Usporedba analize glavnih komponenti i faktorske analize

Rajčić, Krunoslav

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Krunoslav Rajčić

USPOREDBA ANALIZE GLAVNIH KOMPONENTI I FAKTORSKE ANALIZE

Diplomski rad

Voditelj rada:

Prof. dr. sc. Anamarija Jazbec

Zagreb, rujan 2015.

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

1. ______________________________, predsjednik
2. ______________________________, član
3. ______________________________, član

Povjerenstvo je rad ocijenilo ocjenom ___________________.

Potpisi članova povjerenstva:

1. ________________________________
2. ________________________________
3. ________________________________

Zahvala:

Zavaljujem se svojoj mentorici prof. dr. sc. Anamariji Jazbec na velikoj pomoći i brojnim savijetima koje mi je pružila prilikom izrade diplomskog rada.

Također, zahvaljujem se svim svojim prijateljima i prijateljicama, koji su uvijek bili uz mene i bez kojih ovo moje studiranje ne bi prošlo tako lako.

Posebnu zahvalnost iskazujem svojoj obitelji koja me uvijek podržavala.

Najveća zahvala ide mojim roditeljima i sestri, koji su mi velikim odricanjima omogućili da studiram, koji su uvijek bili tu, uz mene, bilo da se radilo o sretnim ili teškim trenutcima i bez kojih sve ovo što sam do sada postigao ne bi bilo moguće.

Veliko HVALA svima.

Sadržaj

Uvod

Statistika je znanost o principima prikupljanja, organizacije, analize, sažetog prikaza i interpretacije podataka dobivenih promatranjem ili mjerenjem vrijednosti varijable osnovnog skupa, uzorka ili pojedinca.

Postoje brojni statistički načini obrade podataka, kao i brojni statistički programski jezici za obradu tih istih podataka. U ovom diplomskom radu bazirat ćemo se na programski jezik SAS (statistical analysis sistem), te na obradu podataka koristeći PCA (principal components analysis – analiza glavnih komponenti) i FA (factor analysis – faktorska analiza).

Analiza glavnih komponenti (engl. principal components analysis - PCA) i faktorska analiza (engl. factor analysis - FA) statističke su tehnike koje se primjenjuju na jedan skup varijabli kada istraživača zanima koje varijable u tom skupu čine koherentne podskupove koji su međusobno relativno nezavisni. Varijable koje su međusobno povezane, no uglavnom nezavisne od drugih podskupova varijabli, povezuju se u faktore[1]. Smatra se da faktori odražavaju temeljne procese koji stvaraju veze između varijabli.

Uzmimo, primjerice, istraživača koji želi ispitati karakteristike radnika na platformi. Istraživač na velikom uzorku radnika mjeri osobne karakteristike, akademsku povijest, obiteljsku povijest, motivaciju, intelektualne sposobnosti, zdravlje, tjelesne karakteristike itd. Svako od ovih područja opisano je nizom varijabli; varijable ulaze u analizu jedna po jedna te se proučavaju njihove međusobne veze. Analiza otkriva obrasce povezanosti između varijabli za koje se smatra da odražavaju temeljne procese koji utječu na ponašanje radnika. Primjerice, nekoliko pojedinačnih varijabli iz vrijednosti dobivenih za osobne karakteristike povezuju se s određenim varijablama iz vrijednosti dobivenih za motivaciju i

[1] Analiza glavnih komponenti daje komponente, dok faktorska analiza daje faktore, no ovdje rezultate obiju analiza zovemo faktorima radi jednostavnosti.

akademsku povijest te se dobiva faktor koji mjeri stupanj do kojeg osoba preferira raditi samostalno - faktor samostalnosti. Nekoliko varijabli iz vrijednosti za intelektualne sposobnosti povezuju se s drugima iz akademske povijesti te se dobiva mogući faktor inteligencije.

POGLAVLJE 1

1. Opis i namjena PCA i FA

Specifični su ciljevi analize glavnih komponenti i faktorske analize dobivanje glavnih obrazaca odnosa između promatranih varijabli, svođenje velikog broja promatranih varijabli na manji broj faktora, dobivanje operacijske definicije (regresijske jednadžbe) za temeljni proces uporabom promatranih varijabli ili testiranje hipoteze o prirodi temeljnih procesa. U središtu određenog istraživačkog projekta mogu biti neki ili svi ovi ciljevi.

Analiza glavnih komponenti i faktorska analiza od značajne su koristi jer svode brojne varijable na nekoliko faktora. Matematički gledano, ove analize daju nekoliko linearnih kombinacija promatranih varijabli, od kojih je svaka linearna kombinacija faktor. Faktori sumiraju obrasce korelacija u promatranoj korelacijskoj matrici i mogu se, s različitim stupnjevima uspjeha, koristiti za reprodukciju promatrane korelacijske matrice. Međutim, budući da je broj faktora obično puno manji od broja promatranih varijabli, uporaba faktorske analize podrazumijeva značajnu „škrtost". Nadalje, vrijednosti faktora utvrđene za svaki subjekt često su pouzdanije od vrijednosti za pojedinačne promatrane varijable.

Koraci u ovim analizama uključuju odabir i mjerenje skupa varijabli, pripremu korelacijske matrice (za izvođenje bilo faktorske, bilo analize glavnih komponenti), ekstrakciju skupa faktora iz korelacijske matrice, određivanje broja faktora, rotiranje faktora kako bi se povećala interpretabilnost te, konačno, interpretacija rezultata. Iako se pri svakom od ovih koraka u obzir uzimaju važni statistički uvjeti, važan test analize također je njena interpretabilnost.

POGLAVLJE 1: OPIS I NAMJENA PCA I FA

Dobra faktorska ili analiza glavnih komponenti „ima smisla“, loša nema. Interpretacija i imenovanje faktora ovise o značenju onih kombinacija promatranih varijabli koje su u visokoj korelaciji s pojedinim faktorom. Određeni faktor lakše je interpretirati kad je nekoliko promatranih varijabli u visokoj korelaciji s njim i te varijable nisu u korelaciji s drugim faktorima.

Jednom kad je postignuta odgovarajuća interpretabilnost, posljednji je, i vrlo velik, korak provjera strukture faktora utvrđivanjem konstruktivne valjanosti faktora. Istraživaču je cilj dokazati da rezultati za latentne varijable (faktore) kovariraju s rezultatima za druge varijable, ili da se rezultati za latentne varijable mijenjaju zajedno s eksperimentalnim uvjetima u skladu s teoretskim predviđanjima.

Jedan od problema ovih dviju analiza jest nedostatak kriterijske varijable u usporedbi s kojom bi se moglo testirati rješenje. U regresijskoj analizi, primjerice, zavisna varijabla jest kriterij te korelacija između promatrane i predviđene vrijednosti zavisne varijable služi kao testiranje rješenja – slično kao i za dva skupa varijabli u kanonskoj korelaciji. U diskriminantnoj funkcijskoj analizi, logističkoj regresiji, analizi profila i multivarijatnoj analizi varijance, rješenje se ocjenjuje prema tome koliko dobro se predvidi pripadnost grupi. No u faktorskoj i analizi glavnih komponenti ne postoji vanjski kriterij, kao što je pripadnost grupi, prema kojem bi se testiralo rješenje.

Drugi problem s ovim analizama sastoji se u tome što se nakon ekstrakcije pruža beskonačan broj mogućih rotacija, od kojih svaka odgovara jednakom iznosu varijance u izvornim podacima, no s nešto različito definiranim faktorima. Konačni izbor između alternativa ovisi o istraživačevoj procjeni njihove interpretabilnosti i znanstvene primjenjivosti. U situaciji u kojoj postoji beskonačan broj matematički identičnih rješenja, očekivano je da se istraživači razilaze u ocjeni koje je najbolje. Budući da se razlike ne mogu prevladati pozivanjem na objektivne kriterije, rasprave oko najboljeg rješenja znaju postati bučne. Međutim, oni koji očekuju djelomičnu neodređenost u izboru najboljeg rješenja faktorske analize neće biti iznenađeni kada drugi istraživači izaberu neko drugo rješenje. Također, neće im biti iznenađujuće ni kada se rezultati ponovljene analize ne poklapaju savršeno, ako se različite odluke donesu pri jednom ili više koraka provedbe faktorske analize.

Treći je problem činjenica da se faktorska analiza često koristi u pokušaju „spašavanja“ loše osmišljenog istraživanja. Ako nijedan drugi statistički postupak nije primjenjiv, podaci se obično barem mogu faktorski analizirati. Stoga su razni oblici faktorske analize za mnoge povezani s loše provedenim istraživanjem.

Postoje dvije glavne vrste faktorske analize: eksploratorna i konfirmatorna. U eksploratornoj faktorskoj analizi, cilj je opisati i sumirati podatke tako što se zajedno grupiraju varijable koje su u korelaciji. Same varijable mogu i ne moraju biti odabrane s potencijalnim temeljnim procesima na umu. Eksploratorna faktorska analiza obično se provodi u ranim fazama istraživanja, kada omogućuje konsolidiranje varijabli i postavljanje hipoteza o temeljnim procesima. Konfirmatorna faktorska analiza mnogo je sofisticiranija tehnika koja se koristi u naprednim stadijima istraživačkog procesa za testiranje hipoteze o latentnim procesima. Varijable se pomno i ciljano odabiru kako bi se otkrili temeljni procesi. Ova vrsta analize često se provodi kroz modeliranje strukturnih jednadžbi.

McCleod, Brown i Becker (1977.) koristili su analizu glavnih komponenti kako bi odgovore na 17 stavki sveli na četiri faktora u istraživanju odnosa između percipiranog lokusa krivnje za skandal *Watergate* i nekoliko različitih indikatora političke orijentacije i ponašanja. Četiri faktora lokusa krivnje bili su: mediji, režim (političke stranke, Kongres), Nixon sa suradnicima i sustav (stranačke politike, politički sustav, ekonomski sustav). Razlike u vrijednostima faktora našli su oni koji jesu ili nisu podržali Nixona 1972. g.[2]

U istraživanju spremnosti na doniranje dijelova tijela (Pessemier, Bemmoar i Hanssens, 1977.), faktorska analiza korištena je za svođenje deset faktora koji su mjerili spremnost na doniranje na tri faktora: doniranje krvi, kože i koštane srži; doniranje u slučaju smrti; doniranje bubrega. Faktorska analiza provedena je nakon

[2] B. G. Tabachnick, L. S. Fidell, *Using multivariate statistics,* Needham Heights, MA, 1996. by Allyn & Beacon

što je analiza Guttmanovom ljestvicom otkrila da spremnost na doniranje nije jednodimenzionalna. Odgovori na varijable koje su činile faktore bili su sumirani tako da daju tri vrijednosti faktora za svakog ispitanika koji je služio kao zavisna varijabla u zasebnim analizama varijanci.[3] Sredovječne ispitanice iznadprosječnih prihoda pokazale su se najsklonijima doniranju regenerativnog tkiva ili dijelova tijela nakon smrti.[4]

Wernimont i Fitzpatrick u istraživanju 1972. g. ispitali su značenje novca. Analiza glavnih komponenti rabljena je kako bi se utvrdilo sedam komponenti koje su bile u temelju uzoraka odgovora 533 subjekta na 40 parova pridjeva, ocijenjenih na ljestvici od 1 do 7. Sedam komponenti označeno je nazivima „sramotan neuspjeh“, „društvena prihvatljivost“, „podcjenjivački stav“, „moralno zlo“, „udobna sigurnost“, „društvena neprihvatljivost“ te „konzervativni poslovni stavovi“. Provodeći analizu za grupu u cjelini, subjekti su podijeljeni u 11 skupina na temelju zanimanja, koje je služilo kao nezavisna varijabla u jednosmjernim analizama varijanci, s vrijednostima komponenti kao zavisnim varijablama. Zasebne analize varijanci provedene su za svaku komponentu. Skupine su se značajno razlikovale u svojim vrijednostima komponenti, s prekidom obično između značenja novca za nezaposlene (osobe u procesu osposobljavanja za posao, osobe na fakultetu) i značenja novca za zaposlene (znanstvenike, poslovođe, prodavače i sl.). Općenito, došlo se do zaključka da novac ima vrlo različita značenja za različite dijelove populacije.[5]

Sada ćemo definirati nekoliko pojmova. Prvi pojmovi odnose se na korelacijske matrice. Korelacijsku matricu promatranih varijabli nazivamo *opažanom korelacijskom matricom*. Korelacijsku matricu faktora zovemo *reproduciranom*

---

[3] Autori su mogli koristiti multivarijatnu analizu varijance jednako kao i univarijatnu analizu varijance.

[4] B. G. Tabachnick, L. S. Fidell, *Using multivariate statistics,* Needham Heights, MA, 1996. by Allyn & Beacon

[5] B. G. Tabachnick, L. S. Fidell, *Using multivariate statistics,* Needham Heights, MA, 1996. by Allyn & Beacon

korelacijskom matricom. Razlika između njih jest *rezidualna korelacijska matrica*. U dobroj faktorskoj analizi, korelacije u rezidualnoj matrici malene su, što upućuje na blisko podudaranje između opažane i reproducirane matrice.

Drugi skup pojmova odnosi se na matrice proizvedene i interpretirane kao dio rješenja. Rotacija faktora postupak je kojim se povećava interpretabilnost rješenja bez mijenjanja matematičkih svojstava u osnovi tog rješenja. Postoje dvije glavne vrste rotacija: ortogonalna i kosa. Ako je rotacija *ortogonalna* (tako da su svi faktori međusobno nepovezani), dobiva se *matrica faktorskog opterećenja.* Matrica faktorskog opterećenja tj. matrica faktorske strukture jest matrica korelacija između promatranih varijabli i faktora. Koeficijenti matrice predstavljaju opseg veze između svake promatrane varijable i svakog faktora. Ortogonalna faktorska analiza interpretira se iz matrice opterećenja tako što se gleda koje promatrane varijable koreliraju sa svakim od faktora.

Ako je rotacija *kosa* (tako da su sami faktori u korelaciji), dobiva se nekoliko dodatnih matrica. *Faktorska korelacijska matrica* sadrži korelacije između faktora. Matrica opterećenja iz ortogonalne rotacije dijeli se na dvije matrice za kosu rotaciju: *matricu faktorske strukture*, koja sadrži korelacije između faktora i varijabli, te *matricu faktorskog obrasca* jedinstvenih odnosa (nekontaminiranih preklapanjem među faktorima) između svakog faktora i svake promatrane varijable. Provodeći kosu rotaciju, značenje faktora utvrđuje se iz matrice faktorskog obrasca.

Na kraju, za oba tipa rotacije postoji matrica koeficijenata vrijednosti faktora – tj. matrica koeficijenata koja se koristi za nekoliko vrsta jednadžbi sličnih regresijskima, sa svrhom predviđanja vrijednosti faktora iz vrijednosti promatranih varijabli za svakog pojedinca.

