12 Paskaita. Faktorinė analizė.

12.1 Faktorinės analizės tikslai ir etapai

Faktorinės analizės užduotis – atsižvelgiant į kintamųjų tarpusavio koreliacijas, suskirstyti stebimus kintamuosius į grupes, kurias vienija koks nors tiesiogiai nestebimas (latentinis) faktorius. Pereidami nuo didelio skaičiaus kintamųjų prie faktorių mes koncentruojame informaciją, padarome ją labiau aprėpiamą. Patys faktoriai, pvz., kūribiškumas, altruizmas, erdvinė vaizduotė, lyderiavimas, dažnai negali būti tiesiogiai išmatuoti. Faktorinė analizė turi padėti juos nustatyti. Nagrinėsime atvejį, kai tyrėjas nežino, nei kokie faktoriai slypi už esamų kintamųjų, nei kiek jų yra. Tokia faktorinė analizė vadinama *tiriančiąja*. Taigi faktorinė analizė taikoma, kai pradinių kintamųjų yra labai daug. Faktorinė analizė dažnai taikoma kartu su kitais daugiamatės statistikos metodais, kai latentinių kintamųjų reikšmių įverčiai naudojami, kaip pradiniai duomenys regresinėje arba klasterinėje analizėje. Faktorinės analizės tikslas – pakeisti didelį pradinių kintamųjų skaičių kelių fakrorių, apibūdinančių stebimą reiškinį, rinkiniu. Faktorinės analizės etapai:

1. tikriname, ar duomenys tinka faktorinei analizei;
2. faktorių skaičiaus nustatymas bei faktorių skaičiavimo metodo parinkimas;
3. faktorių sukimas ir interpretavimas;
4. faktorių reikšmių šverčių skaičiavimas.

12.2 Faktorinės analizės matematinis modelis

Tarkime, kad stebime k kintamųjų $X_1, \ldots, X_k$. Modelis grindžiama prielaida, kad kiekvieno kintamojo X_i elgesį sąlygoja m bendrųjų latentinių faktorių $F_1, \ldots, F_m$ ir specifinis latentinis faktorius ε_i. Bendrųjų faktorių turi būti mažiau, nei kintamųjų, t.y. $m < k$. Matematinis faktorinės analizės modelis yra tiesinis:

$$\begin{aligned} X_1 &= \lambda_{11}F_1 + \lambda_{12}F_2 + \ldots + \lambda_{1m}F_m + \varepsilon_1, \\ X_2 &= \lambda_{21}F_1 + \lambda_{22}F_2 + \ldots + \lambda_{2m}F_m + \varepsilon_2, \\ &\vdots \\ X_k &= \lambda_{k1}F_1 + \lambda_{k2}F_2 + \ldots + \lambda_{km}F_m + \varepsilon_k. \end{aligned}$$

Daugikliai λ_{ij} vadinami *faktorių svoriais*. Faktorinės analizės uždavinys yra atvirkštinis tiesinės regresijos uždaviniui, t.y., žinome X_i reikšmes, ir norime išsiaiškinti, ką galima pasakyti apie bendruosius faktorius F_j.

Faktorinės analizės prielaidos:

- stebimi kintamieji turi normalųjį skirstinį $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$;
- bendrieji faktoriai F_j nekoreliuoti ir $\mathbf{D}F_j = 1$;
- charakteringieji faktoriai ε_i nekoreliuoti ir $\mathbf{D}\varepsilon_i = \tau_i$;
- faktoriai F_j ir ε_i nekoreliuoti, $i = 1, 2, \ldots, k, j = 1, 2, \ldots, m$.

