Faktorska analiza

Jurić, Ozana

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ozana Jurić

FAKTORSKA ANALIZA

Diplomski rad

Voditelj rada:

Prof. dr. sc. Anamarija Jazbec

Zagreb, rujan 2021.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

mami i tati

Sadržaj

Uvod

Faktorska analiza (engl. *factor analysis*) je višedimenzionalna statistička metoda koja se koristi za opisivanje varijabilnosti među opaženim varijablama preko potencijalno manjeg broja nepromatranih (latentnih) varijabli koje nazivamo $faktori$. Dvije su vrste faktorske analize: eksplorativna faktorska analiza i konfirmatorna faktorska analiza. Eksplorativna faktorska analiza služi za otkrivanje pozadinske strukture relativno velikog broja varijabli, dok konfirmatorna faktorska analiza služi za verifikaciju faktorske strukture skupa promatranih varijabli. U ovom radu tema je eksplorativna faktorska analiza.

Zadaća faktorske analize je sažimanje većeg broja povezanih izvornih varijabli u manji broj faktora tako da varijablu odaziva možemo gotovo jednako dobro opisati kao i sa većim brojem varijabli poticaja. Sažimanje izvornih varijabli se sastoji od toga da varijable koje su međusobno povezane, a istovremeno su nezavisne od drugih podskupova varijabli, povezujemo u faktore. Polazna pretpostavka je da među varijablama postoji linearna korelacija, a svaki od dobivenih faktora nije u korelaciji s drugim faktorima. Faktorska analiza je korisna jer se njeni rezultati primijenjuju u daljnjim analizama. Tako se umjesto nad velikim brojem koreliranih izvornih varijabli, analiza provodi nad manjim brojem nekoreliranih faktora, od kojih je svaki faktor linearna kombinacija nekoliko promatranih varijabli. Primjenu faktorske analize nalazimo u biologiji, marketingu, financijama, strojnom učenju te brojnim drugim područjima. Glavna motivacija za korištenje faktorske analize je u mogućnosti smislene interpretacije podataka.

U ovom radu bavit ćemo se teorijskom pozadinom faktorske analize i primjenom proučene teorije. Teorijski dio rada započet ćemo detaljnom definicijom modela faktorske analize. Ovdje ćemo vidjeti da je model faktorske analize određen faktorima i matricom koeficijenata, odnosno težinama. Ukratko ćemo se ostvrnuti na nejedinstvenost težina u faktorskom modelu ukoliko u model uvedemo ortogonalnu matricu. Nakon toga ćemo u trećem poglavlju predstaviti načine za određivanje faktorskog modela, odnosno koeficijenata modela, odnosno težina. Načine koje ćemo detaljno predstaviti biti će metoda glavnih komponenata, metoda glavnih faktora, iterativna metoda glavnih faktora i metoda maksimalne vjerodostojnosti.

U četvrtom poglavlju obrađivat ćemo kako odabrati broj faktora u modelu. Vidjet ćemo da postoji više kriterija za odabir broja faktora, a da kod podataka koji su pogodni za pri-

mjenu faktorske analize, većina kriterija ishodi jednakim brojem.

U petom poglavlju ćemo se upoznati s rotacijama faktora modela i različitim tipovima rotacija. Cilj rotacija je pronaći novi prostor u kojem su dobiveni faktori interpretabilniji te postaviti osi prostora blizu što je više točaka, odnosno varijabli, moguće. Promatrat ćemo dva tipa rotacija: ortogonalnu i kosu rotaciju. Naposlijetku, ćemo se baviti faktorskim score-ovim i valjanosti modela faktorske analize.

Na samom kraju, u primijenjenom dijelu radu, ćemo pokazati na konkretnom primjeru sprovedbu faktorske analize i detaljno opisati dobivene rezultate. Svi izvodi u radu koji nemaju referencu dolaze iz [2].

Poglavlje 1

Model faktorske analize

1.1 Definicija modela

Neka je $\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n}$ slučajan uzorak iz homogene populacije sa vektorom očekivanja $\boldsymbol{\mu}$ i kovarijacijskom matricom $\boldsymbol{\Sigma}$. Model fakorske analize prikazuje svaku od varijabli kao linearnu kombinaciju (neopaženih, odnosno latentnih) zajedničkih faktora $f_1, f_2, ..., f_m$ i slučajne greške. U faktorskom modelu faktori nisu zadani, oni se procijenjuju iz podataka. Neka je $y_1, y_2, ..., y_p$ opservacija p slučajnih varijabli, tada je model dan sa:

$$\begin{aligned} y_1 - \mu_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + ...\lambda_{1m} f_m + \epsilon_1 \\ y_2 - \mu_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + ...\lambda_{2m} f_m + \epsilon_2 \\ &\vdots \\ y_p - \mu_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + ...\lambda_{pm} f_m + \epsilon_p \end{aligned}$$

U faktorskom modelu je cilj da je m strogo manji od p, odnosno da je broj faktora manji od broja slučajnih varijabli. Koeficijente λ_{ij} modela nazivamo težine, te one ukazuju koliko j-ti faktor f_j utječe na i-tu varijablu y_i te nam služe za interpretaciju faktora f_j. Faktor f_2 opisujemo analizirajući njegove koeficijente (težine) $\lambda_{12}, \lambda_{22}, ..., \lambda_{p2}$. Veće težine povezuju npr. faktor f_2 s odgovarajućem y_i-om. Preko varijabli modela y_i objašnjavamo i interpretiramo faktor f_2. Nakon procjene koeficijenata, odnosno težina λ_{ij} nadamo se da će oni razdvojiti varijable u grupe koje odgovaraju faktorima. Vektor $\boldsymbol{\epsilon}$ sadrži grešku mjerenja, individualan učinak svake varijable y_i na grešku te grešku uzorkovanja.

Osnovne pretpostavke faktorskog modela su da za $j = 1, 2, ..., m$ vrijedi $E(f_j) = 0, var(f_j) = 1$ i da je $cov(f_j, f_k) = 0, j \neq k$. Pretpostavke za greške su slične, odnosno vrijedi za $i = 1, 2, ..., p$ da je $E(\epsilon_i) = 0$, $cov(\epsilon_i, \epsilon_k) = 0, i \neq k$. Budući da su greške rezidualni dijelovi od pripadnih y_i, vrijedi da imaju različite varijance, odnosno $var(\epsilon_i) = \psi_i$. Dodatno, pretpostavljamo da je $cov(\epsilon_i, f_j) = 0$, za sve i, j.

Gore navedene pretpostavke su prirodna poslijedica modela i ciljeva faktorske analize. Budući da je $E(y_i - \mu_i) = 0$ slijedi da je $E(f_j) = 0$ za $j = 1, 2, ..., m$. Pretpostavka $cov(f_j, f_k) = 0$ proizlazi iz cilja prikazivanja varijabli $\mathbf{y_i}$ kao funkcije što je manje faktora moguće, odnosno razdvajanja varijabli u odvojene "grupe" faktora.
Pretpostavke $var(f_j) = 1$, $var(\epsilon_i) = \Psi_i$, $cov(f_j, f_k) = 0$ i $cov(\epsilon_i, f_j) = 0$ nas dovode do jednostavnog izraza za varijancu od y_i :

$$var(y_i) = {\lambda_{i1}}^2 + {\lambda_{i2}}^2 + ... + {\lambda_{im}}^2 + \psi_i \tag{1.1}$$

koji ima važnu ulogu u izgradnji modela. Također, valja primijetiti da pretpostavka $cov(\epsilon_i, \epsilon_k) = 0$ implicira da faktori sadrže svu korelaciju između **y**-a, odnosno, sve što **y** imaju zajedničko. Odatle je naglasak u faktorskoj analizi na modeliranje korelacija između **y**-a.

Model faktorske analize može biti zapisan u matričnom obliku:

$$\mathbf{y} - \boldsymbol{\mu} = \mathbf{\Lambda f} + \boldsymbol{\epsilon} \tag{1.2}$$

gdje je $\mathbf{y} = (y_1, y_2, ..., y_p)^T$, $\boldsymbol{\mu} = (\mu_1, \mu_2, ...\mu_p)^T$, $\mathbf{f} = (f_1, f_2, ..., f_m)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ...\epsilon_p)^T$ i :

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{bmatrix}. \tag{1.3}$$

Ilustrirajmo sada model za neke p i m. Neka je broj varijabli $p = 5$ i broj faktora $m = 2$. Model sada zapisujemo:

$$\begin{aligned} y_1 - \mu_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \epsilon_1 \\ y_2 - \mu_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \epsilon_2 \\ y_3 - \mu_3 &= \lambda_{31} f_1 + \lambda_{32} f_2 + \epsilon_3 \\ y_4 - \mu_4 &= \lambda_{41} f_1 + \lambda_{42} f_2 + \epsilon_4 \\ y_5 - \mu_5 &= \lambda_{51} f_1 + \lambda_{52} f_2 + \epsilon_5 \end{aligned} \tag{1.4}$$

odnosno, u matričnom zapisu:

$$\begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ y_3 - \mu_3 \\ y_4 - \mu_4 \\ y_5 - \mu_5 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}, \tag{1.5}$$

odnosno, $\mathbf{y} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon}$.

Gore navedene pretpostavke sada možemo, koristeći matričnu notaciju, zapisat na slijedeći način :
$E(f_j) = 0$, za $j = 1, 2, ..., m$ postaje :

$$E(\mathbf{f}) = \mathbf{0} \tag{1.6}$$

iz $var(f_j) = 1$, za $j = 1, 2, ..., m$ i $cov(f_j, f_k) = 0$, $j \neq k$ imamo:

$$cov(\mathbf{f}) = \mathbf{I} \tag{1.7}$$

iz $E(\epsilon_i) = 0$, za $i = 1, 2, ..., p$ imamo:

$$E(\boldsymbol{\epsilon}) = \mathbf{0} \tag{1.8}$$

dok iz $var(\epsilon_i) = \psi_i$, za $i = 1, 2, ..., p$ i $cov(\epsilon_i, \epsilon_k) = 0$, $j \neq k$ imamo :

$$cov(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}, \tag{1.9}$$

te naposlijetku iz $cov(\epsilon_i, f_j) = 0$, za sve i, j imamo :

$$cov(\mathbf{f}, \boldsymbol{\epsilon}) = \mathbf{0}. \tag{1.10}$$

Spomenuli smo već da je naglasak u faktorskoj analizi na modeliranju kovarijanci između y-a. U cilju nam je izraziti $\frac{1}{2}p(1-p)$ kovarijanci (odnosno, p varijanci) varijabli $y_1, y_2, ..., y_p$ preko pojednostavljene strukture koja uključuje pm težina μ_{ij} i p varijanci ψ_i, odnosno $\boldsymbol{\Sigma}$ želimo izraziti preko Λ i $\boldsymbol{\Psi}$. To ćemo učiniti koristeći naš model zapisan u matričnom obliku (1.2) i koristeći pretpostavke (1.7),(1.9) i (1.10).
Budući da $\boldsymbol{\mu}$ ne utječe na varijance i kovarijance od $\mathbf{y}$-a to iz (1.2) imamo :

$$\boldsymbol{\Sigma} = cov(\mathbf{y}) = cov(\boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon}). \tag{1.11}$$

Po (1.10) imamo da su $\boldsymbol{\Lambda}\mathbf{f}$ i $\boldsymbol{\epsilon}$ nekorelirani, stoga je kovarijanca njihovog zbroja jednaka zbroju njihovih kovarijanci, odnosno:

$$\begin{aligned} \boldsymbol{\Sigma} &= cov(\boldsymbol{\Lambda}\mathbf{f}) + cov(\boldsymbol{\epsilon}) \\ &= \boldsymbol{\Lambda} cov(\mathbf{f}) \boldsymbol{\Lambda}^T + \boldsymbol{\Psi} \\ &= \boldsymbol{\Lambda}\mathbf{I}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi} \\ &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi} \end{aligned}$$

Ukoliko $\mathbf{\Lambda}$ ima mali broj kolona, recimo dvije ili tri kolone, onda jednakost $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ predstavlja pojednostavljenu strukturu za $\mathbf{\Sigma}$ u kojoj su kovarijance modelirane pomoću λ_{ij} budući da je $\mathbf{\Psi}$ dijagonalna matrica. Ako se vratimo na primjer (1.4) gdje je m = 2, odnosno broj faktora je jednak dva, onda je σ_{12} produkt prva dva reda matrice $\mathbf{\Lambda}$:

$$\sigma_{12} = cov(y_1, y_2) = \lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22} \tag{1.12}$$

gdje je $(\lambda_{11}, \lambda_{12})$ prvi red matrice $\mathbf{\Lambda}$, a $(\lambda_{21}\lambda_{22})$ drugi red matrice $\mathbf{\Lambda}$. Ukoliko varijable y_1 i y_2 imaju puno toga zajedničkog onda će one imati vrijednosno slične težine na zajedničkim faktorima f_1 i f_2, odnosno $(\lambda_{11}, \lambda_{12})$ će biti slično $(\lambda_{21}, \lambda_{22})$. S druge strane, ako y_1 i y_2 nemaju puno sličnosti, onda će njihove pripadne težine λ_{11} i λ_{21} uz faktor f_1 biti drugačije, kao i pripadne težine λ_{12} i λ_{22} , uz faktor f_2. Odnosno, u ovom slučaju umnošci $\lambda_{11}\lambda_{21}$ i $\lambda_{12}\lambda_{22}$ će biti mali.

