



УНИВЕРЗИТЕТ У НОВОМ САДУ
ПРИРОДНО-МАТЕМАТИЧКИ ФАКУЛТЕТ
ДЕПАРТМАН ЗА МАТЕМАТИКУ И
ИНФОРМАТИКУ



Сања Станковић

ПРИМЕНА ФАКТОРСKE АНАЛИЗЕ У КИНЕЗИОЛОГИЈИ

МАСТЕР РАД

Нови Сад, 2013.

САДРЖАЈ:

ПРЕДГОВОР.....	4
1. МОТИВАЦИЈА ЗА УПОТРЕБУ ФАКТОРСКЕ АНАЛИЗЕ КРОЗ ПРИМЕРЕ	5
1.1. Дефиниција факторске анализе	5
1.2. Преглед проблема које решавамо факторском анализом	6
1.2.1. Седиментарна петрологија	6
1.2.2. Минералологија	6
1.2.3. Морска обала (приобаље)	6
1.2.4. Тешки минерали	6
2. ОСНОВНИ МАТЕМАТИЧКИ И СТАТИСТИЧКИ ПОЈМОВИ	7
2.1. Неке дефиниције	7
2.1.1. Матрица података	7
2.1.2. Матрични производ веће и мање димензије	7
2.2. Матрична нотација дескриптивних статистика	8
2.2.1. Аритметичка средина	8
2.2.2. Девијација	8
2.2.3. Варијанса	8
2.2.4. Коваријанса	9
2.2.5. Стандардна девијација	9
2.2.6. Стандардизовани резултати	9
2.2.7. Корелација	10
2.2.8. Угао између два вектора	12
2.2.9. Регресија, вишеструка и парцијална корелација.....	15
2.3. Ротација координатног система	16
2.3.1. Алгебра ротације	17
2.3.2. Интерпретација ротације	18
2.4. Ранг матричног производа	19
2.5. Карактеристични корени и вектори	20
2.5.1. Алгебарски преглед	20
2.5.2. Геометријска интерпретација карактеристичних вредности и вектора	24
2.6. Апроксимација матрице методом најмањих квадрата	26
3. ЦИЉЕВИ, ИДЕЈЕ И МОДЕЛИ ФАКТОРСКЕ АНАЛИЗЕ	29
3.1. Развој модела	29
3.1.1. Модел за регистроване променљиве	29
3.1.2. Модел за варијансе и коваријансе	31

3.2. Пример факторске анализе	33
3.3. Приказ резултата	37
4. R-MODE METODE	39
4.1. Анализа главних компоненти	39
4.1.1. Број компоненти	45
4.2. Факторска анализа	46
4.2.1. Нескалирана оцена матрица A и Ψ	47
4.2.2. Пример	48
4.3. Анализа главних компоненти и крос-валидација (унакрсна провера ваљаности)	50
5. ПРИМЕНА МЕТОДА ФАКТОРСKE АНАЛИЗЕ У КИНЕЗИОЛОГИЈИ	52
5.1. Поступак факторске анализе у SPSS-у	53
ЗАКЉУЧАК	60
ЛИТЕРАТУРА	61

ПРЕДГОВОР

Предмети истраживања, како у природним, тако и у друштвеним наукама, најчешће се не могу потпуно објаснити једном или двама променљивим, него читавим скупом променљивих. Зато се јавља потреба за применом метода вишедимензионалне анализе. Једна од метода вишедимензионалне анализе је факторска анализа, која се користи за опис међусобне зависности великог броја променљивих, помоћу одређивања скупа нових заједничких променљивих које се називају фактори. Тако је један од основних задатака факторске анализе сажимање већег броја међусобно повезаних, колинеарних, полазних променљивих у мањи број заједничких фактора који ће их описивати и објаснити њихову међусобну повезаност. На тај начин се анализа, уместо над великим бројем корелираних изворних променљивих, спроводи на некорелираним факторима и на тај се начин отклања проблем колинеарности.

У првом делу рада ће се кроз примере дати мотивација за употребу факторске анализе, и биће дати неки појмови матричног рачуна, који су специфични за факторску анализу, као што је ротација координатног система, геометријска интерпретација карактеристичних вредности и корена, апроксимација матрице методом најмањих квадрата. Такође ће бити уведена матрична нотација дескриптивних статистика.

У другом делу рада биће дате математичке основе факторске анализе за узорак. Прво ће бити приказани R-mode методи факторске анализе којима се проучава међусобна повезаност променљивих, као што је анализа главних компоненти. Овај метод ће бити илустрован примерима из природних наука.

У трећем делу ће се приказати примена факторске анализе у кинезиологији (спорту), за утврђивање логичке ваљаности тестова за процену координације код деце средњег школског узраста. Узорак испитаника је састављен од укупно 179 испитаника, девојчица 11-12 година старости. Манифестне варијабле које служе за процену опште координације код деце: полигон уназад или натрашке, слалом са три лопте, скок у даљ из места, скок у даљ итд. се свде на мањи број фактора, Биће спроведена комплетна факторска анализа, рачунањем дескриптивних статистика, матрице интеркорелације манифестних варијабли, КМО тест и Бартлетов тест сферицитетa. Биће израчунате вредности комуналитета, тј. величине које описују векторе манифестних варијабли у простору фактора, укупна количина ваљане варијансе матрице интеркорелације манифестних варијабли објашњене екстрахованим фактором, величине коефицијената корелација између манифестних варијабли и добијеног фактора, итд. Статистичка обрада података ће се радити помоћу статистичког софтвера СПСС.

1. МОТИВАЦИЈА ЗА УПОТРЕБУ ФАКТОРСКЕ АНАЛИЗЕ КРОЗ ПРИМЕРЕ

Факторска анализа је основана од стране научника који се баве изучавањем понашања. Њене саме почетке забележио је Харман приписујући њено порекло чланку о ортогоналним регресионим линијама које је Пирсон објавио 1901. године. Први прави развој постиже Хотелинг 1933. године. На самом почетку, факторску анализу није пратила добра репутација, јер различите варијанте истог метода нису давале исте резултате. То се дешавало због различитих поједностављујућих претпоставки и помоћних услова коришћених у прекомпјутерској ери. Током година, компјутери су у многоме утицали на развој факторске анализе, те је ово поље постало толико велико да је немогуће испитати све методологије. У овом раду, концентрисаћемо се на оне технике које су постале популарне.

1.1. Дефиниција факторске анализе

Ако тражимо дефиницију факторске анализе, од сваког научника добићемо другачији одговор. Дефиниција се мењала током година обухватајући све ширу селекцију техника. Факторска анализа служи за свођење великог скупа променљивих на мањи број фактора, с којима је лакше радити.

Факторска анализа је генерички појам који користимо да бисмо описали неколико метода осмишљених за анализу међусобних односа у оквиру скупа променљивих или објеката. Иако се различите технике доста разликују у својим циљевима и математичким моделима у основи, сви они имају једну заједничку одлику, конструисање неколико хипотетичких променљивих (или објеката), названих фактори, који треба да садрже основне информације о већем скупу посматраних променљивих или објеката. Фактори су конструисани тако да редукују укупну комплексност података користећи зависност међу променљивама. Као резултат, мали број фактора ће дати приближно исту количину информација као и много већи скуп оригиналних посматрања. Према томе, факторска анализа је у овом смислу, мултиваријантни метод за редукцију података.

Психолози су развили и направили широку употребу факторске анализе у својим студијама о људској менталној способности. Метод је првобитно био осмишљен за анализирање одређених резултата великог броја појединаца у мноштву психолошких тестова. Тестови способности и достигнућа су били осмишљени да мере различите аспекте менталне способности, али ускоро је постало очигледно да они показују велики степен међусобне корелације. Факторска анализа покушава да „објасни“ ове корелације анализом, која, када се успешно обави, даје мали број основних фактора, који садрже све суштинске информације о корелацијама међу тестовима. Интерпретација фактора је довела до теорије основних аспеката људске способности.

Једноставно речено, факторска анализа ствара минималан број нових променљивих, које су линеарне комбинације оних оригиналних, тако да нове променљиве садрже већину или све информације.

1.2. Преглед проблема које решавамо факторском анализом

У овом одељку даћемо кратак преглед случајно изабраних научних чланака у којима се користи факторска анализа како би се решио научни проблем. Ово радимо у нади да ћемо имати бољи увид у врсте проблема које решавамо техникама факторских анализа.

1.2.1. Седиментарна петрологија

Озборн је употребио факторску анализу за груписање кречњака из доба Ордовик, на основу одређених карактеристика пронађених у танким деловима. Успео је да придода палеоеколошки значај издвојеним факторима. Користећи истепретпоставке, Мекемон (1968) је урадио упоредну студију метода факторске анализе о груписању фацијакарбоната из садашњег доба са Бахама. Сличну студију је извео Фенинцер за кречњаке из доба Јуре са северних Алпа.

1.2.2. Минералологија

Мидлтон (1964) је користио факторску анализу да би објаснио компликован минералогски проблем у скаполитима. Применом основних компоненти и факторске анализе на главни елемент и елементе који се налази у траговима могао је идентификовати постојање *мариалите-мелиомите* чврстог раствора у скаполитима и понудити могуће постојање независног крајњег члана (у смислу чистоће) који носи *Mg* и *ОН*. Три значајне групације елемената у траговима су биле отписане. Минералогске анализе често захтевају специјалан третман.

1.2.3. Морска обала (приобаље)

Питање палеоеколошког значаја се односи на препознавање заједница плиматских седимената. Кејси и Мајкл (1968) су представили неколико мултиваријантних метода у добро документованом истраживању овог проблема и дошли су до закључка да се главна компонента анализе показала као најсвестранија те дозвољава и препознавање структуре заједнице и веродостојно обликовање заједница у простору.

1.2.4. Тешки минерали

Имбри и Ван Андел (1964) су проучавали појаву тешких метала код Калифорнијског залива и код спруда Ориноко у Гвајани, ***R-mode*** и ***Q-mode*** факторском анализом. Ове две области имају различиту седиментолошку прошлост. Факторска анализа једноставне ситуације коју представља материјал из Калифорније дала је резултате сличне онима добијеним конвенционалним прегледом, мада су откривени значајнији детаљи. Удаљенији петрографски извори спруда Ориноко у Гвајани производе компликованију ситуацију са много мешања минерала. Минерални дистрибутивни образац који је дала факторска анализа у великој мери разликује се од утиска који имамо након чистог прегледа података.

2. ОСНОВНИ МАТЕМАТИЧКИ И СТАТИСТИЧКИ ПОЈМОВИ

2.1. Неке дефиниције

Излагању факторске анализе може се приступити на више начина. Ми ћемо јој приступити комбинујући линеарну алгебру, вишедимензионалну геометрију и статистику. Сврха овог поглавља је да дефинише и илуструје неке од основних појмова техника факторске анализе.

2.1.1. Матрица података

Табела бројева (Табела 2.1) је пример матрице података. Приказано је осам врста, по једна за сваки примерак камена, и четири колоне, по једна за сваки минерал. Елементи, у овом примеру, нам показују количину одређеног минерала у сваком примерку камена. Матрица података се састоји од p карактеристика мерених на обиму узорка N . Врсте матрице података се називају објекти или узорци и на њима се врше мерења, а колоне матрице података се називају променљиве и представљају карактеристике које меримо на примерцима.

Примерак камена	Врста минерала			
	Кварц	Амфибол	Биотит	Фелдспат
1	4.51	2.66	0.42	4.10
2	6.07	0.58	1.77	1.54
3	6.42	1.32	4.65	4.05
4	4.46	2.16	1.41	4.47
5	8.92	2.54	4.66	4.50
6	7.60	2.39	4.14	4.49
7	5.29	1.69	3.66	3.77
8	4.73	2.65	3.29	5.08

Табела 2.1: Пример матрице података

2.1.2. Матрични производ веће и мање димензије

Два типа производа одређених матрица са њима самима су битна у мултиваријантној статистици. Матрични производ веће димензије је дефинисан као производ матрице помножене здесна својом транспонованом. Дакле, за матрицу X димензије $N \times p$, матрични производ веће димензије је

$$C=XX'$$

Матрица C , матрични производ веће димензије, је квадратна симетрична матрица димензије $N \times N$. За множење две матрице, да бисмо добили матрични производ веће димензије $C=XX'$, потребно је:

$$C_{nm} = \sum_{j=1}^p x_{nj}x_{mj}$$

Дакле, елемент C_{nm} садржи суму производа n -те и m -те врсте елемената матрице C . Главна дијагонала матрице C садржи суму одговарајућих квадрираних елемената врсте матрице X .

Матрични производ мање димензије је дефинисан као множење матрице с лева њеном транспонованом. За матрицу X димензије $N \times p$, матрични производ мање димензије је

$$E = X'X$$

E је квадратна матрица димензије $p \times p$. Елементи матрице E су добијени из $e_{ij} = \sum_{k=1}^N x_{ik}x_{kj}$. Елемент e_{ij} садржи суму производа i -те и j -те колоне матрице X , а елементи на главној дијагонали садрже суму одговарајућих квадрираних елемената колоне матрице X .

2.2. Матрична нотација дескриптивних статистика

Као помоћ у разумевању матричног развоја факторске анализе, представимо неке добро познате основне статистичке једначине у матричној форми.

2.2.1. Аритметичка средина

Формула за израчунавање аритметичке средине матрице X димензије $N \times p$, изражена у матричној нотацији, је

$$\bar{x} = \frac{\mathbf{1}'X}{\mathbf{1}'\mathbf{1}}$$

где је $\bar{x}$ вектор димензије $1 \times p$ чији су елементи аритметичке средине, а $\mathbf{1}$ је јединични вектор димензије $N \times 1$.

2.2.2. Девијација

Разлика између регистроване вредности променљиве и аритметичке средине те променљиве назива се **девијација**. Дакле, девијација, y_{nj} , је дефинисана као

$$y_{nj} = x_{nj} - \bar{x}_j \quad (2.1)$$

Збир девијација за променљиву је нула.

$$\sum_{j=1}^N y_{nj} = \sum_{j=1}^N x_{nj} - \sum_{j=1}^N \bar{x}_j = \sum_{j=1}^N x_{nj} - N\bar{x}_j = \sum_{j=1}^N x_{nj} - N \sum_{j=1}^N \frac{x_{nj}}{N} = 0$$

2.2.3. Варијанса

Варијанса неке променљиве је мера расипања појединачних вредности око аритметичке средине. Дефинисана је као просечна вредност збира квадрата девијација. Користећи (2.1) за девијацију, за j -ту променљиву, варијанса је

$$s_j^2 = \frac{1}{N} \sum_{n=1}^N y_{nj}^2$$

У претходној једнакости користимо дељење са N , укупним обимом узорка, а не са $N - 1$ која нам даје непристрасну оцену варијансе и коваријансе. Требали бисмо да будемо свесни чињенице да се у статистици скоро увек користи дељење са $N - 1$.

Горња једначина може се записати у векторској нотацији на следећи начин

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

2.2.4. Коваријанса

Коваријанса исказује однос између две променљиве. За променљиве x_i и x_j , коваријанса се дефинише као просечна вредност збира производа девијација за све објекте:

$$s_{ij} = \frac{1}{N} \sum_{n=1}^N y_{ni} y_{nj}.$$

Упоредјујући је са формулом за варијансу, видимо да је варијанса специјалан случај формуле за коваријансу. Матрица података, чији су елементи варијансе и коваријансе, назива се **коваријансна матрица** и добија се из

$$\mathbf{S} = \frac{\mathbf{Y}' \mathbf{Y}}{\mathbf{1}' \mathbf{1}} \quad (2.2)$$

Коваријансна матрица је матрични производ мање димензије девијационих матрица података, чији је сваки елемент подељен са N . Дакле, коваријансна матрица је квадратна, симетрична матрица реда p . Елементи на главној дијагонали су варијансе променљивих.

2.2.5. Стандардна девијација

Стандардна девијација је дефинисана као квадратни корен варијансе. Дужину вектора можемо дефинисати као $|\mathbf{y}| = (\mathbf{y}' \mathbf{y})^{1/2}$, где је $\mathbf{y}$ вектор девијације променљиве. Видимо да је стандардна девијација променљиве пропорционална дужини вектора регистроване променљиве. Дакле,

$$s_j = |\mathbf{y}| N^{-\frac{1}{2}} \quad (2.3)$$

Ова веза дозвољава интерпретацију стандардне девијације преко дужине вектора девијационих променљивих.

2.2.6. Стандардизовани резултати

Корисно је изразити регистроване вредности преко њихових одступања од аритметичке средине, користећи стандардну девијацију као јединицу дивергенције.

Стандардизовани резултат за одређени објект променљиве j је дат са:

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

Стандардизоване променљиве имају аритметичку средину једнаку нули, а варијансу (и стандардну девијацију) једнаку један. Дужине вектора променљивих су $\sqrt{N}$. Формула за стандардизацију свих променљивих у матрици података је

$$\mathbf{Z} = \mathbf{Y}\mathbf{D}^{-\frac{1}{2}},$$

где је $\mathbf{D}$ дијагонална матрица формирана од дијагоналних елемената коваријансне матрице $\mathbf{S}$, а $\mathbf{Y}$ је матрица девијације. Употреба ове формуле и других значајних прорачуна биће приказана у раду.

2.2.7. Корелација

Од великог значаја у факторској анализи је мера асоцијације (повезаности) између две променљиве. Пирсонов коефицијент корелације (Пирсоново r) је дефинисан као количник коваријансе две променљиве и производа њихових стандардних девијација

$$r_{ij} = \frac{s_{ij}}{s_i s_j} = \frac{\sum_{n=1}^N y_{ni} y_{nj}}{\left(\sum_{n=1}^N y_{ni}^2 \sum_{n=1}^N y_{nj}^2 \right)^{\frac{1}{2}}}.$$

За матрицу стандардизованих података, формула се може представити као

$$\mathbf{R} = \frac{1}{N} \mathbf{Z}' \mathbf{Z}. \quad (2.4)$$

Корелацијска матрица је квадратна, симетрична матрица. Кораци при рачунању корелацијске и коваријансне матрице из матрице података су следећи:

Почињемо са матрицом девијације, њени елементи су добијени помоћу (2.1). Коваријансну матрицу $\mathbf{S}$ добијамо из

$$\mathbf{S} = \frac{1}{N} \mathbf{Y}' \mathbf{Y}$$

$$\begin{bmatrix} -1.49 & 0.66 & -2.58 & 0.10 \\ 0.07 & -1.42 & -1.23 & -2.46 \\ 0.42 & -0.68 & 1.65 & 0.05 \\ -1.54 & 0.16 & -1.59 & 0.47 \\ 2.92 & 0.54 & 1.66 & 0.50 \\ 1.60 & 0.39 & 1.14 & 0.49 \\ -0.71 & -0.31 & 0.66 & -0.23 \\ -1.27 & 0.65 & 0.29 & 1.08 \end{bmatrix} \begin{bmatrix} -1.49 & 0.07 & 0.42 & -1.54 & 2.92 & 1.60 & -0.71 & -1.27 \\ 0.66 & -1.42 & -0.68 & 0.16 & 0.54 & 0.39 & -0.31 & 0.65 \\ -2.58 & -1.23 & 1.65 & -1.59 & 1.66 & 1.14 & 0.66 & 0.29 \\ 0.10 & -2.46 & 0.05 & 0.47 & 0.50 & 0.49 & -0.23 & 1.08 \end{bmatrix} \begin{bmatrix} 18.00 & 0.00 & 12.73 & 0.00 \\ 0.00 & 3.92 & 0.00 & 4.85 \\ 12.73 & 0.00 & 18.00 & 3.63 \\ 0.00 & 4.85 & 3.63 & 8.00 \end{bmatrix}$$

$$\mathbf{S} = \frac{1}{8} \mathbf{Y}' \mathbf{Y} = \begin{bmatrix} 2.25 & 0.00 & 1.59 & 0.00 \\ 0.00 & 0.49 & 0.00 & 0.61 \\ 1.59 & 0.00 & 2.25 & 0.45 \\ 0.00 & 0.61 & 0.45 & 1.00 \end{bmatrix}$$

Следећи корак, израчунавање корелацијска матрица $\mathbf{R}$, захтева дијагоналну матрицу формирану од реципрочних вредности квадратних корена дијагоналних елемената матрице $\mathbf{S}$

$$\mathbf{D}^{-\frac{1}{2}} = (\text{diag}\mathbf{S})^{-\frac{1}{2}} = \begin{bmatrix} 0.67 & 0 & 0 & 0 \\ 0 & 1.43 & 0 & 0 \\ 0 & 0 & 0.67 & 0 \\ 0 & 0 & 0 & 1.00 \end{bmatrix}$$

Користећи $\mathbf{Y}$, рачунамо матрицу стандардизованих резултата $\mathbf{Z} = \mathbf{Y}\mathbf{D}^{-1/2}$:

$$\mathbf{Y} = \begin{bmatrix} -1.49 & 0.66 & -2.58 & 0.10 \\ 0.07 & -1.42 & -1.23 & -2.46 \\ 0.42 & -0.68 & 1.65 & 0.05 \\ -1.54 & 0.16 & -1.59 & 0.47 \\ 2.92 & 0.54 & 1.66 & 0.50 \\ 1.60 & 0.39 & 1.14 & 0.49 \\ -0.71 & -0.31 & 0.66 & -0.23 \\ -1.27 & 0.65 & 0.29 & 1.08 \end{bmatrix},$$

$$\mathbf{D}^{-\frac{1}{2}} = \begin{bmatrix} 0.67 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.43 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.67 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix},$$

$$\mathbf{Z} = \begin{bmatrix} -0.99 & 0.95 & -1.72 & 0.10 \\ 0.05 & -2.03 & -0.82 & -2.46 \\ 0.28 & -0.97 & 1.10 & 0.05 \\ -1.03 & 0.23 & -1.06 & 0.47 \\ 1.95 & 0.78 & 1.11 & 0.50 \\ 1.07 & 0.56 & 0.76 & 0.49 \\ -0.48 & -0.44 & 0.44 & -0.23 \\ -0.85 & 0.93 & -0.20 & 1.08 \end{bmatrix}.$$

Корелацијска матрица је

$$\mathbf{R} = \frac{1}{8} \mathbf{Z}'\mathbf{Z} = \begin{bmatrix} 1.00 & 0 & 0.71 & 0 \\ 0 & 1 & 0 & 0.87 \\ 0.71 & 0 & 1 & 0.3 \\ 0 & 0.87 & 0.3 & 1 \end{bmatrix}$$

Геометријска интерпретација корелацијске матрице има важну улогу у факторској анализи.