Pasinaudoję šiomis prielaidomis galime apskaičiuoti stebimų kintamųjų X_i kovariacijas ir dispersijas:

$$
\begin{aligned}
&cov(X_i, X_j) = \lambda_{i1}\lambda_{j1} + \ldots + \lambda_{im}\lambda_{jm}, i \neq j,\\
&\mathbf{D}X_i = \sigma_i^2 = \lambda_{i1}^2 + \ldots + \lambda_{im}^2 + \tau_i = h_i^2 + \tau_i, i = 1, 2, \ldots, k,\\
&cov(X_i, F_j) = \lambda_{ij}, i = 1, 2, \ldots, k, j = 1, 2, \ldots, m.
\end{aligned}
$$

Dydis $h_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2$ vadinamas kintamojo X_i *bendrumu*, o dydis τ_i – *specifiškumu*. Matome, kad kiekvieno kintamojo X_i dispersija skaidoma į dvi dalis – dispersiją, kurią sąlygoja bendrieji faktoriai $F_1, \ldots, F_m$ (bendrumas h_i^2) ir dispersiją, kurią lemia bendraisiais faktoriais nepaaiškinama dalis (specifiškumas τ_i). Kuo didesnis h_i^2, palyginti su σ_i^2, tuo daugiau informacijos apie kintamąjį X_i išsaugoma pereinant nuo pradinių kintamųjų prie bendrųjų faktorių.

Matrica, kurios elementai yra $cov(X_i, X_j), i \neq j$, o pagrindinėje įstrižainėje yra bendrumai h_i^2 vadinama *redukuotąją* kovariacijų matrica. Redukuotosios ir pradinės kovariacijų matricų panašumas yra modelio adekvatumo indikatorius. Jei visi $\varepsilon_i = 0$, tai redukuotoji ir pradinė kovariacijų matricos sutampa ir bendrieji faktoriai F_j išsaugo visą informaciją apie kintamuosius X_i.

Faktorinė analizė turi išspręsti šiuos uždavinius:

- įvertinti faktorių svorius λ_{ij} ir specifines dispersijas τ_i;
- įvertinti bendrųjų faktorių $F_1, \ldots, F_m$ reikšmes kiekvienam kintamųjų X_i stebėjimų rinkiniui.

Iš pradžių randama pradinių duomenų matricos $\mathcal{X}$ kovariacijų matrica $\mathcal{S}$ arba koreliacijų matrica $\mathcal{R}$. Dėl paprastumo pradinius duomenis rekomenduojama standartizuoti (imti z-reikšmes vietoje x-reikšmių). Tada kovariacijų matrica $\mathcal{S}$ sutampa su koreliacijų matrica $\mathcal{R}$, visi $\sigma_i^2 = 0$, todėl palengvėja h_i^2 įverčio interpretacija. Po to skaičiuojama redukuotoji koreliacijų matrica, randami faktorių svorių įverčiai. Kitas etapas – faktorių svorių matricos sukimas ir paskutinis etapas – faktorių reikšmių įverčių skaičiavimas.

12.3 Duomenų tikimas faktorinė analizei. Modelio adekvatumo tyrimas

Faktorinė analizė neturi prasmės nekoreliuotiems kintamiesiems. Todėl visų pirma reikia įsitikinti, ar stebimi kintamieji tarpusavyje koreliuoja. Tai padeda nustatyti Bartlett'o sferiškumo kriterijus, pagal kurį yra tikrinama hipotezė, kad kintamųjų koreliacijų matrica yra vienetinė, t. y. visi stebimi kintamieji yra

nekoreliuoti. Jeigu taikant Bartlett'o sferiškumo kriterijų p-reikšmė yra didesnė už pasirinktąjį reikšmingumo lygmenį α, t. y. minėta hipotezė neatmetama, tai turimiems duomenims faktorinė analizė yra netaikytina. 1 pav. Bartleto

```
r <- cor(data)
cortest.bartlett(r, n = nrow(data))
$chisq
[1] 247.7095

$p.value
[1] 1.930071e-16

$df
[1] 91
```

1 pav.: Bartleto kriterijaus taikymas.

kriterijaus p-reikšmė labai maža. Modelis yra adekvatus. Be Bartleto testo yra dar Tucker Lewis faktorinės analizės patikimumo indeksas, kurio reikšmė turi būti kuo artimesnė vienetui esant FA tinkamiems duomenims.