Također možemo izraziti kovarijance y-a i faktora f preko težina λ. Učinimo to za $cov(y_1, f_2)$. Iz $y_1 - \mu_1 = \lambda_{11}f_1 + \lambda_{12}f_2 + ...\lambda_{1m}f_m + \epsilon_1$, zatim iz $cov(\mathbf{f}) = \mathbf{I}$ (1.7) imamo da je f_2 nekoreliran s ostalim f_j , a iz $cov(\mathbf{f}, \boldsymbol{\epsilon}) = \mathbf{0}$ (1.10) imamo da f_2 nije korelirana s $\boldsymbol{\epsilon}$. Dakle, vrijedi:

$$\begin{aligned} cov(y_1, f_2) &= E[(y_1 - \mu_1)(f_2 - \mu f_2)] \\ &= E[(\lambda_{11}f_1 + \lambda_{12}f_2 + ...\lambda_{1m}f_m)f_2] \\ &= E[\lambda_{11}f_1f_2 + \lambda_{12}f_2f_2 + ... + \lambda_{1m}f_mf_2] \\ &= \lambda_{11}cov(f_1, f_2) + \lambda_{12}var(f_2) + ... + \lambda_{1m}cov(f_m, f_2) \\ &= \lambda_{12} \end{aligned}$$

jer je $var(f_2) = 1$. Iz toga vidimo da su težine zapravo jednake kovarijanci varijabli i faktora. Općenito vrijedi:

$$\begin{aligned} cov(y_i, f_j) &= \lambda_{ij} \\ i &= 1, 2, ..., p \\ j &= 1, 2, ..., m \end{aligned} \tag{1.13}$$

Budući da je λ_{ij} (ij)-ti element matrice $\mathbf{\Lambda}$, gornji izvod zapisujemo u matričnom obliku:

$$cov(\mathbf{y}, \mathbf{f}) = \mathbf{\Lambda} \tag{1.14}$$

U (1.1) smo prikazali varijancu od y_i kao zbroj dvaju dijela, prvi dio koji dolazi od težina, tj. zajedničkih koeficijenata, taj dio zovemo *komunalitet* te drugog dijela koji je specifičan za svaki y_i, taj dio zovemo *specifična varijanca*:

$$\begin{aligned} \sigma_{ii} &= var(y_i) \\ &= (\lambda_{i1}{}^2 + \lambda_{i2}{}^2 + ... + \lambda_{im}{}^2) + \psi_i \\ &= h_i{}^2 + \psi_i \\ &= \textit{komunalitet + specifična varijanca} \end{aligned} \tag{1.15}$$

Komunalitet ($h_i{}^2$) određene varijable nam dakle govori koliko je varijance te varijable objašnjeno zajedničkim faktorima.
Jednadžbe (1.6)-(1.10) su nas dovele do jednostavne jednadžbe za varijancu $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ koja je temeljni dio faktorskog modela. Dijagonalni elementi od $\mathbf{\Sigma}$ mogu lagano biti modelirani prilagođavajući dijagonalne elemente od $\mathbf{\Psi}$, dok uz pomoć $\mathbf{\Lambda}\mathbf{\Lambda}^T$ dolazimo do nedijagonalnih elemenata.
U rijetkim slučajevima se populacijska kovarijacijska matrica može prikazati u obliku $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ gdje je $\mathbf{\Psi}$ dijagonalna matirica, a $\mathbf{\Lambda}$ $p \times m$ matrica, uz relativno mali m. U praksi rijetko imamo uzoračke kovarijacijske matrice koje zadovolje taj idealni model, no tu pretpostavku ne izostavljamo jer je struktura $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ esencijalna za procjenu $\mathbf{\Lambda}$.

Jedna od prednosti modela faktorske analize je da ukoliko podaci ne odgovaraju modelu, to se jasno vidi u procjeni od $\mathbf{\Lambda}$. U situacijama kada model ne odgovara podacima, dva su problema : nejasno je koliko faktora je potrebno i nejasno je koji su faktori.

1.2 Nejedinstvenost težina

Težine modela $\mathbf{y} - \boldsymbol{\mu} = \Lambda\mathbf{f} + \epsilon$ mogu biti pomnožene ortogonalnom matricom bez da to utječe na faktorizaciju $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$. Neka je $\mathbf{T}$ proizvoljna ortogonalna matrica. Budući da za ortogonalne matrice vrijedi $\mathbf{T}\mathbf{T}^T = \mathbf{I}$ to vrijedi slijedeće:

$$\mathbf{y} - \boldsymbol{\mu} = \Lambda\mathbf{T}\mathbf{T}^T\mathbf{f} + \boldsymbol{\epsilon} \tag{1.16}$$

Neka su sada:

$$\begin{aligned} \mathbf{\Lambda}^* &= \mathbf{\Lambda}\mathbf{T}, \\ \mathbf{f}^* &= \mathbf{T}^T\mathbf{f}. \end{aligned}$$

Napravimo sada zamjenu, umjesto $\mathbf{\Lambda}$ stavimo u temeljnu jednadžbu $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ transformiranu matricu težina $\mathbf{\Lambda}^*$. Tada imamo:

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{\Lambda}^*\mathbf{\Lambda}^{*T} + \mathbf{\Psi} \\ &= \mathbf{\Lambda}\mathbf{T}(\mathbf{\Lambda}\mathbf{T})^T + \mathbf{\Psi} \\ &= \mathbf{\Lambda}\mathbf{T}\mathbf{T}^T\mathbf{\Lambda}^T + \mathbf{\Psi} \\ &= \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi} \end{aligned}$$

Odnosno, transformirana matrica težina $\mathbf{\Lambda}^*$ reproducira istu kovarijacijsku matricu kao i matrica težina $\mathbf{\Lambda}$:

$$\mathbf{\Sigma} = \mathbf{\Lambda}^*\mathbf{\Lambda}^{*T} + \mathbf{\Psi} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi} \tag{1.17}$$

Transformirani faktori $\mathbf{f}^* = \mathbf{T}^T\mathbf{f}$ zadovoljavaju pretpostavke (1.6), (1.7) i (1.10).

Transformacijom $\mathbf{\Lambda}^* = \mathbf{\Lambda T}$ nije ni zajednička varijanca, odnosno komunalitet ${h_i}^2 = {\lambda_{i1}}^2 + {\lambda_{i2}}^2 + ... + {\lambda_{im}}^2$ za $i = 1, 2, ..., p$ promijenjen. Zajednička varijanca je suma kvadrata i-tog reda matrice težina $\mathbf{\Lambda}$. Ako označimo i-ti red matrice $\mathbf{\Lambda}$ sa ${\lambda_\mathbf{i}}^T$ onda je suma kvadrata težina u matričnom obliku zapisana kao ${h_i}^2 = {\lambda_\mathbf{i}}^T\lambda_\mathbf{i}$. Sada je i-ti red matrice $\mathbf{\Lambda}^* = \mathbf{\Lambda T}$ jednak ${\lambda_\mathbf{i}}^* = {\lambda_\mathbf{i}}^T\mathbf{T}$, a pripadni komunalitet je:

$$\begin{aligned} {h_i}^{*2} &= {\lambda_\mathbf{i}}^{*T}\lambda_\mathbf{i}^* \\ &= {\lambda_\mathbf{i}}^T\mathbf{T}\mathbf{T}^T\lambda_\mathbf{i} \\ &= {\lambda_\mathbf{i}}^T\lambda_\mathbf{i} \\ &= {h_i}^2. \end{aligned}$$

Dakle, komunalitet ostaje jednak. Važno je uočiti da je ${h_i}^2 = {\lambda_{i1}}^2 + {\lambda_{i2}}^2 + ... + {\lambda_{im}}^2$ udaljenost od ishodišta do točke $(\lambda_{i1}, \lambda_{i2}, ..., \lambda_{im})$ u m-dimenzionalnom prostoru. Kako je udaljenost ${\lambda_\mathbf{i}}^T\lambda_\mathbf{i}$ jednaka kao udaljenost ${\lambda_\mathbf{i}}^{*T}\lambda_\mathbf{i}^*$ to su točke λ_i^* rotirane od točke λ_i (ovo slijedi i iz činjenice da je množenje vektora ortogonalnom matricom ekvivalentno rotaciji osi). Ova upravo diskutirana mmogućnost, rotacije težina, a pritom zadržavanje svih pretostavki i svojstava je vrlo korisna u interpretaciji faktora i o tome će biti govora više u poglavlju Rotacija.

1.3 Ekstrakcije faktora

Među tehnikama ekstrakcije faktora modela faktorske analize nalaze se analiza glavnih komponenata (eng. *principal component analysis* – PCA), analiza zajedničkih faktora (eng. *principal factor analysis* – PFA), metoda maksimalne vjerodostojnosti, image ekstrakcija, alfa ekstrakcija i metoda neponderiranih i ponderiranih najmanjih kvadrata [3]. Od navedenih najčešće se primjenjuju metoda analize glavnih komponenata i analize zajedničkih faktora. Sve tehnike ekstrakcije računaju skup ortogonalnih komponenti, odnosno faktora. U slijedećim potpoglavljima diskutirat ćemo četiri različita pristupa procijene, odnosno ekstrakcije težina u modelu faktorske analize.

1.3.1 Metoda glavnih komponenata

Prva metoda procijene težina koju ćemo diskutirati zove se *metoda glavnih komponenti*. Cilj ove metode je ekstrahirati maksimalnu varijancu iz podatkovnog skupa sa svakom ortogonalnom komponentom. Prva glavna komponenta je linearna kombinacija promatranih varijabli koja uzima najviše varijabilnosti među podacima. Druga glavna komponenta je opet linearna kombinacija promatranih varijabli koja uzima drugi najveći udio varijabilnosti među podacima i ortogonalna je s prvom. Nastavljajući tako, dobivamo skup glavnih

komponenti. Glavne komponente su poredane, gdje prva komponenta ekstrahira najveći dio varijabilnosti, a posljednja ekstrahira najmanji dio varijabilnosti. Rezultat je matematički jedinstven i ukoliko su zadržane sve komponente, možemo reproducirati promatranu korelacijsku matricu. Ova metoda je najpogodnija za istraživače kojima je od primarnog interesa redukcija velikog broja varijabli na manji broj komponenti. PCA metoda je korisna i kao inicijalni korak u FA metodi gdje pomaže u otkrivanju maksimalnog broja i prirode faktora.

Za slučajni uzorak $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ izračunamo uzoračku kovarijacijsku matricu $\mathbf{S}$ i želimo pronaći procijenitelja $\hat{\mathbf{\Lambda}}$ koji aproksimira temeljnu faktorsku jednadžbu $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ uz zamijenu $\mathbf{\Sigma}$ s $\mathbf{S}$:

$$\mathbf{S} \cong \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}}. \tag{1.18}$$

Kako bi faktorizirali $\mathbf{S}$ koristimo se spektralnom dekompozicijom:

$$\mathbf{S} = \mathbf{C}\mathbf{D}\mathbf{C}^T \tag{1.19}$$

gdje je $\mathbf{C}$ ortogonalna matrica koja se sastoji od normaliziranih svojstvenih vektora (normalnizirani vektori su oni kojima je norma jednaka jedan) matrice $\mathbf{S}$, a $\mathbf{D}$ je dijagonalna matrica koja na dijagonali ima svojstvene vrijednosti $\theta_1, \theta_2, ..., \theta_p$ matrice $\mathbf{S}$:

$$\mathbf{D} = \begin{bmatrix} \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_p \end{bmatrix}. \tag{1.20}$$

Budući da su svojstvene vrijednosti θ_i od pozitivno semidefinitne matrice nenegative (veće ili jednake od 0), možemo faktorizirati matricu $\mathbf{D}$:

$$\mathbf{D} = \mathbf{D}^{1/2}\mathbf{D}^{1/2} \tag{1.21}$$

gdje vrijedi:

$$\mathbf{D}^{1/2} = \begin{bmatrix} \sqrt{\theta_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\theta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\theta_p} \end{bmatrix}. \tag{1.22}$$

Iskoristimo tako faktoriziran $\mathbf{D}$ u (1.19) i imamo:

$$\begin{aligned} \mathbf{S} = \mathbf{C}\mathbf{D}\mathbf{C}^T &= \mathbf{C}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{C}^T = \\ &(\mathbf{C}\mathbf{D}^{1/2})(\mathbf{C}\mathbf{D}^{1/2})^T \end{aligned} \tag{1.23}$$

Sada smo dobili traženu formu $\mathbf{S} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T$, ali ne definiramo da je $\hat{\mathbf{\Lambda}}$ jednako $\mathbf{CD}^{1/2}$ jer je $\mathbf{CD}^{1/2}$ $p \times p$ dimezija, a nama je u cilju pronaći $\hat{\mathbf{\Lambda}}$ koji je $p \times m$ dimenzija, pri čemu je $m < p$. Iz navedenih razloga definiramo $\mathbf{D}_1 = diag(\theta_1, \theta_2, ..., \theta_m)$ sa m najvećih svojstvenih vrijednosti $\theta_1 > \theta_2 > ... > \theta_m$ i $\mathbf{C}_1 = (\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_m)$ koji sarži pripadnih m svojstvenih vektora. Sada procijenjujemo $\mathbf{\Lambda}$ sa prvih m stupaca matrice $\mathbf{CD}^{1/2}$, odnosno:

$$\hat{\mathbf{\Lambda}} = \mathbf{C}_1\mathbf{D}_1^{1/2} = (\sqrt{\theta_1}\mathbf{c}_1, \sqrt{\theta_2}\mathbf{c}_2, ..., \sqrt{\theta_m}\mathbf{c}_m) \tag{1.24}$$

gdje su $\hat{\mathbf{\Lambda}}$ $p \times m$, $\mathbf{C}_1$ $p \times m$ i $\mathbf{D}_1^{1/2}$ $m \times m$ dimenzija.

Pokažimo sada strukturu iz (1.24) na primjeru gdje je $p = 5$ i $m = 2$:

$$\begin{bmatrix} \hat{\lambda_{11}} & \hat{\lambda_{12}} \\ \hat{\lambda_{21}} & \hat{\lambda_{22}} \\ \hat{\lambda_{31}} & \hat{\lambda_{32}} \\ \hat{\lambda_{41}} & \hat{\lambda_{42}} \\ \hat{\lambda_{51}} & \hat{\lambda_{52}} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \\ c_{41} & c_{42} \\ c_{51} & c_{52} \end{bmatrix} \begin{bmatrix} \sqrt{\theta_1} & 0 \\ 0 & \sqrt{\theta_2} \end{bmatrix} = \begin{bmatrix} \sqrt{\theta_1}c_{11} & \sqrt{\theta_2}c_{12} \\ \sqrt{\theta_1}c_{21} & \sqrt{\theta_2}c_{22} \\ \sqrt{\theta_1}c_{31} & \sqrt{\theta_2}c_{32} \\ \sqrt{\theta_1}c_{41} & \sqrt{\theta_2}c_{42} \\ \sqrt{\theta_1}c_{51} & \sqrt{\theta_2}c_{52} \end{bmatrix} \tag{1.25}$$

Iz gornje jednadžbe (1.25) vidi se odakle dolazi ime ove metode. Kolone od $\hat{\mathbf{\Lambda}}$ su proporcionalne sa svojstvenim vektorima od $\mathbf{S}$, dakle težine na j-tom faktoru su proporcionalne koeficijentima uz j-tu glavnu komponentu.

Faktori su stoga povezani sa prvih m glavnih komponenti te je stoga za očekivati da je interpretacija faktora ista kao interpretacija glavnih komponenti. No, nakon rotacije težina, interpretacija faktora je najčešće drugačija. Znamo da je i-ti dijagonalni element od $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T$ jednak sumi kvadrata i-tog reda matrice $\hat{\mathbf{\Lambda}}$, odnosno $\hat{\lambda}_i^T\hat{\lambda}_i = \sum_{j=1}^m \hat{\lambda}_{ij}^2$. Sada definiramo:

$$\hat{\psi}_i = s_{ii} - \sum_{j=1}^{m} \hat{\lambda}_{ij}^2 \tag{1.26}$$

te imamo

$$\mathbf{S} \cong \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}} \tag{1.27}$$

gdje je $\hat{\mathbf{\Psi}}$ dijagonalna matrica. U (1.27) su varijance na dijagonali od $\mathbf{S}$ modelirane egzaktno dok su nedijagonalne varijance aproksimirane. Ovo je jedan od izazova faktorske analize.

