Idősorok és többdimenziós statisztika

11. előadás

Zempléni András

2019.11.26

Két várható érték vektor összehasonlítása

- Legyen
  $\underline{X}_1, \underline{X}_2, \ldots \underline{X}_{n_1} \sim N(\mu_1, \Sigma)$, és
  $\underline{Y}_1, \underline{Y}_2, \ldots \underline{Y}_{n_2} \sim N(\mu_2, \Sigma)$
  két független d-dimenziós normális eloszlású minta, megegyező, de ismeretlen kovariancia mátrixszal.
- H_0: $\{\mu_1 = \mu_2\}$ tesztje teljes alternatíva mellett.
- Mindkét mintából megbecsüljük a kovariancia mátrixot,

$$\widehat{\Sigma}_1 = \sum_{i=1}^{n_1} (X_i - \bar{\underline{X}}) (X_i - \bar{\underline{X}})^T .$$

$$\widehat{\Sigma}_2 = \sum_{i=1}^{n_2} (Y_i - \bar{\underline{Y}}) (Y_i - \bar{\underline{Y}})^T .$$

- A kovariancia mátrix összevont (pooled) becslése:

$$\widehat{\Sigma}_{pl} = \frac{1}{n_1 + n_2 - 2} \left(\widehat{\Sigma}_1 + \widehat{\Sigma}_2 \right)$$

Két várható érték vektor összehasonlítása/2

- A próbastatisztika

$$T^2 = \frac{n_1 \cdot n_2}{n_1 + n_2} \cdot (\bar{\underline{X}} - \bar{\underline{Y}})^T \widehat{\Sigma}_{pl}^{-1} (\bar{\underline{X}} - \bar{\underline{Y}})$$

- Ebben az esetben is Hotelling féle T^2 eloszlású lesz a próbastatisztika, mégpedig

$$T^2_{d, n_1+n_2-2}$$

.

- Irodalom (a lentebbiekhez is): T.W.Anderson: *Introduction to Multivariate Statistical Analysis* Wiley & Sons, New York, 2003. (pp.77-82,170-181,251-257,459-466, 569-575.)

Próbák kovariancia mátrixra: ismert várható érték vektor

- Legyen megint $\underline{X}_1, \underline{X}_2, \ldots \underline{X}_n \sim N(\mu, \Sigma)$ eloszlású független d-dimenziós minta, ismert μ várható érték vektorral .
- A H_0: $\{\Sigma = \Sigma_0\}$ nullhipotézist teszteljük teljes alternatíva mellett.
- Felhasználva az ismert várható érték vektort, megbecsüljük a kovariancia mátrix n-1-szeresét,
  $C = \sum_{i=1}^{n} (X_i - \mu) (X_i - \mu)^T$.
- Legyen

$$\lambda = \left(\frac{e}{n}\right)^{\frac{dn}{2}} |C\Sigma_0^{-1}|^{\frac{n}{2}} exp\{ - tr(C\Sigma_0^{-1})/2\}$$

  ezzel a próbastatisztika:

$$T = -2 \log(\lambda)$$

- amelynek 0-hipotézis melletti eloszlása

$$T \sim \chi^2_{d(d+1)/2}$$

Próbák kovariancia mátrixra: ismeretlen várható érték vektor

- Változatlanul független d-dimenziós $N(\mu, \Sigma)$ eloszlású a minta, de most μ-t nem ismerjük.
- Újfent a H_0: $\{\Sigma = \Sigma_0\}$ nullhipotézist teszteljük teljes alternatíva mellett.
- Mivel a várható érték vektor nem ismert, csak a kovariancia mátrix becslésére $\widehat{\Sigma}$-ra támaszkodunk.
- Legyen most

$$\lambda = (n-1)^{(1-d)\cdot\frac{n-1}{2}} |\widehat{\Sigma}\Sigma_0^{-1}|^{\frac{n-1}{2}} exp\left\{\frac{n-1}{2}\left(d - tr(\widehat{\Sigma}\Sigma_0^{-1})\right)\right\}$$

amivel a próbastatisztikát változatlan alakban kapjuk:

$$T = -2\log(\lambda)$$

- és amelynek 0-hipotézis melletti eloszlása az ismert várható értékkel megegyezően

