Faktorinė analizė

Doc. Dr. Rūta Simanavičienė

Įvadas

- Pasitaiko tokių užduočių, kada tiriamąjį objektą geriausiai apibūdina ne stebimi kintamieji, bet jų kombinacijos.
- Pvz., vietoj stebimų kintamųjų „ūgis“ ir „svoris“ įvedamas abstraktesnis kintamasis — „dydis“.
- Kartais objektui apibūdinti iš viso nėra tokių kintamųjų, kuriuos būtų galima tiesiogiai išmatuoti, todėl iš duomenų tenka išskirti taip vadinamus faktorius.
- Patys ***faktoriai*** dažnai neturi kiekybinio mato, pvz. *kūrybiškumas*, *agresija*, *altruizmas* negali būti išmatuoti betarpiškai, bet šias sąvokas galime įsivaizduoti kaip atitinkamas ***požymių grupes*** vienijančias kategorijas.

Faktorinės analizės uždavinio pavyzdys (idėja)

- Tarkime, tiriama, kodėl dalis pirmakursių neigiamai žiūri į dalyką „Matematinė analizė“, t.y. kokie faktoriai sąlygoja neigiamą požiūrį.
- Respondentams pateikiama apie 30 klausimų apimančių įvairius neigiamo požiūrio aspektus.
- Atsakymai vertinami penkių balų sistema nuo „griežtai nesutinku“ iki „pilnai sutinku“.
- Faktorinėje analizėje pagal ***respondentų vertinimų koreliacijas*** studentai yra suskirstomi į kelias grupes. Tada sprendžiama, koks **faktorius** galėtų **vienyti** konkrečios grupės studentus.
- Pavadinimą **faktoriui** suteikia pats tyrėjas, išanalizavęs grupės sudėtį. Šiuo atveju, tai gali būti ***silpnos mokyklinės matematikos žinios, lėtas naujos medžiagos įsisavinimo greitis*** ir t.t.

Faktorinės analizės tikslas ir etapai

- **Faktorinės analizes užduotis** – atsižvelgiant į kintamųjų tarpusavio koreliacijas, suskirstyti stebimus kintamuosius į grupes, kurias vienija koks nors tiesiogiai nestebimas (*latentinis*) faktorius. Koks tas faktorius, sprendžiame patys nagrinėdami grupes sudarančius kintamuosius.
- **Faktorines analizes tikslas** – minimaliai prarandant informacijos pakeisti stebimą reiškinį charakterizuojančių požymių aibę kelių faktorių rinkiniu.
- **Faktorinės analizes 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 įverčių skaičiavimas.

Latentinis - (lot. *latens*, kilm. *latentis* - paslėptas, nematomas): nematomas, išoriškai nepastebimas.

Latentiniai faktoriai

- ***Pavyzdžiui***, psichologas prieš tyrimą sudaro klausimyną žmogaus lyderio savybėms matuoti. Klausimyne yra dvi klausimų grupės: viena - dalykiniams gebėjimams nustatyti, kita - bendravimo gebėjimams įvertinti.
- Tiek dalykiniai gebėjimai, tiek bendravimo gebėjimai betarpiškai neišmatuojami (*latentiniai*) faktoriai.
- **Ko siekiame taikydami faktorinę analizę?** Sociologinių apklausų, medicininių tyrimų, psichologinių testų ir pan. rezultatai – dešimčių ir šimtų požymių matavimų aibės.
- Faktorinė analizė padeda **didelio** skaičiaus kintamųjų tarpusavio koreliacijas paaiškinti **tam tikrų** bendrųjų **faktorių** įtaka.
- **Nuo kintamųjų** pereidami **prie faktorių** kondensuojame informaciją, padarome ją labiau aprėpiamą.
- Būtent dėl to faktorinė analizė dažnai taikoma kartu su kitais daugiamatės statistikos metodais (pavyzdžiui, latentinių faktorių reikšmių įverčiai gali būti naudojami kaip pradinių duomenų pakaitalas ***klasterinėje*** ar ***regresinėje analizėje***).

Faktorinė analizė savotiškai prieštaringa

- **Faktorinė analizė – gana sudėtinga ir prieštaringa daugiamatės statistinės analizės dalis,** nes:
  a) ne visada latentiniai faktoriai realiai egzistuoja ir ne visada patikimai pagal turimus duomenis juos galima išskirti;
  b) tiems patiems duomenims taikydami skirtingus faktorinės analizės metodus, gauname keletą galimų faktorių rinkinių;
  c) išskirtieji faktoriai ne visada lengvai interpretuojami.
- Tipinė faktorinė analizė „pasufleruoja" atsakymus į tokius klausimus:
  a) kiek latentinių faktorių paaiškina tiriamų kintamųjų priklausomybės struktūrą;
  b) kokie tie faktoriai;
  c) kaip gerai faktoriai paaiškina duomenis.