U ovoj metodi procjene, suma kvadrata redova i stupaca matrice $\hat{\mathbf{\Lambda}}$ je jednaka težinama i svojstvenim vrijednostima, respektivno. Ovo se lako pokaže. Iz (1.26) i-ti komunalitet je procijenjen sa :

$$\hat{h}_i^2 = \sum_{j=1}^{m} \hat{\lambda}_{ij}^2 \tag{1.28}$$

što je suma kvadrata i-tog reda matrice $\hat{\mathbf{\Lambda}}$. Suma kvadrata j-te kolone matrice $\hat{\mathbf{\Lambda}}$ jednaka je j-toj svojstvenoj vrijednosti matrice $\mathbf{S}$:

$$\sum_{i=1}^{p} \hat{\lambda}_{ij}^2 = \sum_{i=1}^{p} (\sqrt{\theta_j} c_{ij})^2 = \theta_j \sum_{i=1}^{p} c_{ij}^2 = \theta_j \tag{1.29}$$

pri čemu je korištena činjenica da normalizirani svojstveni vektori (kolone matrice $\mathbf{C}$) imaju duljinu jednaku jedan. Sada iz (1.26) i (1.28) imamo da je varijanca i-te varijable particionirana na dio koji dolazi od faktora i dio koji dolazi od same te varijable:

$$s_{ii} = \hat{h}_i^2 + \hat{\psi}_i = \hat{h}_{i1}^2 + \hat{h}_{i2}^2 + ... + \hat{h}_{im}^2 + \hat{\psi}_i \tag{1.30}$$

Vidimo dakle, da j-ti faktor pridonosi $\hat{h}_{ij}^2$ u izrazu od s_{ii}. Ulogu u ukupnoj uzoračkoj varijanci, $tr(\mathbf{S}) = s_{11} + s_{22} + ... + s_{pp}$, j-ti faktor ima:

$$\textit{varijanca j-tog faktora} = \sum_{i=1}^{p} \hat{\lambda}_{ij}^2 = \hat{\lambda}_{1j}^2 + \hat{\lambda}_{2j}^2 + ... + \hat{\lambda}_{pj}^2 \tag{1.31}$$

što je suma kvadrata težina u j-tom stupcu matrice $\hat{\mathbf{\Lambda}}$. Po (1.29) imamo da je to ekvivalentno j-toj svojstvenoj vrijednosti θ_j. Na kraju imamo:

$$\frac{\sum_{i=1}^{p} \hat{\lambda}_{ij}^2}{tr(\mathbf{S})} = \frac{\theta_j}{tr(\mathbf{S})} \tag{1.32}$$

Ukoliko varijable nisu proporcionalne, možemo se koristiti sa standardiziranim varijablama i korelacijskom matricom $\mathbf{R}$. Svojstvene vrijednosti i svojstveni vektori od $\mathbf{R}$ se koriste umjesto svojstvenih vrijednosti i svojstvenih vektora matrice $\mathbf{S}$ u (1.24) kako bi došli do procjena težina.
U praksi se češće koristi matrica $\mathbf{R}$ i većina statističkih programskih paketa koristi tu matricu kao standardnu matricu. Budući da je naglasak u faktorskoj analizi na reprodukciji kovarijanci ili korelacija više nego varijanci, korištenje matrice $\mathbf{R}$ je prikladnije u faktorskoj analizi nego u analizi glavnih komponenti. Također, u praksi $\mathbf{R}$ daje bolje rezultate od $\mathbf{S}$. Ukoliko radimo s matricom $\mathbf{R}$ onda (1.32) postaje:

$$\frac{\sum_{i=1}^{p} \hat{\lambda}_{ij}^2}{tr(\mathbf{R})} = \frac{\theta_j}{p} \tag{1.33}$$

gdje je p broj varijabli.

Adekvatnost faktorskog modela možemo dobiti uspoređujući strane u (1.27). Matrica pogreške:

$$\mathbf{E} = \mathbf{S} - (\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}}) \tag{1.34}$$

ima nule na dijagonali, a nedijagonalne elemente nenegativne. Slijedeća nejednakost daje ogradu na veličinu elemenata matrice **E**:

$$\sum_{ij} e_{ij}^2 \leq \theta_{m+1}^2 + \theta_{m+2}^2 + ... + \theta_p^2 \tag{1.35}$$

odnosno, suma kvadrata elemenata matrice **E** je najviše jednaka sumi kvadrata "zanemarenih" svojstvenih vrijednosti matrice **S**. Ukoliko su svojstvene vrijednosti male, reziduali u matrici greške $\mathbf{S} - (\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}})$ su mali i model je dobar.

1.3.2 Metoda glavnih faktora

Cilj i ove metode je kao i kod PCA, ekstrahirati maksimalnu ortogonalnu varijancu iz skupa podataka sa svakim slijedećim faktorom. U metodi glavnih komponenti kako bi procijenili težine, zanemarili smo $\mathbf{\Psi}$ i faktorizirali $\mathbf{S}$ ili $\mathbf{R}$. Metoda glavnih faktora koristi inicijalnog procijenitelja $\hat{\mathbf{\Psi}}$ i faktore $\mathbf{S} - \hat{\mathbf{\Psi}}$ ili $\mathbf{R} - \hat{\mathbf{\Psi}}$ za izračun:

$$\begin{aligned} \mathbf{S} - \hat{\mathbf{\Psi}} &= \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T, \\ \mathbf{R} - \hat{\mathbf{\Psi}} &= \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T \end{aligned} \tag{1.36}$$

gdje je $\hat{\mathbf{\Lambda}}$ $p \times m$ matrica koja se računa kao u (1.24) koristeći svojstvene vrijednosti i svojstvene vektore od $\mathbf{S} - \hat{\mathbf{\Psi}}$ ili $\mathbf{R} - \hat{\mathbf{\Psi}}$. Sada imamo da je i-ti dijagonalni element od $\mathbf{S} - \hat{\mathbf{\Psi}}$ dan sa $s_{ii} - \hat{\psi}_i$, što je ujedno i i-ti komunalitet, $\hat{h}_i^2 = s_{ii} - \hat{\psi}_i$. Analogno, dijagonalni elementi od $\mathbf{R} - \hat{\mathbf{\Psi}}$ su komunaliteti $\hat{h}_i^2 = 1 - \hat{\psi}_i$. Matrična forma od $\mathbf{S} - \hat{\mathbf{\Psi}}$ i $\mathbf{R} - \hat{\mathbf{\Psi}}$ je:

$$\mathbf{S} - \hat{\mathbf{\Psi}} = \begin{bmatrix} \hat{h}_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & \hat{h}_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & \hat{h}_p^2 \end{bmatrix}, \tag{1.37}$$

$$\mathbf{R} - \hat{\mathbf{\Psi}} = \begin{bmatrix} \hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & \hat{h}_2^2 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & \hat{h}_p^2 \end{bmatrix}. \tag{1.38}$$

Popularna inicijalna procjena komunaliteta u $\mathbf{R} - \hat{\mathbf{\Psi}}$ je $\hat{h}_i^2 = R_i^2$, odnosno kvadrat višetruke korelacije između y_i i drugih $p - 1$ varijabli. Također, to možemo pisati pišemo:

$$\hat{h}_i^2 = R_i^2 = 1 - \frac{1}{r^{ii}} \tag{1.39}$$

gdje je r^{ii} i-ti dijagonalni element matrice $\mathbf{R}^{-1}$.

Za $\mathbf{S} - \hat{\mathbf{\Psi}}$ je analogno (1.39) inicijalna procjena komunaliteta :

$$\hat{h}_i^2 = s_{ii} - \frac{1}{s^{ii}} \tag{1.40}$$

gdje je s_{ii} i-ti dijagonalni element matrice $\mathbf{S}$, a s^{ii} i-ti dijagonalni element matrice $\mathbf{S}^{-1}$. Može se pokazati da je (1.40) ekvivalentno sa:

$$\hat{h}_i^2 = s_{ii} - \frac{1}{s^{ii}} = s_{ii} R_i^2 \tag{1.41}$$

Da bi koristili (1.39) i (1.40), matrice $\mathbf{S}$ i $\mathbf{R}$ ne smiju biti singularne, jer kao takve nemaju inverz. Ukoliko je $\mathbf{R}$ singularna kao procjenu komunaliteta možemo koristiti apsolutnu vrijednost ili kvadrat najveće korelacije u i-tom redu matrice $\mathbf{R}$.
Jednom kada imamo procijenjene komunalitete, računamo svojstvene vrijednosti i svojstvene vektore od $\mathbf{S} - \hat{\mathbf{\Psi}}$ i $\mathbf{R} - \hat{\mathbf{\Psi}}$ te koristimo (1.24) kako bi dobili procjene težina, $\hat{\mathbf{\Lambda}}$. Onda pomoću stupaca i redaka matrice $\hat{\mathbf{\Lambda}}$ računamo nove svojstvene vrijednosti (objašnjeni dio varijance) i komunalitete.
Suma kvadrata j-tog stupca matrice $\hat{\mathbf{\Lambda}}$ je j-ta svojstvena vrijednost od $\mathbf{S} - \hat{\mathbf{\Psi}}$ ili $\mathbf{R} - \hat{\mathbf{\Psi}}$, a suma kvadrata i-tog reda matrice $\hat{\mathbf{\Lambda}}$ je komunalitet od y_i. Proporcija varijace koja je objašnjena sa j-tim faktorom je:

$$\frac{\theta j}{tr(\mathbf{S} - \hat{\mathbf{\Psi}})} = \frac{\theta_j}{\sum_{i=1}^{p} \theta_i} \tag{1.42}$$

odnosno,

$$\frac{\theta j}{tr(\mathbf{R} - \hat{\mathbf{\Psi}})} = \frac{\theta_j}{\sum_{i=1}^{p} \theta_i} \tag{1.43}$$

gdje je θ_j j-ta svojstvena vrijednost od $\mathbf{S} - \hat{\mathbf{\Psi}}$ ili $\mathbf{R} - \hat{\mathbf{\Psi}}$. Matrice $\mathbf{S} - \hat{\mathbf{\Psi}}$ i $\mathbf{R} - \hat{\mathbf{\Psi}}$ nisu nužno pozitivno semidefinitne i često imaju male pozitivne svojstvene vrijednosti. U takvom slučaju, kumulativna proporcija varijance će preći vrijednost od 1 i zatim pasti prema 1 kako se nadodaju negativne svojstvene vrijednosti.

1.3.3 Iterativna metoda glavnih faktora

Metoda glavnih faktora se lako može iterirati kako bi poboljšali procjene komunaliteta. Nakon što izračunamo $\hat{\mathbf{\Lambda}}$ iz $\mathbf{S} - \hat{\mathbf{\Psi}}$ ili $\mathbf{R} - \hat{\mathbf{\Psi}}$ u (1.36) koristeći inicijalalne procjene komunaliteta , možemo izračunati nove procjene komunaliteta iz težina od $\hat{\mathbf{\Lambda}}$ koristeći (1.28):

$$\hat{h}_i^2 = \sum_{j=1}^{m} \hat{\lambda}_{ij}^2 \tag{1.44}$$

Te novodobivene vrijednosti (procjene komunaliteta) $\hat{h}_i^2$ se supstituiraju na dijagonalu od $\mathbf{S} - \hat{\mathbf{\Psi}}$ ili $\mathbf{R} - \hat{\mathbf{\Psi}}$, iz kojih izračunamo novu vrijednost od $\hat{\mathbf{\Lambda}}$ koristeći (1.24).
Opisani proces se ponavlja dokle procjene komunaliteta ne iskonvergiraju, no treba imati na umu da za neke podatke, iterativan proces neće iskonvergirati. Nakon što imamo konvergenciju, koristimo svojstvene vrijednosti i svojstvene vektore od $\mathbf{S} - \hat{\mathbf{\Psi}}$ ili $\mathbf{R} - \hat{\mathbf{\Psi}}$ u (1.24) kako bi izračunali težine.

Metoda glavnih faktora i iterativna metoda glavnih faktora često daju rezultate blizu rezultatima medote glavnih komponenata ako je bilo koji od sljedećih uvjeta istinit:

1. Korelacije su dovoljno velike, što rezultira malom vrijednosti od m
2. Broj varijabli, p, je velik

Loša strana iterativne metode je što ponekad dovede do procjena komunaliteta $\hat{h}_i^2$ koje su veće od 1 (kad se faktorizira $\mathbf{R}$). Takav rezultat je poznat kao *Heywood case* (Heywood 1931). Ukoliko je $\hat{h}_i^2 > 1$, onda je po (1.26) i (1.28) $\hat{\psi}_i < 0$, što je očito krivo jer ne možemo imati negativnu varijancu.
U takvim situacijama, kada komunalitet pređe vrijednost 1, iterativni proces bi trebao stati, te bi program trebao javiti da se ne može doći do rješenja.
Neki statistički programski paketi imaju opciju nastavka s iteracijama tako što se komunaliteti postave na 1 u svim slijedećim iteracijama. Rješenje u kojem je $\hat{\psi}_i = 0$ je upitno jer implicira na egzaktnu ovisnost varijabli o faktorima, mogući, ali rijetki ishod.

1.3.4 Metoda maksimalne vjerodostojnosti

Ako pretpostavimo da opažanja $\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n}$ čine slučajan uzorak iz $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, onda $\mathbf{\Lambda}$ i $\mathbf{\Psi}$ mogu biti procijenjeni metodom maksimalne vjerodostojnosti. Može se pokazati da procijenitelji $\hat{\mathbf{\Lambda}}$ i $\hat{\mathbf{\Psi}}$ zadovoljavaju slijedeće jednadžbe:

$$\begin{aligned} \mathbf{S}\hat{\mathbf{\Psi}}\hat{\mathbf{\Lambda}} &= \hat{\mathbf{\Lambda}}(\mathbf{I} + \hat{\mathbf{\Lambda}}^T\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{\Lambda}}), \\ \hat{\mathbf{\Psi}} &= diag(\mathbf{S} - \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T), \\ \hat{\mathbf{\Lambda}}^T\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{\Lambda}} &\text{ je dijagonalna.} \end{aligned} \tag{1.45}$$

Ove jednadžbe moraju se rješavati iterativno i u praksi postupak možda ne iskonvergira ili ishodi rezultatom poznatim kao *Heywood case* [2]. Također, proporcija varijance objašnjene faktorima, kao što je dano u 1.32 i 1.33, neće nužno biti u padajućem poretku u ovoj metodi kao što je za faktore dobivene metodom glavnih komponenata ili metodom glavnih faktora.

1.4 Odabir broja faktora

Cilj istraživanja koja se koriste faktorskom analizom je smanjiti veliki broj varijabli na manji broj faktora. Uz odabir broja faktora veže se niz pitanja. Koliko se pouzdanih i interpretabilnih faktora nalazi u promatranom skupu podataka? Da li je dobiven broj faktora pouzdan i postoji li možda još pouzdanih faktora?
Uključivanje većeg broja faktora u rješenje poboljšava sličnost između promatrane i reproducirane matrice korelacija, stoga je adekvatnost ekstrakcije vezana uz odabir broja faktora. S druge strane, što je veći broj faktora ekstrahiran to je rješenje slabije. Kako bi objasnili čitavu kovarijancu u skupu podataka trebali bi imati jednak broj faktora kao i broj promatranih varijabli (PCA). Dakle, jasno je da je potrebno učiniti kompromis, odnosno želimo zadržati dovoljan broj faktora za adekvatno odgovaranje modela podacima, ali ne i previše. Odabir broja faktora je obično delikatniji od odabira tehnike za ekstrakciju i rotaciju ili vrijednosti komunaliteta.

