbol{\alpha}_1 = 0$, sledi da je $\boldsymbol{\alpha}_1$ odgovarajući karakteristični vektor pridružen karakterističnom korenu λ_1. Njegovim normiranjem $(\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1)$ dobićemo traženi vektor $\boldsymbol{\alpha}_1$.

Ako nam je zadatak da odredimo više od jedne linearne kombinacije tada postupamo kao u slučaju određivanja prve glavne komponente uz dodatni uslov da kovarijansa prve i druge glavne komponente bude jednaka nuli. Neka je $Y_2 = \boldsymbol{\alpha}_2'\mathbf{X} = \alpha_{21}X_1 + \alpha_{22}X_2 + ... + \alpha_{2p}X_p$ linearna kombinacija čije koeficijente $\alpha_{21}, \alpha_{22}, ..., \alpha_{2p}$ treba odrediti uz normirajući uslov $\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 = 1$, pri čemu se dodatni uslov nekorelisanosti prve i druge glavne komponente svodi na uslov $\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1 = 1$. Rezultat sledi na osnovu toga što je

$$Cov(Y_2, Y_1) = Cov(\boldsymbol{\alpha}_2'\mathbf{X}, \boldsymbol{\alpha}_1'\mathbf{X}) = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1\lambda_1 = \boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2\lambda_1 \tag{7.5}$$

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

$$\boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda(\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 - 1) - \phi\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1 \tag{7.6}$$

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

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1 = 0 \tag{7.7}$$

Ako pomnožimo dobijeni izraz s leve strane sa $\boldsymbol{\alpha}_1'$ dobijamo

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 0 \tag{7.8}$$

Pošto su prva dva člana u ovom izrazu jednaka nuli, a $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$ sledi da je $\phi = 0$. Prema tome je $\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 = \mathbf{0}$ ili $(\boldsymbol{\Sigma} - \lambda\mathbf{I})\boldsymbol{\alpha}_2 = 0$, što znači da je λ karakteristični koren kovarijacione matrice $\boldsymbol{\Sigma}$, a $\boldsymbol{\alpha}_2$ odgovarajući karakteristični vektor. Kao i u slučaju prve glavne komponente biramo za λ što je moguće veću vrednost, jer je $\lambda = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$. Drugi po veličini karakteristični koren označimo sa λ_2, njemu pridruženi karakteristični vektor je $\boldsymbol{\alpha}_2$, a linearna kombinacija $Y_2 = \boldsymbol{\alpha}_2'\mathbf{X}$ predstavlja *drugu glavnu komponentu*.

Na izloženi način može se doći do svih glavnih komponenata kojih ima onoliko koliko ima različitih karakterističnih korena kovarijacione matrice. Ukažimo ukratko na osnovne rezultate u vezi glavnih komponenata. Ako su svi karakteristični koreni matrice $\boldsymbol{\Sigma}$ međusobom različiti (uredili smo ih u opadajući niz $\lambda_1 > \lambda_2 > ... > \lambda_p \geq 0$), tada postoji p glavnih komponenata Y_1, Y_2, …, Y_p ($Y_j = \boldsymbol{\alpha}_j'\mathbf{X}$, $j = 1, 2, ..., p$). Vektori koeficijenata $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, …, $\boldsymbol{\alpha}_p$ predstavljaju karakteristične vektore matrice $\boldsymbol{\Sigma}$ koji su pridruženi karakterističnim korenima λ_j. Očekivana vrednost glavnih komponenata je $E(Y_j) = 0$, varijansa $Var(Y_j) = \lambda_j$, a kovarijansa svakog para glavnih komponenata jednaka je nuli.

Primer 7.1 Za datu kovarijacionu matricu

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

odrediti sve glavne komponente. Pokazati da je varijansa dobijenih glavnih komponenata jednaka odgovarajućim karakterističnim korenima, a da je kovarijansa između svakog para glavnih komponenata jednaka nuli.

Na osnovu determinantne jednačine $|\boldsymbol{\Sigma} - \lambda\mathbf{I}| = 0$ računamo karakteristične korene: $\lambda_1 = 5.2361$, $\lambda_2 = 4$ i $\lambda_3 = 0.7639$. Pridruženi karakteristični

vektori su: $\boldsymbol{\alpha}_1' = [0 \quad -0.5257 \quad 0.8507]$, $\boldsymbol{\alpha}_2' = [1 \quad 0 \quad 0]$ i $\boldsymbol{\alpha}_3' = [0 \quad -0.8507 \quad -0.5257]$. Prema tome, glavne komponente su:

$$Y_1 = \boldsymbol{\alpha}_1'\mathbf{X} = -0.5257X_2 + 0.8507X_3$$

$$Y_2 = \boldsymbol{\alpha}_2'\mathbf{X} = X_1$$

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

Varijanse glavnih komponenata jednake su odgovarajućim karakterističnim korenima kovarijacione matrice. Pokažimo to na primeru prve glavne komponente

$$Var(Y_1) = Var(-0.5257X_2 + 0.8507X_3) = (-0.5257)^2 Var(X_2)$$

$$+(0.8507)^2 Var(X_3) + 2(-0.5257)(0.8507)Cov(X_2, X_3)$$

$$= 0.2764 \cdot (2) + 0.7236 \cdot (4) - 0.8944 \cdot (-2) = 5.2361 = \lambda_1$$

Pokazati navedenu osobinu za preostale dve glavne komponente.

Nekorelisanost glavnih komponenti pokazaćemo na primeru prve i druge glavne komponente, a provera za preostale parove ostavlja se čitaocu.

$$Cov(Y_1, Y_2) = Cov(-0.5257X_2 + 0.8507X_3, X_1)$$

$$= -0.5257Cov(X_2, X_1) + 0.8507Cov(X_3, X_1)$$

$$= -0.5257 \cdot (0) + 0.8507 \cdot (0) = 0\,.$$

Dobijen je rezultat da je druga glavna komponenta jednaka prvoj originalnoj promenljivoj. Kako je u originalnom skupu promenljiva X_1 nekorelisana sa preostale dve promenljive ovo je bio očekivan rezultat, s obzirom da se glavne komponente tako konstruišu da su međusobom nekorelisane. ■

7.2.2 Osobine glavnih komponenata

Definišući glavne komponente ukazali smo na njihove sledeće osobine koje direktno slede na osnovu definicije:

$$E(Y_j) = 0\,,\ Var(Y_j) = \lambda_j\,,\ Cov(Y_i, Y_j) = 0,\ i \neq j$$

$$Var(Y_1) \geq Var(Y_2) \geq ... \geq Var(Y_p) \geq 0 \tag{7.9}$$

Sada ćemo, pozivajući se na dve alternativne definicije generalizovane varijanse (trag i determinanta kovarijacione matrice), pokazati da su generalizovane varijanse glavnih komponenata jednake generalizovanim varijansama originalnog skupa promenljivih. Prethodno ćemo se podsetiti na rezultate iz matrične algebre potrebne za naš dokaz.

Neka je $\mathbf{Y}$ vektor glavnih komponenata takav da je $\mathbf{Y}' = [Y_1, Y_2, ..., Y_p]$. Sada se transformacija originalnog skupa promenljivih sadržanog u vektoru $\mathbf{X}$ može pisati na sledeći način: $\mathbf{Y} = \mathbf{AX}$, gde je $\mathbf{A}$ $(p \times p)$ matrica čiji su redovi karakteristični vektori kovarijacione matrice $\boldsymbol{\Sigma}$, odnosno $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_p$, pridruženi odgovarajućim karakterističnim korenima $\lambda_1, \lambda_2, ..., \lambda_p$. Na osnovu osobina karakterističnih vektora ($\boldsymbol{\alpha}_j' \boldsymbol{\alpha}_j = 1$ i $\boldsymbol{\alpha}_i' \boldsymbol{\alpha}_j = 0$, $i \neq j$) matrica $\mathbf{A}$ ima osobinu da je $\mathbf{A}' = \mathbf{A}^{-1}$, pa se $\mathbf{Y} = \mathbf{AX}$ naziva *ortogonalna transformacija* ili *rotacija*, a sama matrica $\mathbf{A}$ ortogonalna matrica. Njena osobina je i da je $|\mathbf{A}| = \pm 1$. Transformacija se naziva ortogonalna jer se njome vrši rotacija koordinatnih osa za izvestan ugao, pri čemu ose ostaje upravne međusobom, a ugao između ma koja dva vektora ostaje isti nakon transformacije.

Primer 7.2 Na osnovu podataka i rezultata iz prethodnog primera formirati matricu $\mathbf{A}$ i pokazati da se zaista radi o ortogonalnoj matrici.

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{\alpha}_1' \\ \boldsymbol{\alpha}_2' \\ \boldsymbol{\alpha}_3' \end{bmatrix} = \begin{bmatrix} 0 & -0.5257 & 0.8507 \\ 1 & 0 & 0 \\ 0 & -0.8507 & -0.5257 \end{bmatrix}.$$

Vrednost determinante matrice $\mathbf{A}$ je

$$|\mathbf{A}| = (-1)^{2+1} \begin{vmatrix} -0.5257 & 0.8507 \\ -0.8507 & -0.5257 \end{vmatrix} = (-1)(0.2763 + 0.7237) = -1$$

a njena inverzna matrica je

$$\mathbf{A}^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ -0.5257 & 0 & -0.8507 \\ 0.8507 & 0 & -0.5257 \end{bmatrix}.$$

Očigledno je ispunjen uslov da je $\mathbf{A}' = \mathbf{A}^{-1}$, pa je matrica $\mathbf{A}$ zaista ortogonalna matrica. ■

Korišćenjem ortogonalne matrice $\mathbf{A}$ možemo izvršiti ortogonalnu dekompoziciju kvadratne simetrične matrice čiji su koreni različiti. Imamo da je $\boldsymbol{\Sigma} = \mathbf{A}'\boldsymbol{\Lambda}\mathbf{A}$, gde je

$\mathbf{\Lambda}$ dijagonalna matrica čiji su elementi karakteristični koreni matrice $\mathbf{\Sigma}$, a matrica $\mathbf{A}$ je ranije definisana ortogonalna matrica čiji su redovi karakteristični vektori kovarijacione matrice $\mathbf{\Sigma}$. Kako je vektor glavnih komponenata $\mathbf{Y} = \mathbf{AX}$, to je njegova kovarijaciona matrica $Var(\mathbf{Y}) = \mathbf{A\Sigma A}'$. Ako sada zamenimo $\mathbf{\Sigma}$ dobićemo $Var(\mathbf{Y}) = \mathbf{A}(\mathbf{A}'\mathbf{\Lambda A})\mathbf{A}' = \mathbf{\Lambda}$, pošto je $\mathbf{A}$ ortogonalna matrica, te je $\mathbf{A}'\mathbf{A} = \mathbf{I}$, čime smo na drugačiji način izveli glavne komponente.

Na osnovu dobijenog rezultata možemo odrediti generalizovanu varijansu vektora $\mathbf{Y}$. Prema prvoj definiciji genaralizovana varijansa jednaka je determinanti kovarijacione matrice. Kao što smo videli, kovarijaciona matrica glavnih komponenata je $\mathbf{\Lambda}$. Njena determinanta je $|\mathbf{\Lambda}|$ i jednaka je proizvodu karakterističnih korena λ_j. Na osnovu izraza ortogonalne dekompozicije matrice $\mathbf{\Sigma}$ dobijamo da je $\mathbf{\Lambda} = \mathbf{A\Sigma A}'$. Prema osobini determinanti, da je determinanta proizvoda dve matrice jednaka proizvodu njihovih determinanti, imamo da je $|\mathbf{\Lambda}| = |\mathbf{A\Sigma A}'| = |\mathbf{A}||\mathbf{\Sigma}||\mathbf{A}'| = |\mathbf{\Sigma}|$, pošto je $|\mathbf{A}| = \pm 1$. Znači da su prema prvoj definiciji generalizovane varijanse originalnog i transformisanog skupa podataka međusobom jednake.

Prema drugoj definiciji generalizovana varijansa jednaka je tragu kovarijacione matrice. Trag kovarijacione matrice glavnih komponenata jednak je zbiru karakterističnih korena λ_j. Prema izrazu ortogonalne dekompozicije matrice $\mathbf{\Sigma} = \mathbf{A}'\mathbf{\Lambda A}$, dobijamo da je $\mathbf{\Lambda} = \mathbf{A\Sigma A}'$. Uz korišćenje osobine traga matrice ($\mathrm{tr}(\mathbf{BC}) = \mathrm{tr}(\mathbf{CB})$) imamo da je $\mathrm{tr}(\mathbf{\Lambda}) = \mathrm{tr}(\mathbf{A\Sigma A}') = \mathrm{tr}(\mathbf{A}'\mathbf{A\Sigma}) = \mathrm{tr}(\mathbf{\Sigma})$, pošto je $\mathbf{A}'\mathbf{A} = \mathbf{I}$. Znači da su i prema drugoj definiciji generalizovane varijanse originalnog i transformisanog skupa podataka međusobom jednake.

Relativan doprinos $j-$te glavne komponente "objašnjenju" ukupne varijanse određujemo stavljajući u međusobni odnos odgovarajući karakteristični koren λ_j i generalizovanu varijansu (druga definicija), tj. ovaj doprinos određujemo na osnovu izraza

$$\frac{\lambda_j}{\sum_{k=1}^{p} \lambda_k}, \quad j = 1, 2, \ldots, p \tag{7.10}$$

Primer 7.3 Na osnovu podataka iz Primera 7.1 odrediti generalizovane varijanse (obe definicije) originalnog i transformisanog skupa podataka kao i relativan doprinos svake od glavnih komponenti.

Determinante kovarijacionih matrica su

$$|\mathbf{\Sigma}| = 4 \cdot \begin{vmatrix} 2 & -2 \\ -2 & 4 \end{vmatrix} = 16 \text{ i } |\mathbf{\Lambda}| = (5.2361)(4)(0.7639) = 16$$

Tragovi kovarijacionih matrica su

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

Generalizovane varijanse (obe definicije) vektora $\mathbf{X}$ i $\mathbf{Y}$ međusobom su jednake. Relativan doprinos prve glavne komponente ukupnom varijabilitetu je

$$\frac{\lambda_1}{\text{tr}(\boldsymbol{\Lambda})} = \frac{5.2361}{10} = 0.5236 \quad \text{ili} \quad 52.36\%,$$

druge glavne komponente

$$\frac{\lambda_2}{\text{tr}(\boldsymbol{\Lambda})} = \frac{4}{10} = 0.4 \quad \text{ili} \quad 40\%$$

i treće glavne komponente

$$\frac{\lambda_3}{\text{tr}(\boldsymbol{\Lambda})} = \frac{0.7639}{10} = 0.0764 \quad \text{ili } 7.64\%. \qquad \blacksquare$$

Ukoliko u analizi glavnih komponenata dobijemo relativno visok doprinos jedne ili nekoliko prvih glavnih komponenata ukupnoj varijansi, tada je moguće dalju analizu zasnovati na njima, a ne na svim glavnim komponentama. Tako u Primeru 7.3 prve dve glavne komponente doprinose zajedno preko 90% ukupnoj varijansi trodimenzionog skupa podataka. U ovom kontekstu interesantno je ukazati na razlaganje kovarijacione matrice $\boldsymbol{\Sigma}$ na matrice doprinosa svake glavne komponente kovarijacionoj strukturi originalnog skupa podataka.

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

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

$$= \lambda_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1' + \lambda_2\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2' + \ldots + \lambda_p\boldsymbol{\alpha}_p\boldsymbol{\alpha}_p' = \sum_{j=1}^{p} \lambda_j\boldsymbol{\alpha}_j\boldsymbol{\alpha}_j' \tag{7.11}$$

Znači da je doprinos $j-$te glavne komponente kovarijacionoj matrici $\boldsymbol{\Sigma}$ dat matricom $\lambda_j\boldsymbol{\alpha}_j\boldsymbol{\alpha}_j'$. Zadržavajući manji broj glavnih komponenata od p, kovarijacionu matricu $\boldsymbol{\Sigma}$ aproksimiramo zbirom matrica doprinosa zadržanih glavnih komponenata. Ukoliko doprinos ukupnoj varijansi zadržanih glavnih komponenata prelazi neku unapred fiksiranu vrednost, na primer 80-90%, tada je za očekivati da će ta aproksimacija kovarijacione matrice $\boldsymbol{\Sigma}$ relativno dobro reprezentovati kovarijacionu

strukturu originalnog skupa podataka. Međutim, ovo nije glavni, nego tek prateći rezultat analize. Glavne komponente su skoncentrisane na aproksimaciju varijanse, a ne kovarijanse promenljivih, kao što je to slučaj sa faktorskom analizom.

Primer 7.4 Koristeći podatke i rezultate iz Primera 7.1 aproksimirati kovarijacionu matricu $\mathbf{\Sigma}$: a) prvom, b) prvom i drugom i c) svim glavnim komponentama.

Doprinos glavnih komponenata kovarijacionoj matrici dat je sledećim matricama:

$$\lambda_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1' = 5.2361\begin{bmatrix}0\\-0.5257\\0.8507\end{bmatrix}\begin{bmatrix}0 & -0.5257 & 0.8507\end{bmatrix} = \begin{bmatrix}0 & 0 & 0\\0 & 1.4472 & -2.3417\\0 & -2.3417 & 3.7889\end{bmatrix}$$

$$\lambda_2\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2' = 4\begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}1 & 0 & 0\end{bmatrix} = \begin{bmatrix}4 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}$$

$$\lambda_3\boldsymbol{\alpha}_3\boldsymbol{\alpha}_3' = 0.7639\begin{bmatrix}0\\-0.8507\\-0.5257\end{bmatrix}\begin{bmatrix}0 & -0.8507 & -0.5257\end{bmatrix} = \begin{bmatrix}0 & 0 & 0\\0 & 0.5528 & 0.3416\\0 & 0.3416 & 0.2111\end{bmatrix}$$

a) Prva glavna komponenta doprinosi 52.36% ukupnoj varijansi. Ukoliko samo nju koristimo tada je matricom $\lambda_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1'$ data aproksimacija kovarijacione matrice $\mathbf{\Sigma}$.

b) Prve dve glavne komponente zajedno doprinose 92.36% ukupnoj varijansi. Njihovim korišćenjem kovarijacionu matricu $\mathbf{\Sigma}$ aproksimiramo matricom $\lambda_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1' + \lambda_2\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2'$:

$$\lambda_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1' + \lambda_2\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2' = \begin{bmatrix}4 & 0 & 0\\0 & 1.4472 & -2.3417\\0 & -2.3417 & 3.7889\end{bmatrix}$$

U poređenju sa originalnom matricom $\mathbf{\Sigma}$ vidi se da prve dve glavne komponente, odnosno dobijena matrica, dobro aproksimiraju kovarijacionu strukturu.

c) Ukoliko sve tri glavne komponente uključimo u analizu, tada ćemo verno reprodukovati kovarijacionu matricu $\mathbf{\Sigma}$. Naime, sve tri glavne komponente "objašnjavaju" 100% ukupne varijanse originalnog skupa podataka, pa je $\mathbf{\Sigma} = \lambda_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1' + \lambda_2\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2' + \lambda_3\boldsymbol{\alpha}_3\boldsymbol{\alpha}_3'$. ■

7.2.3 Interpretacija glavnih komponenata

Metod glavnih komponenata je statistički postupak transformacije originalnog, korelisanog skupa podataka u skup nekorelisanih promenljivih sa opadajućom vrednošću varijansi. Ova konstatacija istovremeno predstavlja kako sažet opis karakteristika glavnih komponenata, tako i opravdanje često neuspešnih pokušaja njihovih smislenih interpretacija. Uprkos činjenici da su glavne komponente proizvod mehaničkog postupka transformacije jednog u drugi skup podataka, pokušaj njihove interpretacije možemo zasnovati na apsolutnoj veličini i predznaku koeficijenata linearne kombinacije. Mada je metod suštinski različit od metoda regresione analize u interpretaciji postupamo slično kao kod ovog potonjeg. Koeficijenti bliski nuli sugerišu da odgovarajuća originalna promenljiva ne učestvuje značajno u formiranju odnosne glavne komponente.

Do sada smo analizu glavnih komponenata bazirali na kovarijacionoj matrici $\mathbf{\Sigma}$. Problem koji se javlja u interpretaciji glavnih komponenata posledica je njihove osetljivosti na različite merne skale originalnih promenljivih. Ako u analizi jedna od promenljivih ima znatno veću varijansu od ostalih, tada će ona dominirati prvom glavnom komponentom bez obzira na korelacionu strukturu podataka. Jedna mogućnost je da u tom slučaju ne koristimo direktno koeficijente linearne kombinacije u cilju interpretacije glavnih komponenata, nego da analizu zasnivamo na koeficijentima korelacije originalnih promenljivih i glavnih komponenata. Druga mogućnost je da celu analizu baziramo na korelacionoj, a ne kovarijacionoj matrici originalnih podataka. Osim u slučaju kada su originalne promenljive istovrsne (merene na istoj mernoj skali), rezultati analize u ova dva slučaja razlikovaće se međusobno.

Odredimo koeficijente korelacije između originalnih promenljivih i glavnih komponenata. Ranije smo odredili kovarijacionu matricu od $\mathbf{Y}$, tj. $Var(\mathbf{Y}) = \mathbf{\Lambda}$, a $Var(\mathbf{X}) = \mathbf{\Sigma}$. Kovarijansa između $\mathbf{X}$ i $\mathbf{Y}$ je

$$Cov(\mathbf{X}, \mathbf{Y}) = Cov(\mathbf{X}, \mathbf{AX}) = \mathbf{\Sigma A}' = (\mathbf{A}'\mathbf{\Lambda A})\mathbf{A}' = \mathbf{A}'\mathbf{\Lambda} = \left[\boldsymbol{\alpha}_1\lambda_1, \boldsymbol{\alpha}_2\lambda_2, ..., \boldsymbol{\alpha}_3\lambda_3\right] \tag{7.12}$$

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

$$\rho_{X_k Y_j} = \frac{Cov(X_k, Y_j)}{\sqrt{Var(X_k)}\sqrt{Var(Y_j)}} = \frac{\lambda_j \alpha_{jk}}{\sqrt{\sigma_{kk}}\sqrt{\lambda_j}} = \alpha_{jk}\frac{\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}, \quad j,k = 1,2,...,p \tag{7.13}$$

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

$$\boldsymbol{\rho}_{\mathbf{XY}} = \mathbf{\Lambda}^{1/2}\mathbf{A}\mathbf{D}^{-1/2} \tag{7.14}$$

gde smo sa **D** označili dijagonalnu matricu čiji su elementi varijanse originalnih promenljivih.

Primer 7.5 Korišćenjem podataka i rezultata iz Primera 7.1 odrediti koeficijente korelacije originalnih promenljivih i glavnih komponenata.

Elementi potrebni za određivanje koeficijenata korelacije sadržani su u sledećim matricama:

$$\mathbf{\Sigma}=\begin{bmatrix}4 & 0 & 0\\ 0 & 2 & -2\\ 0 & -2 & 4\end{bmatrix},\ \mathbf{A}=\begin{bmatrix}0 & -0.5257 & 0.8507\\ 1 & 0 & 0\\ 0 & -0.8507 & -0.5257\end{bmatrix},\ \mathbf{\Lambda}=\begin{bmatrix}5.2361 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 0.7639\end{bmatrix}$$

Koeficijent korelacije između druge originalne promenljive i prve glavne komponente je

$$\rho_{X_2Y_1}=\alpha_{12}\frac{\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}}=-0.5257\frac{\sqrt{5.2361}}{\sqrt{2}}=-0.8506$$

Ostali koeficijenti korelacije sadržani su u donjoj tabeli.

Koeficijenti korelacije		Glavne komponente		
		Y_1	Y_2	Y_3
Originalne promenljive	X_1	0	1	0
	X_2	-0.8506	0	-0.5258
	X_3	0.9733	0	-0.2297

U pogledu interpretacije glavnih komponenata dobijeni rezultati ukazuju na saglasnost zaključaka izvedenih u analizi zasnovanoj na koeficijentima linearne kombinacije i koeficijentima korelacije originalnih promenljivih i glavnih komponenata. Ovo je očekivani rezultat uzimajući u obzir da su varijanse tri originalne promenljive istog reda veličine. ■

Drugi pristup problemu osetljivosti rezultata analize glavnih komponenata na mernu skalu zasnovan je na korišćenju korelacione matrice originalnih promenljivih umesto njihove kovarijacione matrice. S obzirom da se korelaciona matrica može smatrati kovarijacionom matricom standardizovanih promenljivih, na osnovu dosadašnjih rezultata, ukažimo na osnovne pokazatelje analize u ovom specijalnom slučaju: ukupan varijabilitet meren generalizovanom varijansom (trag kovarijacione matrice) jednak je p, tj. dimenziji korelacione matrice, a koeficijent korelacije između $k-$te originalne promenljive i $j-$te glavne komponente jednak je $\alpha_{jk}\sqrt{\lambda_j}$. U matričnom zapisu odgovarajuća korelaciona matrica je data izrazom: $\mathbf{A\Lambda}^{1/2}$. Rezultati analize glavnih

komponenata na bazi kovarijacione, odnosno korelacione matrice istog skupa podataka međusobom se mogu znatno razlikovati. Istovremeno ne postoji jednostavan put za prevođenje jednog skupa glavnih komponenata u drugi. Mada se glavne komponente računate na osnovu kovarijacione matrice razlikuju od onih koje su izvedene na osnovu korelacione matrice sam računski postupak u oba slučaja je u suštini identičan.

Primer 7.6 Ilustrujmo izrečeno o analizi glavnih komponenata zasnovanoj na kovarijacionoj i korelacionoj matrici na primeru kovarijacione matrice iz Primera 7.1.

Na osnovu kovarijacione matrice dobijamo korelacionu matricu

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

Svi rezultati analize prezentirani su u donjoj tabeli, gde su poređenja radi dati i rezultati analize na osnovu kovarijacione matrice.

Glavne komponente:		**Kovarijaciona matrica**			**Korelaciona matrica**		
		prva	druga	treća	prva	druga	treća
		Koeficijenti linearne kombinacije					
Originalne promenljive	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 komponenata 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

Dobijeni rezultati se u ova dva slučaja međusobno ne razlikuju u većoj meri. S obzirom na to da su varijanse originalnih promenljivih istog reda veličine ovo je bilo očekivano. Uočava se da prva i treća glavna komponenta računate na osnovu korelacione matrice ukazuju na podjednak doprinos druge i treće originalne promenljive, bez obzira da li je reč o koeficijentima linearne kombinacije ili o koeficijentima korelacije glavnih komponenata i originalnih promenljivih. ■

Naposletku, ukažimo na iznos varijanse originalnih promenljivih koji je "objašnjen" zadržanim skupom glavnih komponenata. On pokazuje u kom stepenu zadržane glavne komponente dobro aproksimiraju varijansu svake originalne promenljive ponaosob. Na osnovu izrazu za ortogonalnu dekompoziciju kovarijacione matrice ($\boldsymbol{\Sigma} = \mathbf{A}'\boldsymbol{\Lambda}\mathbf{A}$) imamo da je varijansa $k-$te promenljive

$$\sigma_{kk}^2 = \sum_{j=1}^{p} \lambda_j \alpha_{jk}^2 \ , \ k = 1, 2, ..., p \tag{7.15}$$

Znači da je doprinos svake glavne komponente varijansi $k-$te promenljive jednak kvadratu koeficijenta korelacije odnosne glavne komponente i te originalne promenljive. Doprinos svih glavnih komponenata računamo na osnovu korelacione matrice $\mathbf{A}\mathbf{\Lambda}^{1/2}$ tako što ćemo sabrati kvadrate elemenata u njenom $k-$tom redu. Ukoliko smo u analizi zadržali nekoliko prvih glavnih komponenata tada stavljanjem u odnos dobijene sume i odgovarajuće varijanse originalne promenljive dobijamo proporciju varijanse te promenljive koja je "objašnjena" zadržanim glavnim komponentama. Ova proporcija u analizi glavnih komponenata i u faktorskoj analizi naziva se *komunalitet* promenljive i može se uslovno čitati kao koeficijent determinacije (procenat "objašnjenja" varijanse originalne promenljive zadržanim glavnim komponentama). Korišćenjem korelacione matrice umesto kovarijacione matrice originalnih promenljivih odmah dobijamo proporciju varijanse originalne promenljive "objašnjene" zadržanim glavnim komponentama jer je standardizacijom promenljivih vrednost varijanse jednaka jedinici.

Primer 7.7 Pretpostavimo da smo u Primeru 7.1 za potrebe dalje analize zadržali prve dve glavne komponente. Odrediti koja proporcija varijanse svake od tri originalne promenljive je "objašnjena" dvema zadržanim glavnim komponentama.

Prema navedenom izrazu deo varijanse prve originalne promenljive "objašnjen" dvema glavnim komponentama iznosi

$$\lambda_1 \alpha_{11}^2 + \lambda_2 \alpha_{21}^2 = 5.2361 \cdot (0)^2 + 4 \cdot (1)^2 = 4 \ ,$$

tj. prve dve glavne komponente "objašnjavaju" svih 100% varijanse prve originalne promenljive. Razumljivo, pošto je druga glavna komponenta jednaka drugoj originalnoj promenljivoj.

Deo varijanse druge originalne promenljive "objašnjen" dvema glavnim komponentama iznosi

$$\lambda_1 \alpha_{12}^2 + \lambda_2 \alpha_{22}^2 = 5.2361 \cdot (-0.5257)^2 + 4 \cdot (0)^2 = 1.4471$$

odn. 72.35% varijanse druge originalne promenljive "objašnjeno" je prvim dvema glavnim komponentama.

Deo varijanse treće originalne promenljive "objašnjen" dvema glavnim komponentama iznosi

$$\lambda_1 \alpha_{13}^2 + \lambda_2 \alpha_{23}^2 = 5.2361 \cdot (0.8507)^2 + 4 \cdot (0)^2 = 3.7893$$

odn. 94.73% varijanse treće originalne promenljive "objašnjeno" je prvim dvema glavnim komponentama. ■

7.3 UZORAČKE GLAVNE KOMPONENTE

Teorijski okvir definisan u prethodnom poglavlju koristimo u analizi glavnih komponenata kada raspolažemo uzorkom iz $p-$dimenzione populacije. Sa analitičko-računske strane ne postoje razlike u izvođenju i korišćenju populacionih i uzoračkih glavnih komponenata. Zato ćemo izostaviti ponavljanje analitičkog okvira u kontekstu slučajnog uzorka, a pažnju ćemo posvetiti novim elementima koji se javljaju pri njegovom korišćenju.

7.3.1 Ocena glavnih komponenata

Neka je uzet slučajan uzorak od n elemenata $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ iz $p-$dimenzione populacije sa sredinom $\boldsymbol{\mu}$ i kovarijacionom matricom $\boldsymbol{\Sigma}$. Analizu glavnih komponenata zasnivamo na uzoračkoj kovarijacionoj matrici $\mathbf{S}$ ili na uzoračkoj korelacionoj matrici $\mathbf{R}$. U oba slučaja metod glavnih komponenata koristi se u deskriptivne svrhe. Međutim, ukoliko uvedemo dodatnu pretpostavku, da je populacija odakle je uzet uzorak normalna $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, tada je moguće izvesti brojne rezultate koji se tiču asimptotskih svojstava glavnih komponenata i pratećih pokazatelja. Na ovom mestu ukratko ćemo navesti, bez dokaza, neke od najvažnijih rezultata (za dokaz videti: Anderson (1972, ss. 279-284)).

Ocena najveće verodostojnosti kovarijacione matrice $\boldsymbol{\Sigma}$, odn. $\hat{\boldsymbol{\Sigma}} = ((n-1)/n)\mathbf{S}$, predstavlja polaznu veličinu u uzoračkoj analizi glavnih komponenata. Uz pretpostavku da su svi različiti, njeni karakteristični koreni i vektori predstavljaju ocenu najveće verodostojnosti odgovarajućih populacionih karakterističnih korena i vektora. Umesto ocene najveće verodostojnosti kovarijacione matrice $\boldsymbol{\Sigma}$, tj. $\hat{\boldsymbol{\Sigma}}$, možemo koristiti i nepristrasnu ocenu kovarijacione matrice $\boldsymbol{\Sigma}$, tj. uzoračku kovarijacionu matricu $\mathbf{S}$. Bez obzira koja ocena od se koristi, glavne komponente su identične u oba slučaja, kao i proporcija objašnjene varijanse. Istovremeno, $\hat{\boldsymbol{\Sigma}}$ i $\mathbf{S}$ daju istu uzoračku korelacionu matricu $\mathbf{R}$, pa izbor između ocena postaje bespredmetan.

Na osnovu ocenjenih koeficijenata linearne kombinacije možemo izračunati vrednosti opservacija uzoračkih glavnih komponenata koje nazivamo *skorovi glavnih komponenata*. Oni se za $j-$tu glavnu komponentu računaju prema izrazu: $\hat{Y}_j = \mathbf{a}_j'\mathbf{X}$, gde smo sa $\mathbf{a}_j$ označili ocenu vektora koeficijenata linearne kombinacije promenljive $X_1, X_2, ..., X_p$, a računamo ih za svaki element uzorka. Skorove glavnih komponenata koristimo na primer u grafičkom prikazu opservacija u dvodimenzionom prostoru generisanom parovima glavnih komponenata. Na taj način vizuelno se može zaključiti o prisustvu nestandardnih opservacija u uzorku.

Prema dosadašnjem izlaganju sledi zaključak, da u osnovi analize glavnih komponenata leže upravo karakteristični koreni i vektori kovarijacione ili korelacione matrice. Korišćenje uzoračkih ocena odgovarajućih populacionih veličina implicitno podrazumeva da će se one, zbog slučajnih varijacija, razlikovati od svojih populacionih pandana. Zato navodimo njihov asimptotski raspored, pod pretpostavkom da je iz višedimenzionog normalnog rasporeda uzet slučajan uzorak od n elemenata, a u cilju određivanja intervala poverenja i testiranja hipoteza o vrednosti karakterističnih korena.

Neka je za kovarijacionu matricu $\mathbf{\Sigma}$, sa $\mathbf{\Lambda}$ označena dijagonalna matrica karakterističnih korena $\lambda_1, \lambda_2, ..., \lambda_p$. Imamo asimptotski raspored uzoračkih karakterističnih korena

$$\sqrt{n}\left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\right) \sim N_p(\mathbf{0}, 2\mathbf{\Lambda}^2) \tag{7.16}$$

gde smo sa $\hat{\boldsymbol{\lambda}}' = [\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_p]$ označili ocenu vektora karakterističnih korena $\boldsymbol{\lambda}$ zasnovanu na uzoračkoj kovarijacionoj matrici. Znači da su karakteristični koreni u velikim uzorcima međusobom nezavisni. Ako formiramo matricu

$$\mathbf{\Omega}_j = \lambda_j \sum_{\substack{k=1 \\ k \neq j}}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_j)^2} \boldsymbol{\alpha}_j \boldsymbol{\alpha}'_j \tag{7.17}$$

tada imamo asimptotski raspored karakterističnih vektora

$$\sqrt{n}\left(\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j\right) \sim N_p(\mathbf{0}, \mathbf{\Omega}_j) \tag{7.18}$$

pri čemu je $\hat{\lambda}_j$ nezavisno raspoređeno od elemenata pridruženog karakterističnog vektora $\hat{\boldsymbol{\alpha}}_j$. Prema navedenim rezultatima možemo formirati interval poverenja svakog karakterističnog korena s obzirom na njihov raspored. Naime, raspored $\hat{\lambda}_j \sim N_p(\lambda_j, 2\lambda_j^2 / n)$ implicira asimptotski $100(1-\alpha)\%$ interval poverenja za ocenu $j-$tog karakterističnog korena

$$\frac{\hat{\lambda}_j}{\left(1 + z_{\alpha/2}\sqrt{\frac{2}{n}}\right)} \leq \lambda_j \leq \frac{\hat{\lambda}_j}{\left(1 - z_{\alpha/2}\sqrt{\frac{2}{n}}\right)} \tag{7.19}$$

gde je $z_{\alpha/2}$ gornji $(100\alpha/2)$ percentil standardizovanog normalnog rasporeda. Napominjemo da dobijeni interval poverenja može biti veoma širok čak i za velike vrednosti n, tj. veličine uzorka. Iz tog razloga postupak određivanja broja glavnih komponenti koje treba zadržati u daljoj analizi ne možemo zasnivati na korišćenju dobijene intervalne ocene karakterističnog korena.

Primer 7.8 Pretpostavimo da je uzet slučajan uzorak od 100 elemenata iz populacije $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, gde je $\boldsymbol{\Sigma}$ pozitivno definitna matrica sa različitim karakterističnim korenima. Konstruisati 95% interval poverenja prvog karakterističnog korena, ako je izračunata uzoračka kovarijaciona matrica

$$\mathbf{S} = \begin{bmatrix} 10 & 3 & -1 \\ 3 & 6 & 4 \\ -1 & 4 & 8 \end{bmatrix}$$

Prvi uzorački karakteristični koren je $\hat{\lambda}_1 = 12.1039$, a gornjih 97.5% standardizovanog normalnog rasporeda je $z_{0.025} = 1.96$. 95% interval poverenja je dat nejednakošću

$$\frac{12.1039}{\left(1+1.96\sqrt{\frac{2}{100}}\right)} \le \lambda_1 \le \frac{12.1039}{\left(1-1.96\sqrt{\frac{2}{100}}\right)} \text{ ili } 9.4770 \le \hat{\lambda}_1 \le 16.7455 \qquad \blacksquare$$

7.3.2 Testiranje značajnosti glavnih komponenata

Analiza glavnih komponenata predstavlja metod za redukciju podataka i kao takva nije zasnovana na teorijskom modelu, nasuprot njoj bliske faktorske analize. Apriori nije nam sugerisano sa koliko je glavnih komponenata potrebno izvršiti analizu da bi se obuhvatio relevantan iznos ukupne varijanse. Iz dosadašnjeg izlaganja proizilazi da smo u analizi zainteresovani za one glavne komponente koje imaju najveće karakteristične korene. Odatle ne sledi da su analitički manje interesantne glavne komponente sa manjim karakterističnim korenima. One su od koristi u postupku utvrđivanja otklona od pretpostavke o normalnosti. Zbog svega toga, kada govorimo o značajnosti glavnih komponenata imamo u vidu testove karakterističnih korena.

Za karakteristične korene kovarijacione matrice najpoznatiji je test koji se pripisuje Bartlettu, za testiranje hipoteze da su poslednja $(p-k)$ karakteristična korena međusobom jednaka. Dakle, nulta hipoteza je $H_0 : \lambda_{k+1} = \lambda_{k+2} = ... = \lambda_p$, protiv alternativne hipoteze H_1 da su barem dva od poslednjih $(p-k)$ karakteristična korena međusobom različiti. Ako prihvatimo nultu hipotezu, tada u analizi koristimo samo prvih k glavnih komponenata jer za njih pretpostavljamo da obuhvataju relevantan iznos ukupnog varijabiliteta, a da poslednje $(p-k)$ glavne komponente jednakog varijabiliteta mere samo "šum" u podacima.

Statistika testa za testiranje navedene hipoteze konstruisana je uz pretpostavku o normalnosti, a zasnovana je na korišćenju principa količnika najveće verodostojnosti (eng. *Likelihood Ratio test* - otuda skraćena oznaka testova ovog tipa: *LR* testovi)

$$LR = \left(\frac{\prod_{j=k+1}^{p} \hat{\lambda}_j}{\left[\frac{1}{p-k} \sum_{j=k+1}^{p} \hat{\lambda}_j \right]^{p-k}} \right)^{\frac{n}{2}} \tag{7.20}$$

gde su $\hat{\lambda}_j$ karakteristični koreni uzoračke kovarijacione matrice. Prema navedenom izrazu zaključujemo da je test zasnovan na poređenju geometrijske i aritmetičke sredine poslednja $(p-k)$ karakteristična korena. Poznato je da se geometrijska sredina koristi ako u podacima ima ekstremnih vrednosti, jer izravnava proporcionalne promene podataka, a da je uvek manja ili jednaka aritmetičkoj sredini (Žižić, Lovrić i Pavličić (1992, ss. 44-46)). Ukoliko je tačna nulta hipoteza *LR* statistika ima vrednost jednaku nuli. U suprotnom, udaljavajući se od nulte hipoteze razlika između aritmetičke i geometrijske sredine postajaće sve veća, što znači da će vrednost *LR* statistike biti sve manja. U tom slučaju odbacili bismo nultu hipotezu o jednakosti poslednja $(p-k)$ karakteristična korena. U primeni ovog testa koristimo asimptotski raspored statistike $-2\ln LR$ koja ima asimptotski χ^2-raspored (ako je tačna nulta hipoteza) sa $\nu = 1/2(p-k+2)(p-k-1)$ stepeni slobode, pri čemu je izvršena modifikacija prvobitne aproksimacije u cilju njenog poboljšanja

$$\left[n - \frac{1}{6}(2p+11) \right] \left[(p-k)\ln \bar{\lambda}_{p-k} - \sum_{j=k+1}^{p} \ln \hat{\lambda}_j \right] \sim \chi^2_\nu \tag{7.21}$$

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

Primer 7.9 Koristeći podatke i rezultate iz Primera 7.8 testirati hipotezu da su poslednja dva karakteristična korena međusobom jednaka.