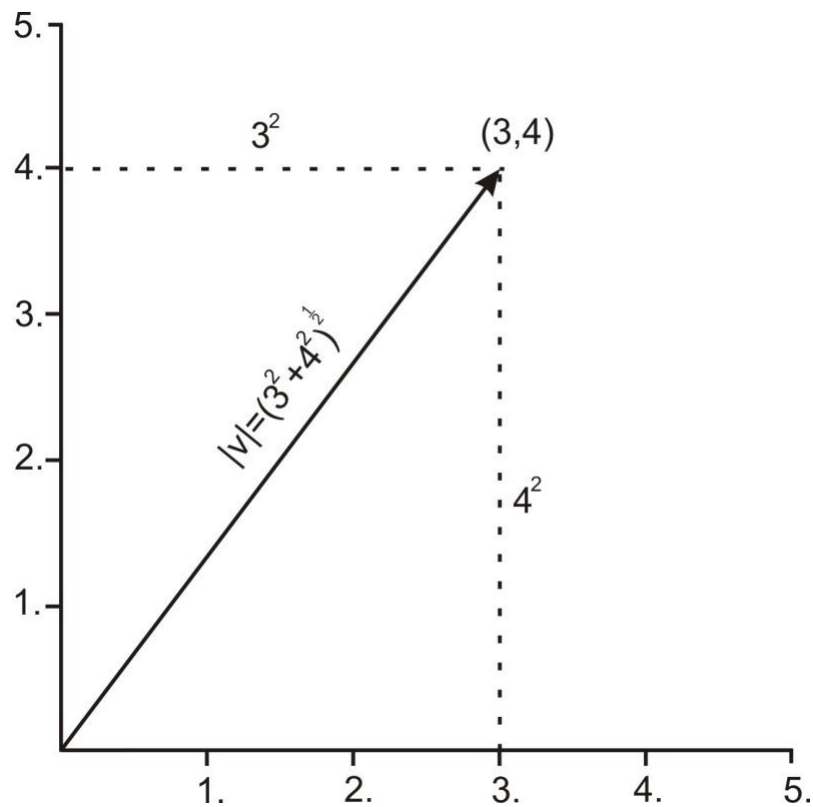
$$r_{ij} = \frac{y'_i y_j}{\left[(y'_i y_i)^{\frac{1}{2}} (y'_j y_j)^{\frac{1}{2}} \right]}, \quad (2.5)$$

Из (2.5), која представља корелацијски коефицијент преко вектора девијације, видимо да је корелацијски коефицијент матрични производ мање димензије два вектора, подељен производом дужина истих вектора. Ово је иста једначина коју користимо за израчунавање угла између два вектора.

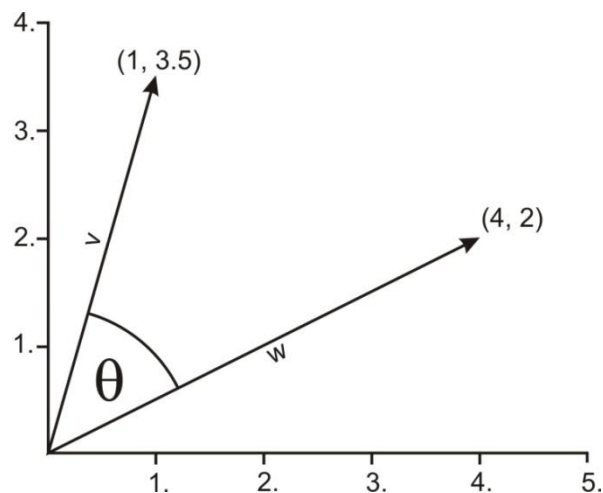
2.2.8. Угао између два вектора

Косинус угла између два вектора, $\mathbf{v}$ и $\mathbf{w}$, добијамо из унутрашњег производа два вектора који је подељен производом њихових дужина:

$$\cos \theta = \frac{\mathbf{v}'\mathbf{w}}{(\mathbf{v}'\mathbf{v})^{\frac{1}{2}}(\mathbf{w}'\mathbf{w})^{\frac{1}{2}}}. \quad (2.6)$$



Слика 2.1: Питагорина теорема примењена на дво-димензионални простор, ради проналажења дужине вектора: $\mathbf{v} = [3,4]$ и $|\mathbf{v}| = [3^2 + 4^2]^{1/2} = 5$.



Слика 2.2: Угао између два вектора

Другим речима, косинус угла између два вектора је количник унутрашњег производа вектора и производа квадратног корена матричних производа мање димензије (дужине). На примеру са Сlike 2.2 видимо да је $\mathbf{v}' = [1 \ 3.5]$ и $\mathbf{w}' = [4 \ 2]$, $\mathbf{v}'\mathbf{w} = 11$, $(\mathbf{v}'\mathbf{v})^{1/2} = 3.64$, $(\mathbf{w}'\mathbf{w})^{1/2} = 4.47$. Сада користећи (2.6), рачунамо

$$\cos \theta = \frac{11}{(3.64)(4.47)} = 0.67,$$

$$\theta = 47.5^\circ.$$

Косинус угла раздвајања може се користити за илустрацију узајамног положаја два вектора. Косинус 1 указује на то да су вектори колинеарни, а косинус 0 да су вектори ортогонални.

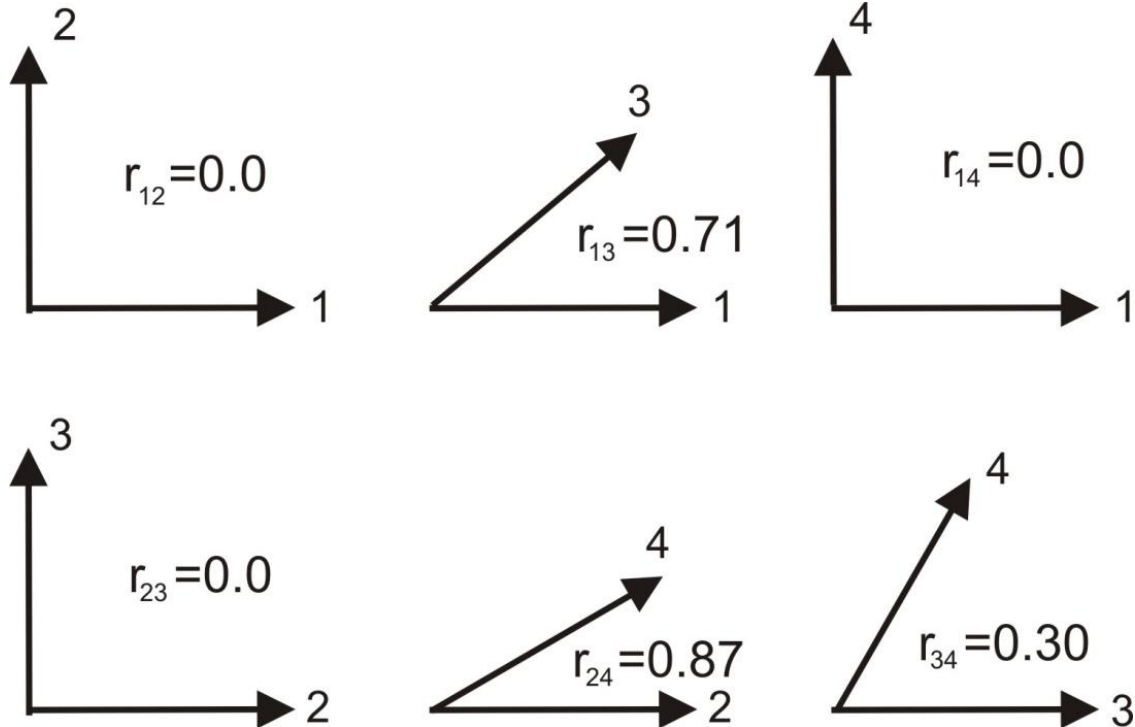
Из (2.2), бројилац из (2.5) је

$$y'_i y_j = N s_{ij}.$$

Из (2.3), $|y_i| = N^{1/2} s_i$ и $|y_j| = N^{1/2} s_j$.

Дакле,

$$\cos \theta = \frac{N s_{ij}}{N^{1/2} s_i N^{1/2} s_j} = \frac{s_{ij}}{s_i s_j} = r_{ij}.$$



Слика 2.3: Неки примери корелације израчунатих из матрице података Табеле 2.1

За стандардизоване променљиве, једначина гласи

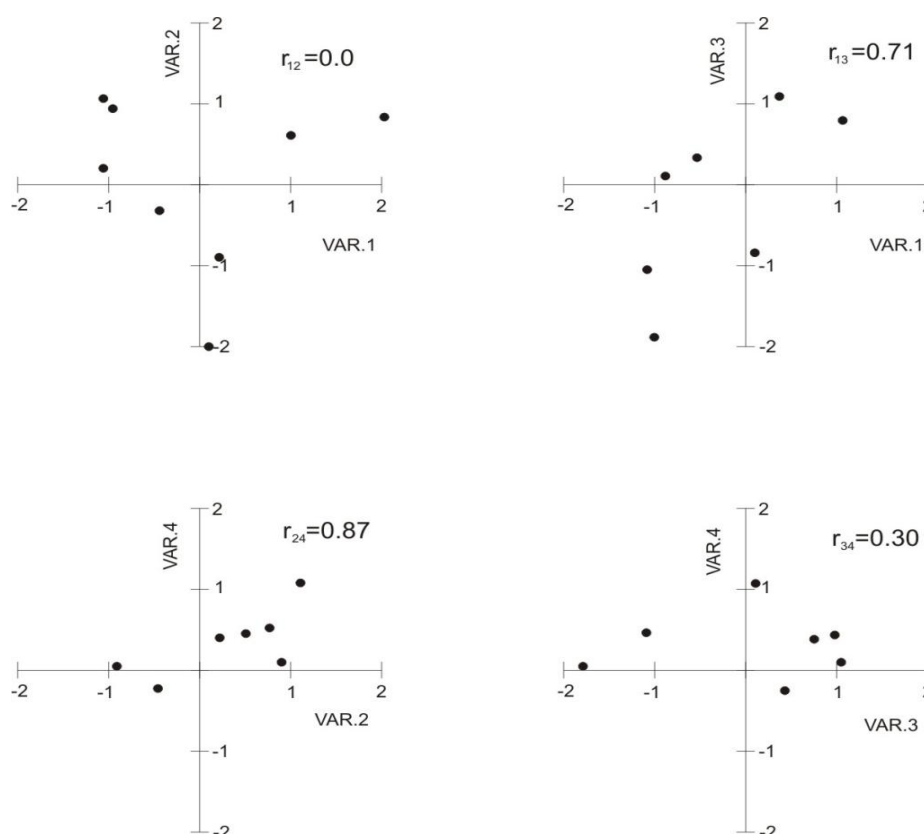
$$\cos \theta = \frac{z'_i z_j}{(|z_i| |z_j|)}$$

Ово се смањује на $\cos \theta = \frac{1}{N} z'_i z_j$ зато што је дужина стандардизованог вектора $\sqrt{N}$. Због (2.4), $z'_i z_j = N r_{ij}$. Дакле,

$$\cos \theta = \frac{1}{N} z'_i z_j = r_{ij}$$

Ове формуле воде до интерпретације корелације између променљивих, представљених у простору као углови између вектора променљивих. Корелацијска матрица R , садржи косинусе свих могућих углова раздвајања, између свих променљивих. Ова идеја је представљена на Слици 2.3.

Други начин за интерпретирање корелацијског коефицијента је помоћу дијаграма расипања. На Слици 2.4 објекти матрице података из Табеле 2.1 се цртају као тачке постављене у простору стандардизованих парова променљивих. Корелацијски коефицијент мери тежњу тачака да се групишу уз линију. Ове интерпретације су корисне за факторску анализу.



Слика 2.4: Дијаграми расипања за табелу 2.1 са приказаним корелацијским коефицијентима

2.2.9. Регресија, вишеструка и парцијална корелација

Неке од основних идеја факторске анализе ослањају се на концепте као што су регресија, вишеструка корелација и парцијална корелација. Ово су веома важни појмови у мултиваријантној статистици. У овом раду даћемо само њихов кратак преглед.

Разматраћемо p променљивих $y_1, y_2, \dots, y_p$ са по N регистрованих вредности. Због једноставности, претпоставићемо да је свака променљива мерена као девијација своје аритметичке средине, тако да је матрица података Y димензије $N \times p$ матрица девијације. Уобичајена процедура за изучавање линеарних међуодноса над променљивама је разматрање **линеарне регресије** сваке променљиве у односу на остале. Претпоставимо да је $p=4$ и размотримо y_1 . Ову регресију записујемо

$$y_1 = b_{12}y_2 + b_{13}y_3 + b_{14}y_4 + u_1 \quad (2.7)$$

где са u_1 означавамо резидуал (остатак).

Коефицијенти регресије b_{12}, b_{13} и b_{14} се обично одређују методом најмањег квадрата, одатле следи:

$$\sum_{n=1}^N (y_{n1} - b_{12}y_{n2} - b_{13}y_{n3} - b_{14}y_{n4})^2$$

Може се показати да је

$$b_{1j} = \frac{s^{1j}}{s^{11}} ; \quad j = 2, 3, 4,$$

где су s^{1j} и s^{11} елементи инверзне матрице S^{-1} коваријансне матрице S . Линеарна комбинација

$$y_1^* = b_{12}y_2 + b_{13}y_3 + b_{14}y_4$$

је најбоља линеарна оцена y_1 у смислу најмањих квадрата која се може добити из y_2, y_3 , и y_4 . **Резидуал** u_1 из (2.7) представља онај део y_1 који се не може објаснити или предвидети из y_2, y_3 , и y_4 . Може се показати да y_1^* и u_1 нису у корелацији, само ако је варијанса од y_1 сума варијансе од y_1^* и варијансе од u_1 . Пропорција варијансе променљиве y_1 изражена са y_2, y_3 , и y_4 је

$$R_1^2 = \frac{s_{11} - \frac{1}{s^{11}}}{s_{11}} = 1 - \frac{1}{s_{11}s^{11}}$$

и зове се **вишеструка квадратна корелација** променљиве y_1 . Позитиван квадратни корен овога је корелација између y_1 и y_1^* .

До сада смо пажњу усмеравали на прву променљиву y_1 . Међутим, на сличан начин може се наставити са разматрањем регресије y_2 у односу на остале променљиве (укључујући и y_1), y_3 у односу на остале променљиве итд. У општем облику можемо записати:

$$y = By + u$$

где је y вектор који се састоји од свих p променљивих, u је вектор који се састоји од свих p резидуала и B је матрица чији су дијагонални елементи нуле (пошто y_i не постоји са десне стране i -те једначине) а елементи ван дијагонале су коефицијенти регресије. Дакле,

$$B = I - (diag S^{-1})^{-1} S$$

и варијансе резидуала су дате као реципрочне вредности дијагоналних елемената матрице S^{-1} . Да бисмо увели појам парцијалне корелације, разматрамо поново случај кад је $p=4$, а регресију променљивих y_1 и y_2 преко y_3 и y_4 , записујемо

$$y_1 = b_{13}y_3 + b_{14}y_4 + u_1, \quad (2.8)$$

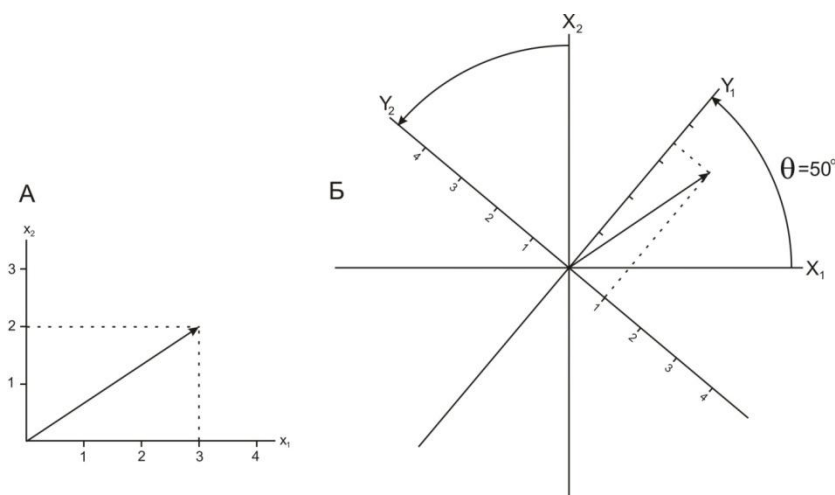
$$y_2 = b_{23}y_3 + b_{24}y_4 + u_2.$$

Корелација између u_1 и u_2 се назива **парцијална корелација** између y_1 и y_2 када су y_3 и y_4 елиминисани. То је мера повезаности између y_1 и y_2 с обзиром да су y_3 и y_4 непромењене.

Треба нагласити да коефицијенти регресије b_{13} и b_{14} из (2.8) нису исти они из (2.7). На пример, b_{13} из (2.7) представља ефекат y_3 на y_1 када се ефекти y_2 и y_4 узимају у обзир, а b_{13} из (2.8) представља ефекат y_3 на y_1 када се у обзир узима само ефекат y_4 , односно када је y_2 искључен. Ако y_2 има било какав ефекат на y_1 , резидуална променљива из (2.7) је мања од оне из (2.8).

2.3. Ротација координатног система

Координатни системи се могу ротирати матричним операцијама. Упознаћемо се са неколико ситуација у којима је неопходно ротирати координатни систем.



