Eksploratorna faktorska analiza

Erdeljac, Barbara

Master's thesis / Diplomski rad

2021

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:910335

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Barbara Erdeljac

EKSPLORATORNA FAKTORSKA ANALIZA

Diplomski rad

Voditelj rada:
doc. dr. sc. Azra Tafro

Zagreb, srpanj 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. _____

Zahvaljujem svojoj mentorici, doc. dr. sc. Azri Tafro, na pomoći, savjetima, uloženom vremenu i strpljenju prilikom pisanja ovoga rada.
Hvala mojim najmilijima, koji su bezuvjetno vjerovali u mene,
i bili velika podrška.
Neizmjerno hvala mojim kolegicama, uz njih su sve brige postajale manje, a svaka sreća zajednička.

Sadržaj

Uvod

Faktorska analiza je metoda koja je inicijalno razvijena od strane psihologa, u nastojanju za redukcijom dimezionalnosti velikog skupa podataka. I danas se smatra jednom od najkorištenijih metoda na području psihometrije, no ima sve veću primjenu i u mnogim prirodnim znanostima, poput molekularne biologije, biokemije ili astrofizike.

Faktorska analiza dio je multivarijatne analize, područja multivarijatne statistike koje se bavi istovremenim promatranjem većeg broja varijabli i nastoji uočiti jesu li one, i na koji način, međusobno povezane. Upravo iz tog razloga, faktorska analiza polazi od jedne od mjera povezanosti među varijablama, kovarijance ili korelacije. Cilj analize je pronaći manji broj neopaženih varijabli koje objašnjavaju što veći dio varijabilnosti. Opažene varijable modeliraju se kao linearne kombinacije neopaženih faktora, a razlikujemo eksploratornu i konfirmatornu faktorsku analizu, ovisno o tome jesu li nam faktori unaprijed poznati.

U prvom poglavlju definiramo kovarijancu i korelaciju – mjere međusobne povezanosti varijabli koje su potrebne za daljnje definiranje pojmova vezano uz samu metodu faktorske analize. U drugom poglavlju objašnjen je faktorski model i korelacija unutar njega, što nastojimo geometrijski interpretirati u trećem poglavlju. U četvrtom poglavlju objašnjene su neke od metoda određivanja broja faktora u faktorskom modelu – Guttman-Kaiserov kriterij, scree plot, paralelna analiza i metoda B-koeficijenta, dok su u petom poglavlju definirane neke od metoda procjene težina faktora – metoda glavnih komponenata i metoda maksimalne vjerodostojnosti, i objašnjen pojam rotacije faktora koja omogućava lakšu interpretaciju modela. Na samom kraju, u šestom poglavlju, provodena je eksploratorna faktorska analizu na skupu varijabli koje mjere emocije pobuđene u osobi prilikom čitanja naslova novinskih članaka, kao i ”vrijednost” tih naslova, odnosno određuju pripadnost vijesti određenom području.

Poglavlje 1

Mjere međusobne povezanosti varijabli

Kovarijanca

Kovarijanca je mjera kojom izražavamo ovisnost dviju slučajnih varijabli. Uzmimo da su X i Y dvije slučajne varijeble. Njihovu kovarijancu definiramo s

$$\sigma_{XY}^2 = \mathrm{Cov}\,(X, Y) = \mathrm{E}\,(XY) - \mathrm{E}\,(X)\,\mathrm{E}\,(Y).$$

Kovarijancu varijable X same sa sobom nazivamo *varijancom varijable* X,

$$\sigma_X^2 = \mathrm{Var}\,(X) = \mathrm{Cov}\,(X, X),$$

dok korijen vrijednosti varijance nazivamo *standardnom devijacijom varijable*.

Prethodno definirane pojmove ćemo u radu računati prema formuli za *uzoračku kovarijancu*

$$s_{XY}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}), \tag{1.1}$$

odnosno *uzoračku varijancu*

$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2, \tag{1.2}$$

ili prema formuli za *populacijsku kovarijancu*

$$\sigma_{XY}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y), \tag{1.3}$$

odnosno *populacijsku varijancu*

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2, \tag{1.4}$$

ovisno o potrebi, pri čemu je $x_1, \ldots, x_N$ realizacija slučajne varijable X (N broj elemenata u uzorku/populaciji), $\overline{x}$ srednja vrijednost uzorka, odnosno μ_x očekivanje varijable X, te analogno za Y.

U slučaju kada je $\mathbf{X}$ n-dimenzionalni slučajni vektor, odnosno

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix},$$

tada vrijednosti kovarijanci između varijabli X_i prikazujemo u matričnoj formi, unutar tzv. kovarijacijske matrice $\mathbf{\Sigma}$,

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1n}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \cdots & \sigma_{nn}^2 \end{bmatrix},$$

pri čemu koristimo oznaku $\sigma_{ij}^2 = \mathrm{Cov}\,(X_i, X_j)$. Uzoračku kovarijacijsku matricu označit ćemo sa $\mathbf{S}$, i njezini su elementi uzoračke kovarijance, odnosno varijance, definirane u (1.1) i (1.2).

Koeficijent korelacije

Korelacija mjeri stupanj međusobne povezanosti varijabli, odnosno mogućnost predviđanja vrijednosti jedne varijable na osnovu saznanja vrijednosti druge varijable. Jedna od mogućih mjera za korelaciju je Pearsonov koeficijent korelacije. Općenito, za slučajne varijable X i Y, računamo ga prema formuli

$$r = \frac{s_{xy}^2}{s_x \, s_y}, \tag{1.5}$$

ukoliko se radi o koeficijentu korelacije uzorka, odnosno formulom

$$\rho = \frac{\sigma_{xy}^2}{\sigma_x \, \sigma_y} \tag{1.6}$$

za populacijski koeficijent korelacije, pri čemu su u prvoj formuli s_x i s_y prethodno definirane empirijske vrijednosti standardne devijacije varijabli X i Y, te s_{xy}^2 empirijska vrijednost kovarijance, odnosno σ_X, σ_Y populacijske standardne devijacije, a σ_{xy}^2 populacijska kovarijanca.

Poglavlje 2

Faktorski model

Faktorska analiza je statistička metoda koja nastoji smanjiti broj početno danih međusobno koreliranih varijabli na manji broj varijabli, tzv. faktora, odnosno pronalazi model koji će jednako dobro objašnjavati korelacije početnog skupa varijabli, ali pomoću manjeg broja varijabli. Najjednostavnija veza koju je moguće ostvariti između početnih varijabli i faktora jest linearna, stoga je cilj prikazati svaku od početnih varijabli kao linearnu kombinaciju faktora. Kako bi faktorski model uopće imao svrhu, broj faktora bi uvijek trebao biti značajno manji od broja početno danih varijabli.

Dvije su osnovne vrste faktorske analize, *eksploratorna (EFA)* i *konfirmatorna (CFA*[1]*)*. Eksploratorna faktorska analiza nastoji utvrditi faktore na temelju dobivenih podataka i koristi se kada faktori nisu unaprijed definirani, dok konfirmatorna testira već poznatu ili unaprijed postavljenu teoretsku strukturu faktora. U ovom radu ćemo se, u dijelu primjene faktorske analize, koristiti metodama eksploratorne faktorske analize.

Standardizirane varijable

Označimo sa $X_1, X_2, \ldots, X_n$ početno dane varijable, pri čemu ćemo sa n označavati njihov ukupan broj. Ukoliko želimo osigurati da vrijednost varijable ne ovisi o njezinoj mjernoj jedinici, svaku od varijabli moguće je svesti na njezin standardizirani oblik - da je njezino očekivanje nula, a standardna devijacija jednaka jedan. Općenito, standardizacija varijable X podrazumijeva oduzimanje njene srednje vrijednosti, $\overline{X}$, te dijeljenje s njenom standardnom devijacijom, σ,

$$\frac{X - \overline{X}}{\sigma}.$$

[1] eng. *confirmatory factor analysis*

Faktorski model

Označimo sa $F_1, F_2, \ldots, F_m$ tzv. *zajedničke faktore*, njih m, i pretpostavimo da su dani u standardiziranom obliku. Kako je ideja faktorske analize opisati vrijednosti n varijabli kao linearnu kombinaciju vrijednosti m faktora, faktorski model zapisujemo kao

$$X_j = a_{j1}F_1 + a_{j2}F_2 + a_{j3}F_3 + \cdots + a_{jm}F_m + \varepsilon_j + \mu_j, \quad (j = 1, 2, \ldots, n), \tag{2.1}$$

gdje su sa $a_{j1}, a_{j2}, \ldots, a_{jm}$ označene tzv. *težine zajedničkih faktora*, pri čemu je, primjerice, a_{j1} težina j-te varijable u odnosu na prvi faktor. Iz težina faktora vidimo koliki je utjecaj određenog faktora na neku od varijabli X_j: što je veća težina, veći je i njen utjecaj na varijablu. S μ_j označeno je očekivanje varijable X_j. Rezidualne veličine, označene s ε_j, $(j = 1, \ldots, n)$, koristimo za opisivanje onog dijela korelacije među početno danim varijablama koji zajednički faktori nisu uspjeli objasniti. Kako bismo i taj rezidualni dio objasnili preko faktora, uvodimo pojam specifičnih faktora, $U_1, U_2, \ldots, U_n$, njih n. Pretpostavljamo, kao i za zajedničke faktore, da su specifični faktori standardizirane veličine. Njihov broj jednak je broju početno danih varijabli, jer je svaki od njih vezan isključivo za jednu od početnih varijabli. Sada je faktorski model oblika

$$X_j = a_{j1}F_1 + a_{j2}F_2 + a_{j3}F_3 + \cdots + a_{jm}F_m + b_jU_j + \mu_j, \quad (j = 1, 2, \ldots, n), \tag{2.2}$$

gdje b_j označava težinu specifičnog faktora U_j u prikazu varijable X_j.

U idealnom slučaju, zajednički faktori bi mogli u potpunosti objasniti korelacije početnog skupa varijabli, ali to najčešće nije moguće, pa upravo iz tog razloga uz njihovu vrijednost dodajemo i vrijednosti specifičnih faktora.

Pretpostavimo da je $x_{j1}, x_{j2}, \ldots, x_{jN}$ realizacija slučajnog uzorka za svaku od slučajnih varijabli X_j, pri čemu je N broj mjerenja, te neka su u_{ji} i $f_{j1}, f_{j2}, \ldots, f_{jN}$ pripadajuće vrijednosti faktora U_i i $F_1, F_2, \ldots, F_m$. Tada izraz (2.2) možemo zapisati za svaku pojedinu vrijednost realizacije slučajnog uzorka:

$$x_{ji} = a_{j1}f_{1i} + a_{j2}f_{2i} + \cdots + a_{jm}f_{mi} + b_ju_{ji} + \mu_j, \quad (j = 1, 2, \ldots, n;\ i = 1, 2, \ldots, N). \tag{2.3}$$

Sada je jasno da je glavni problem kod određivanja faktorskog modela određivanje broja faktora i procjena vrijednosti njihovih težina, a o pojedinim metodama određivanja istih bit će riječ kasnije u ovom radu.

Matrični model

Prikažimo sada faktorski model matrično. Neka je $\mathbf{X}$ slučajni vektor definiran s

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix},$$

pri čemu su $X_1, X_2, \ldots, X_n$ početno dane varijable, te neka su $\mathbf{F}$ i $\mathbf{U}$ također slučajni vektori, $\mathbf{F}$ slučajni vektor zajedničkih faktora, dimenzije m, te $\mathbf{U}$ slučajni vektor specifičnih faktora, dimenzije n, definirani analogno kao slučajni vektor $\mathbf{X}$. Pripadne težine zajedničkih faktora nalaze se u $n \times m$ matrici $\mathbf{A}$, dok se težine specifičnih faktora nalaze na dijagonali matrice $\mathbf{B}$, $\mathbf{B} = \text{diag}(b_1, \ldots, b_n)$. Tada je matrični zapis faktorskog modela (2.2) dan s

$$\mathbf{X} = \mathbf{AF} + \mathbf{BU} + \mu, \tag{2.4}$$

pri čemu je μ n - dimenzionalni vektor očekivanja od $\mathbf{X}$.

Ukoliko želimo standardizirati vrijednosti početno danih varijabli, matricu $\mathbf{X}$ moguće je transformirati u matricu standardiziranih vrijednosti $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{HXD}^{-1/2},$$

gdje je $\mathbf{D}$ dijagonalna matrica s vrijednostima uzoračkih varijanca varijabli $X_1, \ldots, X_n$ na dijagonali, tj. $\mathbf{D} = \text{diag}(s_{11}, s_{22}, \ldots, s_{nn})$, a $\mathbf{H}$ tzv. centrirajuća matrica[2] definirana s $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^T$, pri čemu je $\mathbf{I}_n$ jedinična matrica dimenzije n, a $\mathbf{1_n}$ kvadratna matrica dimenzije n čiji su elementi jedinice.