Faktorska analiza daje *faktore,* dok analiza glavnih komponenti daje *komponente*. Međutim, ovi su postupci međusobno slični; razlika je jedino u pripremi opažane korelacijske matrice za ekstrakciju, te u osnovnoj teoriji. Matematički, razlika između ovih dviju analiza leži u varijanci koja se analizira. U analizi glavnih komponenti, cjelokupna se varijanca u promatranim varijablama analizira, dok se u faktorskoj analizi analizira samo zajednička varijanca; pokušava se utvrditi i eliminirati varijanca zbog pogreške i varijanca specifična za svaku pojedinu varijablu. Termin *faktor* ovdje se odnosi i na komponente i na faktore, osim ako razlika nije presudna, kada se koristi odgovarajući od dva termina.

Teoretski, razlika između faktorske analize i analize glavnih komponenti leži u razlogu zbog kojeg se varijable povezuju s faktorom ili s komponentom. Smatra se da faktori „uzrokuju“ varijable – konstrukt u osnovi (faktor) ono je što proizvodi vrijednosti na varijablama. Stoga, eksploratorna faktorska analiza povezana je s razvijanjem hipoteze, a konfirmatorna faktorska analiza – s njenim testiranjem. Pitanje u eksploratornoj faktorskoj analizi glasi: koji su temeljni procesi mogli uzrokovati korelacije između ovih varijabli? Pitanje u konfirmatornoj faktorskoj analizi jest: jesu li korelacije između varijabli konzistentne s hipotetskom strukturom faktora? Komponente su jednostavno agregati varijabli u korelaciji. U tom smislu, varijable „uzrokuju“, ili proizvode, komponente. Ne postoji nikakva temeljna teorija koja bi nalagala koje bi se varijable trebale povezivati s kojim faktorima; oni su jednostavno empirijski povezani. Podrazumijeva se da su oznake koje se daju dobivenim komponentama jednostavno pogodni opisi kombinacija varijabli u njima, te ne odražavaju nužno nikakve temeljne ishodišne procese.

POGLAVLJE 2

2. Pitanja koja se pojavljuju tijekom ispitivanja

Cilj istraživanja koja koriste analizu glavnih faktora i faktorsku analizu jest smanjivanje velikog broja varijabli na manji broj faktora, jezgrovito opisivanje (te možda razumijevanje) veza između promatranih varijabli ili testiranje hipoteze o temeljnim ishodišnim procesima.

2.1. Broj faktora

Koliko se pouzdanih i interpretabilnih faktora nalazi u skupu podataka? Koliko je faktora potrebno za sumariziranje obrasca korelacija u korelacijskoj matrici? Koliko je faktora pouzdano?

2.2. Priroda faktora

Koje je značenje faktora? Kako faktore treba interpretirati?

2.3. Važnost rješenja i faktora

Koliko varijance u skupu podataka objašnjavaju faktori? Koji faktori objašnjavaju najveći dio varijance? U dobroj faktorskoj analizi, nekoliko prvih faktora objašnjava visok postotak varijance u promatranim varijablama. Budući da se faktori računaju od najvećega prema najmanjemu, prvi faktor objašnjava najveći dio varijance, sljedeći faktori sve manje i manje varijance, sve dok ne prestanu biti pouzdanima.

2.4. Testiranje hipoteze u faktorskoj analizi

Koliko se dobiveno faktorsko rješenje podudara s očekivanim faktorskim rješenjem? Ako je istraživač postavio hipoteze o broju i prirodi faktora očekivanih od studenata na diplomskom studiju, usporedbe između hipotetskih faktora i faktorskih rješenja omogućuju testiranje postavljenih hipoteza.

2.5. Procjenjivanje rezultata za faktore

Da su faktori bili točno izmjereni, koje bi rezultate subjekti ostvarili za svaki od njih? Na primjer, kada bi se izravno mjerila neovisnost i inteligencija svakog radnika na platformi, kakve bi rezultate svaki radnik ostvario za te faktore?

Odgovorene na ova pitanja dati ćemo kasnije u radu.

POGLAVLJE 3

3. Ograničenja

3.1. Teoretski problemi

Većina primjena analize glavnih komponenti i faktorske analize po prirodi su istraživačke; faktorska analiza prvenstveno služi kao oruđe za smanjenje broja varijabli ili za ispitivanje obrazaca korelacija između varijabli. U ovakvim okolnostima, i teoretska i praktična ograničenja faktorske analize stavljena su po strani u korist izravnog istraživanja podataka. Odluke o broju faktora i rotacijskoj shemi donose se na temelju pragmatičkih, a ne teoretskih kriterija.

Istraživački projekt koji je osmišljen s namjerom da bude faktorski analiziran bitno se razlikuje od ostalih projekata na nekoliko načina. Jedna od najtemeljitijih analiza ovih razlika jest ona Comreyja i Leeja (1992.), odakle je uzet dio razmatranja koja slijede.

Prvi je zadatak istraživača proizvesti hipotezu o faktorima za koje vjeruje da su u osnovi domene istraživanja. Statistički, važno je da je interes istraživanja dovoljno širok da uključi pet ili šest hipotetskih faktora, kako bi rješenje bilo stabilno. Kako bi se otkrili procesi u osnovi područja istraživanja, logično je da svi relevantni faktori moraju biti uključeni. Izostavljanje nekog važnog faktora u mjerenju može dati iskrivljenu sliku odnosa između mjerenih faktora. Uključivanje svih relevantnih faktora predstavlja logički, no ne i statistički, problem za istraživača.

Istraživač zatim odabire varijable koje će promatrati. Za svaki hipotetski faktor odabire se pet ili šest varijabli za koje se smatra da su razmjerno čiste mjere faktora. Čiste mjere zovemo *varijablama markerima*. Varijable markeri u visokoj su korelaciji s jednim, i samo s jednim, faktorom i imaju visoko opterećenje za taj faktor bez obzira na metodu ekstrakcije ili rotacije. Varijable markeri korisne su jer jasno definiraju prirodu faktora; dodavanje potencijalnih varijabli faktoru kako bismo ga dopunili puno je smislenije ako je faktor već nedvosmisleno definiran varijablama markerima.

Složenost varijabli također se uzima u obzir. Složenost je određena brojem faktora s kojima je varijabla u korelaciji. Čista varijabla, koja se preferira, u korelaciji je samo s jednim faktorom, dok je složena varijabla u korelaciji s nekoliko faktora. Ako su varijable različite složenosti uključene u analizu, one sa sličnim razinama složenosti mogu „uloviti" jedna drugu u faktorima koji nemaju puno veze s procesima u osnovi. Varijable slične složenosti mogu korelirati jedna s drugom zbog svoje složenosti, a ne zato što su povezane s istim faktorom. Procjenjivanje (ili izbjegavanje) složenosti varijabli dio je postavljanja hipoteza o faktorima i odabira varijabli kojima ćemo ih mjeriti.

Istraživač koji planira istraživanje faktorskom analizom mora uzeti u obzir još nekoliko stavki. Važno je, primjerice, da subjekti u odabranom uzorku pokazuju različite rezultate za varijable i faktore koje varijable mjere. Ako svi subjekti ostvare približno jednak rezultat za neki faktor, korelacije su između promatranih varijabli niske i faktor se možda neće pojaviti u analizi. Odabir subjekata za koje se očekuje da će dati različite rezultate za promatrane varijable važna je stavka u planiranju istraživanja.

Također treba biti oprezan sa združivanjem rezultata nekoliko uzoraka, ili istoga uzorka mjerenog nekoliko puta s vremenskim odmacima, za potrebe faktorske analize. Prvo, uzorci za koje se zna da su različiti po nekom kriteriju (npr. socioekonomski status) možda imaju i različite faktore. Ispitivanje grupnih razlika često otkrije puno. Drugo, struktura faktora u osnovi može se s vremenom mijenjati za iste subjekte, kao posljedica učenja ili iskustva u eksperimentalnom okruženju, te ove razlike također mogu otkriti puno. Združivanje rezultata iz različitih skupina u faktorskoj analizi može otežati razumijevanje razlika, umjesto da ih objasni. S druge strane, ako različiti uzorci daju iste faktore, poželjno je združiti ih jer se povećava veličina uzorka. Primjerice, ako muškarci i žene daju iste faktore, treba spojiti uzorke i koristiti rezultate zajedničke faktorske analize.

3.2. Praktični problemi

Budući da su faktorska analiza i analiza glavnih komponenti iznimno osjetljive na veličine korelacija, presudno je da se koriste stvarne, pouzdane korelacije. Osjetljivost na slučajeve koji puno odskaču, problemi zbog podataka koji nedostaju te degradacija korelacija između loše distribuiranih varijabli problemi su koji prate faktorsku i analizu glavnih komponenti.

3.2.1. Veličina uzorka i podaci koji nedostaju

Koeficijenti korelacija često su manje pouzdani kada se utvrđuju iz malih uzoraka. Stoga, važno je da su uzorci dovoljno veliki za pouzdano utvrđivanje korelacija. Potrebna veličina uzorka također ovisi o veličini korelacija u populaciji te o broju faktora: ako postoje jake, pouzdane korelacije i nekoliko jasnih faktora, primjereni su manji uzorci.

Comrey i Lee (1992.) navode opće smjernice za veličinu uzorka: 50 je premalo, 100 je malo, 200 je prolazno, 300 je dobro, 500 je vrlo dobro te je 1000 odlično. Kao opće pravilo prakse, minimalno 300 slučajeva pruža relativnu pouzdanost u faktorskoj analizi. Rješenja koja imaju nekoliko varijabli markera s visokim opterećenjem (> 0.80) ne zahtijevaju tako velike uzorke (približno 150 slučajeva trebalo bi biti dovoljno), za razliku od rješenja s nižim opterećenjem i/ili manje varijabli markera (Guadagnoli i Velicer, 1988.)

Ako nedostaju podaci za neki slučaj, tada se slučaj briše ili se analizira korelacijska matrica (oznaka $\boldsymbol{R}$) s procjenjenim podacima za podatke koji nedostaju. Ako uzorku vrijednosti nedostaju po određenom uzorku ili bi preostali uzorak nakon brisanja postao premalen, preporučljiva je procjena. Međutim, treba pripaziti na korištenje postupaka procjene (kao što je regresija) koji će vjerojatno previše urediti podatke i proizvesti previsoke korelacije. Ovakvi postupci mogu „stvoriti“ faktore.

3.2.2. Normalitet

Sve dok se analiza glavnih komponenti i faktorska analiza koriste opisno kao prikladni načini sumariziranja odnosa u velikom skupu promatranih varijabli, pretpostavke o distribuciji varijabli nisu na snazi. Ako su varijable normalno distribuirane, rješenje se poboljšava. U onoj mjeri u kojoj se normalitet ne uspostavlja, rješenje se degradira, no još uvijek može biti od značaja.

3.2.3. Linearnost

Multivarijatni normalitet također implicira da je veza između parova varijabli linearna. Analiza se degradira kada linearnost opada, budući da korelacija mjeri linearne odnose i ne odražava nelinearne. Linearnost između parova varijabli ustanovljuje se uvidom u raspršne dijagrame.

3.2.4. Odsustvo slučajeva izvan prosjeka

Kao u svim multivarijatnim tehnikama, slučajevi mogu odskakati od prosjeka bilo na pojedinim varijablama (univarijatno), bilo na kombinacijama varijabli (multivarijatno). Takvi slučajevi imaju više utjecaja na faktorsko rješenje nego drugi slučajevi.

3.2.5. Odsutnost multikolinearnosti i singularnosti

U analizi glavnih komponenti, multikolinearnost nije problem budući da nema potrebe za invertiranjem matrice. Za većinu oblika faktorske analize i za utvrđivanje faktorskih vrijednosti u bilo kom obliku faktorske analize, singularnost ili ekstremna multikolinearnost predstavljaju problem. Za faktorsku analizu, ako se determinanta $\boldsymbol{R}$ i svojstvene vrijednosti povezane s nekim faktorima približavaju nuli, multikolinearnost ili singularnost mogu biti prisutne.

3.2.6. Faktorabilnost R

Matrica koja je faktorabilna trebala bi uključivati nekoliko značajnih korelacija. Očekivana veličina ovisi, u određenoj mjeri, od *N* (veći uzorci obično daju manje korelacije), ali ako nijedna korelacija ne prelazi 0.30, korištenje faktorske analize upitno je budući da vjerojatno nema ničega što bi se moglo faktorski analizirati.

Visoke bivarijatne korelacije, međutim, nisu nepobitan dokaz da korelacijska matrica sadrži faktore. Moguće je da su korelacije samo između dviju varijabli i ne odražavaju temeljne procese koji istovremeno utječu na nekoliko varijabli. Zbog ovoga je korisno ispitati matrice parcijalnih korelacija gdje su parne korelacije prilagođene za učinke svih drugih varijabli. Ako su prisutni faktori, onda visoke bivarijatne korelacije postaju vrlo niske parcijalne korelacije. SAS daje parcijalne korelacijske matrice.

Testovi značajnosti korelacija u korelacijskoj matrici daju naznaku pouzdanosti veza između parova varijabli. Ako je ***R*** faktorabilan, brojni parovi su značajni. Anti-image korelacijska matrica sadrži negativne vrijednosti parcijalnih korelacija između parova varijabli uz uklonjene efekte drugih varijabli. Ako je ***R*** faktorabilan, među nedijagonalnim elementima anti-image matrice nalaze se većinom male vrijednosti. Konačno, Kaiserova mjera za adekvatnost uzorkovanja jest omjer između zbroja kvadriranih korelacija i zbroja kvadriranih korelacija plus zbroj parcijalnih korelacija. Vrijednost se približava 1 ako su parcijalne korelacije male. Za dobru faktorsku analizu potrebne su vrijednosti iznad 0.6.

3.2.7. Odsutnost rubnih slučajeva

Nakon faktorske analize, i u eksploratornoj i u konfirmatornoj faktorskoj analizi, identificiraju se varijable koje nisu povezane s ostalima u skupu. Ove varijable obično nisu u korelaciji s nekoliko prvih faktora, iako često koreliraju s faktorima dobivenima poslije. Ovi faktori obično su nepouzdani, i zato što su odgovorni za vrlo malo varijance i zato što faktori definirani samo jednom ili dvjema varijablama nisu stabilni. Stoga, nikad se ne može znati jesu li ovi faktori stvarni. Ako je varijanca za koju je odgovoran faktor koji je definiran samo jednom ili dvjema varijablama dovoljno visoka, faktor se interpretira s velikom rezervom ili se ignorira, ovisno kako praktični razlozi nalažu. U konfirmatornoj faktorskoj analizi, faktor predstavlja ili obećavajući trag za budući rad ili (vjerojatno) varijancu pogreške, no za njegovu interpretaciju potrebno je daljnje istraživanje.