$$T \sim \chi^2_{d(d+1)/2}$$

Főkomponens- és faktoranalízis

- A főkomponens- és faktoranalízis olyan statisztikai technika, amelyet változók halmazára alkalmazunk, hogy feltárjuk, közülük melyek tartalmaznak közös fluktuációs mintákat - akár csak részben, más fluktuációkkal kombináltan is -, és meghatározzuk ezeket a közös mintákat.
- A közös változékonyságminta általában a háttérben meghúzódó (*látens*) változó/folyamat hatásának eredményeként áll elő. E hatást a **faktor**változó reprezentálja.
- Mivel a faktor az egyes megfigyelt változók közös additiv komponense (bár súlya az egyes változóban általában eltérő), így ez a megfigyelt változók korrelációjának forrása.
- A faktorok segítségével az összes megfigyelt változó változékonysága leírható, így ezek teljesen jellemzik megfigyeléseinket ezért pusztán ezeket megtartva csökkenthetjük (sokszor jelentősen) a változóink számát

Példák

- Hallgatók adatai: motiváció, intellektuális képességek, iskolatörténet, családtörténet, egészség, fizikai jellemzők, személyiségjegyek. Mindegyiket több változóval is mérik. Néhány személyiségjegy, motivációs és iskolatörténeti változó mutathatja, hogy mennyire szeret önállóan dolgozni a hallgató, kombinálódhat egy önállósági faktorban. Mások egy intelligencia fatort adhatnak ki, stb.
- Talajvízszint mérő kutak adatainak fluktuációja főként a csapadékból történő utánpótlás, esetlegesen fólyóvízből oldalirányú betáplálás és a kommunális vízkivétel eredőjeként alakul, e három hatás kutak százainak adatait jellemezheti globálisan (és e hatások eltávolítása után határozhatók meg a lokális befolyásoló tényezők).

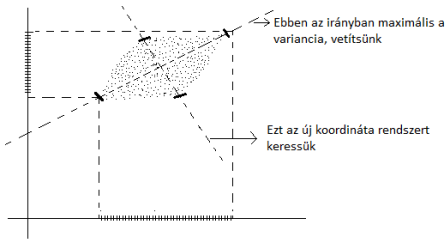
Főkomponens analízis

- A főkomponens analízis (Principal Component Analysis, PCA) a változók kombinációinak varianciáját vizsgálja.
- A PCA jelentősen csökkenti a változók számát. Bizonyos változók a kísérletek, megfigyelések során alig változnak ingadozásuk (szórásuk) kicsi, ezeket elhagyhatjuk. Ám gyakran nem ez vagy az a változó kis szórású, hanem pl. a kettő összege, vagy valamely más lineáris kombinációja. Ezeket keressük. Illetve inkább azokat, amelyeknek nagy a szórása, és ezért nem hagyhatók el.
- Matematikailag: az $X_1, ..., X_n$ minta egy p dimenziós teret feszít ki, ám még véletlenül sem ortogonális bázisként. Mi tehát adatainkat egy $F_1, ..., F_p$ **új, ortogonális** bázisban szeretnénk felírni, melynek össz-hossznégyzete, azaz szórásnégyzet-összege az eredetivel egyező.
- Ha megvannak a megfigyeléseink transzformációi, közülük a legkisebb szórásúakat elhagyhatjuk.

A főkomponens analízis célja, "elméleti" megoldás

- Az első főkomponens megtalálásához maximalizálni akarjuk a változók egy lineáris kombinációjának szórását. Lényegileg egy olyan irányt keresünk, amely mentén a változók maximálisan "szétterülnek", szétszóródnak.
- Néha a PCA a végcél, de inputot generál további elemzéshez.
- Ez az irány a X p-dimenziós val. változónál éppen a Σ kovariancia-mátrix legnagyobb sajátértékhez tartozó β sajátvektora.
- Az erre vonatkozó $\beta' X$ vetület adja a maximális szórást.
- A további főkomponensek irányait rendre a nagyság szerint következő sajátértékekhez tartozó sajátvektorok adják meg
- Ez ortogoális bázis, azaz X megfelelő lineáris kombinációi korrelálatlanok

2 dimenziós szemléltetés



- Eltoljuk a középpontot az új középpontba, majd beforgatjuk a tengelyeket.