Pretpostavke modela

Iako smo već ranije spominjali neke od pretpostavki faktorskog modela, sumirajmo ih sada sve zajedno. Ukoliko je faktorski model dan matrično, kao u (2.4), tada su osnovne pretpostavke faktorskog modela:

$$\begin{aligned} &\text{E}(\mathbf{F}) = 0, \\ &\text{Cov}\,\mathbf{F} = \mathbf{I}_m, \\ &\text{E}(\mathbf{U}) = 0, \\ &\text{Cov}\,\mathbf{U} = \mathbf{I}_n, \\ &\text{Cov}(\mathbf{F}, \mathbf{U}) = 0, \end{aligned} \tag{2.5}$$

Preciznije, pretpostavljamo da su faktori, i zajednički i specifični, standardizirane varijable, tj. da je njihovo očekivanje jednako nuli, a varijanca jedan. Pretpostavljamo i da su zajednički faktori F_i međusobno nekorelirani te da isto vrijedi za specifične faktore U_i. Također, pretpostavljamo da su zajednički i specifični faktori međusobno nekorelirani. Zbog pretpostavki o nekoreliranosti faktora, faktorski model ponekad se naziva i ortogonalnim.

[2] eng. *centering matrix*

2.1 Korelacija unutar faktorskog modela

Glavna statistička veličina na kojoj se temelji faktorska analiza jest koeficijent korelacije. Korelacija opisuje vezu između dvije varijable i faktorski model nije moguće postaviti dok nije izračunata korelacija svakog od parova danih varijabli. Daljnji račun, ako su uočene veze među varijablama, provodi se na temelju matrice korelacija ili kovarijacijske matrice.

Komunalnost i specifična varijanca

Promotrimo sada utjecaj faktora na ukupnu varijancu svake od varijabli X_j u faktorskom modelu (2.2). Pretpostavljamo da su vrijednosti varijable X_j standardizirane. Ideja je izraziti varijancu u terminima težina faktora, stoga, prema formuli za populacijsku varijancu, kvadriramo obje strane izraza (2.3), sumiramo po svim vrijednostima, njih N, i na kraju podijelimo sa N:

$$\begin{aligned}\frac{\sum x_{ji}^2}{N} = a_{j1}^2\frac{\sum f_{1i}^2}{N} + a_{j2}^2\frac{\sum f_{2i}^2}{N} + \cdots + a_{jm}^2\frac{\sum f_{mi}^2}{N} + b_j^2\frac{\sum u_{ji}^2}{N} + \\ + 2\Big(a_{j1}a_{j2}\frac{\sum f_{1i}f_{2i}}{N} + \cdots + a_{j,m-1}a_{jm}\frac{\sum f_{m-1,i}f_{mi}}{N} + \\ + a_{j1}b_j\frac{\sum f_{1i}u_{ji}}{N} + \cdots + a_{jm}b_j\frac{\sum f_{mi}u_{ji}}{N}\Big).\end{aligned}$$

Koristeći se pretpostavkom da su vrijednosti faktora standardizirane, kao i dodatnom pretpostavkom da to isto vrijedi za početne varijable, prethodna jednadžba svodi se na

$$1 = \sigma_{jj}^2 = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 + b_j^2 + 2\Big(a_{j1}a_{j2}r_{F_1F_2} + \cdots + a_{jm}b_jr_{F_mU_j}\Big), \tag{2.6}$$

pri čemu je σ_{jj}^2 varijanca varijable X_j. Kako su, prema pretpostavci, svi faktori međusobno nekorelirani, jednadžba (2.6) pojednostavljuje se na

$$1 = \sigma_{jj}^2 = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 + b_j^2. \tag{2.7}$$

Iz posljednjeg izraza zaključujemo da su kvadrati težina zajedničkih faktora a_{ji}, $i = 1, \ldots, m$ i težine specifičnog faktora b_j zapravo dijelovi varijance varijable X_j koji se mogu objasniti odgovarajućim faktorima. Primjerice, utjecaj faktora F_2 na varijancu varijable X_1 je a_{12}^2, što daje da je ukupan utjecaj faktora F_2 na varijancu svih varijabli jednak

$$a_{12}^2 + a_{22}^2 + a_{32}^2 + \cdots + a_{n2}^2.$$

Poopćenjem prethodnog izraza, dobivamo da ukupan utjecaj faktora F_t na varijancu svih varijabli iznosi

$$\sum_{j=1}^{n} a_{jt}^2.$$

Dodatno, izračun kovarijacijske matrice moguće je zapisati i matrično

$$\begin{aligned}
\mathbf{I}_n = \operatorname{Cov}\mathbf{X} &= \mathrm{E}\,(\mathbf{X}\mathbf{X}^T) \\
&= \mathrm{E}\,[(\mathbf{A}\mathbf{F} + \mathbf{B}\mathbf{U})(\mathbf{A}\mathbf{F} + \mathbf{B}\mathbf{U})^T] \\
&= \mathrm{E}\,[(\mathbf{A}\mathbf{F} + \mathbf{B}\mathbf{U})(\mathbf{F}^T\mathbf{A}^T + \mathbf{U}^T\mathbf{B}^T)] \\
&= \mathbf{A}\mathrm{E}\,(\mathbf{F}\mathbf{F}^T)\mathbf{A}^T + \mathbf{A}\mathrm{E}\,(\mathbf{F}\mathbf{U}^T)\mathbf{B}^T + \mathbf{B}\mathrm{E}\,(\mathbf{U}\mathbf{F}^T)\mathbf{A}^T + \mathbf{B}\mathrm{E}\,(\mathbf{U}\mathbf{U}^T)\mathbf{B}^T \\
&= \mathbf{A}\operatorname{Cov}(\mathbf{F})\mathbf{A}^T + \mathbf{A}\operatorname{Cov}(\mathbf{F},\mathbf{U})\mathbf{B}^T + \mathbf{B}\operatorname{Cov}(\mathbf{U},\mathbf{F})\mathbf{A}^T + \mathbf{B}\operatorname{Cov}(\mathbf{U})\mathbf{B}^T \\
&= \mathbf{A}\operatorname{Cov}(\mathbf{F})\mathbf{A}^T + \mathbf{B}\operatorname{Cov}(\mathbf{U})\mathbf{B}^T \\
&= \mathbf{A}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T.
\end{aligned} \tag{2.8}$$

Uvedimo sada neke pojmove specifične za faktorsku analizu. Sumu kvadrata težina zajedničkih faktora, odnosno

$$h_j^2 = \sum_{l=1}^{m} a_{jl}^2$$

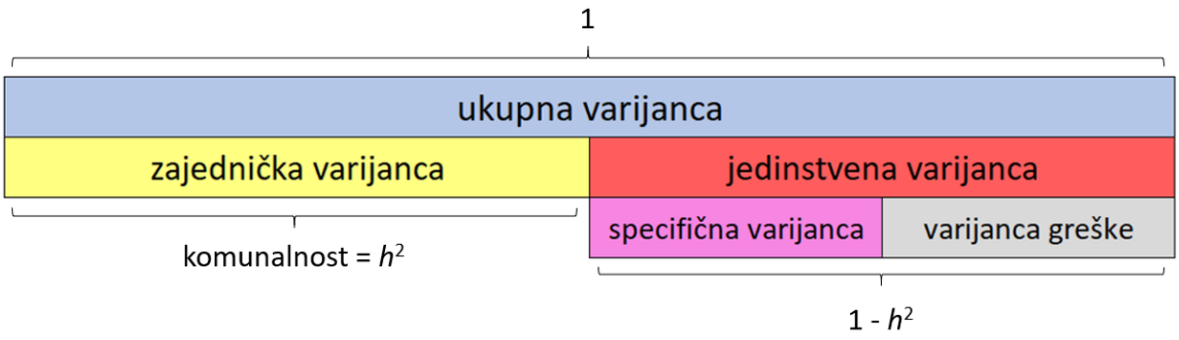
zovemo *komunalnost varijable*, koja u varijanci varijable X_j zauzima udio koji je objašnjen zajedničkim faktorima. Kako oni ne uspjevaju u potpunosti objasniti ukupnu varijancu varijable, preostali dio nazivamo *jedinstvenom varijancom*[3]. Jedinstvena varijanca varijable X_j je upravo b_j^2 (posljednji član zbroja desne strane jednakosti (2.7)). Ukupnu jedinstvenu varijancu u faktorskom modelu često ćemo označavati s $\mathbf{\Psi}$, a kovarijacijsku matricu od $\mathbf{X}$ sa $\mathbf{\Sigma}$, stoga izraz (2.8) možemo zapisati i kao

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^T + \mathbf{\Psi}. \tag{2.9}$$

Jedinstvenu varijancu moguće je razdijeliti na *specifičnu varijancu*, onu koja je vezana uz određenu varijablu, i na *varijancu greške* koja dolazi od grešaka pri uzorkovanju. Raspodjela varijance u faktorskom modelu prikazana je na slici 2.1.

Specifičnu varijancu i varijancu greške ne modeliramo zasebno u faktorskom modelu, već vrijednošću specifičnog faktora nastojimo objasniti ukupnu varijancu koja nije objašnjena zajedničkim faktorima, odnosno jedinstvenu varijancu.

[3] eng. *unique variance*



Slika 2.1: Raspodjela varijance u faktorskom modelu

Poglavlje 3

Geometrijska interpretacija faktorskog modela

Kako bi razumijevanje faktorske analize bilo što potpunije, u ovom poglavlju razmatramo geometrijski pristup faktorskom modelu, odnosno nastojimo geometrijski interpretirati algebarske izraze koji se koriste u metodi. Literatura korištena za pisanje ovog poglavlja navedena je pod [4].

Ideja jest, pomoću rezultata koje ćemo prikazati, odrediti najmanji broj zajedničkih faktora koji je potreban za faktorski model. Varijable interpretiramo kao točke, odnosno radijvektore, u višedimenzionalnom prostoru. Tada standardna devijacija varijable postaje mjera udaljenosti, a korelacija dviju varijabli je zapravo kosinus kuta koji zatvaraju vektori tih dviju varijabli.

3.1 Geometrijska interpretacija korelacije

Neka su $X'_1, X'_2, \ldots, X'_n$ početno dane varijable na kojima želimo postaviti faktorski model. Pretpostavimo da je za svaku od njih dana N-dimenzionalna realizacija slučajnog uzorka, odnosno da za svaku od varijabli imamo vrijednosti N mjerenja. Njihov matrični prikaz je

$$[X'_{ji}] = \begin{bmatrix} x'_{11} & x'_{12} & x'_{13} & \ldots & x'_{1N} \\ x'_{21} & x'_{22} & x'_{23} & \ldots & x_{2N'} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ x'_{n1} & x'_{n2} & x'_{n3} & \ldots & x'_{nN} \end{bmatrix}$$

gdje je element u j-tom retku i i-tom stupcu i-ta vrijednost realizacije varijable X'_j, odnosno svaki redak matrice je N-dimenzionalna realizacija jedne od slučajnih varijabli.

Definirajmo varijable $X_1, X_2, \ldots, X_n$ kao odstupanja varijabli od njihovih sredina, $X_j = X'_j - \overline{X'_j}$. Tada, iz prethodne matrice, oduzimanjem srednje vrijednosti $\overline{x'_j}$ realizacije slučajnog uzorka varijable X'_j od svakog njenog mjerenja,

$$x_{ji} = x'_{ji} - \overline{x'_j},$$

dobivamo matricu odstupanja[1]

$$[x_{ji}] = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1N} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2N} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ x_{n1} & x_{n2} & x_{n3} & \ldots & x_{nN} \end{bmatrix}, \tag{3.1}$$

te iz nje matricu realizacija standardiziranih početnih vrijednosti,

$$[z_{ji}] = \begin{bmatrix} z_{11} & z_{12} & z_{13} & \ldots & z_{1N} \\ z_{21} & z_{22} & z_{23} & \ldots & z_{2N} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ z_{n1} & z_{n2} & z_{n3} & \ldots & z_{nN} \end{bmatrix},$$

gdje je $z_{ji} = \frac{x_{ji}}{s_j}$, pri čemu je s_j oznaka za uzoračku standardnu devijaciju, tj. $s_j = \sqrt{\frac{\sum x_{ji}^2}{N-1}}$.