Faktorinės analizės (*FA*) matematinis modelis

- Tarkime stebime ***k*** kintamųjų $X_1, X_2, \dots, X_k$.
- Modelis grindžiamas prielaida, kad kiekvieno kintamojo X_i elgesį sąlygoja ***m*** bendrųjų latentinių faktorių $F_1, F_2, \dots, F_m$ ir specifinis (charakteringasis) latentinis faktorius $\varepsilon_i, i = \overline{1, k}$.
- Bendrųjų faktorių yra mažiau nei kintamųjų $(m < k)$. Tarkim $X_i, i = \overline{1, k}$ nuo faktorių priklauso tiesiškai. Tada ***FA* matematinis modelis** užrašomas:

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

- Daugikliai $\lambda_{ij}, i = \overline{1, k}, j = \overline{1, m}$ – vadinami ***faktorių svoriais***.
- Nors ***FA*** modelis primena regresinės analizės modelį, tačiau ***FA*** uždavinys – žinant X_i išsiaiškinti ką galima pasakyti apie bendruosius faktorius F_j.

FA modelio prielaidos

- **Faktorinės analizės modelio prielaidos:**
  a) stebimi kintamieji pasiskirstę pagal normalųjį dėsnį, t. y. $X_i \sim N(\mu_i, \sigma_i^2)$;
  b) bendrieji faktoriai F_j **nekoreliuoti** ir $DF_j = 1$;
  c) charakteringieji faktoriai ε_i, **nekoreliuoti** ir $D\varepsilon_i = \tau_i$;
  d) faktoriai F_j ir ε_i **nekoreliuoti**, čia $i = \overline{1,k}$, $j = \overline{1,m}$.

- Kintamųjų pasiskirstymo pagal **normalųjį** dėsnį sąlyga nėra kritinė faktorinei analizei.

  Tikrinant modelio prielaidas turime stebėti tam tikras ***statistines charakteristikas***:

  Stebimų kintamųjų dispersijas, Stebimų kintamųjų kovariacijas, Stebimų kintamųjų ir latentinių faktorių kovariacijas.

FA modelio savybės

Atsižvelgus į prielaidas, stebimų kintamųjų $\boldsymbol{X_i}$ ir $\boldsymbol{X_j}$, $\left(i, j = \overline{1, k}\right)$ dispersijas ir kovariacijas galima užrašyti taip:

- Kai $i \neq j$, tai $cov\left(\boldsymbol{X_i}, \boldsymbol{X_j}\right) = \lambda_{i1}\lambda_{j1} + \cdots + \lambda_{im}\lambda_{jm}$. (*Naudodami $cov(X_i, X_j)$ ieškosime svorių λ_{ij}*).
- Kai $i = j$, tai $cov(\boldsymbol{X_i}, \boldsymbol{X_i}) = \lambda_{i1}^2 + \cdots + \lambda_{im}^2$ (*empirinė dispersija*), nes $cov(\boldsymbol{X_i}, \boldsymbol{X_i}) = \frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{x})^2$.

Vadinasi kovariacijų matricos diagonalės elementai yra kintamųjų $\boldsymbol{X_i}$, $i = \overline{1, k}$ dispersijos:

$$DX_i = \sigma_i^2 = \lambda_{i1}^2 + \cdots + \lambda_{im}^2 + \tau_i = h_i^2 + \tau_i, \qquad i = \overline{1, k}.$$

Atsitiktinių dydžių koreliacija ir kovariacija

- ***Kovariacija* ir koreliacijos koeficientas** – tai skaitinės charakteristikos, įvertinančios dviejų atsitiktinių dydžių ***tiesinę priklausomybę***.
- Atsitiktinių dydžių *X* ir *Y* ***kovariacija*** skaičiuojama pagal formulę:

$$cov(X,Y) = EXY - EXEY$$

- Kovariacija yra skaičius, kuris gali būti ir teigiamas ir neigiamas. Kovariacijos savybės:

1) Jeigu *X* ir *Y* yra nepriklausomi, tai $cov(X,Y) = 0$. Vadinasi *X* ir *Y* yra ***nekoreliuoti***;
2) $|cov(X,Y)| \leq \sqrt{DXDY}$. Vadinasi $|cov(X,X)| \leq DX$.

- Atsitiktinių dydžių *X* ir *Y* ***koreliacijos koeficientas*** skaičiuojamas pagal formulę:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{DXDY}} = \frac{EXY - EXEY}{\sqrt{DXDY}}.$$

Iš kovariacijos apibrėžimo išplaukia:

- jeigu dydžiai koreliuoja, tai jie yra priklausomi;
- jeigu dydžiai nekoreliuoja, jie gali būti ir priklausomi, ir nepriklausomi.

Empiriniai: kovariacija ir koreliacijos koeficientas

- Kintamųjų $X = (x_1, x_2, \dots, x_n)$ ir $Y = (y_1, y_2, \dots, y_n)$ imčių empirine kovariacija vadinamas skaičius:

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Kintamųjų $X = (x_1, x_2, \dots, x_n)$ ir $Y = (y_1, y_2, \dots, y_n)$ imčių empiriniu koreliacijos koeficientu vadinamas skaičius:

$$cor(X,Y) = \frac{cov(X,Y)}{s_x s_y}.$$

Čia $\bar{x}$, $\bar{y}$ – atitinkamai kintamųjų X ir Y empiriniai vidurkiai; s_x, s_y – atitinkamai kintamųjų X ir Y empiriniai standartiniai nuokrypiai.

