Idősorok és többdimenziós statisztika

12. előadás

Zempléni András

2019.12.03

Főkomponens módszer

- Először is $\sum_Y$-t S-sel becsüljük. Keressük $\hat{D}$-ot, amelyre
$$S \cong \hat{D}\hat{D}^T + S_\epsilon$$
újfent spektrálfelbontjuk S-et:
$$S = CEC^T$$
ahol E: diagonális mátrix a sajátértékekből, C: sajátvektorok mátrixa
- Mivel E diagonális, $\Rightarrow$ négyzetgyököt vonhatunk, mert a főátlóban szórásnégyzetek állnak $\Rightarrow S = CE^{\frac{1}{2}}(E^{\frac{1}{2}})^T C^T$
- Most lehetne $\hat{D} = CE^{\frac{1}{2}}$, de ez még nem jó, mert p x p-es mátrix. De ne az összes sajátvektort vegyük, csak az első m-et: $\hat{D}_m = C_m E_m^{\frac{1}{2}}$

FA vs PCA

- Tulajdonképp: az utolsó néhány főkomponenst zajnak tekintjük, és a változó egyéni variációjával "azonosítjuk". A dimenziók nem pontosak így a zajra, az ugyanis n rangú, míg az utsó PC-k (n-m) rangúak. Tehát összefüggés marad a zajban.
- Úgy tűnhet, hogy az interpretáció ugyanaz, mint a PCA-nál, de most forgathatunk, míg a PCA-kat nincs értelme forgatni - elvesztik PC tulajdonságukat. (Más a cél!) (Tetszőleges pozitív definit mátrix diagonálisba forgatható (vissza is!), de I-be már nem $\Rightarrow$ a PC-kat forgatva kaphatok összefüggéseket, de a F-kat forgatva nem)
- Újfent használhatjuk a korreláció mátrixot kovariancia helyett. Most ez teljesen osszeegyeztethető az interpretációval.

Principal Factor vagy Principal Axis módszer (főtengely)

- Először becsüljük meg a zajt, azt vonjuk ki, aztán a maradékból határozzuk meg a faktort. Nem a zajt, hanem annak kovariancia mátrixát, tehát az egyes változók specifikus varianciáit kell becsülnünk.
$$S_Y - S_\epsilon = \begin{pmatrix} \hat{h}_1^2 & s_{1,2} & \cdots & s_{1,p} \\ & \ddots & & \\ s_{p,1} & \cdots & s_{p,p-1} & \hat{h}_p^2 \end{pmatrix}$$
ahol $\hat{h}_i^2$ a kommunalitások. Ezeket kell tehát becsülnünk.

A kommunalitás becslése

- s_{ii} az S^{-1} diagonálisának i-ik eleme $\hat{h}_i^2 = s_{ii} - \frac{1}{s_{ii}} = s_{ii} * R_i^2$
(az utolsó egyenlőség megmutatható)
ahol R_i^2 a squared multiple correlation (- a regresszióból) a maradék p-1 változóval.
- Hasonlóan korreláció mátrix esetén:

$$\hat{h}_i^2 = 1 - \frac{1}{r_{ii}} = R_i^2$$

az r_{ii} az R^{-1} diagonálisának i-ik eleme. Ez akkor jó, ha R nem szinguláris.
- Gyakran negatív sajátértékek is adódnak $S_Y - S_\epsilon$-ból. Ekkor a magyarázott variancia 1 fölé megy és aztán csökken vissza 1-re (normált esetben)

Maximum likelihood

- Tfh $Y_1, ..., Y_n$ $N_p(\eta, (\sum_Y))$
Ekkor D és $\sum_\epsilon$ ML becslése is lehetséges. Megmutatható, hogy ekkor $\hat{D}$ és S_ϵ a következőt elégíti ki:

$$S_Y S_\epsilon \hat{D} = \hat{D}(I + \hat{D}^T S_\epsilon^{-1} \hat{D})$$
$$S_\epsilon = diag(S_Y - \hat{D}\hat{D}^T)$$
$$\hat{D}^T S_\epsilon^{-1} \hat{D} \text{ diagonális mátrix}$$

Ezt kell iteratívan megoldani.
- Ez gyakran nem konvergál, vagy nem ad jó megoldást, a kommunalitások meghaladják 1-et.

A faktorszám vizsgálata

- Elég jó a scree plot is, (gyakran) felfedhet bizonytalanságot m megválasztásában.
- $H_0 : \sum_Y = DD^T + \sum_\epsilon$
$H_1 : \sum_Y \neq DD^T + \sum_\epsilon$
akarjuk tesztelni.
- A teszt stat. likelihood hányadosból:

$$h = (p - \frac{2n-2m+11}{6}) * log(\frac{|\hat{D}\hat{D}^T|}{|S_Y|})$$

$|\ |$ a determináns. Ez közelítőleg χ_d^2 ahol

$$d = \frac{1}{2}[(n-m)^2 - n - m]$$

Ha H_0-t elutasítjuk ($h > \chi_{d,1-p}$) $\Rightarrow$ több faktor kell.
Gyakorlatban gyakran túlbecsüli a faktorszámot.