Kako bismo varijable $X_1, X_2, \ldots, X_n$ interpretirali geometrijski, realizaciju varijable prikazujemo kao točku, odnosno radijvektor realizacije varijable $\overrightarrow{Ox_j}$, u N-dimenzionalnom prostoru, čije su koordinate određene vrijednostima njenih realizacija, odnosno

$$\overrightarrow{Ox_j} = (x_{j1}, x_{j2}, \ldots, x_{jN}).$$

Sada je duljina radijvektora $\overrightarrow{Ox_j}$ dana s

$$\rho_j = \sqrt{\sum x_{ji}^2}, \tag{3.2}$$

što možemo izraziti i kao

$$\rho_j = \sqrt{N-1}\, s_j. \tag{3.3}$$

Prema prethodnom, procijenjenu vrijednost standardne devijacije slučajne varijable moguće je interpretirati kao mjeru proporcionalnu duljini radijvektora realizacije varijable s koeficijentom proporcionalnosti $1/\sqrt{N-1}$.

[1] eng. *the matrix of deviates*

Označimo sada s ϕ_{jk} kut između dva radijvektora $\overrightarrow{Ox_j}$ i $\overrightarrow{Ox_k}$. Prema definiciji skalarnog produkta, dobivamo:

$$\sum_i x_{ji}x_{ki} = \rho_j\,\rho_k \cos\phi_{jk}. \tag{3.4}$$

Korištenjem (3.3), prethodni izraz moguće je zapisati kao

$$\cos\phi_{jk} = \frac{1}{N-1}\cdot\frac{\sum_i x_{ji}x_{ki}}{s_j s_k}.$$

Sada, prema formuli za uzoračku kovarijancu (1.1) dobivamo

$$\cos\phi_{jk} = \frac{s_{jk}^2}{s_j s_k}.$$

Konačno, zaključujemo

$$r_{jk} = \cos\phi_{jk}, \qquad (j,k = 1,2,\ldots,n), \tag{3.5}$$

pri čemu je r_{jk} procijenjena vrijednost koeficijenta korelacije dviju varijabli X_j i X_k, koji je definiran u (1.5), odnosno koeficijent korelacije uzorka tih dviju varijabli. Posljednjim izrazom dana je jasna geometrijska interpretacija procjene korelacije dviju varijabli (mjerenih kao odstupanja od njihovih aritmetičkih sredina) koja je, dakle, jednaka kosinusu kuta između njihovih radijvektora u N-dimenzionalnom prostoru.

3.2 Temeljni potprostor u faktorskoj analizi

Kako bismo došli do rezultata koji su ključni za geometrijsku interpretaciju faktorskog modela, naprije navodimo dva teorema koja ćemo koristiti, preuzeta iz [4].

Teorem 3.2.1. *Neka je* $[x_{ji}]$ *matrica s koordinatama točaka* $P_1,\ldots,P_n$ *u retcima i neka je rang matrice* $[x_{ji}]$ *jednak m. Tada su točke* $P_1,\ldots,P_n$ *sadržane u m-dimenzionalnom prostoru, ali ne i u* μ *-dimenzionalnom prostoru, gdje* $\mu < m$ *.*

Teorem 3.2.2. *Neka je A matrica ranga m. Tada je rang matrice* AA^T *jednak m .*

Promotrimo sada produkt matrice $[z_{ji}]$, definirane na početku ovog poglavlja, i njoj transponirane matrice $[z_{ik}]$,

$$[z_{ji}]\cdot[z_{ik}] = \begin{bmatrix} \sum z_{1i}^2 & \sum z_{1i}z_{2i} & \sum z_{1i}z_{3i} & \cdots & \sum z_{1i}z_{ni} \\ \sum z_{2i}z_{1i} & \sum z_{2i}^2 & \sum z_{2i}z_{3i} & \cdots & \sum z_{2i}z_{ni} \\ \sum z_{3i}z_{1i} & \sum z_{3i}z_{2i} & \sum z_{3i}^2 & \cdots & \sum z_{3i}z_{ni} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum z_{ni}z_{1i} & \sum z_{ni}z_{2i} & \sum z_{ni}z_{3i} & \cdots & \sum z_{ni}^2 \end{bmatrix}.$$

S obzirom na to da su vrijednosti matrice $[z_{ji}]$ prethodno standardizirane, vrijedi $\sum z_{ji}^2 = N-1$ i $\sum z_{ji}z_{ki} = (N-1)\cdot r_{jk}$, pa gornji produkt matrica možemo zapisati i kao

$$[z_{ji}]\cdot[z_{ik}] = (N-1)\cdot\begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ r_{21} & 1 & r_{23} & \dots & r_{2n} \\ r_{31} & r_{32} & 1 & \dots & r_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \dots & 1 \end{bmatrix} = (N-1)\cdot[r_{jk}].$$

Sada iz Teorema 3.2.2 možemo zaključiti da je rang matrice korelacija uzorka $[r_{jk}]$ jednak rangu matrice standardiziranih vrijednosti $[z_{ji}]$. Stoga, svi rezultati izvedeni na temelju ranga matrice $[z_{ji}]$ mogu biti iskazani i u terminu matrice korelacija. Štoviše, iz Teorema 3.2.1 slijedi da je moguće realizaciju svake od n varijabli izraziti kao linearnu kombinaciju minimalno m realizacija faktora, gdje je m rang matrice korelacija.

Kako je rang matrice korelacija modela (2.3) jednak n, a cilj je za broj faktora uzeti broj manji od n, zanemarujemo dio koji doprinose specifični faktori te za dijagonalne vrijednosti matrice korelacija, umjesto jedinica, uzimamo komunalnosti, definirane u (2.1). Rang takve matrice, m, generalno je manji od n, i to je ujedno i minimalan broj zajedničkih faktora koji je potrebno koristiti u faktorskom modelu. Geometrijski, najmanji potprostor koji sadrži n radijvektora realizacije varijabli jest dimenzije m. Takav potprostor nazivamo *prostorom zajedničkih faktora* [2]. Formirajmo prethodne zaključke kao teorem:

Teorem 3.2.3. *Neka je m rang matrice korelacija relizacija početno danih varijabli čiji su dijagonalni elementi zamijenjeni komunalnostima. Najmanji broj linearno nezavisnih faktora koji će većinom uračunati korelaciju početnih varijabli jest m, odnosno prostor zajedničkih faktora je m - dimenzionalan.*

Sada je jasno da je m-dimenzionalan prostor zajedničkih faktora razapet osima koje predstavljaju zajedničke faktore, a realizacije varijabli su točke, ili radijvektori, čije su koordinate jednake vrijednostima težina uz odgovarajuće faktore. Kao oznaku za radijvektor realizacije varijable Z_j nadalje ćemo koristiti $\overrightarrow{z_j}$. Duljina radijvektora tada je jednaka korijenu iz komunalnosti.

Potrebno je, prije konačne interpretacije radijvektora varijabli u prostoru zajedničkih faktora, raspraviti geometrijski prikaz u koordinatnom sustavu koji uključuje i koordinatne osi koje predstavljaju specifične faktore. Takav prostor, kojeg razapinje m osi zajedničkih faktora i n osi specifičnih faktora, nazivamo *potpunim faktorskim prostorom*. Tada su koordinate vektora realizacije varijable dane s

$$\overrightarrow{z_j} = (a_{j1}, a_{j2}, \dots, a_{jm}, 0, \dots, 0, b_j, 0, \dots, 0),$$

[2] eng. *common factor space*

pri čemu se prvih m koordinata odnosi na vrijedosti na osima zajedničkih faktora, a posljednjih n koordinata na vrijednosti na koordinatnim osima specifičnih faktora – samo jedna vrijednost nije jednaka nuli. Koristimo oznaku z'_j specifično za prikaz varijable u potpunom faktorskom prostoru. Zbog jednostavnosti, pretpostavit ćemo da su zajednički faktori međusobno okomiti i, kao i inače, specifični faktori okomiti na njih.

Svi radijvektori realizacija varijabli u potpunom faktorskom prostoru su jedinične duljine, tj. vrijedi

$$|\overrightarrow{z'_j}| = \sqrt{a_{j1}^2 + \cdots + a_{jm}^2 + b_j^2} = 1,$$

a kosinus kuta između dvaju radijvektora realizacija varijabli dan je s

$$\cos\phi'_{jk} = \sum_{s=1}^{m} a_{js}\, a_{ks} = r'_{jk}, \tag{3.6}$$

pri čemu je s r'_{jk} označena procijenjena korelacija varijabli X_j i X_k, kako bismo je razlikovali od stvarne, opažene vrijednosti r_{jk}. Ovisno o tome koliko je dobro faktorski model prilagođen danim podacima, procijenjena korelacija će aproksimirati vrijednost opažene korelacije.

U prostoru zajedničkih faktora ćemo za radijvektore realizacija varijabli uzimati ortogonalne projekcije radijvektora realizacija varijabli iz potpunog faktorskog prostora, u oznaci $\overrightarrow{z''_j}$, čije su koordinate dane s

$$\overrightarrow{z''_j} = (a_{j1}, a_{j2}, \ldots, a_{jm}).$$

Kako je duljina radijvektora realizacije varijable u prostoru zajedničkih faktora jednaka korijenu iz komunalnosti, odnosno

$$|\overrightarrow{z''_j}| = \sqrt{a_{j1}^2 + \cdots + a_{jm}^2} = h_j,$$

zaključujemo da je duljina radijvektora realizacije varijabli u prostoru zajedničkih faktora manja od onih u potpunom faktorskom prostoru, što je očekivano obzirom da su projekcije vektora uvijek manje ili jednake duljine od vektora koji je projiciran.

Kosinus kuta između dvaju radijvektora realizacija varijabli u prostoru zajedničkih faktora dan je s

$$\cos\phi''_{jk} = \frac{\sum_{s=1}^{m} a_{js}\, a_{ks}}{h_j h_k} = \frac{r'_{jk}}{h_j h_k}. \tag{3.7}$$

Jasno je da je prethodni izraz generalno veći od onoga u (3.6), što je posljedica manjeg kuta između dva radijvektora realizacija varijabli u prostoru zajedničkih faktora. Kosinus kuta između dvaju radijvektora realizacija varijabli u prostoru zajedničkih faktora možemo

smatrati procjenom korelacije koju bi te varijable imale kada ne bi bilo utjecaja specifične varijance.

Iz prethodnog izraza dobivamo da je procijenjena korelacija između dviju varijabli jednaka skalarnom produktu njihovih radijvektora u prostoru zajedničkih faktora, tj.

$$r'_{jk} = h_j h_k \cos \phi''_{jk}.$$

Primjer 3.2.4. *Ilustrirajmo sada prethodno dobivene rezultate na primjeru dvije početne varijable, čiji je faktorski model dan s*

$$Z_1 = a_{11}F_1 + a_{12}F_2 + b_1U_1,$$
$$Z_2 = a_{21}F_1 + a_{22}F_2 + b_2U_2.$$

S obzirom na to da je potpuni faktorski prostor četverodimenzionalan, koordinate radijvektora (standardiziranih) realizacija varijabli, koji su tada jedinične duljine, dane su s

$$\overrightarrow{z'_1} = (a_{11}, a_{12}, b_1, 0),$$
$$\overrightarrow{z'_2} = (a_{21}, a_{22}, 0, b_2).$$

Procijenjenu korelaciju realizacija tih dviju varijabli u potpunom faktorskom prostoru računamo kao

$$r'_{12} = a_{11}a_{21} + a_{12}a_{22}. \tag{3.8}$$

Uzimanjem projekcija radijvektora realizacija varijabli iz potpunog faktorskog prostora na prostor zajedničkih faktora, koji je dvodimenzionalan, dobivamo radijvektore $\overrightarrow{z''_1}$ *i* $\overrightarrow{z''_2}$ *s koordinatama*

