Butstrapa metodes atkarīgiem datiem

Mārcis Bratka

2010

Tapered block bootstrap

$X_1, X_2, \ldots$ ir stacionāri un vienādi sadalīti gadījuma lielumi no R^m ar sadalījumu F. $X_1, \ldots, X_n$ ir dota izlase un $T_n = T_n(X_1, \ldots, X_n)$ ir novērtējums parametram $T(F)$. Piemēram,

$$T_n = f(n^{-1} \sum_{t=1}^{n} \phi(X_t)),$$

kur $f : R^d \to R$ un $\phi : R^m \to R^d$.
MBB (Kunsch 1989) novērtējums $Var(\sqrt{n} T_n)$ bloka garumam b ir $\hat{\sigma}^2_{b,MBB}$. Pie dažiem momentu un jaukto procesu nosacījumiem

$$Bias(\hat{\sigma}^2_{b,MBB}) = O(1/b) \text{ un } Var(\hat{\sigma}^2_{b,MBB}) = O(b/n).$$

T_n ir lineāra, ja

$$T_n = T(F) + n^{-1} \sum_{t=1}^{n} IF(X_t, F) + R_n,$$

kur R_n ir mazs. Funkcija IF statistikai T_n ir definēta

$$IF(y, F) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon\delta_y) - T(F)}{\epsilon},$$

kur δ_y punktā y ir 1. Praktiksi $IF(y, F)$ netiek lietota, jo F nav zināms, bet lieto empīrisko $IF(y, \hat{F}_n)$, kur $\hat{F}_n$ ir empīrisks sadalījums ar $1/n$ katrā $X_1, \ldots, X_n$.

$$\sup_y |IF(y, \hat{F}_n) - IF(y, F)| = O_P(1/\sqrt{n}).$$

$$Y_t := \mathrm{IF}(X_t, \hat{F}_n),\ t = 1, \ldots, n.$$

Piemērs

$T_n = \bar{X} = n^{-1} \sum_{t=1}^{n} X_t$, kurai $m = 1$ un f un ϕ ir identitātes funckijas, $R_m = 0$. $\mathrm{IF}(y, F) = y - \int x dF(x)$ un $\mathrm{IF}(y, \hat{F}_n) = y - \hat{X}$, tātad $Y_t = \mathrm{IF}(X_t, \hat{F}_n) = X_t - \hat{X}$.

Tapered block bootstrap (TBB) ideja ir lietot svarus $w_n(\cdot)$, $n = 1, 2, \ldots$, lai samazinātu bloka sākuma un beigu elementu nozīmi. $w_n(t)$ vētības ir $[0, 1]$ un $w_n(t) = 0, t \notin \{1, \ldots, n\}$. $\|w_n\|_1 = \sum_{t=1}^{n} |w_n(t)|$ un $\|w_n\|_2 = \sqrt{\sum_{t=1}^{n} w_n^2(t)}$.

$$w_n(t) = w\left(\frac{t-0.5}{n}\right), \text{ kur}$$

$w(\cdot)$ apmierina nosacījumus:

- $w(t) \in [0, 1]$ visiem $t \in R$, $w(t) = 0$, ja $t \notin [0, 1]$, un $w(t) > 0$ pie $t = 1/2$;
- $w(t)$ ir simetriska pret $t = 1/2$ un nedilstoša $t \in [0, 1/2]$.

Ja $w(t) = 1_{[0,1]}(t)$, tad TBB ir ekvivalents MBB. Lai TBB iegūtu novirzes uzlabojumu w ir jābūt gludai, t.i., $w * w(t) = \int_{-1}^{1} w(x)w(x + |t|)dx$ ir divreiz nepārtaukti diferencējama $t = 0$.

TBB procedūra:

- izvēlamies bloka garumu b un $i_0, i_1, \ldots, i_{k-1}$ ir vienmērīgi sadalīti no $\{1, \ldots Q\}$, $Q = n - b + 1$, $k = [n/b]$;
- konstruējam $Y_1^*, \ldots, Y_l^*$ ar $l = kb$

$$Y_{mb+j}^* := w_b(j) \frac{\sqrt{b}}{\|w_b\|_2} Y_{i_m+j-1}, \text{ kur } j = 1, \ldots, b,$$
$$m = 0, 1, \ldots, k-1;$$

- $\bar{Y}_l^* = l^{-1} \sum_{i=1}^{l} Y_i^*$.

TBB dispersijas novērtējums $\hat{\sigma}_{b,TBB}^2 := \mathrm{Var}^*(\sqrt{l}\bar{Y}_l^*)$ un sadalījuma $P(\sqrt{n}(T_n - T(F)) \leq x)$ novērtējums $P^*(\sqrt{l}\bar{Y}_l^* \leq x)$.