Nulta hipoteza je $H_0 : \lambda_2 = \lambda_3$, protiv alternativne $H_1 : \lambda_2 \neq \lambda_3$. Kako je $\hat{\lambda}_2 = 10.1157$ i $\hat{\lambda}_3 = 1.7805$, to je $\bar{\lambda}_{3-1} = 5.9481$, gde smo koristili $p = 3$, $k = 1$, a prema zadatku je $n = 100$. Vrednost statistike testa je

$$[100 - \frac{1}{6}(2\cdot 3+11)][(3-1)\ln(5.9481) - \ln(10.1157) - \ln(1.7805)] = 66.6033$$

Kako je kritična vrednost testa $\chi^2_{2;0.05} = 5.99$ manja od izračunate, na nivou značajnosti od 5% odbacujemo nultu hipotezu o jednakosti poslednja dva karakteristična korena. ■

Testovi karakterističnih korena mogu biti od pomoći u izboru broja glavnih komponenata. Međutim, oni kao i intervali poverenja sugerišu značajnost velikog broja karakterističnih korena, a time i glavnih komponenata. Stoga se pri izboru češće koriste drugi kriterijumi.

Ako analizu glavnih komponenata obavljamo na korelacionoj matrici, tada izložena statistika testa za testiranje jednakosti poslednja $(p-k)$ karakteristična korena kovarijacione matrice nema svog parnjaka. U ovom slučaju koristi se ranije izloženi test o nezavisnosti promenljivih, pa je nulta hipoteza $H_0:\boldsymbol{\rho}=\mathbf{I}$, gde je $\boldsymbol{\rho}$ korelaciona matrica originalnih promenljivih. Drugim rečima, testiramo hipotezu da su karakteristični koreni jednaki međusobom. Statistika testa

$$-\left[n-1-\frac{1}{6}(2p+5)\right]\ln|\mathbf{R}| \tag{7.22}$$

ima asimptotski χ^2-raspored sa $1/2p(p-1)$ stepeni slobode, gde je sa $\mathbf{R}$ označena uzoračka korelaciona matrica.

Primer 7.10 Koristeći podatke iz Primera 7.8 testirati hipotezu $H_0:\boldsymbol{\rho}=\mathbf{I}$.

Izračunata uzoračka korelaciona matrica je

$$\mathbf{R}=\begin{bmatrix} 1 & 0.3873 & -0.1118 \\ 0.3873 & 1 & 0.5774 \\ -0.1118 & 0.5774 & 1 \end{bmatrix},$$

njeni karakteristični koreni: $\hat{\lambda}_1=1.647$, $\hat{\lambda}_2=1.103$ i $\hat{\lambda}_3=0.25$, a determinanta $|\mathbf{R}|=0.4542$. Izračunata vrednost statistike testa je

$$-\left[100-1-\frac{1}{6}(2\cdot 3+5)\right]\ln(0.4542)=76.6928.$$

Kako je kritična vrednost testa $\chi^2_{3;0.05}=7.81$ manja od izračunate, na nivou značajnosti od 5% odbacujemo nultu hipotezu o nezavisnosti originalnih promenljivih ili, drugim rečima, karakteristični koreni korelacione matrice nisu međusobom jednaki. ■

Praktičan postupak primene navedenog testa za testiranje jednakosti poslednja $(p-k)$ karakteristična korena kovarijacione matrice je sledeći. Prvo se testira hipoteza da su svi karakteristični koreni jednaki međusobom $(k=0)$. Ovo je poznata hipoteza sferičnosti. Ako se odbaci ova hipoteza, postavlja se nova prema kojoj su svi karakteristični koreni, osim prvog, međusobom jednaki $(k=1)$. Ako se odbaci ova

hipoteza, postupak testiranja nastavljamo, ali sada nulta hipoteza sugeriše da su svi karakteristični koreni, osim prva dva, međusobom jednaki ($k = 2$). Odbacivanjem ove nulte hipoteze nastavljamo postupak testiranja sve dok se ne prihvati hipoteza o jednakosti poslednja $(p - k)$ karakteristična korena.

7.3.3 Izbor broja glavnih komponenata

Jedan od ciljeva analize glavnih komponenata jeste redukcija početnog skupa podataka. Umesto velikog broja promenljivih u daljoj analizi koristimo manji broj glavnih komponenata koje u najvećoj mogućoj meri treba da obuhvate varijansu polaznog skupa podataka. Odgovor na pitanje o broju glavnih komponenata koje ćemo zadržati u analizi možemo potražiti u formalnom postupku testiranja značajnosti poslednjih karakterističnih korena. Ukazali smo na manjkavost tog pristupa problemu izbora koja se ogleda u zadržavanju suviše velikog broja glavnih komponenata. Stoga ćemo izložiti i druge, manje formalne postupke koji su našli svoje mesto u brojnim primenama ovog metoda multivarijacione analize.

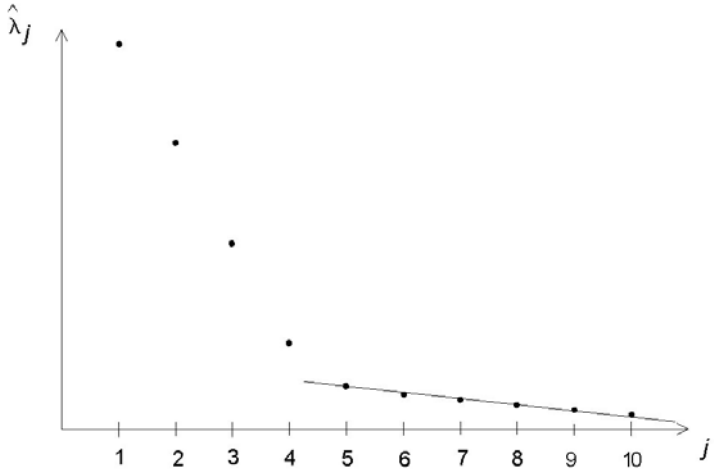
Prvi pristup polazi od apriornog fiksiranja kumulativne proporcije ukupne varijanse "objašnjene" izdvojenim skupom glavnih komponenata. U zavisnosti od proučavanog problema izabere se kumulativna proporcija od na primer 80% ili 90% ukupne varijanse, pa se broj zadržanih glavnih komponenti povećava sve dok se ne postigne ova, unapred zadata, granična vrednost kriterijuma. Očita je subjektivnost ovog načina određivanja broja glavnih komponenti jer se on određuje na bazi proizvoljno fiksirane vrednosti kriterijuma kumulativne proporcije "objašnjene" varijanse.

Drugi pristup sugeriše zadržavanje onih glavnih komponenata čija varijansa (λ_j) je veća od prosečne vrednosti $\bar{\lambda} = \sum_{j=1}^{p} \lambda_j / p$. Ako umesto kovarijacione koristimo korelacionu matricu, tada je prosečna vrednost varijanse jednaka jedinici, što znači da naš kriterijum glasi: zadržati one glavne komponente kod kojih je varijansa (karakteristični koren) veći od jedinice. Ovaj kriterijum je u širokoj upotrebi u faktorskoj analizi u fazi određivanja broja faktora i poznat je pod nazivom "kriterijum jediničnog korena" ili Kaiserov kriterijum (Kaiser (1958)).

Prema trećem pristupu u kriterijumu izbora koristi se geometrijska sredina. Naime, generalizovana varijansa (prva definicija) jednaka je proizvodu karakterističnih korena, tj. $\prod_{j=1}^{p} \lambda_j$. Ako dobijenu vrednost dignemo na stepen $1/p$ dobićemo geometrijsku sredinu karakterističnih korena. Dakle, prosečna generalizovana varijansa data je geometrijskom sredinom karakterističnih korena, pa prema ovom pristupu zadržavamo one glavne komponente čiji karakteristični koren je veći od geometrijske sredine svih karakterističnih korena.

Poslednji pristup zasniva se na grafičkom prikazu vrednosti karakterističnih korena prema njihovom rednom broju. Ovaj dijagram naziva se "*scree* test", a predložio ga je Cattell (1966).

Prelom ili lakat na krivoj određuje se tako što se prisloni lenjir uz poslednje vrednosti karakterističnih korena proveravajući da li one leže na pravoj liniji. Broj glavnih komponenata određujemo tako što uočavamo tačku nakon koje spomenuta prava linija ima prelom, pri čemu se krećemo od većeg ka manjem rednom broju glavne komponente. Broj glavnih komponenata predstavlja upravo redni broj glavne komponente čija vrednost karakterističnog korena kao poslednja leži na pravoj liniji. Na Slici 7.2, na kojoj je prikazan hipotetski primer kada je $p=10$, to je slučaj sa petom glavnom komponentom, jer osim nje sve ostale glavne komponente (četvrta, treća, druga i prva) imaju karakteristične korene koji ne leže na spomenutoj pravoj. Znači da *scree* test sugeriše zadržavanje pet glavnih komponenata. *Scree* test nije od pomoći ukoliko na grafikonu nema očiglednog preloma ili ukoliko ih ima više od jednog.



Slika 7.2 *Scree test*

Primer 7.11 Na osnovu slučajnog uzorka iz petodimenzione populacije izračunata je uzoračka kovarijaciona matrica

$$\mathbf{S}=\begin{bmatrix} 40 & & & & \\ -1 & 18 & & & \\ 18 & 6 & 16 & & \\ 20 & 2 & 11 & 20 & \\ -3 & 2 & -2 & -3 & 6 \end{bmatrix}.$$

Odrediti karakteristične korene, a potom, korišćenjem različitih kriterijuma, izabrati broj glavnih komponenata za dalju analizu.

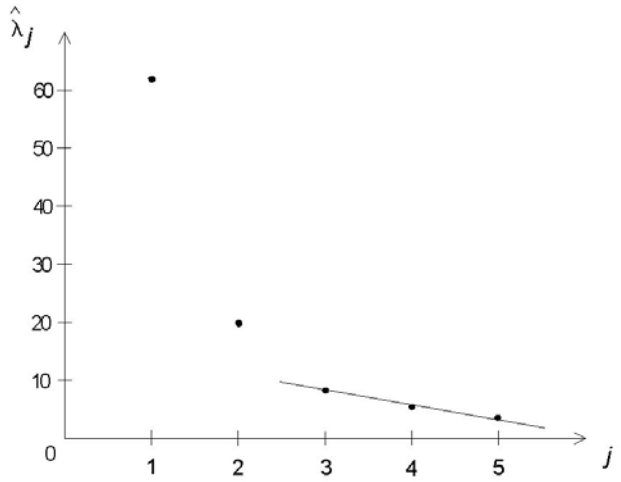
Karakteristični koreni uzoračke kovarijacione matrice su: $\hat{\lambda}_1=62.5455$, $\hat{\lambda}_2=20.5252$, $\hat{\lambda}_3=7.9973$, $\hat{\lambda}_4=5.6761$ i $\hat{\lambda}_5=3.2558$. Kumulativna proporcija "objašnjene" varijanse iznosi redom: za prvu glavnu komponentu 62.55%, za prve dve glavne komponente 83.07%, za prve tri

glavne komponente 91.07%, za prve četiri glavne komponente 96.74% i za sve glave komponente 100%. Ako unapred fiksiramo iznos ukupne "objašnjene" varijanse na 90%, onda prema ovom kriterijumu treba zadržati tri glavne komponente.

Aritmetička sredina karakterističnih korena iznosi $\bar{\lambda} = 20$, pa prema drugom kriterijumu zadržavamo one glavne komponente kod kojih je $\hat{\lambda}_j > \bar{\lambda}$. Kako je to slučaj kod prve dve komponente, prema ovom kriterijumu toliko ćemo ih i zadržati u daljoj analizi.

Kriterijum zasnovan na geometrijskoj sredini sugeriše zadržavanje dve glavne komponente, pošto je geometrijska sredina jednaka 11.3665, a upravo prva dva karakteristična korena su veća od ove srednje vrednosti.

Naposletku, korišćenje *scree* testa zasnivamo na sledećoj slici



Sa slike čitamo da za potrebe dalje analizi treba zadržati tri glavne komponente.

Korišćenjem različitih kriterijuma izbora broja glavnih komponenata sugerisano je izdvajanje dve (kriterijum aritmetičke i geometrijske sredine), odnosno tri (kriterijum kumulativne proporcije "objašnjene" varijanse i *scree* test) glavne komponente. Zbog sklonosti kriterijuma aritmetičke sredine da podcenjuje broj glavnih komponenata koje treba zadržati u daljoj analizi opredelili bismo se za tri glavne komponente. Ovaj izbor možemo usloviti i daljim korišćenjem izdvojenih glavnih komponenata u drugim metodama multivarijacione analize. ■

U opštem slučaju pokazalo se da kada je $p \geq 20$, Kaiserov kriterijum je restriktivan u tom smislu što uključuje suviše mali broj glavnih komponenti. Nasuprot njemu, *scree* test zadržava veliki broj glavnih komponenti u daljoj analizi. Zato se sugeriše nalaženje kompromisnog broja glavnih komponenata među brojevima koji su određeni primenom navedenih kriterijuma.

7.4 PRIMENA GLAVNIH KOMPONENATA

Dobijene glavne komponente najčešće predstavljaju tek polazni podatak za druge metode multivarijacione analize. U istraživanju u kome broj promenljivih prevazilazi broj opservacija, problem singulariteta izbija u prvi plan, pa je nužna redukcija dimenzije skupa promenljivih koju obavljamo korišćenjem glavnih komponenata. Tako na primer, u diskriminacionoj analizi umesto originalnog skupa podataka koristimo izvestan broj njihovih glavnih komponenata (sa visokom varijansom) u cilju formiranja diskriminacionih funkcija za razdvajanje grupa. Ukoliko prve dve glavne komponente imaju visok udeo u varijansi originalnog skupa, mogu biti korišćene u dvodimenzionom grafičkom prikazu podataka za ilustracija kvaliteta međusobnog razdvajanja grupa. Ozbiljan problem koji se pri korišćenju glavnih komponenata u diskriminacionoj analizi javlja ogleda se u tome što prve glavne komponente ne moraju biti dobri diskriminatori između grupa. One će to biti samo ako su varijacije unutar i između grupa duž istoga pravca. U opštem slučaju nema garancije da će razdvajanje između grupa biti u pravcu glavne komponente sa visokom varijansom. Tako se izostavljanjem glavnih komponenata sa niskom varijansom može odbaciti važna informacija koja se tiče varijacije između grupa.

U kanoničkoj korelacionoj analizi odredimo glavne komponente za svaki skup originalnih promenljivih zasebno, a potom se analiza obavi na osnovu dobijenih glavnih komponenata dveju grupa. Ako se zadrže sve glavne komponente rezultati analize kanoničke korelacije lakše se mogu interpretirati, a i sama teorijska osnova ove analize postaje razumljivija. Međutim, ukoliko u analizi zadržimo nekoliko glavnih komponenata javlja se problem slične prirode onom kod diskriminacione analize. Linearna kombinacija jednog skupa promenljivih koje su visoko korelisane sa linearnim kombinacijama drugog skupa ne moraju nužno biti sadržane u prvih nekoliko glavnih komponenata tog originalnog skupa promenljivih. Činjenica da linearna kombinacija promenljivih prvog skupa ima nisku varijansu ne sprečava ovu da ima visok koeficijent korelacije sa nekom linearnom kombinacijom drugog skupa promenljivih.

O primeni analize glavnih komponenata u ovim i drugim metodama multivarijacione analize opširnije kod Jolliffe (1986, Glava 7-9). U ovom poglavlju ukratko samo o korišćenju glavnih komponenata u regresionoj analizi, te otkrivanju nestandardnih opservacija i o robustnoj analizi glavnih komponenata.

7.4.1 Primena u regresionoj analizi

Pojava visoke linearne zavisnosti između regresora u višestrukom regresionom modelu naziva se multikolinearnost. Poznati rezultat iz regresione analize (vidi npr. Jovičić (1981, ss. 149-155)) sugeriše da su parametri regresionog modela neprecizno ocenjeni kada je problem multikolinearnosti znatno izražen. Jedno od rešenja ovog problema zasnovano je na korišćenju metode glavnih komponenata u regresionoj analizi. Naime, na osnovu skupa nezavisnih promenljivih formiraju se glavne komponente, a

zatim se obavlja regresiona analiza korišćenjem glavnih komponenata kao regresora. Ukoliko se u regresionom modelu zadrže sve glavne komponente, tada problem multikolinearnosti nije umanjen u odnosu na model u kome smo koristili originalne regresore. Međutim, osnovna ideja primene glavnih komponenata u ovom kontekstu nije u korišćenju svih glavnih komponenata nego manjeg broja njih. Analiza rezultata regresione analize u modelu sa glavnim komponentama koje imaju smislenu interpretaciju, možemo dati na osnovu ocenjenog modela. U suprotnom, ocenjeni ćemo model sa glavnim komponentama, koristeći vezu koja se može uspostaviti između originalnih promenljivih i glavnih komponenata, izraziti preko originalnih promenljivih.

Korišćenje manjeg broja glavnih komponenata od broja originalnih regresora u regresionom modelu zahteva da se prethodno izvrši izbor glavnih komponenata koje ćemo koristi u analizi. Za razliku od prethodnih slučajeva kada smo izbor glavnih komponenti vršili prema visini varijanse (birajući one sa navećom vrednošću varijanse) ovde je izbor diktiran visinom korelacije glavnih komponenata i zavisne promenljive. U regresioni model uključujemo one glavne komponente koje imaju najveću korelaciju sa zavisnom promenljivom, zato što je cilj regresione analize da objasni varijacije zavisne promenljive. Pri izboru glavnih komponenata u multivarijacionoj regresiji koristimo isti kriterijum kao i u višestrukom regresionom modelu, samo što sada ispitujemo korelaciju glavnih komponenata sa svakom zavisnom promenljivom.

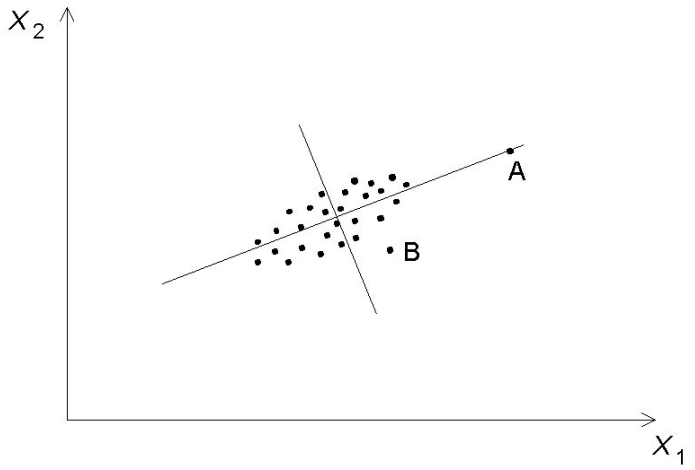
Nasuprot ocena regresionog modela sa originalnim promenljivama dobijenih metodom običnih najmanjih kvadrata, ocene modela u koji smo uključili manji broj glavnih komponenata su pristrasne ocene. Međutim, korišćenjem regresije glavnih komponenata očekujemo da ćemo znatno preciznije oceniti parametre modela, po cenu da te ocene budu u izvesnom stepenu pristrasne.

7.4.2 Identifikacija nestandardnih opservacija

Nestandardne opservacije (eng. *outlier*) nemaju preciznu definiciju u statistici. Pod njima se podrazumevaju one opservacije koje su na neki način nekonzistentne sa preostalim delom podataka. Statistički postupci utvrđivanja prisustva ekstremnih vrednosti u jednodimenzionom slučaju mogu poslužiti za identifikaciju višedimenzione nestandardne opservacije. Umesto originalnih promenljivih u tim postupcima koristimo glavnu komponentu (kao jednodimenzionu promenljivu). Na primer, poređenjem distribucije frekvencije standardizovanih skorova glavnih komponenti i standardizovanog normalnog rasporeda otkrivamo ekstremne vrednosti na levom i desnom kraju rasporeda.

Dva su karakteristična primera višedimenzione nestandardne opservacije. Njih smo prikazali kao tačke A i B na dvodimenzionom dijagramu rasturanja na Slici 7.3. Tačkom A predstavljena je nestandardna opservacija čije prisustvo možemo utvrditi na osnovu distribucija frekvencija originalnih promenljivih X_1 i X_2. Odgovarajuće koordinate tačke A izvan su intervala varijacija ostalih vrednosti kako prve tako i druge originalne promenljive. Međutim, tačka A je saglasna sa korelacionom strukturom ovog skupa dvodimenzionih podataka. U slučaju tačke B situacija je obratna.

Jednodimenzione distribucije frekvencija originalnih promenljivih sugerišu da tačka B ne predstavlja nestandardnu opservaciju, ali je zato nesaglasna sa korelacionom strukturom podataka, te zato predstavlja dvodimenzionu nestandardnu opservaciju. Na Slici 7.3 ucrtane su i dve prave koje prolaze kroz roj tačaka. One su tako postavljene da se poklapaju sa osama zamišljene elipse koja se može opisati oko datog roja tačaka. Od ranije poznati rezultat prvu glavnu komponentu identifikuje sa glavnom osom elipse, a drugu glavnu komponentu sa sporednom osom elipse.



Slika 7.3 Dva primera nestandardnih opservacija

U ovom dvodimenzionom primeru koji koristimo kao ilustraciju, a važi i u opštem slučaju, prve glavne komponente imaće visoku vrednost varijanse i težiće da budu u bliskoj vezi sa promenljivama koje takođe imaju visoku varijansu i međusobom su visoko korelisane. Opservacije koje su nestandardne sa stanovišta tih prvih glavnih komponenti mogu biti otkrivene i na osnovu distribucije frekvencija originalnih promenljivih, kao što je to slučaj sa tačkom A. S druge strane, poslednje glavne komponente imaju najmanju varijansu i osetljive su na one opservacije koje su nesaglasne sa korelacionom strukturom podataka. Ove nestandardne opservacije ne mogu se otkriti na osnovu rasporeda pojedinih originalnih promenljivih. Na Slici 7.3 to je slučaj sa tačkom B. Dakle, u našem primeru na Slici 7.3, prva glavna komponenta (glavna osa elipse) otkriva nestandardne opservacije tipa A, a poslednja, u dvodimenzionom slučaju druga glavna komponenta (sporedna osa elipse), tipa B. Na osnovu rečenog, očigledno je da u primeni metoda glavnih komponenata, čak i u slučaju da smo zadržali samo nekoliko prvih glavnih komponenata, a ostale odbacili, postoji interes za korišćenjem poslednjih glavnih komponenata radi identifikacije nestandardnih opservacija.

Pored upotrebe grafičkog prikaza glavnih komponenti u identifikaciji nestandardnih opservacija koristi se i Mahalanobisovo odstojanje. Skorovi poslednjih glavnih komponenti koje su odbačene iz dalje analize služi za računanje Mahalanobisovog odstojanja svake opservacije od sredine uzorka. Visoka vrednost ovog odstojanja sugeriše da je odnosna opservacija nestandardna s obzirom na korelacionu strukturu

podataka. Ukoliko želimo da ispitamo stepen uticaja neke opservacije (ne mora biti nestandardna) na rezultat analize, ponovićemo postupak određivanja glavnih komponenti na redukovanom skupu podataka iz koga smo eliminisali tu opservaciju. Poredeći rezultate u ova dva slučaja možemo utvrditi uticaj izostavljene opservacije. Za formalan postupak testiranja hipoteze o prisustvu nestandardne opservacije zasnovane na Mahalanobisovom odstojanju kao i ostale postupke njihove identifikacije videti kod Hawkinsa (1980), Jolliffea (1986), Mardia, Kent i Bibbya (1979) i Jobsona (1992).

7.4.3 Robustna analiza glavnih komponenata

U bliskoj vezi sa identifikacijom nestandardnih opservacija jeste robustna analiza glavnih komponenata. Naime, ukoliko je nestandardna opservacija određenog tipa prisutna u podacima, tada dolazi do povećanja vrednosti varijansi i kovarijansi, pa se dešava da uzoračka kovarijaciona ili korelaciona matrica nisu pozitivno definitne. Ovo ima direktne posledice u pogledu određivanja glavnih komponenata, posebno prvih nekoliko. Da bi se analiza glavnih komponenti učinila manje osetljivom na prisustvo nestandardnih opservacija preporučuje se korišćenje robustnih ocena kovarijacione i korelacione matrice (videti kod Devlina i dr. (1981)). Postupkom iterativnog prilagođavanja koji je u osnovu ovih ocena, apsolutne vrednosti vandijagonalnih elemenata kovarijacione matrice, "podešavaju" se u odnosu prema dijagonalnim ne bi li se postigla pozitivna definitnost matrice. Flury (1988, s. 162) sugeriše da se, ako nas već interesuju glavne komponente, princip robustnosti primeni direktno na njih, a ne da se čini međukorak time što će se robustno oceniti kovarijaciona ili korelaciona matrica koje kasnije služe određivanju glavne komponente. Međutim, ovaj pristup nije u punoj meri teorijski ispitan i empirijski proveren da bi se našao u široj upotrebi.

Drugi pristup zasnovan je na korišćenju rangova umesto originalnih opservacija promenljivih i formiranju matrice čiji su elementi koeficijenti korelacije ranga. Analiza glavnih komponenata obavlja se potom na tako definisanoj korelacionoj matrici. Ovaj pristup preporučuje se i u slučaju kada su originalne promenljive merene na različitim mernim skalama. Pregled različitih postupaka robustnih ocena glavnih komponenata dat je kod Jolliffea (1986, ss. 195-197).

8 FAKTORSKA ANALIZA

Metod multivarijacione analize koji se koristi za opis međusobne zavisnosti velikog broja promenljivih korišćenjem manjeg broja osnovnih, ali neopažljivih slučajnih promenljivih poznatih kao *faktori*, naziva se *faktorska analiza*. Na prvi pogled faktorska analiza je podudarna sa analizom glavnih komponenata. Oba metoda multivarijacione analize koristi se u svrhe redukcije skupa podataka, aproksimirajući kovarijacionu ili korelacionu matricu originalnih promenljivih. Međutim, dok je analiza glavnih komponenata koncentrisana na dijagonalne elemente kovarijacione matrice (varijanse), dotle nas u faktorskoj analizi interesuju vandijagonalni elementi (kovarijanse). Takođe, za razliku od analize glavnih komponenata, faktorska analiza pretpostavlja postojanje teorijskog modela kojim se uspostavlja relacija između opservacija dimenzione promenljive i manjeg broja zajedničkih faktora. Upravo zato, nakon uvoda definišemo osnovni model faktorske analize i etape u postupku njegovog statističkog ocenjivanja. Zatim se bavimo tzv. rotacijom faktora, odnosno postupkom kojim transformišemo dobijene ocene koeficijenata u cilju lakše interpretacije ocenjenog modela faktorske analize. Pored ocena parametara faktorskog modela interes nam je da odredimo i vrednost faktora za svaku jedinicu posmatranja, tj. *faktorske skorove*, kako zbog lakše interpretacije rezultata analize, tako i zbog korišćenja vrednosti faktora u daljoj analizi. Dvema metodama određivanja faktorskih skorova: metod ponderisanih najmanjih kvadrata i regresioni metod, bavimo se u poslednjem poglavlju.

8.1 UVOD

Osnovna ideja faktorske analize, onako kako je izložena krajem prošlog stoleća od strane Francis Galtona, Charles Spearmana i drugih, proizašla je uglavnom iz napora psihologa da bolje razumeju i istraže kompleksan fenomen "inteligencije". U tim istraživanjima korišćeni su brojni testovi inteligencije sa velikim brojem pitanja. Ona su trebala da istaknu različite dimenzije opšte sposobnosti ispitanika, kao što su verbalna sposobnost, matematička sposobnost, memorija itd. Faktorska analiza je razvijena za

analizu rezultata tih testova. Njen zadatak je bio da utvrdi da li je "inteligencija" komponovana od jednog bazičnog, opšteg faktora ili od nekoliko zajedničkih faktora koji mere osobine kao što su na primer, "verbalna", "matematička" i druge sposobnosti. Ili, drugi primer iz sociologije, gde prikupljamo informacije u širem krugu ljudi o njihovom poslu, vrsti obrazovanja, visini dohotka, o tome da li poseduju vlastitu kuću ili ne, itd. Postavlja se pitanje da li je koncept društvenih klasa višedimenzioni ili je moguće konstruisati jedan "indeks" klase na osnovu podataka, drugim rečima, interesuje nas da li postoji jedinstven osnovni faktor.

Utemeljenje faktorske analize dugujemo Charles Spearmanu. U svom radu iz 1904. godine, izučavajući korelaciju između rezultata različitih testova inteligencije, ukazao je na mogućnost njihovog iskazivanja jednostavnim modelom. Tako na primer, u istraživanju pripremljenosti dece za školu, dobijena je korelaciona matrica rezultata testova iz Klasike (X_1), Francuskog (X_2), Engleskog (X_3) i Matematike (X_4):

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

Spearman uočava proporcionalnost ma koja dva reda ili kolone u ovoj matrici, ako se zanemare elementi na glavnoj dijagonali. Tako za elemente prvog i poslednjeg reda imamo količnike

$$\frac{0.83}{0.67} \cong \frac{0.78}{0.64} \cong 1.2$$

Stoga Spearman predlaže redukciju dimenzije problema sa $p = 4$ na $p = 1$, tako što ćemo rezultate svih testova (X_i) iskazati u obliku sledećeg modela

$$X_i = \beta_i F + \varepsilon_i, \; i = 1, 2, 3, 4 \tag{8.1}$$

U ovom modelu F je zajednički faktor, β_i su koeficijenti koje nazivamo *faktorska opterećenja*, a ε_i su slučajne greške ili, kako se nazivaju u faktorskoj analizi, *specifični faktori*. Uz izvesne pretpostavke vezane za model, koje ćemo eksplicitno iskazati u narednom poglavlju, dobijamo da su elementi u dva reda korelacione matrice proporcionalni međusobom. Ovo znači da predloženi model dobro objašnjava korelacionu strukturu podataka. Na osnovu svojih istraživanja Spearman formuliše dvofaktorsku teoriju testova inteligencije prema kojoj se rezultat svakog testa može dekomponovati na dva dela, prvi, koji je zajednički za sve testove (F) i drugi koji je specifičan za svaki test (ε_i). Zajednički faktor može se interpretirati kao "opšta sposobnost" ili "inteligencija".

Kasnija istraživanja na ovom polju proširila su prvobitni model u tom smislu što je uvedeno nekoliko zajedničkih faktora (pretpostavlja se dakle da postoje posebne vrste inteligencija, odnosno sposobnosti). Takođe je specifičan faktor razložen na dva dela od kojih prvi pokazuje u kom stepenu se nečija individualna sposobnost u na primer, Matematici razlikuje od opšte sposobnosti, a drugi deo reprezentuje činjenicu da su rezultati testiranja samo aproksimacija sposobnosti individue u konkretnoj oblasti.

Na osnovu rečenog možemo konstatovati da faktorska analiza ima dvojaku ulogu u analizi višedimenzionih podataka. Ona, kao i analiza glavnih komponenata, služi za redukciju originalnog skupa podataka. U takvoj vrsti analize nemamo na umu neku određenu teorijsku hipotezu, nego faktorsku analizu koristimo da bismo identifikovali zajedničku strukturu koja je generisala dobijeni skup korelisanih originalnih promenljivih. To je tzv. istraživačka primena faktorske analize, odnosno kažemo da se ona koristi u deskriptivne svrhe. Druga primena faktorske analize tiče se onih istraživanja u kojima raspolažemo apriornom teorijskom informacijom o zajedničkoj strukturi, a faktorsku analizu koristimo da bismo testirali hipoteze o broju zajedničkih faktora koji su latentno prisutni u skupu podataka ili o prirodi tih faktora. Drugim rečima, faktorska analiza, u svojoj drugoj primeni, koristi se u svrhe testiranja kojim bi se potvrdila ili negirala hipotetska struktura podataka.

Mnogi autori tretiraju analizu glavnih komponenata kao specifičan slučaj ili prvu fazu u faktorskoj analizi, pa se ide čak dotle da se koristi i naziv faktorska za analizu koja po svojoj suštini pripada analizi glavnih komponenata. Opravdanje za ovakav tretman dve metode multivarijacione analize nalazimo u sličnosti rezultata koje dobijamo njihovom primenom. Dodatnim ukazivanjem na razlike između analize glavnih komponenata i faktorske analize istićemo specifičnosti ove potonje.

Analiza glavnih komponenata izučava ukupan varijabilitet skupa podataka. Nasuprot njoj, faktorska analiza polazi od razlaganja promenljive na dva dela: zajednički i specifičan deo. Zajednički deo je onaj deo varijacija promenljive koji ona deli sa ostalim promenljivama, dok je specifičan deo onaj deo varijacija promenljive koji je poseban za tu promenljivu. Faktorska analiza izučava deo varijacija koji je zajednički za sve promenljive, a analiza glavnih komponenata ukupan varijabilitet.

U oba metoda multivarijacione analize javljaju se opažljive i neopažljive ili latentne promenljive. Međutim, dok se u analizi glavnih komponenata na osnovu linearne kombinacije opažljivih promenljivih (originalan skup podataka) formiraju neopažljive promenljive (glavne komponente), dotle prema faktorskom modelu na osnovu latentnih promenljivih (faktora) izražavamo originalne promenljive. Primetimo da u opštem slučaju izražavanje neopažljivih glavnih komponenata kao funkcije originalnih promenljivih, nije ekvivalentno izražavanju tih originalnih promenljivih kao funkcije neopažljivih faktora. U prvom slučaju originalne promenljive formiraju latentne komponente i to egzaktno, bez greške. U drugom slučaju originalne promenljive predstavljaju samo manifestaciju, tj. pojavni oblik latentnih faktora i pri tome su merene sa greškom. Primer koji smo koristili u izlaganju sugeriše da se "opšta sposobnost" ili "inteligencija" može smatrati faktorom koji je našao svoga odraza u rezultatima testova

inteligencije, pa su prema tome originalne promenljive odraz latentnog faktora. Nasuprot tome, kada na primer u demografiji izučavamo fenomen "promene populacije", originalne promenljive koje uslovljavaju ove promene, kao što su natalitet, mortalitet i migracije, tretiramo kao promenljive koje formiraju latentnu komponentu.

8.2 MODEL FAKTORSKE ANALIZE I OCENJIVANJE

8.2.1 Model faktorske analize

Pretpostavimo, kao i u analizi glavnih komponenata, da je $\mathbf{X}$ $p-$dimenzioni vektor opažljivih promenljivih sa sredinom $\boldsymbol{\mu}$ i kovarijacionom matricom $\boldsymbol{\Sigma}$. Model faktorske analize pretpostavlja da se $\mathbf{X}$, vektor opažljivih promenljivih, može izraziti preko skupa od m neopažljivih promenljivih, koje nazivamo *zajednički faktori*, u oznaci $F_1, F_2, ..., F_m$, gde je $m \ll p$ i p specifičnih, ali neopažljivih faktora, u oznaci $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$. Model u razvijenom obliku dat je sledećim jednačinama

$$\begin{aligned}
(X_1 - \mu_1) &= \beta_{11}F_1 + \beta_{12}F_2 + ... + \beta_{1m}F_m + \varepsilon_1 \\
(X_2 - \mu_2) &= \beta_{21}F_1 + \beta_{22}F_2 + ... + \beta_{2m}F_m + \varepsilon_2 \\
&\cdots \\
(X_p - \mu_p) &= \beta_{p1}F_1 + \beta_{p2}F_2 + ... + \beta_{pm}F_m + \varepsilon_p
\end{aligned} \tag{8.2}$$

ili ekvivalentno u matričnoj notaciji

$$\underset{(p\times 1)}{\mathbf{X}-\boldsymbol{\mu}} = \underset{(p\times m)}{\mathbf{B}}\,\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}} \tag{8.3}$$

gde je

$$\mathbf{X}-\boldsymbol{\mu} = \begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \vdots \\ X_p-\mu_p \end{bmatrix},\ \mathbf{F} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix},\ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix},\ \mathbf{B} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pm} \end{bmatrix}.$$

Elementi matrice $\mathbf{B}$, tj. β_{ij} nazivaju se *faktorska opterećenja* $i-$te promenljive na $j-$ti faktor, a sama matrica naziva se *matrica faktorskih opterećenja*. Na prvi pogled model faktorske analize liči na model višestruke regresije. Međutim, ovde p odstupanja $(X_1-\mu_1)$, …, $(X_p-\mu_p)$ izražavamo preko $m+p$ slučajnih promenljivih $F_1, F_2, ..., F_m$ i $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$ koje su neopažljive, za razliku od regresionog modela gde su nezavisne promenljive opažljive.

Nametanjem dodatnih ograničenja na elemente modela obezbeđujemo uslove da na osnovu opservacija promenljivih $X_1, X_2, ..., X_p$ empirijski proverimo model faktorske analize. Ta ograničenja tiču se vektora zajedničkih faktora. Naime, pretpostavljamo da je

$$E(\mathbf{F}) = \mathbf{0}\,,\ Cov(\mathbf{F}) = E(\mathbf{F}\mathbf{F}') = \mathbf{\Phi} \tag{8.4}$$

Izlaganje u ovoj glavi baziramo na specijalnom slučaju modela faktorske analize kada je $\mathbf{\Phi} = \mathbf{I}$ (radi se o tzv. ortogonalnom modelu kod koga su faktori nezavisni međusobom). Što se specifičnih faktora tiče, njihova očekivana vrednost jednaka je nuli, a kovarijaciona matrica je dijagonalna

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}\,,\ Cov(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \Psi = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \tag{8.5}$$

Takođe pretpostavljamo da su zajednički faktori nezavisni od specifičnih, tj. da je

$$Cov(\boldsymbol{\varepsilon}, \mathbf{F}) = E(\boldsymbol{\varepsilon}\mathbf{F}') = \mathbf{0} \tag{8.6}$$

Vezu između odstupanja opažljivih promenljivih od njihove sredine i neopažljivih faktora, zajedno sa navedenim pretpostavkama nazivamo *model faktorske analize*. Ovaj model omogućava sledeće razlaganje kovarijacione matrice $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{\Psi} \tag{8.7}$$

Pokažimo ovaj rezultat polazeći od

$$(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' = (\mathbf{B}\mathbf{F} + \boldsymbol{\varepsilon})(\mathbf{B}\mathbf{F} + \boldsymbol{\varepsilon})' = \mathbf{B}\mathbf{F}(\mathbf{B}\mathbf{F})' + \boldsymbol{\varepsilon}(\mathbf{B}\mathbf{F})' + \mathbf{B}\mathbf{F}\boldsymbol{\varepsilon}' + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \tag{8.8}$$

a zatim odredimo očekivanu vrednost dobijenog izraza

$$\mathbf{\Sigma} = E\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'\right] = \mathbf{B}\,E(\mathbf{F}\mathbf{F}')\mathbf{B}' + E(\boldsymbol{\varepsilon}\mathbf{F}')\mathbf{B}' + \mathbf{B}\,E(\mathbf{F}\boldsymbol{\varepsilon}') + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \tag{8.9}$$

Na osnovu pretpostavki modela sledi da je $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{\Psi}$. U opštijem slučaju, uz korelisanost faktora ($\mathbf{\Phi} \neq \mathbf{I}$), razlaganje kovarijacione matrice glasi: $\mathbf{\Sigma} = \mathbf{B}\mathbf{\Phi}\mathbf{B}' + \mathbf{\Psi}$.