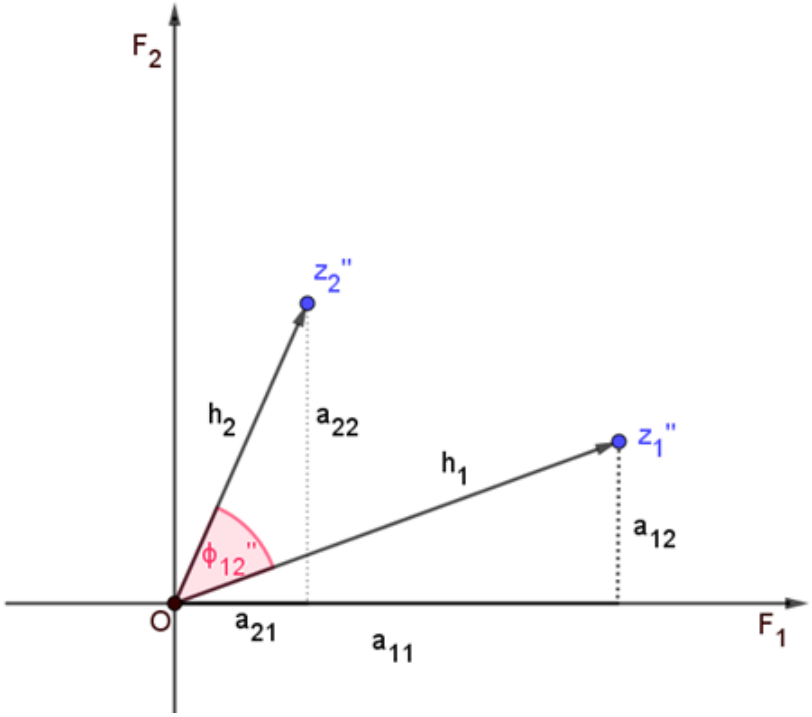
$$\overrightarrow{z''_1} = (a_{11}, a_{12}),$$
$$\overrightarrow{z''_2} = (a_{21}, a_{22}),$$

čije su duljine jednake drugom korijenu iz pripadajuće komunalnosti,

$$\rho_1 = \sqrt{a_{11}^2 + a_{12}^2} = \sqrt{h_1^2} = h_1,$$
$$\rho_2 = \sqrt{a_{21}^2 + a_{22}^2} = \sqrt{h_2^2} = h_2.$$

Prostor zajedničkih faktora, razapet faktorima F_1 *i* F_2*, koji su međusobno nekorelirani, prikazan je na slici 3.1. Faktori su prikazani kao međusobno okomiti jedinični vektori.*

Kosinus kuta ϕ''_{12} *kojeg zatvaraju vektori* $\overrightarrow{z''_1}$ *i* $\overrightarrow{z''_2}$ *računamo prema formuli* (3.7) *i dobivamo*



Slika 3.1: Dvofaktorski model

$$\cos\phi''_{12} = \frac{a_{11}}{h_1}\cdot\frac{a_{21}}{h_2} + \frac{a_{12}}{h_1}\cdot\frac{a_{22}}{h_2} = \frac{1}{h_1h_2}(a_{11}a_{21} + a_{12}a_{22}).$$

Uvrštavanjem (3.8) *u prethodni izraz, nastaje*

$$r'_{12} = h_1h_2\cos\phi''_{12}.$$

Time je jasno da je procjena vrijednosti korelacije dviju varijabli jednaka skalarnom produktu radijvektora realizacija varijabli u prostoru zajedničkih faktora.

Poglavlje 4

Određivanje broja faktora u modelu

Već smo ranije utvrdili da je jedan od glavnih problema faktorske analize određivanje broja faktora, odnosno dimenzije faktorskog prostora, stoga ćemo u ovo poglavlju definirati neke od metoda kojima je moguće odrediti dimenziju faktorskog modela na podacima koje promatramo.

4.1 Guttman-Kaiserov kriterij

Jedan od čestih kriterija za određivanja broja faktora u faktorskom modelu jest tzv. *Guttman-Kaiserov kriterij*. Temelji se na svojstvenim vrijednostima matrice korelacija, odnosno njenim karakterističnim korijenima, koji upućuju na to koliko je određeni faktor značajan pri objašnjavanju podataka na temelju kojih procjenjujemo faktorski model.

Faktor čija je svojstvena vrijednost jednaka jedan objašnjava jednaku količinu varijance u modelu kao i pojedina početna varijabla [10]. Ideja ove metode je zadržati u modelu one faktore koji imaju svojstvene vrijednosti veće ili jednake jedan.

Česta primjedba Guttman-Kaiserovom kriteriju jest zadržavanje prevelikog broja faktora u modelu, odnosno precjenjivanje broja dimenzija [5], stoga ga ne bi trebalo koristiti kao samostalan kriterij.

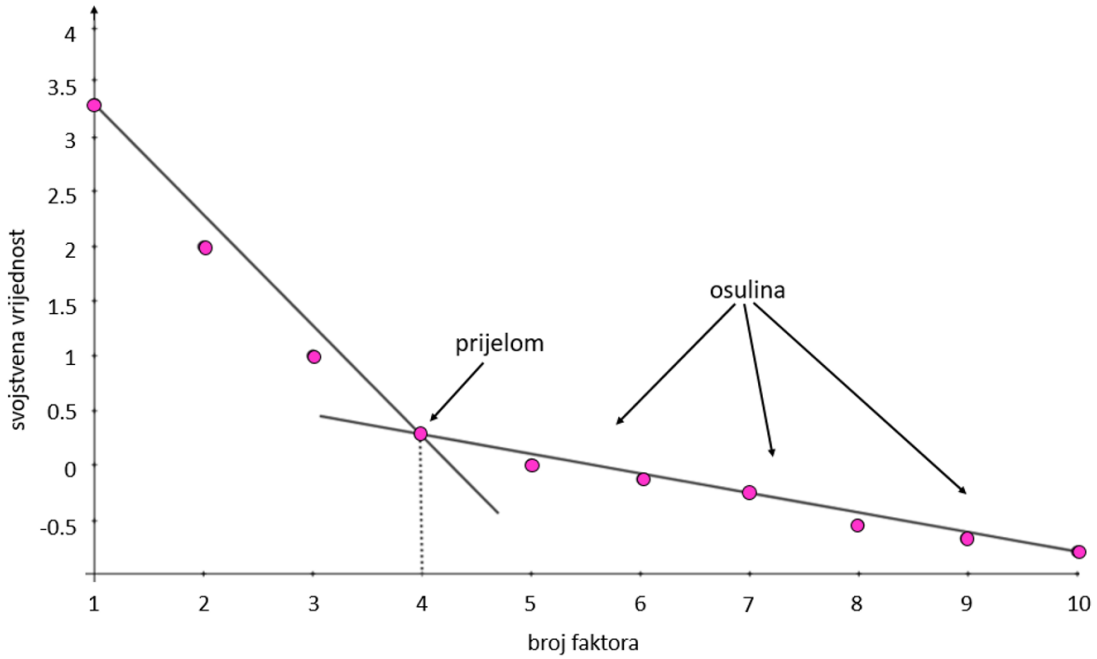
4.2 Scree plot

Scree plot[1] naziv je za graf koji na osi apscisa ima prikaz broja faktora u faktorskom modelu, a na ordinatnoj osi svojstvene vrijednosti matrice korelacija. Faktori se nanose redom prema veličini svojstvene vrijednosti, od najveće, a svaka sljedeća ima vrijednost manju od prethodne [6].

[1] eng. *scree*: stožasta nakupina osutog kamenja u podnožju strme padine, tzv. osulina

Ideja je promatranjem dobivenog grafa procijeniti posljednji osjetni pad svojstvenih vrijednosti, tj. identificirati *točku prijeloma* [2] [9]. One točke grafa koje slijede nakon prijeloma nazivamo *osulinom* [3]. U modelu zadržavamo onaj broj faktora koji je jednak broju točaka prije prijeloma, interpretirajući ga kao broj faktora koji su uspjeli objasniti značajan dio (zajedničke) varijance, a doprinos svakog idućeg faktora nije značajan.

Ova metoda često je korištena zbog svoje jednostavnosti, no, nailazi na kritike upravo zbog subjektivnosti koju zahtjeva, primjerice u situacijama poput postepenog pada svojstvenih vrijednosti, bez očiglednog prijeloma, ili kada postoji više od jednog prijeloma [9].



Slika 4.1: Primjer scree plota na hipotetskoj faktorskoj analizi s 10 varijabli

Na slici 4.1 dan je primjer scree plota na 10 hipotetskih varijabli. Točka prijeloma identificirana je kada je broj faktora jednak četiri, iz čega zaključujemo da je većina varijance objašnjena s prva tri faktora - stoga bismo, prema dobivenim rezultatima ove metode, odabrali faktorski model s tri faktora.

[2] eng. *elbow*
[3] eng. *scree*

4.3 Paralelna analiza

Paralelna analiza je metoda koja je također, kao i prethodne dvije, zasnovana na promatranju svojstvenih vrijednosti matrice korelacija, i zapravo je svojevrsna modifikacija Guttman-Kaiserovog kriterija. Pretpostavka ove metode jest da u modelu treba zadržati samo one dimenzije čije su svojstvene vrijednosti veće od svojstvenih vrijednosti koje dobivamo iz slučajnih podataka generiranih s analognim karakteristikama - tzv. *paralelnih uzoraka* [9].

Ideja u pozadini ove metode jest, u kontekstu promatranja cijele populacije, da su svojstvene vrijednosti matrice korelacija nekoreliranih varijabli jednake 1. Kada se radi o slučajnom uzorku iz te populacije, kao rezultat greške uzorkovanja, svojstvene vrijednosti korelacijske matrice variraju oko 1, neke imaju veću, a neke manju vrijednost. Na temelju više slučajnih uzoraka iz iste populacije, teoretski je moguće konstruirati empirijsku uzoračku distribuciju očekivanih svojstvenih vrijednosti i svaku opaženu svojstvenu vrijednost usporediti s dobivenom distribucijom pripadne svojstvene vrijednosti. Ukupan broj opaženih svojstvenih vrijednosti koje su značajno veće od onih očekivanih, dobivenih ponovnim uzorkovanjem, uzimamo za broj faktora u faktorskom modelu [5].

Prvo je potrebno izračunati svojstvene vrijednosti matrice korelacija dobivene na stvarnim podacima. Nakon toga, slijedi simulacija K paralelnih slučajnih uzoraka i računanje svojstvenih vrijednosti njihovih korelacijskih matrica. Kako je u praksi $K > 1$, odnosno generiramo više paralelnih uzoraka, za "teoretsku" svojstvenu vrijednost možemo uzeti aritmetičku sredinu odgovarajućih svojstvenih vrijednosti - tzv. *kriterij aritmetičke sredine*. Kako je za kriterij aritmetičke sredine utvrđeno da ima tendenciju ka precjenjivanju broja dimenzija, jedna od mogućih alternativa jest *kriterij 95. percentila*, koji podrazumijeva uspoređivanje svojstvenih vrijedosti stvarnih podataka sa svojstvenim vrijednostima slučajnih podataka sa 95. percentila (rjeđe, u praksi se koriste druge vrijednosti percentila, primjerice 90. ili 99.). Konačno, za dimenziju faktorskog modela uzima se broj stvarnih svojstvenih vrijednosti koje su veće od pripadnih slučajnih svojestvenih vrijednosti [9].

Generiranje paralelnih slučajnih uzoraka uzima u obzir pretpostavku o normalnoj distribuiranosti podataka. U slučaju kada distribucija stvarnih podataka značajnije odstupa od normalne, moguće je generiranje paralelnih slučajnih uzoraka kao permutacija stvarnih podataka, što osigurava da paralelni slučajni uzorci zadrže istu distribuciju kao i stvarni podaci (tzv. *neparametarska paralelna analiza*) [9].

4.4 *B*-koeficijent

Metoda *B*-koeficijenta (eng. *B-Coefficient, Coefficient of belonging*) koristi se za raspodjelu danih varijabli po grupama temeljenu na pretpostavci da su korelacije varijabli unutar same grupe veće od korelacija tih varijabli s preostalima. *B*-koeficijent definiramo kao 100

puta omjer prosječne vrijednosti korelacije varijabli unutar grupe (podskupa) i prosječne korelacije s preostalim varijablama. Metoda je preuzeta iz [4].

Uvodimo pojam argumenta od B pod kojim podrazumijevamo podskup varijabli za koje računamo B-koeficijent, te komplementa od B, koji predstavlja skup komplementaran argumentu. Vrijednost B-koeficijenta skupa koji sadrži varijable Z_1, Z_2, Z_3 označavamo s $B(Z_1, Z_2, Z_3)$, ili skraćeno $B(1, 2, 3)$.

Kako bismo precizno iskazali formulu za izračun B-koeficijenta nekog skupa varijabli, bit će korištena sljedeća notacija:

n = ukupan broj varijabli u skupu,
U = podskup varijabli koje su sadržane u argumentu od B,
p = broj varijabli u skupu U,
j, k = koristimo za indeksiranje elemenata skupa U ,
U^C = skup komplementaran skupu U,
a = koristimo za indeksiranje elemenata skupa U^C.