Слика 2.5: (А) Позиција вектора врсте $x' = [3 \ 2]$ дата координатама на ортогоналним осама x_1 и x_2 . (Б) Ротација за 50° којом остаје очувана ортогоналност оса. Координате вектора врсте y' су дате рачуном

$$y_1 = (0.6428 \times 3) + (0.7660 \times 2) = 3.46,$$

$$y_2 = (-0.7660 \times 3) + (0.6428 \times 2) = -1.01.$$

2.3.1. Алгебра ротације

На Слици 2.4(А), елементи вектора x дају координате тачке у координатном систему са две променљиве. Претпоставимо да из неког разлога желимо да ротирамо осе за Φ° у смеру супротном од кретања казаљке на сату, а да притом остане очувана ортогоналност оса, као што је приказано на Слици 2.4(Б). Тада је проблем наћи координате тачке у односу на нове осе y_1 и y_2 .

Користећи основне тригонометријске функције, тражене координате се добијају једначинама

$$y_1 = \cos \Phi x_1 + \sin \Phi x_2,$$

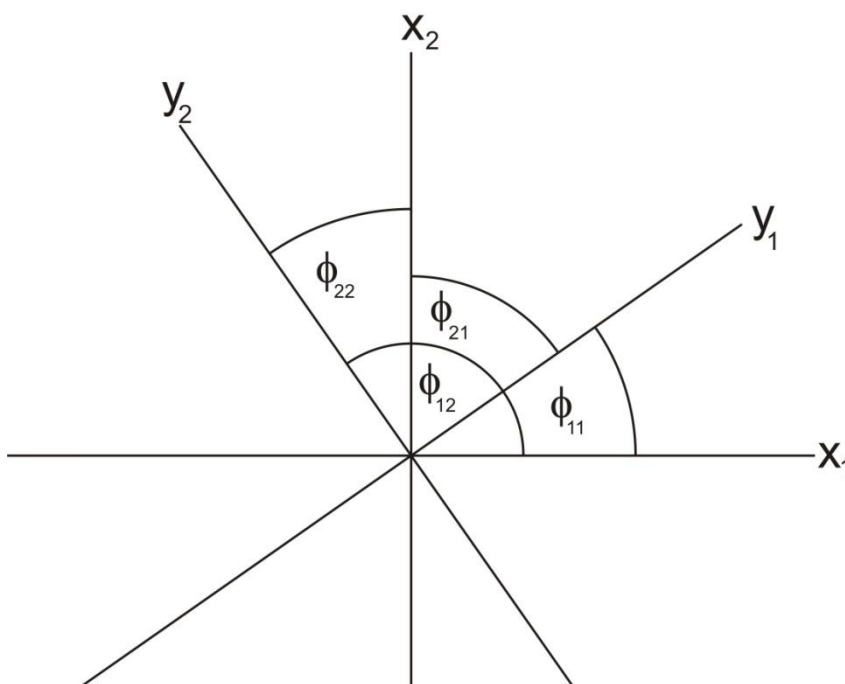
$$y_2 = -\sin \Phi x_1 + \cos \Phi x_2.$$

У матричној нотацији, ове две једначине могу бити записане као

$$y' = xT \quad (2.9)$$

где је

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



Слика 2.6: Угаоне везе између пара оригиналних координатних оса x_1 и x_2 и ротираних координатних оса y_1 и y_2 .

Да бисмо проширили ову корисну везу на општији случај, уводимо обележавање приказано на Слици 2.5. Углови Φ_{ij} представљају углове између i -те старе координатне

осе и j -те нове осе. За ротацију при којој остаје очувана ортогоналност оса и илустровану дво-димензионалну шему, примећујемо следеће везе између углова:

$$\Phi_{12} = \Phi_{11} + 90^\circ,$$

$$\Phi_{21} = \Phi_{11} - 90^\circ,$$

$$\Phi_{22} = \Phi_{11}.$$

Применом тригонометријских функција добијамо

$$\sin \Phi_{11} = \sin(\Phi_{21} + 90^\circ) = \cos \Phi_{21},$$

$$-\sin \Phi_{11} = -\sin(\Phi_{12} - 90^\circ) = \sin(90^\circ - \Phi_{12}) = \cos \Phi_{12}.$$

Заменом у (2.9) добијамо

$$y_1 = \cos \Phi_{11} x_1 + \cos \Phi_{21} x_2,$$

$$y_2 = \cos \Phi_{12} x_1 + \cos \Phi_{22} x_2.$$

Цела процедура може бити уопштена за било који број ортогоналних оса и њихових ротираних еквивалената. Систем једначина за p оса је:

$$y_1 = \cos \Phi_{11} x_1 + \cos \Phi_{21} x_2 + \dots + \cos \Phi_{p1} x_p$$

$$y_p = \cos \Phi_{1p} x_1 + \cos \Phi_{2p} x_2 + \dots + \cos \Phi_{pp} x_p$$

Сменом $t_{ij} = \cos \Phi_{ij}$, где је i индекс који означава осе оригиналног координатног система, а j означава осе ротираног система, једначине се у матричној нотацији могу краће записати као

$$\mathbf{y}' = \mathbf{x}'\mathbf{T}$$

За скуп од N вектора врста у матрици $\mathbf{X}_{(N \times p)}$, једначина

$$\mathbf{Y} = \mathbf{X}\mathbf{T} \tag{2.10}$$

ће у $\mathbf{Y}$ дати координате свих N вектора врста у односу на p ротираних оса. Матрица $\mathbf{T}$ се назива **матрица трансформације**. У циљу да трансформација одговара ротацији, $\mathbf{T}$ мора бити ортонормирана матрица, што значи да мора да важи $\mathbf{T}'\mathbf{T} = \mathbf{I}$.

Очигледно, растојање и углови између вектора се не мењају за време ротације при којој остаје очувана ортогоналност поменутих оса.

2.3.2. Интерпретација ротације

Множење матрице података здесна ортонормираном матрицом може се сматрати једноставним, механичким поступком ротације при којој остаје очувана ортогоналност оса. Две одлике овог поступка су корисне за пружање додатног увида за коришћење података помоћу матричних операција. Прва, нове, ротиране осе могу и саме бити интерпретиране као променљиве. Вектори колоне у $\mathbf{Y}$ из (2.10) су нове променљиве које су линеарне комбинације вектора колоне (променљивих) од $\mathbf{X}$. Елементи матрице $\mathbf{T}$ су према томе коефицијенти из линеарних комбинација. Друга, вектори врсте матрице $\mathbf{X}$

дају саставе објеката при мерењу регистрованих променљивих. Два вектора матрице Y дају саставе објеката у односу на ново изведене променљиве.

2.4. Ранг матричног производа

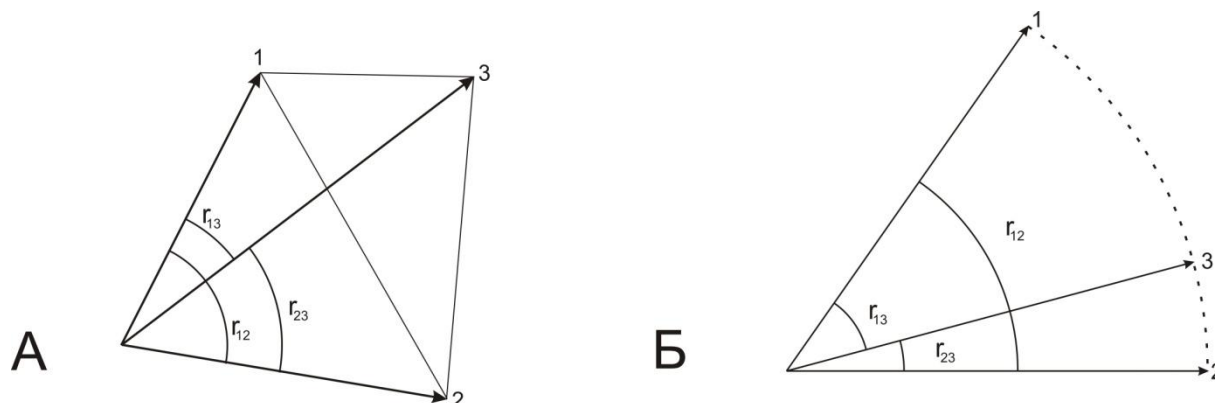
Један од циљева факторске анализе је проналажење ранга матрице података. То чинимо тако што утврђујемо најмањи број линеарно независних вектора који могу да чине базу простора који садржи и векторе колоне и векторе врсте чинећи матрицу. Многе методе факторске анализе одређују ранг анализом матричног производа матрице података пре него анализом саме матрице података. Дакле, ранг матричног производа матрице је једнак броју фактора. Ако је $X_{(N \times p)}$ ранга r и

$$B_{(p \times p)} = X'_{(p \times N)} X_{(N \times p)},$$

онда је B ранга r . Слично важи, ако

$$C_{(N \times N)} = X_{(N \times p)} X'_{(p \times N)},$$

C је ранга r .



Слика 2.7: (А) Линеарно независни вектори конструишу тетраедар чија је запремина пропорционална са детерминантом матрице. (Б) Сингуларна корелацијска матрица је сачињена од копланарних вектора. Запремина тако конструисаног тетраедра је нула.

Корелацијска матрица је почетна тачка факторске анализе. Као што је већ речено, ова матрица је матрични производ мање димензије стандардизоване матрице података. Штавише, елементи ове матрице садрже косинусе углова између свих могућих парова стандардизованих вектора колоне. Могуће је геометријски приказати везу између вектора чији су косинуси представљени корелацијама. На Слици 2.7(А), на пример, приказана су три вектора која одговарају корелацијској матрици

$$R_1 = \begin{bmatrix} 1 & 0.5 & 0.866 \\ 0.5 & 1 & 0.707 \\ 0.866 & 0.707 & 1 \end{bmatrix}.$$

Да бисмо одражали доследност углова датих елементима матрице, вектори морају бити представљени у тро-димензионалном простору. Даље, ови вектори су линеарно независни, па је ранг матрице једнак 3. На Слици 2.7(Б), су представљени вектори друге корелацијске матрице R_2

$$R_2 = \begin{bmatrix} 1 & 0.5 & 0.707 \\ 0.5 & 1 & 0.966 \\ 0.707 & 0.966 & 1 \end{bmatrix}.$$

Геометријски услови приморавају три вектора да леже у дво-димензионалном простору. Дакле, сваки од ова три вектора је линеарно зависан са преостала два, па је ранг ове матрице једнак 2.

Приметимо да тетраедар можемо конструисати спајањем врхова вектора, као што је приказано на Слици 2.7(А). Запремина овог тетраедра је пропорционална са детерминантом корелацијске матрице. У овим примерима, имамо

$$|R_1| = 0.5 - (-0.0562) + (-0.4438) = 0.1124,$$

$$|R_2| = 0.0668 - (-0.0915) + (-0.1583) = 0.$$

С обзиром да су вектори приказани на Слици 2.7(Б) копланарни, запремина тетраедра конструисаног на тај начин је једнака нули. У вези са тим, детерминанта корелацијске матрице је нула.

Веза између линеарно независних вектора, ранга, детерминанте и геометријске конфигурације вектора може се уопштити за простор било које димензије. Претходна дискусија о рангу матричног производа доводи нас до следећих закључака:

- 1) Свака матрица, чији је ранг r мањи од њене мање димензије p , може се посматрати кроз два типа вектора које садржи: скуп од $(p - r)$ линеарно зависних вектора и скуп од r линеарно независних вектора.
- 2) Ранг одређује димензију простора који садржи све векторе, а r линеарно независних вектора чине једну базу од бесконачно много могућих база тог простора.

Ранг матрице се може одредити било којом од ових метода.

2.5. Карактеристични корени и вектори

Ранг ће нам дати димензију матрице која одрђује број линеарно независних вектора потребних за генерисање простора који садржи векторе матрице. Карактеристични корени и вектори матрице нам неће само одредити ранг, већ и скуп линеарно независних вектора база. Нашли су своју примену у многим научним темама и због разноврсности тема на које се примењују, отворени су за велики избор интерпретација.

2.5.1. Алгебарски преглед

Ограничимо нашу дискусију на карактеристичне вредности реалних, симетричних матрица R , које могу бити, на пример, матрични производ мање или веће димензије матрице података, коваријансна или корелацијска матрица.

Карактеристични вектор матрице R је вектор u , тако да важи

$$Ru = u\lambda, \quad (2.11)$$

где је λ непознати скалар, а вектор u мора имати бар један елемент различит од нуле.

У суштини, треба пронаћи вектор такав да је вектор $\mathbf{Ru}$ пропорционалан са $\mathbf{u}$. Други облик ове једначине је

$$\mathbf{Ru} - \mathbf{u}\lambda = \mathbf{0},$$

односно,

$$(\mathbf{R} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0}, \quad (2.12)$$

где $\mathbf{0}$ представља нула вектор.

Одавде следи да је непознати вектор $\mathbf{u}$ ортогоналан на све векторе врсте матрице $(\mathbf{R} - \lambda\mathbf{I})$. Једначина (2.12) представља систем хомогених једначина. За матрицу $\mathbf{R}$ димензије 2×2 , важи

$$(r_{11} - \lambda)u_1 + r_{12}u_2 = 0,$$

$$r_{21}u_1 + (r_{22} - \lambda)u_2 = 0.$$

Ако $\mathbf{R} - \lambda\mathbf{I}$ није сингуларна, (2.12) се може решити множењем обе стране једначине слева са $(\mathbf{R} - \lambda\mathbf{I})^{-1}$; међутим, ово би нам донело тривијално решење $\mathbf{u} = \mathbf{0}$. Због тога, једина могућност да добијемо нетривијално решење је да $(\mathbf{R} - \lambda\mathbf{I})$ буде сингуларна, односно, детерминанта мора бити једнака нули. Дакле,

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

За матрицу 2×2 , општи облик **карактеристичне једначине** се добија на следећи начин:

$$\begin{vmatrix} r_{11} - \lambda & r_{12} \\ r_{21} & r_{22} - \lambda \end{vmatrix} = 0,$$

одакле добијамо

$$(r_{11} - \lambda)(r_{22} - \lambda) - r_{12}r_{21} = 0,$$

$$\lambda^2 - \lambda(r_{11} + r_{22}) + (r_{11}r_{22} - r_{12}r_{21}) = 0.$$

Последња једначина је квадратна (знамо да је $r_{12} = r_{21}$). Добијамо две вредности за λ , λ_1 и λ_2 . Ова два корена називамо **карактеристичним вредностима** матрице $\mathbf{R}$. Треба запамтити да је

$$\lambda_1 + \lambda_2 = r_{11} + r_{22},$$

$$\lambda_1\lambda_2 = r_{11}r_{22} - r_{12}r_{21} = |\mathbf{R}|.$$

У општем случају $\mathbf{R}_{(p \times p)}$, карактеристична једначина биће полином степена p из које добијамо p корена или карактеристичних вредности $\lambda_1, \lambda_2, \dots, \lambda_p$. Ако је $\mathbf{R}$ реална, квадратна, симетрична матрица, онда су и карактеристичне вредности реалне. Међутим, ових p карактеристичних вредности не морају бити међусобно различите и неке могу бити једнаке нули. Ако су две или више карактеристичних вредности једнаке, имамо вишеструку карактеристичну вредност; у супротном, карактеристична вредност је јединствена.

Када добијемо карактеристичну вредност, карактеристичне векторе можемо добити из (2.12). Ако је u решење, онда је и cu решење (c је скалар). Карактеристични вектори су увек нормирани.

Нека су карактеристичне вредности $\lambda_i (i = 1, \dots, p)$ поређане по главној дијагонали матрице Λ , и нека колоне матрица U представљају добијене карактеристичне векторе, тада (2.11) записујемо у матричној нотацији

$$RU = U\Lambda. \quad (2.13)$$

Матрица U је квадратна и ортонормирана, па важи

$$U'U = UU' = I.$$

Множењем (2.13) здесна матрицом U' добијамо

$$R = U\Lambda U',$$

једначину која има фундаменталну улогу у многим мултиваријантним методама. Она нам показује да симетрична матрица R може бити представљена преко својих карактеристичних корена и вектора. Други начин да је запишемо је

$$R = \lambda_1 u_1 u_1' + \lambda_2 u_2 u_2' + \dots + \lambda_p u_p u_p'.$$

Матрица R је линеарна комбинација производа λ_i и матрица $u_i u_i'$ димензије $p \times p$ и ранга 1.

Множењем с лева (2.13) са U' добијамо

$$\Lambda = U'RU,$$

из које видимо да је U матрица која своди R на дијагоналну матрицу.

Навешћемо још корисних особина карактеристичних вредности:

- 1) $tr \Lambda = tr R$; сума карактеристичних вредности једнака је суми елемената са главне дијагонале матрице.
- 2) $\prod_{i=1}^p \lambda_i = |R|$; производ карактеристичних вредности једнак је детерминанти матрице.
- 3) Број ненула карактеристичних вредности једнак је рангу матрице R .

Сада следи пример рачуна за једноставан проблем карактеристичне вредности.

Да бисмо илустровали израчунавање карактеристичних вредности и вектора, посматраћемо матрицу података $X_{(8 \times 2)}$:

$$X = \begin{bmatrix} -8 & -1 \\ 6 & 10 \\ -2 & -10 \\ 8 & 1 \\ 0 & 3 \\ -6 & -6 \\ 0 & -3 \\ 2 & 6 \end{bmatrix}$$

Корак 1. Рачунамо матрични производ мање димензије, $\mathbf{X}'\mathbf{X} = \mathbf{C}$:

$$\mathbf{C} = \begin{bmatrix} 208 & 144 \\ 144 & 292 \end{bmatrix}$$

Корак 2. Записујемо карактеристичну једначину:

$$\begin{aligned} \begin{vmatrix} 208 - \lambda & 144 \\ 144 & 292 - \lambda \end{vmatrix} &= 0, \\ (208 - \lambda)(292 - \lambda) - 144 \cdot 144 &= 0, \\ \lambda^2 - 500\lambda + 40\,000 &= 0. \end{aligned}$$

Корак 3. Рачунамо корене:

$$\lambda_1 = 400 \text{ и } \lambda_2 = 100.$$

Корак 4. Убацујемо ове вредности у следећу једначину да бисмо добили карактеристичне векторе за λ_1 :

$$(\mathbf{C} - \lambda_1 \mathbf{I})\mathbf{u}_1 = \mathbf{0};$$

$$\begin{aligned} \left[\begin{bmatrix} 208 & 144 \\ 144 & 292 \end{bmatrix} - \begin{bmatrix} 400 & 0 \\ 0 & 400 \end{bmatrix} \right] \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \\ -192u_{11} + 144u_{21} &= 0, \\ 144u_{11} - 108u_{21} &= 0, \\ -192u_{11} + 144u_{21} &= 144u_{11} - 108u_{21}, \\ -336u_{11} &= -252u_{21}, \\ u_{11} &= 0.75u_{21}. \end{aligned}$$

Нека је $u_{11} = 1$, добијамо $u_{21} = 1.33$ и нормализујемо их

$$\begin{aligned} u_{11} &= \frac{1}{1.66} = 0.6, \\ u_{21} &= \frac{1.33}{1.66} = 0.8. \end{aligned}$$

Дакле, први карактеристични вектор је

$$\mathbf{u}_1 = \begin{bmatrix} 0.6 \\ 0.8 \end{bmatrix}.$$

За λ_2 ,

$$\left[\begin{bmatrix} 208 & 144 \\ 144 & 292 \end{bmatrix} - \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix} \right] \begin{bmatrix} u_{12} \\ u_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$108u_{12} + 144u_{22} = 0,$$

$$144u_{12} + 192u_{22} = 0,$$

$$u_{12} = -1.33u_{22};$$

Дакле, други карактеристични вектор је

$$u_2 = \begin{bmatrix} -0.8 \\ 0.6 \end{bmatrix}$$

Корак 5. Проверавамо резултате,

$$Cu = \lambda u :$$

За λ_1 :

$$\begin{bmatrix} 208 & 144 \\ 144 & 292 \end{bmatrix} \begin{bmatrix} 0.6 \\ 0.8 \end{bmatrix} = \begin{bmatrix} 240 \\ 320 \end{bmatrix},$$

$$400 \begin{bmatrix} 0.6 \\ 0.8 \end{bmatrix} = \begin{bmatrix} 240 \\ 320 \end{bmatrix}.$$