Za odabir broja faktora, m, postoji nekoliko kriterija. Ovdje će biti uzeta u obzir četiri kriterija.

1. Biramo m koji je jednak broju faktora koji je potreban da postignemo određeni postotak objašenje varijance, recimo 80% ukupne varijance tr($\mathbf{S}$) ili tr($\mathbf{R}$).
2. Biramo m koji je jednak broju svojstvenih vrijednosti većih od prosječne svojstvene vrijednosti. Za $\mathbf{R}$ prosjek je 1, a za $\mathbf{S}$ prosjek je $\frac{\sum_{j=1}^{p} \theta_j}{p}$.
3. Koristeći *scree test*, odnosno metodu lakta, koji je baziran na grafičkom prikazu svojstvenih vrijednosti od $\mathbf{R}$ ili $\mathbf{S}$. Ukoliko graf oštro padne te nakon toga imamo ravnu liniju malog nagiba, biramo m koji je jednak broju svojstvenih vrijednosti koje se nalaze prije ravne linije blagog nagiba.
4. Testiramo hipotezu da je m dobar broj faktora, odnosno testiramo nultu hipotezu $H_0 : \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$, gdje je $\mathbf{\Lambda}$ $p \times m$ matrica.

Metoda jedan se primijenjuje u metodi glavnih komponenti. Po (1.32) proporcija ukupne uzoračke varijance uzrokovane j-tim faktorom iz $\mathbf{S}$ je jednako $\frac{\sum_{i=1}^{p} \hat{\lambda}_{ij}^2}{tr(\mathbf{S})}$. Odgovarajuća proporcija iz $\mathbf{R}$ je jednaka $\frac{\sum_{i=1}^{p} \hat{\lambda}_{ij}^2}{p}$.

Doprinos svih m faktora u tr($\mathbf{S}$) ili p je stoga $\sum_{i=1}^{p}\sum_{j=1}^{m}\hat{\lambda}_{ij}^2$, što je suma kvadrata svih elemenata matrice $\hat{\mathbf{\Lambda}}$. Za metodu glavnih komponenti , iz (1.28) i (1.29) , vidimo da je ta suma također jednaka sumi prvih m svojstvenih vrijednosti ili sumi svih p komunaliteta:

$$\sum_{i=1}^{p}\sum_{j=1}^{m}\hat{\lambda}_{ij}^2 = \sum_{i=1}^{p}\hat{h}_i^2 = \sum_{i=1}^{p}\theta_j \tag{1.46}$$

Stoga biramo m dovoljno velik tako da suma komunaliteta ili suma svojstvenih vrijednosti (objašnjena varijabilnost) čini relativno veliki dio od tr($\mathbf{S}$) ili p.

Metoda 1 može se proširiti na metodu glavnih faktora, gdje se procjene komunaliteta koriste za formiranje $\mathbf{S} - \hat{\mathbf{\Psi}}$ ili $\mathbf{R} - \hat{\mathbf{\Psi}}$. Budući da $\mathbf{S} - \hat{\mathbf{\Psi}}$ i $\mathbf{R} - \hat{\mathbf{\Psi}}$ često imaju negativne svojstvene vrijednosti i budući da je m broj između 1 i p, kumulativna proporcija svojstvenih vrijednosti $\frac{\sum_{j=1}^{m}\theta_j}{\sum_{j=1}^{p}\theta_j}$ će preći vrijednost 1.0 i onda biti reducirana na 1.0 kako se zbrajaju negativne svojstvene vrijednosti. Stoga će postotak jednak npr. 80% biti dostignut za manju vrijednost m nego što i to bilo kod $\mathbf{S}$ ili $\mathbf{R}$ i bolja strategija bi mogla biti odabir m koji je jednak vrijednosti za koju postotak prvi put prijeđe 100%.

U iterativnoj metodi m se bira prije iteriranja i $\sum_{i=1}\hat{h}_i^2$ se računa nakon iteracija kao $\sum_{i=1}\hat{h}_i^2 = tr(\mathbf{S} - \hat{\mathbf{\Psi}})$. Kako bi odlučili koji m odabrati prije početka iteriranja možemo se poslužiti sa prije navedenim metodama ili svojstvenim vrijednostima od $\mathbf{S}$ ili $\mathbf{R}$ kao u metodi glavnih komponenti.

Metoda 2 je standardna metoda u mnogim statističkim programskim paketima. Iako je heuristički bazirana, u praksi se pokazala dobrom. Predložena je varijacija metode 2 kada se koristi za $\mathbf{R} - \hat{\mathbf{\Psi}}$, u kojoj se m bira tako da je on jednak broju pozitivinih svojstvenih vrijednosti. Loša strana upravo navedene varijacije je što ta metoda često ishodi velikom broju faktora, budući da će suma pozitivih svojstvenih vrijednosti biti veća od sume komunaliteta.

Metoda 3, odnosno metoda lakta, se također pokazala dobrom u praksi.

U metodi 4 želimo testirati hipotezu:

$$H_0 : \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi} \tag{1.47}$$

naspram alternative

$$H_0 : \mathbf{\Sigma} \neq \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi} \tag{1.48}$$

Testna statistika je:

$$\left(n - \frac{2p + 4m + 11}{6}\right) ln\left(\frac{|\ \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}\ |}{|\ \mathbf{\Psi}\ |}\right), \tag{1.49}$$

što ima aproksimativno χ^2_υ razdiobu kada je nulta hipoteza istinita
te gdje je $\upsilon = \frac{1}{2}\left[(p-m)^2 - p - m\right]$. Odbacivanje nulte hipoteze implcira da je odabrani m premali i da je potrebno uzeti više faktora.

U praksi, kada je n velik, test u metodi 4 često pokazuje da je više faktora značajno nego ostale tri metode, stoga možemo vrijednost m koju dobijemo metodom 4 smatratim gornjom međom za stvarni broj potrebnih faktora u praktičnoj primjeni.

Za mnoge skupine podataka odabir vrijednosti m neće biti očit. Ta neodlučnost zato mnoge statističare ostavlja skeptičnima prema faktorskoj analizi. O valjanosti faktorske analize nešto kasnije.

Kada se model koji smo dobili faktorskom analizom dobro podudara s podcima, prve tri navedene metode gotovo uvijek daju istu vrijednost broja m, te u takvim situacijama nema sumnje u odabir broja m. Dakle, za "dobar" podatkovni skup cijeli proces faktorske analize i odabira m postaje objektivniji.

1.5 Rotacije

Rezultat ekstrakcije faktora koji nije kombiniran s rotacijama je često težak za interpretaciju bez obzira koja je metoda ekstrakcije korištena. Nakon ekstrakcije, koristimo se rotacijom kako bi poboljšali interpretabilnost i korist rješenja. Važno je shvatiti da se rotacije ne koriste za poboljšavanje kvalitete podudarnosti između promatrane i reproducirane korelacijske matrice zato što su sva ortogonalno rotirana rješenja matematički ekvivalentna jedna drugom i jednaka su rješenju prije primjene rotacija.

Isto kao što različite metode ekstrakcije daju slične rezultate za "dobre" skupove podataka, tako i različite metode rotacija daju slične rezultate ukoliko je uzorak korelacija u skupu podataka relativno jasan. Potrebno je odlučiti se da li ćemo primijeniti ortogonalnu ili kosu rotaciju. U ortogonalnoj rotaciji faktori su nekorelirani, a rješenja dobivena tom rotacijom su jednostavna za interpretaciju i izvještavanje rezultata, ali ne odražavaju pravu stvarnost osim ako je istraživač siguran da su latentni procesi gotovo nezavisni. Istraživač koji sumnja da su latentni, pozadinski procesi korelirani koristi se kosom rotacijom. U kosoj rotaciji faktori će možda biti korelirani, imamo konceptualne prednosti, a nedostatke u interpretaciji i izvještavanju rezultata.

Kao što znamo iz poglavlja 1.2, težine faktora (retci matrice $\mathbf{\Lambda}$) populacijskog modela su jedinstvene do na množenje ortogonalnom matricom koja rotira težine. Rotirane težine imaju sačuvana svojstva originalnih težina; reproduciraju kovarijacijsku matricu i zadovoljavaju sve početne prepostavke. Procijenjena matrica težina $\hat{\mathbf{\Lambda}}$ također može biti rotirana tako da se dobije $\hat{\mathbf{\Lambda}}^* = \hat{\mathbf{\Lambda}}\mathbf{T}$, pri čemu je $\mathbf{T}$ ortogonalna matrica. Kako vrijedi $\mathbf{T}\mathbf{T}^T = \mathbf{I}$ to rotirane težine daju istu procjenu kovarijacijske matrice kao i prije:

$$\mathbf{S} \cong \hat{\mathbf{\Lambda}}^*\hat{\mathbf{\Lambda}}^{*\prime} + \hat{\mathbf{\Psi}} = \hat{\mathbf{\Lambda}}\mathbf{T}\mathbf{T}^T\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}} \tag{1.50}$$

Geometrijski, težine u i-tom redu matrice $\hat{\mathbf{\Lambda}}$ sudjeluju u kreiranju koordinata točke u prostoru težina koja odgovara y_i. Rotacija p točaka daje im nove koordinata u odnosu na nove osi (faktore), ali ostavlja njihova geometrijska stvojstva netaknuta. U cilju nam je pronaći novi okvir/prostor u kojem su faktori interpretabilniji. Još jedan od ciljeva rotiranja je postaviti osi blizu što je više točaka moguće. Ukoliko imamo nakupine točaka (grupiranja y-a), želimo pomaknuti osi tako da prolaze kroz ili blizu tih nakupina. Takav pristup pridružuje svaku grupu varijabli nekom faktoru (osi) i čini interpretaciju objektivnijom.

Ukoliko možemo doći do rotacije u kojoj će svaka točka biti blizu nekoj od osi onda svaka varijabla ima veliku težinu koja odgovara pripadnom faktoru (osi) i ima male težine na ostalim faktorima. Takav ishod naziva se *jednostavna struktura* i njegova interpretacija je uvelike pojednostavljena. Jednostavno promatramo koje su varijable pridužene kojem faktoru i faktor definiramo ili imenujemo sukladno tome.

Kako bi identificirali prirodna grupiranja varijabli tražimo rotaciju u kojoj varijable imaju veliku težinu na samo jednom faktoru. Broj faktora na kojima varijabla ima srednju do veliku težinu naziva se *kompleksnost* varijable. U idealnim uvjetima koji su malo prije spomenuti koje nazivamo *jadnostavim strukturama* sve varijable imaju kompleksnost jednaku 1. U takvim situacijama varijable su grupirane u grupe koje odgovaraju faktorima.

Promatrat ćemo dva tipa rotacija: ortogonalnu i kosu rotaciju. Gore spomenuta rotacija sa ortogonalnom matricom je ortogonalna rotacija; originalne okomite osi se rotiraju i ostaju okomite nakon rotacije. U ortogonalnoj rotaciji kutevi i udaljenosti ostaju sačuvani te komunaliteti nepromijenjeni. U kosoj rotaciji nema zahtjeva da osi moraju nakon rotacije ostati okomite te su time slobodne prolaziti bliže nakupinama točaka.

1.5.1 Ortogonalna rotacija

Kao što je spomenuto gore, ortogonalna rotacija čuva komunalitete. Razlog tomu je što su retci matrice $\hat{\mathbf{\Lambda}}$ rotirani i udaljenost od ishodišta ostaje nepromijenjena, što su po (1.28) komunaliteti. Ono što će se promijeniti je varijanca zbog j-tog faktora (1.31) kao i pripadne proporcije (1.32) i (1.33). U narednim potpoglavljima ćemo diskutirati dva pristupa ortogonalnim rotacijama.

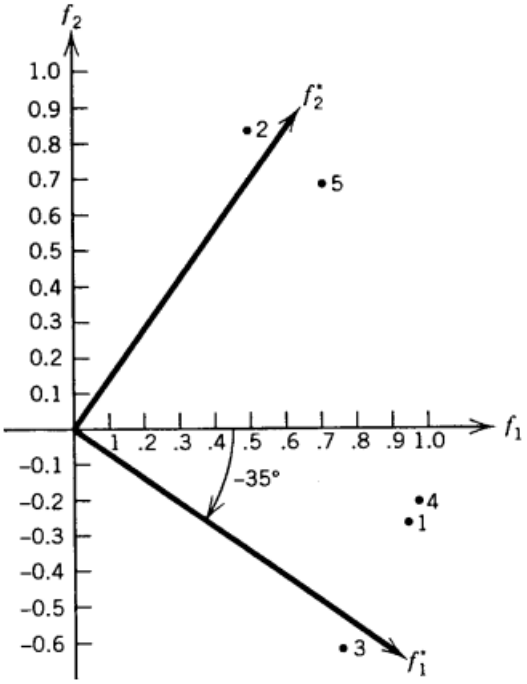
1.5.1.1 Grafički pristup

Ukoliko imamo samo dva faktora ($m = 2$), možemo koristiti *grafičku* rotaciju baziranu na vizualnim kontrolama prikaza težina faktora. U ovom slučaju, retci matrice $\hat{\mathbf{\Lambda}}$ su parovi težina $(\hat{\lambda}_{i1}, \hat{\lambda}_{i2})$, $i = 1, 2, ..., p$, što odgovara $y_1, y_2, ..., y_p$ varijablama. Biramo kut ϕ za koji osi mogu biti rotirane tako da budu bliže grupiranim točkama. Nove rotirane težine

$(\hat{\lambda}^*_{i1}, \hat{\lambda}^*_{i2})$ prikazujemo na grafu uz pomoć $\hat{\mathbf{\Lambda}}^* = \hat{\mathbf{\Lambda}}\mathbf{T}$ pri čemu je:

$$\mathbf{T} = \begin{bmatrix} cos\phi & -sin\phi \\ sin\phi & cos\phi \end{bmatrix} \tag{1.51}$$

Primjer 1.5.1. Na donjoj slici vidimo prikaz pet pari težina $(\hat{\lambda}_{i1}, \hat{\lambda}_{i2})$ koji odgovaraju pet imaginarnih varijabli. Ortogonalna rotacija od ϕ = -35°rotira osi (faktore) bliže dvama nakupinama točaka (varijabli). Nakon rotacije obje nakupine varijabli odgovaraju puno bliže novim faktorima.



Slika 1.1: Ortogonalna rotacija [2]

1.5.1.2 Varimax rotacija

Grafički pristup rotaciji generalno je limitiran na $m = 2$. Za $m > 2$ su predožene brojne analitičke metode. Najpopularnija metoda je *Varimax* tehnika. Ona traži rotirane težine koje maksimiziraju varijancu kvadrata težina u svakom stupcu matrice $\hat{\mathbf{\Lambda}}^*$. Ako su težine u stupcu približno jednake, varijanca je blizu 0. Kako se kvadrati težina približavaju 0 i 1 (ako se faktorizira **R**), varijanca će se približiti maksimumu. Dakle, varimax tehnika

nastoji učiniti težine ili velikima ili malima kako bi došla do interpretacije.