Varijabla s niskom kvadriranom višestrukom korelacijom sa svim ostalim varijablama i niskim korelacijama sa svim važnim faktorima jest rubni slučaj među varijablama. Varijabla se obično ignorira u trenutnoj faktorskoj analizi te ili briše, ili se spaja s drugim varijablama u budućem istraživanju.

POGLAVLJE 4

4. Fundamentalne jednadžbe za faktorsku analizu

Zbog brojnosti i složenosti proračuna uključenih u pripremu korelacijske matrice, dobivanje faktora i njihovog rotiranja, te zato što se, po našem sudu, ne dobivaju značajni novi uvidi demonstracijom nekih od ovih postupaka, ovaj odjeljak ne bavi se njima.

Tablica 1 navodi mnoge važne matrice u faktorskoj analizi i analizi glavnih komponenti. Iako je popis podugačak, sastoji se većinom od *matrica faktorskih korelacija* (između varijabli, između faktora, i između varijabli i faktora), *matrica standardiziranih vrijednosti* (za varijable i za faktore), *matrica faktorskih vrijednosti* (za dobivanje vrijednosti za faktore iz vrijednosti za varijable) i *matrica uzoraka* jedinstvenih odnosa između faktora i varijabli nakon kose rotacije.

Tablica 1

Oznaka	Naziv	Red matrice
$\boldsymbol{R}$	korelacijska matrica	$p \times p$
$\boldsymbol{Z}$	matrica standardiziranih vrijednosti	$N \times p$
$\boldsymbol{F}$	matrica faktorskih vrijednosti	$N \times m$
$\boldsymbol{A}$	matrica uzoraka matrica opterećenja	$p \times m$
$\boldsymbol{B}$	matrica koeficijenata faktorskih vrijednosti	$p \times m$
$\boldsymbol{C}$	matrica faktorske strukture	$p \times m$
$\boldsymbol{\Phi}$	faktorska korelacijska matrica	$m \times m$
$\boldsymbol{L}$	matrica svojstvenih vrijednosti	$m \times m$
$\boldsymbol{V}$	matrica svojstvenih vektora	$p \times m$

p – broj varijabli

N – broj ispitanika

m – broj faktora ili komponenti

K1 (poznato kao i Kaiser-Gutmanovo pravilo) najpoznatija je procjena za određivanje broja faktora. Ovo pravilo sugerira da sve komponente sa svojstvenim vrijednostima većima od 1 treba zadržati u analizi. Nakon što se odredi broj komponenti vrši se daljna analiza.

4.1. Ekstrakcija

Važan teorem iz algebre matrica upućuje na to da se, pod određenim uvjetima, matrice mogu dijagonalizirati. Korelacijske matrice među onima su koje se često mogu dijagonalizirati. Kada se matrica dijagonalizira, transformira se u matricu s brojevima na pozitivnoj dijagonali[6] i nulama na ostalim mjestima. U ovoj primjeni, brojevi na pozitivnoj dijagonali predstavljaju varijancu iz korelacijske matrice koja se dobiva na sljedeći način:

$$\boldsymbol{L} = \boldsymbol{V}'\boldsymbol{R}\boldsymbol{V} \tag{1}$$

Stupce u $\boldsymbol{V}$ zovemo svojstvenim vektorima, a vrijednosti na glavnoj dijagonali $\boldsymbol{L}$ svojstvenim vrijednostima. Prvi svojstveni vektor odgovara prvoj svojstvenoj vrijednosti, itd.

Matrica svojstvenih vektora pomnožena svojom transponiranom matricom daje matricu identiteta s jedinicama na pozitivnoj dijagonali[3] i nulama drugdje. Stoga, prethodno i naknadno množenje korelacijske matrice svojstvenim vektorima više je preuređuje, nego što je mijenja.

$$\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I} \tag{2}$$

Važno je to da, zato što se korelacijske matrice često mogu dijagonalizirati, moguće je na njima primijeniti algebru matrica svojstvenih vektora i svojstvenih vrijednosti s faktorskom analizom kao rezultatom. Kada se matrica dijagonalizira, informacija u njoj se preuređuje. U faktorskoj analizi, varijanca u korelacijskoj matrici kondenzira se u svojstvene vrijednosti. Faktor s najvećom svojstvenom vrijednosti ima najviše varijance itd., sve do faktora s malim ili negativnim svojstvenim vrijednostima koji se obično ispuštaju iz rješenja.

Proračuni za svojstvene vektore i svojstvene vrijednosti zahtijevaju puno rada i ne donose posebno važne uvide. Zahtijevaju rješavanje p jednadžbi s p nepoznanica

[3] Pozitivna dijagonala ona je od gornjeg lijevog do donjeg desnog kuta u matrici.

s dodatnim ograničenjima i rijetko se provode ručno. Međutim, jednom kad su poznate svojstvene vrijednosti i svojstveni vektori, ostatak faktorske analize (ili analize glavnih komponenti) više-manje „ispada“, kao što je vidljivo iz jednadžbi (3) do (6).

Jednadžba (1) može se zapisati:

$$\boldsymbol{R} = \boldsymbol{V}\boldsymbol{L}\boldsymbol{V}' \tag{3}$$

Korelacijska matrica može se smatrati umnoškom triju matrica – matrica svojstvenih vrijednosti i odgovarajućih svojstvenih vektora.

Nakon reorganizacije, uzima se drugi korijen matrice svojstvenih vrijednosti.

$$\boldsymbol{R} = \boldsymbol{V}\sqrt{\boldsymbol{L}}\sqrt{\boldsymbol{L}}\boldsymbol{V}' \tag{4}$$

Označimo $\boldsymbol{V}\sqrt{\boldsymbol{L}}$ s $\boldsymbol{A}$, i $\sqrt{\boldsymbol{L}}\boldsymbol{V}'$ s $\boldsymbol{A}'$, tada je

$$\boldsymbol{R} = \boldsymbol{A}\boldsymbol{A}' \tag{5}$$

Korelacijska matrica može se također smatrati umnoškom dviju matrica, od kojih je svaka kombinacija svojstvenih vektora i drugog korijena svojstvenih vrijednosti.

Jednadžba (5) često se naziva fundamentalnom jednadžbom faktorske analize[4]. Predstavlja tvrdnju da je korelacijska matrica umnožak matrice uzorka, $\boldsymbol{A}$, i njene transponirane matrice.

Jednadžbe (4) i (5) također otkrivaju da je glavni posao faktorske analize (i analize glavnih komponenti) računanje svojstvenih vrijednosti i svojstvenih vektora. Jednom kada su poznati, (nerotirana) matrica faktorskog opterećenja dobiva se izravnim množenjem matrica, kao što slijedi:

$$\boldsymbol{A} = \boldsymbol{V}\sqrt{\boldsymbol{L}} \tag{6}$$

[4]Za točnu reprodukciju matrice, kao što je naznačeno u jednadžbama (4) i (5), potrebne su sve svojstvene vrijednosti i svojstveni vektori, ne samo prvih nekoliko.

4.2. Ortogonalna rotacija

Rotacija se redovno koristi nakon ekstrakcije za maksimiziranje visokih korelacija i minimiziranje niskih. Brojne su metode rotacije dostupne, no najčešće korištena jest *varimax*. Varimax je postupak za maksimiziranje varijance. Cilj je varimax rotacije maksimizirati varijancu opterećenja faktora povišenjem visokih opterećenja i snižavanjem niskih za svaki faktor.

Ovo se postiže uz pomoć transformacijske matrice $\boldsymbol{\Lambda}$ (definirane u jednadžbi (8)) gdje je

$$\boldsymbol{A_{nerotirano}\Lambda = A_{rotirano}} \tag{7}$$

Nerotirana matrica faktorskog opterećenja množi se s transformacijskom matricom kako bi se dobila rotirana matrica opterećenja.

Brojevi u transformacijskoj matrici imaju prostornu interpretaciju.

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{cos\psi} & \boldsymbol{-sin\psi} \\ \boldsymbol{sin\psi} & \boldsymbol{cos\psi} \end{bmatrix} \tag{8}$$

Transformacijska matrica jest matrica sinusa i kosinusa kuta $\boldsymbol{\psi}$.

4.3. Vrijednosti faktora

Vrijednosti za faktore mogu se predvidjeti za svaki slučaj jednom kada je dostupna matrica opterećenja. Koeficijenti nalik na regresijske računaju se za težinske vrijednosti varijabli da bi se dobile vrijednosti faktora. Zato što je $\boldsymbol{R^{-1}}$ inverzna matrica korelacijske matrice među varijablama i $\boldsymbol{A}$ je matrica korelacija između faktora i varijabli, jednadžba (9) za koeficijente faktorskih vrijednosti slična je jednadžbi (6) za regresijske koeficijente u višestrukoj regresiji.

$$\boldsymbol{B = R^{-1}A} \tag{9}$$

Koeficijenti faktorskih vrijednosti za određivanje vrijednosti faktora iz vrijednosti varijabli umnožak su inverza korelacijske matrice i matrice faktorskih opterećenja.

Vrijednosti faktora umnožak su standardiziranih vrijednosti za varijable i koeficijenata faktorskih vrijednosti.

$$\boldsymbol{F} = \boldsymbol{ZB} \tag{10}$$

Također je moguće predvidjeti vrijednosti varijabli iz vrijednosti za faktore. Jednadžba za ovo je

$$\boldsymbol{Z} = \boldsymbol{FA'} \tag{11}$$

4.4. Kosa rotacija

Svi dosad spomenuti odnosi vrijede za ortogonalnu rotaciju. Kada se koristi kosa (korelirana) rotacija, većina složenosti ortogonalne rotacije ostaju te ih se stvara još nekoliko.

U kosoj rotaciji, matrica opterećenja postaje matrica uzoraka. Vrijednosti u matrici uzoraka, kada se kvadriraju, predstavljaju jedinstven doprinos svakog faktora varijanci svake varijable, no ne uključuju segmente varijance koji proizlaze iz preklapanja između koreliranih faktora.

: Jednom kada su faktorski rezultati utvrđeni, mogu se dobiti korelacije između faktora. Među jednadžbama korištenima u tu svrhu je

$$\boldsymbol{\phi} = \left(\frac{\boldsymbol{1}}{\boldsymbol{N-1}}\right)\boldsymbol{F'F} \tag{12}$$

Matrica faktorskih korelacija standardni je dio računalnog rješenja nakon kose rotacije.

Međutim, ako se koristi kosa rotacija, matrica faktorske strukture $\boldsymbol{C}$, predstavlja korelacije između varijabli i faktora. Ove korelacije utvrđuju jedinstvenu vezu između varijable i faktora (u matrici uzoraka) plus odnos između varijable i preklapanja varijance među faktorima. Jednadžba za matricu faktorske strukture je

$$\boldsymbol{C} = \boldsymbol{A\phi} \tag{13}$$

Strukturalna matrica umnožak je matrice uzoraka i matrice faktorskih korelacija.

Postoje neka neslaganja oko toga treba li interpretirati matricu uzorka ili strukturalnu matricu nakon kose rotacije. Strukturalna je matrica pogodna jer je lako razumljiva. Međutim, korelacije između varijabli i faktora umjetno se povećavaju u slučaju bilo kakvog preklapanja između faktora. Problem postaje još ozbiljniji kada se korelacije među faktorima povećaju i može biti teško odrediti koje su varijable povezane s faktorom. S druge strane, matrica uzoraka sadrži vrijednosti koje predstavljaju jedinstvene doprinose svakog faktora varijanci u varijablama. Zajednička se varijanca izbacuje (kao i u standardnoj višestrukoj regeresiji), no skup varijabli koje čine faktor obično je lakše vidjeti. Ako su faktori u vrlo visokoj korelaciji, može se činiti da nema varijabli povezanih s njima, jer jednom kada se preklapanje eliminira, gotovo da nema jedinstvene varijance.

Većina istraživača radije interpretira i dostavlja matricu uzoraka nego strukturalnu matricu. Međutim, ako istraživač dostavi bilo koju od matrica, strukturalnu ili matricu uzoraka te $\boldsymbol{\phi}$, zainteresirani čitatelj može dobiti onu drugu uz pomoć jednadžbe (13).

U kosoj rotaciji, reducirana korelacijska matrica $\overline{\boldsymbol{R}}$ dobiva se na sljedeći način:

$$\overline{\boldsymbol{R}} = \boldsymbol{CA'} \tag{14}$$

Reducirana korelacijska matrica umnožak je strukturalne matrice i transponirane matrice uzoraka.

POGLAVLJE 5

5. Glavne vrste faktorske analize

5.1. Tehnike ekstrakcije faktora

Među tehnikama ekstrakcije nalaze se analiza glavnih komponenata (eng. *primary component analysis* – PCA), analiza zajedničkih faktora (eng. *principal factor analysis* – PFA), metoda maksimalne vjerodostojnosti, *image* ekstrakcija, alfa ekstrakcija i metoda neponderiranih i ponderiranih najmanjih kvadrata. Najčešće se primjenjuju metode PCA i analiza zajedničkih faktora.

Sve ekstrakcijske tehnike izračunavaju set ortogonalnih komponenata ili faktora koji u kombinaciji reproduciraju $\boldsymbol{R}$. Kriteriji koji se koriste za uspostavljanje rješenja, poput maksimiziranja varijance ili minimiziranja rezidualnih korelacija, razlikuju se od metode do metode. No razlike u rješenjima male su u slučajevima setova podataka s velikim uzorkom, brojnim varijablama i sličnim procjenama komunaliteta. Komunalitet je dio ukupne varijance koji je uvjetovan zajedničkim faktorima tj. onaj dio varijance koji varijabla dijeli s drugim varijablama. Komunalitet $i-$te varijable se računa na način da se zbroji suma kvadrata opterećenja za tu varijablu. Štoviše, jedan od testova stabilnosti rješenja za faktorsku analizu jest da se rješenje treba pojaviti bez obzira na primijenjenu tehniku ekstrakcije.

Nijedna ekstrakcijska tehnika rutinski ne daje rješenje bez rotacije. Sve vrste ekstrakcije mogu se rotirati bilo kojom od procedura objašnjenih u dijelu 5.2.

Konačno, prilikom korištenja faktorske analize istraživač bi trebao privremeno zanemariti poznate propise protiv višestrukog korištenja podataka radi odabira konačnog modela. Sasvim je uobičajeno koristiti PCA ili PFA kao preliminarnu tehniku za ekstrakciju pa potom jednu ili više drugih procedura, mijenjajući možda broj faktora, procjene komunaliteta i metode rotacije sa svakim pokušajem. Analiza završava kad se istraživač odluči za rješenje koje mu najviše odgovara.