Sa S označimo sumu svih korelacija elemenata skupa U,

$$S = \sum_{j<k} r_{jk}. \tag{4.1}$$

Ova oznaka podrazumijeva sumu po svim korelacijama r_{jk} takvima da je j uvijek strogo manji od k (kako bismo svaku željenu korelaciju ubrojali točno jednom). Primjerice, kod izračuna $B(1, 2, 3, 4)$ koristit ćemo sumu $S = r_{12} + r_{13} + r_{14} + r_{23} + r_{24} + r_{34}$.

Sumu svih korelacija varijabli iz skupa U s varijablama iz skupa U^C označimo s T,

$$T = \sum_{j,a} r_{ja}. \tag{4.2}$$

Tako ćemo, za isti primjer kao prethodni, uz podskupove $U = \{Z_1, Z_2, Z_3, Z_4\}$ i $U^C = \{Z_5, Z_6\}$, T računati formulom $T = r_{15} + r_{16} + r_{25} + r_{26} + r_{35} + r_{36} + r_{45} + r_{46}$.

Broj elemenata sume S jednak je broju dvočlanih podskupova skupa s p elemenata, odnosno

$$\binom{p}{2} = \frac{p(p-1)}{2},$$

dok je broj korelacija koje zbrajamo u sumi T jednak

$$p(n-p).$$

Konačno, formula za izračun B-koeficijenta varijabli sadržanih u skupu U jest:

$$B(U) = 100 \cdot \frac{\frac{S}{\binom{p}{2}}}{\frac{T}{p(n-p)}} = \frac{200(n-p)S}{(p-1)T}. \tag{4.3}$$

Jasno je da vrijednost B-koeficijenta ne ovisi o redoslijedu ulaska varijabli u argument od B, no trenutak ulaska određene varijable u argument od B ima utjecaj. Primjerice, vrijedi $B(1,3,4) = B(1,4,3)$, dok $B(1,3,2)$ ne mora nužno biti jednako kao $B(1,3,4,2)$.

U slučaju kada je vrijednost B-koeficijenta skupa U jednaka 100, zaključujemo da je prosječna vrijednost korelacija varijabli sadržanih unutar skupa U jednaka prosječnoj vrijednosti korelacija s ostalim varijablama, onima iz skupa U^C. Jasno je da takav način grupiranja varijabli u skupove U i U^C nije dobar, stoga se za uobičajenu praksu uzima mogućnost pripadanja varijabli u određeni skup samo u slučaju kada je vrijednost B-koeficijenta skupa minimalno 130.

Grupiranje varijabli

Proces grupiranja varijabli započinje izračunom svih korelacija i uzimanjem dviju varijabli s najvećom međusobnom korelacijom. Uz njih dodajemo varijablu čija je suma korelacija s prethodne dvije varijable najviša. Postupak se nastavlja na isti način, u argument od B uvijek dodajemo onu varijablu koja ima najveće korelacije s varijablama koje su već u argumentu, dok ne dođe do naglog pada vrijednosti B-koeficijenta. Tada izbacujemo varijablu nakon koje je došlo do pada vrijednosti, te možemo pokušati s dodavanjem neke druge varijable, ili završiti sa formiranjem grupe. Kada odredimo elemente prve grupe, od preostalih uzimamo dvije varijable koje imaju najvišu korelaciju te ponovimo isti postupak prilikom formiranja druge grupe varijabli. Proces se nastavlja dok nisu grupirane sve početno dane varijable.

Primjer 4.4.1. *Neka je* $\mathbf{R}$ *matrica korelacija 5 varijabli,*

$$\mathbf{R} = \begin{pmatrix} 1 & 0.02 & 0.96 & 0.42 & 0.01 \\ 0.02 & 1 & 0.13 & 0.71 & 0.85 \\ 0.96 & 0.13 & 1 & 0.50 & 0.11 \\ 0.42 & 0.71 & 0.50 & 1 & 0.79 \\ 0.01 & 0.85 & 0.11 & 0.79 & 1 \end{pmatrix}.$$

Radi jednostavnosti, varijable ćemo označavati redom s Z_1, Z_2, Z_3, Z_4, Z_5.

Pokušajmo podijeliti varijable u grupe koristeći se metodom B-koeficijenta. Vrijedi: $n = 5$. *Najveća korelacija jest između varijable* Z_1 *i* Z_3 *, stoga tražimo još jednu varijablu čija je suma korelacija s prethodne dvije najveća. Taj uvjet ispunjava varijabla* Z_4*, čije korelacije s prvom i trećom varijablom u sumi daju* 0.92.

Sada je $p = 3$ *i vrijedi:*

$$U = \{Z_1, Z_3, Z_4\},$$
$$U^C = \{Z_2, Z_5\}.$$

Izračunajmo sada S *i* T *prema formulama* (4.1) *i* (4.2).

$$S = 0.96 + 0.42 + 0.5 = 1.88,$$
$$T = 0.02 + 0.01 + 0.13 + 0.11 + 0.71 + 0.79 = 1.77.$$

Uvrštavanjem prethodno dobivenih vrijednosti u formulu (4.3)*, dobivamo vrijednost B-koeficijenta:*

$$B(U) = \frac{200 \cdot 2 \cdot 1.88}{2 \cdot 1.77} = 212,43.$$

Kako smo dobili vrijednost veću od 130, varijablu Z_4 *ostavljamo u skupu U i dalje dodajemo varijablu* Z_5 *koja u sumi ima najveće korelacije s varijablama* Z_1, Z_3 *i* Z_4*.*

Za skupove

$$U = \{Z_1, Z_3, Z_4, Z_5\},$$
$$U^C = \{Z_2\},$$

računom dobivamo da je $S = 2.79$ *i* $T = 1.71$*, iz čega slijedi da je vrijednost B-koeficijenta* $B(U) = 108.77$*, što je manje od 130, pa varijablu* Z_5 *izbacujemo iz skupa U. Također, dodavanjem varijable* Z_2 *u skup U, vrijednost B-koeficijenta nije dovoljna kako bismo je zadržali u skupu (iznosi* 103.79*), stoga završavamo s formiranjem prve grupe, i zaključujemo da jednu grupu čine varijable* Z_1, Z_3, Z_4*, a preostale dvije varijable* Z_2 *i* Z_5 *drugu grupu.*

Poglavlje 5

Izračun faktorskog modela

Drugi osnovni problem faktorske analize, nakon što smo odredili broj faktora u modelu, je procjena težina faktora. U ovom poglavlju raspravljamo broj mogućih rješenja, navodimo neke od metoda procjene faktorskih težina koje su u praksi najčešće korištene i spominjemo mogućnost rotacije faktorskog prostora, preuzeto iz [3].

5.1 Mnogostrukost rješenja

Pretpostavimo da je $\mathbf{G}$ ortogonalna matrica. Tada faktorski model (2.4) možemo zapisati i kao

$$\mathbf{X} = (\mathbf{AG})(\mathbf{G}^T\mathbf{F}) + \mathbf{BU} + \mu, \tag{5.1}$$

iz čega je jasno da on nije jedinstveno određen. Dakle, ukoliko je model s faktorima $\mathbf{F}$ i njihovim težinama $\mathbf{A}$ adekvatan za početno dane varijable $\mathbf{X}$, svaki drugi model s faktorima $\mathbf{G}^T\mathbf{F}$ i njihovim težinama $\mathbf{AG}$ je jednako valjano rješenje. Iako nejedinstvenost rješenja otežava izračun samih težina faktora, ona dozvoljava odabir onog rješenja koje ćemo najlakše interpretirati. Množenje vektora $\mathbf{F}$ ortogonalnom matricom geometrijski interpretiramo kao rotaciju koordinatnog sustava koji je razapet faktorima $\mathbf{F}$.

Nejedinstvenost rješenja onemogućuje izračun faktorskog modela s obzirom na to da postoji beskonačno rješenja. Iz tog razloga postavljamo neke restrikcije kako bismo došli do (jedinstvenog) rješenja koje će zadovoljavati

$$\mathbf{\Sigma} = \mathbf{AA}^T + \mathbf{\Psi}, \tag{5.2}$$

a tek onda iskorištavamo mogućnost rotiranja faktora. Primjerice, možemo zahtijevati da je matrica

$$\mathbf{A}^T\mathbf{\Psi}^{-1}\mathbf{A} \tag{5.3}$$

dijagonalna [3].

Kako je m broj faktora, a n broj početno danih varijabli, broj slobodnih parametara u (5.2) je $mn + n$. Ograničenje dano u (5.3), obzirom na to da zahtijevamo da je matrica dijagonalna, utječe na $\frac{1}{2}m(m-1)$ parametara. Time dobivamo da je broj stupnjeva slobode faktorskog modela s m faktora jednak:

$$\begin{aligned} d &= \Big[\text{broj neograničenih parametara unutar } \mathbf{\Sigma}\Big] - \Big[\text{broj ograničenih parametara unutar } \mathbf{\Sigma}\Big] \\ &= \frac{1}{2}n(n+1) - (mn + n - \frac{1}{2}m(m-1)) \\ &= \frac{1}{2}(n-m)^2 - \frac{1}{2}(m+n). \end{aligned}$$

U slučaju kada je $d < 0$, faktorski model je neodređen - postoji beskonačno rješenja jer taj slučaj podrazumijeva da je broj faktora veći od broja početnih varijabli. Kada je $d = 0$ faktorski model je jedinstven, do na rotacije faktora. Posljednji slučaj, $d > 0$, koji je i najčešći, ne određuje egzaktno rješenje, stoga se uzimaju aproksimativna rješenja.

Broj stupnjeva slobode, d, daje gornju granicu broja faktora u modelu. Primjerice, ukoliko imamo šest početnih varijabli, modeli s jednim ili dva faktora dobiveni su aproksimativno, dok je model s tri faktora jedinstveno određen ($d = 0$), ako izuzmemo mogućnost rotacije faktora. Modeli s četiri i više faktora nisu mogući, što je i smisleno u kontekstu osnovne ideje faktorske analize o smanjivanju dimenzije podataka koji su dani kroz početne varijable.

Primjer 5.1.1. *Želimo pronaći jednofaktorski model prilagođen za slučaj tri početne varijable, odnosno vrijedi $n = 3$ i $m = 1$. Računamo*

$$d = \frac{1}{2}(n-m)^2 - \frac{1}{2}(m+n) = 2 - 2 = 0,$$

iz čega znamo da će model biti jedinstveno određen, do na rotacije faktora. Iz matrica

$$\mathbf{A} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} \quad i \quad \mathbf{\Psi} = \begin{bmatrix} \psi_{11} & 0 & 0 \\ 0 & \psi_{22} & 0 \\ 0 & 0 & \psi_{33} \end{bmatrix}$$

dobivamo

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}s & \sigma_{13}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \sigma_{23}^2 \\ \sigma_{31}^2 & \sigma_{32}^2 & \sigma_{33}^2 \end{bmatrix} = \begin{bmatrix} a_{11}^2 + \psi_{11} & a_{11}a_{21} & a_{11}a_{31} \\ a_{11}a_{21} & a_{21}^2 + \psi_{22} & a_{21}a_{31} \\ a_{11}a_{31} & a_{21}a_{31} & a_{31}^2 + \psi_{33} \end{bmatrix}.$$

Primijetimo da je ograničenje zadano u (5.3) *trivijalno zadovoljeno zbog $m = 1$ (matrica $\mathbf{A}^T\mathbf{\Psi}^{-1}\mathbf{A}$ je dimenzije 1×1 čime je osigurano da je dijagonalna). Izjednačavanjem odgovarajućih elemenata matrica, dobivamo da su formule za težine faktora jedinstveno dane s*

$$a_{11}^2 = \frac{\sigma_{12}^2\sigma_{13}^2}{\sigma_{23}^2},\ \ a_{21}^2 = \frac{\sigma_{12}^2\sigma_{23}^2}{\sigma_{13}^2},\ \ a_{31}^2 = \frac{\sigma_{13}^2\sigma_{23}^2}{\sigma_{12}^2},$$

iz kojih dalje računamo jedinstvene varijance,

$$\psi_{11} = \sigma_{11}^2 - a_{11}^2,\quad \psi_{22} = \sigma_{22}^2 - a_{21}^2,\quad \psi_{33} = \sigma_{33}^2 - a_{31}^2.$$

Kako se radi o jednofaktorskom modelu, jedina moguća rotacija definirana je za $\mathbf{G} = -1$ *pa je drugo rješenje dano matricom težina faktora* $-\mathbf{A}$.