Varimax tehnika ne garantira da će sve varijable imati veliku težinu na samo jednom faktoru. Zapravo, niti jedna metoda ovo ne može postići za sve moguće podatkovne skupove. Konfiguracija točaka u prostoru težina ostaje fiksirana; sve što mi radimo je rotiranje osi (fakora) kako bi bili bliže što je više točaka moguće. U mnogim slučajevima točke nisu u lijepim nakupinama te je nemoguće zarotirati osi da budu blizu svim točkama. Ovaj je problem sakriven u odabiru m. Ako je m promijenjen, koordinate se mijenjaju te se time i relativne pozicije točaka mijenjaju.

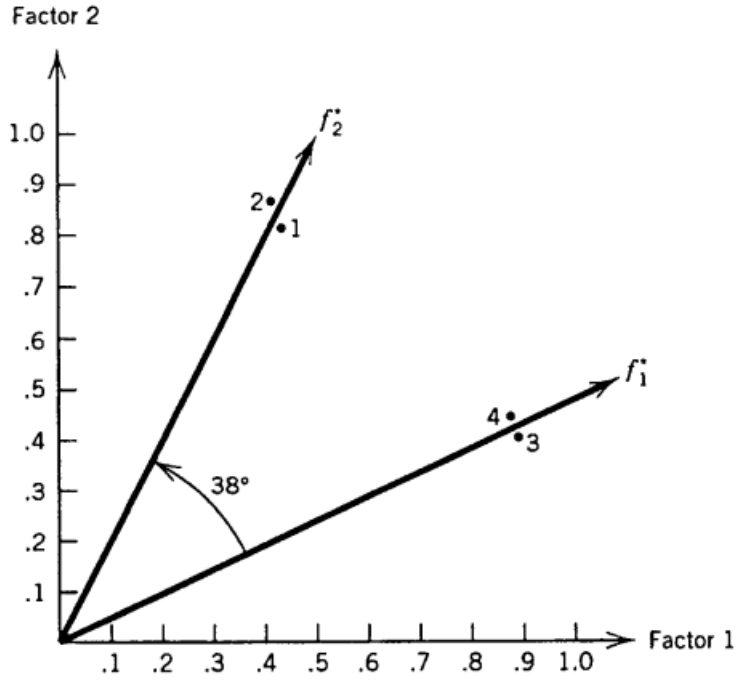
Varimax metoda je dostupna u gotovo svim statističkim programskim paketima koji podržavaju faktorsku analizu. Nakon primjene metode izlaz koji dobijemo u većini statističkih programa je rotirana matrica težina $\hat{\mathbf{\Lambda}}^*$, udio objašnjene varijabilnosti (suma kvadrata pojedinih stupaca matrice $\hat{\mathbf{\Lambda}}^*$), komunalitete (suma kvadrata pojedinih redaka matrice $\hat{\mathbf{\Lambda}}^*$) te ortogonalna matrica $\mathbf{T}$ koja služi kako bi dobili $\hat{\mathbf{\Lambda}}^*$, odsnosno $\hat{\mathbf{\Lambda}}^* = \hat{\mathbf{\Lambda}}\mathbf{T}$.

1.5.2 Kosa rotacija

Kosa rotacija u terminima faktorske analize je ona rotacija u kojoj osi ne ostaju okomite. Umjesto ortogonalne matrice koju smo koristili u poglavlju nejedinstvenost težina, kosa rotacija koristi nesingularnu matricu $\mathbf{Q}$ kako bi izračunali $\mathbf{f}^* = \mathbf{Q}^T\mathbf{f}$. Vrijedi:

$$cov(\mathbf{f}^*) = \mathbf{Q}^T\mathbf{I}\mathbf{Q} = \mathbf{Q}^T\mathbf{Q} \neq \mathbf{I} \tag{1.52}$$

Dakle, novi faktori su korelirani. Budući da udaljenosti i kutevi nisu sačuvani komunaliteti od $\mathbf{f}^*$ su drugačiji od komunaliteta od $\mathbf{f}$. Budući da u kosoj rotaciji nemamo zahtjev da osi ostaju okomite, možemo se lakše približiti nakupinama točaka sa osima (ukoliko nakupine postoje). Na slijedećoj slici, koja služi kao primjer, imamo grafički prikaz kose rotacije od 38°, odnosno kut između novih osi je 38°. Na slici vidimo da koso rotirane osi prolaze puno bliže nakupinama točaka te da su nove težine jako blizu 0 i 1. Međutim interpretacija se nebi promijenila budući da su iste točke (varijable) koje su povezane sa kosim osima, povezane i s okomitim osima.



Slika 1.2: Kosa rotacija [2]

Brojne analitičke metode za postizanje kose rotacije su predložene i dostupne u programskim paketima. Izlaz koji se dobije kao rezultat tih metoda su najčešće: *matrica*

uzoraka, *matrica strukture* i matrica korelacija među kosim faktorima (osima). Za interpretaciju najčešće se koristi *matrica uzoraka*. Težine u retcima matrice uzoraka su prirodne koordinate točaka (varijabli) na kosim osima i služe kao koeficijenti u modelu koji povezuje varijable s faktorima.

Jedna od koristi kose rotacije je provjera ortogonalnosti faktora. Ortogonalnost između originalnih faktora je nametnuta modelom i sačuvana ortogonalnim rotacijama. Ako kosa rotacija rezultira matricom korelacija koja je približno dijagonalna možemo biti sigurniji da su faktori zaista ortogonalni.

1.5.3 Interpretacija

U poglavljina 4.1, 4.2 i 4.3 komentirali smo tipove i korist rotacija faktora (osi) kao pomoć u interpretaciji. Naš je cilj postići jednostavnu strukturu u kojoj svaka varijabla ima veliku težinu na samo jednom faktoru, dok na ostalim faktorima postiže male težine. U praksi, često ne uspjevamo ostvariti taj cilj, ali nam u tome pomažu rotacije uz pomoć kojih dobivamo težine koje su bliže ostvarivanju željene jednostavne strukture.

Ovdje ćemo predstaviti generalne smjernice za interpretaciju faktora ispitivajući matricu rotiranih težina. U svakom retku matrice težina, krenuvši slijeva te se kretajući nadesno detektiramo, po apsolutnoj vrijednosti, najveću vrijednost, odnosno težinu. Ukoliko je najveća težina značajne veličine (subjektivna odluka, više o tome u idućem paragrafu) pocrtamo ju ili zaokružimo. Ovo radimo za svaku od p varijabli. Ukoliko ima u jednom retku više značajnih težina koje će biti uzete u obzir, interpretacija postaje manje jednostavna. S druge strane moguće je i da postoje varijable s malim komunalitetima da se ne pojavljuje niti jedna značajna težina na nekom faktoru. U ovom slučaju istraživač možda želi povećati broj faktora i ponovno pokrenuti program kako bi se takve varijable pridružile nekom novom faktoru.

Kako bi procijenili značajnost težine na nekom faktoru, vrijednost od 0.3 je predložena. Za brojne uspješne primjene vrijednost od 0.3 je premala i rezultirala je varijablama kompleksnosti veće od jedan. Zbog navedenog, vrijednosti od 0.5 i 0.6 su korisnije u praksi.

Nakon što smo identificirali potencijalno značajne težine, istraživač nastoji otkriti neko značenje faktora te im u idealnim slučajevima dodijeliti smisleno ime. U mnogim situacijama, grupiranja nisu logična i analiza može biti sprovedena ponovno, mijenjajući m, prilagođavajući razinu značajnosti težina, koristeći novu metodu za procjenu težina ili korištenje druge vrste rotacije [2].

1.6 Faktorski score-ovi

U mnogim primjenama, cilj istraživača je utvrditi da li model faktorske analize odgovara podacima i identificirati faktore. Postoje primjene u kojima istraživač želi dobiti $faktorske\ score-ove\ \hat{\mathbf{f}}_\mathbf{i} = (\hat{f}_{i1}, \hat{f}_{i2}, ..., \hat{f}_{im})$, $i = 1, ...n$, koji se definiraju kao procjene of pozadinskih faktorskih vrijednosti za svaku obzervaciju. Dvije su potencijalne upotrebe za tako definirane score-ove:

1. Od interesa nam je analizirati ponašanje obzervacija u terminima faktora
2. Želimo iskoristiti faktosrske score-ove kao ulazne vrijednosti za neku drugu analizu, recimo multivarijatnu analizu varijance (MANOVA)

Budući da f-ovi nisu promatrani/izmjereni moramo ih procijeniti kao funkcije opaženih y-a. Najpopularniji pristup procjeni faktora je baziran na regresiji. Ukratko ćemo opisati ovu metodu i spomenuti što je potrebno učiniti ukoliko je $\mathbf{R}$(ili $\mathbf{S}$) singularna.
Budući da je $E(f_i) = 0$, f-ove povezujemo sa y-ima preko centriranog regresijskog modela:

$$\begin{aligned} f_1 &= \beta_{11}(y_1 - \bar{y}_1) + \beta_{12}(y_2 - \bar{y}_2) + ... + \beta_{1p}(y_p - \bar{y}_p) + \epsilon_1 \\ f_2 &= \beta_{21}(y_1 - \bar{y}_1) + \beta_{22}(y_2 - \bar{y}_2) + ... + \beta_{2p}(y_p - \bar{y}_p) + \epsilon_2 \\ &\vdots \\ f_m &= \beta_{m1}(y_1 - \bar{y}_1) + \beta_{m2}(y_2 - \bar{y}_2) + ... + \beta_{mp}(y_p - \bar{y}_p) + \epsilon_m \end{aligned}$$

što u matričnom obliku zapisujemo kao:

$$\mathbf{f} = \mathbf{B}_1^T(\mathbf{y} - \bar{\mathbf{y}}) + \boldsymbol{\epsilon} \tag{1.53}$$

Imajmo na umu važnu činjenicu, greška $\boldsymbol{\epsilon}$ u (1.53) se razlikuje od greške u originalnom faktorskom modelu $\mathbf{y} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon}$. Naš je pristup prvo procijeniti $\mathbf{B}_1$ te onda iskoristiti predviđenu vrijednost $\hat{\mathbf{f}} = \hat{\mathbf{B}}_1^T(\mathbf{y} - \bar{\mathbf{y}})$ za procjenu od $\mathbf{f}$.
Model (1.53) za pojedinu obzervaciju izgleda:

$$\mathbf{f}_i = \mathbf{B}_1^T(\mathbf{y}_i - \bar{\mathbf{y}}) + \boldsymbol{\epsilon}_i, i = 1, 2, ..., n \tag{1.54}$$

Odnosno, nakon transponiranja imamo:

$$\mathbf{f}_i^T = (\mathbf{y}_i - \bar{\mathbf{y}})^T\mathbf{B}_1 + \boldsymbol{\epsilon}_i^T, i = 1, 2, ..., n \tag{1.55}$$

te tih n jednadžbi mogu biti spojene u jedan model koji onda izgleda kao:

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}_1^T \\ \mathbf{f}_2^T \\ \vdots \\ \mathbf{f}_n^T \end{bmatrix} = \begin{bmatrix} (\mathbf{y}_1 - \bar{\mathbf{y}})^T\mathbf{B}_1 \\ (\mathbf{y}_2 - \bar{\mathbf{y}})^T\mathbf{B}_1 \\ \vdots \\ (\mathbf{y}_n - \bar{\mathbf{y}})^T\mathbf{B}_1 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \boldsymbol{\epsilon}_2^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix} = \begin{bmatrix} (\mathbf{y}_1 - \bar{\mathbf{y}})^T \\ (\mathbf{y}_2 - \bar{\mathbf{y}})^T \\ \vdots \\ (\mathbf{y}_n - \bar{\mathbf{y}})^T \end{bmatrix} \mathbf{B}_1 + \boldsymbol{\Xi} = \mathbf{Y}_c\mathbf{B}_1 + \boldsymbol{\Xi} \tag{1.56}$$

Model (5.4) izgleda kao model centrirane multivarijatne višestruke regresije kojemu je $\mathbf{Y}_c$ na mjestu $\mathbf{x}_c$ na što smo inače navikli. Procjenitelj za $\mathbf{B}_1$ bi bio:

$$\hat{\mathbf{B}}_1 = (\mathbf{Y}_c^T\mathbf{Y}_c)^{-1}\mathbf{Y}_c^T\mathbf{F} \tag{1.57}$$

No, kao što znamo $\mathbf{F}$ nije obzerviran. Kako bi svejedno izračunali $\hat{\mathbf{B}}_1$ model (1.57) zapisujemo u terminima kovarijacijskih matrica

$$\hat{\mathbf{B}}_1 = \mathbf{S}_{yy}^{-1}\mathbf{S}_{yf} \tag{1.58}$$

pri čemu je $\mathbf{S}_{yy}$ reprezentirano sa $\mathbf{S}$; a za $\mathbf{S}_{yf}$ koristimo $\hat{\mathbf{\Lambda}}$, budući da $\hat{\mathbf{\Lambda}}$ procijenjuje $cov(\mathbf{y}, \mathbf{f}) = \mathbf{\Lambda}$. Sada (1.58) pišemo kao:

$$\hat{\mathbf{B}}_1 = \mathbf{S}^{-1}\hat{\mathbf{\Lambda}} \tag{1.59}$$

Sada po (1.56) , procijenjene $\mathbf{f}_i$ vrijednosti dane su sa:

$$\hat{\mathbf{F}} = \begin{bmatrix} \hat{\mathbf{f}}_1^T \\ \hat{\mathbf{f}}_2^T \\ \vdots \\ \hat{\mathbf{f}}_n^T \end{bmatrix} = \mathbf{Y}_c\hat{\mathbf{B}}_1 = \mathbf{Y}_c\mathbf{S}^{-1}\hat{\mathbf{\Lambda}} \tag{1.60}$$

Ukoliko umjesto $\mathbf{S}$ faktoriziramo $\mathbf{R}$ onda gornje jednadžbe postaju:

$$\hat{\mathbf{B}}_1 = \mathbf{R}^{-1}\hat{\mathbf{\Lambda}}$$
$$\hat{\mathbf{F}} = \mathbf{Y}_s\mathbf{R}^{-1}\hat{\mathbf{\Lambda}}$$

pri čemu je $\mathbf{Y}_s$ matrica standardiziranih varijabli $\frac{y_{ij}-\bar{y}_j}{s_j}$. Uobičajeno je računati faktorske score-ove za rotirane faktore, a ne za originalno dobivene faktore. Iz tog razloga u gornjim jednadžbama mijenjamo $\hat{\mathbf{\Lambda}}$ s $\hat{\mathbf{\Lambda}}^*$.

