Butstrapa metožu pielietojumi testa statistikām atkarīgiem datiem

Mārcis Bratka
Latvijas Universitāte

2012

SATURS

U un V statistikas

U (unbiased) statistikas teoriju izveidoja W. Hoeffding (1948).

Dotai izlasei $X_1, \ldots, X_n$ un $n \geq m$

$$U_n(h) = \frac{1}{\binom{n}{m}} \sum_{C_{m,n}} h(X_{i_1}, \ldots, X_{i_m}),$$

kodols h ir simetriska funkcija.

R. von Mises (1947) ieviesa V statistiku un izveidoja tās robežsadalījumu teoriju.

U un V statistiku piemēri

Piemērs

- Vidējā vērtība : $U_n = \bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$, kur $h(x) = x$,
- Dispersija : $U_n = \frac{2}{n(n-1)} \sum_{i \leq j} \frac{(X_i - X_j)^2}{2}$ un

$V_n = \frac{1}{n^2} \sum_{i,j} \frac{(X_i - X_j)^2}{2}$, kur $h(x_1, x_2) = \frac{(x_1 - x_2)^2}{2}$.

Mežonīgais atkarīgais butstraps

Mežonīgo atkarīgo butstrapu (dependent wild bootstrap) ieviesa X. Shao (2010), kas darbojas gludiem funkcionāļiem no izlases vidējās vērtības.

Doti $X_1,\ldots,X_n$ ir stacionāri ar $\mu = E(X_t)$ un $\gamma_k = \mathrm{cov}(X_0,X_k)$. Butstrapa izlases elementi ir

$$X_i^* = \bar{X}_n + (X_i - \bar{X}_n)W_i,\ i = 1,\ldots,n,$$

$\bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t$ un $\{W_i\}_{i=1}^{n}$ ir gadījuma lielumi.

Gadījuma lielumi $\{W_t\}_{t=1}^{n}$ ir neatkarīgi no $X_1,\ldots,X_n$, $E(W_t) = 0$ un $\mathrm{var}(W_t) = 1$, $t = 1,\ldots,n$. W_t ir stacionāri ar $\mathrm{cov}(W_t,W_{t'}) = a((t-t')/l)$, kur $a(\cdot)$ ir kodola funkcija un $l = l_n$

Mežonīgā atkarīgā butstrapa konsistence

X. Shao parādīja, ka pie attiecīgiem nosacījumiem

$$\sup_{x\in R} |P(\sqrt{n}[H(\bar{X}_n) - H(EX_1)] \le x) - P^*(\sqrt{n}[H(n^{-1}\sum_{t=1}^{n} X_t^*) - H(\bar{X}_n)] \le x)| \to^P 0,$$

H ir gluda funkcija.

U un V statistiku robežsadalījumi atkarīgiem datiem

$(X_t)_{t\in Z}$ ir stacionārs τ - atkarīgs process (Dedecker and Prieur, 2005).

Statistiku

$$U_n = n^{-1}\sum_{s,t=1,s\neq t}^{n} h(X_s,X_t),$$

$$V_n = n^{-1}\sum_{s,t=1}^{n} h(X_s,X_t)$$

robežsadalījumi ir

$$V_n \to^d Z := \sum_k \lambda_k Z_k^2,$$

$$U_n \to^d Z - Eh(X_0,X_0)$$

$(Z_k)_k$ ir Normāli sadalīti gadījuma lielumi un λ_k nenulles īpašvērtības.