Redukuotoji kovariacijų matrica

- Matrica, kurios elementai a_{ij} yra $cov\left(X_i, X_j\right), i \neq j$, o pagrindinėje įstrižainėje yra bendrumai h_i^2, vadinama ***redukuotąja* kovariacijų matrica**.

Čia dydis $h_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2$ - vadinami kintamojo X_i ***bendrumu***, o dydis τ_i - vadinamas kintamojo X_i ***specifiškumu***.

$$cov\left(X_i, X_j\right) = \lambda_{i1}\lambda_{j1} + \cdots + \lambda_{im}\lambda_{jm}, \qquad i \neq j.$$

$$DX_i = \sigma_i^2 = \lambda_{i1}^2 + \cdots + \lambda_{im}^2 + \tau_i = h_i^2 + \tau_i, \quad |cov(X_i, X_i)| \leq DX_i.$$

$$cov\left(X_i, F_j\right) = \lambda_{ij}, i = \overline{1, k}; j = \overline{1, m}.$$

- 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.
- Jeigu visi $\varepsilon_i = 0$, tai ***redukuotoji kovariacijų matrica*** sutampa su pradine kovariacijų matrica ir bendrieji faktoriai F_j išsaugo visą informaciją apie kintamuosius X_i.

FA turi išspręsti šiuos uždavinius

- Matematiniai faktorinės analizės uždaviniai:

1) Rasti faktorių svorių $\boldsymbol{\lambda_{ij}}$ ir specifinių dispersijų $\boldsymbol{\tau_i}$ įverčius;
2) Rasti kiekvieno kintamojo $\boldsymbol{X_i}$ stebėjimų rinkinio latentinių faktorių $\boldsymbol{F_1}, \boldsymbol{F_2}, \dots, \boldsymbol{F_m}$ įverčius.

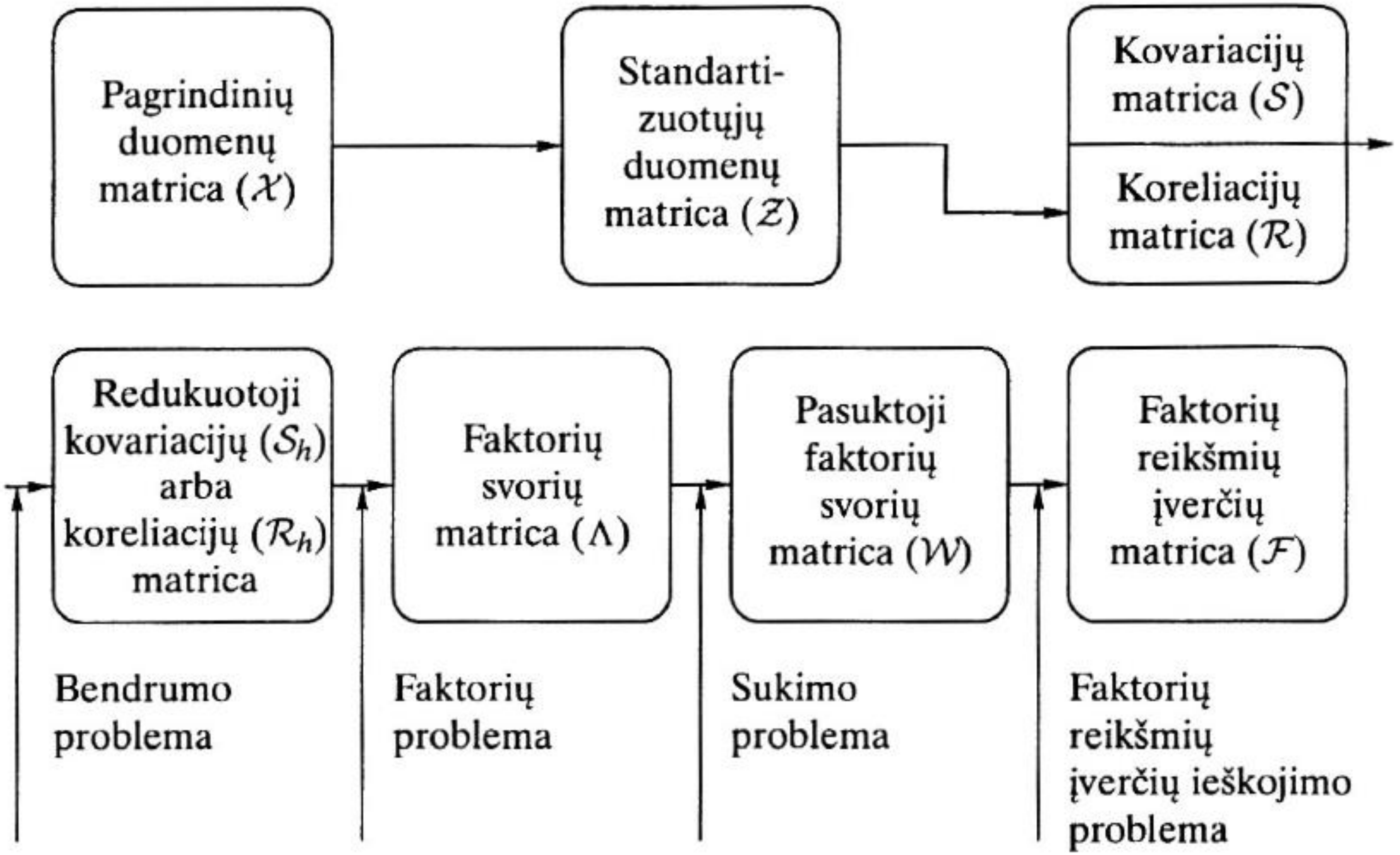
Šių uždavinių sprendimui atliekami tokie veiksmai:

1. Imame pradinių duomenų matricą $X = \left(X_{ij}\right), i = \overline{1, n},\ j = \overline{1, k}$.
2. Pradinius duomenis *standartizuojame* (*rekomenduojama*, nes tada bus $\sigma_i^2 = 1$);
3. Skaičiuojame kovariacijų matricą ***S*** ir koreliacinę (koreliacijų) matricą ***R***.

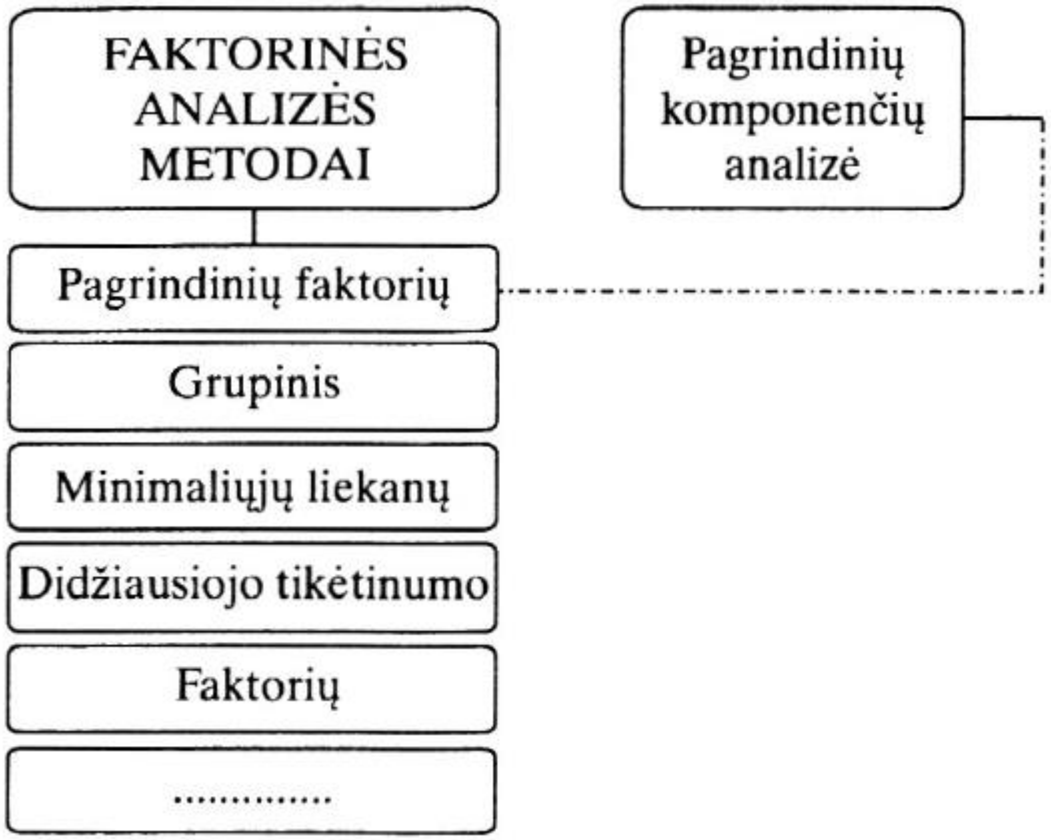
   (stand. duomenų kovariacijų matrica = stand. Koreliacijų matricai = nestand. koreliacijų matricai).

3. Skaičiuojama **redukuotoji kovariacijų** matrica ir randami **faktorių svorių** įverčiai.
4. Faktorių svorių matricos sukimas ir faktorių reikšmių įverčių skaičiavimas.

Bendroji faktorinės analizės algoritmo schema



FA metodų klasifikacija



Duomenų tikimas faktorinei analizei

Ar duomenys tinka faktorinei analizei tikrinama:

1) Sudarant koreliacijų matrica;

2) Skaičiuojant *Kaiserio-Meyerio-Olkino (KMO) matą;*

3) Skaičiuojant *i-ojo* ***kintamojo*** stebėjimų ***tinkamumo*** matą MSA_i.

Duomenų tikimas faktorinei analizei (1)

1) Faktorinė analizė *neturi prasmės nekoreliuotiems* kintamiesiems.