Forgatás

- Tegyük fel, hogy a centrálás már megtörtént.
- A forgatás egy $\mathbf{A}$ ortogonális mátrixszal: $\mathbf{A}^T\mathbf{A} = I$ való szorzás. X az adataink mátrixa, Z a főkomponenseké
$$Z = \mathbf{A}X$$
- A ellipszoid tengelyeit megtalálni pont az $\mathbf{A}$ mátrix megtalálásával ekvivalens, amely úgy forgatja el a változókat, hogy azok korrelálatlanok legyenek, vagyis a kovariancia mátrix **diagonális**:
$$S_Z = diag(\sigma^2_{Z_1}, ..., \sigma^2_{Z_p})$$
- Másfelől:
$$S_Z = EZZ^T = E(\mathbf{A}X)(\mathbf{A}X)^T = \mathbf{A}S_x\mathbf{A}^T$$
- Szimmetrikus mátrixok spektrálfelbontásának S_X-re alkalmazásához vegyük az S_X p db normált sajátvektorából $(v_1, ..., v_p)$-ből mint oszlopokból álló $\mathbf{V}$ mátrixot.

Spektrálfelbontás

Ekkor $I = \mathbf{V}\mathbf{V}^T \Rightarrow$:

$$S_X = S_X\mathbf{V}\mathbf{V}^T = S_X(v_1, ..., v_p)\mathbf{V}^T =$$

$$= (S_X v_1, ..., S_X v_p)\mathbf{V}^T = (\lambda_1 v_1, ..., \lambda_p v_p)\mathbf{V}^T = \mathbf{V}\Lambda\mathbf{V}^T$$

Ez a spektrálfelbontás, ahol Λ a sajátértékek diagonális mátrixa:

$$\Lambda = diag(\lambda_1, ..., \lambda_p).$$

Innen $S_X = \mathbf{V}\Lambda\mathbf{V}^T$ miatt

$$\Rightarrow \mathbf{V}^T S_X \mathbf{V} = \mathbf{V}^T\mathbf{V}\Lambda\mathbf{V}^T\mathbf{V} = \Lambda$$

- Tehát az $\mathbf{A} = \mathbf{V}^T$ választással kapott $Z = \mathbf{A}X$ bázisváltozók S_Z kovariancia mátrixa diagonális lesz, ahogy a főkomponensekétől megkívántuk. A keresett forgatás tehát az $\mathbf{A}$ mátrixszal adható meg, az $\mathbf{A}$ meghatározásához pedig az S_X sajátvektorainak és sajátértékeinek számítása szükséges.

A megmagyarázott variancia

- Egyszersmind az S_X mátrix sajátértékei a főkomponensek szórásnégyzetei is lesznek. Nagyságrend szerint rendezzük őket.
- S_X és S_Z nyoma ($\Rightarrow$ az összes szórásnégyzet összege) megegyezik, ezért van értelme az első k főkomponens által "megmagyarázott" varianciáról beszélni, ami

$$\text{Proportion of variance} = \frac{\sigma^2_{Z_1}+\ldots+\sigma^2_{Z_k}}{\lambda_1+\ldots+\lambda_p} = \frac{\sigma^2_{Z_1}+\ldots+\sigma^2_{Z_k}}{\sigma^2_{X_1}+\ldots+\sigma^2_{X_p}} = \frac{\sigma^2_{Z_1}+\ldots+\sigma^2_{Z_k}}{\sigma^2_{Z_1}+\ldots+\sigma^2_{Z_p}}$$

- Ha az eredeti változóink korreláltak (erősen), akkor az első néhány főkomponens "sok" varianciát magyaráz, míg az utolsó (jó)néhány keveset, így ez utóbbiak akár el is dobhatóak. Tehát az első néhányat megtartva redukálhatjuk a dimenziót, miközben megőrizzük a változékonyságot.

Megjegyzések

- Ha függetlenek (vagy inkább korrelálatlanok) a változóink, akkor ők maguk főkomponensek is $\Rightarrow$ nincs mit keresni.
- Vigyázni kell a skálával. A főkomponensek nem skálainvariánsok. Ha g/l helyett mg/l-ben mérünk egy változót $\Rightarrow$ jóval nagyobb lesz a súlya a főkomponensek előállításában.
- A megoldás, hogy a kovariancia mátrix helyett a korrelációkkal dolgozunk, azaz pl. minden változónk szórását 1-re normáljuk.
- Eredetileg Z_1 szórásnégyzetét akartuk maximalizálni, aztán a rá ortogonális altérben Z_2-t, és így tovább. De Z_i szórásnégyzete: $\underline{a}^T S_x \underline{a}$, és tetszőleges $\underline{a}$-ra nincs maximum, ezért $\lambda = \frac{\underline{a}^T S_X \underline{a}}{\underline{a}^T \underline{a}}$-t maximalizáljuk.
- λ_1 a legnagyobb sajátérték az $(S_X - \lambda I)a = 0$ egyenletben
- Itt nem kell invertálni $\Rightarrow$ szinguláris S_X mátrix is megengedhető.