Kako je $Cov(\mathbf{X}, \mathbf{F}) = E(\mathbf{B}\mathbf{F} + \boldsymbol{\varepsilon})\mathbf{F}' = \mathbf{B}\,E(\mathbf{F}\mathbf{F}') + E(\boldsymbol{\varepsilon}\mathbf{F}') = \mathbf{B}$, znači da su elementi matrice faktorskih opterećenja kovarijanse između originalnih promenljivih i faktora. Korelacionu matricu promenljivih X i faktora F nazivamo *matrica faktorske strukture*. Ove dve matrice u opštem slučaju razlikuju se međusobom. Međutim, one su identične aku su faktori nezavisni jedan od drugog (slučaj ortogonalnog modela faktorske analize). Obe matrice, faktorskih opterećenja i faktorske strukture, koristimo prilikom interpretacije dobijenih faktora u istraživačkoj primeni faktorske analize.

Na osnovu razlaganja kovarijacione matrice kod ortogonalnog modela faktorske analize imamo da je varijansa $i-$te promenljive

$$Var(X_i)=\sigma_{ii}=\beta_{i1}^2+\beta_{i2}^2+...+\beta_{im}^2+\psi_i$$

$$=\sum_{j=1}^{m}\beta_{ij}^2+\psi_i,\ i=1,2,...,p \tag{8.10}$$

Znači da je varijansa $i-$te originalne promenljive podeljena na dva dela. Prvi deo $\sum_{j=1}^{m}\beta_{ij}^2$ je varijansa objašnjena zajedničkim faktorima i nazivamo ga *zajednička varijansa* ili *komunalitet* (u oznaci h_i^2), a drugi deo nazivamo *specifična varijansa* (u oznaci ψ_i). Kovarijansa između $i-$te i $j-$te originalne promenljive data je vandijagonalnim elementima matrice $\mathbf{BB}'$, tj. sa $\sum_{j=1}^{m}\beta_{ij}\beta_{kj}$ $(i\neq j)$.

Primer 8.1 Data je kovarijaciona matrica

$$\mathbf{\Sigma}=\begin{bmatrix}15 & 8 & 3 & 16\\ 8 & 10 & 1 & 10\\ 3 & 1 & 16 & 2\\ 16 & 10 & 2 & 24\end{bmatrix}.$$

Pokazati da važi jednakost $\mathbf{\Sigma}=\mathbf{BB}'+\mathbf{\Psi}$, odn.

$$\begin{bmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14}\\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24}\\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34}\\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44}\end{bmatrix}=\begin{bmatrix}\beta_{11} & \beta_{12}\\ \beta_{21} & \beta_{22}\\ \beta_{31} & \beta_{32}\\ \beta_{41} & \beta_{42}\end{bmatrix}\begin{bmatrix}\beta_{11} & \beta_{21} & \beta_{31} & \beta_{41}\\ \beta_{12} & \beta_{22} & \beta_{32} & \beta_{42}\end{bmatrix}+\begin{bmatrix}\psi_1 & 0 & 0 & 0\\ 0 & \psi_2 & 0 & 0\\ 0 & 0 & \psi_3 & 0\\ 0 & 0 & 0 & \psi_4\end{bmatrix}$$

gde je

$$\mathbf{B}=\begin{bmatrix}3 & 2\\ 2 & 1\\ -1 & 3\\ 4 & 2\end{bmatrix}\ \text{i}\ \mathbf{\Psi}=\begin{bmatrix}2 & 0 & 0 & 0\\ 0 & 5 & 0 & 0\\ 0 & 0 & 6 & 0\\ 0 & 0 & 0 & 4\end{bmatrix}.$$

Uz pretpostavku postojanja dva zajednička faktora $(m=2)$ na osnovu ortogonalnog faktorskog modela odrediti komunalitete sve četiri originalne promenljive i razložiti njihove varijanse.

Komunalitet prve promenljive je

$$h_1^2=\beta_{11}^2+\beta_{12}^2=3^2+2^2=13$$

a razlaganje varijanse σ_{11} na komunalitet i specifičnu varijansu je

$$\sigma_{11} = h_1^2 + \psi_1 \text{ ili } 15 = 13 + 2$$

Komunalitet druge, treće i četvrte promenljive je $h_2^2 = 5$, $h_3^2 = 10$ i $h_4^2 = 20$ pa je razlaganje odnosnih varijansi

$$\sigma_{22} = h_2^2 + \psi_2 \text{ ili } 10 = 5 + 5$$

$$\sigma_{33} = h_3^2 + \psi_3 \text{ ili } 16 = 10 + 6$$

$$\sigma_{44} = h_4^2 + \psi_4 \text{ ili } 24 = 20 + 4 \quad \blacksquare$$

Korišćenjem korelacione umesto kovarijacione matrice imamo da se jedinična varijansa standardizovane promenljive sastoji iz dva dela

$$Var(X_i) = 1 = h_i^2 + \psi_i \tag{8.11}$$

u ovom slučaju h_i^2 predstavlja proporciju varijanse $i-$te promenljive objašnjene zajedničkim faktorima, a ψ_i proporciju varijanse $i-$te promenljive koja se pripisuje specifičnom faktoru. Najčešće se faktorska analiza obavlja na standardizovanim promenljivama, odnosno vršimo razlaganje korelacione matrice ($\boldsymbol{\rho} = \mathbf{BB}' + \boldsymbol{\Psi}$). Opravdanje za korišćenje standardizovanih promenljivih nalazimo u tome što model faktorske analize važi i za njih kao i za originalne promenljive. Kažemo da su rezultati faktorske analize nepromenljivi s obzirom na postupak standardizacije promenljivih. Međutim, pri ocenjivanju modela faktorske analize neke od metoda ocenjivanja osetljive su na postupak standardizacije promenljivih. Do matrice faktorskih opterećenja i matrice specifičnih varijansi standardizovanih promenljivih možemo doći na osnovu odnosnih matrica originalnih promenljivih (množeći $i-$ti red matrice $\mathbf{B}$ recipročnom vrednošću standardne devijacije $i-$te originalne promenljive kod matrice faktorskih opterećenja, a recipročnom vrednošću varijanse $i-$te originalne promenljive kod matrice specifičnih varijansi). Ova osobina faktorske analize predstavlja još jedan elemenat po kome se ona razlikuje od analize glavnih komponenata. Za ovu potonju smo konstatovali da daje različite rezultate ako se na originalne promenljive primeni postupak standardizacije.

Ukažimo na vrednost ukupne varijanse originalnog skupa promenljivih i njeno razlaganje. Generalizovana varijansa od $\mathbf{X}$ je

$$\text{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \sigma_{ii} = + \sum_{i=1}^{p} \sum_{j=1}^{m} \beta_{ij}^2 + \sum_{i=1}^{p} \psi_i \tag{8.12}$$

Ako sa h označimo ukupan komunalitet od $\mathbf{X}$, gde je

$$h = \sum_{i=1}^{p} h_i^2 = \sum_{i=1}^{p} \sum_{j=1}^{m} \beta_{ij}^2 \tag{8.13}$$

tada je

$$\text{tr}(\mathbf{\Sigma}) = h + \text{tr}(\mathbf{\Psi}) \tag{8.14}$$

Znači da je ukupna ili generalizovana varijansa od $\mathbf{X}$ jednaka zbiru dve komponente: ukupnog komunaliteta i ukupne varijanse specifičnih faktora. Formiranjem količnika

$$\frac{\sum_{i=1}^{p} \beta_{ij}^2}{h}, \quad j = 1, 2, ..., m \tag{8.15}$$

dobićemo proporciju ukupnog komunaliteta koja se može pripisati $j-$tom zajedničkom faktoru.

Primer 8.2 Na osnovu ortogonalnog modela faktorske analize iz Primera 8.1 odrediti ukupnu varijansu, ukupan komunalitet, ukupnu specifičnu varijansu i doprinose svakog faktora ukupnom komunalitetu.

Ukupna varijansa je

$$\text{tr}(\mathbf{\Sigma}) = \sigma_{11} + \sigma_{22} + \sigma_{33} + \sigma_{44} = 15 + 10 + 16 + 24 = 65$$

Ukupan komunalitet je

$$h = h_1^2 + h_2^2 + h_3^2 + h_4^2 = 13 + 5 + 10 + 20 = 48$$

Ukupna specifična varijansa je

$$\text{tr}(\mathbf{\Psi}) = \psi_1 + \psi_2 + \psi_3 + \psi_4 = 2 + 5 + 6 + 4 = 17$$

Doprinos prvog zajedničkog faktora ukupnom komunalitetu je

$$\frac{\beta_{11}^2 + \beta_{21}^2 + \beta_{31}^2 + \beta_{41}^2}{h} = \frac{30}{48} \text{ ili } 62.5\%$$

Doprinos drugog zajedničkog faktora ukupnom komunalitetu je

$$\frac{\beta_{12}^2 + \beta_{22}^2 + \beta_{32}^2 + \beta_{42}^2}{h} = \frac{18}{48} \text{ ili } 37.5\%$$

■

Konstatovali smo da model faktorske analize reprezentuje p promenljivih X sa ukupno $m + p$ zajedničkih faktora F i specifičnih faktora ε. Odgovarajuće razlaganje kovarijacione matrice sugeriše postojanje $p(p+1)/2$ jednačine, jedna za svaki element

od $\boldsymbol{\Sigma}$, u kojima figurišu elementi matrice $\mathbf{B}$, $\boldsymbol{\Phi}$ i $\boldsymbol{\Psi}$, odnosno ukupno $mp + m(m+1)/2 + p$ koeficijenata. Uslov identifikacije zahteva da je $p(p+1)/2 \geq mp + m(m+1)/2 + p$, jer tada na osnovu poznatih elemenata matrice $\boldsymbol{\Sigma}$ možemo odrediti elemente matrice $\mathbf{B}$, $\boldsymbol{\Phi}$ i $\boldsymbol{\Psi}$. Kod ortogonalnog modela faktorske analize uslov identifikacije svodi se na zahtev da bude $p(p+1)/2 \geq mp + p$ ili $(p-1)/2 \geq m$. Tako na primer, ako raspolažemo sa $p = 16$ originalnih promenljivih, maksimalan broj zajedničkih faktora koji možemo izdvojiti iznosi $m = 7$. Ova situacija poznata je kao problem strukture faktorske matrice. Korišćenjem faktorske analize u cilju redukcije dimenzije problema težimo dobijanju m smislenih zajedničkih faktora koji se mogu lako interpretirati pri čemu je $m \ll p$. Na žalost, uslov identifikacije ne garantuje postojanje razlaganja kovarijacione matrice ($\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}' + \boldsymbol{\Psi}$) ili njenu jednoznačnost. Postoje kovarijacione matrice koje se ne mogu na navedeni način razložiti, a da bude m znatno manje od p. Nejednoznačnost dobijenog razlaganja kovarijacione matrice u slučaju kada rešenje postoji i kada je $m > 1$ pokazaćemo primenom ortogonalne transformacije ili rotacije zajedničkih faktora.

Neka je $\mathbf{T}$ $(m \times m)$ ortogonalna matrica ($\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}$) kojom ćemo izvršiti ortogonalnu rotaciju zajedničkih faktora. Prema ortogonalnom modelu faktorske analize imamo

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{B}\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{B}\mathbf{T}\mathbf{T}'\mathbf{F} = \boldsymbol{\Gamma}\mathbf{F}^* + \boldsymbol{\varepsilon} \tag{8.16}$$

gde je

$$\boldsymbol{\Gamma} = \mathbf{B}\mathbf{T} \quad \text{i} \quad \mathbf{F}^* = \mathbf{T}'\mathbf{F} \tag{8.17}$$

Prema pretpostavkama modela imamo $E(\mathbf{F}^*) = \mathbf{T}'E(\mathbf{F}) = \mathbf{0}$ i $Cov(\mathbf{F}^*) = \mathbf{T}'Cov(\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{I}$. Znači da na osnovu opservacija originalnih promenljivih ne možemo razlučiti matricu faktorskih opterećenja $\mathbf{B}$ od matrice faktorskih opterećenja $\boldsymbol{\Gamma}$. Štaviše, obe matrice, u opštem slučaju različite međusobom, generišu istu kovarijacionu matricu $\boldsymbol{\Sigma}$, pošto je

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}' + \boldsymbol{\Psi} = \mathbf{B}\mathbf{T}\mathbf{T}'\mathbf{B}' + \boldsymbol{\Psi} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \boldsymbol{\Psi} \tag{8.18}$$

Problem nejednoznačnosti faktorskog rešenja predstavlja glavni problem u ocenjivanju modela faktorske analize, pored ostalih kao što su problem strukture, identifikacije i nepostojanja razlaganja kovarijacione matrice.

8.2.2 Metodi ocenjivanja

Na osnovu uzetog uzorka iz $p-$dimenzione populacije ocenjujemo[1] model faktorske analize, odnosno proveravamo da li važi razlaganje kovarijacione (ili korelacione) matrice koje je implicirano tim modelom. Među brojnim metodama ocenjivanja modela

[1] Neki autori (npr. Dillon i Goldstein (1984)) postupke ocenjivanja nazivaju metodama izdvajanja (ekstrakcije) faktora, dok ih drugi (npr. Harman (1970)) nazivaju metode rešavanja faktorskog modela, a dobijene ocene, faktorska rešenja.

faktorske analize izdvajamo dva najpoznatija: metod glavnih komponenata (glavnih faktora) i metod najveće verodostojnosti. Prvi metod ocenjivanja javlja se u dve varijante: 1) direktno korišćenje metode glavnih komponenata na kovarijacionu ili korelacionu matricu radi istovremenog ocenjivanja komunaliteta i matrice faktorskih opterećenja i 2) korelaciona matrica se modifikuje u duhu modela faktorske analize, a zatim se iterativno i odvojeno ocenjuju komunaliteti i faktorska opterećenja primenom metode glavnih komponenata na tu modifikovanu korelacionu matricu. Delimično opravdanje za poistovećivanje metode glavnih komponenata i metode faktorske analize nalazimo u korišćenju prve varijante ove prve metode u oceni modela faktorske analize.

Podsećamo da smo vektor glavnih komponenata $\mathbf{Y}$ definisali na osnovu vektora $\mathbf{X}$ kao $\mathbf{Y} = \mathbf{AX}$, gde su redovi matrice $\mathbf{A}$ normalizovani karakteristični vektori kovarijacione matrice $\mathbf{\Sigma}$ (pri čemu je $\mathbf{AA}' = \mathbf{I}$ i $Cov(\mathbf{Y}) = \mathbf{\Lambda}$, gde je $\mathbf{\Lambda}$ dijagonalna matrica čiji su elementi karakteristični koreni kovarijacione matrice $\mathbf{\Sigma}$). U modelu faktorske analize opažljive promenljive $\mathbf{X}$ izražavamo preko neopažljivih zajedničkih faktora F. Zato ćemo vektor $\mathbf{X}$ u analizi glavnih komponenata izraziti preko vektora $\mathbf{Y}$, tj. $\mathbf{X} = \mathbf{A}^{-1}\mathbf{Y} = \mathbf{A}'\mathbf{Y}$. Ako sada matricu $\mathbf{A}$ pomnožimo s leve strane sa $\mathbf{\Lambda}^{-1/2}$ i ako je $\mathbf{F} = \mathbf{\Lambda}^{-1/2}\mathbf{Y} = \mathbf{\Lambda}^{-1/2}\mathbf{AX}$, tada je $Cov(\mathbf{F}) = \mathbf{I}$ i $\mathbf{X} = \mathbf{A}^{-1}\mathbf{\Lambda}^{1/2}\mathbf{F}$. Razmotrimo model faktorske analize za vektor $\mathbf{X}$ iz koga su isključeni specifični faktori, tada je $\mathbf{X} = \mathbf{BF}$, gde je matrica $\mathbf{B}$ jednaka $\mathbf{A}^{-1}\mathbf{\Lambda}^{1/2}$ na osnovu gornje analize glavnih komponenata. Znači da je

$$\mathbf{B} = \mathbf{A}^{-1}\mathbf{\Lambda}^{1/2} = \left[\sqrt{\lambda_1}\boldsymbol{\alpha}_1 \vdots \sqrt{\lambda_2}\boldsymbol{\alpha}_2 \vdots \cdots \vdots \sqrt{\lambda_p}\boldsymbol{\alpha}_p\right] \tag{8.19}$$

Poznata ortogonalna dekompozicija kovarijacione (ili korelacione) matrice (videti Odeljak 7.2.2) sugeriše da je $\mathbf{\Sigma} = \lambda_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1' + \lambda_2\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2' + \ldots + \lambda_p\boldsymbol{\alpha}_p\boldsymbol{\alpha}_p'$, pa prema gornjim razmatranjima proizilazi da smo razložili kovarijacionu matricu prema modelu faktorske analize sa $m = p$ zajednička faktora i bez specifičnih faktora (matrica specifičnih varijansi jednaka je nula matrici) na sledeći način

$$\underset{(p\times p)}{\mathbf{\Sigma}} = \underset{(p\times p)}{\mathbf{B}}\ \underset{(p\times p)}{\mathbf{B}'} + \underset{(p\times p)}{\mathbf{0}} = \mathbf{BB}' \tag{8.20}$$

Kao i u analizi glavnih komponenata, ni ovde nam nije cilj da zadržimo sve zajedničke faktore. Suprotno od toga, težimo da sa što manjim brojem zajedničkih faktora "objasnimo" dobijenu kovarijacionu strukturu. U glavi posvećenoj analizi glavnih komponenata sugerisano je da se odbace one glavne komponente koje imaju u statističkom smislu zanemarljivo male vrednosti karakterističnih korena. Zadržavanjem samo prvih m, a odbacivanjem preostalih $p - m$ sabiraka u ortogonalnoj dekompoziciji kovarijacione matrice dobićemo njenu aproksimaciju. Ukoliko u analizi ne zanemarujemo značajnost specifičnih faktora, tada elemente matrice specifičnih varijansi dobijamo kao dijagonalne elemente razlike $\mathbf{\Sigma} - \mathbf{BB}'$, gde je matrica $\mathbf{BB}'$ formirana na osnovu prvih m sabiraka u ortogonalnoj dekompoziciji kovarijacione matrice.

U uzoračkom slučaju postupak primene metode glavnih komponenata u oceni faktorskog modela zahteva određivanje karakterističnih korena ($\hat{\lambda}_j$) i vektora ($\hat{\boldsymbol{\alpha}}_j$) uzoračke kovarijacione ($\mathbf{S}$) ili korelacione ($\mathbf{R}$) matrice. Na osnovu prvih m faktora formiramo matricu ocenjenih faktorskih opterećenja prema sledećem izrazu

$$\hat{\mathbf{B}} = \left[\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{\alpha}}_1 \vdots \sqrt{\hat{\lambda}_2}\hat{\boldsymbol{\alpha}}_2 \vdots \cdots \vdots \sqrt{\hat{\lambda}_p}\hat{\boldsymbol{\alpha}}_p \right] \tag{8.21}$$

Ocenjene specifične varijanse dobijamo kao dijagonalne elemente matrice $\mathbf{S} - \hat{\mathbf{B}}\hat{\mathbf{B}}'$, pa je $i-$ti elemenat

$$\hat{\psi}_i = s_{ii} - \sum_{j=1}^{m} \hat{\beta}_{ij}^2 , \ i = 1, 2, ..., p \tag{8.22}$$

a ocenjeni komunaliteti su

$$\hat{h}_i^2 = \hat{\beta}_{i1}^2 + \hat{\beta}_{i2}^2 + ... + \hat{\beta}_{im}^2 , \ i = 1, 2, ..., p \tag{8.23}$$

Broj zadržanih faktora možemo odrediti apriori, korišćenjem postupaka koji su navedeni kod izbora broja glavnih komponenata. Alternativa je da se posmatraju elementi *matrice reziduala*

$$\mathbf{S} - (\hat{\mathbf{B}}\hat{\mathbf{B}}' + \hat{\boldsymbol{\Psi}}) \tag{8.24}$$

koja je dobijena kao rezultat aproksimacije uzoračke kovarijacione matrice sa prvih m faktora. Broj faktora koje smo uključili u aproksimaciju povećavamo sve dotle dok ne procenimo da su elementi matrice reziduala dovoljno mali. Pri tome možemo definisati i formalniji postupak za određivanje kvaliteta aproksimacije kovarijacione strukture. Na primer, poredeći sume kvadrata elemenata matrice reziduala dobijene na osnovu primene alternativnih metoda ocenjivanja modela faktorske analize, zaključujemo da je metod ocenjivanja koji daje manju sumu kvadrata elemenata matrice reziduala bolji u odnosu na svoje konkurente. Interesantno je istaći da se korišćenjem ovog metoda ocene faktorskog modela ocenjena faktorska opterećenja za dati faktor ne menjaju povećanjem broja faktora.

Kada je reč o drugoj varijanti metode glavnih komponenata, poznatoj pod nazivom: *metod glavnih faktora*, podsećamo da je interesovanje u faktorskoj analizi usmereno ka kovarijansi između originalnih promenljivih i onom delu varijanse svake promenljive koji se može "objasniti" zajedničkim faktorima (komunalitet), tj. deo koji promenljiva deli sa ostalim promenljivama. Ako analizu zasnivamo na korelacionoj matrici, nju razlažemo prema faktorskom modelu: $\boldsymbol{\rho} = \mathbf{B}\mathbf{B}' + \boldsymbol{\Psi}$, gde su dijagonalni elementi

$$\rho_{ii} = 1 = h_i^2 + \psi_i , \ i = 1, 2, ..., p \tag{8.25}$$

Zajednički faktori "objašnjavaju" vandijagonalne elemente matrice $\boldsymbol{\rho}$ i deo dijagonalnih elemenata (h_i^2). Stoga formiramo novu matricu

$$\boldsymbol{\rho}-\boldsymbol{\Psi}=\mathbf{B}\mathbf{B}'=\begin{bmatrix} h_1^2 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & h_2^2 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & h_p^2 \end{bmatrix} \tag{8.26}$$

koja predstavlja korelacionu matricu zajedničkih faktora, a nazivamo je *redukovana korelaciona matrica*. Nju dakle dobijamo zamenom jedinica na glavnoj dijagonali korelacione matrice odgovarajućim komunalitetima. Ona se, a ne korelaciona matrica originalnih promenljivih (sa jedinicama na glavnoj dijagonali), razlaže u ortogonalnom modelu faktorske analize na matricu $\mathbf{B}$. Do ocene korelacione matrice zajedničkih faktora dolazimo na osnovu razlike ocenjene korelacione matrice i matrice specifičnih varijansi. Kako nam je pri tome potrebno da ocenimo specifične varijanse $\psi_i = 1 - h_i^2$, to pred nama otvara problem ocene komunaliteta pre nego što se primeni metod glavnih faktora radi ocene samog faktorskog modela.

Pri oceni matrice specifičnih varijansi $\boldsymbol{\Psi}$, moramo voditi računa o dva međusobom suprotstavljena zahteva. Naime, za datu uzoračku korelacionu matricu $\mathbf{R}$ ako ocenimo $\boldsymbol{\Psi}$ tako da su njeni elementi suviše veliki, ocenjena redukovana korelaciona matrica $\hat{\mathbf{B}}\hat{\mathbf{B}}' = \mathbf{R} - \hat{\boldsymbol{\Psi}}$ može postati negativno definitna, što je u suprotnosti sa pozitivnom definitnošću matrice $\mathbf{B}\mathbf{B}' = \boldsymbol{\rho} - \boldsymbol{\Psi}$. Zato se teži minimiziranju broja zajedničkih faktora m (predstavlja rang matrice $\mathbf{B}\mathbf{B}'$), ali uz istovremeno vođenje računa da se obezbedi pozitivna definitnost matrice $\hat{\mathbf{B}}\hat{\mathbf{B}}'$. Tako imamo da za male vrednosti elemenata matrice $\hat{\boldsymbol{\Psi}}$ broj faktora postaje suviše velik, a ako su velike vrednosti ocena specifične varijanse, matrica $\hat{\mathbf{B}}\hat{\mathbf{B}}'$ može postati negativno definitna.

Različite pristupe ocenjivanju komunaliteta, a time i specifične varijanse, možemo klasifikovati u dve grupe već prema tome da li koriste delimično ili u potpunosti elemente uzoračke korelacione matrice. Jedan metod iz prve grupe za ocenu $i-$tog komunaliteta uzima maksimalnu vrednost koeficijenata korelacije $i-$te promenljive i ma koje od preostalih $p-1$ promenljive. Apsolutna vrednost tog koeficijenata korelacije zamenjuje zatim $i-$tu jedinicu na glavnoj dijagonali korelacione matrice. Pokazalo se da ovaj metod ocene komunaliteta ima dobre osobine za velike vrednosti p. Druga grupa metoda koristi celokupnu korelacionu matricu u cilju formiranja ocene komunaliteta. U ovu grupu spada i pristup koji se najčešće koristi u empirijskim istraživanjima. Prema njemu ocena h_i^2 predstavljena je kvadratom višestrukog koeficijenata korelacije promenljive X_i i svih preostalih $p-1$ promenljivih. Ovaj metod daje donju granicu ocene komunaliteta. Praktičan postupak primene ovog metoda ocene specifičnih varijansi i komunaliteta bazira se na inverznoj matrici korelacione matrice $\mathbf{R}$. Tako je ocena $i-$te specifične varijanse $\hat{\psi}_i = 1/r^{ii}$, gde je r^{ii} $i-$ti dijagonalni element matrice $\mathbf{R}^{-1}$. Prema tome, inicijalna ocena komunaliteta je

$$\hat{h}_i^2 = 1 - \hat{\psi}_i = 1 - \frac{1}{r^{ii}}, \ i = 1, 2, ..., p \tag{8.27}$$

Nakon ocene komunaliteta i formiranja redukovane korelacione matrice vrši se izbor broja faktora primenom postupaka koji su korišćeni pri izboru broja glavnih komponenata. Na njih ćemo se podsetiti u narednom odeljku. Pretpostavivši da je taj izbor obavljen, određujemo karakteristične vektore redukovane korelacione matrice koji odgovaraju prvim m pozitivnim karakterističnim korenima te matrice. Prema tome, ocena matrice faktorskih opterećenja dobijena metodom glavnih faktora je

$$\hat{\mathbf{B}}^* = \left[\sqrt{\hat{\lambda}_1^*}\hat{\boldsymbol{\alpha}}_1^* \vdots \sqrt{\hat{\lambda}_2^*}\hat{\boldsymbol{\alpha}}_2^* \vdots \cdots \vdots \sqrt{\hat{\lambda}_p^*}\hat{\boldsymbol{\alpha}}_p^* \right] \tag{8.28}$$

gde su $(\hat{\lambda}_j^*, \hat{\alpha}_j^*)$, $j = 1, 2, ..., m$, parovi karakterističnih korena i pridruženih karakterističnih vektora redukovane korelacione matrice. Ocena specifične varijanse je zatim

$$\hat{\psi}_i^* = 1 - \sum_{j=1}^{m} \hat{\beta}_{ij}^{*2}, \ i = 1, 2, ..., p \tag{8.29}$$

gde su $\left\{ \hat{\beta}_{ij}^{*2} \right\}$ ocenjena faktorska opterećenja. Na osnovu ovako ocenjene specifične varijanse možemo ponovo oceniti komunalitete

$$\hat{h}_i^{*2} = \sum_{j=1}^{m} \hat{\beta}_{ij}^{*2}, \ i = 1, 2, ..., p \tag{8.30}$$

a zatim ih u narednoj iteraciji koristiti za formiranje redukovane korelacione matrice i određivanje njenih karakterističnih vektora. Ovaj iterativan proces ocenjivanja modela faktorske analize nastavljamo dalje do momenta kada promene u sukcesivnim ocenama komunaliteta budu zanemarive.

Dva pristupa oceni modela faktorske analize: metod glavnih faktora i metod glavnih komponenata su međusobom identični ako kod prvog za inicijalnu ocenu komunaliteta uzmemo jedinice. U tom smislu drugi metod možemo smatrati specijalnim slučajem prvog. Međutim, ideja koja leži u osnovi svakog od ova dva pristupa čini ih suštinski različitima metodama ocenjivanja modela faktorske analize.

Primer 8.3 Neka je na osnovu uzorka iz dimenzione populacije dobijena uzoračka korelaciona matrica $\mathbf{R}$

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0.8 & 0.3 & 0 \\ 0 & 1 & 0.1 & 0.6 & 0.8 \\ 0.8 & 0.1 & 1 & 0.4 & 0.1 \\ 0.3 & 0.6 & 0.4 & 1 & 0.7 \\ 0 & 0.8 & 0.1 & 0.7 & 1 \end{bmatrix}.$$

Korišćenjem metode glavnih komponenata i metode glavnih faktora oceniti model faktorske analize.

Karakteristični koreni i vektori matrice **R** prikazani su u donjoj tabeli

Karakteristični vektori				
prvi	drugi	treći	četvrti	peti
0.2775	0.6365	-0.1965	0.6831	-0.1123
0.4893	-0.3508	-0.6102	-0.1294	-0.4985
0.3426	0.5871	-0.1201	-0.6802	0.2468
0.5510	-0.0451	0.7568	-0.0213	-0.3481
0.5124	-0.3538	-0.0444	0.2314	0.7462
Karakteristični koreni				
2.5835	1.7166	0.3338	0.1941	0.1720

Prema kriterijumu jediničnog korena u daljem postupku zadržavamo prve dve glavne komponente, jer su njihovi karakteristični koreni veći od jedinice. Dva zajednička faktora "objašnjavaju" 86% ukupne uzoračke varijanse pošto je

$$\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p} = \frac{2.5835 + 1.7166}{5} = 0.8$$

Ocena matrice faktorskih opterećenja je

$$\hat{\mathbf{B}} = \left[\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{\alpha}}_1 \vdots \sqrt{\hat{\lambda}_2}\hat{\boldsymbol{\alpha}}_2 \right] = \begin{bmatrix} 0.4460 & 0.8339 \\ 0.7865 & -0.4596 \\ 0.5507 & 0.7692 \\ 0.8857 & -0.0591 \\ 0.8236 & -0.4635 \end{bmatrix}$$

Ocene komunaliteta date dijagonalnim elementima matrice $\hat{\mathbf{B}}\hat{\mathbf{B}}'$ i ostale elemenata ocenjenog faktorskog modela prikazujemo u donjoj tabeli.

Promenljive	**Ocena faktorskih opterećenja**		**Komunalitet**	**Specifična varijansa**
	F_1	F_2		
prva	0.4460	0.8339	0.8944	0.1056
druga	0.7865	-0.4596	0.8297	0.1703
treća	0.5507	0.7692	0.8949	0.1051
četvrta	0.8857	-0.0591	0.7879	0.2121
peta	0.8236	-0.4635	0.8932	0.1068
Karakteristični koreni	2.5835	1.7166		
Kumulativna proporcija ukupne uzoračke varijanse	51.67%	86%		

Razlaganje uzoračke korelacione matrice na osnovu modela faktorske analize sa dva faktora, korišćenjem glavnih komponenata, dato je izrazom

$$\hat{\mathbf{B}}\hat{\mathbf{B}}' + \hat{\mathbf{\Psi}} = \begin{bmatrix} 0.4460 & 0.8339 \\ 0.7865 & -0.4596 \\ 0.5507 & 0.7692 \\ 0.8857 & -0.0591 \\ 0.8236 & -0.4635 \end{bmatrix} \begin{bmatrix} 0.4460 & 0.7865 & 0.5507 & 0.8857 & 0.8236 \\ 0.8339 & -0.4596 & 0.7692 & -0.0591 & -0.4635 \end{bmatrix} +$$

$$\begin{bmatrix} 0.1056 & 0 & 0 & 0 & 0 \\ 0 & 0.1703 & 0 & 0 & 0 \\ 0 & 0 & 0.1051 & 0 & 0 \\ 0 & 0 & 0 & 0.2121 & 0 \\ 0 & 0 & 0 & 0 & 0.1068 \end{bmatrix} = \begin{bmatrix} 1 & -0.0325 & 0.8871 & 0.3457 & -0.0192 \\ & 1 & 0.0796 & 0.7237 & 0.8608 \\ & & 1 & 0.4423 & 0.0971 \\ & & & 1 & 0.7569 \\ & & & & 1 \end{bmatrix}$$

Poredeći dobijenu aproksimaciju korelacione matrice na osnovu dva zajednička faktora i uzoračku korelacionu matricu konstatujemo da zajednički faktori relativno uspešno reprodukuju korelacionu strukturu pet originalnih promenljivih. Kvalitet aproksimacije možemo sagledati i na osnovu matrice reziduala

$$\mathbf{R} + (\hat{\mathbf{B}}\hat{\mathbf{B}}' + \hat{\mathbf{\Psi}}) = \begin{bmatrix} 0 & 0.0325 & -0.0871 & -0.0457 & 0.0192 \\ & 0 & 0.0204 & -0.1237 & -0.0608 \\ & & 0 & -0.0423 & 0.0029 \\ & & & 0 & -0.0569 \\ & & & & 0 \end{bmatrix}$$

Najveće apsolutne vrednosti reziduala su u četvrtoj koloni matrice što nam sugeriše da dva zajednička faktora najslabije "objašnjavaju" varijabilitet četvrte promenljive. Ovo je vidljivo i na osnovu ocenjenih komunaliteta, jer je on najniži kod četvrte promenljive i iznosi 78.79%, odnosno upravo u tom iznosu ova dva faktora obuhvataju varijabilitet četvrte promenljive. Kod prve, treće i pete promenljive preko 89% njihovog varijabiliteta "objašnjeno" je sa dva zajednička faktora.

Postupak primene metoda glavnih faktora za ocenu modela faktorske analize zahteva inicijalnu ocenu komunaliteta. Opredelili smo se za ocenu h_i^2 na osnovu kvadrata višestrukog koeficijenta korelacije promenljive X_i i svih preostalih $p-1$ promenljivih. Inverzna matrica korelacione matrice je

$$\mathbf{R}^{-1} = \begin{bmatrix} 2.8592 & 0.1518 & -2.2299 & -0.2509 & 0.2772 \\ & 2.8106 & -0.0971 & -0.2469 & -2.0660 \\ & & 3.0271 & -0.6394 & 0.2226 \\ & & & 2.5415 & -1.5176 \\ & & & & 3.6928 \end{bmatrix}$$

pa su ocene komunaliteta $h_1^2 = 1 - \frac{1}{r^{11}} = 1 - \frac{1}{2.8592} = 0.6502$, $h_2^2 = 0.6442$, $h_3^2 = 0.6697$, $h_4^2 = 0.6065$ i $h_5^2 = 0.7292$. Zamenom jedinica na glavnoj dijagonali korelacione matrice sa upravo određenim komunalitetima dobijamo redukovanu korelacionu matricu

$$\begin{bmatrix} 0.6502 & 0 & 0.8 & 0.3 & 0 \\ & 0.6442 & 0.1 & 0.6 & 0.8 \\ & & 0.6697 & 0.4 & 0.1 \\ & & & 0.6065 & 0.7 \\ & & & & 0.7292 \end{bmatrix}$$

Ocena faktorskog modela na osnovu redukovane korelacione matrice data je u donjoj tabeli.

Promenljive	**Ocena faktorskih opterećenja**		**Komunalitet**	**Specifična varijansa**
	F_1	F_2		
prva	0.4080	0.7458	0.7226	0.2774
druga	0.7319	-0.3975	0.6936	0.3064
treća	0.5096	0.6984	0.7474	0.2526
četvrta	0.8071	-0.0410	0.6531	0.3469
peta	0.7936	-0.4235	0.8092	0.1908
Karakteristični koreni	2.2431	1.3829		
Kumulativna proporcija ukupne uzoračke varijanse	44.86%	72.52%		

Korišćenjem ocenjenih faktorskih opterećenja i specifičnih varijansi imamo

$$\hat{\mathbf{B}}^*\hat{\mathbf{B}}^{*\prime} + \hat{\mathbf{\Psi}}^* = \begin{bmatrix} 0.41 & 0.75 \\ 0.73 & -0.40 \\ 0.51 & 0.70 \\ 0.81 & -0.04 \\ 0.79 & -0.42 \end{bmatrix} \begin{bmatrix} 0.41 & 0.73 & 0.51 & 0.81 & 0.79 \\ 0.75 & -0.40 & 0.70 & -0.04 & -0.42 \end{bmatrix} +$$

$$\begin{bmatrix} 0.2774 & 0 & 0 & 0 & 0 \\ 0 & 0.3064 & 0 & 0 & 0 \\ 0 & 0 & 0,2526 & 0 & 0 \\ 0 & 0 & 0 & 0.3469 & 0 \\ 0 & 0 & 0 & 0 & 0.1908 \end{bmatrix} = \begin{bmatrix} 1 & 0.0022 & 0.7287 & 0.2987 & 0.0079 \\ & 1 & 0.0954 & 0.6070 & 0.7492 \\ & & 1 & 0.3826 & 0.1087 \\ & & & 1 & 0.6579 \\ & & & & 1 \end{bmatrix}$$

U poređenju sa originalnom korelacionom matricom i prethodno određenom aproksimacijom, konstatujemo za poslednju aproksimaciju da je nešto bolja u odnosu na prethodnu. Do ovog zaključka dolazimo i na osnovu poređenja matrica reziduala kod prve i druge aproksimacije. Matrica reziduala kod druge aproksimacije je

$$\mathbf{R} - \left(\hat{\mathbf{B}}^{*} \hat{\mathbf{B}}^{*\prime} + \hat{\mathbf{\Psi}}^{*} \right) = \begin{bmatrix} 0 & -0.0022 & 0.0713 & 0.0013 & -0.0079 \\ & 0 & 0.0046 & -0.0070 & 0.0508 \\ & & 0 & -0.0174 & -0.0087 \\ & & & 0 & 0.0421 \\ & & & & 0 \end{bmatrix}$$

Suma kvadrata elemenata matrice reziduala kod metode glavnih komponenata iznosi 0.0711, a kod metode glavnih faktora 0.0199. Znači da se u ovom primeru, pri ocenjivanju modela faktorske analize, bolja aproksimacija korelacione matrice postiže korišćenjem metode glavnih faktora u odnosu na metod glavnih komponenata. ■

Metod najveće verodostojnosti zahteva pretpostavku o rasporedu slučajnog vektora $\mathbf{X}$. Pretpostavlja se da je on raspoređen po $p-$dimenzionom normalnom rasporedu sa sredinom $\boldsymbol{\mu}$ i kovarijacionom matricom $\mathbf{\Sigma} = \mathbf{BB}' + \mathbf{\Psi}$. Matrica opterećenja $\mathbf{B}$ ima m zajedničkih faktora, a ovaj broj je određen pre nego što smo pristupili ocenjivanju. Na osnovu slučajnog uzorka od n elemenata $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ izračunavamo uzoračku kovarijacionu matricu $\mathbf{S}$ koja ima Wishartov raspored. Njegovim korišćenjem formiramo funkciju verodostojnosti čiji logaritam glasi

$$\ln L = \frac{1}{2} n \left[\ln \left| \mathbf{\Sigma} \right| + \operatorname{tr} \left(\mathbf{\Sigma}^{-1} \mathbf{S} \right) \right]$$

$$= \frac{1}{2} n \left\{ \ln \left| \mathbf{BB}' + \mathbf{\Psi} \right| + \operatorname{tr} \left[\left(\mathbf{BB}' + \mathbf{\Psi} \right)^{-1} \mathbf{S} \right] \right\} \tag{8.31}$$

pri čemu smo zanemarili članove koji ne sadrže parametre za ocenjivanje. Naš zadatak je da maksimiziramo $\ln L$ po elementima matrica $\mathbf{B}$ i $\mathbf{\Psi}$. Tako određene veličine $\tilde{\mathbf{B}}$ i $\tilde{\mathbf{\Psi}}$ nazivamo *ocene najveće verodostojnosti*. Ranije spomenuti problem nejednoznačnosti matrice faktorskih opterećenja s obzirom na ortogonalnu transformaciju, rešavamo nametanjem *uslova jednoznačnosti* prema kome je $\mathbf{B}'\mathbf{\Psi}^{-1}\mathbf{B}$ dijagonalna matrica. Do

ocena najveće verodostojnosti dolazimo složenim postupkom numeričke maksimizacije (videti: Lawley i Maxwell (1971) ili Morrison (1976)) prema kojima je napisan odgovarajući kompjuterski program u poznatijim statističkim programskim paketima opšte namene kao što su SAS, SPSS, BMDP i dr.