5.2 Procjena težina faktora

Kako bismo postavili faktorski model, potrebno je procijeniti vrijednosti težina faktora $\mathbf{A}$ i jedinstvenih varijanci $\mathbf{\Psi}$, tj. procijeniti vrijednosti matrica $\hat{\mathbf{A}}$ i $\hat{\mathbf{\Psi}}$, koje će zadovoljavati izraz

$$\mathbf{S} = \hat{\mathbf{A}}\hat{\mathbf{A}}^T + \hat{\mathbf{\Psi}}, \tag{5.4}$$

gdje je $\mathbf{S}$ uzoračka kovarijacijska matrica od $\mathbf{X}$ definirana s

$$\mathbf{S} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^T.$$

Procjenom vrijednosti matrice $\hat{\mathbf{A}}$, jedinstvene varijance određene su izrazom

$$\hat{\psi}_{jj} = s_{jj}^2 - \sum_{l=1}^{m}\hat{a}_{jl}^2, \tag{5.5}$$

pri čemu je $\sum_{l=1}^{m}\hat{a}_{jl}^2$ zapravo procjenitelj komunalnosti varijable X_j, tj. $\hat{h}_j^2$, pa izraz (5.5) možemo zapisati i kao

$$\hat{\psi}_{jj} = s_{jj}^2 - \hat{h}_j^2.$$

Ukoliko su s $\mathbf{Y}$ dane standardizirane vrijednosti početnih varijabli $\mathbf{X}$, veza procjenitelja faktorskih težina i jedinstvenih varijanci dana je s

$$\hat{\mathbf{A}}_Y = \mathbf{D}^{-1/2}\hat{\mathbf{A}}_X \quad \text{i} \quad \hat{\mathbf{\Psi}}_Y = \mathbf{D}^{-1}\hat{\mathbf{\Psi}}_X,$$

pri čemu je $\mathbf{D} = \operatorname{diag}(s_{11}, s_{22}, \ldots, s_{nn})$, gdje s_{jj} označava uzoračku varijancu varijable X_j. Tada za korelacijsku matricu od $\mathbf{X}$, matricu $\mathbf{R}$, vrijedi

$$\mathbf{R} = \hat{\mathbf{A}}_Y\hat{\mathbf{A}}_Y^T + \hat{\mathbf{\Psi}}_Y,$$

odnosno kod standardiziranih varijabli izrazi za kovarijacijsku matricu i matricu korelacija se podudaraju.

Metoda maksimalne vjerodostojnosti (MLE)

Jedna od često korištenih metoda procjene težina faktora jest *metoda maksimalne vjerodostojnosti* ili *MLE - metoda*[1].

Općenito, metoda maksimalne vjerodostojnosti koristi se u svrhu procjene nepoznatih parametara distribucije vjerojatnosti. Neka je $\mathbf{X} = (X_1, \ldots, X_N)$ slučajni uzorak duljine N, $(n \geq 1)$, s funkcijom gustoće $f(x; \theta)$. Cilj je procijeniti vrijednost nepoznatog parametra θ dimenzije m $(m \geq 1)$.

Ako je $\mathbf{x} = (x_1, \ldots, x_N)$ jedna realizacija od $\mathbf{X}$, tada vjerodostojnost (eng. *likelihood*) definiramo kao funkciju

$$L(\mathbf{X}; \theta) = \prod_{i=1}^{N} f(x_i; \theta). \tag{5.6}$$

Statistiku $\hat{\theta} \equiv \hat{\theta}(\mathbf{X})$ nazivamo *procjeniteljem maksimalne vjerodostojnosti* (eng. *maximum likelihood estimator, MLE*) ako vrijedi

$$L(\hat{\theta}) = \max_{\theta} L(\mathbf{X}; \theta). \tag{5.7}$$

Najčešće, radi jednostavnijeg maksimiziranja, promatramo log-vjerodostojnost

$$l(\mathbf{X}; \theta) = \ln L(\mathbf{X}; \theta), \tag{5.8}$$

budući da je prirodni logaritam strogo rastuća injektivna funkcija.

Specijalno kod procjene težina faktora metodom maksimalne vjerodostojnosti, pretpostavljamo da slučajni uzorak dolazi iz višedimenzionalnog normalnog modela $N_n(\mu, \Sigma)$, gdje je $\theta = (\mu, \Sigma)$ nepoznati parametar, tj. da su početne varijable nomalano distribuirane. Tada je funkcija log-vjerodostojnosti dana s

$$l(\mathbf{X}; \mu, \Sigma) = -\frac{N}{2} \ln |2\pi\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)\Sigma^{-1}(x_i - \mu)^T, \tag{5.9}$$

što još možemo zapisati kao

$$l(\mathbf{X}; \mu, \Sigma) = -\frac{N}{2} \ln |2\pi\Sigma| - \frac{N}{2} tr(\Sigma^{-1} S) - \frac{N}{2} (\overline{x} - \mu)\Sigma^{-1}(\overline{x} - \mu)^T, \tag{5.10}$$

pri čemu je S empirijska kovarijacijska matrica [3].

Zamijenimo li μ sa pripadajućim procjeniteljem maksimalne vjerodostojnosti $\hat{\mu} = \overline{x}$, a Σ zapišemo kao $\Sigma = AA^T + \Psi$, prema (2.9), konačno dobivamo

$$l(\mathbf{X}; \hat{\mu}, A, \Psi) = -\frac{N}{2} \Big[\ln \{|2\pi(AA^T + \Psi)|\} + \mathrm{tr}\{(AA^T + \Psi)^{-1} S\} \Big]. \tag{5.11}$$

Maksimizacija prethodne funkcije poprilično je kompliciranа već i za jednofaktorski model $(m = 1)$, pa se za izračun najčešće koriste iterativni numerički algoritmi [3].

[1] eng. *the maximum likelihood method*

Metoda glavnih komponenata

Metoda glavnih komponenata[2] (preuzeta iz [3]) jest metoda čiji je glavni cilj aproksimirati matricu težina $\hat{\mathbf{A}}$. Najprije, uzoračku kovarijacijsku matricu definiranu s

$$\mathbf{S} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^T,$$

dijagonaliziramo, odnosno pronalazimo njezinu spektralnu dekompoziciju,

$$\mathbf{S} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T,$$

pri čemu je $\mathbf{\Lambda}$ dijagonalna matrica čiji elementi predstavljaju svojstvene vrijednosti matrice $\mathbf{S}$ koje ćemo označavati s λ_i, dok su stupci matrice $\mathbf{\Gamma}$ svojstveni vektori matrice $\mathbf{S}$, označeni s γ_i. Također, prema (2.9) znamo da vrijedi

$$\mathbf{S} = \hat{\mathbf{A}}\hat{\mathbf{A}}^T + \hat{\mathbf{\Psi}}, \tag{5.12}$$

odnosno

$$\hat{\mathbf{A}}\hat{\mathbf{A}}^T = \mathbf{S} - \hat{\mathbf{\Psi}}, \tag{5.13}$$

Pod pretpostavkom da je m broj faktora u modelu koji želimo postaviti, pomoću vrijednosti prvih m svojstvenih vrijednosti i vektora procjenjujemo matricu $\mathbf{A}$,

$$\hat{\mathbf{A}} = \left[\sqrt{\lambda_1}\gamma_1 \ \ldots \ \sqrt{\lambda_p}\gamma_p\right].$$

Tada su dijagonalni elementi matrice $\mathbf{S} - \hat{\mathbf{A}}\hat{\mathbf{A}}^T$ procijenjene vrijednosti specifičnih varijanci $\hat{\psi}_j$, odnosno $\hat{\psi}_j = s_{jj} - \sum_{l=1}^{m}\hat{a}_{jl}^2$. Time je dana i aproksimacija matrice jedinstvene varijance u faktorskom modelu

$$\hat{\mathbf{\Psi}} = \begin{bmatrix} \hat{\psi}_1 & & & 0 \\ & \hat{\psi}_2 & & \\ & & \ddots & \\ 0 & & & \hat{\psi}_n \end{bmatrix}.$$

Sada znamo da su, prema definiciji, dijagonalni elementi matrice $\mathbf{S}$ isti kao i dijagonalni elementi matrice $\hat{\mathbf{A}}\hat{\mathbf{A}}^T + \hat{\mathbf{\Psi}}$, dok elementi koji su van dijagonale ne moraju nužno biti procijenjeni.

Rezidualna matrica faktorskog modela dobivenog ovom metodom dana je s

$$\mathbf{S} - (\hat{\mathbf{A}}\hat{\mathbf{A}}^T + \hat{\mathbf{\Psi}}), \tag{5.14}$$

[2] eng. *the principal component method, PC-method*

čije elemente možemo ograničiti vrijednostima

$$\sum_{i,j}(\mathbf{S} - \hat{\mathbf{A}}\hat{\mathbf{A}}^T - \hat{\mathbf{\Psi}})_{ij}^2 \leq \lambda_{p+1}^2 + \cdots + \lambda_n^2. \tag{5.15}$$

Tada je jasno da greška aproksimacije ovisi o veličini zanemarenih $n - p$ svojstvenih vrijednosti.

Dodatno, u slučaju kada su početne varijable dane u standardiziranom obliku, provodimo isti račun, ali na empirijskoj korelacijskoj matrici $\mathbf{R}$ umjesto kovarijacijske matrice $\mathbf{S}$.

5.3 Interpretacija faktora

Postavljanjem faktorskog modela, potrebna je interpretacija istog, kako bismo objasnili značenje dobivenih faktora i uvidjeli ima li on smisla, ovisno o sadržaju podataka koji su bili polazna točka za model. Računanjem korelacija dobivenih faktora s početno danim varijablama dobivamo uvid u njihov međusobni odnos. Neka je $\mathbf{P}_{XF}$ matrica čiji su elementi upravo korelacije između faktora i početnih varijabli. Kako je njihova kovarijacijska matrica jednaka

$$\mathbf{\Sigma_{XF}} = \mathrm{E}\{(\mathbf{AF} + \mathbf{BU})\mathbf{F}^T\} = \mathbf{A},$$

slijedi da je matrica $\mathbf{P}_{XF}$ dana s

$$\mathbf{P}_{XF} = \mathbf{D}^{-1/2}\mathbf{A},$$

pri čemu je $\mathbf{D} = \mathrm{diag}(\sigma_{X_1X_1}, \ldots, \sigma_{X_pX_p})$. Specijalno, ukoliko su početne varijable dane u standardiziranom obliku, vrijedi

$$\mathbf{P}_{XF} = \mathbf{A},$$

tj. faktore je moguće interpretirati kroz vrijednosti njihovih težina [3].

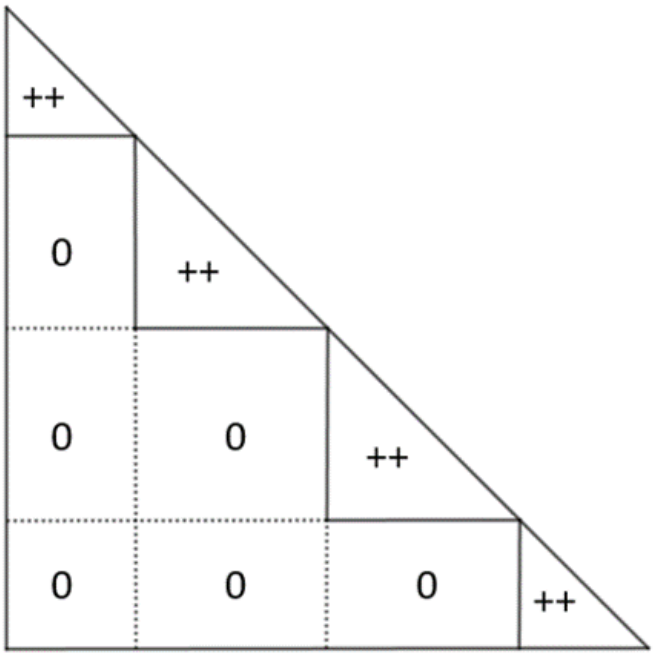
Kako je prostor zajedničkih faktora m-dimenzionalan, korelacije između faktora i početnih varijabli najbolje uočavamo na dvodimenzionalnim presjecima - ravninama, koje razapinju dva faktora, i projekcijama točaka varijabli na tu ravninu.

