6. előadás, 2022. március 17.

Zempléni András

Valószínűségelméleti és Statisztika Tanszék
Természettudományi Kar
Eötvös Loránd Tudományegyetem

Áringadozások előadás

Bootstrap (Efron, 1979)

- Újramintavételezési eljárás, a becsléseink szórásának vizsgálatára, modell-illeszkedés ellenőrzésére
- Számtalan változatát dolgozták ki azóta, az egyik leggyorsabban fejlődő részterülete a statisztikának
- Előnye: rugalmas a minta (a statisztika) eloszlására vonatkozó feltételek változására

Bootstrap módszer - bevezetés

- $\mathbf{X}_i^* = \{X_1^*, \ldots, X_m^*\}$ visszatevéses mintavétellel az eredeti mintából
- általában $m = n$
- Nehézségek a gyakorlatban:
  1. $\underline{x} \Longrightarrow \hat{P}$ minden modellnél más és más
  2. $\hat{P} \longrightarrow \underline{x}^*$ a sok ismétlés megterheli a számítógépet

Az i.i.d. bootstrap

- Legyenek $X_1, X_2, \ldots$ i.i.d. valószínűségi változók, F (ismeretlen) közös eloszlással
- $T_n = t_n(\mathcal{X}_n; F)$ minket érdeklő val.változó, az eloszlása: G_n
- Cél: G_n eloszlásának becslése
- Bootstrap módszer:
  - Adott $\mathcal{X}$-re, visszatevéssel m elemű mintát veszünk: $\mathcal{X}_m^* = \{X_1^*, \ldots, X_m^*\}$
  - az X_i^*-ok közös eloszlása: $F_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$
  - $T_{m,n}^* = t_m(\mathcal{X}_m^*; F_n)$
  - Ismétlések $\Rightarrow \hat{G}_{m,n}$

Megjegyzések

- Az ötlet a módszer mögött nagyon egyszerű: jó lenne, ha sok mintánk lenne a populációból, de csak egy van. Ezért vegyünk mintát a becsléséből: ez a tapasztalati eloszlás.
- Az ismétlések száma legyen elég nagy ahhoz, hogy a mintavételi hiba elhanyagolható legyen (legalább 500, de 10000 is elképzelhető)
- A naív "középső 95%" konfidencia intervallum túl szűk kicsi minták esetén (például a várható érték becslésénél: a tapasztalati eloszlás szórásnégyzete $(n-1)/n$-szerese a ténylegesnek, ez öröklődik a bootstrap mintákra
- Nagyon könnyű a programozása (vannak R-es csomagok, de általában nincs is szükség a használatukra)

Alaptétel (Efron)

- A fenti esetben, ha $\sigma^2 = Var(X_i)$ véges és a statisztika a standardizált mintaátlag

$$T_n = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$$

akkor

$$\lim_{n\to\infty} \sup_x |P_*(T^*_{n,n} \le x) - \Phi(x)| = 0$$

m.m.

- A bizonyítás a Berry-Esséen tételen alapul (megadja a konvergencia sebességét a centrális határeloszlás tételnél), a konvergencia még gyorsabb is lehet, mint a klasszikus normális eloszlással történő approximációnál

Egy ellenpélda

- Bizonyos esetekben a becslés nem kozisztens (Singh, 1981):

Definíció

$\{X_n\} n\ge 1$ m-összefüggő egy $m \ge 0$-ra, ha $\{X_1, \dots, X_k\}$ és $\{X_{k+m+1}, \dots\}$ függetlenek minden $k \ge 0$.

- Jelölés: $\sigma_m^2 = Var(X_1) + 2\sum_{i=1}^{m-1} Cov(X_1, X_{1+i})$
- Legyen a becsülendő statisztika: $T_n = \sqrt{n}(\overline{X}_n - \mu)$
- A bootstrap megfelelője: $T^*_{n,n} = \sqrt{n}(\overline{X}^*_n - \overline{X}_n)$

Theorem

Legyen $\{X_n\} n\ge 1$ egy stacionárius m-összefüggő sorozat, $EX_1 = \mu$, $\sigma^2 = Var(X_1) \in (0,\infty)$, $\sum_{i=1}^{m-1} Cov(X_1, X_{1+i}) \ne 0$ és $\sigma_m^2 \ne 0$. Ekkor

$$\lim_{n\to\infty} \sup_x |P_*(T^*_{n,n} \le x) - P(T_n \le x)| \ne 0$$

m.m.