Prednost metode najveće verodostojnosti u odnosu na metod glavnih komponenata (faktora) ogleda se u tome da su njene ocene modela faktorske analize nezavisne od merne skale. Tako je ocena najveće verodostojnosti faktorskih opterećenja date promenljive proporcionalna njenoj standardnoj devijaciji. Na primer, faktorska opterećenja dobijena na osnovu korelacione matrice razlikuju se od faktorskih opterećenja ocenjenih na osnovu kovarijacione matrice za faktor $1/\sqrt{S_{ii}}$. Formalno, ovo sledi na osnovu poznate veze između kovarijacione i korelacione matrice: $\boldsymbol{\rho} = \mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2}$ (Odeljak 2.1.5), gde smo sa $\mathbf{D}$ označili dijagonalnu matricu sa varijansama promenljivih na glavnoj dijagonali. Prema modelu faktorske analize imamo

$$\boldsymbol{\rho} = \mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2} = (\mathbf{D}^{-1/2}\mathbf{B})(\mathbf{D}^{-1/2}\mathbf{B})' + \mathbf{D}^{-1/2}\boldsymbol{\Psi}\mathbf{D}^{-1/2} \tag{8.32}$$

što znači da se korelaciona matrica razlaže analogno razlaganju kovarijacione matrice. Ovde je matrica faktorskih opterećenja $\mathbf{B}_z = \mathbf{D}^{-1/2}\mathbf{B}$, a matrica specifičnih varijansi $\boldsymbol{\Psi}_z = \mathbf{D}^{-1/2}\boldsymbol{\Psi}\mathbf{D}^{-1/2}$. Prema osobini invarijantnosti ocene najveće verodostojnosti, ocena najveće verodostojnosti korelacione matrice je

$$\tilde{\boldsymbol{\rho}} = (\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{B}})(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{B}})' + \tilde{\mathbf{D}}^{-1/2}\tilde{\boldsymbol{\Psi}}\tilde{\mathbf{D}}^{-1/2} = \tilde{\mathbf{B}}_z\tilde{\mathbf{B}}_z' + \tilde{\boldsymbol{\Psi}}_z \tag{8.33}$$

gde su $\tilde{\mathbf{B}}$, $\tilde{\boldsymbol{\Psi}}$ i $\tilde{\mathbf{D}}^{-1/2}$ ocene najveće verodostojnosti od $\mathbf{B}$, $\boldsymbol{\Psi}$ i $\mathbf{D}^{-1/2}$ respektivno.

Na žalost, tokom iterativnog postupka ocenjivanja može se desiti da se dobiju negativne vrednosti specifičnih varijansi, što je statistički besmisleno. Zato se uvode dodatna ograničenja kojima zahtevamo da ove varijanse budu pozitivne. Slučaj negativnih specifičnih varijansi, a time pojava komunaliteta koji su veći od jedinice, naziva se *Heywoodov slučaj*.

Primer 8.4 Na osnovu podataka iz Primera 8.3 korišćenjem statističkog programskog paketa SAS ocenili smo model faktorske analize metodom najveće verodostojnosti. Rezultati ocenjivanja prikazani su u donjoj tabeli.

Promenljive	Ocena faktorskih opterećenja		Komunalitet	Specifična varijansa
	F_1	F_2		
prva	0.1477	0.8478	0.7406	0.2594
druga	0.8161	-0.1325	0.6836	0.3164
treća	0.2635	0.8977	0.8753	0.1247
četvrta	0.7727	0.2189	0.6450	0.3550
peta	0.9531	-0.1677	0.9365	0.0635
Kumulativna proporcija ukupne uzoračke varijanse	45.25%	77.62%		

Ocena korelacione matrice na osnovu modela faktorske analize je

$$\tilde{\mathbf{B}}\tilde{\mathbf{B}}' + \tilde{\mathbf{\Psi}} = \begin{bmatrix} 1 & 0.008 & 0.8 & 0.2997 & -0.0014 \\ & 1 & 0.0961 & 0.6016 & 0.8 \\ & & 1 & 0.4001 & 0.1007 \\ & & & 1 & 0.6997 \\ & & & & 1 \end{bmatrix}$$

a matrica reziduala

$$\mathbf{R} - \left(\tilde{\mathbf{B}}\tilde{\mathbf{B}}' + \tilde{\mathbf{\Psi}}\right) = \begin{bmatrix} 0 & -0.0082 & 0 & 0.0003 & 0.0014 \\ & 0 & 0.0039 & -0.0016 & 0 \\ & & 0 & -0.0001 & -0.0007 \\ & & & 0 & 0.0003 \\ & & & & 0 \end{bmatrix}$$

Poredeći sve tri ocene korelacione matrice zaključujemo da je ocena po metodu najveće verodostojnosti dala najbolju aproksimaciju korelacione matrice. Suma kvadrata elemenata matrice reziduala kod metode najveće verodostojnosti iznosi svega 0.000175, što je preko 113 puta manje u odnosu na odgovarajuću vrednost kod metode glavnih faktora, a preko 406 puta manje u odnosu na odgovarajuću vrednost kod metode glavnih komponenata.

Poređenjem kod različitih metoda kumulativne proporcije varijanse koju "objašnjavaju" faktori, uočavamo da se one menjaju od metode do metode. Najveće su kod metode glavnih komponenti: 51.67% za prvi i 86% za oba faktora, nešto niže kod metode najveće verodostojnosti: 45.25% za prvi i 77.62% za oba faktora, a najniže su kod metode glavnih faktora: 44.86% za prvi i 72.52% za oba faktora. Dobijeni rezultat je očekivan, jer se kod metoda glavnih komponenata faktori izdvajaju prema najvećem doprinosu ukupnoj uzoračkoj varijansu ($\mathrm{tr}(\mathbf{\Sigma})$), pa će ovaj kriterijum davati prednost upravo metodu glavnih komponenata u odnosu na ostale metode ocenjivanja. ■

8.2.3 Odredjivanje broja faktora

Dosadašnje izlaganje faktorske analize polazilo je od modela sa m zajedničkih faktora. Pretpostavljalo se da je broj faktora unapred tačno određen, pre ocenjivanja ostalih parametara modela. Međutim, ukoliko smo prilikom određivanja broja faktora pogrešili izabravši suviše mali broj faktora, tada će značajni zajednički faktori biti izostavljeni iz analize, a u suprotnom slučaju, izborom suviše velikog broja faktora neki od specifičnih faktora biće pomešani sa zajedničkim faktorima. Zato se određivanju

broja zajedničkih faktora mora prići sa puno pažnje, jer taj izbor u velikoj meri opredeljuje kvalitet zaključaka na osnovu ocenjenog modela faktorske analize.

Mada se faktorska analiza suštinski razlikuje od metode glavnih komponenata postupci izbora broja glavnih komponenata koriste se i prilikom određivanja broja faktora (videti Odeljak 7.3.3). Najpoznatiji je svakako kriterijum jediničnog korena prema kome u modelu zadržavamo onoliko zajedničkih faktora koliko ima karakterističnih korena uzoračke korelacione matrice koji su veći od jedinice (vrednost aritmetičke sredine karakterističnih korena). Znači da se zadržavaju oni faktori koji doprinose objašnjenju barem koliko i jedna promenljiva. Dobre osobine ovog kriterijuma su u slučaju kada je broj promenljivih između 20 i 50. Ako je broj promenljivih veći od 50, tada ovaj kriterijum izdvaja suviše veliki broj zajedničkih faktora, a ako je broj promenljivih manji od 20, tada je kriterijum konzervativan jer izdvaja suviše mali broj zajedničkih faktora. Po ideji sličan pristup zasnovan je na korišćenju broja pozitivnih karakterističnih korena matrice $\mathbf{R}-\hat{\boldsymbol{\Psi}}$ kao ocene donje granice broja faktora koje zadržavamo u analizi.

Svoju primenu u određivanju broja faktora našao je i Cattellov *scree* test. Mada svi faktori sadrže u izvesnom stepenu specifičnu varijansu, ona dominira nad zajedničkom varijansom kod poslednje izdvojenih faktora. Na osnovu *scree* testa određujemo optimalan broj faktora koje treba izdvojiti pre nego što specifična varijansa počne da dominira nad zajedničkom. Grafički prikaz karakterističnih korena prema redosledu izdvajanja faktora rezultira u krivoj na kojoj identifikujemo tačku preloma nakon koje ona aproksimativno postaje horizontalna prava. Redni broj faktora sugerisan tom tačkom preloma predstavlja broj faktora koji ćemo izdvojiti na osnovu *scree* testa. U odnosu na kriterijum jediničnog korena *scree* test izdvaja veći broj faktora.

Preporuka je da se koristi više od jednog kriterijuma za određivanje broja faktora. Konačna odluka o broju izdvojenih faktora zavisi i od mogućnosti smislene interpretacije rezultata faktorske analize. Tako se izdvojeni faktori, prema na primer kriterijumu jediničnog korena, podvrgavaju daljem postupku analize. Ukoliko se pokaže da dobijena faktorska struktura dobro reprezentuje međusobne veze promenljivih i omogućava interpretaciju svakog od izdvojenih faktora, tada je broj izdvojenih faktora optimalan. U suprotnom, pokušali bismo u narednoj iteraciji da ponovimo analizu sa drugim brojem izdvojenih faktora sve dok se ne ispune napred navedeni uslovi.

U faktorskoj analizi su pored navedenih kriterijuma razvijeni i asimptotski statistički postupci za testiranje broja zajedničkih faktora. Polazna pretpostavka tiče se rasporeda, kao i u slučaju ocene modela faktorske analize metodom najveće verodostojnosti. Testiramo hipotezu da je broj zajedničkih faktora m. Drugim rečima, na osnovu modela faktorske analize, sa tim brojem faktora na odgovarajući način smo reprodukovali kovarijacionu matricu. Nulta hipoteza je $H_0:\mathbf{BB}'+\boldsymbol{\Psi}$, protiv alternativne hipoteze $H_1:\boldsymbol{\Sigma}$ je ma koja pozitivno definitna matrica.

Test zasnovan na principu količnika verodostojnosti zahteva ocenu kovarijacione matrice bez i sa ograničenjem na parametre. Maksimalna vrednost funkcije

verodostojnosti bez ograničenja na parametre postiže se kada za ocenu $\boldsymbol{\Sigma}$ uzmemo ocenu najveće verodostojnosti $\frac{n-1}{n}\mathbf{S}$. Ako je tačna nulta hipoteza, tada se kovarijaciona matrica $\boldsymbol{\Sigma}$ može razložiti prema modelu faktorske analize: $\boldsymbol{\Sigma} = \mathbf{BB}' + \boldsymbol{\Psi}$. Na osnovu ocena najveće verodostojnosti $\tilde{\mathbf{B}}$ i $\tilde{\boldsymbol{\Psi}}$ dobijamo ocenu najveće verodostojnosti kovarijacione matrice, u oznaci $\tilde{\boldsymbol{\Sigma}}$. Stavljanjem u međusobni odnos maksimalne vrednosti funkcije verodostojnosti dobijene na osnovu $\tilde{\boldsymbol{\Sigma}}$ (uz ograničenje na kovarijacionu matricu definisano modelom faktorske analize) i maksimalne vrednosti funkcije verodostojnosti dobijene na osnovu $\frac{n-1}{n}\mathbf{S}$ (bez ograničenja na kovarijacionu matricu $\boldsymbol{\Sigma}$) definišemo Wilksovu lambdu (Λ). Na osnovu poznate transformacije ove statistike ($-2\ln\Lambda$) i uz Bartlettovu korekciju, statistika testa glasi

$$\left[n-\left(\frac{2p+11}{6}\right)-\frac{2m}{3}\right]\ln\frac{\left|\tilde{\mathbf{B}}\tilde{\mathbf{B}}'+\tilde{\boldsymbol{\Psi}}\right|}{\left|\frac{n-1}{n}\mathbf{S}\right|} \tag{8.34}$$

Ako je tačna nulta hipoteza ova statistika u velikim uzorcima ima χ^2-raspored sa $\frac{1}{2}[(p-m)^2-(p+m)]$ stepeni slobode. Na nivou značajnosti α odbacujemo hipotezu H_0 ako je na osnovu realizovanih vrednosti slučajnog uzorka

$$\left[n-\left(\frac{2p+11}{6}\right)-\frac{2m}{3}\right]\ln\frac{\left|\tilde{\mathbf{B}}\tilde{\mathbf{B}}'+\tilde{\boldsymbol{\Psi}}\right|}{\left|\frac{n-1}{n}\mathbf{S}\right|} > \chi_{[(p-m)^2-(p+m)]/2;\alpha} \tag{8.35}$$

Ukoliko smo odbacili nultu hipotezu, tada zaključujemo da matrica reziduala sadrži značajnu varijansu, pa je potrebno uključiti dodatne zajedničke faktore da bi se adekvatno aproksimirala kovarijansa između promenljivih. Da bismo bili u mogućnosti primeniti ovaj test, veličina uzorka n i $n-p$ moraju biti veliki. Takođe mora biti zadovoljena nejednakost

$$m < \frac{1}{2}\left(2p+1-\sqrt{8p+1}\right) \tag{8.36}$$

koja je dobijena na osnovu zahteva da broj stepeni slobode mora biti pozitivan.

Za veliki broj zajedničkih faktora statistika testa nije pouzdana. Zato se predlaže postupak korak po korak, tako što će analiza započeti sa jednim zajedničkim faktorom. Zatim povećavamo broj faktora za jedan sve dok se ne prihvati nulta hipoteza ili dok broj stepeni slobode ne postane negativan. Pre otpočinjanja ovog postupka proveravamo da li je uopšte potrebno sprovoditi faktorsku analizu. Naime, ako je kovarijaciona matrica dijagonalna to znači da su originalne promenljive međusobom nekorelisane, pa nema potrebe za faktorskom analizom. Zato se prvo proverava da li se može odbaciti hipoteza o sferičnosti.

Primer 8.5 Na osnovu ocene najveće verodostojnosti modela faktorske analize iz Primera 8.4 testirati hipotezu $H_0 : \boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}' + \boldsymbol{\Psi}$ sa $m = 2$ na nivou značajnosti $\alpha = 0.05$.

Statistika testa bazirana je na generalizovanim varijansama, tj. iskazana je preko kovarijacionih matrica. Nama je u zadatku data uzoračka korelaciona matrica, što znači da se i nulta hipoteza odnosi na populacionu korelacionu matricu ($H_0 : \boldsymbol{\rho} = \mathbf{B}_z\mathbf{B}_z' + \boldsymbol{\Psi}_z$). Tokom izlaganja postupka ocenjivanja modela nismo pravili razliku u oznaci između matrica faktorskih opterećenja dobijenih pri razlaganju korelacione i kovarijacione matrice. Na ovom mestu tu razliku pravimo, jer želimo pokazati da se ista statistika testa može koristiti na bazi ili jedne ili druge matrice faktorskih opterećenja. S obzirom na osobinu invarijantnosti ocene najveće verodostojnosti i osobine determinante, imamo da je

$$\frac{\left|\tilde{\boldsymbol{\Sigma}}\right|}{\left|\frac{n-1}{n}\mathbf{S}\right|} = \frac{\left|\tilde{\mathbf{D}}^{-1/2}\right|}{\left|\tilde{\mathbf{D}}^{-1/2}\right|} \frac{\left|\tilde{\mathbf{B}}\tilde{\mathbf{B}}' + \tilde{\boldsymbol{\Psi}}\right|}{\left|\frac{n-1}{n}\mathbf{S}\right|} \frac{\left|\tilde{\mathbf{D}}^{-1/2}\right|}{\left|\tilde{\mathbf{D}}^{-1/2}\right|} =$$

$$= \frac{\left|\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{B}}\tilde{\mathbf{B}}'\tilde{\mathbf{D}}^{-1/2} + \tilde{\mathbf{D}}^{-1/2}\tilde{\boldsymbol{\Psi}}\tilde{\mathbf{D}}^{-1/2}\right|}{\left|\tilde{\mathbf{D}}^{-1/2}\frac{n-1}{n}\mathbf{S}\tilde{\mathbf{D}}^{-1/2}\right|} = \frac{\left|\tilde{\mathbf{B}}_z\tilde{\mathbf{B}}_z' + \tilde{\boldsymbol{\Psi}}_z\right|}{\left|\mathbf{R}\right|}$$

pa se determinante kovarijacionih matrica mogu zameniti determinantama odgovarajućih korelacionih matrica. Tako dobijamo

$$\frac{\left|\tilde{\mathbf{B}}_z\tilde{\mathbf{B}}_z' + \tilde{\boldsymbol{\Psi}}_z\right|}{\left|\mathbf{R}\right|} = \frac{\begin{vmatrix} 1 & 0.008 & 0.8 & 0.2997 & -0.0014 \\ & 1 & 0.0961 & 0.6016 & 0.8 \\ & & 1 & 0.4001 & 0.1007 \\ & & & 1 & 0.6997 \\ & & & & 1 \end{vmatrix}}{\begin{vmatrix} 1 & 0 & 0.8 & 0.3 & 0 \\ & 1 & 0.1 & 0.6 & 0.8 \\ & & 1 & 0.4 & 0.1 \\ & & & 1 & 0.7 \\ & & & & 1 \end{vmatrix}} = \frac{0.0495}{0.0494} = 1.0014$$

a izračunata vrednost statistike testa je

$$\left[n - \left(\frac{2p+11}{6}\right) - \frac{2m}{3}\right] \ln \frac{\left|\tilde{\mathbf{B}}_z\tilde{\mathbf{B}}_z' + \tilde{\boldsymbol{\Psi}}_z\right|}{\left|\mathbf{R}\right|} = \left[100 - \left(\frac{10+11}{6}\right) - \frac{4}{3}\right] \ln(1.0014) = 0.1327$$

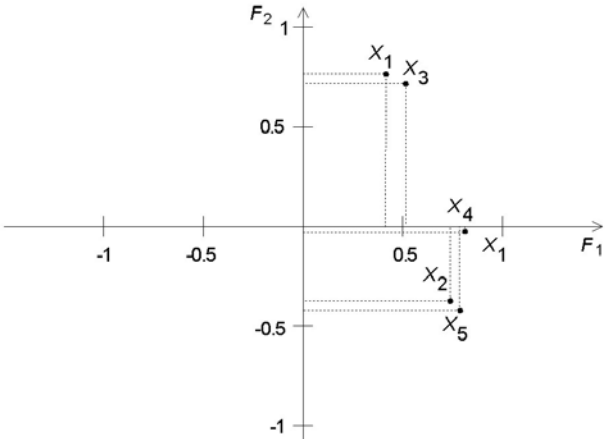
gde smo pretpostavili da je veličina uzorka 100. Pošto je broj stepeni slobode $\frac{1}{2}[(p-m)^2-(p+m)]=\frac{1}{2}[(5-2)^2-(5+2)]=1$, to je na nivou značajnosti od 5% kritična vrednost testa $\chi^2_{1;0.05}=3.841$. Kako je izračunata vrednost statistike testa manja od kritične vrednosti, prihvatamo nultu hipotezu. Zaključujemo, dakle, o adekvatnosti modela faktorske analize sa dva zajednička faktora. ■

Pored izloženog statističkog postupka određivanja broja zajedničkih faktora u literaturi su navedeni i drugi. O njima videti kod Jobsona (1992, ss. 415-419).

8.3 ROTACIJA FAKTORA

Kao što je pokazano na kraju Odeljak 8.2.1 ne postoji jednoznačno rešenje modela faktorske analize, odnosno postoji više različitih matrica faktorskih opterećenja koje generišu istu kovarijacionu matricu. Primenjeni postupak ortogonalne transformacije matrice u geometrijskom smislu predstavlja ustvari rotaciju koordinatnih osa za izvestan ugao, pri čemu njihov međusobni odnos ostaje nepromenjen, što znači da su ose i dalje međusobom ortogonalne, tj. pod uglom od 90^0. U faktorskoj analizi ortogonalnu transformaciju matrice faktorskih opterećenja i time impliciranu ortogonalnu transformaciju faktora (faktorskih osa) nazivamo *rotacija faktora* ili preciznije *ortogonalna rotacija faktora*. Napuštanjem zahteva da rotirani faktori moraju biti međusobom ortogonalni, razvijeni su postupci tzv. *neortogonalne rotacije faktora* ili kako se u literaturi nazivaju: *oblique rotacije*. U opštem slučaju, bez obzira na tip rotacije, postupak rotacije faktora primenjujemo u cilju dobijanja takve matrice faktorskih opterećenja koja će olakšati interpretaciju faktora. Izbor ugla za koji ćemo rotirati faktore opredeljen je jednim od kriterijuma, a najčešće korišćeni poznat je pod nazivom *jednostavna struktura*.

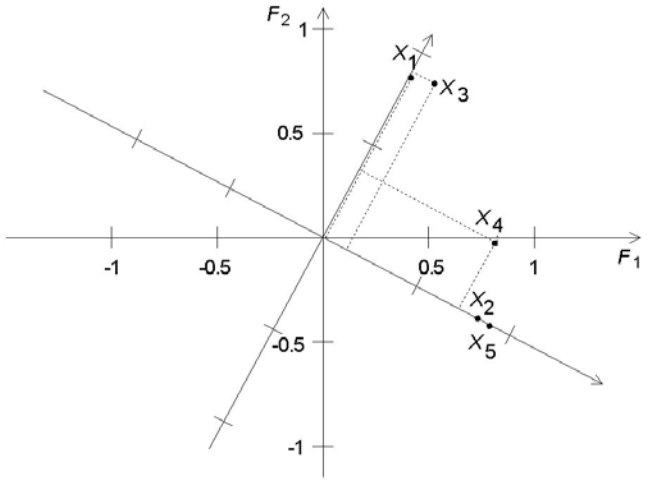
Pre nego što definišemo jednostavnu strukturu, ilustrujmo potrebu za rotacijom faktora na osnovu grafičkog prikaza rezultata ocene modela faktorske analize metodom glavnih faktora iz Primera 8.3.



Slika 8.1 Nerotirana faktorska opterećenja

Na Slici 8.1 za promenljive X_1, X_2, X_3, X_4 i X_5 prikazana su pripadajuća faktorska opterećenja u dvodimenzionom faktorskom prostoru koji grade dva faktora (faktorske ose) F_1 i F_2. Na osnovu grafikona je očigledno da sem kod četvrte, kod svih ostalih promenljivih imamo relativno visoke vrednosti faktorskih opterećenja kod oba faktora (projekcije tačaka na prvu i drugu faktorsku osu prikazane su kao isprekidane linije). To nam otežava interpretaciju dobijenog rešenja modela faktorske analize, jer nismo u mogućnosti nedvosmisleno zaključiti koje promenljive opredeljuju prvi, odnosno drugi faktor. Istovremeno se na Slici 8.1 jasno uočava grupisanje promenljivih u dve grupe. Prva grupa sa promenljivama: X_2, X_4 i X_5 i druga sa preostale dve promenljive: X_1 i X_3.

Primenom ortogonalne transformacije matrice faktorskih opterećenja, rotiramo faktorske ose tako da one u svom novom položaju prolaze što bliže tačkama koje predstavljaju faktorska opterećenja pet originalnih promenljivih. Time ćemo dobiti situaciju sa Slike 8.2 na kojoj su prikazane rotirane faktorske ose, odnosno rotirana faktorska opterećenja.



Slika 8.2 Rotirana faktorska opterećenja

Situacija u pogledu interpretacije faktora je sada u manjoj meri nejasna. Druga i peta promenljiva opredeljuju prvi faktor, a prva i treća opredeljuju drugi faktor. Za četvrtu promenljivu se može reći da je bliža prvom no drugom faktoru. Međusobni položaj pet tačaka na Slici 8.2 nije se promenio nakon rotacije faktora, nego se promenio samo referentni koordinatni sistem u odnosu na koji te tačke posmatramo. Dakle, promenom ugla gledanja na faktorska opterećenja, odnosno rotacijom faktora, jasnije sagledavamo prirodu faktora. Projekcija tačaka na rotirane faktorske ose ukazuje na promenjene vrednosti faktorskih opterećenja svake promenljive u odnosu na prvi i drugi faktor. Tako na primer, prvobitno visoka vrednost oba faktorska opterećenja kod treće promenljive, nakon rotacije faktora ukazuje na visoku vrednost opterećenja na drugi faktor, a nisku vrednost opterećenja na prvi faktor. Problemom interpretacije faktora pozabavićemo se detaljnije na kraju ovog poglavlja.

8.3.1 Jednostavna struktura

Jednostavna struktura predstavlja jedan od kriterijuma koji se koriste u faktorskoj analizi u cilju lakše interpretacije rešenja modela faktorske analize. Nju je definisao Thurstone (1945) vezujući je za numeričke osobine matrice faktorskih opterećenja. Kod jednostavne strukture pokušavamo da postignemo mali broj visokih vrednosti faktorskih opterećenja i veliki broj niskih vrednosti faktorskih opterećenja. Istraživač zatim interpretira niske vrednosti opterećenja kao nule, a visoke vrednosti kao vrednosti različite od nule. Za matricu faktorskih opterećenja kažemo da poseduje jednostavnu strukturu ako važi da:

1. svaki njen red (promenljiva) ima barem jednu nultu vrednost opterećenja,
2. svaka njena kolona (faktor) ima najmanje m nultih vrednosti opterećenja,
3. svaki par kolona (faktora) sadrži nekoliko redova (promenljivih) čija su opterećenja jednaka nuli u jednoj od posmatranih kolona, a različita od nule u ostalim kolonama,
4. kada je $m \geq 4$ svaki par kolona (faktora) sadrži veći broj redova (promenljivih) sa nultim vrednostima opterećenja u obe posmatrane kolone, i
5. svaki par kolona (faktora) ima mali broj opterećenja koja su različita od nule u svakoj od posmatranih kolona.

Rotacijom faktora želi se dobiti nova, rotirana, matrica faktorskih opterećenja sa navedenim osobinama jednostavne strukture, jer se time olakšava interpretacija ocene modela faktorske analize. Prvobitno se do jednostavne strukture dolazilo geometrijskim pristupom rotaciji. U dvodimenzionom prostoru vršila se rotacija faktora (kao što smo prikazali na Slici 8.2) za svaki par od m faktora, korišćenjem relativno složenog postupka (videti kod Harmana (1970, ss. 254-259)). Posle nekoliko uzastopnih rotacija po parovima dobija se ortogonalna matrica kojom se vrši transformacija kompletnog rešenja modela faktorske analize.

Iskustva nekih autora (na primer, Morrison (1976, s. 322)) su negativna u pogledu mogućnosti ispunjenja svih pet uslova jednostavne strukture u istraživanjima na podacima iz stvarnog života. No, većini autora ona je poslužila za definisanje analitičkih mera (kriterijuma) za merenje jednostavnosti strukture na osnovu koje će se vršiti rotacija faktora.

8.3.2 Metodi ortogonalne rotacije

Ortogonalna rotacija faktora ne menja međusobni odnos faktorskih osa, one su i dalje ortogonalne. Ona se po tome razlikuje od neortogonalne rotacije faktora kod koje nema tog ograničenja, jer se faktorske ose rotiraju nezavisno jedna od druge. Nakon neortogonalne rotacije faktorske ose u opštem slučaju zaklapaju međusobom ugao različit od 90^0.

Svi metodi ortogonalne rotacije koriste ortogonalnu matricu kojom transformišu matricu faktorskih opterećenja. Novodobijena matrica faktorskih opterećenja trebalo bi u većoj meri da izađe u susret zahtevanim osobinama jednostavne strukture. Metodi ortogonalne rotacije razlikuju se međusobom prema načinu definisanja pokazatelja, tj. analitičke mere jednostavnosti strukture. Određivanje ortogonalne matrice vrši se zatim maksimiziranjem izabrane analitičke mere.

Neka je $\mathbf{T}$ ortogonalna matrica kojom smo transformisali ocenjenu matricu faktorskih opterećenja $\hat{\mathbf{B}}$. Tokom izlaganja koristimo ocenu dobijenu metodom glavnih komponenata, ali izloženi rezultati važe i za ma koji drugi metod ocenjivanja. Znači da je $\hat{\mathbf{\Gamma}} = \hat{\mathbf{B}}\mathbf{T}$, pri čemu je $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}$, gde matricu $\hat{\mathbf{\Gamma}}$ nazivamo ocenjena matrica rotiranih faktorskih opterećenja. Ortogonalna rotacija faktora ima poželjne osobine koje su navele istraživače da češće koriste ove postupke rotacije u faktorskoj analizi od neortogonalnih. Na osnovu osobine da ocenjena kovarijaciona matrica ostaje nepromenjena ortogonalnom rotacijom faktora

$$\hat{\mathbf{B}}\hat{\mathbf{B}}' + \hat{\mathbf{\Psi}} = \hat{\mathbf{B}}\mathbf{T}\mathbf{T}'\hat{\mathbf{B}}' + \hat{\mathbf{\Psi}} = \hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}' + \hat{\mathbf{\Psi}} \tag{8.37}$$

sledi i osobina da matrica reziduala ostaje nepromenjena

$$\frac{n-1}{n}\mathbf{S} - \hat{\mathbf{B}}\hat{\mathbf{B}}' + \hat{\mathbf{\Psi}} = \frac{n-1}{n}\mathbf{S} - \hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}' - \hat{\mathbf{\Psi}} \tag{8.38}$$

Takođe, ortogonalnom rotacijom faktora ostaje nepromenjena ocena pojedinačnih i ukupnog komunaliteta jer je

$$\sum_{j=1}^{m} \hat{\gamma}_{ij}^2 = \sum_{j=1}^{m} \hat{\beta}_{ij}^2 = \hat{h}_j^2 \quad \text{i} \quad \hat{h} = \sum_{i=1}^{p} \hat{h}_i^2 \tag{8.39}$$

gde su $\hat{\gamma}_{ij}$ elementi matrice $\hat{\mathbf{\Gamma}}$. Time ostaju neizmenjene i ocene specifičnih varijansi. Ostali elementi modela faktorske analize menjaju svoju vrednost rotacijom faktora. Tu se pre svega misli na matricu faktorskih opterećenja, položaj faktorskih osa u $m-$dimenzionom faktorskom prostoru, kao i doprinos svakog faktora ukupnom komunalitetu.

Najčešće korišćeni analitički metod ortogonalne rotacije faktora svakako je Kaiserov (1958) *varimax metod*. Posmatrajmo kvadrate elemenata matrice $\hat{\mathbf{\Gamma}}$ u $j-$toj koloni (uz faktor F_j). Tretirajući ih kao populaciju od p brojeva odredimo njihovu varijansu

$$S_j^2 = \frac{1}{p}\sum_{i=1}^{p} \hat{\gamma}_{ij}^4 - \frac{1}{p^2}\left(\sum_{i=1}^{p} \hat{\gamma}_{ij}^2\right)^2, \quad j = 1, 2, ..., m \tag{8.40}$$

Sabirajući vrednosti ovih varijansi kod svih m faktora dobijamo *sirov varimax kriterijum* kao meru jednostavnosti strukture

$$V = \sum_{j=1}^{m} S_j^2 = \frac{1}{p} \sum_{j=1}^{m} \left[\sum_{i=1}^{p} \hat{\gamma}_{ij}^4 - \frac{1}{p} \left(\sum_{i=1}^{p} \hat{\gamma}_{ij}^2 \right)^2 \right] \tag{8.41}$$

koji težimo maksimizirati izborom matrice $\mathbf{T}$. Na osnovu njega dobijamo *normalan varimax kriterijum* ili kraće *varimax kriterijum*, tako što u izrazu za $\mathbf{V}$ zamenimo $\hat{\gamma}_{ij}^2$ sa $\hat{\gamma}_{ij}^2 / \hat{h}_i^2$ (vršimo normalizaciju kvadrata rotiranih faktorskih opterećenja). Znači da varimax kriterijum predstavlja zbir varijansi od $\hat{\gamma}_{ij}^2 / \hat{h}_i^2$ za svaki od m faktora. Kako težimo da maksimiziramo varijansu od $\hat{\gamma}_{ij}^2 / \hat{h}_i^2$, to će rezultirati uglavnom u nultim ili jediničnim vrednostima koeficijenata po kolonama matrice faktorskih opterećenja, pošto je $\left| \hat{\gamma}_{ij} / \sqrt{\hat{h}_i} \right| < 1$, i malim brojem koeficijenata čija će vrednost biti između ovih ekstrema. Na taj način ispunjavamo zahteve iskazane kod jednostavne strukture.

Postupak primene varimax i drugih kriterijuma jednostavnosti strukture predstavlja iterativan proces. Izdvojeni faktori posmatraju se po parovima i vrši se njihova rotacija, dok se ne postigne maksimalna vrednost varimax kriterijuma za prvi par faktora. Zatim se prvi rotirani faktor u paru sa trećim, nerotiranim faktorom, rotira do postizanja maksimuma varimax kriterijuma. Postupak se ponavlja dalje sve dok se svih $m(m-1)/2$ parova faktora na navedeni način ne rotiraju. Ovaj niz rotacija naziva se *ciklus*. On se ponavlja sve dok se ne postigne da su svi uglovi dobijeni za parove faktora manji od unapred izabrane vrednosti, koja predstavlja kriterijum konvergencije. Većina statističkih programskih paketa opšte namene sadrži postupke za rotaciju faktora. Zbog iterativne prirode postupka rotacije faktora, on je računski veoma zahtevan. Zato ćemo u narednim primerima koristiti izlazne rezultate primene programskog paketa SAS (PROC FACTOR) za rotaciju faktora.

Primer 8.6 Na osnovu podataka iz Primera 8.3 i ocenjene matrice faktorskih opterećenja metodom glavnih komponenata (GK), glavnih faktora (GF) i najveće verodostojnosti (NV) (Primer 8.4), izvršiti rotaciju faktora korišćenjem varimax kriterijuma.

Primenom SAS programskog paketa dobijene su ortogonalne matrice $\mathbf{T}$ u sva tri slučaja:

$$\mathbf{T}_{\text{GK}} = \begin{bmatrix} 0.8822 & 0.4709 \\ -0.4709 & 0.8822 \end{bmatrix}, \quad \mathbf{T}_{\text{GF}} = \begin{bmatrix} 0.8852 & 0.4653 \\ -0.4653 & 0.8852 \end{bmatrix}, \quad \mathbf{T}_{\text{NV}} = \begin{bmatrix} 0.9885 & 0.1515 \\ -0.1515 & 0.9885 \end{bmatrix}$$

Korišćenjem ovih matrica izračunati matricu rotiranih faktorskih opterećenja. Pokazati da se vrednost komunaliteta i specifičnih varijansi ne menja rotacijom faktora (vrednosti ovih pokazatelja u nerotiranom slučaju date su u tabelama u Primerima 8.3 i 8.4). Odrediti doprinos rotiranih faktora ukupnoj varijansi i komunalitetu.

Zahtevani rezultati prikazani su u donjoj tabeli

Promenljive	Rotirana matrica faktorskih opterećenja (varimax metod)					
	Glavne komponenete		Glavni faktori		Najveća verodostojnost	
	F_1	F_2	F_1	F_2	F_1	F_2
prva	0.0008	**0.9457**	0.0141	**0.8499**	0.0176	**0.8604**
druga	**0.9102**	-0.0351	**0.8328**	-0.0113	**0.8267**	-0.0073
treća	0.1236	**0.9379**	0.1261	**0.8553**	0.1245	**0.9273**
četvrta	**0.8092**	0.3649	**0.7335**	0.3392	**0.7306**	0.3334
peta	**0.9448**	-0.0210	**0.8996**	-0.0056	**0.9675**	-0.0214
Objašnjena varijansa	2.3913	1.9089	2.0569	1.5691	2.1691	1.7118
Proporcija	47.83%	38.17%	41.14%	31.38%	43.38%	34.24%

Mada su komunaliteti neizmenjeni, kao i ukupan komunalitet, došlo je do preraspodele objašnjene varijanse pojedinim faktorima. Tako na primer, kod ocene metodom glavnih komponenata proporcija objašnjene ukupne varijanse prvim faktorom iznosi 51.67% kod nerotirane matrice faktorskih opterećenja, a 47.83% kod rotirane matrice.

Napomena: Igrom slučaja su kod različitih metodi ocenjivanja dobijene rotirane matrice faktorskih opterećenja slične međusobom. Pre bismo mogli reći da je izuzetak nego pravilo da su pri korišćenju različitih metodi ocenjivanja faktori opredeljeni istim promenljivama (faktorska opterećenja koja opredeljuju pojedine faktore su podvučena u prethodnoj tabeli).

Komentar: U dvodimenzionom slučaju ortogonalna matrica transformacije je

$$\mathbf{T} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

gde je θ ugao za koji se rotiraju faktorske ose u pravcu kretanja kazaljke na satu. Tako na primer za matricu transformacije $\mathbf{T}$ kod metode glavnih faktora, imamo da je $\cos\theta = 0.8852$, odakle je $\theta = 27.8^0$. Znači da se upravo za ovaj ugao vrši rotacija faktorskih osa. Na Slici 8.2 za ovaj ugao smo rotirali faktorske ose i u odnosu na te, rotirane ose prikazali rotirana faktorska opterećenja. ■

Mada je varimax kriterijum najčešće korišćen, za rotaciju faktora koriste se i druge metode ortogonalne rotacije. Navedimo još jedan kriterijum. To je tzv. *quartimax kriterijum* prema kome se kao indikator jednostavnosti strukture uzima suma varijansi kvadrata svih elemenata matrice $\hat{\mathbf{\Gamma}}$, pa se maksimizira izraz:

$$\sum_{j=1}^{m}\sum_{i=1}^{p}\hat{\gamma}_{ij}^{4}-\frac{1}{pm}\left(\sum_{j=1}^{m}\sum_{i=1}^{p}\hat{\gamma}_{ij}^{2}\right)^{2}. \tag{8.42}$$

U drugom članu navedenog izraza figuriše zbir komunaliteta koji se ne menja rotacijom faktora. Zato je maksimiziranje ovog kriterijuma ekvivalentno maksimiziranju prvog člana izraza, tj. $\sum_{j=1}^{m}\sum_{i=1}^{p}\hat{\gamma}_{ij}^{4}$, pa otuda i naziv ovoga kriterijuma.

Za razliku od varimax metoda primena quartimax metoda obično rezultira u opštem faktoru, jer se varijansa računa na osnovu svih elemenata matrice faktorskih opterećenja. Kod varimax metoda varijansu opterećenja računamo za svaku kolonu posebno, pa se zakonito dobija veći broj zajedničkih faktora. Zato će nakon prvog opšteg faktora, faktorska opterećenja na preostalim faktorima biti niža kod quartimax rotacije nego što je to slučaj sa varimax rotacijom.

Ponderisani prosek sirovog varimax i quartimax kriterijuma predstavlja svojevrsnu generalizaciju ortogonalnih kriterijuma rotacije. Tzv. *orthomax metod* zasnovan je na maksimiziranju sledećeg izraza:

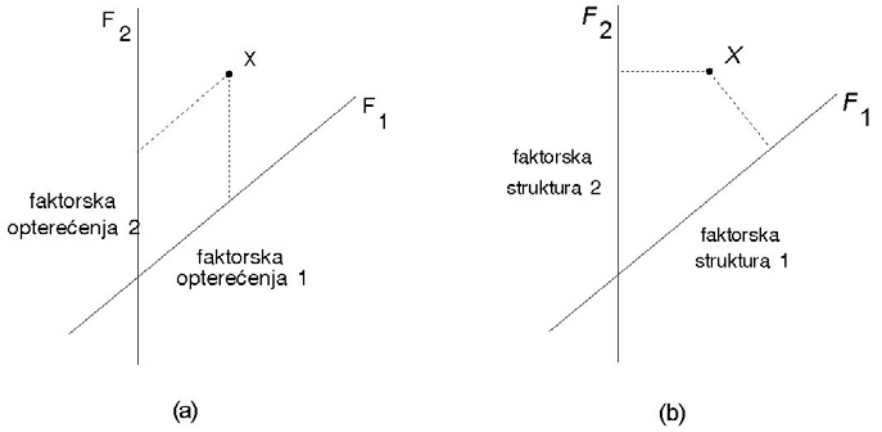
$$\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\hat{\gamma}_{ij}^{4}-\frac{\delta}{p}\left(\sum_{i=1}^{p}\hat{\gamma}_{ij}^{2}\right)^{2}\right], \tag{8.43}$$

gde je $0\le\delta\le 1$. Kada je $\delta=0$ ovaj kriterijum svodi se na quartimax kriterujum, a za $\delta=1$ na sirovi varimax kriterijum. Interesantno je istaći još dve vrednosti za δ, tj. $\delta=0.5$ generiše tzv. *biquartimax* kriterijum, a za $\delta=m/2$ orthomax kriterijum je ekvivalentan *equamax* kriterijumu.