- Todėl reikia isitikinti, ***ar stebimi kintamieji tarpusavyje koreliuoja***. Tai padeda nustatyti ***Bartlett'o sferiškumo kriterijus***, pagal kuri yra tikrinama hipotezė, kad kintamųjų koreliacijų matrica yra vienetine, t. y. visi stebimi kintamieji yra nekoreliuoti.
- Jeigu taikant Bartlett'o sferiškumo kriterijų *p*-reikšmė $p \geq \alpha$, tai turimiems duomenims faktorinė analizė yra netaikytina (α – pasirinktas reikšmingumo lygmuo).

Duomenų tikimas faktorinei analizei (2)

2) Ar kintamieji tinka faktorinei analizei, *įvertina Kaiserio-Meyerio-Olkino (KMO) matas*. Tai – empirinių koreliacijos koeficientų didumų ir dalinių koreliacijos koeficientų didumų palyginamasis indeksas. Kuo arčiau vieneto, tuo kintamieji labiau tinka faktorinei analizei.

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

Čia r_{ij} – kintamųjų X_i ir X_j koreliacijos koeficientas; $\tilde{r}_{ij}$ yra X_i ir X_j dalinės koreliacijos koeficientas. Dalinės koreliacijos koeficientai leidžia įvertinti dviejų tiriamų kintamųjų tarpusavio ryšį, kai kitų kintamųjų įtaka yra ***eliminuojama***.

- Jeigu $KMO > 0{,}9$ – ***FA*** tinka puikiai;
- Jeigu $0{,}8 < KMO \leq 0{,}9$ – ***FA*** tinka gerai;
- Jeigu $0{,}7 < KMO \leq 0{,}8$ – ***FA*** tinka patenkinamai;
- Jeigu $0{,}6 < KMO \leq 0{,}7$ – ***FA*** tinka pakenčiamai;
- Jeigu $0{,}5 < KMO \leq 0{,}6$ – ***FA*** tinka blogai;
- Jeigu $KMO < 0{,}5$ – ***FA*** nepriimtina.

Duomenų tikimas faktorinei analizei (3)

3) Kiekvieno *i-ojo* kintamojo stebėjimų tinkamumo matą MSA_i galima apskaičiuoti pagal formulę:

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

- Kintamuosius, kurių MSA_i reikšmės mažos, t.y. $MSA_i < 0{,}5$, reikia iš faktorinės analizės pašalinti, nes jie netinka ***FA***.

Faktorių išskyrimas (1)

- **Pagrindinių komponenčių metodas** – vienas iš dažniausiai naudojamų faktorių išskyrimo metodų, grindžiamų pagrindinių komponenčių analize (angl. *Principal components analysis – PCA*).
- Tarkim turim ***k*** kintamųjų $X_1, X_2, \dots, X_k$. Daugelio kintamųjų tarpusavio priklausomybė gali būti įvertinta jų ***koreliacijomis*** arba ***kovariacijomis***, bei ***dispersijomis***.
- Taikant pagrindinių komponenčių analizę randamos tarpusavyje nekoreliuojančių kintamųjų $X_1, X_2, \dots, X_k$ ***tiesinės daugdaros*** (kombinacijos) $Y_1, Y_2, \dots, Y_k$, t.y.

$$\boldsymbol{Y_1} = \sum_{j=1}^{k} \alpha_{1j} X_j,\ \ \boldsymbol{Y_2} = \sum_{j=1}^{k} \alpha_{2j} X_j, \dots, \boldsymbol{Y_k} = \sum_{j=1}^{k} \alpha_{kj} X_j\,,$$

tenkinančios sąlygas:

1) $\boldsymbol{cov(Y_i, Y_j) = 0, i, j = 1, 2, \dots, k, i \neq j;}$

2) $DY_1 \geq DY_2 \geq \cdots \geq DY_k;$

3) $\sum_{i=1}^{k} DY_i = \sum_{i=1}^{k} DX_i = D.$

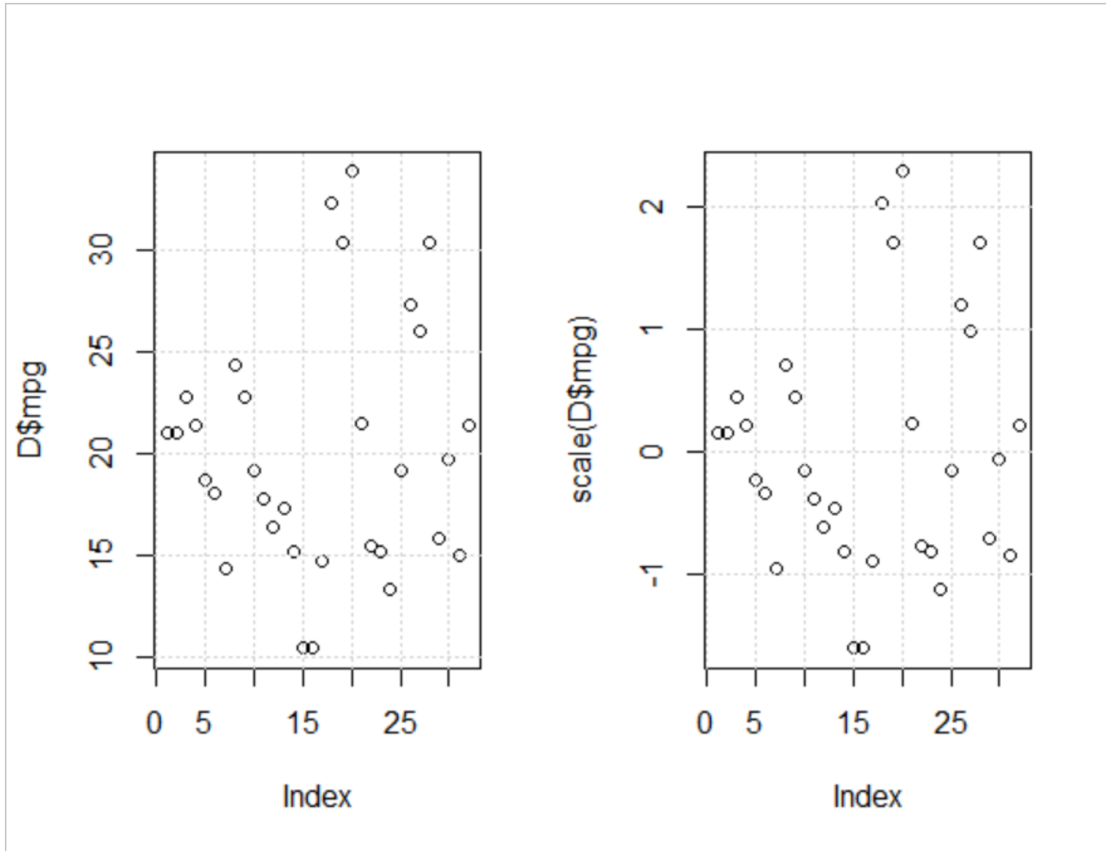
Šios tiesinės daugdaros vadinamos ***pagrindinėmis komponentėmis***.

Faktorių išskyrimas (2)

- Jeigu PCA naudotume standartizuotąsias kintamųjų $X_1, X_2, \dots, X_k$ reikšmes (z-reikšmes), tai visos dispersijos būtų $DX_i = 1$, o bendroji dispersijų suma:

$$\sum_{i=1}^{k} DY_i = \sum_{i=1}^{k} DX_i = D = k.$$

- Dažniausiai PCA rekomenduojama naudoti ne pradines kintamųjų reikšmes, o kintamųjų standartizuotąsias reikšmes.



Faktorių išskyrimas (3)

- Matome, kad pagrindinių komponenčių paieška susiveda į jų koeficientų $\alpha_{ij}, i, j = 1, 2, \dots, k$ paiešką.
- Tarkim, $Y_1 = \alpha_{11}X_1 + \cdots + \alpha_{1k}X_k$.
- Pradžioje ieškome $\alpha_{11}, \alpha_{12}, \dots, \alpha_{1k}$, su kuriais būtų DY_1 maksimizuojama (*nes Y_j išdėstyti dispersijų mažėjimo tvarka*), t.y.

$$DY_1 = \sum_{i=1}^{k}\sum_{j=1}^{k} \boldsymbol{\alpha_{1i}\alpha_{1j}}\sigma_{ij}^2,$$

su sąlyga, kad $\sum_{j=1}^{k} \boldsymbol{\alpha_{1j}^2} = 1$ (sąlyga reikalinga, norint gauti vienintelį sprendinį), čia $\sigma_{ij}^2 = cov(X_i, X_j)$.

- Taikant matricų algebros operacijas, galima įrodyti, kad šio **uždavinio sprendinys $\boldsymbol{\alpha_1} = (\boldsymbol{\alpha_{11}}, \boldsymbol{\alpha_{12}}, \dots, \boldsymbol{\alpha_{1k}})$ yra pradinių kintamųjų kovariacijų matricos *S tikrinis vektorius*,** kuris atitinka maksimalią matricos S *tikrinę reikšmę*. Ši tikrinė reikšmė lygi DY_1.
- Taip gauta tiesinė daugdara $Y_1 = \alpha_{11}X_1 + \cdots + \alpha_{1k}X_k$ vadinama kintamųjų $X_1, X_2, \dots, X_k$ pirmąja ***pagrindine komponente***. Ji paaiškina $100 \cdot DY_1/D$ procentų bendrosios dispersijos.
- Analogiškai apskaičiuojama $Y_2 = \alpha_{21}X_1 + \cdots + \alpha_{2k}X_k$ – antroji pagrindinė komponentė, kuris paaiškina $100 \cdot DY_2/D$ procentų bendrosios dispersijos ir t.t.

Tikrinis vektorius ir tikrinė reikšmė

- Sakome, kad kvadratinė matrica C turi **tikrinę** reikšmę (angl. *eigenvalue*) $\boldsymbol{\lambda}$, atitinkančią tikrinį vektorių (angl. *eigenvector*) $\vec{\boldsymbol{\alpha}}$**,** jei

$$C\vec{\alpha} = \boldsymbol{\lambda}\vec{\alpha} \qquad (1)$$

Reikšmė λ randama iš charakteringosios lygties $|C - \lambda I| = 0$, čia I yra vienetinė matrica, kurios matmenys sutampa su matricos *C* matmenimis.