POGLAVLJE 5:GLAVNE VRSTE FAKTORSKE ANALIZE

5.1.1. PCA i FA

Jedna od najvažnijih odluka izbor je između PCA i FA. Matematički, razlika je u sadržaju pozitivne dijagonale u korelacijskoj matrici (dijagonala koja sadrži korelaciju između varijable i sebe same). U slučaju i PCA i FA varijanca koja se analizira zbroj je vrijednosti na pozitivnoj dijagonali. Sva varijanca raspoređena je na komponente, uključujući varijancu pogreške i jedinstvenu varijancu za svaku promatranu varijablu. Ako zadržimo sve komponente, PCA kopira promatranu korelacijsku matricu i standardne vrijednosti promatranih varijabli.

Kod faktorske analize za analizu je dostupna samo varijanca koju svaka promatrana varijabla dijeli s drugim promatranim varijablama. Isključivanje varijance pogreške i jedinstvene varijance iz FA temelji se na vjerovanju da takva varijanca samo remeti sliku o temeljnim procesima. Dijeljena varijanca procjenjuje se po komunalitetima, vrijednostima između 0 i 1 koje se ubacuju na pozitivnu dijagonalu korelacijske matrice (ekstrakcija metodom maksimalne vjerodostojnosti manipulira elementima izvan dijagonale umjesto vrijednostima na dijagonali). Rješenje kod FA koncentrira se na varijable s visokim vrijednostima komunaliteta. Zbroj je komunaliteta (zbroj SSL-ova) varijanca koja je distribuirana među faktorima i iznosi manje od ukupne varijance u setu promatranih varijabli. Zbog izostavljanja jedinstvene varijance i varijance pogreške, linearna kombinacija faktora aproksimira, ali ne duplicira promatranu korelacijsku matricu i vrijednosti na promatranim varijablama.

PCA analizira varijancu; FA analizira kovarijancu (komunalitet). Cilj je PCA ekstrahirati maksimalnu varijancu iz seta podataka s nekoliko ortogonalnih komponenata. Cilj je FA reproducirati korelacijsku matricu s nekoliko ortogonalnih faktora. PCA je jedinstveno matematičko rješenje, dok većina oblika FA nije.

Izbor između PCA i FA ovisi o procjeni prikladnosti modela, o setu podataka i ciljevima istraživanja. Ako ste zainteresirani za teorijsko rješenje, a istraživanje se bazira na temeljnim konstruktima koji bi trebali proizvesti rezultate na vašim promatranim varijablama, FA je bolji izbor. Ako samo želite empirijski sažetak seta podataka, PCA je bolje rješenje.

5.1.2. Analiza glavnih komponenata

Cilj je PCA ekstrahirati maksimalnu varijancu iz seta podataka sa svakom komponentom. Prva glavna komponenta linearna je kombinacija promatranih vrijednosti koja maksimalno razdvaja subjekte maksimizirajući varijancu njihovih komponentnih vrijednosti. Druga se komponenta oblikuje iz rezidualnih korelacija; to je linearna kombinacija promatranih varijabli koja ekstrahira maksimalnu varijabilnost koja ne korelira s prvom komponentom. Sljedeće komponente također ekstrahiraju maksimalnu varijabilnost iz rezidualnih korelacija i ortogonalne su na sve prethodno ekstrahirane komponente.

Glavne se komponente poredaju, tako da prva komponenta ekstrahira najviše varijance, a posljednja komponenta najmanje. Rješenje je matematički jedinstveno i, ako se zadrže sve komponente, točno reproducira promatranu korelacijsku matricu. Nadalje, s obzirom da su komponente ortogonalne, njihovo korištenje u drugim analizama može uvelike olakšati tumačenje rezultata.

PCA je optimalno rješenje za istraživača koji je prvenstveno zainteresiran za smanjenje velikog broja varijabli na manji broj komponenata. PCA je korisna i kao početni korak u FA gdje otkriva mnogo o maksimalnom broju i prirodi faktora.

5.1.3. Analiza zajedničkih faktora

Analiza zajedničkih faktora razlikuje se od PCA u tome što su procjene komunaliteta, umjesto jedinica, na pozitivnoj dijagonali u promatranoj korelacijskoj matrici. Ove se procjene dobivaju iz iterativnog postupka, sa SMC-ovima (kvadrati višestruke korelacije – eng. *squared multiple correlations* – svake varijable sa svim ostalim varijablama) koji se koriste kao početne vrijednosti u iteraciji. Cilj je analize, kao i u PCA, ekstrahirati maksimalnu ortogonalnu varijancu iz seta podataka sa svakim sljedećim faktorom. Prednosti su ove analize što se naširoko koristi (i razumije) i što je sukladna s faktorskim analitičkim modelom u kojem se zajednička varijanca analizira bez jedinstvene i varijance pogreške. Međutim, s obzirom da je cilj maksimizirati ekstrahiranu varijancu, analiza zajedničkih faktora ponekad u reprodukciji korelacijske matrice nije jednako dobra kao ostale ekstrakcijske tehnike. Osim toga komunaliteti se moraju procijeniti, a rješenje je donekle određeno tim procjenama.

5.1.4. Faktorska analiza uporabom image ekstrakcije

Image ekstrakcija je zanimljiv kompromis između PCA i FA. Poput PCA *image* ekstrakcija pruža matematički jedinstveno rješenje jer postoje fiksne vrijednosti na pozitivnoj dijagonali ***R***. Poput analize zajedničkih faktora, vrijednosti na dijagonali su komunaliteti bez jedinstvene varijabilnosti i varijabilnosti pogreške.

Odraženi rezultati za svaku varijablu proizvode se višestrukom regresijom u kojoj svaka varijabla služi kao zavisna varijabla. Matrica kovarijance izračunava se iz takvih odraženih (predviđenih) rezultata. Varijance iz matrice kovarijance odraženog rezultata komunaliteti su za ekstrakciju faktora. Potrebna je pažnja pri tumačenju rezultata *image* analize jer opterećenja predstavljaju kovarijance između varijabli i faktora, umjesto korelacija.

5.1.5. Ekstrakcija faktora metodom maksimalne vjerodostojnosti

Ekstrakcija faktora metodom maksimalne vjerodostojnosti procjenjuje vrijednosti populacije za faktorska opterećenja računajući opterećenja koja maksimiziraju vjerojatnost uzimanja uzorka upravo promatrane korelacijske matrice iz populacije. Unutar ograničenja koja nameću korelacije između varijabli izračunavaju se procjene populacije za faktorska opterećenja koje imaju najveću vjerojatnost da daju uzorak s promatrane korelacijske matrice. Ova metoda ujedno i maksimizira kanonske korelacije između varijabli i faktora.

5.1.6. Metoda neponderiranih najmanjih kvadrata

Cilj je ove metode minimizirati kvadrirane razlike između promatrane i reproducirane korelacijske matrice. Razmatraju se samo razlike izvan dijagonale; komunaliteti se dobivaju iz rješenja umjesto da se procjenjuju kao dio rješenja. Iz tog razloga možemo na metodu neponderiranih najmanjih kvadrata gledati kao na poseban slučaj analize zajedničkih faktora u kojoj se komunaliteti procjenjuju nakon rješenja.

5.1.7. Metoda ponderiranih najmanjih kvadrata

Metoda ponderiranih najmanjih kvadrata također pokušava minimizirati kvadrirane razlike (izvan dijagonale) između promatrane i reproducirane korelacijske matrice, ali u ovom slučaju ponderi se primjenjuju na varijable. Varijable koje imaju značajnu dijeljenu varijancu s drugim varijablama dobivaju teže pondere od varijabli koje imaju značajnu jedinstvenu varijancu. Drugim riječima, varijable koje nisu toliko povezane s drugim varijablama u setu nisu toliko važne za rješenje.

5.1.8. Alfa ekstrakcija

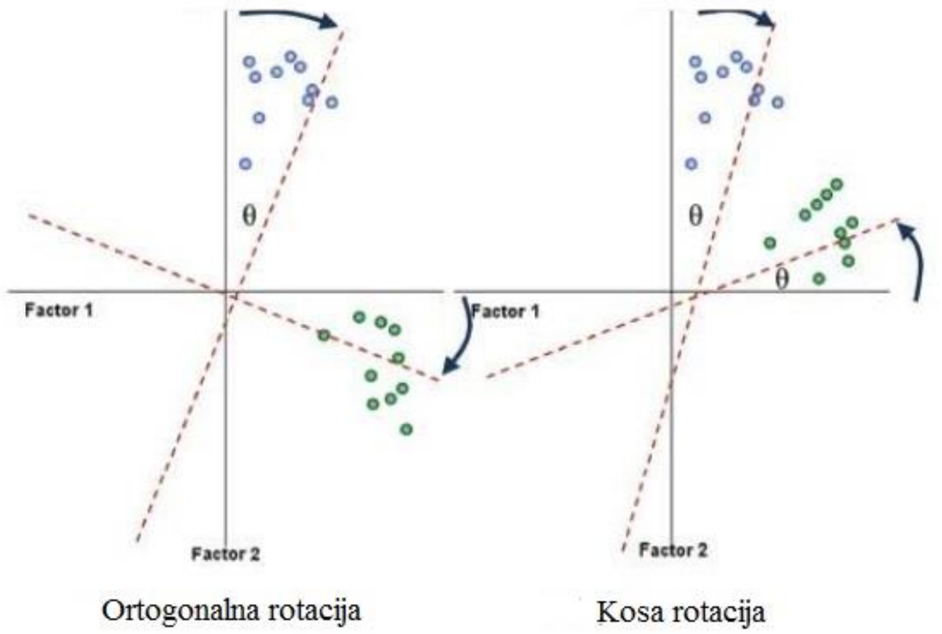
Alfa ekstrakcija faktora razvila se iz psihometrijskih istraživanja u kojima je glavni fokus na otkrivanju činjenice koji se zajednički faktori dosljedno pronalaze kad se ponovljeni uzorci varijabli uzmu iz populacija varijabli. Problem je jednak prepoznavanju srednjih razlika koje se dosljedno pronalaze među uzorcima subjekata uzetim iz populacije subjekata – središnje pitanje većine univarijatne i multivarijatne statistike.

Kod alfa ekstrakcije problem nije pouzdanost grupnih razlika, nego pouzdanost zajedničkih faktora. Koeficijent alfa mjera je proizašla iz psihometrije radi pouzdanosti rezultata koji se uzimaju u različitim situacijama. Komunaliteti koji maksimiziraju koeficijent alfa za faktore procjenjuju se upotrebom iterativnih postupaka (i ponekad prelaze 1.0).

Vjerojatno je najveća prednost ovog postupka to što usmjerava pažnju istraživača izravno na problem uzimanja uzorka varijabli iz domene varijabli od interesa. Nedostatak je relativne neupoznatost većine istraživača s postupkom i primjenom.

5.2. Rotacija

Slika 1. Prikaz ortogonane i kose rotacije



Rezultati ekstrakcije faktora bez rotacije su komplicirani za tumačenje neovisno o metodi ekstrakcije. Nakon ekstrakcije koristi se rotacija koja poboljšava mogućnost interpretacije i znanstvenu korist od rješenja. Ne koristi se za poboljšanje kvalitete matematičkog slaganja između promatrane i reproducirane korelacijske matrice jer su sva ortogonalno rotirana rješenja matematički ekvivalentna jedno drugom, kao i rješenju prije rotacije.

Kao što različite metode ekstrakcije obično daju slične rezultate uz dobar set podataka, tako i različite metode rotacije obično daju slične rezultate ako je uzorak korelacija u podacima relativno jasan. Drugim riječima, stabilno rješenje pojavljuje se neovisno o metodi rotacije.

Potrebno je odlučiti se između ortogonalne i kose rotacije. Kod ortogonalne rotacije faktori ne koreliraju. Ortogonalna rješenja jednostavna su za tumačenje, opisivanje i izvještavanje rezultata; ipak, ne odražavaju pravu stvarnost osim ako je istraživač uvjeren da su ishodišni procesi gotovo neovisni. Istraživač koji smatra da ishodišni procesi koreliraju koristit će kosokutnu rotaciju. Kod takve rotacije faktori mogu biti u korelaciji, uz konceptualne prednosti, ali i praktične nedostatke pri tumačenju, opisivanju i izvještavanju rezultata.

5.2.1. Ortogonalna rotacija

Varimax, quartimax i equamax – tri ortogonalne tehnike . Varimax je daleko najčešće upotrebljavana od svih dostupnih tehnika.

Kao što ekstrakcijske tehnike imaju neznatno različite statističke ciljeve, tako i rotacijske tehnike maksimiziraju ili minimiziraju određene statistike. Cilj je varimax rotacije pojednostaviti faktore maksimiziranjem varijance opterećenja unutar faktora, preko varijabli. Opterećenja koja su visoka nakon ekstrakcije postaju viša nakon rotacije, a niska opterećenja postanu niža. Tumačenje faktora je jednostavnije jer je očigledno koja varijabla s njim korelira. Osim toga varimax rotacija obično distribuira varijancu po faktorima tako da oni postanu relativno jednaki po važnosti; varijanca se uzima od prvih ekstrahiranih faktora i distribuira među kasnijima.

Quartimax djeluje na varijable na isti način kao što varimax djeluje na faktore. Pojednostavljuje varijable povećanjem disperzije opterećenja unutar varijabli, preko faktora. Varimax djeluje na stupce matrice opterećenja, quartimax na redove. Quartimax nije ni približno popularna kao varimax jer su istraživači obično više zainteresirani za jednostavne faktore nego za jednostavne varijable.

Equamax je hibrid između varimaxa i quartimaxa koji pokušava istovremeno pojednostaviti i faktore i varijable. Mulaik (1972) izvještava da se equamax može ponašati nepredvidivo ako istraživač ne može pouzdano navesti broj faktora.

Pojednostavljivanje faktora i varijabli vrši se postavljanjem razina na kriterij jednostavnosti – primjerice, Γ (gama) – od 1, 0, 1/2. Gama može i konstantno varirati između 0 (pojednostavljene su varijable) i 1 (pojednostavljeni su faktori) pomoću ortogonalne rotacije koja omogućuje korisniku da unese razinu Γ.

5.2.2. Kosa rotacija

Kosa rotacija daje kontinuirani raspon korelacija između faktora.. Kad je vrijednost manja od nule, rješenja su sve više ortogonalna; na otprilike − 4 rješenje je ortogonalno. Kad je vrijednost nula, rješenja mogu imati prilično visoke korelacije. Vrijednosti blizu 1 mogu proizvesti faktore s vrlo visokom korelacijom.

Treba istaknuti da faktori ne koreliraju nužno kad se koristi kosa rotacija. Zapravo vrlo često ne koreliraju pa istraživač podnese izvješće o jednostavnijoj ortogonalnoj rotaciji.

Kosa rotacija koristi algoritam quartimax za proizvodnju ortogonalnog rješenja na prilagođenim faktorskim opterećenjima; stoga rješenje može biti koso s obzirom na izvorna faktorska opterećenja.