8.3.3 Metodi neortogonalne rotacije

Napuštanjem zahteva za ortogonalnošću faktora dolazimo do metoda koje pri rotaciji faktora dozvoljavaju mogućnost da rotirani faktori zaklapaju ugao različit od . No, sada se, za razliku od ortogonalnog slučaja, matrica faktorskih opterećenja i matrica faktorske strukture razlikuju međusobom. Ove dve matrice zajedno sa korelacionom matricom faktora omogućavaju interpretaciju rezultata faktorske analize. Zbog razlike između matrica opterećenja i strukture pored osnovnih faktorskih osa, koje ovde nazivamo *primarne ose*, uvodimo dodatni par osa, pod nazivom *referentne ose* u odnosu na koje posmatramo faktorska opterećenja.

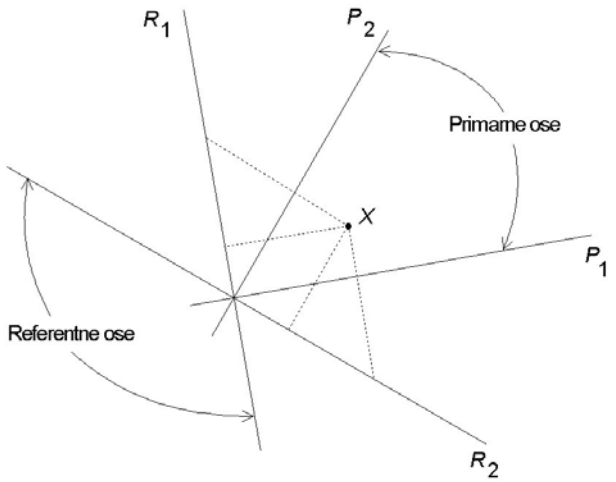
U ortogonalnom slučaju na Slici 8.1 i 8.2 prikazali smo ortogonalne projekcije faktorskih opterećenja. Međutim, kod neortogonalnog slučaja postoje dve vrste projekcija kojima prikazujemo položaj faktorskog opterećenja u odnosu na neortogonalno rotirane faktorske ose. Na Slici 8.3 prikazane su: (a) paralelna i (b) ortogonalna projekcija faktorskih opterećenja originalne promenljive X.



Slika 8.3 Projekcija faktorskih opterećenja: (a) paralelna i (b) ortogonalna

Vrednosti koje paralelne projekcije promenljive X imaju na faktorskim osama nazivamo paralelne koordinate. Faktorska matrica čiji su elementi paralelne koordinate pojedinih promenljivih je matrica faktorskog opterećenja. Zato smo na Slici 8.3(a) te paralelne koordinate označili kod prvog faktora sa "faktorska opterećenja 1", a kod drugog faktora sa "faktorska opterećenja 2". Vrednosti koje ortogonalne projekcije promenljive X imaju na faktorskim osama nazivamo ortogonalne koordinate. Faktorska matrica čiji su elementi ortogonalne koordinate pojedinih promenljivih je matrica faktorske strukture. Zato smo na Slici 8.3(b) te ortogonalne koordinate označili kod prvog faktora sa "faktorska struktura 1", a kod drugog sa "faktorska struktura 2". Ranije smo ukazali na to da elementi matrice faktorskih opterećenja (matrica faktorskog obrasca) mere doprinos svakog faktora varijansi promenljivih, a da elementi matrice faktorske strukture predstavljaju koeficijente korelacije svake promenljive sa faktorima.

Na Slici 8.4 prikazan je međusobni odnos primarnih i referentnih osa kao i projekcija (paralelna i ortogonalna) promenljive X na njih.



Slika 8.4 Međusobni odnos primarnih i referentnih osa

Sa Slike 8.4 se vidi da je referentna osa R_1 (R_2) ortogonalna na primarnu osu P_1 (P_2) tako da je međusobni ugao između primarnih osa P_1 i P_2 različit od ugla između referentnih osa R_1 i R_2. Kada su ta dva ugla identična, primarne i referentne ose se poklapaju i zaklapaju ugao od 90^0, odnosno dobijamo slučaj ortogonalnih faktora. Sa Slike 8.4 uočavamo da je paralelna projekcija promenljive X na na primer, primarnu osu P_1, istovremeno ortogonalna projekcija na referentnu osu R_2. Obrnuto važi za paralelnu projekciju na referentnu osu R_1 koja je istovremeno ortogonalna projekcija na primarnu osu P_2. Stoga kažemo da se uzajamna veza primarnih i referentnih osa ispoljava tako da su koeficijenti faktorskih opterećenja primarnih osa koeficijenti faktorske strukture refrentnih osa, a da su koeficijenti faktorske strukture primarnih osa koeficijenti faktorskih opterećenja referentnih osa.

Metode neortogonalne rotacije faktora zasnovane su na različitim kriterijumima. Njihovim se korišćenjem dobijaju rešenja modela faktorske analize koja korespondiraju jednostavnoj strukturi, a pri tome faktori nisu suviše korelisani međusobom. Navodimo samo imena, a detaljnije o njihovoj definiciji kao i o primenama može se naći u literaturi posvećenoj faktorskoj analizi. Među poznatijim metodama neortogonalne rotacije su: oblimax, quartimin, biquartimin, oblimin, Harris-Kaiserov metod i dr. O načinu njihovog konstruisanja i primeni može se videti kod Harmana (1970) i Fulgosija (1979).

Primer 8.7 Na osnovu podataka iz Primera 8.3 i ocenjene matrice faktorskih opterećenja metodom glavnih faktora, korišćenjem Harris-Kaiserovog metoda izvršiti rotaciju faktora.

Primenom SAS programskog paketa dobijena je oblique matrica transformacije:

$$\begin{bmatrix} 0.8457 & 0.3919 \\ -0.5706 & 0.9420 \end{bmatrix}$$

Pokazati da ova matrica nije ortogonalna.

Korišćenjem ove matrice izvršena je rotacija matrice faktorskih opterećenja. Ona, zajedno sa matricom faktorske strukture, matricom referentne strukture, matricom korelacije faktora i matricom korelacije referentnih osa prikazana je u narednoj tabeli.

Dobijeno rešenje, po strukturi faktora, identično je prethodno određenom rešenju koje je dobijeno korišćenjem varimax metoda rotacije faktora. Na osnovu vrednosti koeficijenta korelacije faktora možemo izračunati ugao koji faktori zaklapaju. Kako je kosinus tog ugla jednak upravo koeficijentu korelacije: 0.198, to je ugao među faktorima jednak 78.6^0.

Promenljive	Harris-Kaiserov metod					
	Rotirana matrica faktorskih opterećenja		Matrica faktorske strukture		Referentna struktura	
	F_1	F_2	F_1	F_2	F_1	F_2
prva	-0.0806	**0.8623**	0.0902	0.8464	-0.0790	0.8453
druga	**0.8457**	-0.0876	0.8284	0.0799	0.8290	-0.0859
treća	0.0324	**0.8575**	0.2023	0.8639	0.0318	0.8405
četvrta	**0.7060**	0.2776	0.7610	0.4175	0.6920	0.2721
peta	**0.9128**	-0.0879	0.8954	0.0928	0.8948	-0.0862
Faktori	**Korelaciona matrica faktora**		**Korelaciona matrica referentnih osa**			
F_1	1	0.1980	1	-0.1980		
F_2	0.1980	1	-0.1980	1		

■

8.3.4 Interpretacija faktora

Pre nego se pristupi interpretaciji faktora mora se odgovoriti na sledeće pitanje: koji se od ocenjenih elemenata matrice faktorskih opterećenja mogu smatrati statistički značajnim? Na žalost, u odgovoru na postavljeno pitanje ne možemo posegnuti za formalnom statističkim postupkom, ali je na raspolaganju nekoliko iskustvenih kriterijuma.

Prvi kriterijum je proistekao iz iskustva velikog broja istraživača u primeni modela faktorske analize. Sučeljavajući apriorna saznanja o prirodi izučavanog fenomena i rezultate brojnih empirijskih istraživanja istog (broj i struktura faktora), oni su sugerisali da se svi koeficijenti faktorskih opterećenja čija je apsolutna vrednost veća od 0.30, smatraju statistički značajno različitim od nule. Kod uzoraka veličine 50 i više elemenata, ovaj kriterijum se pokazao prihvatljivim u poređenju sa drugima.

Sledeći kriterijum zasnovan je na činjenici da je kod ortogonalnog modela faktorske analize matrica faktorskih opterećenja identična matrici faktorske strukture. Kako su elementi ove potonje koeficijenti korelacije promenljivih sa faktorima, to nam njihova visoka vrednost sugeriše da odnosna promenljiva opredeljuje faktor sa kojim je korelisana. Zato se testovi statističke značajnosti koeficijenata korelacije direktno primenjuju na elemente matrice faktorskih opterećenja. Tako na primer, standardni $t-$test za testiranje hipoteze o nultoj vrednosti koeficijenta korelacije sugeriše, za uzorke veličine 100 elemenata i na nivou značajnosti od 5% i 1%, da se smatraju statistički značajnim ona faktorska opterećenja čija je apsolutna vrednost veća od 0.19 i 0.26 respektivno. Pri tome se preporučuje korišćenje višeg nivoa značajnosti (od na primer 1%) jer je teško proceniti veličinu greške koja se čini u faktorskoj analizi. Što se uticaja veličine uzorka tiče, poznato je iz primene ovog $t-$testa da će se kritična vrednost, od na primer 0.26, snižavati sa povećanjem veličine uzorka.

Navedeni kriterijumi ne uzimaju u obzir broj promenljivih u analizi kao i redosled faktora čija opterećenja preispitujemo sa stanovišta značajnosti. U tom smislu se

sugeriše da sa povećanjem broja promenljivih u analizi snižavamo nivo značajnosti. S druge strane, što je viši redni broj izdvojenog faktora, to bi trebalo uzimati viši nivo značajnosti pri testiranju značajnosti opterećenja vezanih za taj faktor. Ove savete praktične prirode direktno koristimo u narednom etapnom postupku interpretacije matrice faktorskih opterećenja.

Povezujući teorijske rezultate sa naraslim praktičnim iskustvom u primeni modela faktorske analize predlaže se sledeći postupak koji nam olakšava kretanje u pravcu smislene interpretacije faktora.

Posmatrajmo matricu faktorskih opterećenja po redovima (promenljivama). Zaokružimo koeficijenata sa najvećom apsolutnom vrednošću u prvom redu. Zatim prelazimo u drugi red i u njemu tražimo najveću apsolutnu vrednost faktorskog opterećenja te promenljive za ma koji faktor, pa je zaokružimo. Na isti način postupamo sa preostalim od ukupno p redova matrice.

Nakon toga proveravamo značajnost zaokruženih faktorskih opterećenja korišćenjem gore izloženih kriterijuma, pa podvučemo statistički značajna faktorska opterećenja. Idealna situacija je kada se broj zaokruženih i podvučenih koeficijenata poklapa. Tada svaka promenljiva pripada samo jednom faktoru. Sada je naš zadatak da svakom faktoru pridružimo odgovarajući naziv, što činimo s obzirom na strukturu faktora, odnosno listu onih promenljivih koje su visoko korelisane sa tim faktorom.

Moguća su dva otklona od ove idealne situacije. Prvi, kada je manji broj podvučenih (statistički značajnih koeficijenata) od broja zaokruženih koeficijenata. Znači da se neka od promenljivih nije "pridružila" jednom od izdvojenih faktora, tj. nije ni sa jednim od njih statistički značajno korelisana. U takvoj situaciji možemo interpretirati dobijeno rešenje jednostavno zanemarujući tu promenljivu. Ili ćemo preispitati značaj te promenljive sa stanovišta konkretnog istraživanja, koristeći se njenim komunalitetom. Ako nam preispitivanje promenljive sugeriše (preko niske vrednosti komunaliteta), možemo je izostaviti iz analize i izvesti novo rešenje modela faktorske analize bez nje.

Druga, u praksi češća situacija jeste pojava većeg broja statistički značajnih faktorskih opterećenja u jednom redu. To znači da je ta promenljiva korelisana sa više faktora, što otežava interpretaciju rešenja. Ovakav rezultat obično se dobija ako analizu baziramo na nerotiranoj, umesto rotiranoj, matrici faktorskih opterećenja. Upravo to je bio razlog za korišćenje postupka rotacije faktora, jer smo rotacijom minimizirali broj značajnih faktorskih opterećenja po redovima matrice, a maksimizirali broj neznačajnih (zahtev iskazan preko koncepta jednostavne strukture).

Kada smo svaku promenljivu pridružili jednom od faktora, naziv faktoru dajemo prema listi promenljivih koji su mu pridruženi i to vodeći računa o visini i predznaku opterećenja. Dajemo veći naglasak onoj promenljivoj koja je sa posmatranim faktorom u najčvršćoj korelacionoj vezi.

Primer 8.8 Korišćenjem rezultata ocene modela faktorske analize iz Primera 8.3 i rotacije dobijenog rešenja iz Primera 8.6 i 8.7 interpretirati dobijeno rešenje modela faktorske analize.

Na osnovu ocenjenih matrica faktorskih opterećenja iz Primera 8.3 očigledno je da su sva faktorska opterećenja statistički značajna (osim kod četvrte promenljive na drugom faktoru), što nam onemogućava interpretaciju faktora. Ovaj zaključak važi bez obzira na primenjeni metod ocene modela. Ortogonalnom i neortogonalnom rotacijom matrica faktorskih opterećenja dobijamo matrice koje smo prikazali u tabelama u okviru Primera 8.6 i 8.7. Kriterijumi statističke značajnosti faktorskih opterećenja nedvosmisleno izdvajaju u svakom redu (osim u četvrtom) samo jedan koeficijenat. Kod svih rotiranih matrica dobijamo identičan rezultat sa stanovišta liste promenljivih koje su pridružene prvom, odnosno drugom faktoru. Tako prvi faktor opredeljuju promenljive X_2, X_4 i X_5 (sve tri sa pozitivnim predznakom), a drugi faktor promenljive X_1 i X_3 (obe sa pozitivnim predznakom). Prema nazivu promenljivih pridružili bismo naziv svakom od izdvojenih faktora. ■

Napominjemo da je postupak pridruživanja imena faktorima u većoj meri zasnovan na poznavanju fenomena koji se istražuje i subjektivnoj proceni istraživača, nego na formalnom pristupu. Uostalom, u svakoj fazi primene faktorske analize vrši se subjektivna procena u kom pravcu dalje usmeriti analizu, pa ni faza davanja imena faktorima nije izuzetak. Ako se podsetimo na odsustvo formalnih kriterijuma izbora broja faktora, metode ocene modela, rotacije faktora, provere značajnosti faktorskih opterećenja, tada možemo izraziti svoje slaganje sa stavom onih istraživača koji kažu da je faktorska analiza u većoj meri veština, no naučni metod analize podataka. Neki autori, kao što su Chatfield i Collins (1980, ss. 88-9) idu dotle da na osnovu iscrpnih kritika upućenih faktorskoj analizi, istu ne preporučuju u većini konkretnih situacija.

8.4 FAKTORSKI SKOROVI

Faktorska analiza je metod za redukciju podataka. Kao takva ona često predstavlja samo međukorak u složenijem istraživanju. To znači da dobijeni faktori predstavljaju polazni podatak za neku od metodi multivarijacione analize. Zato postoji interes da se na osnovu ocenjenog modela faktorske analize izračuna ili oceni vrednost faktora za svaku jedinicu posmatranja. Te realizovane vrednosti faktora nazivamo *faktorski skorovi*. Oni se mogu koristiti u regresionoj analizi kao opservacije objašnjavajućih promenljivih, u analizi grupisanja, diskriminacionoj analizi ili kanoničkoj korelacionoj analizi.

Kod metode glavnih komponenata jednostavno se za svaku jedinicu posmatranja određuje vrednost glavnih komponenata, jer su one linearna kombinacija originalnih promenljivih. Međutim, faktorski skorovi se u opštem slučaju ne mogu direktno izračunati, nego se moraju oceniti. Direktno izračunavanje faktorskih skorova je moguće

samo ako smo pri ocenjivanju modela faktorske analize koristili metod glavnih komponenata. Primena ostalih metoda ocenjivanja modela faktorske analize zahteva korišćenje jednog od postupaka za ocenu faktorskih skorova. U ovom poglavlju navodimo dva takva postupka, od kojih je prvi zasnovan na regresionoj metodi, a drugi na metodi ponderisanih najmanjih kvadrata. Prethodno pokažimo kako se direktno izračunavaju faktorski skorovi na osnovu ocenjenog modela faktorske analize metodom glavnih komponenata.

8.4.1 Direktno izračunavanje faktorskih skorova

Koristeći metod glavnih komponenata radi ocene modela faktorske analize, pretpostavili smo da u modelu figurišu samo zajednički, a ne i specifični faktori. Drugim rečima, model faktorske analize pišemo u obliku: $\mathbf{X} = \mathbf{BF}$, gde je $\mathbf{B}$ $(p \times m)$ matrica faktorskih opterećenja, $\mathbf{X}$ $(p \times 1)$ opažljiv slučajni vektor, a $\mathbf{F}$ $(m \times 1)$ slučajan vektor neopažljivih faktora. U opštem slučaju je $m \leq p$. Bez umanjenja opštosti pretpostavimo da je sredina $\boldsymbol{\mu} = \mathbf{0}$.

Ako je broj zadržanih faktora u modelu jednak broju promenljivih $(m = p)$, tada je matrica $\mathbf{B}$ kvadratna matrica. Rešavajući model po $\mathbf{F}$, dobijamo relaciju na osnovu koje direktno možemo izračunati vrednosti faktorskih skorova. Naime, $\mathbf{F} = \mathbf{B}^{-1}\mathbf{X}$, a prema izvođenju iz Odeljka 8.2.2 sledi da je $\mathbf{F} = \boldsymbol{\Lambda}^{-1/2}\mathbf{AX}$, gde je $\boldsymbol{\Lambda}$ $(p \times p)$ dijagonalna matrica čiji su elementi karakteristični koreni kovarijacione matrice $\boldsymbol{\Sigma}$ i $\mathbf{A}$ $(p \times p)$ matrica čiji su redovi karakteristični vektori kovarijacione matrice $\boldsymbol{\Sigma}$. Dakle, faktorski skorovi se direktno izračunavaju kao linearna kombinacija originalnih promenljivih, a ne ocenjuju se statistički.

Za analizu je relevantniji slučaj kada je broj zadržanih faktora u modelu manji od broja promenljivih $(m < p)$. Tada uočavamo da u međusobnoj vezi faktora i originalnih promenljivih matrice imaju sledeće dimenzije:

$$\underset{(m\times 1)}{\mathbf{F}} = \underset{(m\times m)}{\boldsymbol{\Lambda}^{-1/2}}\ \underset{(m\times p)}{\mathbf{A}}\ \underset{(p\times 1)}{\mathbf{X}} \tag{8.44}$$

Dakle, matrica $\boldsymbol{\Lambda}$ je dijagonalna matrica čiji su elementi prvih m karakterističnih korena kovarijacione matrice $\boldsymbol{\Sigma}$, redovi matrice $\mathbf{A}$ predstavljaju prvih m pridruženih karakterističnih vektora kovarijacione matrice $\boldsymbol{\Sigma}$. Prema tome, raspolažući realizovanim vrednostima promenljive $\mathbf{X}$ i ocenom matrica $\boldsymbol{\Lambda}$ i $\mathbf{A}$ možemo izračunati faktorske skorove (realizovane vrednosti faktora $\mathbf{F}$). Naime, ta realizovana vrednost faktora je

$$\hat{\mathbf{f}}_j = \hat{\boldsymbol{\Lambda}}^{-1/2}\hat{\mathbf{A}}\mathbf{x}_j, \quad j = 1, 2, ..., m \tag{8.45}$$

gde su $\hat{\boldsymbol{\Lambda}}$ i $\hat{\mathbf{A}}$ matrice ocenjene na osnovu uzoračke kovarijacione matrice $\mathbf{S}$, a $\mathbf{x}_j$ je $j-$ta realizovana vrednost promenljive $\mathbf{X}$. Ako umesto uzoračke kovarijacione matrice $\mathbf{S}$ koristimo uzoračku korelacionu matricu $\mathbf{R}$, tada će se ovaj izraz na odgovarajući

način modifikovati da bi se odnosio na originalnu promenljivu koja je standardizovana. Sada tu realizovanu vrednost faktora izračunavamo preko izraza:

$$\hat{\mathbf{f}}_j = \hat{\mathbf{\Lambda}}_z^{-1/2}\hat{\mathbf{A}}_z\mathbf{z}_j, \quad j = 1,2,...,m \tag{8.46}$$

gde su $\hat{\mathbf{\Lambda}}_z$ i $\hat{\mathbf{A}}_z$ matrice ocenjene na osnovu uzoračke korelacione matrice $\mathbf{R}$, a $\mathbf{z}_j$ $j-$ta realizovana vrednost standardizovane promenljive $\mathbf{Z}$.

Prednost metoda glavnih komponenata u oceni modela faktorske analize ogleda se i u tome da je relativno jednostavno doći do rotiranih faktorskih skorova. Pretpostavimo da smo jednim od metoda rotacije izvršili rotaciju matrice faktorskih opterećenja: $\mathbf{\Gamma} = \mathbf{BT}$, gde je $\mathbf{B}$ nerotirana matrica faktorskih opterećenja, $\mathbf{\Gamma}$ je odgovarajuća rotirana matrica, a $\mathbf{T}$ je matrica transformacije. Pomnožimo model $\mathbf{X} = \mathbf{\Gamma F}^*$, gde je $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$, s leve strane matricom $\mathbf{\Gamma}'$ i rešimo ga po $\mathbf{F}^*$. Dobićemo da je $\mathbf{F}^* = \left(\mathbf{\Gamma}'\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}'\mathbf{X}$. Zamenom $\mathbf{\Gamma}$ odgovarajućom ocenom po metodu glavnih komponenata i realizovanih vrednosti slučajnog vektora $\mathbf{X}$, dobili bismo izračunate vrednosti rotiranih faktorskih skorova.

8.4.2 Ocena faktorskih skorova regresionom metodom

Osnovu ocene faktorskih skorova regresionom metodom postavio je Thompson (1951). Pošavši od modela faktorske analize i pratećih pretpostavki on je posmatrao združeni raspored zajedničkih faktora $\mathbf{F}$ i specifičnih faktora $\boldsymbol{\varepsilon}$. Ako je taj raspored normalan sa $Cov(\mathbf{F}) = \mathbf{I}$, $Cov(\boldsymbol{\varepsilon}) = \mathbf{\Psi}$ i $Cov(\boldsymbol{\varepsilon}, \mathbf{F}) = 0$ tada slučajan vektor $\mathbf{X}$ i vektor zajedničkih faktora $\mathbf{F}$ imaju $(m+p)-$dimenzioni normalan raspored sa kovarijacionom matricom

$$\begin{bmatrix} \mathbf{BB}' + \mathbf{\Psi} & \vdots & \mathbf{B} \\ \cdots\cdots & \vdots & \cdots \\ \mathbf{B}' & \vdots & \mathbf{I} \end{bmatrix} \tag{8.47}$$

Na osnovu rezultata iz Poglavlja 3.2 uslovni raspored od $\mathbf{F}$ pod uslovom $\mathbf{X} = \mathbf{x}$ je normalan sa sredinom

$$E(\mathbf{F} \mid \mathbf{x}) = \mathbf{B}'\left(\mathbf{BB}' + \mathbf{\Psi}\right)^{-1}\mathbf{x} \tag{8.48}$$

i kovarijacionom matricom

$$Cov(\mathbf{F} \mid \mathbf{x}) = \mathbf{I} - \mathbf{B}'\left(\mathbf{BB}' + \mathbf{\Psi}\right)^{-1}\mathbf{B} \tag{8.49}$$

gde smo, bez umanjenja opštosti, pretpostavili da je $\boldsymbol{\mu} = \mathbf{0}$. Sredina uslovnog rasporeda predstavlja multivarijacionu regresiju zajedničkih faktora $\mathbf{F}$ na originalne promenljive. Koeficijenti u toj regresiji su $\mathbf{B}'\left(\mathbf{BB}' + \mathbf{\Psi}\right)^{-1}$, pa na osnovu ocena matrice faktorskih opterećenja $\mathbf{B}$ i specifičnih varijansi $\mathbf{\Psi}$, kao i realizovanih vrednosti slučajnog vektora $\mathbf{X}$ ocenjujemo vrednost faktorskih skorova:

$$\mathbf{f}_j = \tilde{\mathbf{B}}'\left(\tilde{\mathbf{B}}\tilde{\mathbf{B}}' + \tilde{\boldsymbol{\Psi}}\right)^{-1}\mathbf{x}_j, \quad j = 1,2,...,n \tag{8.50}$$

gde smo sa $\mathbf{f}_j$ i $\mathbf{x}_j$ označili $j-$tu realizovanu vrednost faktora i realizovanu vrednost $j-$tog elementa slučajnog uzorka $(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$ respektivno. U gornjem izrazu korišćene su izračunate vrednosti ocena najveće verodostojnosti $\tilde{\mathbf{B}}$ i $\tilde{\boldsymbol{\Psi}}$. Umesto njih mogu se uvrstiti ocene $\hat{\mathbf{B}}^*$ i $\hat{\boldsymbol{\Psi}}^*$ dobijene metodom glavnih faktora.

Ako smo preko uzoračke korelacione matrice $\mathbf{R}$ ocenili $\mathbf{B}$ i $\boldsymbol{\Psi}$ tada umesto realizovane vrednosti slučajnog vektora $\mathbf{X} = \mathbf{x}$ koristimo njegovu standardizovanu vrednost $\mathbf{Z} = \mathbf{z}$. Sada ocena faktorskih skorova glasi:

$$\mathbf{f}_j = \tilde{\mathbf{B}}_z'\left(\tilde{\mathbf{B}}_z\tilde{\mathbf{B}}_z' + \tilde{\boldsymbol{\Psi}}_z\right)^{-1}\mathbf{z}_j, \quad j = 1,2,...,n \tag{8.51}$$

gde smo sa $\tilde{\mathbf{B}}_z$ i $\tilde{\boldsymbol{\Psi}}_z$ označili ocene najveće verodostojnosti matrice faktorskih opterećenja i specifičnih varijansi dobijene na osnovu korelacione matrice, a sa $\mathbf{z}_j$ $j-$tu realizovanu vrednost standardizovane promenljive $\mathbf{Z}$.

Kao što je pokazano[2], do realizovanih vrednosti faktorskih skorova možemo doći jednostavnije, korišćenjem sledeće jednakosti:

$$\underset{(m\times p)}{\tilde{\mathbf{B}}'}\underset{(p\times p)}{\left(\tilde{\mathbf{B}}\tilde{\mathbf{B}}' + \tilde{\boldsymbol{\Psi}}\right)^{-1}} = \underset{(m\times m)}{\left(\mathbf{I} + \tilde{\mathbf{B}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\mathbf{B}}\right)^{-1}}\underset{(m\times p)}{\tilde{\mathbf{B}}'}\underset{(p\times p)}{\tilde{\boldsymbol{\Psi}}^{-1}} \tag{8.52}$$

Kažemo da je izraz s desne strane jednostavniji, jer zahteva izračunavanje inverzne matrice dimenzija $(m\times m)$, a ne $(p\times p)$ kao što je to slučaj sa inverznom matricom na levoj strani ove matrične jednakosti. Istovremeno ova matrična jednakost omogućava poređenje ocena faktorskih skorova dobijenih regresionom metodom i metodom ponderisanih najmanjih kvadrata.

Neki autori koriste direktno originalnu uzoračku kovarijacionu matricu $\mathbf{S}$ u oceni faktorskih skorova, umesto ocene kovarijacione matrice $\boldsymbol{\Sigma}$ na osnovu modela, tj. $\tilde{\mathbf{B}}\tilde{\mathbf{B}}' + \tilde{\boldsymbol{\Psi}}$. Ovime se želi smanjiti uticaj eventualno pogrešno određenog broja faktora. Tada je ocena faktorskih skorova regresionom metodom:

$$\mathbf{f}_j = \tilde{\mathbf{B}}'\mathbf{S}^{-1}\mathbf{x}_j, \quad j = 1,2,...,n \tag{8.53}$$

Ovaj izraz se na odgovarajući način modifikuje ako umesto uzoračke kovarijacione matrice $\mathbf{S}$ koristimo uzoračku korelacionu matricu $\mathbf{R}$. Sada ocena faktorskih skorova glasi:

$$\mathbf{f}_j = \tilde{\mathbf{B}}_z'\mathbf{R}^{-1}\mathbf{z}_j, \quad j = 1,2,...,n \tag{8.54}$$

---

[2] Videti: Morrison (1976, s. 335) ili Johnson i Wichern (1982, s. 434).

Naposletku, ako smo u izrazima za ocenu faktorskih skorova umesto nerotiranih koristili rotirane matrice faktorskih opterećenja, tada dobijamo odgovarajuće rotirane faktorske skorove. Relacija između rotiranih i nerotiranih faktorskih skorova uspostavlja se preko matrice transformacije, tako što se njenom transponovanom matricom množe s leve strane nerotirani faktorski skorovi, kao što je napomenuto kod direktnog izračunavanja faktorskih skorova.

8.4.3 Ocena faktorskih skorova metodom ponderisanih najmanjih kvadrata

Bartlett (1937) je predložio da se do ocene faktorskih skorova dođe minimiziranjem ponderisane sume kvadrata grešaka modela faktorske analize (kako ovde nazivamo specifične faktore). Pri tome su ponderi recipročne vrednosti varijansi grešaka. Koristi se metod ponderisanih najmanjih kvadrata, jer u opštem slučaju varijanse grešaka modela, $Var(\varepsilon_i) = \psi_i$, $i = 1, 2, ..., p$, nisu međusobom jednake. Bez umanjenja opštosti pretpostavimo da je u modelu faktorske analize $\boldsymbol{\mu} = \mathbf{0}$. Ponderisana suma kvadrata grešaka je

$$\sum_{i=1}^{p} \frac{\varepsilon_i^2}{\psi_i} = \boldsymbol{\varepsilon}' \boldsymbol{\Psi}^{-1} \boldsymbol{\varepsilon} = (\mathbf{x} - \mathbf{B}\mathbf{f})' \boldsymbol{\Psi}^{-1} (\mathbf{x} - \mathbf{B}\mathbf{f}) \tag{8.55}$$

Minimiziranjem ponderisane sume kvadrata grešaka[3] dobijamo ocenu faktorskih skorova. Zamenom matrica faktorskih opterećenja $\mathbf{B}$ i specifičnih varijansi $\boldsymbol{\Psi}$ njihovim ocenama po metodu najveće verodostojnosti, dobijamo izraz na osnovu koga ocenjujemo $j-$tu realizovanu vrednost faktora:

$$\mathbf{f}_j = \left(\tilde{\mathbf{B}}' \tilde{\boldsymbol{\Psi}}^{-1} \tilde{\mathbf{B}} \right)^{-1} \tilde{\mathbf{B}}' \tilde{\boldsymbol{\Psi}}^{-1} \mathbf{x}_j, \quad j = 1, 2, ..., n \tag{8.56}$$

Podsećamo da za ocene $\tilde{\mathbf{B}}$ i $\tilde{\boldsymbol{\Psi}}$ dobijene metodom najveće verodostojnosti važi uslov jednoznačnosti prema kome je $\tilde{\mathbf{B}}' \tilde{\boldsymbol{\Psi}}^{-1} \tilde{\mathbf{B}}$ dijagonalna matrica. To pojednostavljuje postupak ocene faktorskih skorova, jer je u gornjem izrazu potrebno odrediti inverznu matricu dijagonalne matrice. Dobijeni izraz za ocenu faktorskih skorova treba na adekvatan način modifikovati ukoliko koristimo standardizovane promenljive, odnosno vršimo razlaganje korelacione umesto kovarijacione matrice. Tada ocena faktorskih skorova glasi:

$$\mathbf{f}_j = \left(\tilde{\mathbf{B}}_z' \tilde{\boldsymbol{\Psi}}_z^{-1} \tilde{\mathbf{B}}_z \right)^{-1} \tilde{\mathbf{B}}_z' \tilde{\boldsymbol{\Psi}}_z^{-1} \mathbf{z}_j, \quad j = 1, 2, ..., n \tag{8.57}$$

Do rotiranih faktorskih skorova dolazimo na isti način kao što je to napomenuto kod direktnog izračunavanja faktorskih skorova i ocene faktorskih skorova regresionom metodom.

---

[3] O metodu ponderisanih najmanjih kvadrata kao specijalnom slučaju uopštenih najmanjih kvadrata videti kod Jovičić (1981, ss. 137-142).

Primer 8.9 Na osnovu podataka iz Primera 8.3 direktno izračunati faktorske skorove za dati vektor standardizovanih opservacija $\mathbf{z}' = \begin{bmatrix} -0.6 & -1.1 & 0.9 & 1.3 & -0.3 \end{bmatrix}$. Uporediti ih sa ocenjenim faktorskim skorovima koji su dobijeni na osnovu ocenjenih matrica faktorskih opterećenja i specifičnih varijansi metodom najveće verodostojnosti (Primer 8.4). Izračunati faktorske skorove za prvi i drugi faktor regresionom metodom (dve varijante) i metodom ponderisanih najmanjih kvadrata. Na osnovu ocenjenih faktorskih skorova izračunati rotirane faktorske skorove.

Relevantni podaci za direktno izračunavanje faktorskih skorova su:

$$\hat{\mathbf{\Lambda}}_z = \begin{bmatrix} 2.5835 & 0 \\ 0 & 1.7166 \end{bmatrix} \text{ i } \hat{\mathbf{A}}_z = \begin{bmatrix} 0.2775 & 0.4893 & 0.3426 & 0.5510 & 0.5124 \\ 0.6365 & -0.3508 & 0.5871 & -0.0451 & -0.3538 \end{bmatrix}$$

Direktno izračunati faktorski skorovi su:

$$\hat{\mathbf{f}} = \hat{\mathbf{\Lambda}}_z^{-1/2} \hat{\mathbf{A}}_z \mathbf{z} = \begin{bmatrix} 0.1034 \\ -0.4425 \end{bmatrix}$$

Ocenjene matrice faktorskih opterećenja i specifičnih varijansi metodom najveće verodostojnosti su:

$$\tilde{\mathbf{B}}_z = \begin{bmatrix} 0.1477 & 0.8478 \\ 0.8161 & -0.1325 \\ 0.2635 & 0.8977 \\ 0.7727 & 0.2189 \\ 0.9531 & -0.1677 \end{bmatrix}, \quad \tilde{\mathbf{\Psi}}_z = \begin{bmatrix} 0.2594 & 0 & 0 & 0 & 0 \\ & 9.3164 & 0 & 0 & 0 \\ & & 0.1247 & 0 & 0 \\ & & & 0.3550 & 0 \\ & & & & 0.0635 \end{bmatrix}$$

Prema prvoj varijanti regresionog metoda dobijamo:

$$\mathbf{f}^R = \tilde{\mathbf{B}}_z' \left(\tilde{\mathbf{B}}_z' \tilde{\mathbf{B}}_z + \tilde{\mathbf{\Psi}} \right)^{-1} \mathbf{z} = \begin{bmatrix} -0.1494 \\ 0.6049 \end{bmatrix}$$

Prema drugoj varijanti regresionog metoda dobijamo:

$$\mathbf{f}^R = \tilde{\mathbf{B}}_z' \mathbf{R}^{-1} \mathbf{z} = \begin{bmatrix} -0.1494 \\ 0.6049 \end{bmatrix}$$

U opštem slučaju ne dolazi do poklapanja ocena faktorskih skorova dobijenih na osnovu prve i druge varijante regresione metode. Do poklapanja u ovom slučaju došlo je zbog zanemarivo male greške u oceni

korelacione matrice na osnovu modela faktorske analize korišćenjem metode najveće verodostojnosti.

Ocena faktorskih skorova metodom ponderisanih najmanjih kvadrata je

$$\mathbf{f}^{NK} = \left(\tilde{\mathbf{B}}_z' \tilde{\mathbf{\Psi}}_z^{-1} \tilde{\mathbf{B}}_z\right)^{-1} \tilde{\mathbf{B}}_z' \tilde{\mathbf{\Psi}}_z^{-1} \mathbf{z} = \begin{bmatrix} -0.1574 \\ 0.6662 \end{bmatrix}$$

Prema dobijenom rezultatu obe metode, regresiona i ponderisanih najmanjih kvadrata, daju slične rezultate u pogledu ocenjenih vrednosti faktorskih skorova za datu standardizovanu opservaciju.

U Primeru 8.6 navedena je matrica transformacije na osnovu koje možemo preko ocenjenih nerotiranih faktorskih skorova doći do rotiranih faktorskih skorova. Ta matrica je:

$$\mathbf{T}_{\text{NV}} = \begin{bmatrix} 0.9885 & 0.1515 \\ -0.1515 & 0.9885 \end{bmatrix}$$

Korišćenjem ocenjenih faktorskih skorova regresionom metodom dobijamo ocenu rotiranih faktorskih skorova na osnovu relacije:

$$\mathbf{f}^{R*} = \mathbf{T}_{\text{NV}}' \tilde{\mathbf{f}}^{R} = \begin{bmatrix} -0.2393 \\ 0.5353 \end{bmatrix}$$

Slično dobijamo rotirane faktorske skorove ocenjene metodom ponderisanih najmanjih kvadrata:

$$\mathbf{f}^{NK*} = \mathbf{T}_{\text{NV}}' \tilde{\mathbf{f}}^{NK} = \begin{bmatrix} -0.2565 \\ 0.6347 \end{bmatrix}$$ ■

Izmedu ocena faktorskih skorova dobijenih regresionom metodom $\mathbf{f}^{R}$ i ocena faktorskih skorova dobijenih metodom ponderisanih najmanjih kvadrata $\mathbf{f}^{NK}$ možemo uspostaviti sledeću relaciju:

$$\mathbf{f}_j^{NK} = \left(\tilde{\mathbf{B}}'\tilde{\mathbf{\Psi}}^{-1}\tilde{\mathbf{B}}\right)^{-1}\left(\mathbf{I} + \tilde{\mathbf{B}}'\tilde{\mathbf{\Psi}}^{-1}\tilde{\mathbf{B}}\right)\tilde{\mathbf{f}}_j^{R} = \left(\mathbf{I} + \left(\tilde{\mathbf{B}}'\tilde{\mathbf{\Psi}}^{-1}\tilde{\mathbf{B}}\right)^{-1}\right)\tilde{\mathbf{f}}_j^{R} \qquad (8.58)$$

Ukoliko su elementi matrice $\tilde{\mathbf{B}}'\tilde{\mathbf{\Psi}}^{-1}\tilde{\mathbf{B}}$ znatno veći od jedinice, tada će ove dve metode ocena faktorskih skorova dati približno istu vrednost faktorskih skorova.

9 ANALIZA GRUPISANJA

Metod multivarijacione analize koji se koristi za grupisanje objekata u grupe, tako da su objekti unutar grupe sličniji međusobom, a između grupa znatno različiti, naziva se *analiza grupisanja*.[1]

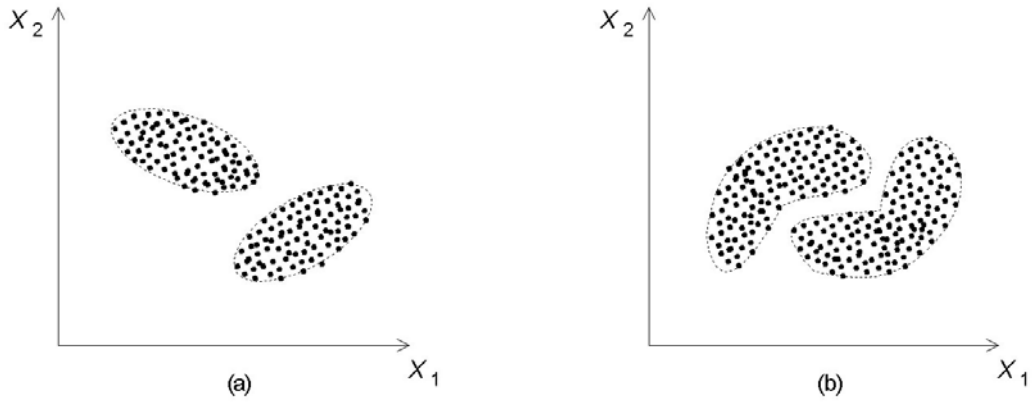
Da bi odgovorila ovom svom osnovnom zadatku analiza grupisanja zahteva definisanje *mere bliskosti* dva objekta na osnovu njihovih karakteristika. Zato se u prvom narednom poglavlju, nakon uvoda, bavimo različitim merama bliskosti. Definišemo mere sličnosti i razlike objekata i razmatramo njihove osobine ukazujući na specifičnosti svake mere. Na bazi mera bliskosti razvijeni su brojni postupci grupisanja objekata. Sve njih možemo svrstati u dve grupe: *hijerarhijski* i *nehijerarhijski metodi*. U osnovi hijerarhijskih metoda leži iterativan proces spajanja objekata u grupe tako da u narednoj etapi spajamo objekte i prethodno formirane grupe. Znači da se jednom formirane grupe samo proširuju novim objektima prema specifičnostima izabranog kriterijuma, a da ne postoji mogućnost prelaska objekata iz jednom formirane grupe u neku drugu, tokom iterativnog formiranja grupa. Mogućnost prelaska objekata iz jedne u drugu grupu dozvoljavaju tzv. nehijerarhijski metodi grupisanja. Ove dve grupe metoda grupisanja razmatramo u trećem i četvrtom poglavlju. Naposletku, ukazujemo na mogućnosti korišćenja analize grupisanja zajedno sa drugim multivarijacionim metodama.