Elnevezések

- faktor/főkomponens mátrix: F *vagy* $Z = AX$, j-ik faktor: F_j *vagy* $Z_j = \sum_{i=1}^{p} a_{i,j} X_i$
- $a_{i,j}$ factor score coefficient
- A factor score coefficient mátrix
- $X_i' = \begin{Bmatrix} X_i(\omega_1) \\ \vdots \\ X_i(\omega_n) \end{Bmatrix}$, $Z_j(\omega_k) = \sum_{i=1}^{p} a_{i,j} X_i(\omega_k)$
- $F_j = Z_j = \begin{Bmatrix} Z_j(\omega_1) \rightarrow \text{(Factor score)} \\ \vdots \\ Z_j(\omega_n) \rightarrow \text{(Factor score)} \end{Bmatrix}$
- (Alternatív elnevezés: Factor score coefficient matrix = loadings, Factor scores = scores)

Tulajdonság, ábrázolás

Factor loadings mátrix: A^T

- $Z = AX \rightarrow A^T Z = A^T AX = X$
- Tehát a faktorokból a megfigyeléseket visszaállíthatjuk. Ez nem érdekes addig, míg pontos az előállítás, nincs zaj.

Főkomponensek ábrázolása

- Az első két vagy néhány főkomponens score-jaira scatterplot párosával. Ezek mutathatnak normalitást, esetleg nemlinearitást. Ez már összefüggés, ami nem jó, mert a PC-k korrelálatlanok és igazából normális eloszlás alapfeltevés mellett $\Rightarrow$ függetlenek is. Outlier is detektálható ezekből a plotokból, illeve csoportok is megfigyelhetőek az "eset"-ekben.
- Itt is igaz, hogy kovariancia mátrix helyett korrelációs mátrixból is lehet dolgozni. Ez ugyanaz, mintha normálnánk a változókat, megszabadulunk a skálázási problémától. Ez azonban nem mindig jogos!

Egy példa

$S = \begin{Bmatrix} 1 & 4 \\ 4 & 25 \end{Bmatrix}$, míg a neki megfelelő korrelációs mátrix:

$R = \begin{Bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{Bmatrix}$

- S-ből $\lambda_1 = 25.65, \lambda_2 = 0.35 \Rightarrow$ Az F_1 98.6% szórást magyaráz
- $F_1 = 0.16X_1 + 0.987X_2$, vagyis F_1 lényegileg X_2
- Ugyanez R-rel:
  - $\lambda_1 = 1.8$
  - $\lambda_2 = 0.2$
  - Az F_1 90% szórást magyaráz.
- $F_1 = 0.707 \cdot X_1 + 0.707 \cdot X_2$ tehát F_1 ugyanannyira X_1, mint X_2.

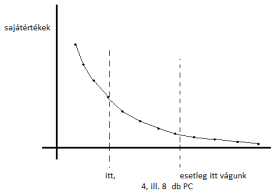
Hány főkomponenst tartsunk meg?

Lehetőségek a döntésre:

- Magyarázzák a szórás rögzített (pl 80) %-át
- Dobjuk ki azokat, melyek az átlagnál kisebb sajátértékhez tartoznak. Korrelációs mátrixra ez az átlag 1, tehát az 1-nél kisebb sajátértékhez tartozó főkomponenseket elhagyjuk.
- Scree plot - kőomlás diagram. (nagyság szerint plottoljuk a sajátértékeket, és ahol az első (vagy második) törést látjuk a közel lineáris csökkenésben, onnantól dobjuk a főkomponenseket.)
- A nagyobb főkomponens szignifikanciáját formálisan teszteljük.
- Értelmezés alapján, a társtudománnyal együttműködve, ez nem statisztikai módszer, de hasznos lehet.

Hány főkomponenst tartsunk meg/2?