Korrekció a konfidencia intervallumoknál

- Az empirikus kvantiliseket finomítani kell. A BC (bias-correction–torzítás korrigáló) módszer a határok megállapítására:

$$\hat{F}^{-1}\left\{\Phi\left(z_0 + \frac{z^\alpha + z_0}{1 - a(z^\alpha + z_0)}\right)\right\}$$

- ahol $\hat{F}^{-1}$ az empirikus kvantilisfüggvénye a bootstrap statisztikának
- z^α a szokásos empirikus kvantilis
- z_0 is a torzítás korrekciós tag
- a a szórásnégyzet növekedésének gyorsulását korrigálja
- Ha $a = 0$ és $z_0 = 0$ és $\hat{F}$ a normális eloszlás, az érték éppen z^α

A BC-formula motivációja és alkalmazása

- Ha monoton transzformációt: $m(\vartheta)$ alkalmazunk a becslésünkre, az eredmény normális eloszlású:

$$m(\hat{\vartheta}) \sim N\left(m(\vartheta) - z_0(1 + am(\vartheta)), 1 + am(\vartheta)\right).$$

- Innen, a monotonitás miatt $P(\hat{\vartheta} < \vartheta) = \Phi(z_0)$, z_0 könnyen becsülhető
- Az a becslését a loglikelihood függvény deriváltjának ferdeségéből kaphatjuk

Példa: konfidencia intervallum a korrelációra

- A standard intervallum (az empirikus korrelációs együttható aszimptotikus normalitásán alapul) szimmetrikus –nem mindig reális kis minták esetén
- A boostrap lehet aszimmetrikus, a lefedési valószínűsége beállítható
- Kérdés: vajon a paraméteres vagy a nemparametéteres bootstrap a jobb (a paraméteres általában szélesebb – konzervatívabb – intervallumot ad)

Az (m, n) bootstrap

- Ha a "szokásos" bootstrap nem működik, általában segít, ha $m < n$ elemű mintákat veszünk
- ekkor a visszatevés nélküli mintavétel (részminta) is lehetséges, gyakran jobb tulajdonságú
- Bickel és Sakov (2008) cikke algoritmust ad az optimális m megválasztására - ez az "igazi" (visszatevéses) bootstrap-re vonatkozik, és az eredmény $m \sim n$, ha az n elemű minta is jó.

Példa

- Legyen X_i i.i.d.μ várható értékkel és σ szórással
- A $\mu = 0$ hipotézist teszteljük a $\sqrt{n}\overline{X}_n$ statisztikával
- Jó bootstrap algoritmus: mintavétel az $X_i - \overline{X}_n$ "reziduálisokból"
- Ha $\sqrt{n}\,\overline{X}_n^*$ bootstrap eloszlását nézzük, ennek kvantilisei nem konzisztensek
- rögzített m-re $n \to \infty$ esetén $\sqrt{m}\,\overline{X}_m^*$ határeloszlása m-től függ (csak a normális eloszlás esetén ugyanaz minden m-re)

Példa/2

- $\sqrt{m}(\overline{X}^*_m - \overline{X}_n) \longrightarrow N(0,\sigma)$ ha $n, m \to \infty$
- Tehát $\sqrt{m}\ \overline{X}^*_m \sim N(\sqrt{m}\ \overline{X}_n, \sigma)$ ha $m \to \infty$
- $\sqrt{m}\ \overline{X}_n = \sqrt{m/n}\sqrt{n}\ \overline{X}_n \longrightarrow N(0, \sqrt{\lambda}\sigma)$ ahol $\lambda = \lim m/n$
- A jó eredményt $m/n \to 0$ esetén kapjuk

Az m kiválasztása

- Az előzőek szerint a jó tartományban a bootstrap eloszlás nem változik lényegesen
- Ha m túl nagy, vagy túl kicsi, akkor a bootstrap eloszlások különbözőek
- Tehát az algoritmus:
  1. Legyen $m_j = \lfloor q^j n \rfloor$ $(0 < q < 1)$
  2. Minden m_j-re határozzuk meg a $T^*_{m_j,n}$ eloszlását (szimulációval)
  3. Válasszuk azt az m-et, amire $\hat{m} = \rho(T^*_{m_j,n}, T^*_{m_{j+1},n})$ (ahol ρ az eloszlásbeli konvergenciával konzisztens metrika - pl. Kolmogorov-Szmirnov távolság)

Alkalmazása az összefüggő esetre

Circular blokk bootstrap (CBB)