Kod promax rotacije ortogonalno rotirano rješenje (obično varimax) rotira se ponovno kako bi se omogućile korelacije između faktora. Ortogonalna opterećenja dižu se na potencije (obično drugu, četvrtu ili šestu) da bi smanjila mala i srednja opterećenja na nulu dok se veća opterećenja smanjuju, ali ne na nulu. Iako faktori koreliraju, jednostavna je struktura maksimizirana pojašnjenjem činjenice koji faktori koreliraju, a koji ne koreliraju sa svakim faktorom. Promax ima dodatnu prednost u vidu brzine i povoljne cijene.

Kod Procrustove metode, dostupne u SAS-u, istraživač navede ciljnu matricu opterećenja (obično nule i jedinice), a traži se transformacijska matrica radi rotacije ekstrahiranih faktora na ciljnu, ako je to moguće. Ako se rješenje može rotirati prema cilju, onda se kaže da je hipotetska struktura faktora potvrđena. Nažalost, kod Procrustove rotacije faktori su često u ekstremno visokoj korelaciji i ponekad korelacijska matrica nastala nasumičnim procesima iznimno lako rotira na ciljnu matricu.

Ekstrakcija faktora daje rješenje u kojem su promatrane varijable vektori koji se protežu od početne točke do točaka označenih koordinatnim sustavom. Faktori služe kao osi sustava. Koordinate svake točke unosi su iz matrice opterećenja pojedine varijable. Ako postoje tri faktora, prostor tada ima tri osi i tri dimenzije, a svaka promatrana varijabla određena je trima koordinatama. Duljina vektora svake varijable njen je komunalitet.

Ako su faktori ortogonalni, sve su osi faktora pod pravim kutovima jedna na drugu, a koordinate točaka varijabli korelacije su između zajedničkih faktora i promatranih varijabli. Korelacije (faktorska opterećenja) očitavaju se izravno iz ovih grafikona projekcijom okomitih linija od svake točke na osi faktora.

Jedan od glavnih ciljeva PCA i PFA te motivacija koja stoji iza ekstrakcije otkriti je minimalni broj faktorskih osi koje su potrebne za pouzdano pozicioniranje varijabli. Drugi glavni cilj i motivacija za rotaciju otkrivanje je značenja faktora koji su baza odgovora na promatrane varijable. Ovaj se cilj ostvaruje tumačenjem faktorskih osi koje se koriste za definiranje prostora. Rotacija faktora premješta faktorske osi na način da ih čini maksimalno protumačivima. Premještanje osi mijenja koordinate točaka varijabli, ali ne i međusobni odnos pozicija točaka.

Faktori se obično mogu protumačiti kad neke od promatranih varijabli imaju visoko opterećenje na njih, a ostale ne. U idealnom slučaju svaka varijabla opterećuje samo jedan faktor. Grafički to znači da točka koja predstavlja svaku varijablu stoji daleko na jednoj osi, ali blizu ishodišta na drugim osima, tj. da su koordinate točke velike za jednu os, a blizu nule na drugim osima.

Ako promatramo samo jednu varijablu, pozicioniranje je faktorske osi vrlo jednostavno. Međutim, u slučaju brojnih varijabli i nekoliko faktorskih osi, potrebni su kompromisi pri pozicioniranju osi. Varijable formiraju „roj" u kojem varijable koje međusobno koreliraju čine skupinu (klaster) točaka. Cilj je postaviti os na roj točaka. Uz malo sreće, rojevi su jedan od drugog na oko 90°, što inducira ortogonalno rješenje. A uz mnogo sreće, varijable se okupljaju u samo nekoliko rojeva uz prazan prostor između njih, pa su faktorske osi lijepo definirane.

Kod kose rotacije situacija je nešto kompliciranija. Kako faktori mogu međusobno korelirati, faktorske osi nisu nužno pod pravim kutovima. Iako je lakše pozicionirati svaku os blizu skupine točaka, osi mogu biti međusobno vrlo blizu (visoka korelacija), što čini rješenje teškim za tumačenje.

5.3. Praktične preporuke

Kada postoji velik broj varijabli s visokom međusobnom korelacijom, uz isti, dobro odabran broj faktora te uz slične vrijednosti komunaliteta, rezultati ekstrakcije slični su bez obzira na korištenu metodu. Nadalje, razlike koje su nakon ekstrakcije očite obično nestaju nakon rotacije.

Većina istraživača započinje faktorsku analizu korištenjem analize glavnih komponenata i varimax rotacije. Iz rezultata se procjenjuje faktorabilnost korelacijske matrice, rang promatrane korelacijske matrice i broj faktora te varijable koje bi se mogle isključiti iz daljnjih analiza.

Tijekom sljedećih nekoliko analiza, istraživači eksperimentiraju s različitim brojem faktora, različitim ekstrakcijskih tehnikama te s ortogonalnom i kosokutnom rotacijom. Određeni broj faktora s određenom kombinacijom ekstrakcije i rotacije daje rješenje koje ima najveću znanstvenu korist, konzistentnost i značenje; to je tada rješenje koje se tumači.

POGLAVLJE 6

6. Važna pitanja

Neka od pitanja koja se pojavljuju u ovom dijelu mogu se riješiti uz nekoliko različitih metoda. Obično različite metode vode do istog zaključka; ponekad i ne. U ovim drugim slučajevima, rezultati se procjenjuju mogućnošću tumačenja i znanstvenom koristi od rješenja.

6.1. Procjene komunaliteta

PFA se razlikuje od PCA u tome što vrijednosti komunaliteta (brojevi između 0 i 1) zamjenjuju jedinice na pozitivnoj dijagonali ***R*** prije ekstrakcije faktora. Vrijednosti komunaliteta koriste se umjesto jedinica kako bi se uklonila jedinstvena varijanca i varijanca pogreške svake promatrane varijable; samo se varijanca koju varijabla dijeli s faktorima koristi u rješenju. No vrijednosti se komunaliteta procjenjuju, a postoje određene rasprave o tome postupku.

SMC svake varijable kao zavisne varijable s ostalima u uzorku kao nezavisne varijeble obično je početna procjena komunaliteta. Kako se rješenje razvija, procjene komunaliteta prilagođavaju se iterativnim procedurama (kojima upravlja istraživač) da bi reducirana matrica odgovarala promatranoj korelacijskoj matrici s najmanjim brojem faktora. Iteracija se prekida kad uzastopne procjene komunaliteta postanu vrlo slične.

Konačne procjene komunaliteta također su SMC-ovi, ali sada između svake varijable kao zavisne i faktora kao nezavisnih. Konačne vrijednosti komunaliteta predstavljaju omjer varijance u varijabli koji je predvidiv iz ishodišnih faktora. Procjene komunaliteta ne mijenjaju se ortogonalnom rotacijom.

Image ekstrakcija i metoda maksimalne vjerodostojnosti neznatno se razlikuju. Kod *image* ekstrakcije varijance iz *image* matrice kovarijance koriste se kao vrijednosti komunaliteta. *Image* ekstrakcija daje matematički jedinstveno rješenje jer se vrijednosti komunaliteta ne promijene. Kod metode maksimalne vjerodostojnosti, procjenjuje se broj faktora umjesto vrijednosti komunaliteta, a korelacije izvan dijagonale „namještaju se" kako bi najbolje odgovarale promatranim i reproduciranim matricama.

Ozbiljnost s kojom gledamo na procjene komunaliteta ovisi o broju promatranih varijabli. Ako broj varijabli, primjerice, prelazi 20, SMC-ovi iz uzorka vjerojatno će dati razumne procjene komunaliteta. Nadalje, uz 20 ili više varijabli, elemenata je na pozitivnoj dijagonali malo u usporedbi s ukupnim brojem elemenata u $\boldsymbol{R}$, a njihove veličine ne utječu značajno na rješenje. Dapače, ako su vrijednosti komunaliteta za sve varijable u PFA otprilike jednake veličine, rezultati PCA i PFA vrlo su slični.

Ako su vrijednosti komunaliteta veće ili jednake 1, to je indikacija problema u rješenju. Ili je premalo podataka, ili su početne vrijednosti komunaliteta krive, ili je broj ekstrahiran faktora pogrešan; dodavanje ili uklanjanje faktora može smanjiti komunalitet na vrijednost ispod 1. S druge strane, vrlo niske razine komunaliteta ukazuju na to da su njihove varijable nepovezane s drugim varijablama u setu.

6.2. Adekvatnost ekstrakcije i broja faktora

Iz razloga što veći broj faktora u rješenju čini da si promatrane i reproducirane korelacijske matrice bolje odgovaraju, adekvatnost ekstrakcije povezuje se s brojem faktora. Veći broj ekstrahiranih faktora dovodi do boljeg odgovaranja matrica i većeg postotka varijance u podacima koji se može objasniti faktorskim rješenjem. Potreban je kompromis između broja faktora i parsimonije rješenja.

Odabir broja faktora vjerojatno je važniji od odabira tehnike za ekstrakciju i rotaciju ili vrijednosti komunaliteta. U konfirmativnoj faktorskoj analizi odabir broja faktora zapravo je odabir broja teoretskih procesa iz određenog područja istraživanja. Djelomično je moguće potvrditi pretpostavljenu strukturu faktora pitajući se je li teoretski broj faktora adekvatan za set podataka.

Prva brza procjena broja faktora dobiva se iz veličina svojstvenih vrijednosti koje se dobiju nakon prvog pokretanja analize glavnih komponenata. Svojstvene vrijednosti predstavljaju varijancu. Iz razloga što varijanca koju svaka standardizirana varijabla donosi ekstrakciji glavnih komponenata iznosi 1, komponenta sa svojstvenom vrijednosti manjom od 1 iz perspektive varijance nije toliko važna kao promatrana varijabla. Broj komponenata sa svojstvenim vrijednostima većim od 1 obično je negdje između broja varijabli podijeljenog sa 3 i broja varijabli podijeljenog sa 5 (npr. 20 varijabli trebalo bi dati između 4 i 7 komponenata sa svojstvenim vrijednostima većim od 1). Ako je to razuman broj faktora za skup podataka, ako je broj varijabli 40 ili manje i ako je uzorak velik, broj faktora na koji ukazuje ovaj kriterij vjerojatno je dobar. U drugim situacijama, ovaj kriterij može ili podcijeniti ili precijeniti broj faktora u setu podataka.

Drugi je kriterij *scree* test svojstvenih vrijednosti po faktorima. Faktori se silaznim redoslijedom slože po apscisi, a svojstvena je vrijednost ordinata. Ovaj se grafički prikaz koristi s PCA i PFA na početku i tijekom analize za pronalaženje broja faktora.

Nažalost, *scree* test nije posve precizan; on uključuje procjenu o tome gdje se pojavljuje nepovezanost u svojstvenim vrijednostima, a istraživači nisu apsolutno pouzdani procjenitelji. Rezultati *scree* testa očitiji su (i pouzdaniji) kad je uzorak velik, kad su vrijednosti komunaliteta visoke, a svaki faktor ima nekoliko varijabli s visokim opterećenjima. Kad uvjeti nisu optimalni, *scree* test obično ima standardnu grešku od jednog do dva faktora.

Ako niste sigurni u broj faktora, obavite nekoliko faktorskih analiza, svaki put unoseći različit broj faktora, ponavljajući *scree* test i analizirajući rezidualnu matricu korelacija. Rezidualna matrica korelacija dobiva se oduzimanjem reproducirane korelacijske matrice od promatrane korelacijske matrice. Brojevi u rezidualnoj matrici zapravo su djelomične korelacije između parova varijabli bez efekata faktora. Ako je analiza dobra, rezidualne korelacije su niske. Nekoliko umjerenih rezidualnih korelacija (npr 0.05 do 0.10) ili nekoliko visokih rezidualnih korelacija (npr.$>$ 0.10) sugeriraju da postoji još jedan faktor.

Nakon određivanja broja faktora po ovim kriterijima važno je promotriti rotiranu matricu opterećenja radi utvrđivanja broja varijabli koje vrše opterećenje na svakom faktoru. Ako samo jedna varijabla vrši veliko opterećenje na neki od faktora, faktor nije dobro određen. Ako dvije varijable vrše opterećenje na faktor, tada njegova pouzdanost ovisi o uzorku korelacija između te dvije varijable međusobno te s drugim varijablama u $\boldsymbol{R}$. Ako su te dvije varijable u visokoj međusobnoj korelaciji (npr. $r > 0.70$) i relativno ne koreliraju s drugim varijablama, faktor bi mogao biti pouzdan. Tumačenje faktora određenih samo jednom ili dvama varijablama štetno je čak i u vrlo deskriptivnoj faktorskoj analizi.

Za analizu glavnih komponenata i analizu maksimalne vjerodostojnosti u konfirmativnoj faktorskoj analizi postoje testovi značajnosti broja faktora. Bartlettov test procjenjuje sve faktore zajedno i svaki faktor pojedinačno u odnosu na hipotezu da ne postoje faktori. Međutim, postoje rasprave o korištenju ovih testova. Zainteresiranog čitatelja upućuje se na Gorsucha (1983) ili neki od drugih, novijih tekstova koji diskutiraju o testiranju značajnosti u faktorskoj analizi.

Postoji debata o tome je li bolje imati previše ili premalo faktora ako je broj dvosmislen. Ponekad istraživač želi rotirati, ali ne i tumačiti, marginalne faktore u statističke svrhe (npr. zadržati neke faktore s vrijednostima komunaliteta manjim od 1). U drugim slučajevima posljednjih nekoliko faktora može predstavljati najzanimljivija, neočekivana otkrića u nekom području istraživanja. Postoje dobri razlozi za zadržavanje faktora koji su marginalno pouzdani. Međutim, ako istraživača zanima korištenje samo dokazivo pouzdanih faktora, zadržava se minimalni broj faktora.

6.3. Adekvatnost rotacije i jednostavne strukture

Odluka između ortogonalne i kose rotacije donosi se čim broj pouzdanih faktora postane jasan. Ponekad se kosa rotacija čini razumnijom jer se čini izglednijim da faktori koreliraju, nego da ne koreliraju. No jednostavnost izvještavanja rezultata ide u korist ortogonalnoj rotaciji. Nadalje, ako će se vrijednosti faktora koristiti kao neyavisne varijable ili yavisne varijable u drugim analizama, ili ako je cilj analize usporedba faktorske strukture u skupinama, tada ortogonalna rotacija ima izražene prednosti.

Možda je najbolji način za odluku između ortogonalne i kose rotacije tražiti kosu rotaciju sa željenim brojem faktora i pogledati korelacije između faktora. Međutim, ako podaci ne podupiru korelacije faktora, rješenje ispada gotovo ortogonalno..

Jednom kad se donese odluka između ortogonalne i kose rotacije, adekvatnost rotacije procjenjuje se na nekoliko načina. Možda je najjednostavniji način usporediti uzorak korelacija u korelacijskoj matrici s faktorima. Jesu li uzorci reprezentirani u rotiranom rješenju? Vrše li visokokorelirane varijable opterećenje na istom faktoru? Ako ste uključili marker varijable, vrše li one opterećenje na predviđene faktore?