- $H_{0,k} : \lambda_{p-k+1} = \cdots = \lambda_p = 0$
  $\bar{\lambda} = \frac{1}{k} \sum_{i=p-k+1}^{p} \log \lambda_i$
  Teszt statisztika:
  $$h = (p - \tfrac{2n+11}{6})(k \log(\bar{\lambda}) - \textstyle\sum_{i=n-k+1}^{n} \log \lambda_i)$$
  Ez közelítőleg χ^2_d, $d = \frac{(k-1)(k+2)}{2}$
  Ez általában kissé túlbecsüli a megtartandó komponensek számát.
- Scree-plot



Faktoranalízis (FA)

- A FA-ban a változókat reprezentálni akarjuk, mint néhány (jóval kevesebb) másik változó (a faktorok) lineáris kombinációja. A faktort általában nem lehet mérni, vagy megfigyelni.
- a FA különbözik a PCA-tól, mert
  - A PC-k az eredeti változók lineáris kombinációi, míg a FA-ban az eredeti változókat fejezzük ki a faktorok lineáris kombinációival.
  - PCA-ban az összes variancia nagy részét magyarázzuk, míg FA-ban a változók közötti kovarianciákat szeretnénk a legjobban reprodukálni.
- Több statisztikus nem szereti - a régebbi számítási módszerek gyakran adtak ellentmondó eredményeket, ezeket ma nem használják. A számítógépes módszerek ma már konsztensebbek. Azonban így is meglehetősen szubjektív az elfogadott modell

A faktormodell egyenlet

$$Y = DF + \epsilon$$

- Most Y a megfigyelés. Y helyett $Y - \mu$ áll(hat), ezért tegyük fel, hogy $\mu = 0$. F a faktorok, ϵ a zaj, D a factor loadings mátrix. ϵ és DF korrelálatlan, a faktorok maguk (F oszlopai) ugyanacsak korrelálatlanok - normálisra függetlenek, és az F_j-ket 1 szórásúnak feltételezzük.
- Ezért:

$$\sum_Y = cov(DF + \epsilon) = cov(DF) + cov\epsilon = E(DFF^T D^T) + \sum_\epsilon = DD^T + \sum_\epsilon$$

- Lényeges, hogy D nem négyzetes mátrix, több sora van, mint oszlopa, míg $\sum_\epsilon = diag(\sigma^2_{1,\epsilon}, \cdots, \sigma^2_{n,\epsilon})$. Így m db faktorunk van. $F = (F_1, \cdots, F_m)$

Tulajdonságok

- Ez a felbontás nem feltétlen létezik $n >> m$-re. De a lényeg, hogy FA-ban ezt keressük, ezt értjük azon, hogy szórásmátrixot szeretnénk minél jobban reprodukálni, kisebb dimenzióból.
- A faktormegoldás nem egyértelmű: ugyanis, ha van egy megoldás tetszőleges m x m-es forgatással:

$$TT^T = I$$
$$\sum_Y = DTT^T D^T + \sum_\epsilon = DD^T + \sum_\epsilon$$

tehát:

$$Y = DTF + \epsilon$$

is jól reprodukálja a szórásmátrixot, így $F^\star = TF$ -fel, mint új faktorokkal:

$$Y = DF^* + \epsilon$$

és mivel T ortogonális, így F^* is faktor tulajdonságú.

A kommunalitás

- A FA modell szerint minden változó varianciáját a faktorok varianciája magyarázza bizonyos mértékig, és van egy, a zajból származó saját, specifikus varianciája.
- A faktorok által magyarázott "arány" az úgynevezett kommunalitás, ez

$$h_i^2 = d_{i,1}^2 + ... + d_{i,m}^2$$

a D mátrix i-ik sorának négyzetösszege.

Tulajdonságok

- Mivel a faktorok korrelálatlanok és standardek, ezért

$$h_i^2 = \sum_{j=1}^m \mathrm{cov}(Y_i, F_j)^2 = \mathfrak{D}^2(\sum_{j=1}^m d_{ij} F_j)$$

- A kommunalitások nem változnak a megoldás forgatásával.
- Megjegyzés: h_i nem más, mint az i-ik sor faktorsúly vektorának hossza az $\Re^m$ -ben. Az a jó, ha közel van 1-hez.

A faktormegoldás előállítása

- Főkomponens módszer
- Principal Factor vagy Principal Axis módszer (főtengely)