---

[1] U literaturi se pod različitim nazivima izlažu metodi koje možemo svrstati u analizu grupisanja. Najčešće se koristi naziv klaster analiza (eng. *cluster analysis*), pa se čak i u domaćoj literaturi odomaćio ovaj naziv, mada nije u duhu srpskog jezika. U skladu sa ovim nazivom metode, grupe se nazivaju klasteri, a postupak grupisanja - klasterisanje. Koriste se i nazivi, odn. u analizu grupisanja svrstavaju se i Q-analiza (eng. *Q-analysis*), tipologija (eng. *typology*), analiza klasifikacije (eng. *classification analysis*), prepoznavanje obrazaca (eng. *pattern recognition*) i numerička taksonomija (eng. *numerical taxonomy*).

9.1 UVOD

Osnovni zadatak analize grupisanja jeste nalaženje "prirodnog" grupisanja skupa objekata ili osoba. Grupisanje objekata u grupe je zasnovano na različitim karakteristikama koje merimo kod svakog objekta. Ako smo na primer merili dve karakteristike kod svakog objekta, tada se možemo poslužiti dijagramom rasturanja u cilju određivanja grupa objekata. Ovaj pristup smo koristili u Poglavlju 1.3, gde se na dijagramu rasturanja (Slika 1.1) jasno izdvajaju tri grupe. Na tom dijagramu rasturanja uočavamo ono što bismo, nešto preciznije nego na početku Uvoda, označili kao osnovni zadatak analize grupisanja: alokacija skupa objekata u skup međusobno isključivih i dovoljnih grupa tako da su objekti unutar grupe slični međusobom (tačke u prostoru su bliže jedna drugoj), a objekti u različitim grupama različiti (tačke u prostoru su na većoj razdaljini). S obzirom da ne postoji opšte prihvaćena definicija grupe, grafički prikaz u vidu dijagrama rasturanja koristimo i da bismo objasnili šta podrazumevamo pod izrazom "prirodnog grupisanja", odnosno "prirodnih grupa". Imajući pred sobom dijagrame rasturanja sa Slike 9.1, opisujemo prirodne grupe ili kraće grupe, kao oblasti u dvodimenzionom prostoru sa velikom gustinom tačaka koje su razdvojene od drugih oblasti, oblastima koje imaju malu gustinu tačaka.



Slika 9.1 Prirodne grupe

Navedena definicija prirodnih grupa dozvoljava pojavu različitih oblika grupa, kao što je to pokazano na Slici 9.1. Alternativna definicija prirodnih grupa zasniva se na kriterijumu bliskosti, pa se prema njemu smatra da objekti u grupi treba da budu bliži jedni drugima, nego objektima u drugim grupama. Grupe prikazane na Slici 9.1(a) i 9.1(b) možemo jasno vizuelno međusobom razdvojiti. Međutim, dok u slučaju sferičnih grupa sa Slike 9.1(a) formalne metode grupisanja nemaju problema pri identifikaciji grupa, u slučaju sa Slike 9.1(b) većina postupaka grupisanja zasnovanih na kriterijumu bliskosti imaće poteškoće da to učine.

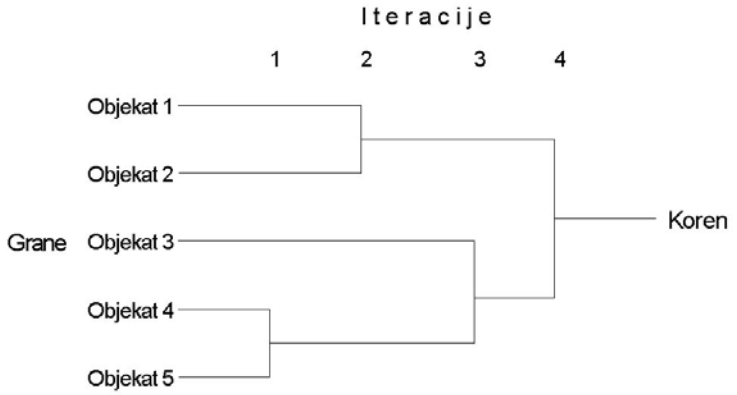
Ukoliko smo kod svakog objekta merili više od dve karakteristike, tada se u svrhe njihovog grupisanja možemo poslužiti grafičkim prikazima koji su izloženi u Poglavlju 1.3. Pokazano je da se na osnovu grafičkih prikaza u obliku zvezde (ili sunčevih zraka),

Andrewovih krivi i Chernoffovih lica veoma uspešno vrši grupisanje objekata u grupe koje su unutar sebe homogene, a međusobom su heterogene.

Pored grafičkih metoda kod kojih se subjektivnom procenom formiraju grupe, definisani su i analitički postupci na osnovu kojih se prema skupu formalnih pravila vrši grupisanje objekata u grupe. Polaznu osnovu različitih metoda grupisanja predstavljaju podaci uređeni u matricu podataka sa n redova (objekata) i p kolona (promenljivih). Elementi u $j-$toj koloni predstavljaju vrednosti $j-$te karakteristike koju različiti objekti uzimaju. Elementi u $i-$tom redu odnose se na različite karakteristike $i-$tog objekta i formiraju njegov *profil*. U 5. Glavi izraz profil koristili smo da označimo elemente sredine slučajnog vektora, a ovde ga formiramo za svaki pojedinačni objekat.

U analizi grupisanja na osnovu $(n\times p)$ matrice podataka formiramo $(n\times n)$ *matricu bliskosti* čiji elementi mere stepen sličnosti ili razlike između svih parova profila iz matrice podataka. Označimo tu matricu bliskosti sa $\mathbf{P}$, čiji su elementi p_{rs}, $r,s=1,2,...,n$. Element predstavlja meru bliskosti između $r-$tog i $s-$tog objekta. Kao što je to opisano u Poglavlju 9.2 definisan je relativno veliki broj mera bliskosti. Zato je u konkretnom istraživanju potrebno veoma pažljivo definisati šta ustvari podrazumevamo pod bliskošću objekata, da bismo potom izabrali odgovarajuću meru bliskosti.

Nakon formiranja matrice bliskosti u narednoj etapi analize grupisanja vršimo izbor metode grupisanja. Metodi grupisanja predstavljaju skup pravila pridruživanja objekata u grupe na osnovu mere bliskosti između objekata. Problem izbora metode grupisanja sličan je problemu izbora mere bliskosti. Naime, u literaturi je predložen veliki broj metoda grupisanja od kojih, u konkretnom istraživanju, treba izabrati onaj koji je primeren proučavanom problemu. Najčešće su korišćeni hijerarhijske metode grupisanja kod kojih se u svakoj iteraciji objekti pridružuju prethodno formiranim grupama ili sa drugim objektom formiraju novu grupu. Svaka naredna iteracija objedinjuje postojeće grupe sve dok se ne kompletira hijerarhijska struktura datog skupa objekata. Ovu strukturu nazivamo *hijerarhijsko drvo*, a grafički smo je prikazali na Slici 9.2, na primeru grupisanja pet objekata.



Slika 9.2 Hijerarhijsko drvo

Kao što smo to objasnili u Odeljku 9.4.1, postoje dva načina formiranja hijerarhijske strukture. Prvi način je *udruživanjem*, kada se pri formiranju grupa krećemo od grana drveta ka njegovom korenu (na Slici 9.2, s leva na desno). Drugi način formiranja hijerarhijske strukture jeste *deobom*, kada se krećemo u obrnutom smeru, od korena ka granama. Jednu grupa kod korena (sadrži sve objekte) delimo sve dok ne formiramo grupe sa po jednim objektom u svakoj grupi.

Kod nekih istraživanja cilj je formiranje kompletne hijerarhijske strukture, dok je cilj kod drugih dobijanje samo jednog grupisanja. U ovom, drugom slučaju, jednostavno ćemo "preseći" hijerarhijsko drvo, dobijajući na taj način jedno rešenje analize grupisanja. Pri tom "sečenju" drveta možemo se poslužiti formalnim kriterijumom grupisanja na osnovu koga ćemo izvršiti izbor broja grupa. Kriterijumom izbora broja grupa bavimo se u Odeljku 9.4.3.

U opštem slučaju kažemo da su ciljevi analize grupisanja:

a) **Istraživanje podataka**. Često ne znamo kako je skup objekata strukturiran, pa analizom grupisanja "otkrivamo" nepoznatu strukturu. Na primer, u marketingu korišćenjem metoda grupisanja odgovaramo na pitanje kako izvršiti segmentaciju tržišta, odnosno kako identifikovati grupe (ljudi, tržišta, organizacija) koje poseduju određene zajedničke karakteristike (stavove, sklonost ka potrošnji, navike u praćenju pojedinih sredstava informisanja, itd.).

b) **Redukcija podataka**. Na primer, u analizi regionalnog razvoja prikupljamo podatke o različitim aspektima dostignutog stepena socioekonomske razvijenosti nekoliko stotina opština. Interes istraživača je u formiranju grupa opština sličnog profila razvijenosti. Na osnovu dobijenog grupisanja opština u svakoj grupi se potom može sugerisati određeni model razvojne politike koji u najvećoj mogućoj meri vodi ka usklađenom regionalnom razvoju. Ovde pretpostavljamo da postoji nekoliko grupa opština različitog nivoa razvijenosti sa različitim privrednim strukturama, gde je težište na primarnim, sekundarnim ili tercijarnim delatnostima.

c) **Generisanje hipoteza**. Analiza grupisanja podataka nepoznate strukture rezultira u grupama čiji nam broj i sastav može pomoći da definišemo hipotezu o strukturi podataka. Tako broj grupa sugerisan prvobitnom analizom može biti hipoteza koja bi se testirala novim skupom podataka.

d) **Predviđanje**. Grupe dobijene analizom grupisanja možemo koristiti u kasnijim istraživanjima u svrhe predviđanja.

Naposletku, ukažimo na sličnosti i razlike analize grupisanja i drugih metoda multivarijacione analize. Kao što smo napomenuli, osnovni zadatak analize grupisanja jeste podela skupa objekata na grupe, tako da su varijacije između grupa znatno veće od varijacija unutar grupa. Zadatak je na prvi pogled sličan zadatku koji se postavlja pred diskriminacionu analizu, kada se ova koristi kao sredstvo za klasifikaciju objekata.

Međutim, dok su kod diskriminacione analize grupe unapred poznate, to kod analize grupisanja nije slučaj. Ovde samo pretpostavljamo da objekti pripadaju jednoj od "prirodnih" grupa ili jednostavno želimo izvršiti grupisanje objekata u izvestan manji broj grupa.

Grupisanje objekata u manji broj grupa sugeriše da se analiza grupisanja, slično metodi glavnih komponenata i faktorskoj analizi, može tretirati i kao metoda za redukciju podataka. No, za razliku od potonje dve, analiza grupisanja vrši redukciju podataka s obzirom na broj objekata, a ne s obzirom na broj promenljivih.

9.2 MERE SLIČNOSTI I RAZLIKE IZMEĐU OBJEKATA

Polazna osnova metode grupisanja predstavlja mera bliskosti između objekata ili promenljivih. Mera bliskosti najčešće iskazuje međusobne razlike između dva objekta, kada nam je zadatak grupisanje objekata. Tada mera bliskosti meri stepen međusobnog rastojanja, tj. kažemo da za grupisanje objekata koristimo *mere odstojanja* među objektima. Primer mere odstojanja predstavlja Euklidsko odstojanje između dva objekta. Za meru bliskosti p_{rs} kažemo da predstavlja meru razlike objekata r i s ako zadovoljava sledeće:

1. $p_{rs} > 0$, ako se objekti r i s razlikuju, a $p_{rs} = 0$, samo ako su objekti identični (uslov ne-negativnosti),
2. $p_{rs} = p_{sr}$ (uslov simetričnosti) i
3. $p_{rs} \le p_{rq} + p_{qs}$, za sve objekte r, s i q (uslov triangularnosti).

Mera bliskosti najčešće meri međusobnu sličnost između promenljivih, kada nam je zadatak grupisanje promenljivih. Tada mera bliskosti meri stepen međusobne sličnosti, pa kažemo da za grupisanje promenljivih koristimo *mere sličnosti* među promenljivama. Mere sličnosti možemo koristiti i prilikom grupisanja objekata, ali se tada praktikuje da se na osnovu međusobne veze mere odstojanje i mere sličnosti ova poslednja preračuna u prvu. Primer mere sličnosti primenjene na grupisanje objekata predstavlja koeficijent korelacije između objekata na osnovu p parova opservacija. Za meru bliskosti p_{rs} kažemo da predstavlja meru sličnosti objekata r i s ako zadovoljava sledeće:

1. $0 \le p_{rs} \le 1$, za sve objekte r i s (uslov normiranosti),
2. $p_{rs} = 1$, samo ako su objekti r i s identični i
3. $p_{rs} = p_{sr}$ (uslov simetričnosti).

Pearsonov koeficijent korelacije uzima vrednosti iz intervala $[-1,+1]$. Da bi se mogao koristiti kao mera sličnosti postupa se na dva načina. Ili se koristi njegova apsolutna vrednost, ili se koeficijentu doda jedinica, pa se dobijena vrednost podeli sa 2, da bi se ispunio uslov normiranosti.

9.2.1 Mere sličnosti i razlike na bazi kvantitativnih promenljivih

Među merama razlike (odstojanja) najpoznatija je tzv. Euklidska mera odstojanja na bazi kvantitativnih promenljivih. Prikazujući u $p-$dimenzionom prostoru realizovane vrednosti slučajnog uzorka, definisali smo u Odeljku 2.2.2, na osnovu $r-$tog i $s-$tog reda matrice podataka $\mathbf{x}_r' = [x_{r1}, x_{r2}, ..., x_{rp}]$ i $\mathbf{x}_s' = [x_{s1}, x_{s2}, ..., x_{sp}]$, *kvadrat Euklidskog odstojanja*:

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s) = \sum_{j=1}^{p} (x_{rj} - x_{sj})^2 \tag{9.1}$$

Mera odstojanja d_{rs}^2 predstavlja opšti element $(n \times n)$ *matrice kvadrata Euklidskog odstojanja* $\mathbf{D}$, ili kraće, *matrice odstojanja*, koja se formira na osnovu n redova date $(n \times p)$ matrice podataka $\mathbf{X}$. Može se pokazati da će Euklidsko odstojanje između r i s ostati neizmenjeno ako se izvrši centriranje promenljivih.

Euklidsko odstojanje predstavlja specijalan slučaj tzv. *odstojanja Minkowskog* koje je dato izrazom

$$M_{rs} = \left[\sum_{j=1}^{p} \left| x_{rj} - x_{sj} \right|^{\lambda} \right]^{1/\lambda} \tag{9.2}$$

Za $\lambda = 2$ odstojanje Minkowskog svodi se na Euklidsko odstojanje. Izdvajamo još jedan tip odstojanja koji definišemo na osnovu odstojanja Minkowskog. To je tzv. *Manhattan odstojanje* ili, kako se još naziva: *odstojanje tipa "gradskog bloka"*[2], koje se dobija za vrednost $\lambda = 1$. U opštem slučaju što je veća vrednost λ veći naglasak se daje razlici koordinata odnosne promenljive pri formiranju vrednosti odstojanja. To znači da će kod Euklidskog odstojanja veći udeo imati promenljiva kod koje se koordinate dva objekta više razlikuju, nego što je to slučaj kod Manhattan odstojanja. Stoga će ovo potonje odstojanje biti manje osetljivo na prisustvo nestandardnih opservacija.

Primer 9.1 Na osnovu podataka iz donje tabele za pet objekata, izračunati matricu kvadrata Euklidskih odstojanja kao i matricu odstojanja na bazi Manhattan odstojanja.

Kvadrat Euklidskog odstojanja između prvog i drugog objekta je

$$d_{12}^2 = (x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 = (5-2)^2 + (4-4)^2 = 9$$

Preostala odstojanja prikazana su u matrici kvadrata Euklidskih odstojanja

[2] Naziv odstojanja "gradski blok" (eng. *city-block*) potiče od načina računanja odstojanja u gradu (na Manhattnu) gde su ulice paralelne ili se seku pod pravim uglom. Odstojanje između dve tačke u gradu meri se ne pravolinijski (vazdušnom linijom), nego se određuje tako što se krećemo po ulicama, dakle izlomljenom putanjom.

$$\mathbf{D} = \begin{bmatrix} 0 & 9 & 2 & 8 & 50 \\ & 0 & 5 & 29 & 89 \\ & & 0 & 18 & 72 \\ & & & 0 & 18 \\ & & & & 0 \end{bmatrix}$$

Redni broj objekta	Naziv objekta	Promenljive	
		X_1	X_2
1	A_r	5	4
2	B_r	2	4
3	C_r	4	3
4	A_s	7	6
5	B_s	10	9

Manhattan odstojanje između prvog i drugog objekta je

$$|x_{11} - x_{21}| + |x_{12} - x_{22}| = |5-2| + |4-4| = 3$$

Preostala odstojanja prikazana su u matrici Manhattan odstojanja

$$\begin{bmatrix} 0 & 3 & 2 & 4 & 10 \\ & 0 & 3 & 7 & 13 \\ & & 0 & 6 & 12 \\ & & & 0 & 6 \\ & & & & 0 \end{bmatrix}$$

■

Odstojanje koje vodi računa i o kovarijacionoj strukturi podataka jeste Mahalanobisovo odstojanje, pa se stoga naziva i multivarijaciona mera odstojanja. Svojstvo Mahalanobisovog odstojanja jeste da eliminiše efekat korelisanosti promenljivih. Ukoliko je u analizi upravo ovaj efekat korelisanosti promenljivih bitan za razlikovanje objekata, tada se prilikom grupisanja objekata Mahalanobisovo odstojanje ređe koristi u odnosu na Euklidsko.

Primer 9.2 Na osnovu podataka iz Tabele 1.2 formirati matricu kvadrata Euklidskih odstojanja između deset osoba: korišćenjem prve dve promenljive i korišćenjem svih pet promenljivih.

U tabeli su date matrice kvadrata Euklidskih odstojanja tako što je na gornjem trouglu iznad glavne dijagonale sa nulama data matrica odstojanja računata na osnovu prve dve promenljive, a na donjem trouglu

ispod glavne dijagonale matrica odstojanja računata na osnovu svih pet promenljivih:

	Stojan	Vladan	Marko	Zorica	Miodrag	Danica	Ljubomir	Bojan	Dubravka	Milena
Stojan	0	0.65	0.85	1.00	1.04	0.64	0.04	0.85	0.89	0.08
Vladan	1.74	0	1.94	0.17	2.21	1.93	0.97	0.04	1.48	1.17
Marko	1.47	3.33	0	3.05	0.01	0.05	0.97	2.50	0.10	0.65
Zorica	2.29	0.23	4.30	0	3.40	2.92	1.28	0.05	2.57	1.64
Miodrag	1.79	3.55	0.06	4.62	0	0.08	1.16	2.81	0.13	0.80
Danica	1.65	2.99	0.18	3.76	0.18	0	0.68	2.45	0.25	0.40
Ljubomir	0.09	1.67	1.46	2.12	1.74	1.40	0	1.17	1.13	0.04
Bojan	2.56	0.14	4.47	0.23	4.61	3.91	2.39	0	2.00	1.45
Dubravka	1.55	3.13	0.12	4.10	0.18	0.46	1.70	4.25	0	0.85
Milena	0.18	2.78	1.19	3.37	1.57	1.45	0.25	3.86	1.41	0

■

Alternativni pristup merenju bliskosti objekata bazira se na merama sličnosti. Posmatrajući dve tačke (objekta) r i s u $p-$dimenzionom prostoru, koristimo veličinu ugla između dva $(p\times 1)$ vektora $\mathbf{x}_r$ i $\mathbf{x}_s$ da bismo izmerili stepen sličnosti između objekata čiji su to vektori. Što je taj ugao manji, objekti r i s su sličniji međusobom. Kao meru sličnosti koristimo kosinus ugla koji zaklapaju dva vektora. Slično razmatranju u Odeljku 2.2.2 kosinus ugla između vektora $\mathbf{x}_r$ i $\mathbf{x}_s$ je

$$c_{rs}=\frac{\sum_{j=1}^{p}x_{rj}x_{sj}}{\sqrt{\sum_{j=1}^{p}x_{rj}^2\sum_{j=1}^{p}x_{sj}^2}}. \tag{9.3}$$

Kako je u gornjem izrazu kvadrat dužine vektora $\sum_{j=1}^{p}x_{rj}^2$ i $\sum_{j=1}^{p}x_{sj}^2$, to znači da mera sličnosti c_{rs} ne zavisi od dužine dva vektora, pa proporcionalne promene u koordinatama vektora $\mathbf{x}_r$ i/ili $\mathbf{x}_s$ neće promeniti vrednost c_{rs}. Mera sličnosti c_{rs} naziva se *kosinusni koeficijent* ili *koeficijent podudarnosti*.

Primer 9.3 Na osnovu podataka iz Primera 9.1 formirati matricu sličnosti čiji su elementi kosinusni koeficijenti.

Kosinusni koeficijent između prvog i drugog objekta je

$$c_{rs}=\frac{x_{11}x_{21}+x_{12}x_{22}}{\sqrt{(x_{11}^2+x_{12}^2)(x_{21}^2+x_{22}^2)}}=\frac{5\cdot 2+4\cdot 4}{\sqrt{(5^2+4^2)(2^2+4^2)}}=\frac{26}{\sqrt{820}}=0.908$$

Preostali kosinusni koeficijenti prikazani su u matrici sličnosti

$$\begin{bmatrix} 1 & 0.908 & 0.9995 & 0.9994 & 0.9983 \\ & 1 & 0.8944 & 0.9216 & 0.9308 \\ & & 1 & 0.9979 & 0.9960 \\ & & & 1 & 0.9997 \\ & & & & 1 \end{bmatrix}$$

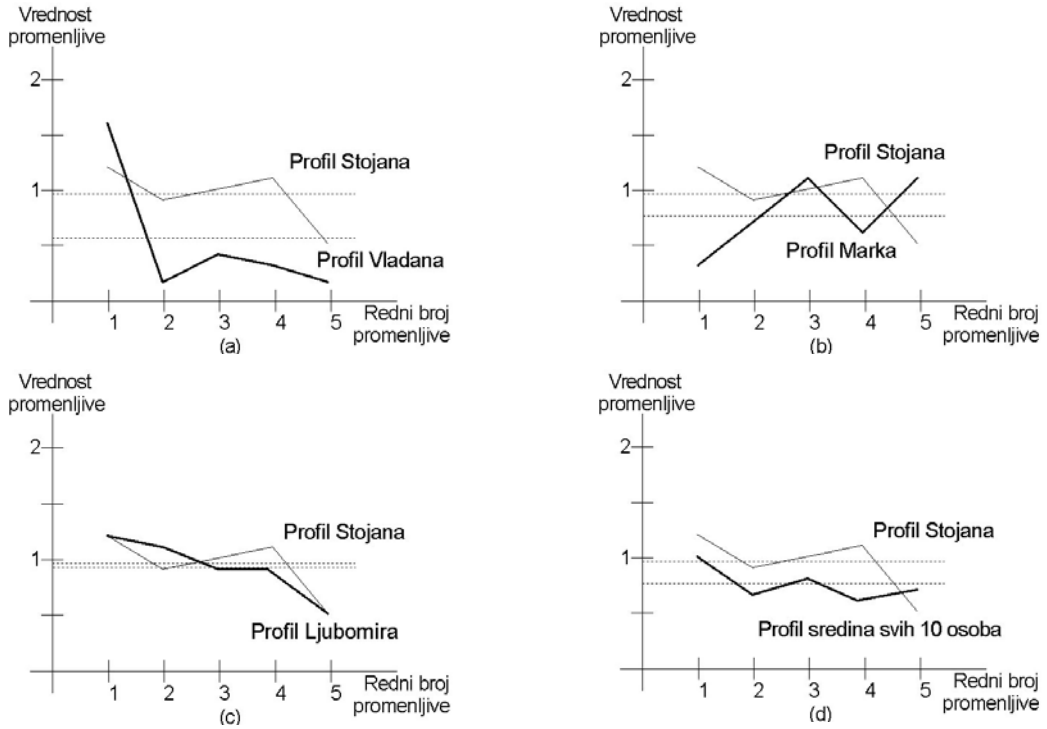
■

Druga mera sličnosti polazi od transponovane matrice podataka u kojoj su objekti zamenili mesta sa promenljivama. Dakle, imamo $(p \times n)$ matricu podataka na osnovu koje računamo $(n \times n)$ korelacionu matricu čiji su elementi koeficijenti korelacije između dva profila. Ova korelaciona matrica objekata predstavlja polaznu osnovu ne samo različitih metoda analize grupisanja, nego se koristi i u tzv. *Q faktorskoj analizi* za grupisanje objekata u različite, međusobno razdvojene grupe.

Pored mere odstojanja kao sumarnog pokazatelja međusobne razlike dva profila, za analizu te razlike koristimo *grafik profila*. Grafički prikaz profila moguće je dati ako su promenljive istovrsne ili je prethodno izvršena njihova standardizacija.

Primer 9.4 Korišćenjem podataka iz Tabele 1.2 formirati grafik profila za sledeće osobe: Stojan, Vladan, Marko i Ljubomir, kao i za sredinu profila svih deset osoba.

Preglednosti radi svaki par profila posebno smo prikazali na Slici 9.3.



Slika 9.3 Poređenje profila

Poređenjem profila za Stojana i ostale osobe uočavamo da je najveće razlika po pojedinim promenljivama između Stojana i Vladana, zatim Stojana i Marka, a najmanja razlika između profila je kod Stojana i Ljubomira. Do kvadrata Euklidskog odstojanja između Stojana i svake osobe dolazimo na osnovu para grafika profila tako što sabiramo kvadrate razlika vrednosti svake promenljive. Isprekidanom linijom prikazali smo sredinu profila. Na sve četiri slike sredina profila za Stojana ima veću vrednost od sredine profila Vladana, Marka, Ljubomira i sredine svih 10 osoba. ■

Pri poređenju dva profila možemo porediti njihove *nivoe*, stepen *varijabiliteta*, a takođe i njihov *oblik*. Štaviše, kvadrat Euklidskog odstojanja možemo razložiti na tri komponente od kojih svaka ukazuje na doprinos: međusobne razlike u nivou, međusobne razlike u varijabilitetu i oblika profila ukupnom odstojanju. Da bismo izvršili spomenuto razlaganje kvadrata Euklidskog odstojanja uvedimo mere za nivo, varijabilitet i oblik profila. Nivo profila se definiše na osnovu sredine p opservacija. Tako za $r-$ti profil imamo sredinu: $\overline{x}_{r\cdot} = \sum_{j=1}^{p} x_{rj} / p$. Varijabilitet profila merimo sa $v_r^2 = \sum_{j=1}^{p} (x_{rj} - \overline{x}_{r\cdot})^2$, tj. sumom kvadrata odstupanja profila od svoje sredine. Slično određujemo sredinu i varijabilitet za $s-$ti profil. Naposletku, oblik jednog profila merimo relativno u odnosu na drugi profil računanjem koeficijenta korelacije q_{rs} između ta dva profila. Što je q_{rs} bliže 1, dva profila teže da pokažu isti oblik. Kvadrat Euklidskog odstojanja između dva profila d_{rs}^2 razlažemo prema sledećem izrazu

$$d_{rs}^2 = p(\overline{x}_{r\cdot} - \overline{x}_{s\cdot})^2 + (v_r - v_s)^2 + 2x_r v_s (1 - q_{rs}) \tag{9.4}$$

gde prvi sabirak ukazuje na doprinos razlike u nivou dva profila, drugi sabirak ukazuje na doprinos razlike u varijabilitetu dva profila i treći sabirak ukazuje na doprinos razlici u obliku dva profila.

Primer 9.5 Na osnovu podataka iz Tabele 1.2 i izračunate matrice odstojanja na bazi svih pet promenljivih (donji trougao ispod glavne dijagonale u tabeli u okviru Primera 9.2) razložiti kvadrat Euklidskog odstojanja od Stojana, do Vladana, Marka, Ljubomira i sredine profila svih deset osoba.

Prema podacima iz Tabele 1.2 računamo sredinu profila svih deset osoba:

Nivo profila Stojana je

$$\overline{x}_{\text{Stojan}\cdot} = \frac{1}{p}\sum_{j=1}^{p} x_{\text{Stojan},j} = \frac{1}{5}(1.2 + 0.9 + 1.0 + 1.1 + 0.5) = 0.94$$

Varijabilitet profila Stojana je

$$v^2_{\text{Stojan}} = \sum_{j=1}^{p} (x_{\text{Stojan},j} - \overline{x}_{\text{Stojan}.})^2 = (1.2-0.94)^2 + ... + (0.5-0.94)^2 = 0.292$$

Kvadratni koren iz varijabiliteta profila Stojana je 0.5404. Svi potrebni elementi za računanje zahtevanog razlaganja kvadrata Euklidskog odstojanja između Stojana i ostalih osoba dati su u narednoj tabeli:

Osoba	$\overline{x}_{s.}$	v_s	q_{rs}	d^2_{rs}	Doprinos (relativni doprinos)		
					Nivoa	Varijabiliteta	Oblika
Vladan	0.540	1.1966	0.606	1.74	0.800	0.431	0.510
				(100%)	(45.97%)	(24.75%)	(29.28%)
Marko	0.760	0.6870	-0.733	1.47	0.162	0.022	1.287
				(100%)	(11.02%)	(1.46%)	(87.52%)
Ljubomir	0.920	0.5367	0.848	0.09	0.002	0.0000138	0.088
				(100%)	(2.22%)	(0.02%)	(97.76%)
Sredina 10 osoba	0.752	0.3342	0.424	0.43	0.177	0.043	0.208
				(100%)	(41.36%)	(9.95%)	(48.69%)

Na osnovu rezultata prezentiranih u tabeli zaključujemo da odstojanju između Stojana i Vladana najveći doprinos daju razlike u nivou dva profila (45.97%), a da kod ostalih osoba dominira doprinos oblika profila (najveći udeo je kod para profila Stojan-Ljubomir i iznosi 97.76%). U odnosu na ova dva izvora razlike između dva profila, doprinos razlike varijabiliteta je zanemariv, osim kod para profila Stojan-Vladan. ■

Meru sličnosti možemo konstruisati na osnovu mere odstojanja. Neka je izračunato Euklidsko odstojanje d_{rs} između dva objekta. Korišćenjem mere odstojanja definišemo meru sličnosti objekata r i s

$$p_{rs} = \frac{1}{1+d_{rs}} \tag{9.5}$$

gde je očigledno $0 \le p_{rs} \le 1$, s obzirom na to da je $d_{rs} \ge 0$.

U suprotnom slučaju imamo ograničenje, jer meru odstojanja možemo konstruisati na osnovu mere sličnosti pod uslovom da je matrica sličnosti nenegativno definitna. Tako se naprimer može uspostaviti relacija između kvadrata Euklidskog odstojanja d^2_{rs} i kosinusnog koeficijenta c_{rs}

$$d^2_{rs} = d^2_r + d^2_s - 2d_r d_s c_{rs} \tag{9.6}$$

gde d^2_r i d^2_s označavaju kvadrate dužine vektora $\mathbf{x}_r$ i $\mathbf{x}_s$. Ako su ovi vektori normirani ($d^2_r = d^2_s = 1$), tada je $d^2_{rs} = 2(1-c_{rs})$.

9.2.2 Mere sličnosti i razlike na bazi kvalitativnih promenljivih

Ako smo promenljive merili na nominalnoj skali, a ne intervalnoj kao u prethodnom Odeljku, tada poređenje objekata, odnosno meru sličnosti zasnivamo na prisustvu ili odsustvu određenih karakteristika. Pretpostavimo da su svih p promenljivih kvalitativnog tipa, pri čemu je broj modaliteta svake promenljive $k_1, k_2, ..., k_p$ respektivno. Tako na primer $j-$ta promenljiva *bračni status* ima četiri modaliteta: neoženjen/neudata, oženjen/udata, razveden/razvedena i udovac/udovica ($k_j = 4$). Neka su kolone matrice podataka $\mathbf{X}$ formirane tako da sadrže veštačke promenljive za svaki modalitet p promenljivih. To znači da ćemo u matrici podataka $\mathbf{X}$ imati ukupno $K = \sum_{j=1}^{p} k_j$ veštačkih promenljivih, pa će matrica podataka biće dimenzija $(n \times K)$. Za svaki profil imaćemo ukupno p jedinica i $(K - p)$ nula, a svaka kvalitativna promenljiva imaće jednu jedinicu po profilu. Na bazi matrice $\mathbf{X}'\mathbf{X}$ formiramo meru sličnosti između profila. Dijagonalni elementi ove matrice su svi jednaki p, a vandijagonalni elementi su jednaki f_{rs}, gde je f_{rs} broj kvalitativnih promenljivih kod kojih posmatrana dva reda sadrže isti kvalitet (1-1 u oba reda).

Sa ovakvom se strukturom matrice $\mathbf{X}'\mathbf{X}$ kosinusni koeficijent, kao mera sličnosti između $r-$tog i $s-$tog reda, svodi na

$$c_{rs} = \frac{\sum_{j=1}^{p} x_{rj} x_{sj}}{\sqrt{\sum_{j=1}^{p} x_{rj}^2 \sum_{j=1}^{p} x_{sj}^2}} = \frac{f_{rs}}{p} \tag{9.7}$$

što predstavlja proporciju pojave para 1-1 u dva reda od ukupnog broja promenljivih. S obzirom na strukturu matrice $\mathbf{X}$ imamo da je suma elemenata svakog reda i suma kvadrata elemenata svakog reda jednaka p/K i $p/K(1-p/K)$ respektivno. Stoga je koeficijent korelacije između $r-$tog i $s-$tog reda matrice $\mathbf{X}$

$$q_{rs} = \frac{\frac{f_{rs}}{K} - \left(\frac{p}{K}\right)^2}{\frac{p}{K}\left(1 - \frac{p}{K}\right)} \tag{9.8}$$

Naposletku, pošto su sredine i varijanse svakog reda jednake, kvadrat Euklidskog odstojanja je

$$d_{rs}^2 = 2K(1 - q_{rs}) = 2K\left(\frac{p}{K} - \frac{f_{rs}}{K}\right) = 2(p - f_{rs}) \tag{9.9}$$

Primer 9.7 U istraživanju tržišta gotovih jela formiran je anketni upitnik prema kome su ispitanici trebali da odgovore, odnosno upišu:

1) pol,

2) koliko često troše gotova jela i

3) bračni status.

Ovde se očigledno radi o tri kvalitativne promenljive. Prva kvalitativna promenljiva, pol, ima dva modaliteta (muško/žensko). Druga kvalitativna promenljiva ima, na primer, tri modaliteta: često troši, ponekad troši i nikad ne troši gotova jela. Treća kvalitativna promenljiva, bračni status, ima četiri modaliteta (neoženjen/neudata, oženjen/udata, razveden/razvedena i udovac/ udovica). Na osnovu ove tri kvalitativne promenljive, sa 2, 3 i 4 modaliteta respektivno, možemo formirati trodimenzionu tabelu kontingenciju dimenzija $2\times3\times4$, sa ukupno 24 ćelije. Znači da postoje 24 različita tipa ispitanika koji se mogu javiti među n anketiranih osoba. Matrica podataka $\mathbf{X}$ sadrži $K=2+3+4=9$ veštačkih promenljivih, pa je u svakom redu po 3 jedinice i 6 nula. Svaka od te tri jedinice pripada jednoj od 3 kvalitativne promenljive. Neka su, na primer, kod prva tri ispitanika dobijeni sledeći odgovori:

1. ispitanik: muško, često troši gotova jela i neoženjen je,

2. ispitanik: žensko, ponekad troši gotova jela i udata je,

3. ispitanik: muško, ponekad troši gotova jela i neoženjen je

Na osnovu njihovih odgovora vrednosti devet veštačkih promenljivih u prva tri reda matrice $\mathbf{X}$ su:

1. ispitanik:	1	0	1	0	0	1	0	0	0
2. ispitanik:	0	1	0	1	0	0	1	0	0
3. ispitanik:	1	0	0	1	0	1	0	0	0

Linijom posle druge i pete kolone razdvojili smo veštačke promenljive koje pripadaju prvoj i drugoj kvalitativnoj promenljivoj. Pri poređenju ma koja dva reda u matrici podataka $\mathbf{X}$ mogu se javiti 3, 2, 1 ili 0 parova jedinica (1-1). Ako postoje tri para jedinica, tada su ta dva reda identična. U našem primeru kod prva dva ispitanika ne postoji nijedan par jedinica, što znači da se ispitanici potpuno razlikuju međusobom. Kod prvog i trećeg ispitanika imamo dva para jedinica (prva i šesta veštačka promenljiva), a kod drugog i trećeg ispitanika jedan par jedinica (četvrta veštačka promenljiva). Kako je $p=3$ i $f_{rs}=3$, 2, 1 ili 0, to kosinusni koeficijent uzima vrednost 1, 2/3, 1/3 ili 0, koeficijent korelacije uzima vrednosti 1, 1/2, 0 ili -1/2, a kvadrat Euklidskog odstojanja uzima vrednosti 0, 2, 4 ili 6. U našem primeru imamo sledeće vrednosti mera sličnosti i odstojanja:

(r, s)	Mere sličnosti i odstojanja		
	c_{rs}	q_{rs}	d_{rs}^2
(1,2)	0	0	6
(1,3)	2/3	1/2	2
(2,3)	1/3	0	4

■

Specijalan slučaj kvalitativnih promenljivih javlja se kada svaka od p promenljivih ima samo dva modaliteta (uzima vrednost 1 kada objekt ima datu karakteristiku, 0 kada je nema). Ove promenljive nazivamo *binarne promenljive*. Za takve promenljive definisane su alternativne mere sličnosti između objekata. Polazna osnova tih mera predstavlja sledeća (2×2) tabela kontingencije:

		Objekat s		Ukupno
		1	0	
Objekat r	1	a	b	$a+b$
	0	c	d	$c+d$
Ukupno		$a+c$	$b+d$	$p=a+b+c+d$

U ovoj tabeli a predstavlja broj promenljivih kod kojih oba objekta imaju posmatrane osobine (oba uzimaju vrednost 1), a d predstavlja broj promenljivih kod kojih oba objekta nemaju posmatrane osobine (oba uzimaju vrednost 0). Zbir $(b+c)$ predstavlja broj promenljivih kod kojih objekti uzimaju različite vrednosti.

Među brojnim merama sličnosti i odstojanja definisanih na osnovu elemenata gornje tabele kontingencije izdvajamo dva najpoznatija. Prvi je tzv. *jednostavni koeficijent podudarnosti* (eng. *simple matching coefficient*) Sokala i Michenera, prema kome je mera sličnosti: $(a+d)/(a+b+c+d)=(a+d)/p$. On daje isti ponder paru 0-0 kao i paru 1-1. Drugi je tzv. *Jaccardov koeficijent* prema kome je mera sličnosti: $a/(a+b+c)$, što znači da on isključuje broj promenljivih kod kojih oba objekta uzimaju vrednost 0. Njegovo korišćenje je opravdano kada želimo utvrditi sličnost objekta s obzirom na prisustvo svojstva (vrednost 1), a zanemariti uticaj odsustva svojstva (vrednost 0). Ova dva i ostali koeficijenti sličnosti dati su u Tabeli 9.1.