За λ_2 :

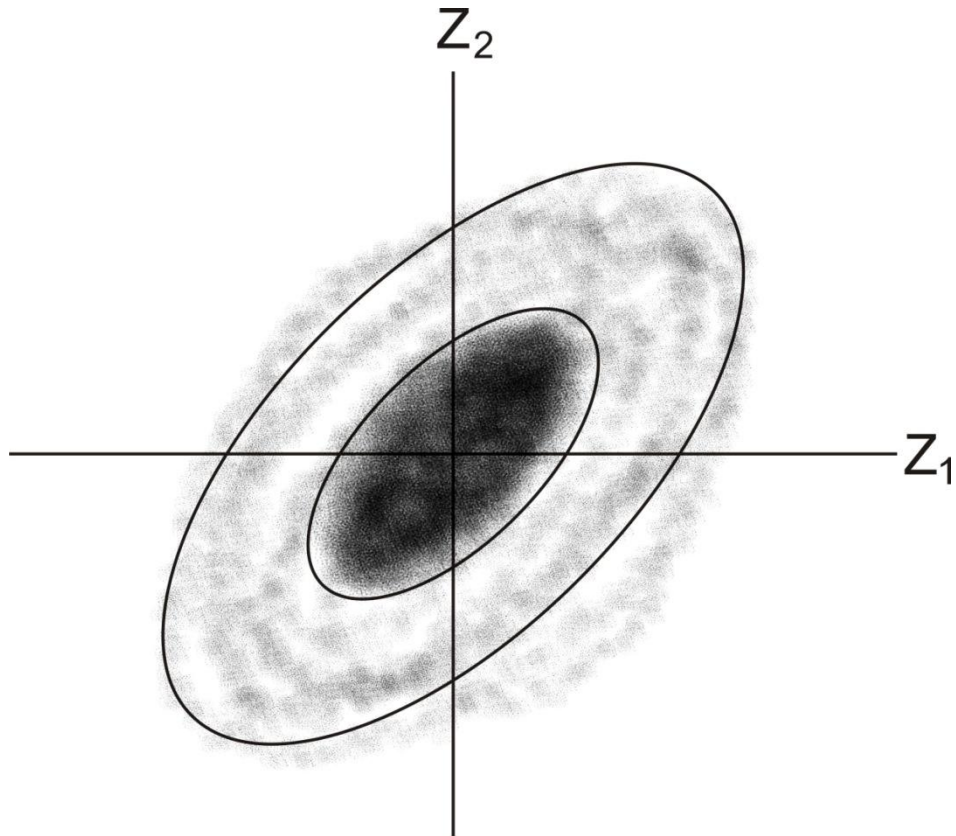
$$\begin{bmatrix} 208 & 144 \\ 144 & 292 \end{bmatrix} \begin{bmatrix} -0.8 \\ 0.6 \end{bmatrix} = \begin{bmatrix} -80 \\ 60 \end{bmatrix},$$

$$100 \begin{bmatrix} -0.8 \\ 0.6 \end{bmatrix} = \begin{bmatrix} -80 \\ 60 \end{bmatrix}.$$

Управо смо показали да је ранг квадратне симетричне матрице једнак броју не нула карактеристичних вредности те матрице. Тачније, ако за квадратну матрицу R димензије $p \times p$ имамо m нула карактеристичних вредности, ранг матрице R је $p - m$. Штавише, скуп од $p - m$ ортогоналних карактеристичних вектора повезаних са $p - m$ не нула карактеристичних вредности формира скуп ортогоналних вектора база који генеришу простор од $p - m$ линеарно независних вектора матрице R .

2.5.2. Геометријска интерпретација карактеристичних вредности и вектора

Слика 2.8 показује дијаграм расипања података две променљиве. Претпоставља се да су подаци у стандардизованом облику и да променљиве имају нормалну расподелу. Тачке података могу бити опасане контурама. Унутрашња контура окружује 66% тачака, а спољашња 95%. Добро је познато да за две променљиве које имају нормалну расподелу ове контуре образују елипсе. Ако су променљиве некорелисане, елипсе ће добити облик круга; ако су потпуно корелисане, елипсе ће се претворити у праву. За више од две променљиве (p променљивих), тачке података ће обликовати p -димензионални хиперелипсоид.



Слика 2.8: Дијаграм расипања две променљиве са контурним линијама једнаке густине. Унутрашња елипса окружује 66% тачака података, спољашња елипса 95% тачака података.

Пирсон (1901) и касније, Хотелинг (1933) показују да се главна и споредна оса хиперелипсоида могу добити из карактеристичних вектора корелацијске матрице. Једначина хиперелипсоида је

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

где је $\mathbf{R}$ корелација коваријансне матрице података, $\mathbf{w}$ је вектор координата за тачке елипсоида, а c је константа. Главна оса елипсоида је одређена тачкама на елипсоиду који се налази најдаље од тежишта. Да бисмо пронашли ове тачке, потребно је само пронаћи тачке на елипсоиду које су најудаљеније од координатног почетка. Квадратно растојање од тачке $\mathbf{w}$ на елипсоиду до координатног почетка је $\mathbf{w}'\mathbf{w}$, па је потребно пронаћи максимум функције $\mathbf{w}'\mathbf{w}$ уз услов $\mathbf{w}'\mathbf{R}^{-1}\mathbf{w} = c$. Користећи једноставан рачун да бисмо решили овај проблем долазимо до следећег резултата:

$$\mathbf{R}\mathbf{w} = \lambda\mathbf{w},$$

где је $\mathbf{w}'\mathbf{w} = \lambda$. Дакле, $\mathbf{w}$ мора бити карактеристични вектор матрице $\mathbf{R}$, који одговара највећој карактеристичној вредности λ . Карактеристични вектор $\mathbf{w}$ дефинише правац главне осе елипсоида. Дужина главне осе, односно, растојање од координатног почетка до тачке на елипсоиду која је најудаљенија од координатног почетка, је $\sqrt{\lambda}$.

Сличним расуђивањем, може се показати да други карактеристични вектор одговара највећој споредној оси елипсоида, трећи карактеристични вектор одговара трећој највећој споредној оси елипсоида, итд.

Интерпретација карактеристичних вредности вектора коваријансних и корелацијских матрица нас доводи до неколико корисних уопштења:

1. Положај карактеристичних вектора на главној оси хиперелипсоида у суштини одређује положај карактеристичних вектора како би се поклапали са правцима максималне варијансе података. Карактеристични вектор чији је карактеристични корен највећи одређује правац максималне варијансе података; карактеристични вектор повезан са другим највећим карактеристичним кореном одређује други правац максималне варијансе који је ортогоналан на први.
2. Положај карактеристичних вектора може се посматрати као проблем ротације или трансформације. Оригиналне осе, представљајући оригиналне променљиве, се ротирају на нови положај према претходно размотреном критеријуму. Врсте матрице карактеристичних вектора су коефицијенти који ротирају осе променљивих на положај дуж главне и споредне осе. Елементи вектора могу се посматрати као косинуси углова ротације од старог до новог система координатних оса. Колоне матрице карактеристичних вектора дају правац косинуса самих карактеристичних вектора.
3. Карактеристични вектори су линеарно независни вектори који представљају линеарне комбинације оригиналних променљивих. Као такви, могу бити посматрани као „нове“ променљиве са пожељном особином да су некорелисане.
4. Квадратни корен карактеристичне вредности може бити коришћен као стандардна девијација „нове“ променљиве представљене карактеристичним вектором.
5. Карактеристична вредност једнака нули означава да је одговарајућа споредна оса дужине нула. Ово нас наводи да је димензија простора који садржи тачке података, мања од оригиналног простора, односно, да је ранг матрице података мањи од њене најмање димензије.

Укратко, карактеристични вектори удружени са не нула карактеристичним вредностима коваријансне или корелацијске матрице представљају скуп ортогоналних оса на које се тачке података могу односити. Они су скуп линеарно независних вектора база чије позиције су постављене тако да би одредили максималну варијансу, варијанса је дата одговарајућим карактеристичним вредностима. Димензија простора је дата бројем не нула карактеристичних вредности.

Претходна интерпретација карактеристичних вредности и вектора може такође бити примењена на векторску интерпретацију корелације.

2.6. Апроксимација матрице методом најмањих квадрата

У делу 2.5, показали смо како свака симетрична матрица S димензије $p \times p$ може бити приказана преко својих карактеристичних вредности $\lambda_1 > \lambda_2 > \dots > \lambda_r$ и карактеристичних вектора $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r$ на следећи начин

$$S = \lambda_1 \mathbf{u}_1 \mathbf{u}_1' + \lambda_2 \mathbf{u}_2 \mathbf{u}_2' + \dots + \lambda_r \mathbf{u}_r \mathbf{u}_r',$$

где је r ранг матрице S . Слично, свака правоугаона матрица X димензије $N \times p$ и ранга r може бити приказана помоћу својих сингуларних вредности $\gamma_1 > \gamma_2 > \dots > \gamma_r$ и вектора $v_1, v_2, \dots, v_r$ и $u_1, u_2, \dots, u_r$ на следећи начин

$$X = \gamma_1 v_1 u'_1 + \gamma_2 v_2 u'_2 + \dots + \gamma_r v_r u'_r.$$

Важније од ова два приказа јесу особине најмањег квадрата и узастопних услова у њима. Решења следећа два проблема су у вези са тим.

Проблем 1.

Дата је симетрична Грамова матрица (Грамова матрица има не-негативне карактеристичне вредности) $S_{(p \times p)}$ ранга r . Наћи симетричну Грамову матрицу $T_{(p \times p)}$ мањег ранга $k < r$, који апроксимира матрицу S методом најмањег квадрата.

Проблем 2.

Дата је правоугаона матрица $X_{(N \times p)}$ ранга r и број $k < r$. Наћи матрицу $W_{(N \times p)}$ ранга k који апроксимира матрицу X методом најмањих квадрата.

Решења ових проблема су:

$$T = \lambda_1 u_1 u'_1 + \lambda_2 u_2 u'_2 + \dots + \lambda_k u_k u'_k,$$

$$W = \gamma_1 v_1 u'_1 + \gamma_2 v_2 u'_2 + \dots + \gamma_k v_k u'_k.$$

респективно. Они су суме првих k чланова у свакој од приказа. Матрице T и W се називају апроксимацијама снижавања ранга методом најмањих квадрата матрица S и X респективно. Особина најмањих квадрата се може речима записати на следећи начин T је близу S колико год је то могуће у смислу да је сума квадрата свих елемената матрице $S-T$ минимална. Слично, W је близу X колико год је то могуће у смислу да је сума квадрата свих елемената матрице $X-W$ минимална. Мера блискости или ваљаности апроксимације је, у оба случаја,

$$\lambda_{k+1} + \lambda_{k+2} + \dots + \lambda_r;$$

односно, ова мера је сума $r - k$ најмањих карактеристичних вредности матрице XX . Понекад се мера односа

$$\frac{\lambda_{k+1} + \lambda_{k+2} + \dots + \lambda_r}{\lambda_1 + \lambda_2 + \dots + \lambda_r}$$

или нека повезана функција користе као мера ваљаности апроксимације.

Важност ова два проблема ћемо продискутовати на најчешћем случај $r = p$. У **Проблему 1**, S може бити коваријансна матрица од p линеарно независних променљивих. Матрица T такође може представљати коваријансну матрицу од p променљивих, али пошто је ранг матрице T једнак $k < p$, ових p променљивих су линеарно зависне. Дакле, оригиналних p променљивих, чија је коваријансна матрица S , може бити генерисано са приближно k променљивих. Ако узмемо у обзир **Проблем 2**, ове идеје ће постати јасније. Оригинална матрица X димензије $N \times p$ може се записати

$$X = VTU',$$

где је $\mathbf{V}$ димензије $N \times p$, са ортонормираним колонама, $\mathbf{\Gamma}$ је дијагонална матрица димензије $p \times p$, и $\mathbf{U}$ је квадратна ортонормирана матрица димензије $p \times p$. Апроксимација снижавања ранга, $\mathbf{W}$, може бити записана

$$\mathbf{W} = \mathbf{V}_k \mathbf{\Gamma}_k \mathbf{U}_k',$$

где се $\mathbf{V}_k$ састоји од првих k колона матрице $\mathbf{V}$, $\mathbf{\Gamma}_k$ се састоји од првих k врста и колона матрице $\mathbf{\Gamma}$, и $\mathbf{U}_k$ се састоји од првих k колона матрице $\mathbf{U}$. Пошто $\mathbf{W} \approx \mathbf{X}$, важи следеће

$$\mathbf{X} \approx \mathbf{V}_k \mathbf{\Gamma}_k \mathbf{U}_k' \quad (2.14)$$

Множењем здесна ове матрице са $\mathbf{U}_k \mathbf{\Gamma}_k^{-1}$ добијамо

$$\mathbf{V}_k = \mathbf{X} \mathbf{U}_k \mathbf{\Gamma}_k^{-1}.$$

Матрица $\mathbf{U}_k \mathbf{\Gamma}_k^{-1}$ димензије $p \times k$ представља трансформацију врста матрице $\mathbf{X}$ из Еуклидовог p -простора у Еуклидов k -простор и (2.14) нам показује да имамо трансформацију из $N \times p$ матрице $\mathbf{X}$ у $N \times k$ матрице $\mathbf{V}_k$. Матрица $\mathbf{X}$ представља N тачака у Еуклидовом p -простору тако да може бити приближно уграђен у Еуклидов k -простор. Матрица $\mathbf{V}_k$ представља координате N тачака Еуклидовог k -простора.

3. ЦИЉЕВИ, ИДЕЈЕ И МОДЕЛИ ФАКТОРСKE АНАЛИЗЕ

У овом поглављу, представимо основне идеје и концепте **R-mode** факторске анализе, односно, проблеме обухваћене анализом односа променљивих. Усредсредимо се на технике факторске анализе.

3.1. Развој модела

Анализа матричног производа мање димензије $X'X$ се односи на **R-mode** анализу, док се анализа матричног производа веће димензије односи на XX' односи на **Q-mode** анализу. **R-mode** методе факторске анализе су засноване на математичким моделима који представљају поједностављене, али никад прецизне, приказе података, на основу одређених претпоставки. Разлике међу техникама факторске анализе су у великој мери повезане са разликама у овим претпоставкама. Представимо модел на два начина:

- Модел за варијансе и коваријансе
- Модел за регистроване променљиве

Користећи ова два модела, можемо развити различите видове метода, али неопходно је да схватимо да је модел исти за оба начина представљања.

3.1.1. Модел за регистроване променљиве

Приказаћемо факторски модел који је повезан са матрицом података. Математички модел гласи

$$X_{(N \times p)} = F_{(N \times k)} A'_{(k \times p)} + E_{(N \times p)},$$

X је матрица података, F матрица фактора, A матрица коефицијената и E матрица резидуала. Овде је k скалар који се односи на број коришћених фактора; скоро увек је мање од p , броја променљивих, а често је и много мање.

Било који елемент x_{ni} , матрице података можемо израчунати следећим моделом у скаларној нотацији

$$x_{ni} = \sum_{j=1}^k f_{nj} a_{ij} + e_{ni}.$$

Било коју врсту x' , матрице X добијамо моделом

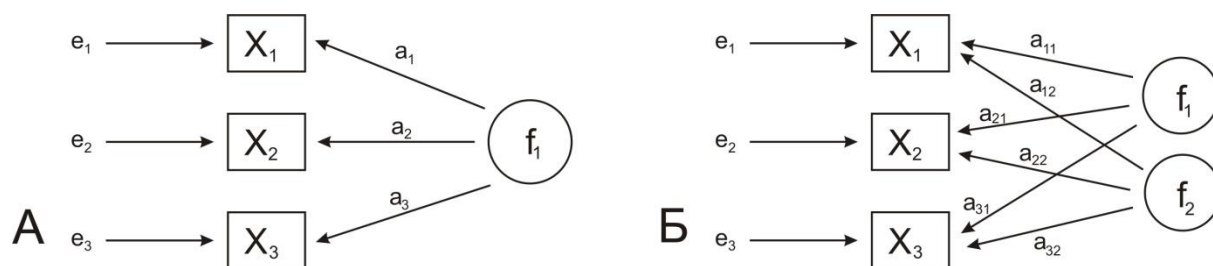
$$x' = f' A' + e' \quad (3.1)$$

где су f' и e' одговарајући вектори врсте матрица F и E . Због једноставности, узимамо транспоновану (3.1) да бисмо добили

$$x = A f + e; \quad (3.2)$$

x је сада вектор колона који представља један од објеката матрице података. Једначина (3.2) је једначина основног модела за све облике **R-mode** факторске анализе. Она нам казује да је свака регистрована променљива (елемент вектора x) пондерисана сума фактора плус резидуал. Производ, $A f$ је вектор оцено за вектор x , док вектор e

представља разлику између оценое и регистрованог вектора. Претпоставка је да су резидуали некорелисани са факторима.



Слика 3.1: Дијаграмски прикази факторских модела

Вектор колона $\mathbf{f}$ садржи факторе, елементи матрице $\mathbf{A}$ су коефицијенти који се множе са сваким фактором одређене променљиве.

На Слици 3.1(А) је представљен једноставан пример са три променљиве, које су на дијаграму приказане као квадрати. Претпоставимо да један фактор помножен коефицијентима a_1, a_2, a_3 , може дати приближне вредности за x_1, x_2, x_3 , респективно. e_1, e_2 , и e_3 представљају разлику између апроксимираних и регистрованих вредности. Модел једначина за овај пример гласи

$$x_1 = a_1 f_1 + e_1,$$

$$x_2 = a_2 f_1 + e_2,$$

$$x_3 = a_3 f_1 + e_3.$$

Уколико проценимо да апроксимација није довољно добра, можемо поставити двофакторски модел као што је приказано на Слици 3.1(Б). Овде је свака променљива повезана са оба фактора преко одговарајућих коефицијената:

$$x_1 = a_{11} f_1 + a_{12} f_2 + e_1,$$

$$x_2 = a_{21} f_1 + a_{22} f_2 + e_2,$$

$$x_3 = a_{31} f_1 + a_{32} f_2 + e_3. \quad (3.3)$$

Значај ових модела једначина није одмах уочљив. Због (3.3) и Сlike 3.1, сетимо се да је било који вектор колона матрице $\mathbf{X}$ променљива над којом бележимо N вредности; вектор врста матрице $\mathbf{X}$ представља објекат над којим је мерено p променљивих. Вектор колона матрице фактора $\mathbf{F}$ је фактор. Такође није одмах уочљиво, N елемената овог вектора представља варирајући број фактора у сваком објекту. Тако да сваки фактор представља неку неприметну особину објеката, а $\mathbf{F}$ представља целину таквих особина за објекте. Чињеница да постоји само k таквих особина ($k < p$) нам казује да оригинална матрица података садржи елементе који су сувишни.

Вектор $\mathbf{f}$ из (3.2) је типична вектор врста матрице $\mathbf{F}$. Вектор од k -променљивих које описују састав одређеног објекта помоћу фактора, односно, вредности појениначног објекта преко фактора.

Матрица коефицијената $\mathbf{A}$, садржи коефицијенте који се морају користити за множење фактора у оцени одређене променљиве. Вектор колона матрице $\mathbf{A}$ може се посматрати као садржалац коефицијената који описују факторе. Ово становиште нас доводи до схватања да је фактор линеарна комбинација регистрованих променљивих, а идеја се користи у анализи главних компоненти.