Postoji još jedan kriterij – kriterij jednostavne strukture (Thurstone, 1947). Ako je prisutna jednostavna struktura (a faktori ne koreliraju previsoko), nekoliko varijabli visoko korelira sa svakim faktorom, a samo jedan faktor visoko korelira sa svakom varijablom. Drugim riječima, stupci $\boldsymbol{A}$, koji određuju faktore nasuprot varijabli, imaju nekoliko visokih i mnogo niskih vrijednosti, dok redovi $\boldsymbol{A}$, koji određuju varijable nasuprot faktora, imaju samo jednu visoku vrijednost. Redovi s više od jedne visoke korelacije odgovaraju varijablama za koje se kaže da su kompleksne jer odražavaju utjecaj više od jednog faktora. Obično je najbolje izbjegavati kompleksne varijable jer one čine tumačenje faktora dvosmislenim.

Adekvatnost rotacije također se utvrđuje PLOT instrukcijama statističkim programima. Na slikama se faktori razmatraju po dva odjednom uz različit par faktora kao osi svakog grafa.

Udaljenost točke varijable od ishodišta odražava veličinu faktorskog opterećenja; varijable koje visoko koreliraju s faktorom daleko su na osi tog faktora. U idealnom slučaju, svaka bi točka varijable bila daleko na jednoj osi i blizu ishodišta na svim drugima osima. Grupiranje (*clustering*, klasteriranje) točaka varijabli otkriva koliko je jasno definiran pojedini faktor. Poželjno bi bilo da su skupine od nekoliko točaka pri kraju svake osi, a da su sve ostale točke blizu ishodišta. Mali broj točaka na različitim udaljenostima po osi ukazuje na faktor koji nije jasno određen, a skupina točaka na pola puta između dvije osi odražava prisutnost još jednog faktora ili potrebu za kosom rotacijom. Smjer skupina/klastera nakon ortogonalne rotacije također može ukazivati na potrebu za kosom rotacijom. Ako skupine točaka padaju između faktorskih osi nakon ortogonalne rotacije, ako kut između skupina u odnosu na ishodište nije 90°, onda bi skupinama bolje odgovarale osi koje nisu ortogonalne. Kosa rotacija može otkriti značajne korelacije između faktora.

6.4. Važnost i unutarnja konzistentnost faktora

Važnost faktora (ili seta faktora) procjenjuje se omjerom varijance ili kovarijance koju daju faktori nakon rotacije. Omjer varijance koji se može pripisati pojedinačnim faktorima razlikuje se prije i nakon rotacije jer rotacija obično u nekoj mjeri preraspodjeljuje varijancu među faktorima. Lakoća utvrđivanja omjera varijance za faktore ovisi o tome je li rotacija bila ortogonalna ili kosa.

Nakon ortogonalne rotacije važnost faktora povezuje se s veličinom njegovih SSL-ova (zbroj kvadrata opterećenja, eng. *Sum of Squared Loadings* iz $\boldsymbol{A}$ nakon rotacije). SSL-ovi se pretvaraju u omjer varijance za faktor dijeljenjem sa p, brojem varijabli. SSL-ovi se pretvaraju u omjer kovarijance za faktor dijeljenjem njegova SSL-a sa zbrojem SSL-ova ili, ekvivalentno, zbrojem komunaliteta.

POGLAVLJE 7

7. Primjena na primjeru

Provedena je anketa među devedeset studenta Ekonomskog fakulteta sveučilišta u Zagrebu. Anketa se odnosila na ostvarene bodove na pismenim ispitima iz kolegija Matematika, Statistika, Mikroekonomija, Makroekonomija, Računovodstvo i Financije..

Tablica 2. Deskriptivna statistika broja bodova na pismenim ispitima

Variable	N	Arit Sred	Std Dev	Minimum	Maksimum
Matematika	90	26.8333333	18.0587369	0	60.0000000
Statistika	90	61.3333333	24.3762640	20.0000000	100.0000000
Mikroekonomija	90	53.4888889	28.6529974	2.0000000	100.0000000
Makroekonomija	90	53.2555556	29.0476405	1.0000000	100.0000000
Racunovodstvo	90	45.3111111	30.2898524	0	97.0000000
Financije	90	48.9444444	27.9901111	0	99.0000000

Na podacima je izvršena PCA i FA u programskom jeziku SAS-u. Cilj je vidjeti možemo li ovih 6 varijabli prikazati pomoću što manje faktora, odnosno komponenti.

Prvo što smo napravili u SAS-u jest PCA. Input u SAS-u:

```
proc import out= work.data
datafile= "/folders/myfolders/kruno1.csv"
dbms=csv replace; getnames=yes; datarow=2;
run;
%let xlist =Matematika Statistika Mikroekonomija Makroekonomija Racunovodstvo
Financije;
proc means data=data;
var &xlist;
run;
proc princomp plots =matrix;
var &xlist;
run;
proc princomp plots =pattern;
var &xlist;
run;
```

U outputu dobivamo (output dobiven u programskom jeziku SAS – cloud verzija):

Tablica 3. Korelacijska matrica

Correlation Matrix						
	Matematika	**Statistika**	**Mikroekonom**	**Makroekonom**	**Racunovodstvo**	**Financije**
Matematika	1.0000	0.1109	0.1439	-.0340	0.0403	0.0150
Statistika	0.1109	1.0000	0.0149	0.1480	0.1508	-.0656
Mikroekonom	0.1439	0.0149	1.0000	-.1697	0.0761	-.0803
Makroekonom	-.0340	0.1480	-.1697	1.0000	-.0616	0.1038
Racunovodstvo	0.0403	0.1508	0.0761	-.0616	1.0000	0.1702
Financije	0.0150	-.0656	-.0803	0.1038	0.1702	1.0000

Kao što možemo vidjeti u korelacijskoj matrici korelacije među varijablama su jako male, što znači da nam podaci nisu baš korelirani. Znači imamo set varijabli koje nisu baš korelirane, što nam i nije najbolje za analizu glavnih komponenti i faktorsku analizu.

Tablica 4. Svojstvene vrijednosti korelacijske matrice

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	1.29277476	0.04276780	0.2155	0.2155
2	1.25000696	0.14933923	0.2083	0.4238
3	1.10066773	0.16960528	0.1834	0.6072
4	0.93106245	0.14289233	0.1552	0.7624
5	0.78817012	0.15085214	0.1314	0.8938
6	0.63731798		0.1062	1.0000

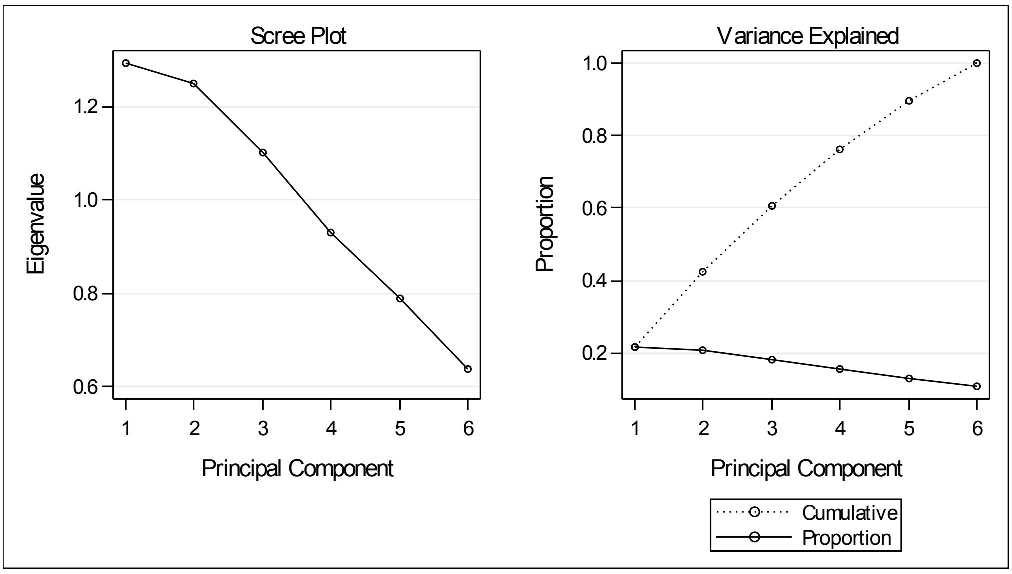
Prvi zanimljivi rezultat su svojstvene vrijednosti korelacijske matrice. Kao što je već ranije spomenuto u radu, jednim od kriterija (K1) zadržati ćemo sve komponente čije su svojstvene vrijednosti veće od 1 (zadržati ćemo samo tri komponente, jer samo tri komponente imaju svojstvenu vrijednost veću od 1) . Kao što možemo vidjeti prvom komponentom je objašnjeno 21.55% varijacije, drugom komponentom 20.83%, a trećom komponentom 18.34%, odnosno prve tri komponente objašnjavu 60.27% varijacije.

Tablica 5. Svojstveni vektori

Eigenvectors						
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**
Matematika	0.471342	0.206635	0.200585	0.714857	-.383194	0.192474
Statistika	0.184703	0.534678	0.536746	-.324021	-.074433	-.530451
Mikroekonomija	0.634874	-.137586	-.015154	0.048878	0.755415	-.068810
Makroekonomija	-.455856	0.441761	0.344276	0.226461	0.494738	0.427161
Racunovodstvo	0.325621	0.508530	-.398913	-.448430	-.113378	0.512145
Financije	-.163776	0.445798	-.627521	0.359481	0.139107	-.481749

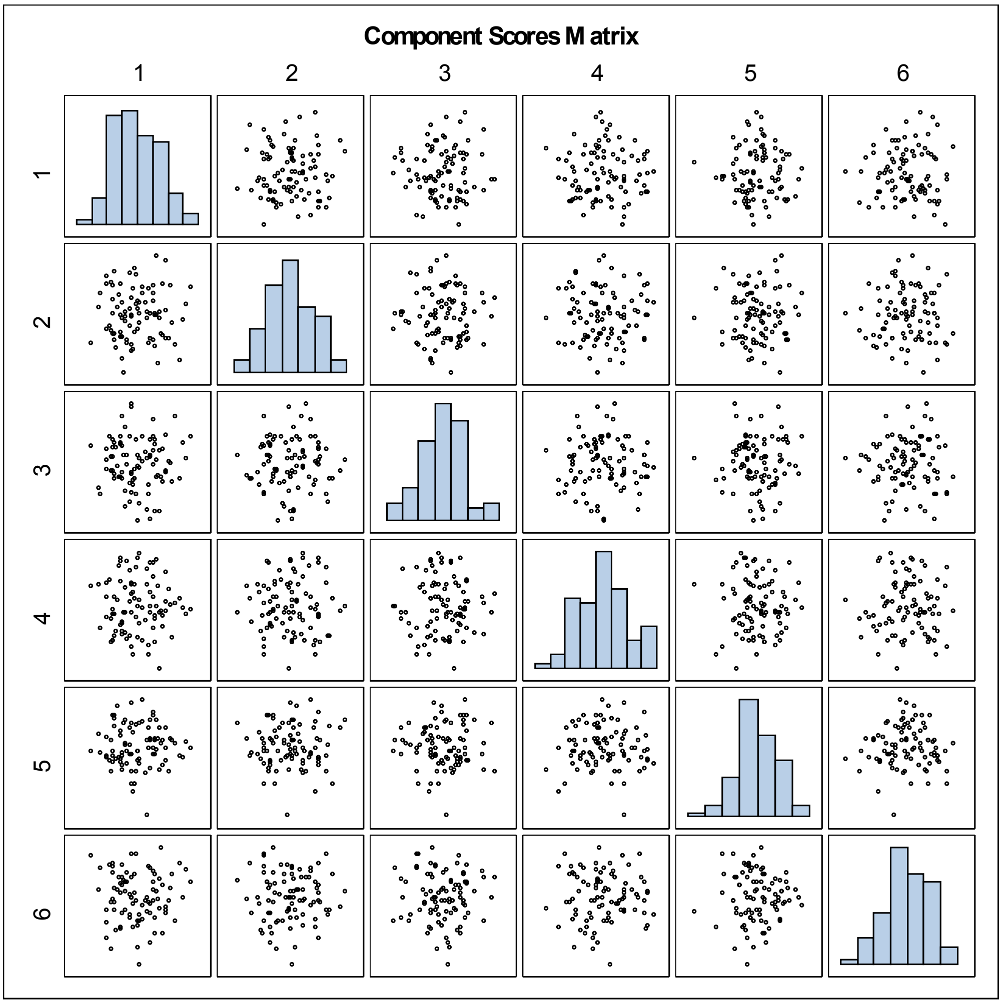
Promatrajući tablicu ne možemo baš točno zaključiti koje varijable čine koju komponentu, razlog tome je što nam podaci nisu dobro korelirani. U prvu komponentu mogli bismo staviti Matematiku i Mikroekonomiju, u drugu komponentu Financije i Računovodstvo, u treću komponentu Makroekonomiju i Statistiku.