Mežonīgais atkarīgais butstraps U un V statistikām

A. Leucht un M. H. Neumann (2011) piedāvāja mežonīgo atkarīgo butstrapa metodi U un V statistikām atkarīgiem datiem.

$$U_n^* = n^{-1} \sum_{s,t=1,s\neq t}^{n} W_s^* h(X_s, X_t) W_t^*,$$

$$V_n^* = n^{-1} \sum_{s,t=1}^{n} W_s^* h(X_s, X_t) W_t^*.$$

Pie attiecīgiem pieņēmumiem tika parādīts, ka

$$\sup_{x\in R} |P^*(U_n^* \leq x) - P(U_n \leq x)| \to^P 0,$$

$$\sup_{x\in R} |P^*(V_n^* \leq x) - P(V_n \leq x)| \to^P 0.$$

Cramer-von Mises tests atkarīgiem datiem

Doti $X_1, \ldots, X_n$ ar kādu sadalījumu F un

$$H_0 : F = F_0 \text{ pret } H_1 : F \neq F_0.$$

Vispārinātais Cramer-von Mises tests

$$T_n = n \int_{R^d} [F_n(z) - F_0(z)]^2 \omega(z) \lambda^d dz,$$

F_n ir empīriskā sadalījuma funkcija un ω ir svaru funkcija.

T_n var pārrakstīt formā

$$T_n = n^{-1} \sum_{s,t=1}^{n} h(X_s, X_t),$$

$$h(x,y) = \int [1_{x \leq z} - F_0(z)][1_{y \leq z} - F_0(z)] \omega(z) \lambda^d dz.$$

Butstrapa procedūra

1. Simulējam $W_1^*, \ldots, W_n^*$;
2. Aprēķinām butstrapa versijas testa statistiku
$$T_n^* = n^{-1} \sum_{s,t=1}^{n} h(X_s, X_t) W_s^* W_t^*;$$
3. Atkārtojam (1) un (2) B reizes un nosakām $100(1-\alpha)\%$ procentīli t_α^* no T_n^*;
4. Noraidām H_0, ja $T_n > t_\alpha^*$.

Simulāciju rezultāti

Izlases ar apjomu n = 200 tiek ģenerētas no AR(1) procesa

$$X_t = \theta X_{t-1} + \epsilon_t,$$

ϵ_t ir iid Normāli sadalīti gadījuma lielumi. Hipotēze

$$H_0 : P_X = N(0,1) \text{ pret } H_1 : P_X \neq N(0,1).$$

Testa noraidīšanas biežumi

θ	N(0,1)	N(0,2)	N(0.3,1)	N(0.5,1)
0.5	0.06	0.34	0.74	0.99
−0.5	0.055	0.95	1	1
0.9	0.325	0.37	0.535	0.7
−0.9	0.12	0.705	1	1

Butstrapa metode baltā trokšņa testam

Baltā trokšņa testi ir klasiska problēma laikrindu analīzē. Ja laikrinda sastāv no elementiem, kas ir veidoti no baltā trokšņa procesa, tad nav vajadzības meklēt datiem atbilstošu ARMA modeli.

Butstrapa metodi baltā trokšņa testam piedāvāja X. Shao (2011), kas balstīta uz mežonīgā bloku butstrapa (blockwise wild bootstrap) metodi.

Pieņemsim, ka $X_t, t \in Z$ ir stacionāri un $E(X_t) = \mu$ un $\gamma(k) = \mathrm{cov}(X_t, X_{t+k})$, tad hipotēze

$$H_0 : \gamma(k) = 0, k \in N \text{ pret } H_1 : \gamma(k) \neq 0 \text{ kādam } k \in N.$$

Spektrālās analīzes elementi

$f(\omega) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\omega}$, $\omega \in [-\pi,\pi]$ ir spektra blīvuma funkcija un $F(\lambda) = \int_0^{\lambda} f(\omega) d\omega$, $\lambda \in [0,\pi]$ ir spektra sadalījuma funkcija.

Pie H_0 funkcija $f(\lambda) = \gamma(0)/(2\pi)$ ir konstanta apgabalā $[-\pi,\pi]$.

Testa ideja ir balstīta uz izlases spektra sadalījuma funkciju $F_n(\lambda) = \int_0^{\lambda} I_n(\omega) d\omega$,

$$I_n(\omega) = (2\pi n)^{-1} \sum_{t=1}^{n} |(X_t - \bar{X}) e^{it\omega}|^2$$

ir periodogramma un $\bar{X}$ ir izlases vidējā vērtība.