За општи случај са p променљивих и k фактора, (3.3) се може уопштити:

$$x_i = a_{i1}f_1 + a_{i2}f_2 + \dots + a_{ik}f_k + e_i, \quad (3.4)$$

где је x_i i -та променљива вектора $\mathbf{x}$. Имаћемо p таквих једначина, по једну за сваку променљиву. Израз (3.4) представља регресију x_i за факторе $f_1, f_2, \dots, f_k$ са резидуалом e_i . Разлика између ове и уобичајне регресије приказане у одељку 2.2 је та што фактори нису директно посматрани. На пример, линеарна комбинација

$$c_i = a_{i1}f_1 + a_{i2}f_2 + \dots + a_{ik}f_k \quad (3.5)$$

је најбоља линеарна оцена за x_i и зовемо је **заједнички део** x_i због тога што овај део променљиве има нешто заједничко са другим променљивама као резултат њихове повезаности са факторима. Резидуал e_i се обично израчунава као сума два некорелисана дела s_i и ϵ_i :

$$e_i = s_i + \epsilon_i, \quad (3.6)$$

где је ϵ_i **грешка мерења**, а s_i се односи на **специфичан део** x_i . Под грешком мерења подразумевамо грешку која се добија када се исти објекат мери више пута; уобичајено је запажање да има разлика у резултатима мерења.

3.1.2. Модел за варијансе и коваријансе

Иако је могуће развити решење за факторски модел директно из матрице података, погодније је приступити овој теми узимајући у обзир варијансе и коваријансе променљивих. У наставку, користимо (3.4), (3.5) и (3.6) да бисмо добили

$$x_i = c_i + e_i = c_i + s_i + \epsilon_i. \quad (3.7)$$

Наравно, могуће је наћи укупну варијансу за променљиву x_i . Једна од претпоставки факторског модела јесте да су сви чланови десне стране једначине (3.7) некорелисани. Због ове претпоставке и познате особине адитивности варијанси, могуће је разложити укупну варијансу променљиве x_i на следећи начин:

$$\sigma_{x_i}^2 = \sigma_{c_i}^2 + \sigma_{e_i}^2 = \sigma_{c_i}^2 + \sigma_{s_i}^2 + \sigma_{\epsilon_i}^2,$$

где:

$\sigma_{c_i}^2$ означава заједничку варијансу или **комуналитет**; представља онај део варијансе од x_i који је заједнички за све променљиве;

$\sigma_{e_i}^2$ је **резидуална варијанса**, често се назива **јединственост**;

$\sigma_{s_i}^2$ је **специфична варијанса** или **специфичност** од x_i ; означава варијансу специфичну за променљиву;

$\sigma_{\epsilon_i}^2$ је варијанса грешке од x_i ;

Компоненте варијансе приказане у (3.7) и компоненте променљиве из (3.4) можемо учинити још разумљивијим помоћу једноставног примера.

Размотримо истраживање неколико магматских стена у којима су одређене количине три микроелемента појединачног минерала и чврстоћа главног примерка стене. Микроелементи могу бити веома повезани услед њихове наводне узајамне зависности од хемијских и физичких услова за време формирања минерала. У факторском моделу, били би приказани велики заједнички делови и низак степен специфичности. Чврстоћа стене би вероватно имала мало тога заједничког са микроелементима и према томе садржала би висок степен специфичности и мало заједничких делова када би се све четири променљиве разматрале истовремено. Ако се овоме дода пета променљива уско повезана са чврстоћом, приметили бисмо да би ово смањило специфичност за чврстоћу, а повећало заједнички део.

У практичној примени, обично је немогуће знати пре анализе услове различитих варијанси. Они морају на неки начин бити процењени, обично уз одвијање анализе. На пример, заједнички делови се могу одредити из (3.5) на следећи начин.

Како је

$$c_i = a_{i1}f_1 + a_{i2}f_2 + \dots + a_{ik}f_k,$$

имамо

$$\begin{aligned}\sigma_{c_i}^2 &= \sum_{n=1}^N \frac{c_{ni}^2}{N} = \sum_{n=1}^N \frac{(a_{i1}f_{n1} + a_{i2}f_{n2} + \dots + a_{ik}f_{nk})^2}{N} \\ &= a_{i1}^2 \frac{\sum f_{n1}^2}{N} + a_{i2}^2 \frac{\sum f_{n2}^2}{N} + \dots + a_{ik}^2 \frac{\sum f_{nk}^2}{N} \\ &\quad + a_{i1}a_{i2} \frac{\sum f_{n1}f_{n2}}{N} + a_{i1}a_{i3} \frac{\sum f_{n1}f_{n3}}{N} \\ &\quad + \dots + a_{ik-1}a_{ik} \frac{\sum f_{nk-1}f_{nk}}{N}.\end{aligned}\tag{3.8}$$

Услови $\sum f_{nj}^2/N$ су варијансе фактора, за које можемо претпоставити да су јединствени, односно, да ће фактори бити стандардизовани. Штавише, услови $\sum f_{nj}f_{ne}/N$ су корелације између фактора. Због тога (3.8) постаје

$$\sigma_{c_i}^2 = a_{i1}^2 + a_{i2}^2 + \dots + a_{ik}^2 + a_{i1}a_{i2}\phi_{12} + \dots + a_{ik-1}a_{ik}\phi_{k-1k},$$

где су ϕ -ови корелације између фактора. Ако су фактори некорелисани, имамо следеће

$$\sigma_{c_i}^2 = a_{i1}^2 + a_{i2}^2 + \dots + a_{ik}^2.$$

У овом случају, заједнички делови променљиве x_i су суме квадратних елемената i -те врсте матрицекоефицијената A .

Други начин за приближавање факторског модела је разматрање коваријансне матрице. Ово ће нам дозволити да развијемо однос између матрице података и популације. Да бисмо поједноставили развој, претпоставићемо да је матрица података у девијацијском облику (Одељак 2.2).

Као што је раније речено, факторски модел је

$$Y = FA' + E,$$

односно,

$$Y' = AF' + E'.$$

Кад су у питању варијансе и коваријансе:

$$\frac{1}{N}Y'Y = A \left[\frac{1}{N}F'F \right] A' + A \left[\frac{1}{N}F'E \right] + \left[\frac{1}{N}E'F \right] A' + \frac{1}{N}E'E. \quad (3.9)$$

Матрица Y се може сматрати узорком (састоји се из N вектора врста) популације од p променљивих. Обим узорка се може повећати убацивањем вектора врста (објеката), што нас, у општем случају, може довести до тога да варијансе и коваријансе буду ближе вредности популације него оне из узорка мањег обима. Дакле, ако дозволимо да обим узорка N иде у бесконачност, сваки члан из (3.9) ће конвергирати ка својој вредности популације. Отуда,

$$\frac{1}{N}Y'Y \rightarrow \Sigma, \quad \frac{1}{N}F'F \rightarrow \Phi, \quad \frac{1}{N}F'E \rightarrow 0, \quad \frac{1}{N}E'E \rightarrow \Psi,$$

и имамо

$$\Sigma = A\Phi A' + \Psi. \quad (3.10)$$

У овој формули, Σ је коваријансна матрица популације димензије $(p \times p)$ регистрованих променљивих, A је матрица коефицијената димензије $(p \times k)$, Φ је коваријансна матрица за факторе димензије $(k \times k)$ (ако су фактори стандардизовани, ово је корелацијска матрица са јединицама на дијагонали), а Ψ је коваријансна матрица за резидуале димензије $(p \times p)$.

Једначина (3.10) представља, у матричној нотацији, комплетан факторски модел за варијансе и коваријансе регистрованих променљивих.

3.2. Пример факторске анализе

Пре него што пређемо на технике решавања факторског модела једначина, користиће нам следећи једноставан пример.

Претпоставимо да су променљиве $x_1, x_2, \dots, x_6$, мерене на великој популацији организама и њихових околина. Нека су променљиве x_1, x_2 и x_6 различите мере морфолошких промена неких организама и нека су x_3, x_4 и x_5 мере одређене врсте услова околине за које се претпоставља да имају утицај на морфологију ових организама. Због једноставности, променљиве су стандардизоване. Корелације између променљивих су следеће:

$$\Sigma = \begin{array}{c|cccccc} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ \hline x_1 & 1 & 0.72 & 0.378 & 0.324 & 0.27 & 0.27 \\ x_2 & 0.72 & 1 & 0.336 & 0.288 & 0.24 & 0.24 \\ x_3 & 0.378 & 0.336 & 1 & 0.42 & 0.35 & 0.126 \\ x_4 & 0.324 & 0.288 & 0.42 & 1 & 0.3 & 0.108 \\ x_5 & 0.27 & 0.24 & 0.35 & 0.3 & 1 & 0.09 \\ x_6 & 0.27 & 0.24 & 0.126 & 0.108 & 0.09 & 1 \end{array}$$

Уколико радимо са великим узорком, елемент ове матрице означавамо са ρ_{ij} и он представља вредност популације. Ово јесте корелацијска матрица, иако би по правилу требала да се означава великим словом ρ , док се ознаком Σ означава коваријансна матрица.

Прво одређујемо факторе корелацијске матрице помоћу (3.10) и користимо претпоставке факторског модела. Факторска анализа покушава да објасни ове корелације увођењем фактора $f_1, f_2, \dots, f_k$. Један од начина да се објасни овај поступак је постављање следећег питања. Постоји ли фактор f_1 , такав да ако би био издвојен, нестане повезаност између променљивих? Ако постоји, парцијалне корелације (Одељак 2.3) између било ког пара променљивих, x_i и x_j , морају нестати након елиминисања f_1 ; другим речима, имаћемо

$$x_i = a_i f_1 + e_i \quad x_j = a_j f_1 + e_j,$$

где су e_i и e_j некорелисане за $i \neq j$. Корелација ρ_{ij} би тада била

$$\rho_{ij} = a_i a_j, \quad i \neq j,$$

тако да би корелације у врстама i и j биле пропорционалне. Како то не важи за матрицу Σ , следи да није довољан само један фактор да објасни све корелације у Σ , па је потребно увести још фактора.

Претпоставимо сада, да су фактори некорелисани, односно, $\Phi = I$, па (3.10) можемо записати на следећи начин

$$\Sigma = AA' + \Psi.$$

Према томе, производ AA' би требао да репродукује вандијагоналне елементе корелацијске матрице Σ . У проширеном облику,

$$\rho_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2} + \dots + a_{ik}a_{jk}. \quad (3.11)$$

Израчунавањем матрице A , заједнички делови за променљиву x_i су дати са

$$\sigma_{c_i}^2 = a_{i1}^2 + a_{i2}^2 + \dots + a_{ik}^2.$$

Коришћењем метода које ћемо касније приказати, израчунавамо матрицу A која задовољава (3.11):

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

На пример, покажимо да важи следеће:

$$\begin{aligned} \rho_{12} &= a_{11}a_{21} + a_{12}a_{22} = 0.889 \times 0.791 + (-0.138) \times (-0.122) \\ &= 0.7032 + 0.0168 = 0.7200. \end{aligned}$$

Такође,

$$\sigma_{c_1}^2 = 0.889^2 + (-0.138)^2 = 0.81,$$

је заједнички део за прву променљиву.

Тако долазимо до чињенице да, за ове податке, постоје два некорелисана фактора f_1 и f_2 , који објашњавају шест променљивих

$$\begin{aligned} x_1 &= 0.889 f_1 - 0.138 f_2 + e_1, \\ x_2 &= 0.791 f_1 - 0.122 f_2 + e_2, \\ x_3 &= 0.501 f_1 + 0.489 f_2 + e_3, \\ x_4 &= 0.429 f_1 + 0.419 f_2 + e_4, \\ x_5 &= 0.358 f_1 + 0.349 f_2 + e_5, \\ x_6 &= 0.296 f_1 - 0.046 f_2 + e_6 \end{aligned} \quad (3.12)$$

и важи $\rho(e_i, e_j) = 0$, за свако i и j .

Сличност између једначине (3.12) и вишеструке регресије (Одељак 2.2) је приметна.

Фактори f_1 и f_2 нису једина два фактора која задовољавају једнакост $\Sigma = AA' + \Psi$. Следећа два фактора, такође, задовољавају једнакост:

$$f_1^* = 0.988 f_1 - 0.153 f_2 \text{ и } f_2^* = -0.153 f_1 + 0.988 f_2,$$

који постају

$$f_1 = 0.988 f_1^* - 0.153 f_2^* \text{ и } f_2 = -0.153 f_1^* + 0.988 f_2^* \quad (3.13)$$

У суштини, да бисмо израчунали f_1^* и f_2^* из (3.13), користимо ортонормирану транспоновану матрицу T :

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

Применом једнакости $A^* = AT^{-1}$ добијамо нову матрицу факторских коефицијената A^* која је једнака

$$\mathbf{A}^* = \begin{bmatrix} 0.9 & 0 \\ 0.8 & 0 \\ 0.42 & 0.56 \\ 0.36 & 0.48 \\ 0.3 & 0.4 \\ 0.3 & 0 \end{bmatrix}.$$

Заменом (3.13) у (3.12), добијамо следеће једначине које представљају оригиналне променљиве:

$$x_1 = 0.90 f_1^* + e_1,$$

$$x_2 = 0.80 f_1^* + e_2,$$

$$x_3 = 0.42 f_1^* + 0.56 f_2^* + e_3,$$

$$x_4 = 0.36 f_1^* + 0.48 f_2^* + e_4,$$

$$x_5 = 0.30 f_1^* + 0.40 f_2^* + e_5,$$

$$x_6 = 0.30 f_1^* + e_6.$$

Очигледно је да важи

$$\rho_{12} = 0.9 \times 0.8 + 0 \times 0 = 0.72$$

$$\sigma_{c_i}^2 = 0.9^2 + 0^2 = 0.81$$

Иако матрице $\mathbf{A}$ и $\mathbf{A}^*$ једнако добро задовољавају факторски модел, $\mathbf{A}^*$ је једноставнија због тога што су јој неки елементи нуле. Факторски образац матрице $\mathbf{A}^*$ се назива **једноставна структура**, управо због тога што су јој многи елементи које садржи једнаки нули. Променљиве ћемо груписати у два скупа.

Генерално, најједноставнију структуру постижемо ако допустимо да фактори постану корелисани, што се постиже коришћењем косих уместо ортогоналних фактора. На пример, два корелисана фактора

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

са $\rho(f_1^{**}, f_2^{**}) = 1 \times 0.6 + 0 \times 0.8 = 0.6$, односно,

$$\mathbf{\Phi} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}.$$

Трансформација из $\mathbf{f}^*$ у $\mathbf{f}^{**}$ је дата матрицом

$$\mathbf{T} = \begin{bmatrix} 1 & 0 \\ 0.6 & 0.8 \end{bmatrix}.$$

Инверзна матрица матрице $\mathbf{T}$ је

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

Користећи $\mathbf{T}^{-1}$, рачунамо факторски образац $\mathbf{A}^{**}$ који одговара $\mathbf{f}^{**}$ помоћу $\mathbf{A}^{**} = \mathbf{A}^* \mathbf{T}^{-1}$ и добијамо

$$A^{**} = \begin{bmatrix} 0.9 & 0 \\ 0.8 & 0 \\ 0 & 0.7 \\ 0 & 0.6 \\ 0 & 0.5 \\ 0.3 & 0 \end{bmatrix}$$

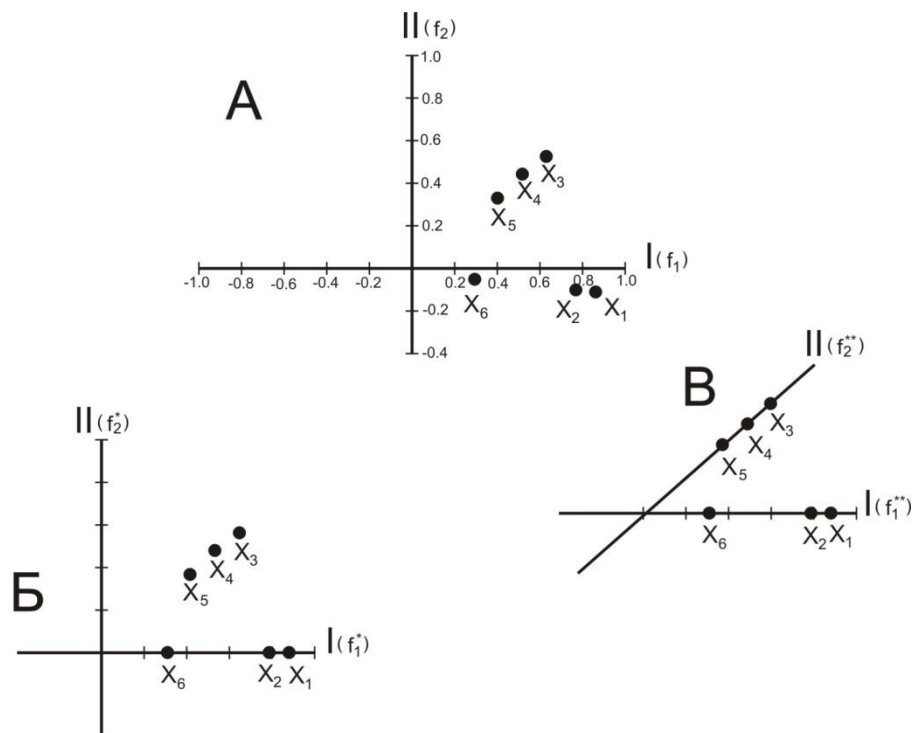
Одговарајућа матрица факторске структуре $A\Phi$ је

$$A\Phi = \begin{bmatrix} 0.9 & 0.54 \\ 0.8 & 0.48 \\ 0.42 & 0.7 \\ 0.36 & 0.6 \\ 0.3 & 0.5 \\ 0.3 & 0.18 \end{bmatrix}$$

Факторски образац матрице A^{**} илуструје груписање променљивих у два скупа. Први фактор комбинује само морфолошке мере, док други фактор комбинује само контроле животне средине.

3.3. Приказ резултата

Користећи геометријску нотацију вектора и матричних операција, можемо илустровати резултате факторске анализе претходног примера.



Слика 3.2: Графичка анализа резултата факторске анализе. (А) Графички приказ неротираних резултата. (Б) Ефекти ротације у смеру казаљки на сату, при којој остаје очувана ортогоналност оса. (В) Ефекат ротирања друге осе.

За **R-mode** модел, прихватили смо векторски приказ корелације. Овде смо, променљиве као и векторе посматрали као смештене у N -димензионалном простору

објеката са њиховим угаоним раздвајањима који су одређени корелацијским коефицијентима. Фактори обезбеђују основу за ове векторе. Резултати претходног примера показују да је димензионалност једнака два.

У ортогоналном случају који је дат са $\mathbf{A}$ и $\mathbf{A}^*$, факторске осе могу бити представљене као два ортогонална вектора. Елемент матрице a_{ij} даје корелацију а тиме и угао између вектора променљивих i и j -те осе фактора. График променљивих у вези са факторским осама је погодан начин за илустрацију односа између променљивих.

Слика 3.2(А) приказује график матрице $\mathbf{A}$. Слика 3.2(Б) приказује график матрице $\mathbf{A}^*$. На Слици 3.2(А) осе представљају систем (f_1, f_2) . Променљиве имају умерено велике пројекције на оба фактора. Слика 3.2(Б) приказује ефекте ротације при којо остаје очувана ортогоналност оса система (f_1, f_2) у смеру кретања казаљки на сату ка систему (f_1^*, f_2^*) ; односно, угао од 90° између оса, остаје сачуван. Променљиве 1, 2 и 6 леже на првој оси (фактор 1) и имају коефицијенте нула у односу на фактор 2 (другу осу). Променљиве 3, 4 и 5 имају не-нула коефицијенте за оба фактора.

Трансформација из (f_1^*, f_2^*) у (f_1^{**}, f_2^{**}) одговара ротацији друге осе тако да прође кроз променљиве 3, 4 и 5, али задржавајући прву осу у истом положају као што је приказано на Слици 3.2(Б). На Слици 3.2(В) угао између оса одговара косинусу од 0.6, корелација између косих фактора.

4. R-MODE METODE

У овом одељку, описаћемо неколико метода које изучавају међуодносе променљивих, односно, **R-mode** методе. Размотрићемо циљеве, идеје, моделе и појмове факторске анализе.