Paprastai reikalaujama, kad tikrinio vektoriaus koordinačių kvadratų suma būtų lygi 1, t.y.

$$\sum_{i=1}^{k} \alpha_i^2 = 1. \qquad (2)$$

- Kvadratinių simetrinių matricų (koreliacijų ir kovariacijų matricos yra simetrinės) **tikrinių** reikšmių skaičius yra lygus matricos eilučių **skaičiui**.

Pavyzdys. Matricos tikrinių reikšmių radimas

Rasti matricos R tikrines reikšmes λ_n, kai $R = \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix}$.

Sprendimas.

Remiantis apibrėžimu:

$|R - \lambda I_n| = 0$, ***n*** yra kvadratinės matricos ***R*** eilučių skaičius.

$$\left|\begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right| = 0$$

$$\left|\begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\right| = 0$$

$$\left|\begin{pmatrix} 1-\lambda & 0.75 \\ 0.75 & 1-\lambda \end{pmatrix}\right| = 0$$

Gauname lygtį: $\lambda^{\mathbf{2}} - 2\lambda + 0{,}4375 = 0$. Išsprendę gauname: $\lambda_1 = 1{,}75 \text{ ir } \lambda_2 = 0{,}25$.

Pavyzdys. Matricos tikrinių vektorių radimas

Rasti matricos R tikrinius vektorius $\vec{\alpha}$, kai $R = \begin{pmatrix} 1 & 0{,}75 \\ 0{,}75 & 1 \end{pmatrix}$.

Sprendimas. Iš ankstesnio etapo turime matricos R tikrines reikšmes. Su kiekviena reikšme randame ją atitinkantį tikrinį vektorių

1) Kai $\lambda_1 = 1{,}75$. Remiantis apibrėžimu: $R\vec{\alpha} = \lambda_n\vec{\alpha}$ sudarysime matricinę lygtį:

$R\vec{\alpha} - \lambda_n\vec{\alpha} = \vec{0},\ \boldsymbol{\vec{0} = \lambda_n\vec{\alpha} - R\vec{\alpha}}$

$$(R - \lambda_n \cdot I_n)\vec{\alpha} = \vec{0}$$

$$\left(\begin{pmatrix} 1 & 0{,}75 \\ 0{,}75 & 1 \end{pmatrix} - \begin{pmatrix} 1{,}75 & 0 \\ 0 & 1{,}75 \end{pmatrix}\right)\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} -0{,}75 & 0{,}75 \\ 0{,}75 & -0{,}75 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Gauname vieną lygtį: $-0{,}75\alpha_1 + 0{,}75\alpha_2 = 0$ iš kurios $\alpha_1 = \alpha_2$. Remiantis sąlyga (2) sudarome lygtį:

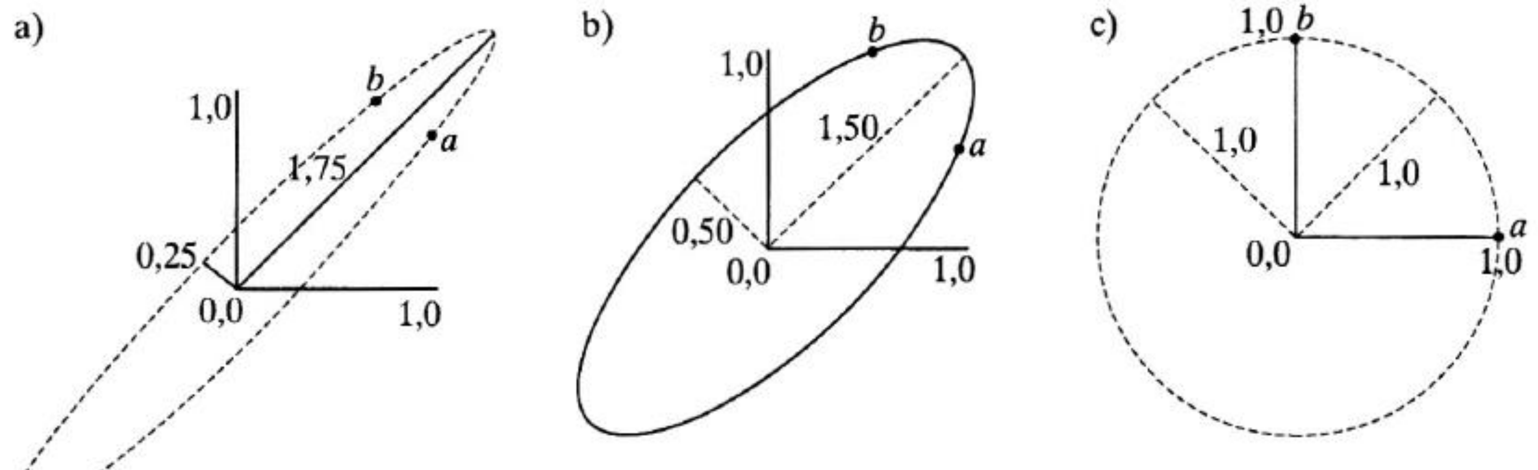
$$\alpha_1^2 + \alpha_2^2 = 1, \quad \rightarrow \quad 2\alpha_1^2 = 1 \quad \rightarrow \quad \alpha_1 = \pm\sqrt{0{,}5} = \pm 0{,}707$$