Slika 2. Grafički prikaz svojstvenih vrijednosti i proporcije objašnjene varijance



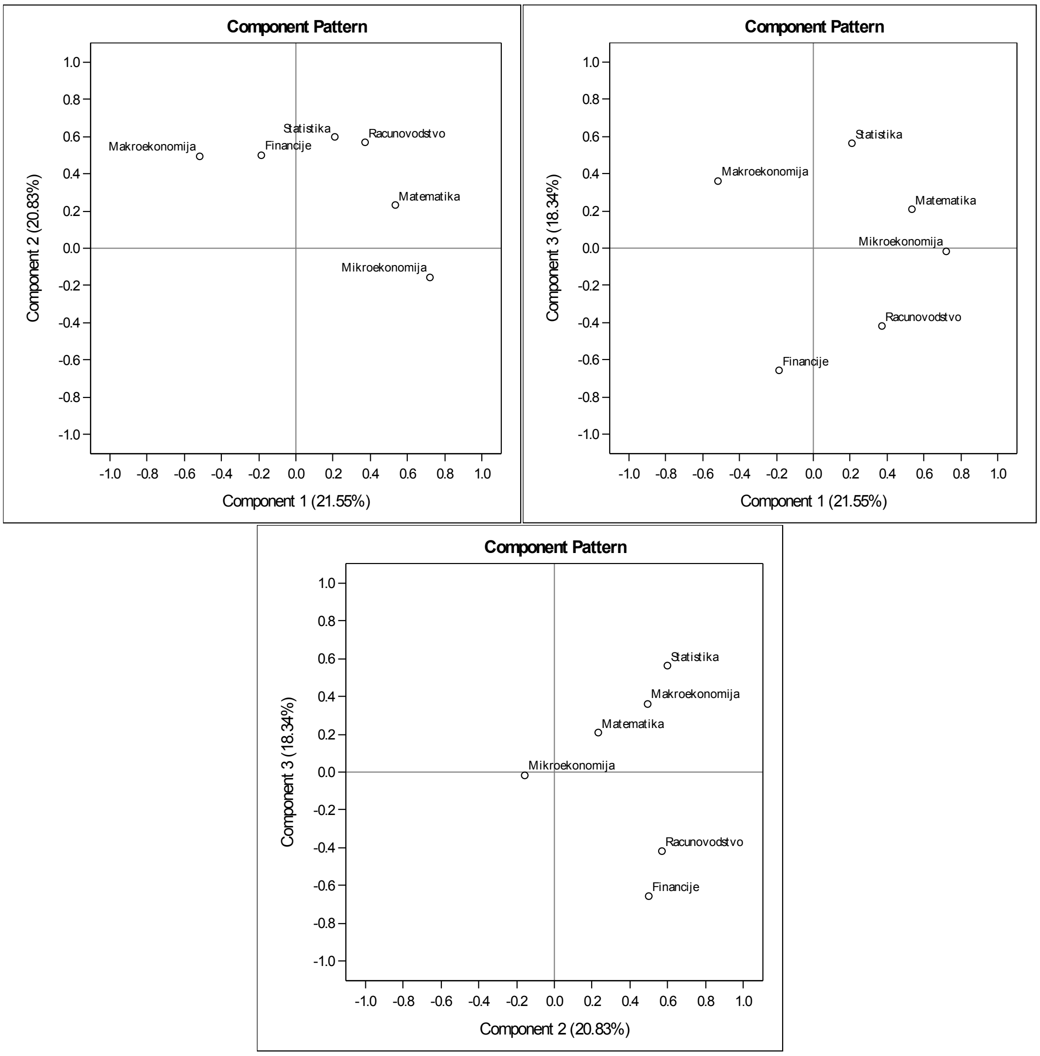
Također i iz slike vidimo da samo tri komponente imaju svojstvenu vrijednost veću od 1, pa i iz grafičkog prikaza zaključujemo da trebamo zadržati samo tri komponente.

Slika 3. Matrica vrijednosti komponenti



Slika 3. prikazuje vrijednosti između svih šest komponenti. Na dijagonali su prikazani histogrami svake komponente. Kao što možemo vidjeti iz histograma komponente su uglavnom simetrično distribuirane (govorimo samo o prve tri komponente, jer samo njih razmatramo u analizi). Također iz slike vidimo da komponente nisu baš dobro međusobno korelirane.

Slika 4. Grafički prikaz komponenti u parovima



Slika 3. nam opet ne daje jasnu sliku koje varijable čine koje komponente. Ponovno napominjemo da je razlog tome nekoreliranost podataka.

Sada u SAS-u vršimo faktorsku analizu:

```
proc import out= work.data

datafile= "/folders/myfolders/kruno1.csv"

dbms=csv replace; getnames=yes; datarow=2;

run;

%let xlist =Matematika Statistika Mikroekonomija Makroekonomija Racunovodstvo Financije;

proc means data=data;

var &xlist;

run;

proc factor data=data corr;     /* opcija corr služi za dobivanje korelacijske matrice*/

run;

proc factor data=data

priors=smc msa          /* opcija smc vrši korelaciu između varijable i svih ostalih varijabli, s opcijom msa zahtjevamo Kaiserovo mjerilo adekvatnosti uzorkovanja */

rotate=promax reorder          /* opcija promax vrši ortogonalnu varimax predrotaciju, nakon koje slijedi kosa Procrustes rotacija, opcija reorder poreda varijable prema najvećem faktoru opterećenja */

plots=(scree initloadings preloadings loadings);     /* scree opcija služi za graf svojstvenih vrijednosti, initloadings za graf nerotiranih varijabli, preloading za graf predrotiranih varijabli i loadings za graf rotiranih varijabli */

run;
```

U outputu dobivamo (output dobiven u programskom jeziku SAS - cloud verzija):

Tablica 6. Matrica korelacija

Correlations						
	Matematika	**Statistika**	**Mikroekonomija**	**Makroekonomija**	**Racunovodstvo**	**Financije**
Matematika	1.00000	0.11090	0.14391	-0.03395	0.04029	0.01496
Statistika	0.11090	1.00000	0.01489	0.14804	0.15076	-0.06563
Mikroekonomija	0.14391	0.01489	1.00000	-0.16972	0.07608	-0.08031
Makroekonomija	-0.03395	0.14804	-0.16972	1.00000	-0.06158	0.10379
Racunovodstvo	0.04029	0.15076	0.07608	-0.06158	1.00000	0.17020
Financije	0.01496	-0.06563	-0.08031	0.10379	0.17020	1.00000

Kao što možemo vidjeti u korelacijskoj matrici korelacije među varijablama su jako male, što znači da nam podaci nisu baš korelirani. Znači imamo set varijabli koje su loše korelirane, što nam i nije najbolje za analizu glavnih komponenti i faktorsku analizu.

Tablica 7. Svojstvene vrijednosti korelacijke matrice

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	1.29277476	0.04276780	0.2155	0.2155
2	1.25000696	0.14933923	0.2083	0.4238
3	1.10066773	0.16960528	0.1834	0.6072
4	0.93106245	0.14289233	0.1552	0.7624
5	0.78817012	0.15085214	0.1314	0.8938
6	0.63731798		0.1062	1.0000

Promatrajući svojstvene vrijednosti, mogli bismo zaključiti da ćemo imati tri faktora (jer su tri svojstvene vrijednosti veće od 1), ali kako smo u kodu koristili opciju MSA, vidjet ćemo kasnije koliko faktora nam je ona ostavila.

Tablica 9. Parcijalne korelacije

Partial Correlations Controlling all other Variables						
	Matematika	**Statistika**	**Mikroekonomija**	**Makroekonomija**	**Racunovodstvo**	**Financije**
Matematika	1.00000	0.11284	0.13788	-0.03022	0.00418	0.03570
Statistika	0.11284	1.00000	0.00618	0.17576	0.17629	-0.11572
Mikroekonomija	0.13788	0.00618	1.00000	-0.15224	0.07302	-0.07841
Makroekonomija	-0.03022	0.17576	-0.15224	1.00000	-0.09578	0.12059
Racunovodstvo	0.00418	0.17629	0.07302	-0.09578	1.00000	0.19861
Financije	0.03570	-0.11572	-0.07841	0.12059	0.19861	1.00000

Parcijalne korelacije bi trebale biti manje u odnosu na korelacije u tablici 6. Ako usporedimo Tablicu 6. i Tablicu 9. vidimo da je apsolutna vrijednost parcijalne korelacije između statistike i financija veća, u odnosu na korelaciju u tablici 6. Razlog tome je ponovno taj što nam podaci nisu dobro korelirani.

Tablica 10. Kaiserova mjera uzorkovanja adekvatnosti

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.45037808					
Matematika	**Statistika**	**Mikroekonomija**	**Makroekonomija**	**Racunovodstvo**	**Financije**
0.51473994	0.41089749	0.53574515	0.45776662	0.42514715	0.40408315

Vrijednosti između 0.8 i 1.0 smatraju se dobrima, dok vrijednosti ispod 0.5 su neprihvatljive. Cjelokupni MSA iznosi 0.45, što je jako loše. Razlog tomu je opet to što nam podaci nisu dobro korelirani. Više varijabli je potrebno za dobru analizu.

Tablica 11. Procjene komunaliteta

Prior Communality Estimates: SMC					
Matematika	**Statistika**	**Mikroekonomija**	**Makroekonomija**	**Racunovodstvo**	**Financije**
0.03460939	0.07247256	0.05762738	0.07105014	0.07177059	0.06103534

U ovoj tablici su prikazani kvadrati višestruke korelacije prije komunalitetne procjene. Budući da su kvadrati višestruke korelacije obično manji od 1, odgovarajuća korelacijska matrica naziva se reduciranom korelacijskom matricom. Budući da su nam SMC-ovi jako mali, rezultati će nam se razlikovati od rezultata dobivenih pomoću PCA.

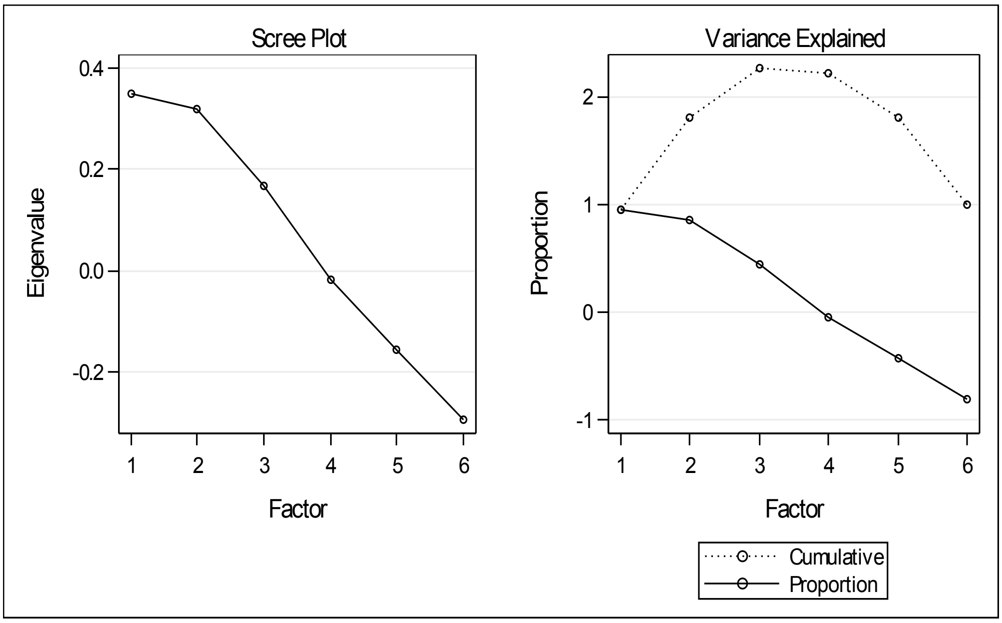
Tablica 12. Svojstvene vrijednosti reducirane korelacijske matrice

Eigenvalues of the Reduced Correlation Matrix: Total = 0.36856539 Average = 0.06142757				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	0.35074095	0.03273191	0.9516	0.9516
2	0.31800903	0.15114233	0.8628	1.8145
3	0.16686670	0.18438033	0.4527	2.2672
4	-.01751363	0.13700470	-0.0475	2.2197
5	-.15451833	0.14050102	-0.4192	1.8005
6	-.29501934		-0.8005	1.0000

Sve svojstvene vrijednosti su manje od 0.5, pa na temelju toga ne možemo zaključiti koliko faktora ćemo imati. Pogledajmo sada kumulativne proporcije. Prve dvije svojstvene vrijednosti reducirane korelacijske matrice objašnjavaju 181.45% zajedničke varijance. Ovo je moguće jer reducirana korelacijska matrica nije pozitivno definitna. Ovakvi rezultati nam ukazuju na to da nam neće biti potrebno više od dva faktora (jer prva dva faktora objašnjavaju više od

100% zajedničke varijance, pa na temelju toga MSA odabire samo dva faktor). Napominjemo da rezultati nisu najbolji, jer nam podaci nisu dobro korelirani.

Slika 5. Grafički prikaz svojstvenih vrijednosti i proporcije objašnjene varijance



Na temelju slike 4. ne možemo baš zaključiti koliko faktora trebamo ostaviti u analizi.

Tablica 13. Faktori

Factor Pattern		
	Factor1	**Factor2**
Mikroekonomija	0.37768	-0.06245
Matematika	0.25739	0.11619
Makroekonomija	-0.28946	0.24082
Statistika	0.09837	0.31162
Racunovodstvo	0.19421	0.29632
Financije	-0.10330	0.24022

Pomoću MSA smo dobili dva faktora. Iz tablice bismo mogli zaključiti da prvi faktor čine varijable Mikroekonomija i Matematika, a drugi faktor varijable Statistika, Računosvodstvo, Makroekonomija i Financije.

Tablica 14. Varijananca objašnjena pojedinim faktorom

Variance Explained by Each Factor	
Factor1	**Factor2**
0.35074095	0.31800903

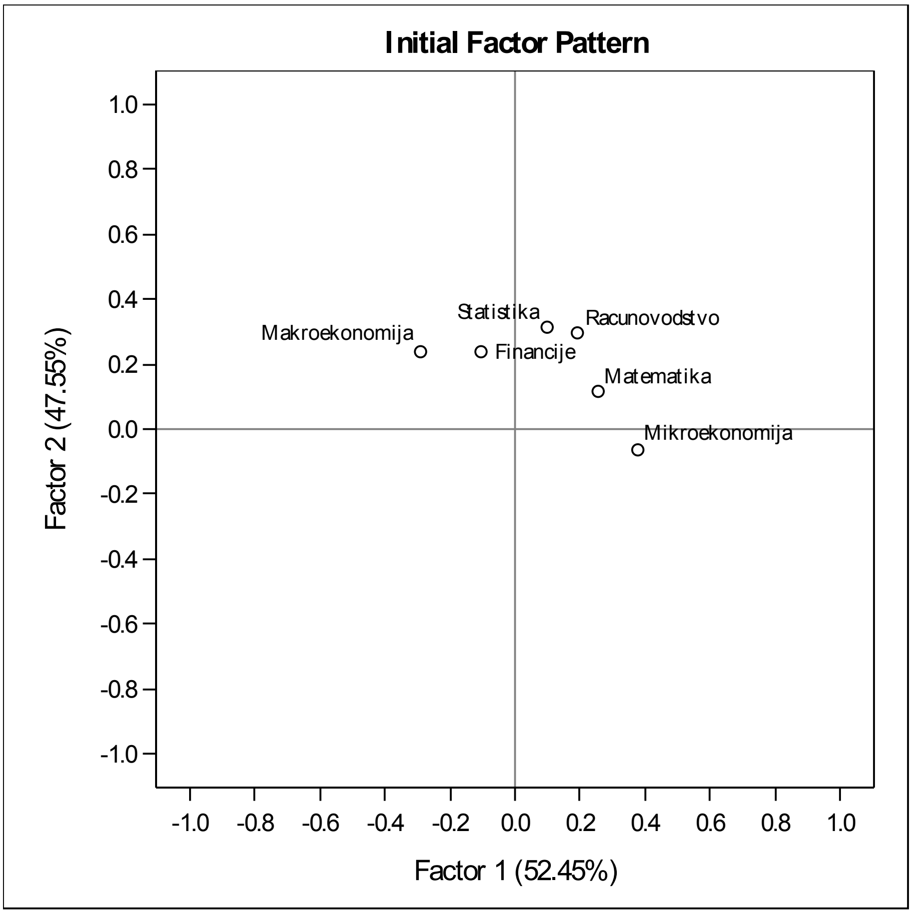
Kao što možemo vidjeti prvim faktorom je objašnjeno 35% varijance, dok je drugim faktorom objašnjeno 32% varijance. Kasnije, pomoću rotacija, ćemo pokušati što više izjednačiti ove varijance, iako su već sada poprilično jednake, razlog tome je već spomenut, podaci nam nisu dobro korelirani.