У пракси, подаци које имамо на располагању састоје се од узорака узетих из популације. Врсте у матрици података $\mathbf{X}$ димензије $N \times p$, представљају случајан узорак узет из популације. Као што смо већ показали у претходним одељцима, из $\mathbf{X}$ можемо израчунати коваријансну матрицу узорка $\mathbf{S}$. Сада статистички проблем постаје, како проценити популацијски модел, из података матрице $\mathbf{X}$ или $\mathbf{S}$? Чак и у овом случају потребно је фитовати модел, пошто ће $\mathbf{X}$ или $\mathbf{S}$ бити апроксимирани матрицама мањег ранга. У овом случају смо заинтересовани за процену оба и матрице факторских коефицијената $\mathbf{A}$ и матрице фактора $\mathbf{F}$. Сада ћемо разматрати анализу главних компоненти и факторску анализу засебно. У даљем тексту, "капа" ($\wedge$) изнад симбола популације означава да је то изведена процена узорка.

4.1. Анализа главних компоненти

Почећемо тако што ћемо поново узети у обзир модел за матрицу података $\mathbf{Y}$ у девијацијском облику:

$$\mathbf{Y} = \mathbf{F}\mathbf{A}' + \mathbf{E}. \quad (4.1)$$

Због једноставности претпосавићемо да су сви елементи матрице $\mathbf{Y}$ подељени са $\sqrt{N}$, тако да је коваријансна матрица $\mathbf{S}$ једноставно дата са

$$\mathbf{S} = \mathbf{Y}'\mathbf{Y},$$

матрични производ мање димензије од $\mathbf{Y}$. Овде је, такође, $\mathbf{F}$ матрица фактора, $\mathbf{A}$ је матрица факторских коефицијената и $\mathbf{E}$ је матрица резидуала.

Можемо фитовати овај модел применом методе најмањих квадрата на матрицу података $\mathbf{Y}$. Ово значи да морамо одредити матрице $\mathbf{F}_{(N \times k)}$ и $\mathbf{A}_{(p \times k)}$, за дато $k < p$, такве да је сума квадрата свих елемената матрице

$$\mathbf{E} = \mathbf{Y} - \mathbf{F}\mathbf{A}'$$

што мања могућа. Као што је већ речено у Одељку 2.6, решење овог проблема је добијено као

$$\widehat{\mathbf{F}\mathbf{A}'} = \gamma_1 \mathbf{v}_1 \mathbf{u}_1' + \gamma_2 \mathbf{v}_2 \mathbf{u}_2' + \dots + \gamma_k \mathbf{v}_k \mathbf{u}_k',$$

k чланова сингуларних вредности декомпозиције матрице $\mathbf{Y}$ које одговарају k највећих сингуларних вредности $\gamma_1, \gamma_2, \dots, \gamma_k$. Нека је

$$\mathbf{V}_k = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k], \quad \mathbf{U}_k = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k] \quad \text{и}$$

$$\boldsymbol{\gamma}_k = \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_k).$$

Тада је решење

$$\widehat{\mathbf{F}\mathbf{A}'} = \mathbf{V}_k \boldsymbol{\Gamma}_k \mathbf{U}_k'.$$

Иако нам ово говори шта је производ матрица $\mathbf{FA}'$, не даје нам јединствено решење за $\mathbf{F}$ и $\mathbf{A}$. Јасно је да има бесконачно много решења за $\mathbf{F}$ и $\mathbf{A}$, пошто можемо заменити $\mathbf{F}$ са $\mathbf{F}^* = \mathbf{FT}$ и $\mathbf{A}$ са $\mathbf{A}^* = \mathbf{AT}^{-1}$, где нам је $\mathbf{T}$ несингуларна матрица.

Сада ћемо размотрити два различита избора решења за $\mathbf{A}$ и $\mathbf{F}$ и продискутовати предности и мане за сваки од њих. Треба нагласити да су ова два решења суштински иста.

Решење 1

$$\hat{\mathbf{F}} = \mathbf{V}_k, \quad \hat{\mathbf{A}} = \mathbf{U}_k \mathbf{\Gamma}_k.$$

Лако се показује да је последица чињенице да $\mathbf{Y}$ има суме колона једнаке нули, и суме колона матрице $\mathbf{V}_k$ ће такође бити нула. Због тога, $\mathbf{F}$ ће бити у девијацијском облику и коваријансна матрица фактора (у узорку) ће бити

$$\hat{\mathbf{F}}' \hat{\mathbf{F}} = \mathbf{V}_k' \mathbf{V}_k = \mathbf{I},$$

односно, фактори су некорелисани и стандардизовани. За матрицу факторских коефицијената имамо

$$\hat{\mathbf{A}}' \hat{\mathbf{A}} = \mathbf{\Gamma}_k \mathbf{U}_k' \mathbf{U}_k \mathbf{\Gamma}_k = \mathbf{\Gamma}_k^2 = \mathbf{\Lambda}_k,$$

где је $\mathbf{\Lambda}_k$ дијагонална матрица димензије $k \times k$ чији су дијагонални елементи $\lambda_1 = \gamma_1^2$, $\lambda_2 = \gamma_2^2$, ..., $\lambda_k = \gamma_k^2$, карактеристичне вредности матрице $\mathbf{S}$, коваријансна матрица. Даље, ако је $\mathbf{E}$ из (4.1) мало тако да имамо

$$\mathbf{Y} \approx \hat{\mathbf{F}} \hat{\mathbf{A}}',$$

коваријансна матрица $\mathbf{S}$ је приближно

$$\mathbf{S} = \mathbf{Y}' \mathbf{Y} \approx \hat{\mathbf{A}} \hat{\mathbf{F}}' \hat{\mathbf{F}} \hat{\mathbf{A}}' = \hat{\mathbf{A}} \hat{\mathbf{A}}',$$

односно, $\hat{\mathbf{A}} \hat{\mathbf{A}}'$ апроксимира $\mathbf{S}$.

Да бисмо израчунали ово решење, претпоставимо да $N \geq p$:

1. Израчунамо коваријансну матрицу $\mathbf{S}$.
2. Израчунамо k највећих карактеристичних вредности $\mathbf{\Lambda}_k$ и њима одговарајуће карактеристичне векторе $\mathbf{U}_k$ од $\mathbf{S}$.
3. Израчунамо $\hat{\mathbf{A}} = \mathbf{U}_k \mathbf{\Lambda}_k^{1/2}$. Скалирамо сваки карактеристични вектор тако да његова квадратна дужина буде једнака карактеристичној вредности.
4. Израчунамо $\hat{\mathbf{F}} = \mathbf{Y} \hat{\mathbf{A}} \mathbf{\Lambda}_k^{-1}$. Ово значи да прво множимо здесна матрицу $\mathbf{Y}$ са $\hat{\mathbf{A}}$ а онда скалирамо колоне реципрочним карактеристичним вредностима.

Решење 2

$$\hat{\mathbf{A}} = \mathbf{U}_k, \quad \hat{\mathbf{F}} = \mathbf{V}_k \mathbf{\Gamma}_k.$$

Овим избором, добијамо коваријансну матрицу фактора

$$\hat{\mathbf{F}}' \hat{\mathbf{F}} = \mathbf{\Gamma}_k \mathbf{V}_k' \mathbf{V}_k \mathbf{\Gamma}_k = \mathbf{\Gamma}_k^2 = \mathbf{\Lambda}_k,$$

која је дијагонална, али са дијагоналним елементима једнаким карактеристичним вредностима матрице $\mathbf{S}$ по опадајућем реду величина. Дакле, фактори су и даље некорелисани, али имају различите варијансе. За матрицу факторских коефицијената $\mathbf{A}$ имамо

$$\widehat{\mathbf{A}}\widehat{\mathbf{A}} = \mathbf{U}'_k \mathbf{U}_k = \mathbf{I},$$

односно, колоне матрице $\widehat{\mathbf{A}}$ су ортонормиране. Даље, у овом случају имамо

$$\mathbf{S} = \mathbf{Y}'\mathbf{Y} \approx \widehat{\mathbf{A}}\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\widehat{\mathbf{A}} = \widehat{\mathbf{A}}\mathbf{\Lambda}_k\widehat{\mathbf{A}}'.$$

Рачунање овог решења следи, претпоставимо као и раније да $N \geq p$:

1. Израчунамо коваријансну матрицу $\mathbf{S}$.
2. Израчунамо k највећих карактеристичних вредности $\mathbf{\Lambda}_k$ и њима одговарајуће карактеристичне векторе $\mathbf{U}_k$ од $\mathbf{S}$.
3. Онда $\widehat{\mathbf{A}} = \mathbf{U}_k$.
4. Израчунамо $\widehat{\mathbf{F}} = \mathbf{Y}\widehat{\mathbf{A}}$.

Приметимо да у овом случају није потребно скалирање колона и због тога овај избор је у једноставнији од другог у погледу рачунања.

У Решењу 1, сви фактори су некорелисани и имају исти варијансу, тако да колоне матрице $\widehat{\mathbf{A}}$ могу бити директно поређене. У овом решењу, матрица $\widehat{\mathbf{A}}$ такође представља коваријансе између променљивих и фактора, односно, структуре фактора. Да бисмо добили корелације између променљивих и фактора, потребно је само поделити сваку врсту матрице $\widehat{\mathbf{A}}$ стандардном девијацијом променљивих. Дакле, матрица корелација између променљивих и фактора је

$$\widehat{\mathbf{C}} = \mathbf{D}_s^{-1}\widehat{\mathbf{A}} = \mathbf{D}_s^{-1}\mathbf{U}_k\mathbf{\Lambda}_k^{1/2},$$

где је $\mathbf{D}_s$ дијагонална матрица са дијагоналним елементима једнаким стандардној девијацији променљивих.

У Решењу 2, с друге стране, фактори су такође некорелисани, али имају различите варијансе, па колоне матрице $\widehat{\mathbf{A}}$ није могуће директно упоређивати. Пошто је коваријансна матрица фактора $\mathbf{\Lambda}_k$, коваријансе између променљивих и фактора су дате са $\widehat{\mathbf{A}}\mathbf{\Lambda}_k$ и одговарајуће корелације су дате следећом формулом

$$\widehat{\mathbf{C}} = \mathbf{D}_s^{-1}\widehat{\mathbf{A}}\mathbf{\Lambda}_k\mathbf{\Lambda}_k^{-1/2} = \mathbf{D}_s^{-1}\mathbf{U}_k\mathbf{\Lambda}_k^{1/2},$$

која је иста као и за Решење 1. Дакле, у Решењу 2, и врсте и колоне матрице $\widehat{\mathbf{A}}$ се морају скалирати да би се добиле корелације између променљивих и фактора.

Да бисмо илустровали Анализу главних компоненти, користимо низ вештачких података, који се састоји од 10 хипотетичких објеката над којим је било мерено 5 особина. Матрица података димензије 10×5 , је направљена тако да је приближно ранга 2. Ова матрица је приказана у Табели 4.1, и задовољава аритметичке средине и стандардне девијације сваке променљиве. Коваријансна матрица $\mathbf{S}$ за ових пет променљивих је дата у Табели 4.2. Запазимо да пошто је N мало овде, $\mathbf{S}$ рачунамо на следећи начин

$$\frac{1}{N-1}Y'Y.$$

Карактеристичне вредности матрице $\mathbf{S}$ су:

$$\lambda_1 = 114.83459, \lambda_2 = 17.72064, \lambda_3 = 0.00588, \lambda_4 = \lambda_5 = 0.00000.$$

Ако узмемо да је $k = 2$, карактеристични вектори који одговарају двома највећим карактеристичним вредностима су

$$\mathbf{U}_2 = \begin{bmatrix} 0.17 & 0.32 \\ 0.75 & 0.19 \\ 0.21 & 0.10 \\ -0.01 & 0.86 \\ -0.61 & 0.35 \end{bmatrix}.$$

Решења методом најмањих квадрата за матрице $\mathbf{A}$ и $\mathbf{F}$ следе.

Решење 1.

$$\hat{\mathbf{A}} = \begin{bmatrix} 1.86 & 1.34 \\ 8.01 & 0.81 \\ 2.25 & 0.40 \\ -0.13 & 3.60 \\ -6.50 & 1.46 \end{bmatrix}, \quad \hat{\mathbf{F}} = \begin{bmatrix} 0.71 & -1.73 \\ 1.00 & 1.61 \\ -1.88 & 0.40 \\ 0.88 & -0.04 \\ 0.19 & -1.31 \\ -0.85 & -0.45 \\ 0.65 & 0.82 \\ 0.28 & -0.30 \\ 0.36 & 0.70 \\ -1.34 & 0.31 \end{bmatrix}.$$

Овде је, $\hat{\mathbf{F}}'\hat{\mathbf{F}} = (N-1)\mathbf{I} = 9\mathbf{I}$.

Решење 2.

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.17 & 0.32 \\ 0.75 & 0.19 \\ 0.21 & 0.10 \\ -0.01 & 0.86 \\ -0.61 & 0.35 \end{bmatrix}, \quad \hat{\mathbf{F}} = \begin{bmatrix} 7.61 & -7.29 \\ 10.67 & 6.77 \\ -20.20 & 1.69 \\ 9.44 & -0.18 \\ 2.05 & -5.49 \\ -9.08 & -1.90 \\ 6.97 & 3.45 \\ 2.97 & -1.28 \\ 3.89 & 2.94 \\ -14.33 & 1.30 \end{bmatrix}.$$

Овде је, $\hat{\mathbf{F}}'\hat{\mathbf{F}} = 9\mathbf{A}_k$.

Табела 4.1 *Вештачки подаци*

Објекат	Променљива				
	x_1	x_2	x_3	x_4	x_5
1	5.0	25.0	15.0	5.0	5.0
2	10.0	30.0	17.0	17.0	8.0
3	3.0	6.0	10.0	13.0	25.0
4	7.5	27.2	16.0	11.0	6.5
5	4.6	21.9	14.0	6.6	9.0
6	3.8	13.6	12.0	9.8	17.0
7	8.3	26.6	15.9	14.2	9.1
8	6.1	22.7	14.6	10.2	9.9
9	7.6	24.2	15.2	13.8	10.8
10	3.9	10.3	11.2	12.6	21.3
Аритметичка средина	5.98	20.75	14.09	11.32	12.16
Стандардна девијација	2.29	8.04	2.29	3.61	6.66

(А) Матрица података X ;

Објекат	Променљива				
	y_1	y_2	y_3	y_4	y_5
1	-0.98	4.25	0.91	6.32	-7.16
2	4.02	9.25	2.91	5.68	-4.16
3	-2.98	-14.75	-4.09	1.68	12.84
4	1.52	7.15	1.91	-0.32	-5.66
5	-1.38	0.45	-0.09	-4.72	-3.16
6	-2.18	-7.15	-2.09	-1.52	4.84
7	2.32	5.85	1.81	2.88	-3.06
8	0.12	1.95	0.51	-1.12	-2.26
9	1.62	3.45	1.11	2.48	-1.36
10	-2.08	-10.45	-2.89	1.28	9.14

(Б) Матрица девијација Y ;

	x_1	x_2	x_3	x_4	x_5
x_1	5.2573				
x_2	15.9933	64.7517			
x_3	4.7387	18.3650	5.2410		
x_4	4.5849	1.9111	1.1769	13.0062	
x_5	-10.1498	-50.8122	-14.0438	6.0676	44.3049

(В) Коваријансна матрица S .

Табела 4.II Вештачки подаци

Објекат	Променљива				
	Z_1	Z_2	Z_3	Z_4	Z_5
1	-0.43	0.53	0.40	-1.75	-1.08
2	1.75	1.15	1.27	1.57	-0.62
3	-1.30	-1.83	-1.79	0.47	1.93
4	0.66	0.89	0.83	-0.09	-0.85
5	-0.60	0.06	-0.04	-1.31	-0.47
6	-0.95	-0.89	-0.91	-0.42	0.73
7	1.01	0.73	0.79	0.80	-0.46
8	0.05	0.24	0.22	-0.31	-0.34
9	0.71	0.43	0.48	0.69	-0.20
10	-0.91	-1.30	-1.26	0.35	1.37

(A) Матрица стандардизованих података $\mathbf{Z}$,

	X_1	X_2	X_3	X_4	X_5
X_1	1.0000				
X_2	0.8668	1.0000			
X_3	0.9028	0.9969	1.0000		
X_4	0.5545	0.0659	0.1426	1.0000	
X_5	-0.6650	-0.9487	-0.9216	0.2528	1.0000

(Б) Корелацијска матрица $\mathbf{R} = \frac{1}{9} \mathbf{Z}' \mathbf{Z}$.

За оба решења корелације између променљивих и фактора су

$$\hat{\mathbf{C}} = \begin{bmatrix} 0.81 & 0.58 \\ 0.99 & 0.10 \\ 0.98 & 0.18 \\ -0.04 & 1.00 \\ -0.98 & 0.22 \end{bmatrix}.$$

Из ове матрице видимо да је фактор 1 јако повезан са променљивама 1, 2 и 3, а негативно са променљивом 5, док је фактор 2 јако повезан са променљивом 4, а слабије са осталим променљивама.

Техника коју смо управо описали назива се **анализа главне компоненте** и у овом случају примењује се директно на матрицу података. Наша илустрација се базира на матрици девијације $\mathbf{Y}$, али техника може бити примењена и на матрицу стандардизованих података $\mathbf{Z}$ и као што ћемо видети, добићемо другачије резултате. Матрица $\mathbf{Z}$ је дата у Табели 4.II(A), а корелацијска матрица $\mathbf{R} = (1/N) \mathbf{Z}' \mathbf{Z}$ у Табели 4.II(B). карактеристичне вредности матрице $\mathbf{R}$ су $\lambda_1 = 3.68237$, $\lambda_2 = 1.31747$, $\lambda_3 = 0.00016$, $\lambda_4 = \lambda_5 = 0.00000$. Карактеристични вектори који одговарају двома највећим карактеристичним вредностима су

$$U_2 = \begin{bmatrix} 0.48 & 0.34 \\ 0.52 & -0.10 \\ 0.52 & -0.04 \\ 0.09 & 0.86 \\ -0.47 & 0.37 \end{bmatrix}.$$

Методом најмањих квадрата Решење 1 је дато са

$$\hat{A} = \begin{bmatrix} 0.92 & 0.39 \\ 0.99 & -0.12 \\ 1.00 & -0.04 \\ 0.18 & 0.98 \\ -0.90 & 0.42 \end{bmatrix}, \quad \hat{F} = \begin{bmatrix} 0.32 & -1.84 \\ 1.32 & 1.36 \\ -1.75 & 0.80 \\ 0.84 & -0.25 \\ -0.09 & -1.31 \\ -0.92 & -0.26 \\ 0.82 & 0.66 \\ 0.21 & -0.35 \\ 0.51 & 0.61 \\ -1.24 & 0.59 \end{bmatrix}.$$

Овде је $\hat{F}'\hat{F} = 9I$.

У овом случају, када су и променљиве и фактори стандардизовани, матрица $\hat{C}$ корелације између променљивих и фактора је идентична матрици $\hat{A}$. Иако се карактеристичне вредности вектори матрице R доста разликују од оних из матрице S , видећемо да је $\hat{C}$ матрица корелација између променљивих и фактора приближно иста у оба случаја. Чињеница да нису идентични нам казује да резултати анализе главних компоненти зависе од скале.

4.1.1. Број компоненти

До сада смо претпостављали да је k , број извучених компоненти, познат. Нормална је ситуација, међутим, када је k непознато, а желимо да га одредимо на такав начин да будемо „задовољни“ постотком укупне варијације, односно, расипања. Не постоји тачно решење овог проблема, али следеће смернице нам често могу бити од велике помоћи.

Почињемо тако што рачунамо процентуалну кумулативну варијансу, коју добијамо за узастопне вредности $k = 1, 2, \dots$. Престајемо када нам је проценат кумулативне варијансе довољно велик, на пример, већи од 75%, 90% или 95%. Колико далеко се може ићи, углавном зависи од природе података, односно, од степена колинеарности и редундантности у њима. Проценат кумулативне варијансе за дато k се рачуна на следећи начин:

$$100 = \frac{\sum_{i=1}^k \lambda_i}{\text{tr } S}.$$

Карактеристичне вредности, $\lambda_1, \lambda_2, \dots, \lambda_k$, треба да буду значајно различите и позитивне, што значи да би опадање у величини, од једног од њих, до другог било приметно. Приближни интервали поверења могу се рачунати да бисмо видели да ли су λ -е значајно различите од нуле.