Tabela 9.1 Koeficijenti sličnosti

1.	$\dfrac{a}{a+b+c+d}$		5.	$\dfrac{a+d}{a+d+2(b+c)}$	Rogers i Tanimoto
2.	$\dfrac{a+d}{a+b+c+d}$	Sokal i Michener	6.	$\dfrac{(a+d)-(b+c)}{a+b+c+d}$	Hamann
3.	$\dfrac{a}{a+b+c}$	Jaccard	7.	$\dfrac{a}{[(a+b)(c+d)]^{1/2}}$	Ochiai
4.	$\dfrac{2(a+d)}{2(a+d)+b+c}$	Sokal i Sneath	8.	$\dfrac{ad-bc}{[(a+b)(a+c)(b+d)(c+d)]^{1/2}}$	fi koeficijent

Izbor koeficijenta ne predstavlja trivijalan zadatak, tim pre što neki od koeficijenata nemaju svojstvo monotonosti. To znači da ako sve vrednosti koeficijenata sličnosti, računatih na osnovu jedne formule za svaki par objekata, uredimo po veličini, dobijeni redosled ne mora se poklapati sa redosledom koji proističe na osnovu korišćenja druge formule za koeficijent sličnosti. Takođe je kod ovih koeficijenata bitno da se vrednost 1, a ne 0, dodeli onom svojstvu koje je za nas važnije. Može se pokazati da na osnovu iste tabele kontingencije, gde su vrednosti 1 i 0 zamenili mesta, dobijamo različite vrednosti koeficijenata sličnosti.

Primer 9.8 Pretpostavimo da smo u nekom istraživanju dobili za dva ispitanika sledeće vrednosti $p = 8$ binarnih promenljivih:

	Promenljive							
	1	2	3	4	5	6	7	8
Osoba A:	0	1	0	0	1	1	0	0
Osoba B:	1	1	0	0	1	0	1	0

Na osnovu datih podataka formirati (2×2) tabelu kontingencije, a zatim izračunati vrednosti svih osam koeficijenata sličnosti iz Tabele 9.1.

(2×2) tabela kontingencije je

		Osoba B		**Ukupno**
		1	0	
Osoba A	1	2	1	3
	0	2	3	5
Ukupno		4	4	8

Na osnovu nje vrednosti koeficijenata sličnosti iz Tabele 9.1 su:

1. $\dfrac{a}{a+b+c+d} = \dfrac{2}{8} 0.25$
2. $\dfrac{a+d}{a+b+c+d} = \dfrac{5}{8} = 0.625$
3. $\dfrac{a}{a+b+c} = \dfrac{2}{5} = 0.4$
4. $\dfrac{2(a+d)}{2(a+d)+b+c} = \dfrac{10}{13} = 0.769$
5. $\dfrac{a+d}{a+d+2(b+c)} = \dfrac{5}{11} = 0.455$
6. $\dfrac{(a+d)-(b+c)}{a+b+c+d} = \dfrac{2}{8} = 0.25$
7. $\dfrac{a}{[(a+b)(c+d)]^{1/2}} = \dfrac{2}{\sqrt{12}} = 0.577$
8. $\dfrac{ad-bc}{[(a+b)(a+c)(b+d)(c+d)]^{1/2}} = 0.258$

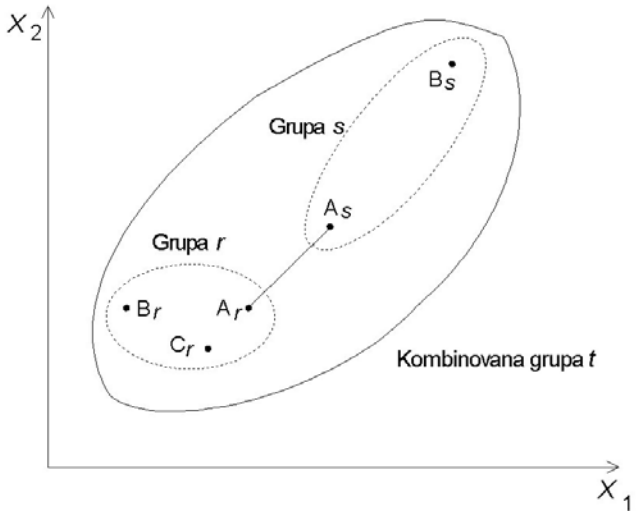
Može se uočiti da vrednosti koeficijenata sličnosti variraju od 0.25 do 0.769. ■

Na kraju poglavlja posvećenog merama sličnosti i razlike, ukažimo na rešenja koja su predložena u slučaju da matrica podataka $\mathbf{X}$ predstavlja mešavinu kvantitativnih i kvalitativnih promenljivih. Tada nijedan od predloženih koeficijenata nije moguće

direktno primeniti. Zato je prva sugestija da se izvrši standardizacija promenljivih pre računanja koeficijenata sličnosti i razlike. Prema drugoj sugestiji, odvojeno se računaju koeficijenti sličnosti za podskup kvantitativnih i podskup kvalitativnih promenljivih, a zatim se ta dva koeficijenta kombinuju u jedan korišćenjem odgovarajućih pondera. Treći pristup polazi od prevođenja kvantitativne u kvalitativnu promenljivu. Definišu se klasni intervali vrednosti kvantitativne promenljive, pa se zatim sve promenljive tretiraju kao kvalitativne. Na primer, promenljiva kao što je godina starosti može se "pretvoriti" u kvalitativnu tako što ćemo definisati tri klasna intervala: do 20 god., od 20 do 65 god. i preko 65 god. Time dobijamo kvalitativnu promenljivu sa tri modaliteta.

9.3 MERE SLIČNOSTI I RAZLIKE IZMEĐU GRUPA

Metodi analize grupisanja razlikuju se prema tome kako mere sličnost ili razliku između grupa. Otuda se i naziv metoda grupisanja poklapa sa nazivom mera bliskosti između grupa. Među brojnim merama izdvajamo pet najpoznatijih: prve tri su u osnovi tzv. *metodi povezivanja* (eng. *linkage methods*) u analizi grupisanja, četvrti je u osnovi *metode centroida*, a peti u osnovi metode *minimalne sume kvadrata* ili *Wardove metode*. U okviru mera, odnosno metodi povezivanja detaljnije ćemo se pozabaviti sledećim merama: *jednostruko* (eng. *single linkage*), *potpuno* (eng. *complete linkage*) i *prosečno povezivanje*[3] (eng. *average linkage*).



Slika 9.4 Jednostruka mera odstojanja između dve grupe

Ideja koja leži u osnovi različitih mera, odnosno metodi povezivanja, najbolje se može razumeti na osnovu grafičkog prikaza problema. Korišćenjem podataka iz Primera 9.1 formirali smo dijagram rasturanja i prezentirali ga na Slici 9.4. Na njoj je prikazan način određivanja odstojanja između dve grupe u skladu sa jednostrukim povezivanjem.

[3] Alternativni nazivi mera, odnosno metoda jednostrukog i potpunog povezivanja su: *metod najbližeg suseda* (eng. *nearest-neighbor method*) i *metod najudaljenijeg suseda* (eng. *further-neighbor method*) respektivno. Ovi nazivi metoda potiču od načina utvrđivanja međusobne bliskosti između dve grupe.

Prema ovom načinu povezivanja, odstojanje između grupa meri se na osnovu najmanjeg odstojanja parova objekata iz posmatrane dve grupe. Na slici je to slučaj sa odstojanjem između objekata A_r i A_s, pa upravo odstojanje između njih reprezentuje odstojanje između $r-$te i $s-$te grupe. Prema potpunom povezivanju, odstojanje između dve grupe određuje se na osnovu najvećeg odstojanja između parova objekata iz dve grupe. Na slici je to slučaj sa odstojanjem između objekata B_r i B_s, pa prema ovoj metodi povezivanja, to odstojanje reprezentuje odstojanje između $r-$te i $s-$te grupe. Naposletku, prema prosečnom povezivanju, odstojanje između dve grupe određuje se na osnovu prosečnog odstojanja svih parova objekata iz dve posmatrane grupe.

Pored navedenih mera odstojanja između grupa, definisane su i mere bazirane na odstojanju između centroida grupa. Da bismo izložili ideju koja je u njihovoj osnovi, pretpostavimo da dve grupe objekata r i s sadrže n_r i n_s objekata respektivno. Označimo opservacije p promenljivih za n_r objekata u $r-$toj grupi sa x_{rjm}, $j=1,2,...,p$, $m=1,2,...,n_r$. Slično imamo za opservacije u $s-$toj grupi, x_{sjm}, $j=1,2,...,p$, $m=1,2,...,n_s$. Ako centroide $r-$te grupe označimo sa $\overline{\mathbf{x}}'_r=[\overline{x}_{r1.},\overline{x}_{r2.},...,\overline{x}_{rp.}]$, a centroid $s-$te grupe sa $\overline{\mathbf{x}}'_s=[\overline{x}_{s1.},\overline{x}_{s2.},...,\overline{x}_{sp.}]$, tada prva mera međusobnog odstojanja ove dve grupe predstavlja kvadrat Euklidskog odstojanja između centroida

$$d_{rs}^2=\sum_{j=1}^{p}(\overline{x}_{rj.}-\overline{x}_{sj.})^2 \tag{9.10}$$

Ova mera leži u osnovi tzv. *metode centroida* za udruživanje grupa. Druga mera odstojanja polazi od činjenice da postoji ukupno $(n_r n_s)$ odstojanja između dve grupe. Mera ukupnog odstojanja između dve grupe je sada $n_r n_s d_{rs}^2$, a prosečno odstojanje je $n_r n_s d_{rs}^2/(n_r+n_s)$, pošto ima ukupno (n_r+n_s) objekata. Može se pokazati da je ova mera odstojanja između grupa ekvivalentna promeni u sumi kvadrata unutar grupa do koje je došlo zbog udruživanja $r-$te i $s-$te grupe.

Za $r-$tu grupu je suma kvadrata odstupanja opservacija od svoje sredine, odnosno suma kvadrata unutar grupe

$$SKW_r=\sum_{m=1}^{n_r}\sum_{j=1}^{p}(x_{rjm}-\overline{x}_{rj.})^2 . \tag{9.11}$$

Slično imamo za sumu kvadrata unutar $s-$te grupe

$$SKW_s=\sum_{m=1}^{n_s}\sum_{j=1}^{p}(x_{sjm}-\overline{x}_{sj.})^2 . \tag{9.12}$$

Udruživanjem ove dve grupe dobijamo kombinovanu grupu t (vidi Sliku 9.4). Možemo posmatrati odstupanja opservacija grupe t od novog centroida $\overline{\mathbf{x}}'_t=[\overline{x}_{t1.},\overline{x}_{t2.},...,\overline{x}_{tp.}]$ i izračunati novu sumu kvadrata unutar te grupe

$$SKW_t = \sum_{m=1}^{n_r+n_s} \sum_{j=1}^{p} (x_{tjm} - \overline{x}_{tj\bullet})^2 . \tag{9.13}$$

Povećanje ukupne sume kvadrata unutar grupe kao rezultat udruživanja $r-$te i $s-$te grupe dato je izrazom: $SKW_t - (SKW_r + SKW_s)$ i ekvivalentno je prosečnom odstojanju između grupa: $n_r n_s d_{rs}^2 / (n_r + n_s)$.

Do ove relacije ekvivalencije možemo doći uspostavljajući vezu između analize varijanse i određivanja odstojanja između grupa. Ako ovo izlaganje smestimo u okvir analize varijanse, tada se ukupna suma kvadrata unutar kombinovane grupe t (u oznaci SKW_t) može posmatrati kao ukupna suma kvadrata u analizi varijanse. Poznato je da se u analizi varijanse ukupna suma kvadrata razlaže na dva dela: sumu kvadrata unutar grupa i sumu kvadrata između grupa. Suma kvadrata unutar grupa je $SKW_r + SKW$, a do sume kvadrata između grupa SKB_t dolazimo na osnovu razlike ukupne sume kvadrata i sume kvadrata unutar grupa, ili direktno

$$SKB_t = \sum_{j=1}^{p} \left[n_r (\overline{x}_{rj\bullet} - \overline{x}_{tj\bullet})^2 + n_s (\overline{x}_{sj\bullet} - \overline{x}_{tj\bullet})^2 \right]$$

$$= \frac{n_r n_s}{(n_r + n_s)} \sum_{j=1}^{p} (\overline{x}_{rj\bullet} - \overline{x}_{sj\bullet})^2 = \frac{n_r n_s}{(n_r + n_s)} d_{rs}^2 \tag{9.14}$$

Pokazano je dakle, da je druga mera odstojanja između grupa ekvivalentna sumi kvadrata između grupa, odnosno priraštaju u sumi kvadrata unutar grupa do koga je došlo udruživanjem $r-$te i $s-$te grupe. Ova druga mera odstojanja u osnovi je Wardove metode hijerarhijskog udruživanja.

Primer 9.9 Korišćenjem podataka i rezultata iz Primera 9.1 odrediti mere odstojanja između grupa koje su u osnovi:

a) jednostrukog povezivanja
b) potpunog povezivanja
c) prosečnog povezivanja
d) metode centroida
e) Wardove metode.

Na osnovu izračunate matrice kvadrata Euklidskog odstojanja u Primeru 9.1 i Slike 9.4 na kojoj su prikazane $r-$ta i $s-$ta grupa, imamo

a) $d_{rs}^2 = \min(8, 50, 29, 89, 18, 72) = 8$, što predstavlja odstojanje između objekata A_r i A_s.

b) $d_{rs}^2 = \max(8, 50, 29, 89, 18, 72) = 89$, što predstavlja odstojanje između objekata B_r i B_s.

c) $d_{rs}^2 = (8+50+29+89+18+72)/6 = 44.33$.

d) Kako su elementi centroida $r-$te grupe: $\overline{x}_{r1\bullet} = \overline{x}_{r2\bullet} = 3.66$, a elementi centroida $s-$te grupe: $\overline{x}_{s1\bullet} = \overline{x}_{s2\bullet} = 7.5$, to je

$$d_{rs}^2 = (3.66-8.5)^2 + (3.66-7.5)^2 = 38.055$$

e) Odredimo sume kvadrata unutar $r-$te, $s-$te i $t-$te grupe:

$$SKW_r = (5-3.66)^2 + (2-3.66)^2 + (4-3.66)^2 + (4-3.66)^2 + (4-3.66)^2 + (3-3.66)^2 = 5.33$$

$$SKW_s = (7-8.5)^2 + (7-8.5)^2 + (10-8.5)^2 + (6-8.5)^2 + (9-8.5)^2 = 9$$

$$SKW_t = (5-5.6)^2 + (2-5.6)^2 + (4-5.6)^2 + (7-5.6)^2 + (10-5.6)^2$$

$$+(4-5.6)^2 + (4-5.6)^2 + (3-5.6)^2 + (6-5.6)^2 + (9-5.6)^2 = 60$$

Priraštaj sume kvadrata unutar grupa je

$$SKB_t = SKW_t - (SKW_r + SKW_s) = 60 - (9+5.33) = 45.66.$$

Dobijena vrednost predstavlja meru odstojanja između dve grupe. Do nje smo mogli doći na osnovu izračunatog odstojanja između centroida: $d_{rs}^2 = 38.055$. Naime,

$$\frac{n_r n_s}{(n_r+n_s)} d_{rs}^2 = \frac{3\cdot 2}{(3+2)} 38.055 = 45.66.$$

■

Posle formiranja nove grupe potrebno je izračunati odstojanja novoformirane grupe i ostalih grupa. U tu svrhu može poslužiti sledeći izraz na osnovu koga izračunavamo elemente nove matrice odstojanja

$$d_{tu}^2 = \alpha_r d_{ru}^2 + \alpha_s d_{su}^2 + \beta d_{rs}^2 + \gamma \left| d_{ru}^2 - d_{su}^2 \right| \qquad (9.15)$$

gde smo sa t označili novoformiranu grupu, u je jedna od ostalih grupa (različita od r i s), a α_r, α_s, β i γ su koeficijenti koji zavise od toga koji se metod udruživanja koristi. U izrazu za reviziju elemenata matrice odstojanja koristili smo kvadrat Euklidskog odstojanja. Umesto njega možemo koristiti neku drugu meru odstojanja. Međutim, metod centroida i Wardov metod zahtevaju da se koristi kvadrat Euklidskog odstojanja.

U zavisnosti od korišćene mere odstojanja između grupa, odnosno metoda udruživanja imamo sledeće vrednosti parametara:

1. jednostruko povezivanje: $\alpha_r = \alpha_s = \frac{1}{2}, \beta = 0$ i $\gamma = -\frac{1}{2}$,
2. potpuno povezivanje: $\alpha_r = \alpha_s = \frac{1}{2}, \beta = 0$ i $\gamma = \frac{1}{2}$,
3. prosečno povezivanje: $\alpha_r = n_r/(n_r + n_s)$, $\alpha_s = n_s/(n_r + n_s)$, $\beta = \gamma = 0$,
4. metod centroida: $\alpha_r = n_r/(n_r + n_s)$, $\alpha_s = n_s/(n_r + n_s)$, $\beta = -n_r n_s/(n_r + n_s)^2$, i $\gamma = 0$,
5. Wardov metod: $\alpha_r = (n_r + n_u)/(n_t + n_u)$, $\alpha_s = (n_s + n_u)/(n_t + n_u)$, $\beta = -n_u/(n_t + n_u)$ i $\gamma = 0$.

Pored revizije elemenata matrice odstojanja na osnovu navedenog izraza mogu se definisati i detaljnije izučiti dodatne osobine mera odstojanja.

9.4 METODI GRUPISANJA

9.4.1 Hijerarhijski metodi grupisanja

Nakon izbora odgovarajuće mere sličnosti ili razlike između objekata, vrši se izbor metode grupisanja. Hijerarhijski metodi grupisanja mogu se svrstati u dve grupe prema tome da li su zasnovane na iterativnom spajanju ili deljenju grupa i objekata. Prva grupa hijerarhijskih metoda polazi od individualnih objekata koje, prema određenom kriterijumu, udružuje u grupe. U narednim koracima formiraju se nove grupe udruživanjem ranije formiranih grupa ili individualnih objekata. Pri tome, u narednim iteracijama, jedanput udružene grupe ostaju zajedno, odnosno u višim etapama nema mogućnosti prelaska objekta iz jedne u drugu grupu. Ovaj hijerarhijski postupak leži u osnovi tzv. *hijerarhijskih metodi udruživanja*. Naziv metoda proizilazi iz iterativnog postupka tokom koga, polazeći od n grupa (objekata), formiramo jednu grupu. Znači da se iz iteracije u iteraciju veličina grupa povećava, a smanjuje njihov broj. Postupak hijerarhijskog udruživanja započinjemo grupisanjem prema kome imamo n grupa sa po jednim objektom. Narednim iteracijama u potpunosti je opisan postupak hijerarhijskog udruživanja:

1. Na osnovu matrice odstojanja biramo dve najbliže grupe i u ovom koraku udružujemo ih u novu grupu. Neka su $r-$ta i $s-$ta grupa udružene u novu grupu koju ćemo označiti sa t.
2. Ponovo izračunavamo matricu odstojanja tako što određujemo odstojanje ostalih grupa i novoformirane grupe. Znači da smo isključili $r-$ti i $s-$ti red, odnosno kolonu matrice odstojanja, a dodali red i kolonu koji sadrže odstojanja grupe t i preostalih grupa.
3. Prethodna dva koraka se ponavljaju $(n-1)$ put sve dok se ne formira jedna grupa čime se završava ovaj iterativan postupak.

Druga grupa metoda tzv. *hijerarhijski metodi* deobe prelaze isti put, ali se kreću u suprotnom smeru. Naime, polazeći od jedne grupe koja sadrži sve objekte, prema određenom kriterijumu, iz iste izdvajamo po jedan objekat ili grupu sve dok se ne formira onoliko grupa koliko ima individualnih objekata. Ovim metodima nećemo se detaljnije baviti.

Najpopularniji metodi grupisanja pripadaju hijerarhijskim metodima udruživanja, a među njima se izdvajaju metodi povezivanja. Metodi hijerarhijskog udruživanja razlikuju se prema tome kako u drugoj fazi gornjeg iterativnog postupka određuju međusobnu bliskost grupa. U prethodnom poglavlju izložili smo nekoliko najpopularnijih mera odstojanja između grupa. U okviru ovog odeljka detaljnije ćemo se pozabaviti njihovim korišćenjem u analizi grupisanja.

Metod jednostrukog povezivanja

Pretpostavimo da smo u primeni ovog metoda pošli od matrice odstojanja. Među elementima ove matrice biramo onaj koji je najmanji i odgovarajuća dva objekta udružujemo u jednu grupu. Napominjemo da u prvoj iteraciji ostale grupe sadrže samo jedan objekat. Sada je potrebno odrediti odstojanje novoformirane grupe i ostalih objekata. Kao što smo u Poglavlju 9.3 definisali, mera odstojanja između dve grupe predstavlja najkraće odstojanje između para objekata koji pripadaju ovim grupama. Upravo su odstojanja formirane grupe i ostalih objekata jednaka najmanjem odstojanju jednog od dva objekta novoformirane grupe i svih ostalih objekata. U drugoj iteraciji u novoformiranoj matrici odstojanja biramo najmanji element. Može se desiti da su neka druga dva objekta bliža međusobom ili da je jedan objekat bliži ranije formiranoj grupi. U prvom slučaju formira se nova grupa, a u drugom, objekat pridružujemo u ranije formiranu grupu koja sada sadrži tri objekta. Iterativan postupak se nastavlja sve dok se svi objekti ne udruže u jednu grupu. Na sledećem primeru bolje će se sagledati postupak jednostrukog povezivanja kao i problemi koji se pri tome mogu javiti.

Primer 9.10 Na osnovu podataka iz Primera 9.1 formirati hijerarhijsku strukturu pet objekata korišćenjem metoda jednostrukog povezivanja.

Matricu odstojanja reprodukujemo iz Primera 9.1 kao matricu $\mathbf{D}^{(1)}$. Označimo $(r,s)-$ti element matrice odstojanja sa d_{rs}. Najmanji element u toj matrici je u prvom redu i trećoj koloni, $d_{13}=2$. Znači da u prvom koraku udružujemo 1. i 3. objekat u jednu grupu.

$$\mathbf{D}^{(1)}=\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 9 & (2) & 8 & 50 \\ & 0 & 5 & 29 & 89 \\ & & 0 & 18 & 72 \\ & & & 0 & 18 \\ & & & & 0 \end{bmatrix} \end{array}$$

Zatim određujemo odstojanje te grupe od ostalih objekata (2., 4. i 5. objekta):

$$d_{(13)2} = \min\{d_{12}, d_{23}\} = \min\{9,5\} = 5$$

$$d_{(13)4} = \min\{d_{14}, d_{34}\} = \min\{8,18\} = 8$$

$$d_{(13)5} = \min\{d_{15}, d_{35}\} = \min\{50,72\} = 50$$

Nova matrica odstojanja je

$$\mathbf{D}^{(2)} = \begin{matrix} (13) \\ 2 \\ 4 \\ 5 \end{matrix} \begin{bmatrix} 0 & (5) & 8 & 50 \\ & 0 & 29 & 89 \\ & & 0 & 72 \\ & & & 0 \end{bmatrix}$$

Kako je u matrici $\mathbf{D}^{(2)}$ najmanji element $d_{(13)2} = 5$, to znači da drugi objekat pripajamo grupi $(1,3)$. Sada određujemo odstojanje grupe $(1,2,3)$ od ostalih objekata (4. i 5. objekat):

$$d_{(123)4} = \min\{d_{(13)4}, d_{24}\} = \min\{8,29\} = 8$$

$$d_{(123)5} = \min\{d_{(13)5}, d_{25}\} = \min\{50,89\} = 50$$

Nova matrica odstojanja je

$$\mathbf{D}^{(3)} = \begin{matrix} (123) \\ 4 \\ 5 \end{matrix} \begin{bmatrix} 0 & (8) & 50 \\ & 0 & 18 \\ & & 0 \end{bmatrix}$$

Kako je u matrici $\mathbf{D}^{(3)}$ najmanji element $d_{(123)4} = 8$, to znači da četvrti objekat pripajamo grupi $(1,2,3)$. Sada određujemo odstojanje grupe $(1,2,3,4)$ od preostalog, 5. objekta:

$$d_{(1234)5} = \min\{d_{(123)5}, d_{45}\} = \min\{50,18\} = 18$$

Nova matrica odstojanja je

$$\mathbf{D}^{(4)} = \begin{matrix} (1234) \\ 5 \end{matrix} \begin{bmatrix} 0 & (18) \\ & 0 \end{bmatrix}$$

Preostaje samo da 5. objekat pripojimo grupi $(1,2,3,4)$, pa da se svi objekti nađu u jednoj grupi. Time se u 4. koraku završava postupak udruživanja objekata u grupe. ■

Metod jednostrukog povezivanja udružuje objekte na osnovu najkraćeg odstojanja između njih. S jedne strane primena ovog metoda na grupe koje su slabije razdvojene, odnosno kod kojih postoje veoma bliski objekti, izrokovaće udruživanje tih grupa tako što će se jedna na drugom nadovezivati kao karike u lancu. Taj efekat se u analizi grupisanja naziva *lančani efekat* (eng. *chaining*). Upravo u Primeru 9.10 se desio takav slučaj, što je bilo za očekivati s obzirom na međusobnu bliskost elemenata A_r i A_s iz dve grupe (videti Sliku 9.4). S druge strane ovaj metod povezivanja je jedan od nekoliko metodi grupisanja koji su u stanju da razdvoje grupe neelipsoidnog oblika.

Metod potpunog povezivanja

Koraci u primeni metoda potpunog povezivanja identični su onima kod metoda jednostrukog povezivanja. Razlika se javlja u delu koji se odnosi na način određivanja odstojanja između grupa. Kod metoda potpunog povezivanja odstojanje se određuje ne prema najkraćem, kao kod metoda jednostrukog povezivanja, nego prema najvećem odstojanju objekata koji pripadaju dvema grupama. Na sledećem primeru bolje će se sagledati postupak potpunog povezivanja kao i problemi koji se pri tome mogu javiti.

Primer 9.11 Na osnovu podataka iz Primera 9.1 formirati hijerarhijsku strukturu pet objekata korišćenjem metoda potpunog povezivanja.

Matrica odstojanja između pet objekata data je u prethodnom primeru. U prvom koraku udružujemo 1. i 3. objekat u jednu grupu pošto je odstojanje između njih najmanje. Zatim određujemo odstojanje te grupe od ostalih objekata (2., 4. i 5. objekta):

$$d_{(13)2} = \max\{d_{12}, d_{23}\} = \max\{9,5\} = 9$$

$$d_{(13)4} = \max\{d_{14}, d_{34}\} = \max\{8,18\} = 18$$

$$d_{(13)5} = \max\{d_{15}, d_{35}\} = \max\{50,72\} = 72$$

Nova matrica odstojanja je

$$\mathbf{D}^{(2)} = \begin{matrix} \\ (13) \\ 2 \\ 4 \\ 5 \end{matrix} \begin{bmatrix} 0 & (9) & 18 & 72 \\ & 0 & 29 & 89 \\ & & 0 & 18 \\ & & & 0 \end{bmatrix}$$

Kako je u matrici $\mathbf{D}^{(2)}$ najmanji element $d_{(13)2} = 9$, to znači da drugi objekat pripajamo grupi $(1,3)$. Sada određujemo odstojanje grupe $(1,2,3)$ od preostalih objekata (4. i 5. objekat):

$$d_{(123)4} = \max\{d_{(13)4}, d_{24}\} = \max\{8, 29\} = 29$$

$$d_{(123)5} = \max\{d_{(13)5}, d_{25}\} = \max\{50, 89\} = 89$$

Nova matrica odstojanja je

$$\mathbf{D}^{(3)} = \begin{matrix} \\ (123) \\ 4 \\ 5 \end{matrix} \begin{bmatrix} 0 & 29 & 89 \\ & 0 & (18) \\ & & 0 \end{bmatrix}$$

Kako je u matrici $\mathbf{D}^{(3)}$ najmanji element $d_{45} = 18$, to znači da četvrti i peti objekat formiraju novu grupu $(4,5)$. Sada određujemo odstojanje te grupe od ranije formirane grupe $(1,2,3)$:

$$d_{(123)(45)} = \max\{d_{(123)4}, d_{(123)5}\} = \max\{29, 89\} = 89$$

Nova matrica odstojanja je

$$\mathbf{D}^{(4)} = \begin{matrix} (123) \\ (45) \end{matrix} \begin{bmatrix} 0 & (89) \\ & 0 \end{bmatrix}$$

Preostaje samo da dve grupe, $(1,2,3)$ i $(4,5)$ spojimo u jednu, pa da se svi objekti nađu u toj jednoj grupi. Time se u 4. koraku završava postupak udruživanja objekata u grupe. ■

Metod jednostrukog povezivanja je u Primeru 9.10 na grupu formiranu u prvom, u svakom narednom koraku dodavao po jedan objekat, što je dovelo do lančanog efekta. Za razliku od njega, na istim podacima je metod potpunog povezivanja (Primer 9.11) formirao u 3. koraku dve kompaktne grupe kao što smo to prikazali na Slici 9.4. Ovo je rezultat korišćene mere odstojanja između grupa kod metoda potpunog povezivanja. Naime, dva najudaljenija objekta unutar te dve grupe predstavljaju meru odstojanja između grupa. Ovo će u primeni metode potpunog povezivanja voditi ka formiranju kompaktnih grupa, kao i u našem primeru.

U poređenju ova dva metoda povezivanja interesantno je istaći uticaj veličine grupa na meru odstojanja između grupa. Pretpostavimo da se povećava broj elemenata koje sadrže dve međusobno razdvojene grupe. Mera odstojanja između grupa kod metoda jednostrukog povezivanja neće se menjati, dok će se kod metoda potpunog povezivanja povećavati. Ako među podacima postoji nestandardna opservacija, mera odstojanja

između nje i grupe čija se veličina povećava kod metoda jednostrukog povezivanja biće fiksna, a kod potpunog povezivanja imaće tendenciju da se povećava. To znači da će ta nestandardna opservacija kod potpunog povezivanja veoma brzo postati bliska postojećoj grupi što kod jednostrukog povezivanja neće biti slučaj. Kod ovog drugog metoda povezivanja nestandardna opservacija teži da ostane izolovana tačka do poslednjih koraka u postupku formiranja hijerarhijske strukture objekata. Zato se kaže da metod jednostrukog povezivanja čuva, odnosno *štedi prostor* (eng. *space conserving*), a da metod potpunog povezivanja razređuje, odnosno *popunjava prostor* (eng. *space diluting*). Pošto oba metoda koriste odstojanje između dva objekta unutar različitih grupa kao meru odstojanja između grupa, razumljivo je zašto su oba veoma osetljiva na prisustvo nestandardnih opservacija u podacima.

Metod prosečnog povezivanja

Koraci u primeni metoda prosečnog povezivanja identični su onima kod prethodne dve metode povezivanja. Razlika se javlja u delu koji se odnosi na način određivanja odstojanja između grupa. Kod metoda prosečnog povezivanja odstojanje se određuju prema prosečnom odstojanju svih objekata koji pripadaju dvema grupama (po jedan objekat iz svake grupe). U Primeru 9.12 dati su rezultati povezivanja objekata korišćenjem metoda prosečnog povezivanja.

*
* *

Preostala dva metoda hijerarhijskog udruživanja su *metod centroida* i *metod minimalne sume kvadrata* ili *Wardov metod*. Kod metoda centroida dve grupe se udružuju u novu grupu ako su njihovi centroidi najmanje udaljeni međusobom u odnosu na međusobnu udaljenost svih mogućih parova grupa koje postoje na posmatranom nivou udruživanja. Kod Wardovog metoda dve grupe se spajaju u jednu, ako je njihovim udruživanjem došlo do najmanjeg povećanja sume kvadrata unutar grupa u odnosu na povećanje sume kvadrata do koga je došlo udruživanjem ma koje druge dve grupe na posmatranom nivou udruživanja. U Primeru 9.12 dati su rezultati povezivanja objekata korišćenjem metoda centroida i Wardovog metoda.

Primer 9.12 Na osnovu podataka iz Primera 9.1 formirati hijerarhijsku strukturu pet objekata korišćenjem metoda prosečnog povezivanja, metoda centroida i Wardovog metoda.

Rezultati analize grupisanja prezentirani su u donjoj tabeli:

Korak	Prosečno povezivanje		Metod centroida		Wardov metod	
	Grupe	Odstojanje	Grupe	Odstojanje	Grupe	Odstojanje
1	(1) (3)	2.00	(1) (3)	2.00	(1) (3)	1.00
2	(13) (2)	7.00	(13) (2)	6.50	(13) (2)	4.33
3	(4) (5)	18.00	(123) (4)	16.55	(4) (5)	9.00
4	(123) (45)	44.33	(1234) (5)	52.81	(123) (45)	45.67

Može se uočiti da metod jednostrukog povezivanja i metod centroida daju identičnu hijerarhijsku strukturu, ali se od koraka do koraka vrednosti odstojanja razlikuju. Sa druge strane, metod potpunog i prosečnog povezivanja i Wardov metod daju identičnu hijerarhijsku strukturu, ali je vrednost odstojanja između grupa u svakom koraku različita. ■

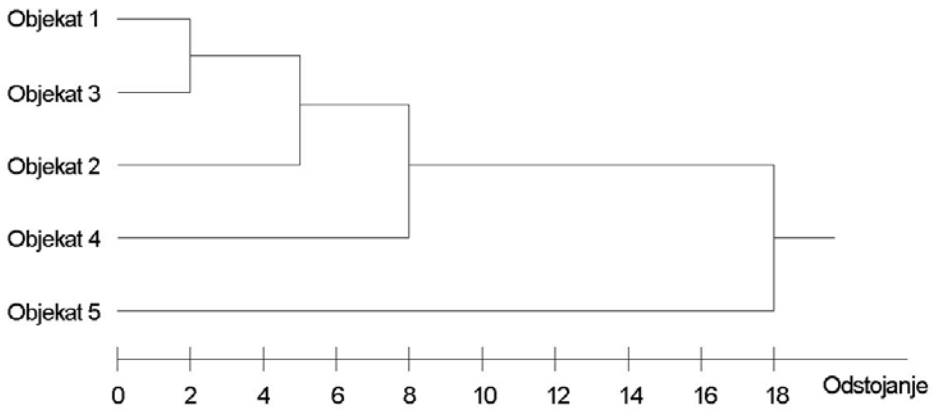
Metod prosečnog povezivanja, metod centroida i Wardov metod se češće koristi u odnosu na metod jednostrukog i potpunog povezivanja, jer su manje osetljivi na prisustvo nestandardnih opservacija. U zavisnosti od oblika grupa ova osobina može predstavljati i njihovu manu.

Međusobno poređenje metoda centroida i Wardovog metoda je olakšano pošto je odstojanje kod drugog metoda dobijeno na osnovu prvog, množenjem sa koeficijentom $n_r n_s /(n_r + n_s)$. Ako napišemo ovaj izraz u obliku $1/(1/n_r + 1/n_s)$, tada sa povećanjem veličine grupa raste i vrednost ovog koeficijenta, a isto se dešava ako n_r raste brže u odnosu na n_s. To znači da će Wardov metod davati veću vrednost mere odstojanja između grupa, ako njihova veličina raste ili ako postaje sve veća razlika u veličini tih grupa. Prema tome, u poređenju sa metodom centroida, Wardov metod iskazuje tendenciju da formira grupe jednake veličine i/ili grupe manje veličine. Dok metod centroida teži da udružuje grupe čiji su centroidi bliski međusobom, dotle Wardov metod teži da udružuje grupe manje veličine sa istim takvim grupama koje su znatno udaljenije od prvih. Zato se često dobija, kao što smo i mi dobili u Primeru 9.12, da metod centroida daje slične rezultate kao metod jednostrukog povezivanja, jer oba metoda teže da ispune prostor. S druge strane, Wardov metod slično metodu potpunog povezivanja iskazuje tendenciju ka štednji prostora formirajući kompaktnije grupe.

9.4.2 Dendrogram i izvedena mera sličnosti

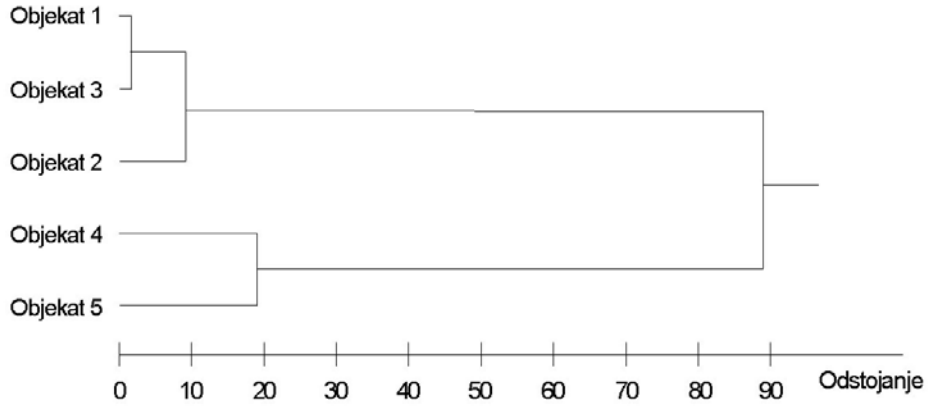
Kao što je napomenuto u Uvodu, hijerarhijsku strukturu objekata možemo grafički prikazati korišćenjem hijerarhijskog drveta. Ako uz hijerarhijsko drvo navedemo i skalu na kojoj su nanete vrednosti mere odstojanja u svakom koraku iterativnog postupka udruživanja grupa, tada dobijamo grafički prikaz hijerarhijske strukture objekata koji se naziva *dendrogram*.

Dendrogram metoda jednostrukog povezivanja na osnovu podataka iz Primera 9.10 prikazali smo na Slici 9.5. Sa njega možemo pročitati da je u prvom koraku formirana grupa koja sadrži prvi i treći objekat (međusobno odstojanje 2). Zatim se u drugom koraku toj grupi pridružuje drugi objekat (na odstojanju 5). Na odstojanju 8, odnosno 18, ovoj grupi se pridružuje četvrti i peti objekat. Dendrogram na Slici 9.5 ukazuje istovremeno na izgled dendrograma hijerarhijske strukture kod koje je prisutan lančani efekat u formiranju grupa.



Slika 9.5 Dendrogram jednostrukog povezivanja

Na Slici 9.6 prikazan je dendrogram metoda potpunog povezivanja. Dendrogrami za ostale metode udruživanja mogu se nacrtati na osnovu podataka iz tabele u okviru Primera 9.12.



Slika 9.6 Dendrogram potpunog povezivanja

Na osnovu dendrograma ili podataka iz tabele u okviru Primera 9.12 možemo formirati *izvedenu matricu odstojanja*. Do elemenata ove matrice dolazimo tako što svim parovima objekata iz dve različite grupe koje se udružuju u jednu, pripisujemo istu vrednost odstojanja, onu pri kojoj smo udružili dve grupe. Međusobnim poređenjem odgovarajućih elemenata originalne i izvedene matrice odstojanja može se utvrditi u kom stepenu formirane grupe predstavljaju dobro rešenje problema grupisanja. Jednostavna mera koja se u te svrhe koristi jeste običan koeficijent korelacije između originalnih i izvedenih mera odstojanja. Ovaj koeficijent naziva se *kofenetički koeficijent* (eng. *cophenetic coefficient*). Njegova vrednost bliska jedinici sugeriše da su formirane grupe objekata kvalitetno rešenje problema grupisanja.

Primer 9.13 Na osnovu rezultata grupisanja objekata u Primerima 9.10-9.12 formirati izvedenu matricu odstojanja za korišćene metode grupisanja. Izračunati odgovarajuće kofenetičke koeficijente.

Izvedene matrice odstojanja za metod centroida i Wardov metod su:

$$\begin{bmatrix} 0 & 6.5 & 2 & 16.55 & 52.81 \\ & 0 & 6.5 & 16.55 & 52.81 \\ & & 0 & 16.55 & 52.81 \\ & & & 0 & 52.81 \\ & & & & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 4.33 & 1 & 45.67 & 45.67 \\ & 0 & 4.33 & 45.67 & 45.67 \\ & & 0 & 45.67 & 45.67 \\ & & & 0 & 45.67 \\ & & & & 0 \end{bmatrix}$$

Čitaocu se ostavlja da formira preostale izvedene matrice odstojanja.