Ar kintamieji tinka faktorinei analizei, įvertina Kaiserio-Meyerio-Olkino (KMO) matas. Tai - empirinių koreliacijos koeficientų didumų ir dalinių koreliacijos koeficientų didumų palyginamasis inteksas. Kuo arčiau vieneto, tuo geriau. Jei $KMO < 0,5$, faktorinė analizė nepriimtina. Kiekvieno kintamojo stebėjimų tinkamumo matas MSA_i (*angl.* Measure of Sampling Adequacy) apskaičiuojamas pagal formulę:

$$MSA_i = \frac{\sum_{j \neq i} r_{ij}}{\sum_{j \neq i} r_{ij} + \sum_{j \neq i} \tilde{r}_{ij}},$$

čia r_{ij} yra kintamųjų X_i ir X_j empirinis koreliacijos koeficientas, $\tilde{r}_{ij}$ – dalinės koreliacijos koeficientas. Kintamuosius, kurių $MSA_i < 0,5$, reikia šalinti, jie netinka faktorinei analizei.

```
KMO(r)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = r)
Overall MSA =  0.61
MSA for each item =
               Price               Safety      Exterior_Looks
                0.72                 0.47                 0.55
      Space_comfort           Technology After_Sales_Service
                0.61                 0.65                 0.62
       Resale_Value           Fuel_Type     Fuel_Efficiency
                0.63                 0.68                 0.62
              Color          Maintenance          Test_drive
                0.56                 0.61                 0.64
   Product_reviews         Testimonials
                0.69                 0.50
```

2 pav.: Kaiserio-Meyerio-Olkino kriterijaus taikymas.

12.4 Faktorių išskyrimas

Panagrinėsime vieną iš dažniausiai nadojamų faktorių išskyrimo metodų, grindžiamą *pagrindinių komponenčių analize*. Tarkime, turime k kintamųjų $X_1, \ldots, X_k$.

Daugelio kintamųjų tarpusavio priklausomybė gali būti įvertinta jų koreliacijomis arba kovariacijomis, iš koreliacijos (kovariacijos) koeficientų suformuojant koreliacinę (kovariacinę) matricą. Taikant pagrindinių komponenčių analizę, randamos tarpusavyje nekoreliuojančios kintamųjų $X_1, \ldots, X_k$ tiesinės kombinacijos $Y_1, \ldots, Y_k$

$$Y_1 = \sum_{j=1}^{k} \alpha_{1j} X_j, \ldots, Y_k = \sum_{j=1}^{k} \alpha_{kj} X_j,$$

tenkinančios šias sąlygas:

1. $cov(Y_i, Y_j) = 0, i, j = 1, \ldots, k, i \neq j$;
2. $DY_1 \geq DY_2 \geq \ldots \geq DY_k$;
3. $\sum_{i=1}^{k} DY_i = \sum_{i=1}^{k} DX_i$.

Kintamieji $Y_1, \ldots, Y_k$ nekoreliuoti ir išdėstyti dispersijų mažėjimo tvarka. Be to, jų dispersijų suma lygi pradinių kintamųjų dispersijų sumai (pažymėkime šią sumą D). Pagrindinių komponenčių paieška susiveda į koeficientų α_{ij} radimą. Įrodyta, kad šie koeficientai yra pradinių kintamųjų kovariacijų matricos tikriniai vektoriai, o dispersijos DY_i lygios atitinkamoms tikrinėms reikšmėms. Y_1 vadinama *pirmąja pagrindine komponente*, ji atitinka didžiausią tikrinę reikšmę ir paaiškina $100 \cdot DY_1/D$ procentų bendrosios kintamųjų $X_1, \ldots, X_k$ dispersijos. Antroji pagrindinė komponentė Y_2 atitinka antrąją pagal didumą tikrinę reikšmę, ji paaiškina $100 \cdot DY_2/D$ procentų bendrosios dispersijos ir t.t.