Резидуалне карактеристичне вредности $\lambda_{k+1}, \dots, \lambda_p$, треба да буду мале и приближно једнаке.

4.2. Факторска анализа

У факторској анализи, једначина модела гласи:

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

где је $\mathbf{\Psi}$ дијагонална матрица и њени дијагонални елементи нису обавезно мали. Дијагонални елементи матрице $\mathbf{\Psi}$ су јединствене варијансе и морају бити процењене из података, заједно са матрицом факторских коефицијената $\mathbf{A}$. Критеријум за одређивање модела са подацима јесте минимизирање суме квадрата свих елемената $\mathbf{S} - \mathbf{A}\mathbf{A}' - \mathbf{\Psi}$, односно, минимизирање

$$\text{tr}(\mathbf{S} - \mathbf{A}\mathbf{A}' - \mathbf{\Psi})^2.$$

Ако би $\mathbf{\Psi}$ била позната, могли бисмо да израчунамо матрицу $\mathbf{A}\mathbf{A}'$ користећи метод анализе главних компоненти. У колоне матрице $\mathbf{A}$ бисмо стављали карактеристичне векторе матрице $\mathbf{S} - \mathbf{\Psi}$ који би одговарали првим k највећим карактеристичним вредностима. Сума квадрата у свакој колони једнака је одређеној карактеристичној вредности. У пракси, $\mathbf{\Psi}$ није познато, па мора бити процењено из података. Сам чин бирања $\mathbf{\Psi}$ је еквивалентан бирању комуналитета, пошто су дијагонални елементи матрице $\mathbf{S} - \mathbf{\Psi}$ оцене комуналитета. У прошлости је било испробано неколико метода за одређивање комуналитета. Најчешће коришћени метод је онај који бира комуналитет сваке променљиве у вези са квадратом вишеструког корелацијског коефицијента (*SMC*) те променљиве са свим осталим променљивама. Гутман је показао да *SMC* метода даје најбољу могућу доњу границу за комуналитет. Матрица $\hat{\mathbf{\Psi}}$ рачуна се као

$$\hat{\mathbf{\Psi}} = (\text{diag } \mathbf{S}^{-1})^{-1}, \quad (4.1)$$

односно, ψ -ови су једнаки реципрочним вредностима дијагоналних елемената инверза матрице $\mathbf{S}$. Са овим избором матрице $\mathbf{\Psi}$ може се наставити са оценом матрице $\mathbf{A}$ као што је претходно објашњено. Ову методу називамо **главна факторска метода** и даје нам оцене које су систематски пристрасне. Метода се може много побољшати применом итеративног поступка. Уколико би се проценила вредност $\hat{\mathbf{\Psi}}$, а тиме и $\hat{\mathbf{A}}$, модел би се могао побољшати новом вредношћу $\hat{\mathbf{\Psi}}$ на следећи начин

$$\hat{\mathbf{\Psi}} = \text{diag } (\mathbf{S} - \hat{\mathbf{A}}\hat{\mathbf{A}}').$$

Овај процес се понавља све док се не добије стабилна вредност матрице $\hat{\mathbf{A}}$. У литератури ово је познато као **итеративна принципална факторска метода**.

Ове методе имају само историјски значај и статистичари их више не користе. Много ефективнија метода факторске анализе се постиже минимизирањем функције

$$\text{tr}(\mathbf{S} - \mathbf{A}\mathbf{A}' - \mathbf{\Psi})^2.$$

Једна од таквих метода јесте *MINRES*, која минимизује суму квадрата свих вандијагоналних елемената матрице $\mathbf{S} - \mathbf{A}\mathbf{A}'$ и тиме се избегавају комуналитети. Друга метода је *ULS* метода, која путем елиминације матрице $\mathbf{A}$, смањује функцију на функцију матрице $\mathbf{\Psi}$ и онда се минимизира путем Њутн-Рафсонове процедуре.

4.2.1. Нескалирана оцена матрица A и Ψ

Један од проблема који се јавља код ULS методе је тај што нам даје скалирано решење. Нескалирани метод је такав да ако је S анализирано да бисмо добили A и Ψ , онда бисмо анализом DSD , где је матрица D дијагонална матрица фактора скале, требали да добијемо оцене DA и $D^2\hat{\Psi}$. Ово је пожељна особина у факторској анализи, уколико су јединице мере код променљиве произвољне. ULS метода није нескалирана и у пракси се обично користи корелациона матрица R уместо матрице S . Нескалирани метод оцене матрица A и Ψ може се развити на следећи начин. Претпоставимо да је Ψ познато. У овом случају, приближно важи

$$S - \Psi \approx AA'.$$

Множењем с лева и десна матрицом $\Psi^{-1/2}$ добијамо

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

Пошто нам је Ψ познато, можемо израчунати $\Psi^{-1/2}S\Psi^{-1/2} - I$ и тако добити вредност за $A^*A'^*$, где нам је $A^* = \Psi^{-1/2}A$. Када добијемо оцену $\hat{A}^*$ из карактеристичних вредности и карактеристичних вектора матрице $\Psi^{-1/2}S\Psi^{-1/2} - I$, онда рачунамо $\hat{A}$ као $\Psi^{1/2}\hat{A}^*$. Избором $\hat{\Psi}$ као у (4.1), рачунамо карактеристичне вредности и векторе матрице

$$(diag S^{-1})^{1/2}S(diag S^{-1})^{1/2} - I,$$

За коју се може показати да је инваријантна.

Пошто је оцена матрице Ψ из (4.1) систематски превелика, Јореског је предложио да је заменимо следећим

$$\hat{\Psi} = \theta(diag S^{-1})^{-1},$$

где је θ непознати скалар, мањи од 1, који оцењујемо из података. Нека су $\lambda_1, \lambda_2, \dots, \lambda_p$ карактеристичне вредности матрице

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

Оцена параметра θ методом најмањих квадрата је

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

односно, средњој вредности $p-k$ најмањих карактеристичних вредности матрице S^* . Штавише, нека је U_k матрица димензије $p \times k$, чије су колоне ортонормирани карактеристични вектори матрице S^* који одговарају k највећим карактеристичним вредностима који су смештени на дијагонали дијагоналне матрице Λ_k . Тада је оцена матрице A методом најмањих квадрата једнака

$$\hat{A} = (diag S^{-1})^{-1/2}U_k(\Lambda_k - \hat{\theta}I)^{1/2}. \quad (4.2)$$

Дакле, прво се скалирају колоне матрице $\mathbf{U}_k$ тако да сума квадрата j -те колоне буде једнака $\lambda_j - \hat{\theta}$, а затим се скалира i -та врста са $1/\sqrt{s^{ii}}$, где је s^{ii} i -ти дијагонални елемент матрице $\mathbf{S}^{-1}$.

4.2.2. Пример

Рачунање обухваћено овом методом ћемо илустровати примером који следи и користећи корелацијска матрицу $\mathbf{S}$

$$\mathbf{S} = \begin{bmatrix} 1.000 & 0.466 & 0.465 & 0.441 & 0.375 & 0.312 & 0.247 & 0.207 \\ 0.466 & 1.000 & 0.311 & 0.296 & 0.521 & 0.286 & 0.483 & 0.314 \\ 0.456 & 0.311 & 1.000 & 0.185 & 0.184 & 0.300 & 0.378 & 0.378 \\ 0.441 & 0.296 & 0.185 & 1.000 & 0.176 & 0.244 & 0.121 & 0.341 \\ 0.375 & 0.521 & 0.184 & 0.176 & 1.000 & 0.389 & 0.211 & 0.153 \\ 0.312 & 0.286 & 0.300 & 0.244 & 0.389 & 1.000 & 0.210 & 0.289 \\ 0.247 & 0.483 & 0.378 & 0.121 & 0.211 & 0.210 & 1.000 & 0.504 \\ 0.207 & 0.314 & 0.378 & 0.341 & 0.153 & 0.289 & 0.504 & 1.000 \end{bmatrix}.$$

Узимамо да нам је матрица $\mathbf{S}$ коваријансна матрица. Дијагонални елементи матрице $\mathbf{S}^{-1}$ и одређене вредности су приказане у табели која следи

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

Табела 4.ИИ: Елементи s^{ii} и потребне вредности

Елементи матрице $\mathbf{S}^*$ се рачунају из елемената матрице $\mathbf{S}$ и елемената $\sqrt{s^{ii}}$ из Табеле 4.ИИ помоћу формуле

$$s_{ij}^* = \sqrt{s^{ii}} s_{ij} \sqrt{s^{jj}}.$$

Матрица $\mathbf{S}^*$ је једнака

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

Ова матрица би била иста и да је почетна матрица била коваријансна матрица уместо корелацијске.

Карактеристичне вредности λ_i матрице $\mathbf{S}^*$ су 5.281, 1.809, 1.507, 1.199, 1.152, 0.703, 0.625, 0.534 и њихова сума је једнака $\text{tr } \mathbf{S}^* = 12.810$. Да бисмо одредили број фактора k , користимо правило које следи. Узимамо најмању карактеристичну вредност, у овом случају то је 0.534 који је за 0.466 мањи од један. Посматраћемо карактеристичне вредности чија је вредност већа од $1+0.466$, односно, од 1.466. Њих има укупно 3, дакле $k = 3$. Да бисмо израчунали матрицу факторских коефицијената, прво морамо израчунати оцену $\hat{\theta}$ од θ . По формули, она је једнака средњој вредности $p - k$ карактеристичних вредности, односно, аритметичкој средини пет најмањих карактеристичних вредности, што је

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

Затим се рачуна вредност матрице $\mathbf{A}_k - \hat{\theta}\mathbf{I}$, па добијамо

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

Квадратни корени ових вредности су

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

Даље рачунамо карактеристичне векторе матрице $\mathbf{S}^*$ који одговарају првим трима највећим карактеристичним вредностима:

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

Коначно, матрицу факторских коефицијената A рачунамо из (4.2)

$$\hat{A} = \begin{bmatrix} 0.65 & 0.27 & 0.22 \\ 0.71 & -0.11 & -0.25 \\ 0.57 & 0.16 & 0.17 \\ 0.47 & -0.12 & 0.38 \\ 0.55 & -0.35 & -0.28 \\ 0.50 & -0.11 & 0.03 \\ 0.59 & 0.40 & -0.20 \\ 0.56 & 0.42 & 0.11 \end{bmatrix}.$$

4.3. Анализа главних компоненти и крос-валидација (унакрсна провера ваљаности)

Крзановски је направио синтезу техника за повећање строгости анализе главних компоненти. Технике крос-валидације су доказале свој значај у примени мултиваријантне статистике.

Најједноставнији приступ процедури крос-валидације је кроз кораке који су потребни за анализу. Крос-валидација је у основи истраживачки метод који тражи занимљив шаблон и претражује матрицу података тражећи редундантност. Кратак преглед који следи дао је Рејмент.

Корак 1. Рачунамо главне компоненте коваријансне матрице S или корелацијске матрице R матрице података X :

$$S = V\Lambda U' \quad \text{и} \quad U'U = I.$$

Друга могућност је да X рачунамо помоћу

$$X = V\Gamma U',$$

где је

$$\gamma_i^2 = (N - 1)\lambda_i.$$

Корак 2. Рачунамо вредности главних компоненти:

$$Z = XU.$$

Корак 3. Одредити критеријум W_m : Овај критеријум се израчунава из средње квадратне разлике између стварних и очекиваних вредности матрице података. Ово се ради методом крос-валидације као што је у наставку наведено:

- а) Поделити X у неколико група.
- б) Бришемо по групу и рачунамо вредности **предиктора** из преосталих података, а онда предвидимо избрисане вредности. У пракси, избрисана група се може наместити да буде само једна врста матрице X .

Корак 4. Одредити утицај сваког објекта: Настављамо са рачунањем **критичних углова** између потпростора и простора података. Критични угао је мера утицаја сваког појединца у узорку, са $t = \arccos \tau$ (где је τ најмањи елемент дијагоналне матрице Γ из

корака 1.). Велике вредности критичног угла означавају висок степен утицаја обсервација у узорку. Још не постоји тест, па се морамо водити искуством за сваку одлуку.

Корак 5. Информативне променљиве: променљиве су избрисане, једна по једна. Мале резидуалне суме квадрата имају благе ефекте анализе главних компоненти, а такве променљиве су добри кандидати за одстрањивање из анализе.

Корак 6. Избор „најбољег“ броја главних компоненти, обично се сматра значајним у факторској анализи. Ако је j -ти елемент матрица $\mathbf{X}$ и $\mathbf{U}$ означен са x_{ij} , односно, u_{ij} , респективно, тада се x_{ij} може израчунати на следећи начин

$$x_{ij} = \sum_{t=1}^p v_{it} \gamma_t u_{tj}.$$

Структура података за m димензија ($p - m$ избрисаних) се моделира следећом формулом

$$x_{ij} = \sum_{t=1}^m v_{it} \gamma_t u_{tj} + e_{ij}, \quad (4.3)$$

где e_{ij} представља грешку.

У вези са датим бројем компоненти m имамо предиктор

$$\hat{x}_{ij}^{(m)} = \sum_{t=1}^m v_{it} \gamma_t u_{tj}$$

за реконструисање елемената матрице података. Средње квадратна разлика између стварне и очекиване вредности је дата следећом једначином

$$PRESS(m) = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p [\hat{x}_{ij}^{(m)} - x_{ij}]^2.$$

Размотримо сада компоненте једначине (4.3) појединачно. Ово ћемо урадити дефинисањем везе

$$W_m = \frac{PRESS(m-1) - PRESS(m)}{D_m} + \frac{PRESS(m)}{D_r}.$$

где је D_m број степени слободе који је потребан за фитовање m -те компоненте, а D_r је број степени слободе који је остао после фитовања m -те компоненте: $D_m n + p - 2m$. PRESS је акроним за *prediction sum of squares*, односно, очекивана сума квадрата. W_m представља раст у очекиваним информацијама у свакој од преосталих компоненти. Дакле, „важне“ компоненте треба да дају вредности матрице W_m веће од један. Међутим, овај критеријум тежи да постане превише строг.

5. ПРИМЕНА МЕТОДА ФАКТОРСКЕ АНАЛИЗЕ У КИНЕЗИОЛОГИЈИ

У овом делу биће дата примена факторске анализе на релане проблеме у кинезиологији, уз примену статистичког пакета *SPSS*.

Факторска анализа није намењена за тестирање хипотеза, нити може рећи да ли се једна група значајно разликује од друге. Сматра се техником за смањење количине података.

Термин факторска анализа, обухвата више различитих, мада сродних техника. По начину спровођења оне се деле на анализу главних компоненти и анализу заједничких фактора. Та два скупа техника слична су по много чему и оба доводе до мањег броја линеарних комбинација првобитних променљивих. Међутим, они се разликују по више основа, иако оба приступа често дају сличне резултате.

У овом поглављу приказаћемо анализу главних компоненти. Факторска анализа се спроводи у три главна корака. (Термин употребљавамо у општем смислу за било коју технику која припада овој породици; дакле, и за анализу главних компоненти).

Корак 1: Оцена прикладности података за факторску анализу

Када се утврђује прикладност одређеног скупа података за факторску анализу, треба размотрити два главна питања: величину узорка и јачину везе између променљивих. Иако се препоруке аутора о величини узорка разликују, сви се слажу у једном: што већи узорак, то бољи. Неки аутори тврде да није толико важна величина узорка, већ количник броја субјеката и броја променљивих. Један од аутора препоручује количник 10 према 1, тј. десет случајева за сваку променљиву коју треба факторски анализирати. Друго питање које треба размотрити јесте јачина корелације међу променљивама. Препоручује се да се у матрицама корелације потраже коефицијенти већи од 0.3. Ако их је мало, ти подаци су неприкладни за факторску анализу. Статистички пакет *SPSS* има два теста оправданости примене факторске анализе: Бартлетов тест сферичности и Кајзер-Мајер-Олкинов показатељ адекватности узорка.

Бартлетов тест сферичности треба да је значајан ($p < 0.05$) да би примена факторске анализе била оправдана.

Кајзер-Мајер-Олкинов показатељ поприма вредности између 0 и 1, па 0.6 препоручује као најмањи износ прихватљив за добру факторску анализу.

Корак 2: Издвајање фактора

Издвајање фактора обухвата одређивање најмањег броја фактора који добро представљају међувезе у скупу променљивих. Као што је већ речено, најчешће се ради анализа главних компоненти. Дакле, истраживач треба сам да одреди број фактора који по његовом мишљењу најбоље описује односе између променљивих, односно, да нађе једноставно решење с најмањим могућим бројем фактора, а да притом објасни што већи део варијансе оригиналног скупа података. За одређивање фактора постоји више техника:

- Кајзеров критеријум (критеријум карактеристичних вредности).
- Дијаграм превоја (*scree test*).

- Паралелна анализа.

Кајзеров критеријум

Кајзеров критеријум или критеријум карактеристичних вредности је једна од најчешће употребљаваних техника. Карактеристична вредност фактора је укупна варијанса свих променљивих објашњена тим фактором. По овом правилу, за даље истраживање задржавају се само они фактори чије су карактеристичне вредности веће од 1, ($\lambda > 1$).

Критеријум дијаграма превоја

Потребно је нацртати карактеристичне вредности свих фактора (SPSS то уме да уради) и на дијаграму наћи тачку у којој се облик криве мења и она прелази у хоризонталу. Препоручује се да се задрже све тачке изнад превоја тог дијаграма, пошто они највише доприносе објашњавању варијансе и скупу података.

Паралелна анализа

Паралелна анализа значи упоредити износе карактеристичних вредности са онима добијеним на једнако великом скупу случајно генерисаних података. Задржавају се само они фактори чије су карактеристичне вредности веће од одговарајућих вредности добијених на насумично генерисаним подацима. Доказано је да је овај начин одређивања броја потребних компоненти најтачнији. У наставку ћемо демонстрирати све три технике у склопу примера.

Корак 3. Ротација и тумачење фактора

Када је број фактора одређен, следећи корак је се покушају протумачити. Да би се овај процес олакшао, прибегава се ротацији фактора. Тиме се решење не мења, али се структура факторских коефицијената корелације између променљивих и фактора-представља тако да се резултати лакше тумаче. SPSS не уме сам да тумачи факторе нити да им додели описе, он само показује које су променљиве „нагомилане заједно“. Сами треба да предложите могућа тумачења на основу сопственог разумевања садржаја променљивих.

Ротирани фактори могу бити ортогонални (некорелирани) или коси (корелирани). Ортогонална ротација даје решења која се лакше тумаче и представљају. У косим ротацијама дозвољени су корелирани фактори, али је њих теже протумачити, описати и представити. У пракси, та два приступа често дају врло слична решења.

У те две широке категорије ротација, спада више различитих техника SPSS-а (ортогоналне: *Varimax*, *Quartimax*, *Equamax*; косе: *Direct Oblimin*, *Promax*). Од ортогоналних ротација најчешће се употребљава метода *Varimax*, док је најчешћа коса ротација *Direct Oblimin*, ипак ми ћемо у примеру користити *Promax*.