Rotacije

Kao što je pokazano u (5.1), na faktore je moguće primijeniti rotaciju, odnosno linearnu transformaciju koja će rezultirati ekvivalentnim podacima, ali potencijalno jednostavnijim za interpretaciju. Jedan od ciljeva rotacije jest maksimalno smanjiti broj negativnih težina faktora, jer je nerijetko lakše interpretirati pozitivne vrijednosti težina. Također, želimo što više težina svesti na nulu, ili blizu nule, i time smanjiti broj parametara koje je potrebno interpretirati. Idealan faktorski model sadržavao bi faktore čije su osi u geometrijskom prikazu ortogonalne. Tada su koordinate točaka koje predstavljaju početne varijable jednake

težinama faktora. Obzirom da je moguća rotacija osi oko ishodišta, same vrijednosti koordinata početnih varijabli moguće je zanemariti u smislu da zanemarimo osi koje prikazuju faktore, i promotrimo ponašanje varijabli, odnosno njihovih reprezentacija u faktorskom prostoru - grupiraju li se neke, kakve su im relativne udaljenosti jedne od drugih i slično. Uočavanjem određenog ponašanja varijabli postavljamo koodinatne osi na način koji će omogućavati najjednostavniju interpretaciju faktora u modelu. Jedna od mogućih rotacija jest *varimax* rotacija – rotacija za kut koja maksimizira sumu varijanci kvadrata težina faktora (prema [2]). Cilj je postići da svaka od varijabli ima značajnu samo jednu težinu, tj. da je povezujemo samo s jednim od faktora, dok su ostale vrijednosti težina faktora zanemarive, blizu nule. Time postižemo i jasniju raspodjelu početnih varijabli na podskupove varijabli koje bi bilo moguće zamijeniti samo jednom varijablom – faktorom.

Matrica korelacija varijabli koje bi originalno zadovoljavale ono što varimax rotacija nastoji postići shematski je prikazana na slici 5.1.



Slika 5.1: Shematski prikaz donjeg trokuta idealnog slučaja matrice korelacija

Znakovi pluseva u trokutima uz dijagonalu predstavljaju pozitivnu korelaciju među varijablama koje pripadaju istoj grupi, tj. njihove vrijednosti može objasniti isti faktor, dok su nule u pravokutnicima na mjestu korelacija među varijablama iz različitih grupa. Kako

je ova vrsta matrice korelacija gotovo nemoguća na stvarnim podacima kojima nastojimo prilagoditi faktorski model, smatramo je idealnom.

Poglavlje 6

Primjer

U ovom poglavlju primjenjujemo neke od prethodno navedenih metoda faktorske analize. Nakon postavljanja faktorskog modela korištenjem softverskog paketa R (vidi [7]), pokušavamo interpretirati faktore i objasniti vezu između početnih varijabli i dobivenih faktora.

6.1 Podaci

Podaci na kojima nastojimo postaviti faktorski model bave se analizom naslova novinskih članaka, odnosno emocijama koje isti pobuđuju u čitateljima. Niz kriterija na temelju kojih se procjenjuje sadržaj neke vijesti - kod nas novinskog članka, i određuju oni aspekti koje je potrebno naglasiti kroz naslov, nazivamo *vrijednost vijesti* [1]. Sva pojašnjenja pojmova vezanih uz temu emocija i vrijednosti vijesti navedena u ovom poglavlju preuzeta su iz [1], kao i ideja o postavljanju faktorskog modela upravo na tim podacima.

Dvije su osnovne uloge koje pridajemo naslovu novinskog članka: sumiranje njegova sadržaja i privlačenje pažnje čitatelja. Kako bi naslov privukao pozornost čitatelja, u njemu je često sadržana informacija koja daje uvid u emocionalni aspekt vijesti. Upravo iz tog razloga, kako bismo proučili postoji li veza između vrijednosti vijesti i emocija koje ona budi u čitatelju, koristimo set podataka koji sadrži 450 naslova novinskih članaka, preuzetih iz novina kao što su *New York Times*, *CNN*, *BBC News* i *Google News*.

Od trinaest početnih varijabli, šest je vezano uz emocije koje naslov pobuđuje u čitatelju (ljutnja, gađenje, strah, radost, tuga, iznenađenje[2]), a preostalih osam procjenjuje vrijednost vijesti (loša vijest, sukob, iznenađenje, zabava, drama, elita moći, dobra vijest[3]). Više o samom procesu mjerenja varijabli koje vrednuju vijesti moguće je pronaći u [1].

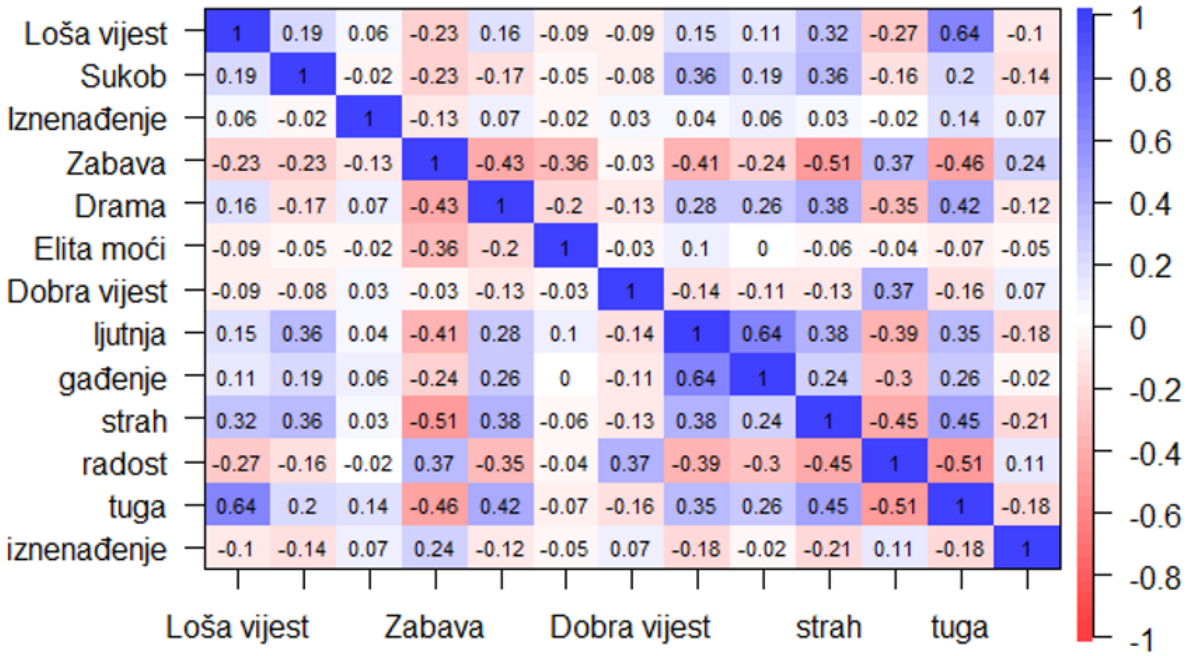
[1] eng. *news value*

[2] eng. *anger, disgust, fear, joy, sadness, surprise*

[3] eng. *bad news, conflict, surprise, entertainment, drama, the power elite, good news*

6.2 Izračun modela

Najprije računamo matricu korelacija varijabli i na slici 6.1 prikazujemo korelacije u bojama, sukladno legendi prikazanoj s desne strane.



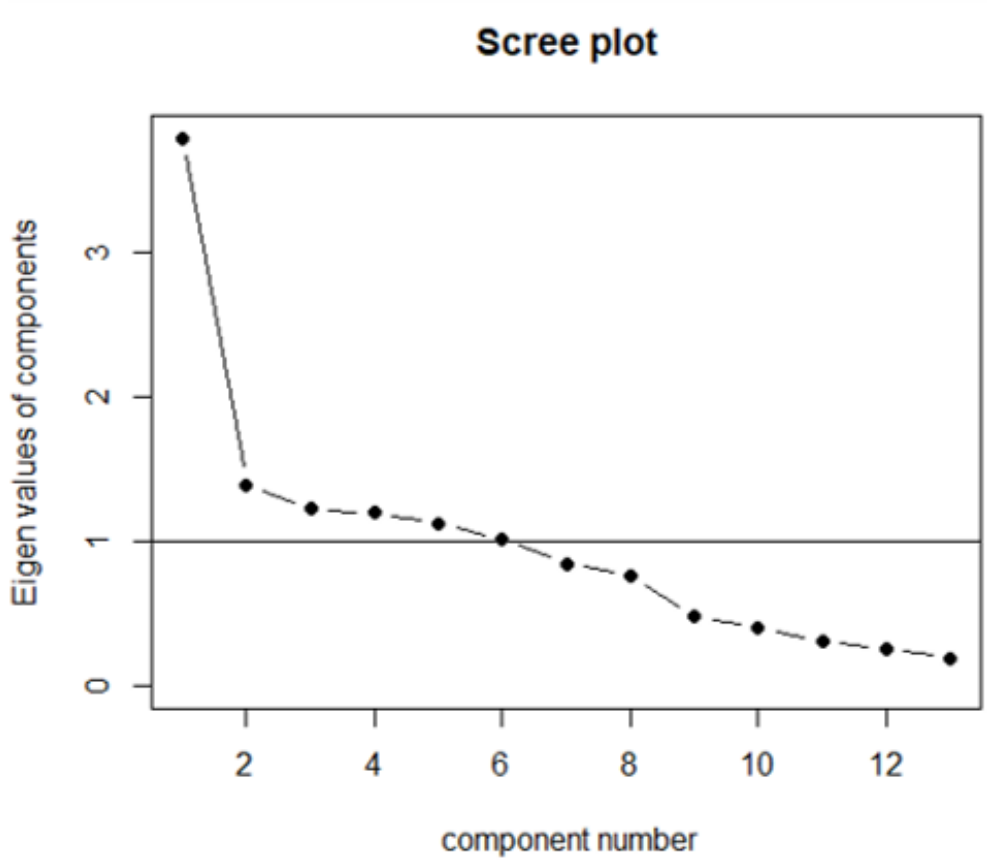
Slika 6.1: Korelacije među varijablama

Dalje, računamo svojstvene vrijednosti matrice korelacija (slika 6.2) i prikazujemo ih grafički na scree plotu (slika 6.3), grafu koji je opisan u 4.2. Na grafu je prikazan i horizontalan pravac za svojstvenu vrijednost jednaku jedan, koji ima ulogu Guttman-Kaiserovog kriterija, tj. dijeli svojstvene vrijedosti na one veće i manje od jedan.

Na grafu, kao i iz numeričkih vrijednosti, vidimo da je broj svojstvenih vrijednosti većih od jedan jednak šest, stoga bismo prema Guttman-Kaiserovom kriteriju za broj faktora uzeli šest. Promatranjem ponašanja svojstvenih vrijednosti nije moguće jednoznačno odrediti potreban broj faktora u modelu jer, osim prvog većeg osjetnog pada svojstvenih vrijednosti, idući veći pad sugerira broj faktora jednak osam, no i za broj faktora jednak šest možemo uočiti manji pad svojstvenih vrijednosti.

```
> eigenvalues$values
 [1] 3.7923367 1.3910891 1.2236750 1.1958446 1.1242452 1.0189775
 [7] 0.8481943 0.7580077 0.4778924 0.4067848 0.3122098 0.2572060
[13] 0.1935369
```

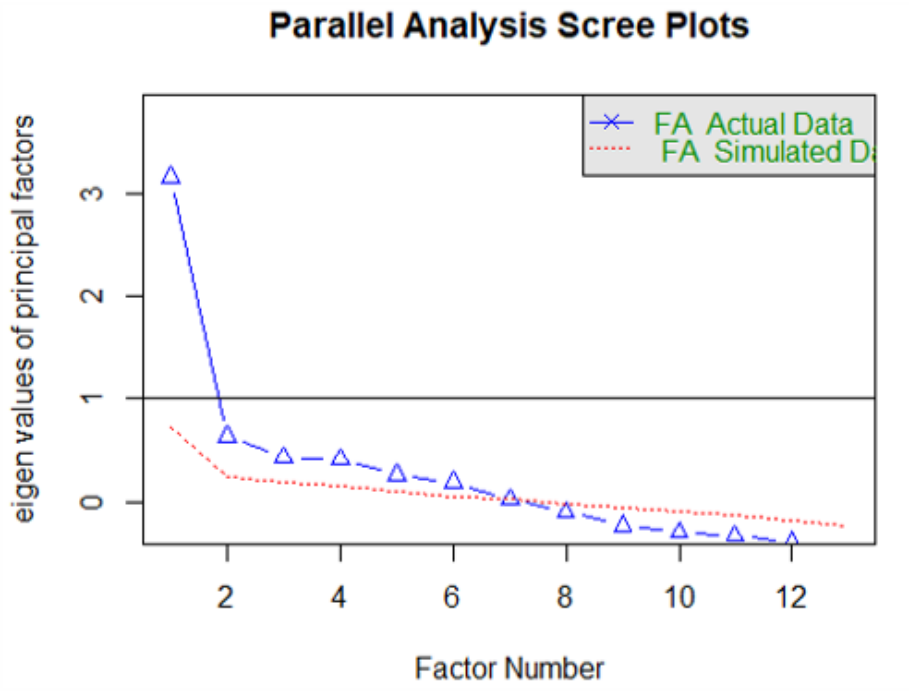
Slika 6.2: Svojstvene vrijednosti



Slika 6.3: Scree plot

Paralelna analiza, objašnjena u 4.3, kao metoda procjene broja faktora sugerira šest faktora u modelu (rezultati na slikama 6.4 i 6.5), što bismo uzeli i prema Guttman-Kaiserovu kriteriju, pa zaključujemo da je većina varijabilnosti objašnjena upravo pomoću šest faktora.