Apibrėžimas 1. *Sakome, kad kvadratinė matrica C turi tikrinę reikšmę (angl. eigenvalue) λ_n, atitinkančią tikrinį vektorių (angl. eigenvector) $\vec{\alpha}_n$, jei*

$$C\vec{\alpha}_n = \lambda_n \vec{\alpha}_n.$$

λ_n reikšmė randama iš charakteringosios lygties $|C - \lambda_n I| = 0$, čia I yra vienetinė matrica, kurios matmenys tokie pat, kaip ir matricos C.

Kuo daugiau bendrosios kintamųjų dispersijos paaiškina pagrindinė komponentė, tuo ji svarbesnė kaip akumuliuojanti informaciją apie kintamuosius. Pavyzdžiui, jei pagrindinė komponentė paaiškina 70% bendrosios dispersijos, galime teigti, kad palikdami vietoje pradinių kintamųjų $X_1, \ldots, X_k$ tik tą vieną komponentę, išlaikysime 70% informacijos apie pradinių kintamųjų reikšmių sklaidą. Visos pagrindinės komponentės (jų yra tiek, kiek ir pradinių kintamųjų) paaiškina visą bendrąją kintamųjų dispersiją, tačiau tik m pirmųjų komponenčių $Y_1, \ldots, Y_m$, paaiškinančių didžiąją dalį bendrosios dispersijos, panaudojamos faktoriams nustatyti. Paprastai, m yra parenkamas lygiu ne mažesnių už vienetą koreliacijos matricos tikrinių reikšmių skaičiui.

Taigi, turint k kintamųjų stebėjimus $(x_{1j}, x_{2j}, \ldots, x_{kj}, j = 1, \ldots, m$ iš pradžių apskaičiuojami k pagrindinių komponenčių įverčiai

$$\hat{Y}_i = \sum_{j=1}^{k} \hat{\alpha}_{ij} X_j, i = 1, \ldots, k.$$

čia $\hat{\alpha}_{ij}$ yra koeficientų α_{ij} empiriniai įverčiai. Latentiniais bendraisiais faktoriais laikomos m pirmųjų pagrindinių komponenčių, normuotų standartiniais nuokrypiais, t. y.

$$\hat{F}_j = \frac{\hat{Y}_j}{\sqrt{s^2(\hat{Y}_j)}}, j = 1, \ldots, m,$$

čia $s^2(\hat{Y}_i)$ yra i-osios pagrindinės komponentės dispersijos įvertis lygus i-ajai pagal dydį koreliacijų matricos tikrinei reikšmei. Faktorių svorių įverčiai išreiškiami lygybe

$$\hat{\lambda}_{ij} = \hat{\alpha}_{ji}\sqrt{s^2(\hat{Y}_j)}, i = 1, \ldots, k, j = 1, \ldots, m.$$

Specifinių faktorių įverčiai išreiškiami lygybe

$$\hat{\varepsilon}_i = \sum_{j=m+1}^{k} \hat{\alpha}_{ji}\hat{Y}_i, i = 1, \ldots, k.$$