Kako bi izračunali faktorske score-ove zahtjev je da $\mathbf{S}$ ili $\mathbf{R}$ nisu singularne. Ukoliko je $\mathbf{S}$ ili $\mathbf{R}$ singularna možemo izračunati faktorske score-ove primijenjujući jednostavnu metodu direktno na rotirane težine. Grupiramo varijable u grupe (faktore) sukladno težinama i pronađemo score za svaki faktor uprosječavanjem varijabli koje su dodijeljene tom faktoru. Ukoliko varijable nisu razmjerne trebale bi biti standardizirane prije uprosječavanja.

1.7 Valjanost modela faktorske analize

Za mnoge statističare, faktorska analiza je problematična i ne spada u klasične multivarijatne metode. Razlog nepovjerenja prema faktorskoj analizi leži u: upitnom načinu na koji

se bira m, odnosno poteškoća izbora m, brojne metode za ekstrakciju faktora, različite tehnike rotacija te u subjektivnoj interpretaciji. Neki statističari kritiziraju faktorsku analizu zbog nejedinstvenosti matrice težina $\mathbf{\Lambda}$ ili $\hat{\mathbf{\Lambda}}$, odnosno jedinstvenosti do na množenje ortogonalnom matricom. Međutim, mogućnost rotiranja daje korist faktorske analize. Glavno pitanje je da li faktori zaista postoje. Model za kovarijacijsku matricu glasi $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ pri čemu je $\mathbf{\Lambda}\mathbf{\Lambda}^T$ ranga m. Mnoge populacije nemaju ovakav uzorak u pogledu kovarijacijske matrice osim ako m nije dovoljno velik. Prema tome u takvim populacijama model neće odgovarati podacima kada pokušamo propagirati malu vrijednost za m. S druge strane, u onim populacijama u kojima je $\mathbf{\Sigma}$ dovoljno blizu $\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ i za malu vrijednost m, procedura sampliranja koja vodi do $\mathbf{S}$ može narušiti taj uzorak. U mnogim slučajevima, temeljni je problem što $\mathbf{S}$ ili $\mathbf{R}$ sadrže model i grešku, a koraci faktorske analize nisu u mogućnosti odvojiti navedeno.

Česta je sljedeća situacija. Istraživač dizajnira dugi upitnik, čiji su odgovori na primjer u skali od jedan do pet ili Likertovoj skali. Osobe koje ispunjavaju upitnik, koje variraju od nezainteresiranih do ogorčenih, u žurbi odgovore na pitanja, a odgovori koje su dali često nisu ni dobri subjektivni odgovori. Zatim istraživač svoje podatke unese i sprovede u određenom alatu faktorsku analizu nad prikupljenim podacima. Bivajući razočaran dobivenim rezultatima obraća se statistčaru za pomoć. Statističar pokuša poboljšati rezultate isprobavajući različite metode ekstrakcije faktora, različite rotacije, različite vrijednosti m i tako dalje. Sve biva uzaludno, moraju ekstrahirati 10 do 12 faktora da bi objasnili, recimo, 60% varijabilnosti, a interpretacija tog velikog broja faktora je beznadna. Ukolko u pozadini postoji par dimenzija, one su neprepoznate što zbog semantičkih što zbog slučajnih grešaka u označavanju upitnika. Model dobiven faktorskom analizom ne odgovara takvim podacima, osim ako nije uzet veliki m, što daje beskorisne rezultate.

U situacijama kada pronađeni faktori zadovoljavajuće podudaraju s podacima, trebamo biti strpljivi u interpretaciji sve dok ne utvrdimo stvarno postojanje faktora. Ukoliko se isti faktori pojavljuju kada radimo ponovno uzorkovanje iz iste populacije onda onda možemo biti uvjereni da je primjena modela otkrila neke prave faktore. Dakle, poželjno je u praksi ponoviti ekspreiment kako bi se provjerila stabilnost faktora.

Ukoliko je podatkovni skup dovoljno velik, možemo ga prepoloviti i primijeniti faktorsku analizu na pojedine polovice. Dva dobivena rezultata možemo zatim usporediti jedan s drugim i sa rezultatom koji dobijemo nakon primjene faktorske analize na čitavi podatkovni skup.

Postoje preporuke da $\mathbf{R}^{-1}$ treba biti približno dijagonalna matrica kako bi dobili model faktorske analize koji uspješno odgovara podacima. Kako bi odredili koliko je $\mathbf{R}^{-1}$ blizu dijagonalne matrice, Kaiser(1970) predlaže *mjeru uzoračke adekvatnosti* (MSA), koja se još naziva i Kaiser-Meyer-Olkinova mjera (KMO):

$$MSA = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} q_{ij}^2} \tag{1.61}$$

gdje je r_{ij}^2 kvadrirani element matrice $\mathbf{R}$, q_{ij}^2 je kvadrirani element od $\mathbf{Q} = \mathbf{D}\mathbf{R}^{-1}\mathbf{D}$, a $\mathbf{D} = \left[\left(diag\ \mathbf{R}^{-1}\right)^{\frac{1}{2}}\right]^{-1}$. Kako se $\mathbf{R}^{-1}$ približava dijagonalnog matrici to se MSA približava 1. Kaiser i Rice (1974) predlažu da bi MSA trebao biti barem 0.8 kako bi mogli očekivati zadovoljavajuće rezultate.

Da zaključimo, postoji mnogo podatkovnih skupova nad kojima se faktorska analiza nebi trebala primjenjivati. Jedna od indikacija da $\mathbf{R}$ nije prikladan za faktorizaciju je neuspjeh metoda navedenih u poglavlju Odabir broja faktora za jasan i objektivan odabir vrijednosti broja m. Ukoliko scree graf nema jasno naglašen pregib ili svojstvene vrijednosti nemaju veliku udaljenost oko 1, onda $\mathbf{R}$ vrlo vjerojatno nije prikladan za faktorizaciju, te dodatno, procjene komunaliteta bi nakon faktorizacije trebale biti dovoljno velike.

Poglavlje 2

Primjena faktorske analize na primjeru

U ovom poglavlju sprovest ćemo faktorsku analizu. Podaci nad kojima će se primijeniti faktorska analiza tiču se istraživanja zadovoljstva putnika aviokompanije koji su preuzeti sa internetske stranice:
`https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction`.
Količina podataka s kojom raspolažemo je 25976. Odnosno 25976 osoba je ocijenilo određene usluge koje nudi aviokompanija brojevima od 0 do 5. Faktorska analiza će se sprovesti koristeći statistički paket SAS OnDemand for Academics: `https://www.sas.com/en_au/software/on-demand-for-academics.html`
Varijable koje sudjeluju u faktorskoj analizi su :

Varijabla	Opis varijable
X1	WIFI usluga tokom leta
X2	Povoljnost vremena polaska/dolaska
X3	Lakoća online rezerviranja
X4	Lokacija gate-a
X5	Hrana i piće
X6	Online ukrcavanja
X7	Udobnost sjedala
X8	Animacije tokom leta
X9	Usluga tokom leta
X10	Usluga prostora za noge
X11	Rukovanje prtljagom
X12	Usluga check-ina
X13	Usluga tokom ukrcaja
X14	Čistoća

Cilj sprovedbe faktorske analize je vidjeti da li možemo 14 ulaznih varijabli sažeti u

manji broj latentnih varijabli, odnosno faktora.
Programski kod u SAS-u kojim to radimo je:

```
PROC IMPORT
DATAFILE= "/home/u58251846/DIPLOMSKI/Data_set_za_dipl_final.xlsx"
OUT= WORK.data
DBMS=XLSX
REPLACE;
SHEET="Sheet1";
GETNAMES=YES;


Proc Means Data = data;
Var X1 - X14;
Run;

proc factor data = data corr;
Run;

proc factor data = data
priors = smc msa
rotate = promax reorder
plots = (scree initloadings preloadings loadings);
run;
```

Nakon pokretanja programskog koda u ispisu dobivamo:

Tablica 2.1: Deskriptivna statistika varijabli (SAS ispis)

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
X1	X1	25976	2.7247459	1.3353841	0	5.0000000
X2	X2	25976	3.0468124	1.5333706	0	5.0000000
X3	X3	25976	2.7567755	1.4129513	0	5.0000000
X4	X4	25976	2.9770942	1.2821327	1.0000000	5.0000000
X5	X5	25976	3.2153526	1.3315064	0	5.0000000
X6	X6	25976	3.2616646	1.3555357	0	5.0000000
X7	X7	25976	3.4492224	1.3200902	1.0000000	5.0000000
X8	X8	25976	3.3577533	1.3382994	0	5.0000000
X9	X9	25976	3.3856637	1.2820884	0	5.0000000
X10	X10	25976	3.3501694	1.3188623	0	5.0000000
X11	X11	25976	3.6332384	1.1765247	1.0000000	5.0000000
X12	X12	25976	3.3141746	1.2693321	1.0000000	5.0000000
X13	X13	25976	3.6492532	1.1806810	0	5.0000000
X14	X14	25976	3.2862257	1.3193301	0	5.0000000

Tablica 2.2: Matrica korelacija (SAS ispis)

Correlations															
		X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
X1	X1	1.00000	0.34914	0.71068	0.34779	0.12231	0.45937	0.11699	0.20178	0.11366	0.15970	0.11820	0.04605	0.10842	0.12577
X2	X2	0.34914	1.00000	0.44023	0.45844	-0.01601	0.08094	-0.00193	-0.02233	0.06098	0.00337	0.06568	0.08246	0.06780	-0.00767
X3	X3	0.71068	0.44023	1.00000	0.46551	0.02514	0.40800	0.02281	0.04471	0.03999	0.11675	0.04069	-0.00011	0.03577	0.01097
X4	X4	0.34779	0.45844	0.46551	1.00000	-0.00969	0.00699	-0.00072	-0.00034	-0.03161	-0.00243	-0.00440	-0.05495	-0.00513	-0.01419
X5	X5	0.12231	-0.01601	0.02514	-0.00969	1.00000	0.22960	0.58097	0.62726	0.05069	0.03587	0.03762	0.07678	0.03999	0.65925
X6	X6	0.45937	0.08094	0.40800	0.00699	0.22960	1.00000	0.41541	0.27939	0.14943	0.12035	0.08458	0.20331	0.07197	0.32091
X7	X7	0.11699	-0.00193	0.02281	-0.00072	0.58097	0.41541	1.00000	0.61682	0.12481	0.09915	0.07493	0.18247	0.06731	0.68392
X8	X8	0.20178	-0.02233	0.04471	-0.00034	0.62726	0.27939	0.61682	1.00000	0.41227	0.30320	0.38278	0.11434	0.41102	0.69527
X9	X9	0.11366	0.06098	0.03999	-0.03161	0.05069	0.14943	0.12481	0.41227	1.00000	0.36666	0.52498	0.24744	0.55473	0.11755
X10	X10	0.15970	0.00337	0.11675	-0.00243	0.03587	0.12035	0.09915	0.30320	0.36666	1.00000	0.37911	0.15093	0.37321	0.09798
X11	X11	0.11820	0.06568	0.04069	-0.00440	0.03762	0.08458	0.07493	0.38278	0.52498	0.37911	1.00000	0.24003	0.63196	0.10218
X12	X12	0.04605	0.08246	-0.00011	-0.05495	0.07678	0.20331	0.18247	0.11434	0.24744	0.15093	0.24003	1.00000	0.23917	0.16506
X13	X13	0.10842	0.06780	0.03577	-0.00513	0.03999	0.07197	0.06731	0.41102	0.55473	0.37321	0.63196	0.23917	1.00000	0.09661
X14	X14	0.12577	-0.00767	0.01097	-0.01419	0.65925	0.32091	0.68392	0.69527	0.11755	0.09798	0.10218	0.16506	0.09661	1.00000

U korelacijskoj matrici vidimo da između nekih varijabli imamo jake korelacije, ali i jako male korelacije. Vidimo da su slijedeći parovi varijabli jako korelirani: X3 i X1, X6 i X1, X7 i X5, X8 i X5, X8 i X7, X11 i X9, X13 i X9 , X13 i X11, X14 i X5, X14 i X7 te X14 i X8. Budući da je dovoljan broj jako koreliranih varijabli zaključujemo da su podaci pogodni za faktorski analizu. Slijedeći korak je analiza vrijednosti MSA i analiza svojstvenih vrijednosti matrice korelacija te odabir broja faktora pomoću istih.

Tablica 2.3: Kaiser-Meyer-Olkinova mjera ili Mjera uzoračke adekvatnosti (MSA) (SAS ispis)

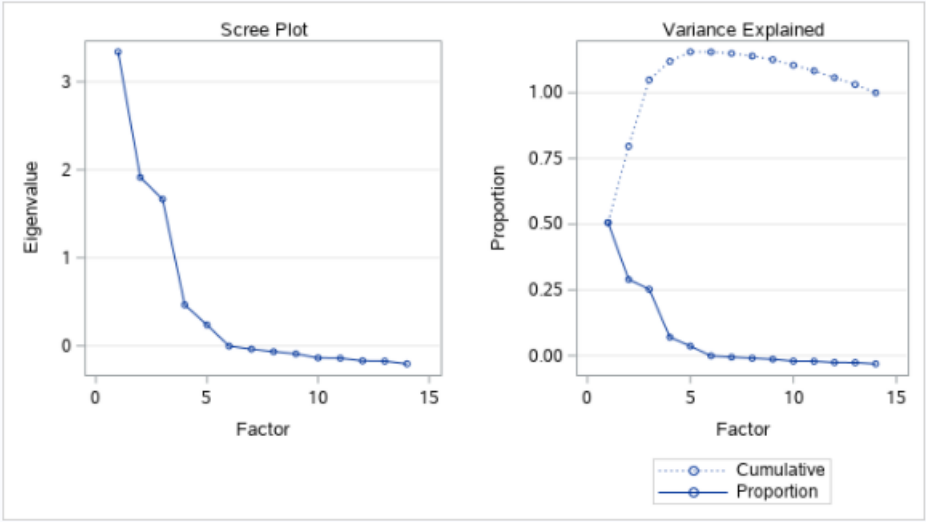
Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.78242746													
X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
0.74700945	0.75431489	0.68851873	0.70504799	0.84121489	0.73260984	0.82664109	0.76807688	0.83058734	0.88729185	0.81727643	0.69441035	0.78625014	0.82261174

Kako je prije diskutirano, ukoliko je vrijednost od MSA manja od 0.5 to se smatra neprihvatljivim te je potrebno više koreliranih varijabli za analizu. Predlaže se da MSA bude barem 0.8 kako bi mogli očekivati zadovoljavajuće rezultate, ovdje smo blizu tome pa možemo očekivati uspješnu sprovedbu analize.