Tablica 15. Konačna pocjena komunaliteta

Final Community Estimates: Total = 0.668750					
Matematika	**Statistika**	**Mikroekonomija**	**Makroekonomija**	**Racunovodstvo**	**Financije**
0.07975166	0.1067814	0.14654362	0.14177743	0.12551971	0.06837615

Opet vidimo da su nam komunaliteti mali, što nam i nije dobro za faktorsku analizu.

Slika 6. Grafički prikaz faktora



Ako pogledamo sliku vidimo da su opterećenja varijabli na oba faktora vrlo blizu nule. Dobra rotacija će ih rasporediti tako da opterećenja budu još bliža nuli.

Tablica 16. Ortogonalna transformacijska matrica

Orthogonal Transformation Matrix		
	1	**2**
1	-0.82987	0.55796
2	0.55796	0.82987

Tablica 17. Rotirani faktori

Rotated Factor Pattern		
	Factor1	**Factor2**
Makroekonomija	0.37458	0.03834
Financije	0.21976	0.14172
Mikroekonomija	-0.34827	0.15890
Racunovodstvo	0.00417	0.35426
Statistika	0.09223	0.31349
Matematika	-0.14877	0.24004

Ako usporedimo tablicu 17. i tablicu 13. vidimo da je došlo do promjene. Prvi faktor sada čine varijable Makroekonomija i Financije, dok ostale varijable čine drugi faktor.

Tablica 18. Varijananca objašnjena pojedinim faktorom nakon rotacije

Variance Explained by Each Factor	
Factor1	**Factor2**
0.34055096	0.32819903

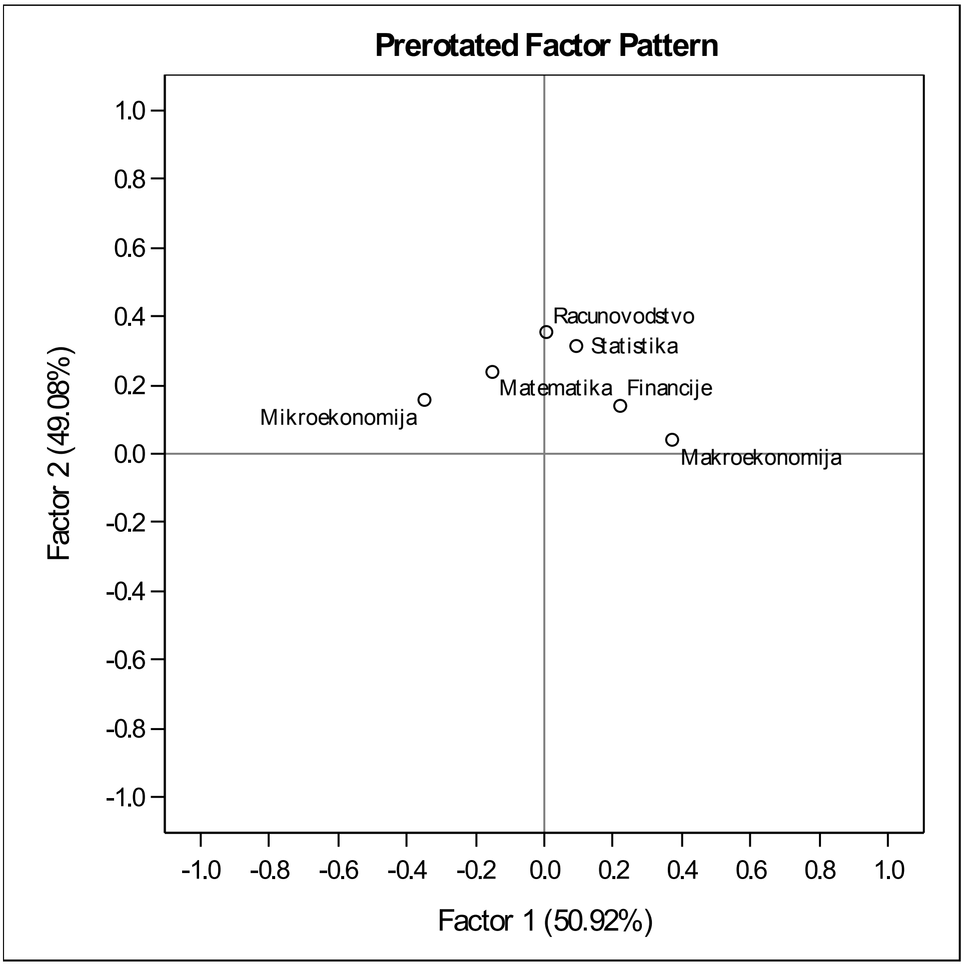
Kao što možemo vidjeti prvim faktorom je objašnjeno 34% varijance nakon rotacije, dok je drugim faktorom objašnjeno 33% varijance nakon rotacije. Još smo više približili vrijednosti, iako su i prije rotacije bile jako blizu.

Tablica 19. Konačna procjena komunaliteta nakon rotacije

Final Communality Estimates: Total = 0.668750					
Matematika	**Statistika**	**Mikroekonomija**	**Makroekonomija**	**Racunovodstvo**	**Financije**
0.07975166	0.10678141	0.14654362	0.14177743	0.12551971	0.06837615

Ako usporedimo tablicu 19. s tablicom 15. uočavamo da je došlo do neznatnih promjena. Naglašavamo da je razlog tome loša koreliranost varijabli.

Slika 7. Grafički prikaz faktora pred rotaciju



Nakon rotacije varijable su raspoređene tako da su ipak malo bliže nuli, nego prije rotacije. Ali ta raspodjela je neznatna, jer su i prije rotacije varijable bile jako blizu nule, tj. opterećenje je bilo malo.

Tablica 20. Ciljna matrica za Procrustean transformaciju (kosu rotaciju)

Target Matrix for Procrustean Transformation		
	Factor1	**Factor2**
Makroekonomija	1.00000	0.00106
Financije	0.60291	0.15922
Mikroekonomija	-0.76489	0.07153
Racunovodstvo	0.00000	1.00000
Statistika	0.02284	0.88309
Matematika	-0.14850	0.61423

Tablica 21. Prorusteanova transformacijska matrica

Procrustean Transformation Matrix		
	1	**2**
1	2.34351745	0.1242668
2	0.02826409	2.48140937

Tablica 22. Normalizirana kosa transformacijska matrica

Normalized Oblique Transformation Matrix		
	1	**2**
1	-0.82467	0.51675
2	0.56902	0.85839

Tablica 23. Međufaktorska korelacija

Inter-Factor Correlations		
	Factor1	**Factor2**
Factor1	1.00000	-0.06206
Factor2	-0.06206	1.00000

Kao što možemo vidjeti, korelacija među faktorima je jako mala. To se moglo i očekivati, jer su nam početne varijable bile slabo korelirane.

Tablica 24. Rotirani faktori (standardizirani regresijski koeficijenti)

Rotated Factor Pattern (Standardized Regression Coefficients)		
	Factor1	**Factor2**
Makroekonomija	0.37574	0.05714
Financije	0.22188	0.15283
Mikroekonomija	-0.34700	0.14156
Racunovodstvo	0.00846	0.35471
Statistika	0.09619	0.31832
Matematika	-0.14615	0.23275

Možemo vidjeti da nema prevelike razlike između tablice 24. i tablice 17.

Tablica 25. Referentne osi korelacije

Reference Axis Correlations		
	Factor1	**Factor2**
Factor1	1.00000	0.06206
Factor2	0.06206	1.00000

Tablica 26. Semiparcijalne korelacije

Reference Structure (Semipartial Correlations)		
	Factor1	**Factor2**
Makroekonomija	0.37501	0.05703
Financije	0.22145	0.15253
Mikroekonomija	-0.34633	0.14128
Racunovodstvo	0.00844	0.35403
Statistika	0.09601	0.31771
Matematika	-0.14586	0.23230

Opet zaključujemo da nema prevelike razlike između tablice 26. i tablice 17. Razlog tomu je loša koreliranost varijabli.

Tablica 27. Varijanca objašnjena pojedinim faktorom bez utjecaja drugih faktora

Variance Explained by Each Factor Eliminating Other Factors	
Factor1	**Factor2**
0.34018363	0.32671573

Prvim faktorom je objašnjeno 34% varijance, dok je drugim faktorom objašnjeno 33% varijance.

Tablica 28. Faktori (korelacije)

Factor Structure (Correlations)		
	Factor1	**Factor2**
Makroekonomija	0.37219	0.03382
Financije	0.21239	0.13906
Mikroekonomija	-0.35578	0.16309
Racunovodstvo	-0.01356	0.35419
Statistika	0.07644	0.31235
Matematika	-0.16059	0.24182

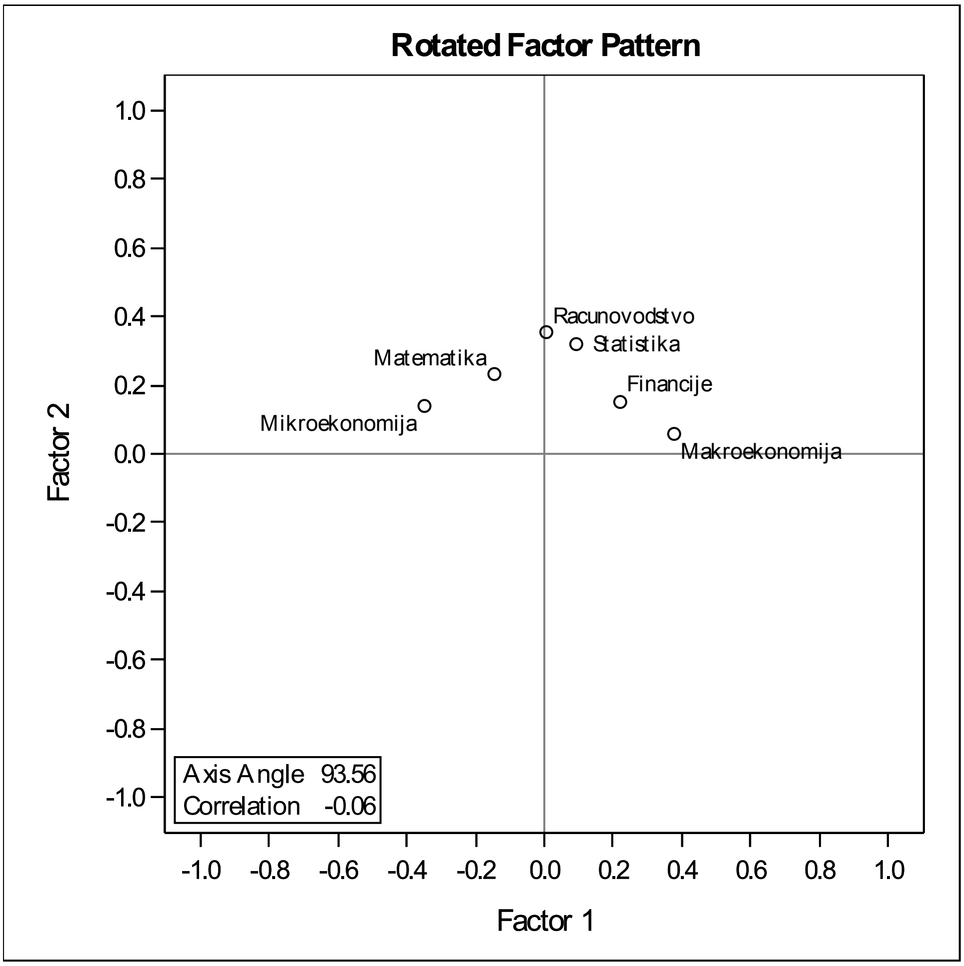
I nakon rotacije vrijednosti se nisu pretjerano promjenile. Prvi faktor čine varijable Makroekonomija, Financije i Mikroekonomija, dok ostale varijable čine drugi faktor.

Tablica 29. Konačna procjena komunalitetna nakon rotacije

Final Communality Estimates: Total = 0.668750					
Matematika	**Statistika**	**Mikroekonomija**	**Makroekonomija**	**Racunovodstvo**	**Financije**
0.07975166	0.10678141	0.14654362	0.14177743	0.12551971	0.06837615

Uočavamo da nema nikakvih promjena.

Slika 8. Grafički prikaz rotiranih faktora



Ako pogledamo sliku, kordinatne osi se sijeku pod kutom od 93.56°, znači, sijeku se skoro pod pravim kutom. Nismo ništa pretjerano dobili rotacijom. Razlog ovome je loša koreliranost varijabli. Primjer koji smo obradili nije dobar i pogodan za faktorsku analizu, a ni za analizu glavnih komponenti. Da bismo dobili bolje rezultate trebali bismo uvesti još varijabli, nadajući se da će među njima biti bolja koreliranost.

Bibliografija

[1] B. G. Tabachnick, L. S. Fidell, *Using multivariate statistics*, Needham Heights, MA, 1996. by Allyn & Beacon

[2] A. Katchova, *Principal component analysis and Factor analysis in SAS*, 2013. (https://www.youtube.com/watch?v=FXU1ktaVcdI)

Sažetak

Ove se tehnike obično koriste za analizu skupine koreliranih varijabli. Analiza glavnih komponenti i faktorska analiza koristite se za svođenje brojnih varijabli na nekoliko faktora. Razlika između ova dva pristupa ima veze s varijancom koja se analizira. Kod PCA, analiziraju se sve promatane varijance, dok se faktorskom analizom analizira samo zajednička varijanca.

Summary

These techniques are typically used to analyze groups of correlated variables. Principal components analysis and factor are used used for reducing the number of variables on several factors. The difference between these two approaches has to do with the variance that is analyzed. In PCA, all of the observed variance is analyzed, while in factor analysis it is only the shared variances that is analyzed.

Životopis

Zovem se Krunoslav Rajčić. Rođen sam 10. srpnja 1987. godine u Vinkovcima.

Svoje obrazovanje započeo sam 1994. godine u osnovnoj školi Ivana Brlić Mažuranić u Andrijaševcima. Nakon završene osnovne škole 2002. godine upisujem se u gimnaziju Matija Antun Reljković u Vinkovcima.

Preddiplomski sveučilišni studij matematike; smjer nastavnički na Prirodoslovno matematičkom fakultetu u Zagrebu upisujem 2008. godine, kojeg završavam 2011. godine. Iste te godine upisujem Diplomski sveučilišni studij Matematička statistika, kojeg upravo završavam.