Testa statistika

Periodogrammu var pārrakstīt formā
$I_n(\omega) = (2\pi)^{-1} \sum_{k=1-n}^{n-1} \hat{\gamma}(k) e^{-ik\omega}$, kur
$\hat{\gamma}(j) = n^{-1} \sum_{t=1+|j|}^{n} (X_t - \bar{X})(X_{t-|j|} - \bar{X})$ ir izlases autokovariāciju funkcija.

Tad $F_n(\lambda) = \sum_{j=0}^{n-1} \hat{\gamma}(j)\Psi_j(\lambda)$ un $F(\lambda) = \sum_{j=0}^{\infty} \gamma\Psi_j(\lambda)$, kur $\Psi_j(\lambda) = \sin(j\lambda)/(j\pi)1_{j\neq 0} + \lambda/(2\pi)1_{j=0}$.

Pie H_0 $F(\lambda) = \gamma(0)\Psi_0(\lambda)$. Tālāk konstruē statistiku kā distanci starp $F_n(\lambda)$ un $\hat{\gamma}(0)\Psi_0(\lambda)$.

$$S_n(\lambda) = \sqrt{n}\{F_n(\lambda) - \hat{\gamma}(0)\Psi_0(\lambda)\} = \sum_{j=1}^{n-1} \sqrt{n}\hat{\gamma}(j)\Psi_j(\lambda).$$

Testa statistikas robežsadalījumi

Pēc teorēmas $S_n(\lambda)$ robežsadalījums dod

- Kolmogorov–Smirnov statistiku

$$\sup_{\lambda\in[0,\pi]}|S_n(\lambda)| \to_D \sup_{\lambda\in[0,\pi]}|S(\lambda)|;$$

- Cramer von–Mises statistiku

$$CM_n = \int_0^\pi S_n^2(\lambda)d\lambda \to_D \int_0^\pi S^2(\lambda)d\lambda,$$

kur $S(\lambda)$ ir Gaussian process ar vidējo vērtību 0 un $\mathrm{cov}(S(\lambda),S(\lambda'))$.

Butstrapa procedūra

Butstrapa procedūrā tiek lietota bloku mežonīgā butstrapa metode, lai aproksimētu CM_n robežsadalījumu.

1. Izvēlas bloka garumu b_n. Apzīmē blokus ar $B_s = \{(s-1)b_n + 1, \ldots, sb_n\}$, $s = 1, \ldots, L_n$, kur L_n ir bloku skaits;
2. Ģenerē iid gadījuma lielumus δ_s, $s = 1, \ldots, L_n$, kas neatkarīgi no izlases elementiem $X_1, \ldots, X_n$, ar sadalījumu W tādu, ka $EW = 0$, $EW^2 = 1$ un $EW^4 \leq \infty$. Definējam $\omega_t = \delta_s$, ja $t \in B_s$, $t = 1, \ldots, n$;
3. $\hat{\gamma}^*(j) = n^{-1} \sum_{t=j+1}^{n} \{(X_t - \bar{X})(X_{t-j} - \bar{X}) - \hat{\gamma}(j)\}\omega_t$ visiem $j = 1, \ldots, n-1$ un definē butstrapoto procesu

$$S_n^*(\lambda) = \sqrt{n} \sum_{j=1}^{n-1} \hat{\gamma}^*(j)\Psi_j(\lambda);$$

Butstrapa procedūra (turp.)

4. Butstrapa testa statistika $\mathrm{CM}_n^* = \int_0^\pi \{\mathrm{S}_n^*(\lambda)\}^2 \mathrm{d}\lambda$;
5. Atkārto soļus 2 un 3 B reizes un nosaka $\mathrm{CM}_{n,\alpha}^*$ $100(1-\alpha)\%$ procentīli no CM_n^*;
6. Noraida H_0, ja $\mathrm{CM}_n > \mathrm{CM}_{n,\alpha}^*$.