Tablica 2.4: Svojstvene vrijednosti reducirane matrice korelacija (SAS ispis)

Eigenvalues of the Reduced Correlation Matrix: Total = 6.59849319 Average = 0.47132094				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.34019624	1.42593415	0.5062	0.5062
2	1.91426209	0.24853784	0.2901	0.7963
3	1.66572424	1.19969671	0.2524	1.0488
4	0.46602753	0.22780917	0.0706	1.1194
5	0.23821835	0.24189374	0.0361	1.1555
6	-.00367539	0.03345554	-0.0006	1.1549
7	-.03713093	0.02888335	-0.0056	1.1493
8	-.06601428	0.02489352	-0.0100	1.1393
9	-.09090780	0.04639070	-0.0138	1.1255
10	-.13729850	0.00408910	-0.0208	1.1047
11	-.14138759	0.02930473	-0.0214	1.0833
12	-.17069232	0.00375473	-0.0259	1.0574
13	-.17444705	0.02993432	-0.0264	1.0310
14	-.20438138		-0.0310	1.0000

Odabir broja faktora moguće je vršiti na više načina. Jedan od kriterija je da je se uzme broj faktora za koji kumulativna proporcija prelazi 1. Drugi kriterij je odabir onoliko faktora koliko je svojstvenih vrijednosti veće od 1 (Kaiserov kriterij).
Iz svojstvenih vrijednosti reducirane korelacijske matrice vidimo da su prve tri svojstvene vrijednosti veće od 1, također vidimo da prve tri svojstvene vrijednosti reducirane matrice korelacija objašnjavaju 104.88% zajedničke varijance. Ovakav postotak je mogući jer reducirana matrica korelacija nije punog ranga, odnosno nije pozitivno definitna. Dakle, potrebna su 3 faktora po prvom i drugom navedenom kriteriju.



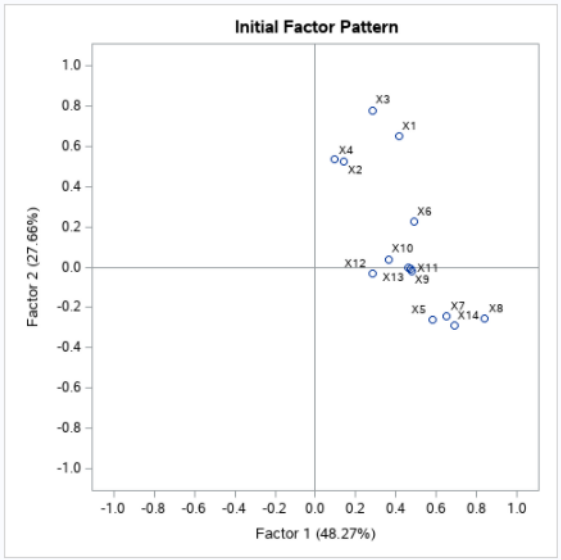
Slika 2.1: Grafički prikaz svojstvenih vrijednosti, proporcija te kumulativne vrijednosti objašnjene varijance (SAS ispis)

Iz slike 2.1 vidimo kako bi metodom lakta očitali broj faktora. Budući da je oštri pad nakon 3. faktora, sa grafa očitavamo da su potrebna tri faktora.

Tablica 2.5: Težine po faktorima, procjene komunaliteta i svojstvene vrijednosti

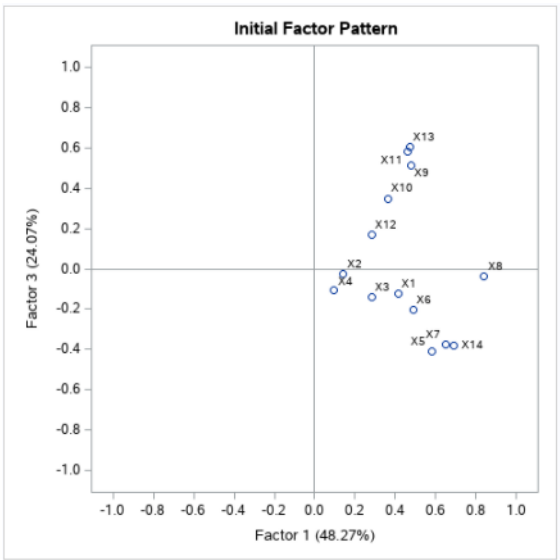
Varijabla	Opis varijable	Faktor 1 Udobnost	Faktor 2 Pogodnost	Faktor 3 Usluge	Komunaliteti
X8	Animacije tokom leta	**0.84006**	-0.25391	-0.03453	0.77135975
X14	Čistoća	**0.68786**	-0.28799	-0.38267	0.70253292
X7	Udobnost sjedala	**0.65176**	-0.24178	-0.37658	0.62506976
X5	Hrana i piće	**0.57900**	-0.25788	-0.40755	0.56783558
X6	Online ukrcavanja	**0.48959**	0.22603	-0.20455	0.33263094
X10	Usluga prostora za noge	**0.36537**	0.04006	0.34923	0.25706139
X12	Usluga check-ina	**0.28318**	-0.03035	0.17080	0.11028313
X3	Lakoća online rezerviranja	0.28400	**0.77805**	-0.14149	0.70603521
X1	WIFI usluga tokom leta	0.41294	**0.65176**	-0.12108	0.60998111
X4	Lokacija gate-a	0.09760	**0.53275**	-0.10799	0.30500980
X2	Povoljnost vremena polaska/dolaska	0.14353	**0.52326**	-0.02484	0.29502111
X13	Usluga tokom ukrcaja	0.47392	-0.01127	**0.60551**	0.59137598
X11	Rukovanje prtljagom	0.46390	-0.00264	**0.58129**	0.55310732
X9	Usluga tokom leta	0.47625	-0.01901	**0.51546**	0.49287857
Svojstvene vrijednosti		3.3401962	1.9142621	1.6657242	6.920183

Faktor 1 smo nazvali Udobnost, faktor 2 Pogodnost, a faktor 3 Usluge. Iz tablice 2.5 očitavamo da prvom faktoru Udobnosti pripadaju varijable X8, X14, X7, X5, X6, X10 i X12. Drugom faktoru Pogodnosti pripadaju varijable X3, X1, X4 i X2. Trećem faktoru Usluge pripadaju varijable X13, X11 i X9. Ovo dodjeljivanje rađeno je na temelju najveće vrijednosti (gledajući apsolutnu vrijednost) u svakom retku matrice faktora. Kao što je prije objašnjeno kako bi procijenili značajnost težine na nekom faktoru vrijednosti od 0.5 su se pokazale dobrima u praksi. Također iz tablice 2.5 vidimo procjene komunaliteta. Komunalitet određene varijable nam govori koliko je varijance te varijable objašnjeno zajedničkim faktorima, možemo reći da su komunaliteti prihvatljivi. Može se vidjeti i da su svojstvene vrijednosti, odnosno varijance objašnjene pojedinim faktorom jednake zbroju kvadrata pripadnih faktorskih težina. Vidimo i da varijabla X10 - Usluga prostora za noge te varijabla X12 - Usluga check-ina imaju vrlo bliske težine na faktorima Udnobnost i Usluge.
Na slijedećim grafičkim prikazima vidjet ćemo ulaznih 14 varijabli prikazane u odnosu na faktore.



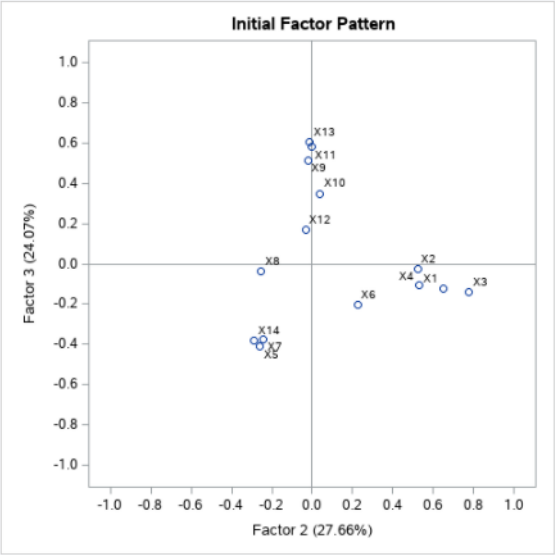
Slika 2.2: Grafički prikaz faktora 1 i faktora 2 (SAS ispis)

Iz slike 2.2 vidimo da je udio varijabilnosti koji objašnjava faktor Udobnost jednak 48.27%, a udio varijabilnosti koji objašnjava faktor Pogodnost jednak 27.66%.



Slika 2.3: Grafički prikaz faktora 1 i faktora 3 (SAS ispis)

Iz slike 2.3 vidimo da je udio varijabilnosti koji objašnjava faktor Usluge jednak 24.07%.



Slika 2.4: Grafički prikaz faktora 2 i faktora 3 (SAS ispis)

Pogledajmo sada što dobivamo rotacijama. Prvo ćemo analizirati ortogonalnu rotaciju faktora.

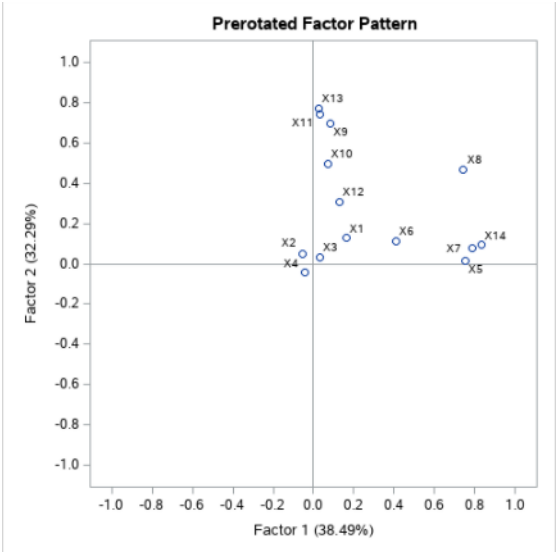
Tablica 2.6: Ortogonalna rotacijska matrica (SAS ispis)

Orthogonal Transformation Matrix			
	1	2	3
1	0.76094	0.58276	0.28523
2	-0.33543	-0.02297	0.94179
3	-0.55539	0.81232	-0.17799

Tablica 2.7: Težine po faktorima nakon ortogonalne rotacije, komunaliteti i svojstvene vrijednosti

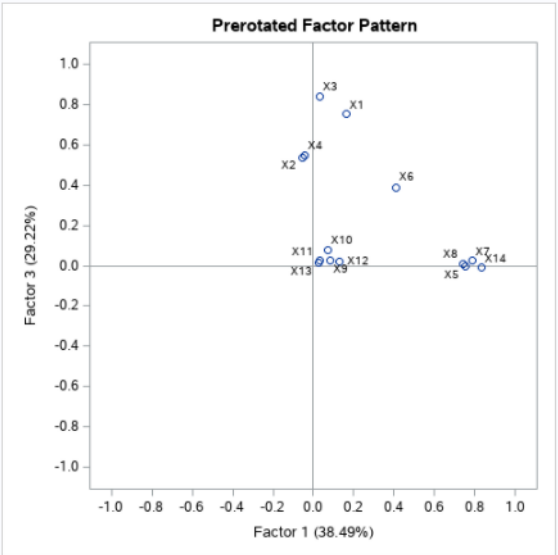
Varijabla	Opis varijable	Faktor 1 Udobnost	Faktor 2 Usluge	Faktor 3 Pogodnost	Komunaliteti
X14	Čistoća	**0.83256**	0.09662	-0.00691	0.70253292
X7	Udobnost sjedala	**0.78620**	0.07947	0.02522	0.62506976
X5	Hrana i piće	**0.75343**	0.01228	-0.00517	0.56783558
X8	Animacije tokom leta	**0.74358**	0.46734	0.00663	0.77135975
X6	Online ukrcavanja	**0.41034**	0.11396	0.38893	0.33263094
X13	Usluga tokom ukrcaja	0.02811	**0.76831**	0.01679	0.59137598
X11	Rukovanje prtljagom	0.03104	**0.74260**	0.02637	0.55310732
X9	Usluga tokom leta	0.08249	**0.69670**	0.02619	0.49287857
X10	Usluga prostora za noge	0.07063	**0.49569**	0.07978	0.25706139
X12	Usluga check-ina	0.13080	**0.30447**	0.02179	0.11028313
X3	Lakoća online rezerviranja	0.03371	0.03270	**0.83895**	0.70603521
X1	WIFI usluga tokom leta	0.16286	0.12731	**0.75316**	0.60998111
X4	Lokacija gate-a	-0.04446	-0.04309	**0.54880**	0.30500980
X2	Povoljnost vremena polaska/dolaska	-0.05250	0.05144	**0.53816**	0.29502111
Svojstvene vrijednosti		2.6632638	2.2345201	2.0223987	6.920183

Kod rotiranih faktora promjene su u varijablama X10 - Usluga prostora za noge i varijabli X12 - Usluga check-ina koje su prije pripadale faktoru Udobnost, a sada su pridružene faktoru Usluga. Faktor Pogodnosti nema promjena. Vidimo da se komunaliteti nisu promijenili nakon ortogonalne rotacije. Varijance su se nakon rotacije po faktorima međusobno prbližno izjednačile u odnosu na varijance koje smo imali na nerotiranim faktorima. Također vidimo da se nakon ortogonalne rotacije ukupna varijabilnost nije promijenila.



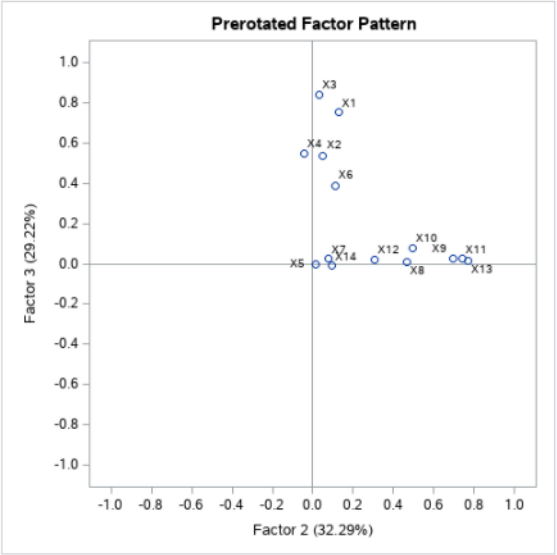
Slika 2.5: Faktori 1 i 2 nakon ortogonalne rotacije (SAS ispis)

Iz slike 2.5 vidimo da je udio varijabilnosti koji objašnjava faktor Udobnost jednak 38.49%, a udio varijabilnosti koji objašnjava faktor Usluga jednak 32.29%.



Slika 2.6: Faktori 1 i 3 nakon ortogonalne rotacije (SAS ispis)

Iz slike 2.6 vidimo da je udio varijabilnosti koji objašnjava faktor Pogodnost jednak 29.22%.



Slika 2.7: Faktori 2 i 3 nakon ortogonalne rotacije (SAS ispis)

Iz slike 2.5 je vidljiv raspored varijabli po faktorima 1 i 2. Vidljivo je da varijable X14, X7, X5, X8 i X6 imaju najveće težine na faktoru 1, a varijable X13, X11, X9, X10 i X12 na faktoru 2, dok preostale varijable imaju težine blizu 0 na faktorima 1 i 2.
Pogledajmo sada što dobivamo kosom rotacijom.