Korištenjem dodatnog paketa 'psych' [8] u R-u, koristimo već implementiranu funkciju



Slika 6.4: Grafički prikaz paralelne analize u R-u

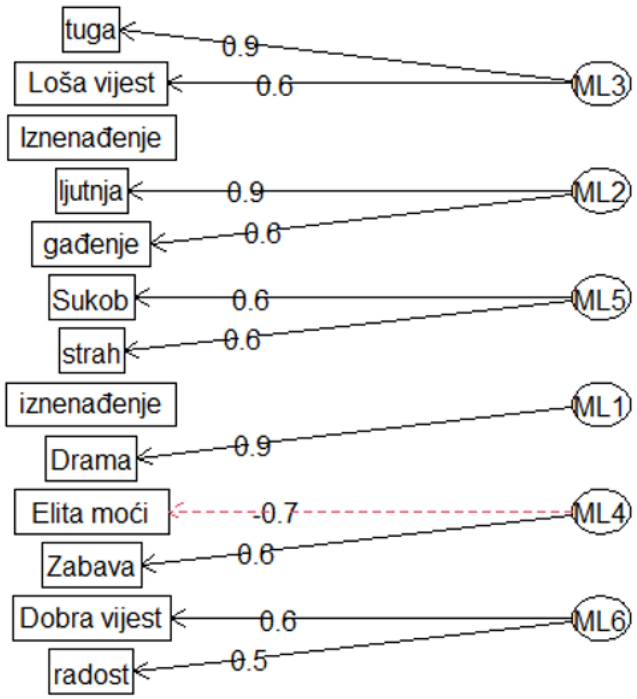
```
> fa.parallel(data_cor,n.obs=450,fm="mle",fa="fa")
Parallel analysis suggests that the number of factors =  6
 and the number of components =  NA
```

Slika 6.5: Rezultati paralelne analize u R-u

'fa' kako bismo izračunali parametre faktorskog modela sa šest faktora na dobivenoj matrici korelacija među varijablama. Računamo dva modela, jedan koji za metodu procjene procjene težina faktora koristi metodu maksimalne vjerodostojnosti, a drugi koji računa težine metodom glavnih komponenata. Dodatan parametar funkcije 'fa' koji koristimo jest vrsta rotacije, za koju odabiremo 'varimax' opciju.

Grafički prikaz veze početnih varijabli sa dobivenih šest faktora odnosno njihove korelacije, prikazujemo pomoću funkcije 'fa.diagram'.

Emocije su na dijagramima napisane malim početnim slovom kako bismo ih mogli razlikovati od vrijednosti vijesti. Na linijama koje spajaju faktor s početnom varijablom napisana je vrijednost njihove međusobne korelacije.

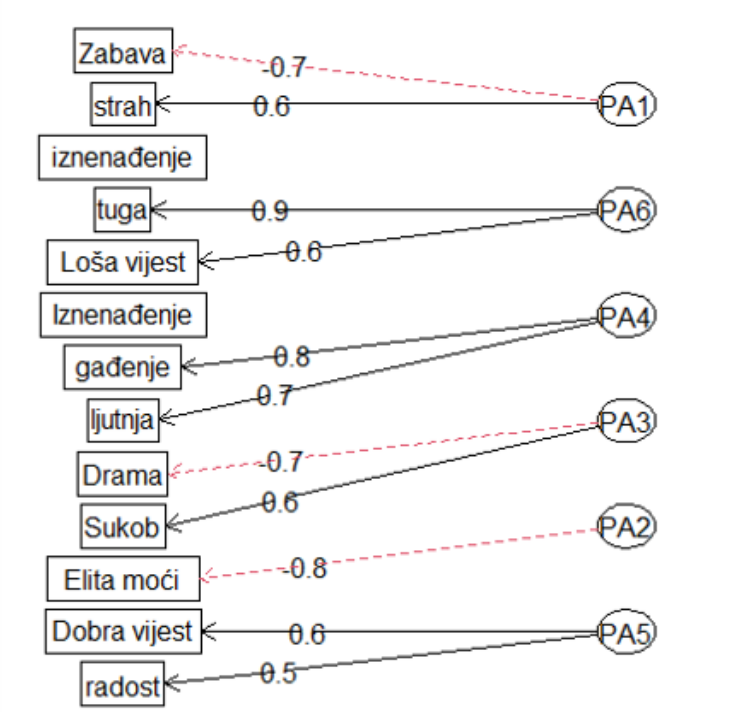


Slika 6.6: Faktorski model dobiven metodom maksimalne vjerodostojnosti

Za kvalitetniju interpretaciju dobivenih modela potrebna je stručnost, ali jasna je, primjerice u prvom modelu, povezanost loše vijesti s tugom, veza između gađenja i ljutnje kao emocija, negativna korelacija zabave sa faktorom s kojim je strah pozitivno koreliran u drugom modelu, radost kao emocija koju izaziva dobra vijest i slično.

Težine faktora, čija je apsolutna vrijednost veća od 0.1, prikazane su na slikama 6.8 i 6.9.

Kod koji je korišten u R-u za dobivanje prethodnih rezultata i izračun faktorskih modela prikazan je na slici 6.10.



Slika 6.7: Faktorski model dobiven metodom glavnih komponenata

```
> EFA_model_MLE$loadings

Loadings:
             ML3    ML2    ML5    ML1    ML4    ML6
Loša vijest   0.619          0.203
Sukob                0.209  0.640 -0.283
Iznenađenje   0.155                              0.106
Zabava       -0.282 -0.154 -0.549 -0.437  0.620 -0.123
Drama         0.191  0.197         0.943  0.119 -0.121
Elita moći                        -0.121 -0.725
Dobra vijest                                     0.645
ljutnja              0.942  0.277         -0.102
gađenje              0.607  0.138  0.110
strah         0.234  0.168  0.597  0.267         -0.175
radost       -0.303 -0.223 -0.284 -0.200  0.115  0.511
tuga          0.923  0.183  0.211  0.181         -0.180
iznenađenje  -0.102 -0.104 -0.216
```

Slika 6.8: Težine faktora dobivene metodom maksimalne vjerodostojnosti

```
> EFA_model_principal$loadings

Loadings:
             PA1    PA6    PA4    PA3    PA2    PA5
Loša vijest   0.193  0.619                       -0.102
Sukob         0.485         0.237  0.570  0.128
Iznenađenje          0.145
Zabava       -0.742 -0.245 -0.204  0.260  0.435 -0.152
Drama         0.356  0.170  0.202 -0.744  0.207 -0.125
Elita moći                                -0.815
Dobra vijest                                     0.616
ljutnja       0.334  0.105  0.730                -0.128
gađenje              0.110  0.819
strah         0.636  0.238  0.188         0.105 -0.162
radost       -0.347 -0.278 -0.247  0.136         0.518
tuga          0.297  0.905  0.165 -0.113         -0.183
iznenađenje  -0.292
```

Slika 6.9: Težine faktora dobivene metodom glavnih komponenata

```
data_cor <- cor(data, use = "pairwise.complete.obs" );
cor.plot(data_cor,numbers=TRUE, show.legend=TRUE);

eigenvalues <- eigen(data_cor);
eigenvalues$values;

scree(data_cor, factors = FALSE);
fa.parallel(data_cor,n.obs=m,fm="mle",fa="fa");

EFA_model_MLE <- fa(data, nfactors=6, fm='mle', rotate="varimax");
EFA_model_pa <- fa(data, nfactors=6, fm='pa',rotate="varimax");

fa.diagram(EFA_model_MLE);
fa.diagram(EFA_model_pa);

EFA_model_MLE$loadings;
EFA_model_pa$loadings;
```

Slika 6.10: Kod u R-u

Bibliografija

[1] M. P. Di Buono, J. Šnajder, B. Dalbelo Bašić, G. Glavaš, M. Tutek i N. Milic-Frayling, *Predicting News Values from Headline Text and Emotions*, Proceedings of the 2017 EMNLP Workshop on Natural Language Processing Meets Journalism (O. Popescu, C. Strapparava), Association for Computational Linguistics, Copenhagen, 2017, 1–6.

[2] S. Glen, *Varimax Rotation: Definition*, dostupno na `https://www.statisticshowto.com/varimax-rotation-definition` (lipanj 2021.).

[3] W. Härdle i L. Simar, *Applied multivariate statistical analysis*, Springer, Berlin, 2007.

[4] K. J. Holzinger i H. H. Harman, *Factor Analysis. A synthesis of factorial methods*, University of Chicago Press, Chicago, 1941.

[5] R. Hoyle i J. Duvall, *Determining the number of factors in exploratory and confirmatory factor analysis*, The SAGE Handbook of Quantitative Methodology for the Social Sciences (D. Kaplan), SAGE Publications, 2004, 302–317.

[6] R. Ledesma, P. Valero-Mora i G. Macbeth, *The Scree Test and the Number of Factors: a Dynamic Graphics Approach*, The Spanish Journal of Psychology **18** (2015).

[7] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021, `https://www.R-project.org/`.

[8] W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois, 2021, `https://CRAN.R-project.org/package=psych`, R package version 2.1.3.

[9] S. Subotić, *Pregled metoda za utvrđivanje broja faktora i komponenti (u EFA i PCA)*, Primenjena psihologija **6** (2013), 203–229.

[10] B. G. Tabachnick i L. S. Fidell, *Using Multivariate Statistics, 6th Edition*, Pearson/Allyn Bacon, Boston, 2007.

Sažetak

Faktorska analiza je statistička metoda čiji je cilj pronaći manji broj neopaženih varijabli (faktora) koje objašnjavaju što veći dio varijabilnosti početnoga skupa podataka. Početne varijable tada možemo prikazati kao linearne kombinacije faktora, iz čijih težina vidimo koliki je utjecaj određenog faktora na neku od varijabli. Razlikujemo eksploratornu i konfirmatornu faktorsku analizu, ovisno o tome jesu li nam faktori unaprijed poznati.

U radu je objašnjena podjela varijance unutar faktorskog modela na komunalnost i jedinstvenu varijancu. Radi boljeg razumijevanja faktorske analize, dana je geometrijska interpretacija faktorskog modela i prikaz faktorskog prostora na primjeru dva zajednička faktora. Dalje, objašnjene su neke od metoda procjene potrebnog broja faktora i njihovih težina te mogućnost rotacije faktora.

Na kraju rada ilustrirana je primjena eksploratorne faktorske analize u programskom jeziku R na skupu podataka koji analiziraju naslove novinskih članaka i osjećaje koje oni pobuđuju u čitatelju.

Summary

Factor analysis is a statistical method that aims to find a smaller number of unobserved variables (factors) that explain as much of the variability of the initial data set as possible. The original variables can be modelled as linear combinations of the factors, whose weights represent the influence of a particular factor on the variable. Depending on our assumptions about factors, there are two types of factor analysis, exploratory and confirmatory.

In this paper, the division of variance within the factor model into communality and unique variance is explained. A geometric interpretation of the factor model and an example of the factor space with two common factors are given. Methods for estimating number of factors and their loadings are explained, as well as rotations of the factors.

In the end, an example of exploratory factor analysis is illustrated on a dataset that analyzes relations among news values and emotions using programming language R.

Životopis

Rođena sam 19. siječnja 1997. godine u Karlovcu. Po završetku osnovnoškolskog obrazovanja u Osnovnoj školi Ivana Gorana Kovačića u Dugoj Resi, upisala sam smjer opće gimnazije u Srednjoj školi Duga Resa. Daljnje školovanje nastavila sam na Prirodoslovno-matematičkom fakultetu u Zagrebu, gdje sam završila preddiplomski sveučilišni studij matematike – nastavnički smjer, i nakon toga upisala diplomski sveučilišni studij Matematička statistika na istom fakultetu.