1. $Y_t = X_{t_{mod(N)}}$ azaz periodikusan kiterjesztjük a mintát
2. Legyen $i_1, i_2, \ldots i_m$ minta az $\{1, \ldots, N\}$ halmazon egyenletes eloszlásból
3. Adott b blokkméretre készítsünk N'=mb ($N' \approx N$) pszeudo-megfigyelést:
$$Y^*_{(k-1)b+j} = Y_{i_m+j-1} \quad \text{ahol } j = 1, \ldots, b; \quad k = 1, \ldots, m$$
4. A minket érdeklő statisztika kiszámítása a pszeudo-megfigyelésekből:
$$\overline{Y}^*_{N'} = (N')^{-1}(Y^*_1 + \ldots + Y^*_{N'})$$

Blokkméret kiválasztása (Politis & White)

Jel. $\mathcal{F}^0_{-\infty} = \sigma\{X_n : n \le 0\}$, $\mathcal{F}^\infty_k = \sigma\{X_n : n \ge k\}$

Def.: $\{X_t$: t$\in \mathbb{Z}\}$ erősen keverő, ha $\alpha_X(k) \longrightarrow 0$ $(k \to \infty)$, ahol
$\alpha_X(k) = sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}^0_{-\infty}, B \in \mathcal{F}^\infty_k\}$

Tétel : Tegyük fel, hogy $E|X_t|^{6+\delta} < \infty$, $\sum_{k=1}^{\infty} k^2(\alpha_X(k))^{\frac{\delta}{6+\delta}} < \infty$ valamely δ>0-ra. Legyen $b = o(N^{1/2})$, N $\to \infty$ esetén b $\to \infty$.
Ekkor $MSE(\sigma^2_{b,\overline{X}}) = \frac{G^2}{b^2} + D\frac{b}{n} + o(b^{-2}) + o(\frac{b}{n})$

ahol D=$\frac{4}{3}g^2(0)$ és $G = \sum_{k=-\infty}^{\infty} |k|R(k)$
g(·): spektrális sűrűségfüggvény
R(·): autokovariancia függvény

Blokkméret kiválasztása/2

- Optimális blokkméret: $b_{opt} = [(\frac{2G^2}{D})n^{1/3}]$
- Kérdés: hogyan becsüljük G-t és D-t
- $\hat{D} = \frac{4}{3}\hat{g}^2(0)$
-

$$\hat{G} = \sum_{k=-M}^{M} \lambda(\frac{k}{M})|k|\hat{R}(k)$$

ahol $\hat{R}(k) = N^{-1}\sum_{k=1}^{N-|k|}(X_i - \overline{X}_N)(X_{i+|k|} - \overline{X}_N)$

$$\lambda(t) = \begin{cases} 1 & \text{ha } |t| \in [0, 1/2] \\ 2(1-|t|) & \text{ha } |t| \in [1/2, 1] \\ 0 & \text{különben} \end{cases}$$

$M = 2\hat{m}$, ahol $\hat{m}$: ahonnan a korrelogram "lényegében" 0

Paraméteres bootstrap

- Eddig semmilyen modellt nem használtunk
- Ha van jó modellünk, akkor azt érdemes a bootstrapnél is alkalmazni
- A legegyszerűbb esetben egyszerűen a becsült modellből vesszük a mintát
- Regressziós modelleknél minta a reziduálisokból, majd ezt adjuk hozzá az illesztett értékhez
- Választás a vizsgálat célja alapján:
  - Modell kiválasztás: nemparaméteres bootstrap
  - Modell megbízhatóság: paraméteres bootstrap

Egyszerű példa a paraméteres bootstrapra

- Kérdés: lehet-e 1 az alakparametere az illesztett gamma eloszlásnak?
- Bootstrap mintákat veszünk az exponenciális eloszlásbó (ez a $\Gamma(1, \lambda)$ eloszlás).
- Statisztika: ezekre a mintákra az alakparaméter ML becslése
- Bootstrap p-érték: azon esetek aránya, ahol távolabb vagyunk 1-től, mint a megfigyelt eset becslése

AR-sieve (szűrő) bootstrap

- Feltétel: a folyamat stacionárius és jól becsülhető AR(p) modellel:

  $X_t - \mu_X = \sum_{j=1}^{p}\phi_j(X_{t-j} - \mu_X) + \varepsilon_t, \quad t \in \mathbb{Z}$

  ahol $\mu_X = EX_t$

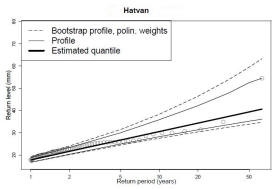
  $(\varepsilon_t)_{t\in\mathbb{Z}}$ i.i.d., E(ε_t)=0 és ε_t független { $X_s; s < t$ }-től
- Paraméterek és hibák becslése:
  - $\hat{p}$=? $\longrightarrow$ AIC
  - $\hat{\mu}_X = n^{-1}\sum_{t=1}^{n} X_t$
  - $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$=? $\longrightarrow$ Yule-Walker módszer
  - $R_t = X_t - \sum_{j=1}^{\hat{p}}\hat{\phi}_j X_{t-j}$, ahol $t = \hat{p}+1, \ldots, n$ ebből pedig $\hat{\varepsilon}_t = R_t - \overline{R}_t$, ahol $t = \hat{p}+1, \ldots, n$
- Bootstrap minta konstruálásának lépései:
  - ε_t^*: véletlen elem { $\hat{\varepsilon}_{\hat{p}+1}, \ldots, \hat{\varepsilon}_n$ } halmazból
  - Nagy u-ra $(X_{-u}^*, \ldots, X_{-u+\hat{p}-1}^*) = (\hat{\mu}_X, \ldots, \hat{\mu}_X)$ (a folyamat indítása)
  - $X_t^* = \mu_X + \sum_{j=1}^{p}\phi_j(X_{t-j}^* - \mu_X) + \varepsilon_t^* \quad t \in \mathbb{Z}$
- Ebből a bootstrap minta: { $X_1^*, \ldots, X_n^*$ }

Súlyozott (vad) bootstrap

- Itt már nem bootstrap mintát veszünk, hanem súlyozunk (például a likelihood függvényt)
- Formálisan: $Z_i^{(k)}$ súlyok, $E(Z_i^{(k)}) = 0$ és $D^2(Z_i^{(k)}) = 1$ ahol $i = 1, \ldots, n$, $k = 1, \ldots, N$ (N a boostrap ismétlések száma).
- A klasszikus esetben $\underline{Z}$ polinomiális eloszlású
- Az első alkalmazás a regressziónál: $\hat{y}_i^* = \hat{y}_i + Z_i \varepsilon_i$
- Heteroszkedasztikus esetben érdemes használni
- További alkalmazási lehetőség: kopulák illeszkedésvizsgálata

Bootstrap az extrém-érték modellekben

- A nemparaméteres bootstrap kis mintákra tipikusan túl szűk konfidenciaintervallumokat ad
- Aszimptotikusan is érdemes $m << n$ elemű bootstrap mintákat venni és ezzel párhuzamosan a feladatot kevésbé extrém kvantilisek becslésére visszavezetni
- Finomhangolni paraméterek (s, t) segítségével lehet
- Itt egy óvatos megközelítést is mutatunk: a bootstrap mintákra számolt profile likelihood intervallumok mediánját is felvettük



ábra: Különböző konfidencia intervallumok a visszatérési szintre

Hall és Weissman módszere

- A cél: $D_1(t, n, x) := E\left\{(F_{\hat{\theta}(t)}(x) - F(x))^2\right\} \to \min_t$
- Ha az $1 - p$-kvantilist becsüljük, akkor átírható: $D_2(t, n, x) := D_1(t, n, \overline{F}^{-1}(p)) = E\left\{(F_{\hat{\theta}(t)}(\overline{F}^{-1}(p)) - p)^2\right\} \to \min_t$
- A bootstrap becslések $\hat{D}_1(t, m, y) = E'\left\{\left(F_{\hat{\theta}^*(t)}(y) - \hat{F}(y)\right)^2\right\}$ és $\hat{D}_2(t, m, q) = E'\left\{\left(F_{\hat{\theta}^*(t)}\left(\hat{\overline{F}}^{-1}(q)\right) - q\right)^2\right\}$.
- Arra kell ügyelni, hogy a transzformációnál a $\log(x)/\log(n)$ hányados legalábbis aszimptotikusan ne változzon, mikor áttérünk (n, x) helyett az (m, y) párra.

Hivatkozások

- Efron, B. and Tibshirani, R.J.: An Introduction to the Bootstrap (1993)
- Kysely, J. : A cautionary note... (2008)
- Bickel, P.J., Götze, F. and van Zwet, W.R.: Resampling fewer than n observations (1997)
- Lahiri, S.N.: Resampling methods for dependent data (Springer, 2003)
- Bickel, P.J. and Sakov, A.: On the Choice of m in the m Out of n Bootstrap and its Application to Confidence Bounds for Extrema (2008)
- Politis, D. N. and White, H.: Automatic Block-Length Selection for the Dependent Bootstrap (2004)