Tablica 2.8: Ciljna matrica za kosu rotaciju (SAS ispis)

Target Matrix for Procrustean Transformation				
		Factor1	Factor2	Factor3
X14	X14	0.98049	0.00154	-0.00000
X7	X7	0.98382	0.00102	0.00003
X5	X5	1.00000	0.00000	-0.00000
X8	X8	0.60716	0.15107	0.00000
X6	X6	0.38032	0.00774	0.30811
X13	X13	0.00005	1.00000	0.00001
X11	X11	0.00007	0.99822	0.00004
X9	X9	0.00162	0.97995	0.00005
X10	X10	0.00270	0.93703	0.00391
X12	X12	0.06113	0.77275	0.00028
X3	X3	0.00006	0.00006	1.00000
X1	X1	0.00907	0.00434	0.90100
X4	X4	-0.00052	-0.00048	0.98583
X2	X2	-0.00090	0.00085	0.97723

Tablica 2.9: Transformacijska matrica kose rotacije (SAS ispis)

Procrustean Transformation Matrix			
	1	2	3
1	1.15400	-0.22668	-0.04736
2	-0.13861	1.41139	-0.06333
3	-0.06242	-0.07553	1.35256

Budući da varijance faktora moraju biti fiksirane na 1 tijekom kose rotacije, normalizirana verzija Procrustean transformacijske matrice je ona koja se zapravo koristi u transformaciji. Korištenje normalizirane transformacijske matrice dovodi do rješenja promax-rotiranih faktora.

Tablica 2.10: Normalizirana transformacijska matrica kose rotacije (SAS ispis)

Normalized Oblique Transformation Matrix			
	1	2	3
1	0.70034	0.45953	0.23340
2	-0.39770	-0.02012	0.96328
3	-0.66698	0.94021	-0.19836

Tablica 2.11: Međufaktorska korelacija (SAS ispis)

Inter-Factor Correlations			
	Factor1	Factor2	Factor3
Factor1	1.00000	0.28148	0.11265
Factor2	0.28148	1.00000	0.12059
Factor3	0.11265	0.12059	1.00000

Vidimo da su nakon kose rotacije faktori korelirani budući da oni nisu više ortogonalni.

Tablica 2.12: Rotirani faktori nakon kose rotacije (standardizirani regresijski koeficijenti)

Varijabla	Opis varijable	Faktor 1 Udobnost	Faktor 2 Usluge	Faktor 3 Pogodnost
X14	Čistoća	**0.85151**	-0.03790	-0.04096
X7	Udobnost sjedala	**0.80379**	-0.04969	-0.00608
X5	Hrana i piće	**0.77988**	-0.11192	-0.03242
X8	Animacije tokom leta	**0.71234**	0.35868	-0.04167
X6	Online ukrcavanja	**0.38942**	0.02811	0.37258
X13	Usluga tokom ukrcaja	-0.06747	**0.78732**	-0.02035
X11	Rukovanje prtljagom	-0.06177	**0.75976**	-0.00957
X9	Usluga tokom leta	-0.00270	**0.70388**	-0.00941
X10	Usluga prostora za noge	0.00703	**0.49544**	0.05459
X12	Usluga check-ina	0.09647	**0.29133**	0.00298
X3	Lakoća online rezerviranja	-0.01617	-0.01817	**0.84383**
X1	WIFI usluga tokom leta	0.11076	0.06280	**0.74823**
X4	Lokacija gate-a	-0.07150	-0.06741	**0.55739**
X2	Povoljnost vremena polaska/dolaska	-0.09101	0.03207	**0.54248**

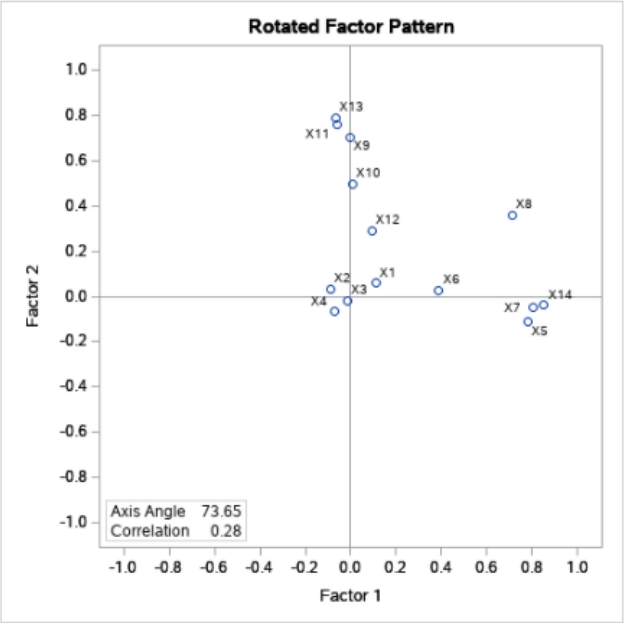
Za razliku od ortogonalnih faktorskih rješenja gdje se faktorske težine interpretiraju kao korelacije između varijabli i faktora, kod kosokutih faktorskih rješenja kao što je ovdje promatrano promax rješenje, potrebno je gledati matricu faktorske strukture kako bi se ispitale korelacije između varijabli i faktora.

Tablica 2.13: Matrica faktorske strukture (matrica korelacija varijabli s faktorima)

Varijabla	Opis varijable	Faktor 1 Udobnost	Faktor 2 Usluge	Faktor 3 Pogodnost
X14	Čistoća	**0.85151**	-0.03790	-0.04096
X7	Udobnost sjedala	**0.80379**	-0.04969	-0.00608
X5	Hrana i piće	**0.77988**	-0.11192	-0.03242
X8	Animacije tokom leta	**0.71234**	0.35868	-0.04167
X6	Online ukrcavanja	**0.38942**	0.02811	0.37258
X13	Usluga tokom ukrcaja	-0.06747	**0.78732**	-0.02035
X11	Rukovanje prtljagom	-0.06177	**0.75976**	-0.00957
X9	Usluga tokom leta	-0.00270	**0.70388**	-0.00941
X10	Usluga prostora za noge	0.00703	**0.49544**	0.05459
X12	Usluga check-ina	0.09647	**0.29133**	0.00298
X3	Lakoća online rezerviranja	-0.01617	-0.01817	**0.84383**
X1	WIFI usluga tokom leta	0.11076	0.06280	**0.74823**
X4	Lokacija gate-a	-0.07150	-0.06741	**0.55739**
X2	Povoljnost vremena polaska/dolaska	-0.09101	0.03207	**0.54248**
Svojstvene vrijednosti		2.9142313	2.4510527	2.1006934

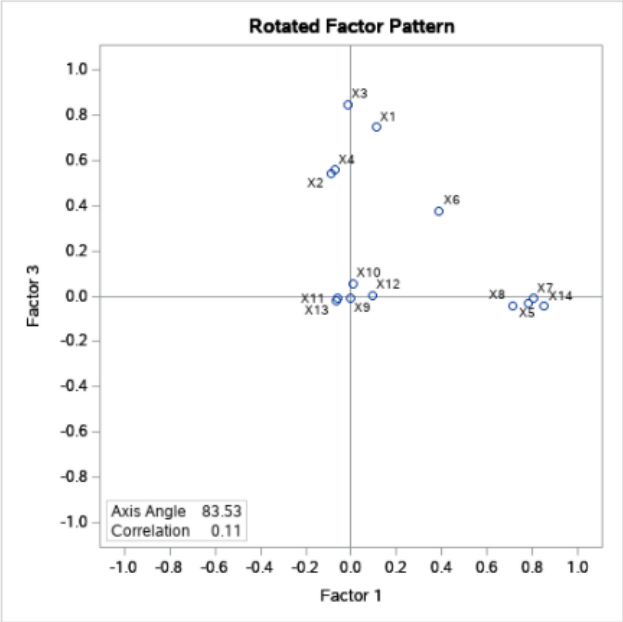
Glavna razlika između tablica 2.12 i 2.13 je što je korelacijsku interpretaciju moguće dobiti koristeći se matricom faktorske strukture. Matrica sturukture je zapravo derivat matrice standardiziranih regresijskih koeficijentata. Ukoliko se matrica standardiziranih regresijskih koeficijentata pomnoži sa međufaktorskom korelacijskom matricom 2.11 dobiva se matrica faktorske strukture [1]. Iz tablice 2.13 vidimo da je korelacija faktora Udobnosti i varijable Čistoća jednaka 0.83623. Također, vidimo da faktor Udobnost objašnjava $(0.83623)^2 = 0.69928 = 69.928\%$ varijance varijable Čistoća.

Vidimo da je raspored varijabli po faktorima nakon kose rotacije ostao isti kao i raspored varijabli po faktorima nakon ortogonalne rotacije, odnosno faktoru Udobnost pripadaju varijable X14, X7, X5, X8 i X6, faktoru Usluge pripadaju X13, X11, X9, X10 i X12, a trećem faktoru Pogodnost pripadaju X3, X1, X4 i X2. Uočavamo da se suma svojstvenih vrijednosti, odnosno ukupna varijabilnost nakon kose rotacije promijenila te je ona sada veća i iznosi 7.465977. Vidimo i da sada za razliku od originalnog i ortogonalnog rješenja suma svojstvenih vrijednosti nije jednaka sumi komunaliteta. Komunaliteti su ostali isti kao i kod faktora bez rotacije i kod faktora nakon ortogonalne rotacije. Ovo je temeljna činjenica o faktorskim rotacijama; rotacije samo redistribuiraju varijancu objašnjenu faktorima dok varijanca objašnjena faktorima za pojedinu varijablu (komunalitet) ostaje nepromijenjen.



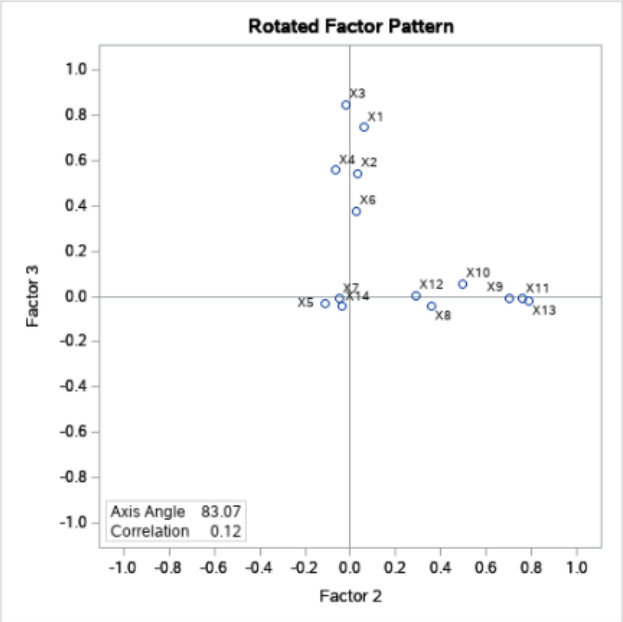
Slika 2.8: Grafički prikaz faktora 1 i 2 nakon kose rotacije (SAS ispis)

Nakon kose rotacije osi, odnosno faktori 1 (Udobnost) i faktor 2 (Usluge) se sijeku pod kutem od 73.65° .



Slika 2.9: Grafički prikaz faktora 1 i 3 nakon kose rotacije (SAS ispis)

Nakon kose rotacije osi, odnosno faktori 1 (Udobnost) i 3 (Pogodnost) se sijeku pod kutem od 83.53° .



Slika 2.10: Grafički prikaz faktora 2 i 3 nakon kose rotacije (SAS ispis)

Nakon kose rotacije osi, odnosno faktori 2 (Usluge) i 3 (Pogodnost) se sijeku pod kutem od 83.07°.

Vidimo da je na ovom primjeru uspješno sprovedena faktorska analiza budući da su se varijable lijepo rasporedile po faktorima, komunalitet je visok te su varijabilnosti objašnjene pojedinim faktorima velike. Uspješno smo od 14 varijabli dobili manji broj, odnosno 3 pozadinska faktora. Prvi faktor (Udobnost) čine varijable Čistoća, Animacije tokom leta, Udobnost sjedala, Hrana i piće i Online ukrcavanja. Drugi faktor (Usluge) čine varijable Usluga tokom leta, Usluga prostora za noge, Rukovanje prtljagom , Usluga check-ina i Usluga tokom ukrcaja. Treći faktor (Pogodnost) čine varijable WIFI usluga tokom leta, Povoljnost vremena polaska/dolaska, Lakoća online rezerviranja i Lokacija gate-a.

Bibliografija

[1] *A practical introduction to factor analysis: exploratory factor analysis*, `https://stats.oarc.ucla.edu/spss/seminars/introduction-to-factor-analysis`, (pristupljeno: prosinac 2021.).

[2] Alvin C. Rencher, *Methods of Multivariate Analysis*, John Wiley & Sons, 2002.

[3] B. G. Tabachnick i L. S. Fidell, *Using multivariate statistics*, Pearson Education, 2013.

Sažetak

U ovom diplomskom radu vidjeli smo teorijsku pozadinu faktorske analize te samu primjenu iste. Na početku smo definirali faktorski model. Vidjeli smo da težine faktorskog modela služe za pridruživanje početnih varijabli faktorima te za interpretaciju dobivenih faktora. Također smo vidjeli da modeliranje započinjemo pretpostavkom da su početne varijable korelirane, a kao izlaz te analize su nekorelirani faktori, odnosno odvojene grupe varijabli. Upoznali smo se sa raznim tehnikama ekstrakcije faktora koje uglavnom imaju isti cilj, a to je ekstrakcija maksimalne varijance iz skupa podataka sa svakim faktorom. Vidjeli smo i kriterije za odabir broja faktora, a u idealnim slučajevima svi kriteriji ishode jednaki brojem. Nakon toga smo vidjeli kako se možemo koristiti različitim tipovima rotacija kako bi poboljšali interpretabilnost i korist rješenja. Na kraju smo na primjeru pokazali kako se provodi faktorska analiza.

Summary

In this thesis, we saw the theoretical background of factor analysis and the application of it. At the beginning, we defined a factor model. We have seen that models factor weights are used to associate initial variables with factors and to interpret the obtained factors. We have also seen that we start modeling by assuming that the initial variables are correlated, and the output of that analysis are uncorrelated factors, i.e. separate groups of variables. We have introduced various factor extraction techniques whom mostly have the same goal, to extract the maximum variance from the data set with each factor. We have also seen the criteria for selecting a number of factors, and ideally all the criteria give the same number as an outcome. After that, we saw how we can use different types of rotation to improve the interpretability and usefulness of the solution. In the end, we showed by example how factor analysis is conducted.

Životopis

Rodena sam 26. listopada 1996. godine u Zagrebu. Osnovnu školu Malešnica u Zagrebu krenula sam pohađati 2003. godine, a svoje srednjoškolsko obrazovanje započela sam 2011. godine u Gimnaziji Lucijana Vranjanina u Zagrebu, Prirodoslovno-matematički smjer. Nakon završene srednje škole, 2015. godine upisujem preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu na kojem 2019. godine stječem titulu univ. bacc. math. Iste sam godine upisala diplomski studij Matematičke statistike, također na Prirodoslovno-matematičkom fakultetu u Zagrebu.