Értelmezés

- A faktormegoldások elforgathatók - ettől megoldások maradnak. A forgatás PCA-ra nem javasolt, csak FA-ra, de Principal Factorból gyakran ugyanazt kapjuk, mintha PCA-t forgattunk volna.
- Az új, forgatott megoldás már korrelál és nem a maximális varianciát határozza meg.
- Úgy forgatjuk a megoldást, hogy minél több együttható a lineáris kombinációban 0 legyen. Így könnyebb értelmezni a megoldást, mert az eredeti változókból csak keveset használunk így fel egy-egy faktor meghatározásához $\Rightarrow$ a különböző faktorok más és más mért változót tartalmaznak (nagy súllyal).

Klasszifikáció

- A feladat: egy mintaelemről megállapítani, hogy az ismert osztályok közül melyikbe tartozik
- Az osztályok száma véges
- Két osztályra hipotézisvizsgálati problémaként is felfogható
- Matematikai eszköz: döntésfüggvény
- Példa: a kölcsönért folyamodó jó adós lesz-e vagy sem
- Az egyes osztályokban a sűrűségfüggvény legyen f_1 és f_2

Két osztály

- A döntés a p dimenziós teret (mintatér) osztja két részre
- Az optimális döntéshez kell a veszteség: $c1|2$ (ha 2 a valóság és mi az 1 mellett döntünk) és $c2|1$ (fordított esetben)
- Ha tudjuk az egyes csoportok apriori valószínűségét π_1 és π_2 és az R_i tartományon döntünk az i-edik csoport mellett ($i = 1, 2$), akkor a várható veszteség

$$\pi_1 c2|1 \int_{R_2} f_1(x)dx + \pi_2 c1|2 \int_{R_1} f_2(x)dx$$

- Az optimális megoldás:

$$R_1 := \{x : \frac{f_1(x)}{f_2(x)} \geq \frac{c1|2\pi_2}{c2|1\pi_1}\}$$

Két normális eloszlású minta

- Legyenek a kovariancia mátrixok azonosak: Σ, a két várható érték vektor pedig μ_1 és μ_2.
- Az előző tétel következménye:

$$R_1 := \{x : x^T\Sigma^{-1}(\mu_1-\mu_2)-\frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) \geq \log(k)\}$$

ahol

$$k = \frac{\pi_2 c1|2}{\pi_1 c2|1}.$$

- Spec. ha $\pi_1 = \pi_2$ és a költségek egyenlőek, akkor $k = 1$ és így $\log(k) = 0$.
- $$X^T\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2)$$

eloszlása normális, ebből a hibás döntések valószínűsége számolható

Ha nem ismertek a paraméterek

- Tegyük fel, hogy van n_1 és n_2 elemű tanuló-mintánk a két eloszlásból
- A mintaátlagokkal meg tudjuk becsülni a várható értéket
- A kovariancia mátrix becslése (S) is a szokásos módon kapható (pooled)
- Behelyettesítve az előző képletbe, a

$$W = x'S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}) - \frac{1}{2}(\bar{x}^{(1)} + \bar{x}^{(2)})'S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)})$$

függvényen alapulhat a döntésünk

- Ha vannak további teszt-adataink (nem tudjuk, hogy melyik eloszlásból), akkor ezek segítségével mindkét hipotézis esetére fel tudjuk írni az ML becsléseket és alkalmazhatjuk a valószínűséghányados próbát

Függetlenségvizsgálat normális eloszlás változó-csoportjaira

- Bonsuk fel a p dimenziós megfigyelés-vektorunkat q részre
- A nullhipotézis: a részek függetlenek, azaz a p dimenziós sűrűségfüggvény felbomlik a megfelelő q tényező szorzatára
- Ekvivalens: a kereszt-kovarianciákra $\Sigma_{ij} = 0$
- A likelihood-hányados módszer próbafüggvénye:

$$\lambda = \frac{\max_{\mu,\Sigma} L(\mu, \Sigma)}{\max_{\mu,\Sigma_0} L(\mu, \Sigma_0)} \sim \frac{|A|}{\prod_{i=1}^{q} |A_{ii}|}$$

  ahol Σ_0 a nullhipotézisnek megfelelő kovariancia mátrix, A a teljes kovarianciamátrix becslése, A_{ii} pedig az i-edik csoport kovarianciamátrixának becslése
- A kritikus tartomány $\lambda > c$ alakú

Nagy adathalmazok (big data)

- Több típus: "széles" (sok változó), "hosszú": sok megfigyelés
- Más-más módszerek kellenek
- Ha sok a megfigyelés: egyre jobban eltér a statisztikai szignifikanciától a gyakorlati szignifikancia (nem minden fontos, ami statisztikailag szignifikáns)
- Sok változó esetén értelmezhetetlen lehet az eredmény pl. egy regresszionál (rengeteg szignifikáns magyarázó változó). Célszerű csökkenteni a számukat
- Reális veszély a túlillesztés: az adatok egyedi véletlen tulajdonságait ragadja meg a modell
- Ellenőrzés: tanuló és tesztadatokra osztás
- Az a jó modell, ami a tesztadatokon pontos

Ridge és Lasso regresszió

- Ridge: L_2 simítás, a módosított minimalizálandó célfüggvény

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Tehát a kisebb β a preferált ("csökkentett hatás")
- Lasso: least absolute shrinkage and selection operator (L_1 simítás)
- $\lambda \sum_{j=1}^{p} |\beta_j|$ a "büntető" tag
- Erőssége: ki is választja a fontos magyarázó változókat (a nem fontosak együtthatói 0-vá válnak)