5.1. Поступак факторске анализе у SPSS-у

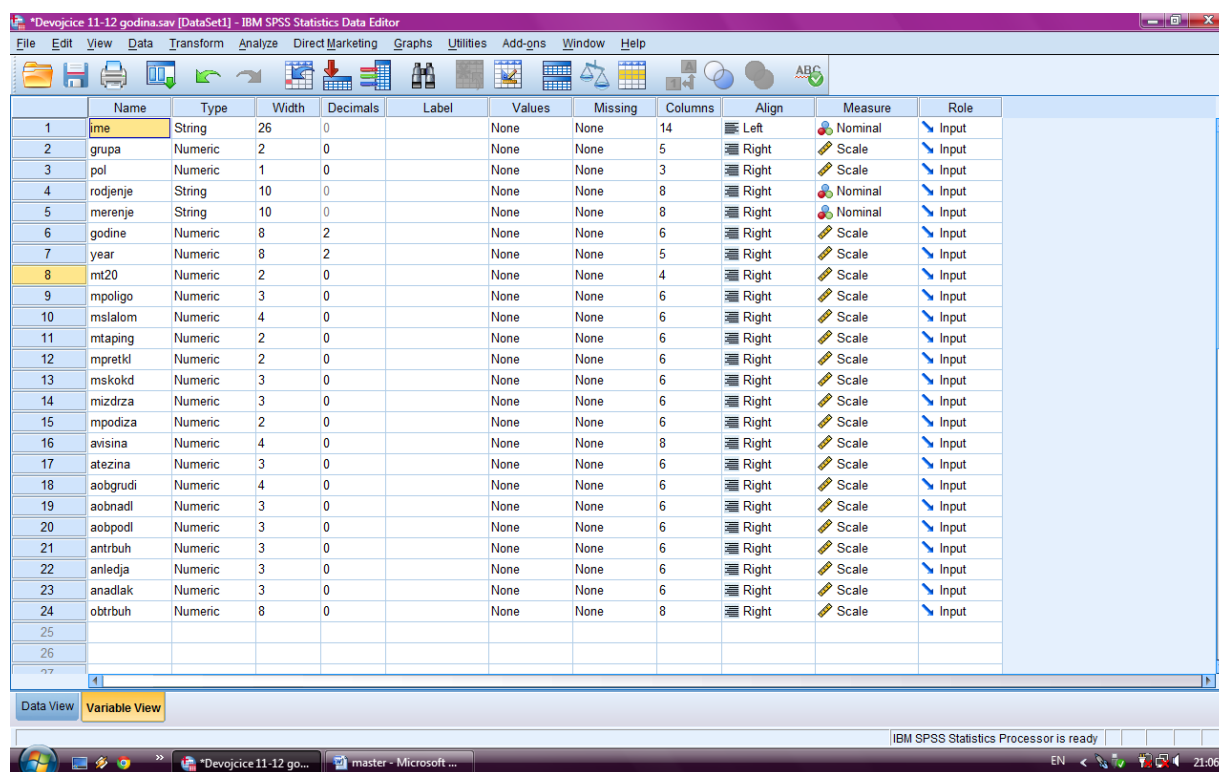
Пример је рађен на узорку од 179 девојчица узраста између 11 и 12 година. Њима су мерене морфолошке карактеристике и моторичке способности. Од морфолошких карактеристика мерене су:

- телесна тежина,
- обим надлактице,
- обим груди,
- обим подлактице,
- кожни набор тбуха,
- набор леђа
- обим трбуха.

Од моторичких карактеристика мерено:

- трчање на 20 метара (спринт),
- полигон унатрашке,
- слалом са три лопте
- скок у даљ.

На слици видимо, главни мени са нашим узорком и подацима.



Слика 5.1: Изглед главног менија у SPSS-у

Correlation Matrix												
	mt20	mpoligo	mslalom	mskokd	atezina	aobnadi	aobgrudi	aobpodl	antrbuh	anledja	anadlak	obtrbuh
Correlation	mt20	1.000	.395	.290	-.516	.091	.161	.066	.100	.276	.248	.209
	mpoligo	.395	1.000	.339	-.528	.420	.315	.207	.213	.488	.417	.400
	mslalom	.290	.339	1.000	-.336	.059	.128	.083	.102	.144	.124	.159
	mskokd	-.516	-.528	-.336	1.000	-.215	-.318	-.192	-.248	-.424	-.369	-.344
	atezina	.091	.420	.059	-.215	1.000	.785	.676	.682	.709	.706	.651
	aobnadi	.161	.315	.128	-.318	.785	1.000	.837	.911	.710	.733	.685
	aobgrudi	.066	.207	.083	-.192	.676	.837	1.000	.813	.523	.547	.464
	aobpodl	.100	.213	.102	-.248	.682	.911	.813	1.000	.548	.578	.527
	antrbuh	.276	.488	.144	-.424	.709	.710	.523	.548	1.000	.869	.820
	anledja	.248	.417	.124	-.369	.706	.733	.547	.578	.869	1.000	.847
	anadlak	.209	.400	.159	-.363	.651	.685	.464	.527	.820	.847	1.000
	obtrbuh	.198	.482	.088	-.344	.848	.767	.668	.642	.803	.803	.698

Табела 5.1: Корелацијска матрица

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.893
Approx. Chi-Square		1922.246
Bartlett's Test of Sphericity		
	df	66
	Sig.	.000

Табела 5.2: Кајзер-Мајер-Олкинов показатељ адекватности узорка и Бартлетов тест сферичности

Како бисмо утврдили да ли је скуп података прикладан за факторску анализу, погледамо да ли је вредност показатеља Кајзер-Мајер-Олкин за адекватност узорка једнака или већа од 0.6 и да ли је вредност Бартлетовог теста сферичности значајна, тј. да ли је вредност Sig. 0.05 или мање. У овом примеру видимо да КМО показатељ износи 0.893, а Бартлетов показатељ је значајан ($p = 0.000$), те је факторска анализа оправдана.

Communalities		
	Initial	Extraction
mt20	1.000	.570
mpoligo	1.000	.603
mslalom	1.000	.374
mskokd	1.000	.650
atezina	1.000	.785
aobnabl	1.000	.880
aobgrudi	1.000	.699
aobpodl	1.000	.723
antrbuh	1.000	.783
anledja	1.000	.781
anadlak	1.000	.687
obtrbuh	1.000	.821

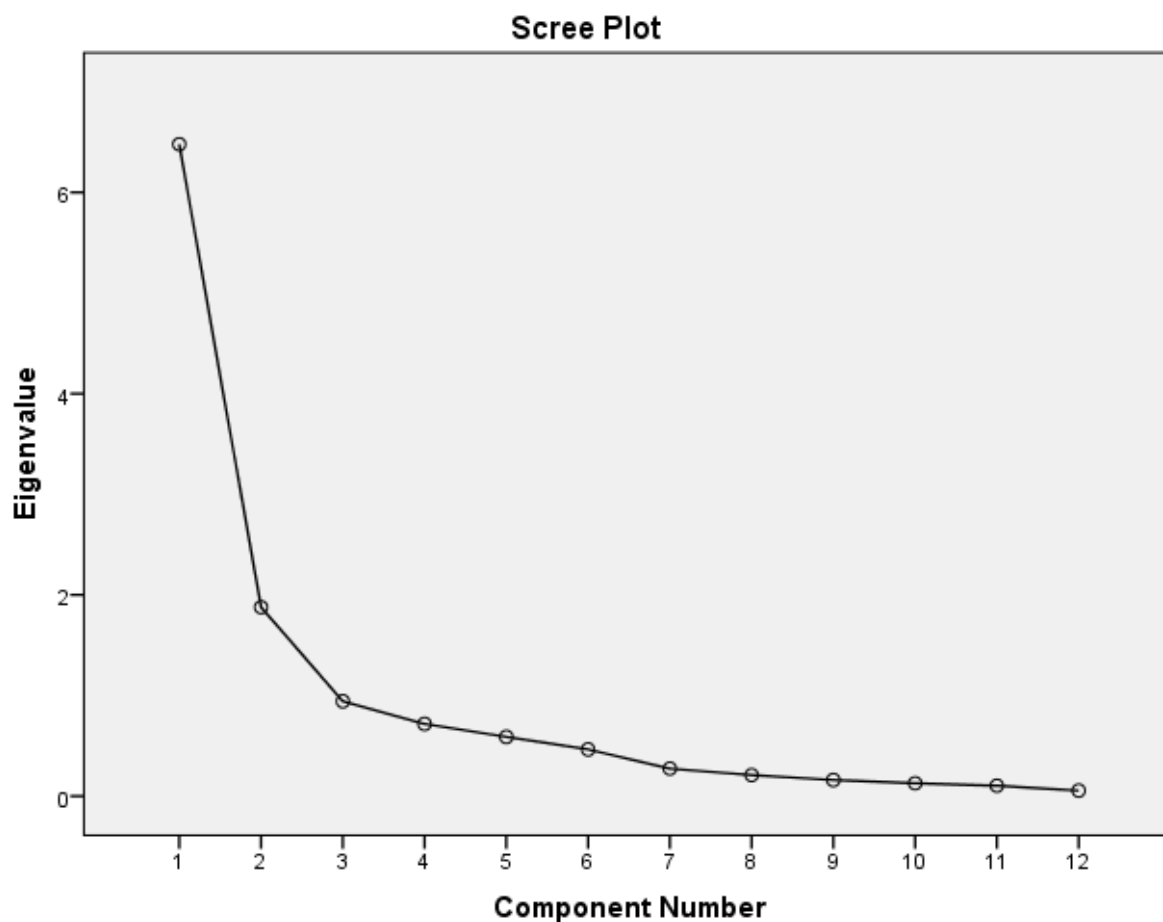
Табела 5.3: Комуналитети

У Табели 5.3 видимо да су све променљиве нормиране. Комуналитет је варијанса променљивих које су објашњене факторима.

Total Variance Explained							
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings ^a
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	6.479	53.990	53.990	6.479	53.990	53.990	6.329
2	1.878	15.649	69.640	1.878	15.649	69.640	3.297
3	.943	7.855	77.494				
4	.718	5.982	83.477				
5	.589	4.908	88.385				
6	.464	3.868	92.253				
7	.273	2.277	94.530				
8	.209	1.743	96.274				
9	.160	1.334	97.607				
10	.128	1.067	98.674				
11	.104	.867	99.541				
12	.055	.459	100.000				

Табела 5.4 Тотал варијансе објашњене факторима

По Крајзеровом критеријуму, занимају нас само компоненте чија је карактеристична вредност један или више. Погледаћемо у Табелу 5.4 и прочитати бројеве у првих неколико колона, под заглављем **Initial Eigenvalues**. Наведене су карактеристичне вредности свих компоненти. У овом примеру само прве две компоненте имају карактеристичне вредности изнад 1 (6.479, 1.878). Те две компоненте објашњавају укупно 69.64 процената варијансе (видимо у колони **Comulative %**).



Слика 5.2: Дијаграм превоја

Некада је број компоненти које задовољавају Кајзеров критеријум превелик, па обавезно треба погледати и дијаграм превоја (ScreePlot) који је *SPSS* нацртао. На њему потражимо превојну тачку, задржавају се само компоненте изнад те тачке. Видимо да је лом на споју друге и треће компоненте. Односно, компоненте 1 и 2 објашњавају много већи део варијансе од преосталих компоненти. На основу овог дијаграма препоручљиво је да се задрже само две компоненте, односно, издвоје само два фактора.

Component Matrix ^a		
	Component	
	1	2
mt20	.309	.689
mpoligo	.545	.553
mslalom	.218	.571
mskokd	-.483	-.646
atezina	.856	-.231
aobnadi	.905	-.246
aobgrudi	.758	-.353
aobpodl	.789	-.318
antrbuh	.883	.066
anledja	.884	-.007
anadlak	.828	.026
obtrbuh	.901	-.096

Табела 5.5 Матрица главних компоненти

У матрици главних компоненти су неротирани факторски коефицијенти сваке од променљивих за те две компоненте (фактора). У Табели 5.5 се види да већина променљивих има пристојно велике факторске коефицијенте (изнад 0.4) за обе компоненте. То нам указује на то да је решење са два фактора примерено.

Pattern Matrix ^a		
	Component	
	1	2
mt20	-.137	.799
mpoligo	.159	.700
mslalom	-.148	.655
mskokd	-.046	-.787
atezina	.914	-.079
aobnadi	.969	-.085
aobgrudi	.900	-.229
aobpodl	.907	-.185
antrbuh	.759	.244
anledja	.804	.166
anadlak	.734	.191
obtrbuh	.874	.075

Табела 5.6: Матрица склопа фактора

Пре него што донесемо коначну одлуку о броју фактора, требало би да погледамо решење са два ротирана фактора приказана у Табели 5.6. Идеално би било да свака

компонента има три или више факторских коефицијената већих од 0.3, па видимо да нам је ово оптимално решење.

Structure Matrix		
	Component	
	1	2
mt20	.180	.745
mpoligo	.436	.763
mslalom	.111	.596
mskokd	-.358	-.805
atezina	.883	.283
aobnadi	.935	.298
aobgrudi	.809	.128
aobpodl	.833	.174
antrbuh	.856	.545
anledja	.870	.485
anadlak	.810	.482
obtrbuh	.904	.421

Табела 5.7: Матрица структуре

Матрица структуре дата у Табели 5.7 даје нам коефицијенте корелације.

Component Correlation Matrix		
Component	1	2
1	1.000	.396
2	.396	1.000

Табела 5.8: Матрица корелација два фактора

Дакле, у овом примеру одлучујемо се за два фактора. Фактор 1 нам представља морфолошке карактеристике, односно, телесну тежину, волуминозност и поткожно масно ткиво, док нам је Фактор 2 генерални моторички фактор, тј. брзина, координација и експлозивна снага.

ЗАКЉУЧАК

Сумирајући добијене резултате у овом раду, видимо да је употреба факторске анализе оправдана. Породица техника факторске анализе има бројне примене. Много је употребљавају истраживачи који се баве развојем и вредновањем тестова и скала. Пројектант скале почиње са великим бројем питања, а затим, техникама факторске анализе, пречишћава и сажима те ставке. Мноштво повезаних променљивих може се факторском анализом свести на мањи број подесан за друге анализе, као што су вишеструка регресија или мултиваријациона анализа варијансе.

У овом раду приказана је примена факторске анализе у кинезиологији за утврђивање логичке ваљаности тестова за процену координације код деце средњег школског узраста. Узорак испитаника је састављен од укупно 179 испитаника, девојчица узраста 11-12 година. Манифестне варијабле које служе за процену опште координације код деце: полигон уназад или натрашке, слалом са три лопте, трчање на 20 метара (спринт), скок у даљ, се свode на мањи број фактора. Сprovedена је комплетна факторска анализа, рачунањем дескриптивних статистика, матрице интеркорелације манифестних варијабли, КМО тест и Бартлетов тест сферицитета. Даљи развој и усавршавање се може вршити у правцу укључивања других променљивих.

Рад завршавамо констатацијом да је факторска анализа веома интересантна и инспиративна за многе научнике. Управо из тог разлога, данас није познат тачан број метода факторских анализа, који се непрестано повећава. Наравно, повећање броја метода прати и повећање броја научних области у којима се она примењује, тако да данас и не постоји грана у којој се не примењује.

ЛИТЕРАТУРА

- [1] Richard A. Reymont, K. G. Joreskog, *Applied Factor Analysis in the Natural Science*, Cabridge University Press, 1993;
- [2] Alexander Basilevsky, *Statistical Factor Analysis and Related Methods*, John Wiley & Sons, Inc., 1994.;
- [3] Bruce Thompson, *Exploratory and Confirmatory Factor Analysis*, American Psychological Association Washington DC, 2004.;
- [4] Julie Pallant, *SPSS Survival Manual, Third Edition*, 2007.;
- [5] Robert Cudeck, Robert C. MacCallum, *Analysis at 100 – Historical Developments and Future Directions*, 2007.;
- [6] David A. Kenny, Series editor, *Metodology in the Social Sciences*, The Guilford Press, 2006.;
- [7] Malinowski R. Edmundi, *Factor Analysis in Chemistry*, 3rd Edition, Wiley-Interscience, 2002.;

КРАТКА БИОГРАФИЈА

Сања Станковић је рођена 10. марта 1987. године у Зрењанину. Основну школу „Соња Маринковић“ је завршила у Зрењанину као носилац Вукове дипломе. 2006. године завршава Зрењанинску гимназију, такође као носилац Вукове дипломе. Током школовања учествовала је на многобројним такмичењима из природних и друштвених наука. У Нови Сад се доселила 2006. године како би студирала на Природно-математичком факултету. На департману за математику и информатику уписала је смер математика финансија. У априлу 2011. завршава основне студије и у октобру исте године уписује мастер студије на Природно-математичком факултету, смер Примењена математика. Положила је све испите предвиђене планом и програмом мастер студија и тиме стекла право на одбрану мастер тезе. Марта 2013. године радила је као професор математике у ОШ „Јован Јовановић Змај“ у Зрењанину, а од јула 2013. године ради као сарадник у заједничким ХР сервисима у НИС а. д. Нови Сад.



Универзитет у Новом Саду
Природно-математички факултет
Кључна документацијска информација

Редни број:

РБР

Идентификациони број:

ИБР

Тип документације: Монографска документација

ТД

Тип записа: Текстуални штампани материјал

ТЗ

Врста рада: Мастер рад

ВР

Аутор: Сања Станковић

АУ

Ментор: Др Загорка Лозанов-Црвенковић

МН

Наслов рада: Примена факторске анализе у кинезологији

НР

Језик публикације: Спрски (ћирилица)

ЈП

Језик извода: с/ен

ЈИ

Земља публикација: Република Србија

ЗП

Уже географско подручје: Војводина

УГП

Година: 2013

ГО

Издавач: Ауторски репринт

ИЗ

Место и адреса: Трг Доситеја Обрадовића 4, 21000 Нови Сад

МА

Физички опис рада: (5/ 62/ 0 / 12 / 12 /0 / 0)

ФО

Научна област: Математика

НО

Научна дисциплина: Статистика

НД

Предметна одредница, кључне речи: Факторска анализа, Анализа главних компоненти

ПО

УДК

Чува се: Библиотека департмана за математику и информатику, Природно- математички факултет,

Трг Доситеја Обрадовића 4, Нови Сад

ЧУ

Важна напомена: Нема

ВН

Извод: У мастер раду су представљене статистичке методе за смањење димензионалности података и то анализа главних компоненти, (стварна) факторска анализа и анализа главних компоненти и крос валидација. Све три методе спадају у статистичку област која се назива факторска анализа. Основна идеја факторске анализе јесте да се број променљивих унутар матрице података смањи на одређени број фактора (компоненти), како би нови модел имао бољу предикциону моћ. Тиме се решава и проблем

мультиколинearности (сингуларне матрице), а и смањује се број променљивих (економичност). Анализу главних компоненти смо применили у кинезиологији, уз примену статистичког пакета *SPSS*.

ИЗ

Датум прихватања теме од стране НН већа:

ДП

Датум одбране:

ДО

Чланови комисије: (научни степен/име и презиме/звање/факултет)

КО

Председник: др Љиљана Гајић, редовни професор Природно-математичког факултета у Новом Саду

Члан: др Загорка Лозанов-Црвенковић, редовни професор Природно-математичког факултета у Новом Саду

Члан: др Ивана Штајнер-Папуга, ванредни професор Природно-математичког факултета у Новом Саду

University of Novi Sad
Faculty of Natural Sciences And Mathematics
Key word documentation

Accession number:

ANO

Identification number:

INO

Document type: Monograph documentation

DT

Type of record: Textual printed material

TR

Contents code: Master's thesis

CC

Author: Sanja Stanković

AU

Mentor: Zagorka Lozanov-Crvenković Ph.D.

MN

Title Applied Factor Analysis in Kinesiology

TI

Language of text: Serbian (cirilic)

LT

Language of abstract: en/s

LA

Country of publication: Republic of Serbia

CP

Locality of publication: Vojvodina

LP

Publication year: 2013

PY

Publisher: Author's reprint

PU

Publ. place: Trg Dositea Obradovic 4, Novi Sad

88

PP

Physical description: (5/ 62/ 0 / 12 / 12 /0 / 0)

PD

Scientific field: Mathematics

SF

Scientific discipline: Statistics

SD

Subject Key words: Factor analysis, Principal component analysis

SKW

UC

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

HD

Abstract: This master's thesis includes three statistical methods for reducing the dimensionality of data: principal component analysis, true factor analysis and principal component analysis and cross-validation. These methods belong to statistical field named factor analysis. The main idea of factor analysis is reducing the number of variables in a data matrix to smaller number of factors (components), because the new model has better prediction power. Furthermore, the new model solves the problem of multicollinearity (singular matrix) and reduces the number of variables (economical). Principal component analysis was applied in kinesiology, using the SPSS statistical software.

AB

Accepted on Scientific Board on:.

AS

Defended:

DE

Thesis Defend board:

DB

President: Ljiljana Gajić Ph. D., Full professor, Faculty of Natural Sciences and Mathematics, Novi Sad

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

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