Tuomet

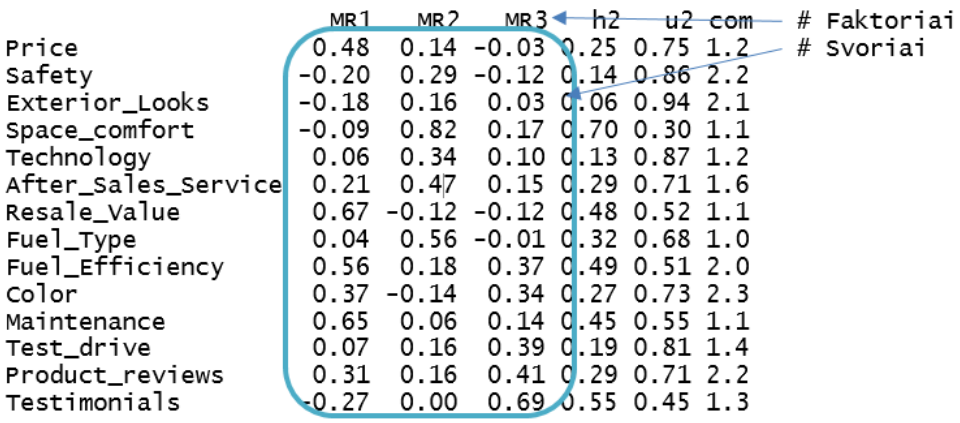
$$\hat{X}_i = \sum_{j=1}^{m} \hat{\lambda}_{ij}\hat{F}_j + \hat{\varepsilon}_i, i = 1, \ldots, k.$$

Faktorių matrica aprašo faktorių ir atskirų kintamųjų priklausomybę. Kaip nustatyti, kokie kinamieji nusako faktorių F_j? Galioja paprasta taisyklė (Čekanavičius, Murauskas, 2002), – faktorius F_j laikomas susijusiu su tais kintamaisiais $X_1, \ldots, X_k$, kurių svorių įverčiai $\hat{\lambda}_{1j}, \ldots, \hat{\lambda}_{kj}$ absoliučiu didumu ne mažesni kaip 0,4. Teigiamas svoris rodo, kad kintamasis su faktoriumi koreliuoja teigiamai, o neigiamas svoris – neigiamai. Kintamieji yra vienodai svarbūs nepriklausomai nuo svorio ženklo.

12.5 Faktorių sukimas ir interpretavimas

Naudojantis pradine faktorių svorių matrica būna sunku interpretuoti faktorius. Taip atsitinka dėl to, kad dažniausiai vyrauja pirmasis faktorius, be to net kelių to paties kintamojo faktorių svoriai gali būti didesni už 0,4. Ką tokiu atveju daryti? Norėdami palengvinti faktorių diferencijaciją bei suteikti jiems lengviau interpretuojamą pavidalą, sudaromos tiesinės gautų faktorių kombinacijos, kurios tarpusavyje nekoreliuoja (yra ortogonalios) ir turi vienetines dispersijas. Naujųjų faktorių rinkinių nustatymo procedūra vadinama *ortogonaliąja pradinių faktorių transformacija arba ortogonaliuoju sukimu.* Populiariausias tarp ortogonalių sukimų yra VARIMAX. Sukimo tikslas – supaprastinti svorių matricos struktūrą, kad kiekvienas kintamasis turėtų tik kelis nenulinius faktorių svorius (idealiu atveju tik vieną). Sukimas nekeičia sprendinio bendrumų ir bendrosios dispersijos paaiškinimo procento. Tačiau keičiasi kiekvieno faktoriaus indelis į bendrosios dispersijos paaiškinamąją dalį.

Interpretuojant faktorius negalima išvengti subjektyvumo, nes faktorių įvardijimas priklauso nuo tyrėjo kompetencijos, jo išsilavinimo.



3 pav.: Trijų faktorių atvejis.

Loadings:

	MR1	MR2	MR3	MR4
Price	0.535			
Safety		0.356		
Exterior_Looks				-0.544
Space_comfort		0.753		
Technology		0.349		
After_Sales_Service		0.528		
Resale_Value	0.724			
Fuel_Type		0.557		
Fuel_Efficiency	0.492			
Color				0.706
Maintenance	0.603			
Test_drive			0.407	
Product_reviews	0.336		0.429	
Testimonials			0.677	

4 pav.: Keturių faktorių atvejis.