Vadinasi, kai $\lambda_1 = 1{,}75$, tai $\vec{\alpha} = (0{,}707; 0{,}707)$ ir $\vec{\alpha} = (-0{,}707; -0{,}707)$ $\rightarrow$ $\boldsymbol{\vec{\alpha}_1 = (0{,}707; 0{,}707)}$

Analogiškai, kai $\lambda_2 = 0{,}25$, tai $\vec{\alpha} = (-0{,}707; 0{,}707)$ ir $\vec{\alpha} = (0{,}707; -0{,}707)$ $\rightarrow$ $\boldsymbol{\vec{\alpha}_2 = (-0{,}707; 0{,}707)}$

Tikrinių reikšmių ir tikrinių vektorių prasmė PCA (1)

- Tikrines reikšmes ir tikrinius vektorių galime stebėti grafike. Kai ***n* = 2**, tai koreliacijos matricos R eilutės reiškia du taškus, kurių koordinatės a=(1; 0,75) ir b=(0,75; 1).
- Tikriniai vektoriai nusako elipsės, einančios per taškus ***a*** ir ***b***, ašių kryptis, tikrinės reikšmės – nusako ašių ilgius.



$a = (1{,}00,\ 0{,}75)$
$b = (0{,}75,\ 1{,}00)$

$a = (1{,}00,\ 0{,}50)$
$b = (0{,}50,\ 1{,}00)$

$a = (1{,}00,\ 0{,}00)$
$b = (0{,}00,\ 1{,}00)$

- Kaip žinome, **koreliacijų matricos tikrinių vektorių koordinatės yra pagrindinių komponenčių koeficientai**, **o tikrinės reikšmės – komponenčių dispersijos** DY_i.
- **Iš grafiko a)** galima pasakyti, kad $\frac{\lambda_1}{\lambda_1+\lambda_2} \cdot 100\% = \frac{1{,}75}{1{,}75+0{,}25} \cdot 100\% = 87{,}5\%$ pirmoji komponentė paaiškina bendrosios kintamųjų dispersijos. Atsisakydami antrosios komponentės, prarastume 12,5% informacijos apie kintamųjų įgyjamų reikšmių sklaidą.

Tikrinių reikšmių ir tikrinių vektorių prasmė PCA (2)

- Kuo kintamieji stipriau koreliuoja, tuo didžioji elipsės ašis ilgesnė, o mažoji trumpesnė. Ilgesnė ašis atitinka svarbesnę komponentę.
- Jeigu kintamųjų koreliacija būtų lygi 1, tuomet didžiosios ašies ilgis = 2, mažosios ilgis = 0.
- Vienetinė koreliacija reikštų, kad visą informaciją apie pradinius kintamuosius suteikia pirmoji pagrindinė komponentė. Vadinasi jeigu koreliacija nėra lygi 1, vienos pagrindinės komponentės nepakanka norint apimti kuo daugiau pradinių kintamųjų informacijos.
- Kuo daugiau bendrosios kintamųjų dispersijos paaiškina pagrindine komponente, tuo ji svarbesne kaip akumuliuojanti informacija apie kintamuosius.
- Visos pagrindines 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 $(m < k)$.

Faktorių išskyrimas (4)

Turint ***k*** kintamųjų stebėjimus $\left(x_{1j}, x_{2j}, \dots, x_{kj}\right), j = \overline{1, m}$ apskaičiuojami ***k*** pagrindinių komponenčių įverčiai:

$$\hat{Y}_i = \sum_{j=1}^{k} \hat{a}_{ij} X_j, i = \overline{1, k}$$

Čia $\hat{a}_{ij}$ yra koeficientų α_{ij} empiriniai įverčiai.

Latentiniais bendraisiais faktoriais laikomos ***m*** pirmųjų pagrindinių komponenčių, normuotų standartiniais nuokrypiais, t.y.

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

Čia $\boldsymbol{s^2\left(\hat{Y}_j\right)}$ yra *i-osios* pagrindinės komponentės dispersijos įvertis **lygus** *i-ajai* pagal dydį koreliacijos matricos **tikrinei reikšmei**.

Faktorių išskyrimas (5)

- 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

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

- Paskutinėje lygtyje pateikta faktorių matrica, kuri aprašo faktorių ir atskirų kintamųjų priklausomybę.
- Faktorius F_j laikomas susijęs su tais kintamaisiais, kurių svorių įverčiai $|\hat{\lambda}_{ij}| \geq 0{,}4$. Teigiamas svoris rodo, jog kintamasis su faktoriumi koreliuoja teigiamai, neigiamas – neigiamai.