U narednoj tabeli izvršeno je poređenje originalnih i izvedenih mera odstojanja na osnovu primene različitih metoda grupisanja.

Par objekata	Originalno odstojanje	Izvedene mere odstojanja za različite metode grupisanja				
		Jednostruko povezivanje	Potpuno povezivanje	Prosečno povezivanje	Metod centroida	Wardov metod
(1,3)	2	2	2	2.00	2.00	1.00
(2,3)	5	8	9	7.00	6.50	4.33
(1,4)	8	8	89	44.33	16.55	45.66
(1,2)	9	5	9	7.00	6.50	4.33
(3,4)	18	8	89	44.33	16.55	45.66
(4,5)	18	18	18	18.00	52.81	9.00
(2,4)	29	8	89	44.33	16.55	45.66
(1,5)	50	18	89	44.33	52.81	45.66
(3.5)	72	18	89	44.33	52.81	45.66
(2,5)	89	18	89	44.33	52.81	45.66
Prosečno odstojanje	30.00	11.10	57.20	30.00	27.59	29.27
Standardna devijacija	28.75	5.91	39.11	17.94	21.10	20.17
Kofenetički koeficijent	1.00	0.78	0.62	0.62	0.79	0.62

Parovi objekata su poređani prema vrednosti originalne mere odstojanja od međusobno najbližeg para $(1,3)$ (odstojanje 2), do najudaljenijeg para $(2,5)$ (odstojanje 89).

Rezultati iz gornje tabele potvrđuju naša očekivanja da jednostruki i potpuni metod povezivanja podcenjuju, odnosno precenjuju originalno prosečno odstojanje. Metod prosečnog povezivanja i Wardov metod su u tom pogledu bliži originalnom prosečnom odstojanju. Ova dva poslednja metoda pokazuju i relativno mali stepen varijabiliteta meren standardnom devijacijom odgovarajućih izvedenih mera odstojanja. S druge strane na osnovu kofenetičkih koeficijenata možemo zaključiti da metod centroida i

jednostrukog povezivanja daju kvalitetnije rešenje problema grupisanja nego ostali metodi. Na nivou pojedinačnih vrednosti izvedenih mera sličnosti uočavamo da za par objekata $(1,4)$, Wardov metod i metodi potpunog i prosečnog povezivanja znatno precenjuju vrednost originalnog odstojanja. Oni sugerišu da se tek u poslednjem koraku ova dva objekta nađu u istoj grupi, mada bi na osnovu originalne vrednosti odstojanja zaključili da je do toga trebalo ranije doći. Ovo zato što je ovaj par objekata po stepenu bliskosti treći na rang listi parova objekata. ■

9.4.3 Izbor broja grupa

Korišćenjem jednog od metodi hijerarhijskog grupisanja dobijamo hijerarhijsku strukturu objekata. Grafički prikaz te hijerarhijske strukture, dendrogram, možemo "preseći" na određenoj visini izborom željenog broja grupa. Time smo dobili jedno od mogućih rešenja problema grupisanja. Poznato je da se u svakom koraku hijerarhijskog postupka grupisanja formira jedno grupisanje objekata, pri čemu se kod metodi udruživanja broj grupa kreće od n do 1. Postavlja se pitanje, da li u odsustvu apriornog izbora broja grupa, možemo odrediti optimalan broj grupa u konkretnom istraživanju.

U svrhe određivanja broja grupa obično se koristi jedan od kriterijuma optimalnosti. Tokom formiranja hijerarhijske strukture objekata, u svakom koraku, izračunava se vrednost izabranog kriterijuma optimalnosti. Prateći iz koraka u korak kretanje vrednosti kriterijuma optimalnosti, donosimo odluku o optimalnom broju grupa.

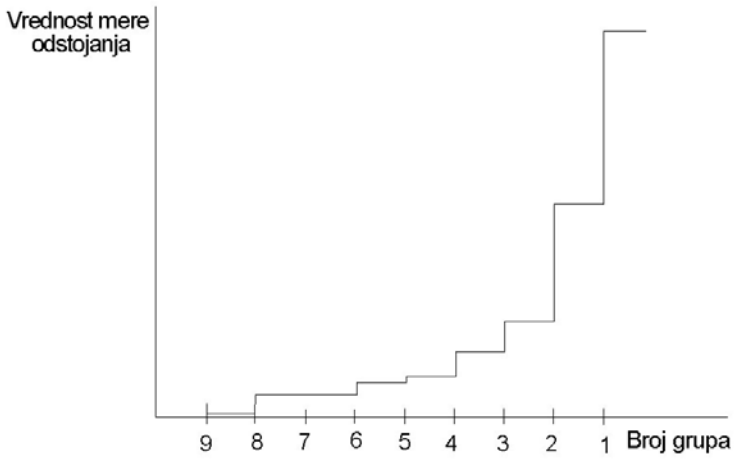
Najjednostavniji pristup problemu izbora broja grupa zasnovan je na praćenju vrednosti mere odstojanja pri kojoj se dve grupe udružuju u jednu. Krećući se od prvog ka $(n-1)$ koraku, vrednost mere odstojanja između grupa će rasti. Vrednost mere odstojanja između grupa će u prvim koracima (kada je broj grupa velik) rasti relativno sporije. Međutim, sa smanjivanjem broja grupa, odnosno povećanjem broja koraka u iterativnom postupku hijerarhijskog udruživanja objekata, vrednost mere odstojanja između grupa će rasti eksponencijalno. Sada se može u okolini očekivanog broja grupa ispitati ponašanje vrednosti mere odstojanja između grupa. Ako se u određenom koraku zabeleži velika promena u vrednosti mere odstojanja između grupa, tada broj grupa koji je prethodio tom koraku proglašavamo optimalnim. U okviru narednog primera, na Slici 9.7, ilustrujemo ovaj način određivanja broja grupa.

Primer 9.14 Na osnovu matrice kvadrata Euklidskih odstojanja iz Primera 9.2 (gornji trougao iznad glavne dijagonale) formirati hijerarhijsku strukturu objekata korišćenjem metoda prosečnog povezivanja. Odrediti optimalan broj grupa korišćenjem grafičkog prikaza vrednosti mera odstojanja između grupa u zavisnosti od broja grupa.

Korišćenjem metoda prosečnog povezivanja, na osnovu matrice odstojanja između deset osoba, dobijena je sledeća hijerarhijska struktura:

Korak	Broj grupa	Grupe	Odstojanje	Priraštaj odstojanja
1	9	(Marko) (Miodrag)	0.010	-
2	8	(Stojan) (Ljubomir)	0.040	0.030
3	7	(Vladan) (Bojan)	0.040	0
4	6	(Stojan, Ljubomir) (Milena)	0.060	0.020
5	5	(Marko, Miodrag) (Danica)	0.065	0.005
6	4	(Vladan, Bojan) (Zorica)	0.110	0.045
7	3	(Marko, Miodrag, Danica) (Dubravka)	0.160	0.050
8	2	(Stojan, Ljubomir, Milena) (Marko, Miodrag, Danica, Dubravka)	0.832	0.678
9	1	(Stojan, Ljubomir, Milena, Marko, Miodrag, Danica, Dubravka) (Vladan, Bojan, Zorica)	1.878	1.040

Na osnovu podataka iz gornje tabele (druga i četvrta kolona) nacrtali smo sledeću sliku.



Slika 9.7 Grafik kriterijuma grupisanja (određivanje broja grupa)

Kretanje vrednosti mere odstojanja na Slici 9.7 ukazuje na to da je u osmom koraku došlo do njenog velikog skoka, odnosno ovde imamo najveći priraštaj u odnosu na prethodne priraštaje. Zato se prema ovom kriterijumu sugeriše da se uzme kao optimalan broj grupa iz prethodnog, sedmog koraka. Znači da se prema korišćenom kriterijumu optimalnosti u ovom zadatku pokazalo da je grupisanje objekata u tri grupe optimalno. ■

Sličan prilaz problemu određivanja broja grupa imamo kada umesto vrednosti mere odstojanja između grupa koristimo njihov priraštaj (elementi iz poslednje kolone tabele u okviru Primera 9.14). U početku ova kriva, slično kao i kriva apsolutnih vrednosti mera odstojanja lagano raste, da bi znatno porasla kada dve međusobno udaljene grupe spojimo u jednu. Dramatična promena u nagibu krive priraštaja predstavlja pokazatelj završetka procesa formiranja grupa.

Pored navedenog kriterijuma optimalnosti, na osnovu koga vršimo izbor broja grupa, u analizi grupisanja razvijeni su i drugi postupci. U cilju definisanja formalnih kriterijuma za određivanje broja grupa, problem grupisanja posmatramo u okviru multivarijacione analize varijanse. Neka je dato g grupa objekata svaka sa $n_1, n_2, ..., n_g$ objekata, pri čemu je na svakom objektu merena $p-$dimenziona promenljiva. Na osnovu modela MANOVA iz 5. Glave i odgovarajuće notacije korišćene za opis razlika između grupa, imamo tri matrice sume kvadrata i uzajamnih proizvoda: $\mathbf{T}$, $\mathbf{W}$ i $\mathbf{B}$. Ove matrice nazvane su matrica ukupne sume kvadrata, matrica sume kvadrata unutar tretmana (grupa) i matrica sume kvadrata između tretmana respektivno i za njih važi relacija: $\mathbf{T} = \mathbf{W} + \mathbf{B}$. Statistike testa za određivanje broja grupa zasnovane su na tragu ovih matrica. Tako veličina

$$\mathrm{tr}(\mathbf{T}) = \sum_{j=1}^{p}\sum_{k=1}^{g}\sum_{i=1}^{n_k}(x_{ijk} - \bar{x}_{\cdot j\cdot})^2 \tag{9.16}$$

predstavlja meru ukupne sume kvadrata,

$$\mathrm{tr}(\mathbf{W}) = \sum_{j=1}^{p}\sum_{k=1}^{g}\sum_{i=1}^{n_k}(x_{ijk} - \bar{x}_{\cdot jk})^2 \tag{9.17}$$

predstavlja meru sume kvadrata unutar grupa i

$$\mathrm{tr}(\mathbf{B}) = \sum_{j=1}^{p}\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\bar{x}_{\cdot jk} - \bar{x}_{\cdot j\cdot})^2 = \sum_{j=1}^{p}\sum_{k=1}^{g} n_k(\bar{x}_{\cdot jk} - \bar{x}_{\cdot j\cdot})^2 \tag{9.18}$$

predstavlja meru sume kvadrata između grupa.

U svakom koraku hijerarhijskog postupka udruživanja $\mathrm{tr}(\mathbf{W})$ raste za iznos $(SKW_t - SKW_r - SKW_s)$ kao što smo sugerisali u Poglavlju 9.3, a za isti iznos se $\mathrm{tr}(\mathbf{B})$ smanjuje. Ukupna suma kvadrata $\mathrm{tr}(\mathbf{T})$ ostaje nepromenjena tokom postupka udruživanja.

Prva statistika koju koristimo pri donošenju odluke o broju grupa je tzv. *pseudo – F statistika*

$$F^* = \frac{\mathrm{tr}(\mathbf{B})/(g-1)}{\mathrm{tr}(\mathbf{W})/(n-g)} \tag{9.19}$$

Uz pretpostavku multivarijacione normalnosti sa sferičnom kovarijacionom matricom, ova statistika je standardna MANOVA statistika sa jednim faktorom za testiranje jednakosti sredina grupa. Uz navedenu pretpostavku F^* ima F raspored sa $p(g-1)$ i $p(n-g)$ stepeni slobode, ako su sredine grupa jednake. Pored uobičajenog postupka testiranja zasnovanog na poređenju realizovane i tablične vrednosti ove statistike, ona se može može koristiti i tako što se prate promene njene vrednosti iz koraka u korak. Prvobitno će sa smanjivanjem broja grupa g doći do smanjivanja vrednosti F^*, jer će

$\mathrm{tr}(\mathbf{W})$ postepeno rasti, a $\mathrm{tr}(\mathbf{B})$ postepeno opadati. Onoga trenutka kada se dve, međusobno znatno različite grupe udruže u jednu, doći će do velike promene u $\mathrm{tr}(\mathbf{W})$ i $\mathrm{tr}(\mathbf{B})$, a time i do znatnih promena u statistici F^*, odnosno do pada njene vrednosti. Veličinu g koja je neposredno prethodila ovoj nagloj promeni vrednosti F^* uzećemo za optimalan broj grupa.

Drugi pokazatelj ukazuje na to kako se vrši podela ukupne sume kvadrata na $\mathrm{tr}(\mathbf{T})$, $\mathrm{tr}(\mathbf{W})$ i $\mathrm{tr}(\mathbf{B})$, i glasi

$$R_g^2 = \frac{\mathrm{tr}(\mathbf{B})}{\mathrm{tr}(\mathbf{T})} \tag{9.20}$$

Ovaj količnik predstavlja proporciju ukupne varijacije između objekata koja se može pripisati varijacijama između grupa. Sa opadanjem broja grupa g, opada i R_g^2. Nagli pad vrednosti R_g^2 sugeriše da je došlo do udruživanja dve grupe koje su znatno različite međusobom. Broj grupa pri kome je došlo do pada vrednosti ovog pokazatelja uzmamo za optimalan broj grupa.

Treći pokazatelj je tzv. *semiparcijalni* R_g^2 i glasi

$$\Delta R_g^2 = R_g^2 - R_{(g-1)}^2 = \frac{SKW_t - SKW_r - SKW_s}{\mathrm{tr}(\mathbf{T})}. \tag{9.21}$$

Slično kao i kod prethodnog pokazatelja, praćenjem ΔR_g^2 u svakom koraku hijerarhijskog postupka grupisanja uočava se za koji broj grupa dolazi do naglog porasta njegove vrednosti. Sa stanovišta ovog pokazatelja taj broj grupa se proglašava optimalnim.

9.4.4 Statističko vrednovanje kvaliteta grupisanja

Upotrebom različitih koeficijenata korelacije merimo kvalitet grupisanja objekata. Mere su zasnovane na poređenju originalne matrice odstojanja i lokacije svakog objekta u grupi. Osnovni princip koji je korišćen u konstrukciji ovih pokazatelja je da objekti koji su u istoj grupi u ma kom koraku, treba da imaju manje međusobno odstojanje nego objekti koji su u različitim grupama. U svakom koraku hijerarhijskog postupka svim parovima objekata dodeljuje se vrednost 0 ako su objekti u istoj grupi, a vrednost 1 ako su u različitim grupama. Koeficijent korelacije između originalnih odstojanja i ovih dodeljenih vrednosti koristi se za određivanje kvaliteta grupa, odnosno interne konzistentnosti rešenja problema grupisanja.

Među brojnim koeficijentima (Milligan (1981)) izdvajamo one sa najboljim osobinama. Prvi je običan koeficijent korelacije izmedu originalnih odstojanja objekata i dodeljenih vrednosti (0 ili 1). Ovaj koeficijent nazvan je *biserijalni koeficijent korelacije*. Njegovu vrednost možemo izračunati i prema izrazu

$$\frac{(\bar{d}_b - \bar{d}_w)\sqrt{n_b n_w / n_d^2}}{s_d} \tag{9.22}$$

gde indeksi b i w označavaju grupe parova koji imaju vrednost 1 i vrednost 0 respektivno u skladu sa gornjim postupkom kodiranja svakog para objekata. Sredine originalnih odstojanja u ove dve grupe parova su označene sa $\bar{d}_b$ i $\bar{d}_w$. Broj parova u svakoj od dve grupe je n_b i n_w, a ukupan broj parova je n_d $(= n_b + n_w)$. Standardna devijacija originalnih odstojanja je označena sa s_d. Visoka vrednost ovog koeficijenta ukazuje da parovi koji su kodirani sa 1 teže da imaju visoku vrednost odstojanja, a parovi koji su kodirani sa 0 teže da imaju nisku vrednost odstojanja.

Drugi pokazatelj je tzv. γ *koeficijent konkordanse*. Izračunavamo ga prema izrazu

$$\gamma = \frac{S(+) - S(-)}{S(+) + S(-)} \tag{9.23}$$

gde $S(+)$ pokazuje koliko puta parovi objekata, koji nisu u istoj grupi, imaju veće odstojanje od drugih parova objekata koji su u istoj grupi. Obrnuto definišemo veličinu $S(-)$. Dakle, poredimo po dva para objekata iz svake od dve grupe, pa ukupno imamo poređenja. Taj broj poređenja se deli na dva dela: $S(+)$ i $S(-)$. Alternativni pokazatelj je $G(+) = 2S(-)/[n_d(n_d - 1)]$. Koeficijent $G(+)$ uzima vrednosti bliske 0 kada postoji veliko slaganje između pripadnosti objekata grupi i veličine originalnog odstojanja. Detaljnije o navedenim i drugim pokazateljima kvaliteta grupa i njihovom međusobnom poređenju videti u radu Milligana (1981).

Primer 9.15 Na osnovu podataka iz Primera 9.14, radi provere kvaliteta grupisanja objekata, izračunati vrednosti sva tri izložena koeficijenta: biserijalni koeficijent korelacije, γ koeficijent konkordanse i $G(+)$ koeficijent.

Kodiranje parova objekata dato je za različit broj grupa, od 9 do 2, u prvoj tabeli zajedno sa originalnim odstojanjima za svaki par objekata. Redosled parova objekata je prema veličini odstojanja, od najmanjeg do najvećeg.

Prema podacima iz te tabele imamo $n_d = 45$, $s_d = 0.935$, a ostali elementi potrebni za izračunavanje pokazatelja dati su u drugoj tabeli.

Naposletku, u poslednjoj tabeli su date izračunate vrednosti traženih koeficijenata.

Parovi	Odstojanje	Broj grupa							
		9	8	7	6	5	4	3	2
(Marko, Miodrag)	0.01	0	0	0	0	0	0	0	0
(Stojan, Ljubomir)	0.04	1	0	0	0	0	0	0	0
(Ljubomir, Milena)	0.04	1	1	1	0	0	0	0	0
(Vladan, Bojan)	0.04	1	1	0	0	0	0	0	0
(Marko, Danica)	0.05	1	1	1	1	0	0	0	0
(Zorica, Bojan)	0.05	1	1	1	1	1	0	0	0
(Stojan, Milena)	0.08	1	1	1	0	0	0	0	0
(Miodrag, Danica)	0.08	1	1	1	1	0	0	0	0
(Marko, Dubravka)	0.10	1	1	1	1	1	1	0	0
(Miodrag, Dubravka)	0.13	1	1	1	1	1	1	0	0
(Vladan, Zorica)	0.17	1	1	1	1	1	0	0	0
(Danica, Dubravka)	0.25	1	1	1	1	1	1	0	0
(Danica, Milena)	0.40	1	1	1	1	1	1	1	0
(Stojan, Danica)	0.64	1	1	1	1	1	1	1	0
(Stojan, Vladan)	0.65	1	1	1	1	1	1	1	1
(Marko, Milena)	0.65	1	1	1	1	1	1	1	0
(Danica, Ljubomir)	0.68	1	1	1	1	1	1	1	0
(Miodrag, Milena)	0.80	1	1	1	1	1	1	1	0
(Dubravka, Milena)	0.85	1	1	1	1	1	1	1	0
(Stojan, Marko)	0.85	1	1	1	1	1	1	1	0
(Stojan, Bojan)	0.85	1	1	1	1	1	1	1	1
(Stojan, Dubravka)	0.89	1	1	1	1	1	1	1	0
(Vladan, Ljubomir)	0.97	1	1	1	1	1	1	1	1
(Marko, Ljubomir)	0.97	1	1	1	1	1	1	1	0
(Stojan, Zorica)	1.00	1	1	1	1	1	1	1	1
(Stojan, Miodrag)	1.04	1	1	1	1	1	1	1	0
(Ljubomir, Dubravka)	1.13	1	1	1	1	1	1	1	0
(Miodrag, Ljubomir)	1.16	1	1	1	1	1	1	1	0
(Vladan, Milena)	1.17	1	1	1	1	1	1	1	1
(Ljubomir, Bojan)	1.17	1	1	1	1	1	1	1	1
(Zorica, Ljubomir)	1.28	1	1	1	1	1	1	1	1
(Bojan, Milena)	1.45	1	1	1	1	1	1	1	1
(Vladan, Dubravka)	1.48	1	1	1	1	1	1	1	1
(Zorica, Milena)	1.64	1	1	1	1	1	1	1	1
(Vladan, Danica)	1.93	1	1	1	1	1	1	1	1
(Vladan, Marko)	1.94	1	1	1	1	1	1	1	1
(Bojan, Dubravka)	2.00	1	1	1	1	1	1	1	1
(Vladan, Miodrag)	2.21	1	1	1	1	1	1	1	1
(Danica, Bojan)	2.45	1	1	1	1	1	1	1	1
(Marko, Bojan)	2.50	1	1	1	1	1	1	1	1
(Zorica, Dubravka)	2.57	1	1	1	1	1	1	1	1
(Miodrag, Bojan)	2.81	1	1	1	1	1	1	1	1
(Zorica, Danica)	2.92	1	1	1	1	1	1	1	1
(Marko, Zorica)	3.05	1	1	1	1	1	1	1	1
(Zorica, Miodrag)	3.40	1	1	1	1	1	1	1	1

	Broj grupa				
	6	5	4	3	2
n_b	40	38	36	33	21
n_w	5	7	9	12	24
$S(+)$	198	264	322	396	482
$S(-)$	2	2	2	0	22

Vrednost izračunatog koeficijenta i koeficijenta sugeriše da je rešenje problema grupisanja sa tri grupe optimalno. Što se biserijalnog koeficijenta tiče on sugeriše da je rešenje sa dve grupe optimalno.

Pokazatelj	Broj grupa				
	6	5	4	3	2
Biserijalni koeficijent	0.409	0.493	0.568	0.669	0.756
Gama koeficijent	0.980	0.985	0.988	1	0.913
$G(+)$	0.002	0.002	0.002	0	0.022

Dakle, korišćenjem kriterijuma optimalnosti pri određivanju broja grupa, kao i koeficijenata za proveru kvaliteta grupisanja dobijeno je da je u slučaju grupisanja deset osoba rešenje problema grupisanja sa tri grupe optimalno. ■

9.4.5 Nehijerarhijski metodi grupisanja

Za razliku od hijerarhijskih, nehijerarhijski metodi grupisanja objekata dozvoljavaju mogućnost premeštanja objekata iz ranije formiranih grupa. Do premeštanja objekata će doći ukoliko to sugeriše izabrani kriterijum optimalnosti. U primeni ovih metoda pretpostavlja se da je broj grupa unapred poznat ili ga, kao kod nekih metoda, variramo tokom postupka grupisanja.

Postupak nehijerarhijskog grupisanja započinje inicijalnom podelom skupa objekata u izabran broj grupa. Alternativa inicijalnoj podeli objekata u grupe je apriori određivanje inicijalne *klice* (eng. *seed*), odnosno inicijalnog centroida za svaku grupu. Potom se odredi odstojanje između svakog objekta i svake grupe (inicijalnog centroida). Objekti se lociraju u grupe kojima su najbliže. Nakon pridruživanja objekta nekoj grupi, izračunava se centroid grupe iz koje je objekat "otišao" i grupe kojoj se objekat "pridružio". Ponovo za svaki objekat izračunavamo njegovo odstojanje od centroida grupa i vršimo preraspodelu objekata između grupa sve dotle, dok izabrana funkcija kriterijuma to sugeriše. Razmotrićemo najpopularniji među nehijerarhijskim metodama, a to je *metod* $k-$*sredina* (eng. $k-$*means method*). Kako ovde koristimo oznaku g za broj grupa, ovaj metod bismo nazvali metod $g-$sredina. Prema njemu objekat pridružujemo grupi koja ima najbliži centroid (sredinu). U sledećem primeru izložićemo postupak primene ovog metoda.

Primer 9.16 Na osnovu podataka iz Primera 9.1 izvršiti grupisanje objekata u dve grupe korišćenjem metoda $k-$sredina.

Pretpostavićemo da smo formirali dve grupe sa sledećom raspodelom objekata između grupa: $(1,4,5)$ i $(2,3)$. Centroidi grupa su dati sledećom tabelom:

Grupe	Promenljive	
	Prva	Druga
(1,4,5)	22/3	19/3
(2,3)	3	3.5

Odstojanje između prvog objekta i svake od grupa je

$$d_{1(145)} = \left(5-\frac{22}{3}\right)^2 + \left(4-\frac{19}{3}\right)^2 = 10.89$$

$$d_{1(23)} = (5-3)^2 + (4-3.5)^2 = 4.25$$

Ostala odstojanja data su u narednoj tabeli:

Grupe	O b j e k t i				
	1	2	3	4	5
(1,4,5)	10.89	33.89	22.22	0.22	14.22
(2,3)	4.25	1.25	1.25	22.27	79.25

Za prvi objekat je dobijen rezultat koji sugeriše da je on bliži grupi kojoj ne pripada $(2,3)$, nego onoj kojoj pripada $(1,4,5)$. Prema tome u narednom koraku premeštamo prvi objekat iz prve u drugi grupu.

Prethodno ćemo izračunati vrednost kriterijuma grupisanja. Pri podeli skupa objekata dešava se da objekte pridružimo grupi kojoj oni, s obzirom na veličinu odstojanja, ne bi trebalo da pripadaju. Stoga se pri podeli skupa objekata javlja greška, a nju merimo zbirom svih odstojanja objekata od centroida grupa kojima pripadaju. Dobijeni zbir predstavlja vrednost kriterijuma grupisanja. U našem primeru vrednost kriterijuma iznosi 27.83 $(=10.89+1.25+1.25+0.22+14.22)$.

Nakon seobe prvog objekta promenjene vrednosti centroida prikazali smo u sledećoj tabeli:

Grupe	Promenljive	
	Prva	Druga
(4,5)	8.5	7.5
(1,2,3)	11/3	11/3

Međusobna odstojanja objekata od centroida grupa data su sledećom tabelom:

Grupe	Objekti				
	1	2	3	4	5
(4,5)	24.50	54.50	40.50	4.50	4.50
(1,2,3)	1.89	2.89	0.56	102.14	125.14

Vrednost kriterijuma grupisanja iznosi 14.34, što znači da je premeštanjem prvog objekta iz prve u drugu grupu došlo do smanjenja ove vrednosti. U gornjoj su tabeli sva odstojanja objekata od centroida svojih grupa manja od odstojanja istih objekata od centroida grupe kojoj ne pripadaju. Znači da daljim premeštanjem objekata iz jedne u drugu grupu ne može doći do smanjivanja vrednosti kriterijuma grupisanja. Time se postupak primene metoda $k-$sredina završava. Poređenjem dobijenog grupisanja uočavamo da je dobijeno isto rešenje kao i primenom Wardove metode ili metode potpunog povezivanja.

Premeštanjem objekata iz jedne u drugu grupu očekujemo da ćemo dobiti sve više i više razdvojene grupe. I u našem primeru došlo je do međusobnog udaljavanja dve grupe. Naime, odstojanje dve grupe (odstojanje njihovih centroida) u prvom koraku, pre selidbe prvog objekta, iznosilo je 26.81, a nakon selidbe centroidi su se međusobom udaljili, odnosno odstojanje sada iznosi 38.06. ■

U primeni nehijerarhijskih metoda grupisanja treba imati u vidu i to da su oni, kao uostalom i drugi metodi grupisanja, osetljivi na prisustvo nestandardnih opservacija. To znači da se u takvim slučajevima može dobiti grupa sa veoma različitim objektima. O ovim i drugim problemima u primeni metoda grupisanja detaljnije videti kod Everitta (1974) i Hartigana (1975).

9.5 KORIŠĆENJE ANALIZE GRUPISANJA SA DRUGIM MULTIVARIJACIONIM METODAMA

Pošto je pronađeno optimalno rešenje problema grupisanja, može se ispitati koje su promenljive doprinele upravo takvom formiranju grupa. Pri interpretaciji dobijenog rešenja problema grupisanja moguće je koristiti više pristupa. Jedan od njih zasnovan je na korišćenju jednodimenzione ili višedimenzione analize varijanse. Njome se ispituje da li su statistički značajne razlike između sredina formiranih grupa.

Drugi pristup interpretaciji dobijenog rešenja problema grupisanja zasnovan je na diskriminacionoj analizi. Empirijski formirane grupe na osnovu postupaka analize grupisanja posmatramo kao apriorno definisane grupe u diskriminacionoj analizi. Zatim se kanoničkom diskriminacionom analizom utvrđuje koje su promenljive u najvećoj meri doprinele međusobnom razdvajanju grupa.

Ukoliko je broj promenljivih suviše velik, tada se može metodom glavnih komponenata ili metodom faktorske analize redukovati njihov broj. Pri tome treba imati u vidu da je metod glavnih komponenata osetljiv na uticaj merne skale. Zato se sugeriše da se pri redukovanju broja promenljivih u analizi grupisanja, metod glavnih komponenata primeni na standardizovane podatke ili što je isto, na korelacionu matricu. Time se eliminiše mogućnost da jedna od promenljivih dominira dobijenim rešenjem zbor svoje merne skale.

Nakon primene metode glavnih komponenata u cilju redukovanja podataka postavlja se pitanje, da li je bolje standardizovati vrednosti dobijenih glavnih komponenata, pre primene analize grupisanja. Može se pokazati da će mera odstojanja između objekata u slučaju korišćenja nestandardizovane i standardizovane vrednosti glavnih komponenata biti međusobom različita. Zato se sugeriše da se ne vrši standardizacija vrednosti glavnih komponenata. Međutim, ako se koristi korelaciona mera sličnosti objekata, tada se standardizacijom vrednosti glavnih komponenata ne menja vrednost mere sličnosti.

LITERATURA

Anderberg, M.R. (1973). *Cluster analysis for applications*. New York: Academic Press.

Anderson, T.W. (1972). *An introduction to multivariate statistical analysis*. New Delhi: Wiley Eastern Private Limited.

Andrews, D.F. (1972). Plots of high dimensional data, *Biometrics* **28**, 125-136.

Arnold, S.J. (1979). A Test for Clusters. *Journal of Marketing Research* **16**, 545-551.

Bartlett, M.S. (1934). The vector representation of a sample. *Proceedings of the Cambridge Philosophical Society* **30**, 327-340.

Bartlett, M.S. (1937). The statistical conception of mental factor. *British Journal of Psychology* **28**, 97-104.

Bartlett, M.S. (1947). Multivariate analysis. *Journal of the Royal Statistical Society* **B9**, 176-197.

Bishop, Y.M.M., S.E. Fienberg i P.W. Holland (1975). *Discrete multivariate analysis: theory and practice*. Cambridge: The MIT Press.

Blashfield, R.K. i Aldenderfer, M.S. (1978). The literature on cluster analysis. *Multivariate Behavioral Research* **13**, 271-295.

Box, G.E.P., W.G. Hunter i J.S. Hunter (1978). *Statistics for experimenters*, New York: John Wiley and Sons.

Carroll, J.B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika* **18**, 23-38.

Cattell, R.B. (1966). The scree test for the number of factors. *Journal of the Multivariate Behavioral Research* **1**, 245-276.

Chambers, J.M., Cleveland, W.S., Kleiner, B. i Tukey, P.A. (1983). *Graphical methods for data analysis*. Belmont: Wadsworth.

Chambers, J.M. i Kleiner, B. (1982). Graphical techniques for multivariate data and for clustering. U *Handbook of Statistics, Volume II: Classification, Pattern Recognition and Reduction of Dimensionality*. (ur. P.R. Krishnaiah i L. Kinal), 209-244, Amsterdam: North-Holland.

Chatfield, C. i Collins, A.J. (1980). *Introduction to multivariate analysis*. London: Chapman and Hall.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically, *Journal of American Statistical Society* **68**, 361-368.

Cleveland, W.S. (1985). *The elements of graphing data*, Monterey, California: Wadsworth.

Devlin, S.J., Gnanadesikan, R. i Kettenring, J.R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association* **76**, 354-362.

Dillon, W.R. (1984). *Multivariate analysis: Methods and application*. New York: John Wiley and Sons.

Duran, B.S. i Odell, P.L. (1974). *Cluster analysis*. New York: Springer-Verlag.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Society for Industrial and Applied Mathematics*, Monograph No. 38, Philadelphia.

Everitt, B.S. (1974). *Cluster analysis*. London: Heinemann.

Everitt, B.S. (1978). *Graphical techniques for multivariate data*. London: Heinemann.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179-188.

Flury, B. (1988). *Common principal components and related multivariate models*. New York: John Wiley and Sons.

Flury, B. i Riedwyl, H. (1981). Graphical representation of multivariate data by means of asymetrical faces, *Journal of American Statistical Society* **76**, 757-765.

Flury, B. i Riedwyl, H. (1988). *Multivariate statistics - A practical approach*, London: Chapman and Hall.

Giri, N.C. (1977). *Multivariate statistical inference*. New York: Academic Press.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. New York: John Wiley and Sons.

Goldstein, M. i Dillon, W.R. (1978). *Discrete discriminant analysis*. New York: John Wiley and Sons.

Gower, J.C. (1967). A comperison of some methods of cluster analysis. *Biometrics* **23**, 623-637.

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857-872.

Graybill, F.A. (1983). *Matrices with applications in statistics*. 2nd edition. Belmont: Wadsworth Publishing Company.

Greenhouse, S.W. i Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika* **32**, 95-112.

Hair, J.F., R.E. Anderson i R.L. Tatham (1987). *Multivariate data analysis - with readings*. 2nd edition. New York: Macmillan Publishing Company.

Hand, D.J. (1981). *Discrimination and classification*. New York: John Wiley and Sons.

Harman, H.H. (1967). *Modern factor analysis*. 2nd edition. Chicago: The University of Chicago Press.

Harris, C.W. i Kaiser H.F. (1964). Oblique factor analytic solutions by orthogonal transformations, *Psychometrika* **29**, 347-62.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: John Wiley and Sons.

Hartigan, J.A. (1978). Asympthotic distributions for clustering criteria. *Annals of Statistics* **6**, 117-131.

Hawkins, D.M. (1980). *Identification of outliers*. London: Chapman and Hall.

Horn, J.L. i Engstrom, R. (1979). Cattel's scree test in relation to bartlett's chi-square test and other observations on the number of factor problem. *Multivariate Behavioral Research* **14**, 283-300.

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology* **26**, 139-142.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika* **28**, 321-377.

Huynh, H. i Feldt, L.S. (1970). Conditions under which mean sqare ratios in repeated measurements designs have exact F-distributions. *Journal of the American Statistical Association* **65**, 1582-1589.

Ivanović B. (1963). *Diskriminaciona analiza sa primenama u ekonomskim istraživanjima*. Beograd: Naučna knjiga.

Ivanović B. (1977). *Teorija klasifikacije*. Beograd: Institut za ekonomiku industrije.

Jambu, M. (1991). *Exploratory and multivariate data analysis*. New York: Academic Press.

Jardine, N. i Sibson, R. (1971). *Mathematical taxonomy*. New York: John Wiley and Sons.

Jobson, J.D. (1992). *Applied multivariate data analysis. Volume II: categorial and multivariate methods*. New York: Springer-Verlag.

Johnson, R.A. and D.W. Wichern (1982). *Applied multivariate statistical analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Johnson, R.A. and D.W. Wichern (1988). *Applied multivariate statistical analysis*. 2nd edition. Englewood Cliffs, NJ: Prentice-Hall.

Jolliffe, I.T. (1986). *Principal component analysis*. New York: Springer-Verlag.

Jöreskog, K.G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika* **31**, 165-178.

Jöreskog, K.G. (1975). Factor analysis by least squares and maximum likelihood. U *Statistical methods for digital computers*. (ur. K. Enslein, A. Ralston i H.S. Wilf), New York: John Wiley and Sons.

Jovičić, M. (1981). *Ekonometrijski metodi*. Beograd: Savremena administracija.

Karson, M.J. (1982). *Multivariate statistical methods - An introduction*. Ames: The Iowa State University Press.

Kaiser, H.F. (1958). The varimax criterion for analytical rotation in factor analysis. *Psychometrika* **23**, 187-200.

Kendall, M.G., Staurt, A. i Ord. J.K. (1983). *The advanced theory of statistics, Vol 3*. 4th edition. New York: Macmillan Publishing Co.

Košmelj, B. (1983). *Uvod v multivariatno analizo*. Ljubljana: Ekonomska fakulteta Borisa Kidriča.

Krisnaiah, P.R. (1982). Selection of variables in discriminant analysis. U *Handbook of Statistics, Volume II: Classification, Pattern Recognition and Reduction of Dimensionality*. (ur. P.R. Krishnaiah i L. Kinal), 169-191, Amsterdam: North-Holland.

Krzanowski, W.J. (1988). *Principles of multivariate analysis: A User's Perspective*. New York: Oxford University Press.

Kshirsagar, A.M. (1972). *Multivariate analysis*. New York: Marcel Dekker.

Lachenbruch, P.A. (1975). *Discriminant analysis*. New York: Hafner.

Lachenbruch, P.A. (1979). Discriminant analysis. *Biometrics* **35**, 69-85.

Lachenbruch, P.A. i Mickey, M.R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1-11.

Lachenbruch, Sneeringer i Revo (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Communications in Statistics* **1**, 39-57.

Lance, G.N. i Williams, W.T. (1967). A general theory of classificatory sorting strategies. I. Hierarchical Systems. *Computer Journal*, **9**, 373-380.

Lawley, D.N. (1959). Tests of significance in canonical analysis. *Biometrika* **46**, 59-66.

Lawley, D.N. i A.E. Maxwell (1971). *Factor analysis as a statistical method*. New York: American Elsevier.

Manly, B.F.J. (1986). *Multivariate statistical methods - A primer*. London: Chapman and Hall.

Mardia, K.V., J.T. Kent i J.M. Bibby (1979). *Multivariate analysis*. London: Academic Press.

Milligan, G.W. (1981). A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika* **46**, 187-199.

Milligan, G.W. i Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**, 159-179.

Milošević, V.M. (1983). *Teorijska statistika*. Beograd: Naučna knjiga.

Momirović, K. (1988). Uvod u analizu nominalnih varijabli. *Metodološke sveske*, 2, Ljubljana: JUS.

Morrison, D.F. (1976). *Multivariate statistical methods*. 2nd edition. Tokyo: McGraw-Hill Kogakusha, Ltd.

Mulaik, S.A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill Book Co.

Press, S.J. (1982). *Applied multivariate analysis: Using Bayesian and frequentist methods of inference*. 2nd edition. Malabar, Florida: Robert E. Krieger Publishing Company.

Rao, C.R. (1948). Tests of significance in multivariate analysis. *Biometrika*, **35**, 58-79.

Rao, C.R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika* **20**, 93-111.

Rao, C.R. (1964). The use and interpretation of pricipal component analysis in applied research. *Sakhya* **A26**, 329-358.

Rao, C.R. (1973). *Linear statistical inference and its applications*. Second edition, New York: John Wiley and Sons.

Scheffe, H. (1959). *The analysis of variance*. New York: John Wiley and Sons.

Searle, S.R. (1982). *Matrix algebra useful for statistics*. New York: John Wiley and Sons.

Smith, H., Gnanadesikan, R. I Hughes J.B. (1962). Multivariate analysis of variance (MANOVA). *Biometrics* **18**, 22-41.

Sneath, P.H.A. i Sokal, R.R. (1973). *Numerical taxonomy*. San Francisco: Freeman.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology* **15**, 201-293.

Stewart, D.K. i W.A. Love (1968). A general canonical correlation index. *Psychological Bulletin* **70**, 160-163.

Thompson, G.H. (1951). *The factorial analysis of human ability*. London: University of London Press.

Thurstone, L.L. (1945). *Multiple factor analysis*. Chicago: Chicago University Press.

Tufte, E. (1983). *The visual display of quantitative information*, Cheshire, Connecticut: Graphics Press.

van den Wollenberg, A.L. (1977). Redundancy analysis - an alternative to canonical correlation analysis. *Psychometrika* **42**, 207-219.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**, 236-244.

Wilks, S.S. (1932). Certain generalizations in the analysis of variance. *Biometrika* **24**, 471-494.

Wilks, S.S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *Annals of Mathematical Statistics* **17**, 257-281.

Wishart, J. (1928). The generalized product moment distribution in samples from a normal multivariate population. *Biometrika* **20**, 32-52.

Žižić, M., Lovrić, M. i Pavličić, D. (1992). *Osnovi statističke analize*. Beograd: Ekonomski fakultet.