Teorēma

$E\hat{\sigma}^2_{b,TBB} = \sigma^2_\infty + \Gamma/b^2 + o(1/b^2)$ un $\mathrm{Var}(\hat{\sigma}^2_{b,TBB}) = \Delta\frac{b}{n} + o(b/n)$,

kur Γ un Δ ir konstantes. Pie tam
$\sup_x |P^*(\sqrt{l}(\bar{Y}^*_l - E^*\bar{Y}^*_l) \leq x) - P(\sqrt{n}(T_n - T(F)) \leq x)| \longrightarrow_p 0.$

$\mathrm{MSE}(\hat{\sigma}^2_{\mathrm{b,TBB}}) = \frac{\Gamma^2}{\mathrm{b}^4} + \Delta\frac{\mathrm{b}}{\mathrm{N}} + \ldots$. MSE tiek minimizēts pie $\mathrm{b_{opt}} = (\frac{4\Gamma^2}{\Delta})^{1/5}\mathrm{n}^{1/5}$. Tad minimālā asimptotiskā MSE

$$\mathrm{MSE_{opt}} = (\Gamma^{2/5}\Delta^{4/5}\frac{5}{4^{4/5}})\mathrm{n}^{-4/5}.$$

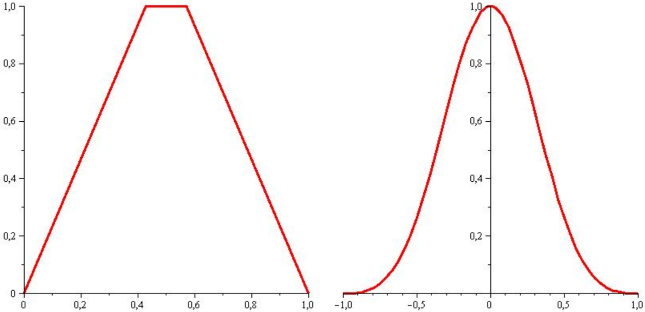
Δ un Γ ir atkarīgas no w un kovariācijām. w tiek izvēlēts tā, lai minimizētu

$$|\tilde{\mathrm{w}}''(0)| * \|\tilde{\mathrm{w}}\|_2^4, \text{ kur}$$

$\tilde{\mathrm{w}}(\mathrm{t}) = (\mathrm{w} * \mathrm{w})(\mathrm{t})/(\mathrm{w} * \mathrm{w})(0)$.

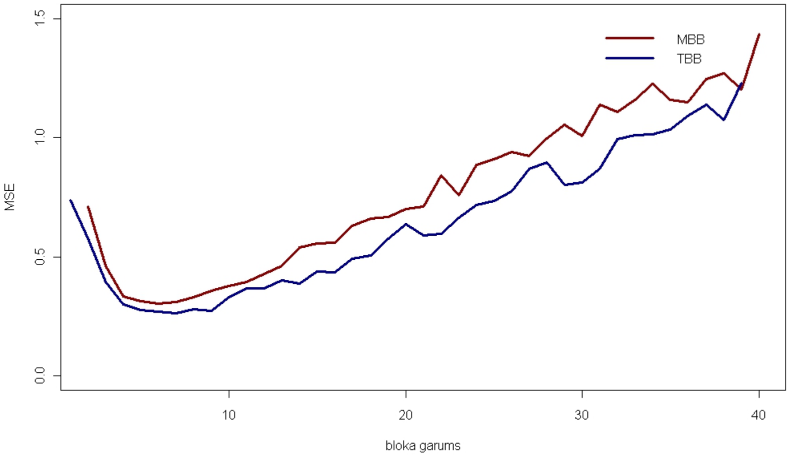
$$w_c^{TRAP}(t) = \begin{cases} t/c & t \in [0, c]; \\ 1 & t \in (c, 1-c];, \\ (1-c)/t & t \in (1-c, 1]. \end{cases}$$

kur $c \simeq 0.43$.



(a) (b)

$X_t = 0.6\sin(X_{t-1}) + Z_t$, kur Z_t ir $N(0,1)$ un $n = 200$. $T_n = \bar{X}$.



Dependent wild bootstrap

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

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

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

Pieņēmums

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 $l = l_n$.

Pie momentu un kovariāciju nosacījumiem
$T_n := \sqrt{n}(\bar{X}_n - \mu) \to_D N(0, \sigma_\infty^2)$, kur $\sigma_\infty^2 = \sum_{j=-\infty}^{\infty} \gamma_j$.

$$\begin{aligned}
\hat{\sigma}^2_{l,DWB} &= n^{-1} \sum_{t,t'=1}^{n} (X_t - \bar{X}_n)(X_{t'} - \bar{X}_n)\mathrm{cov}^*(W_t, W_{t'}) \\
&= n^{-1} \sum_{h=1-n}^{n-1} \sum_{t=\max(1,1-h)}^{\min(n,n-h)} (X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n) a(h/l) \\
&= 2\pi \hat{f}_n(0), \text{ kur}
\end{aligned}$$

$\hat{f}_n(\lambda) = (2\pi)^{-1} \sum_{k=1-n}^{n-1} a(k/l)\hat{\gamma}_k \cos(k\lambda)$ ir spektra novērtējums un $a(\cdot)$ ir lag window funkcija.

Teorēma

$\mathrm{E}(\hat{\sigma}^2_{\mathrm{l,DWB}}) = \sigma^2_\infty + \Gamma/\mathrm{l}^2 + \mathrm{o}(1/\mathrm{l}^2)$ un $\mathrm{var}(\hat{\sigma}^2_{\mathrm{l,DWB}}) = \Delta\frac{\mathrm{l}}{\mathrm{n}} + \mathrm{o}(\mathrm{l}/\mathrm{n})$,

kur Γ un Δ ir konstantes.

- $\hat{\sigma}^2_{\mathrm{l,DWB}}$ un $\hat{\sigma}^2_{\mathrm{l,TBB}}$ ir asimptotiski vienādi, ja $\mathrm{a}(\mathrm{x}) = \mathrm{w} * \mathrm{w}(\mathrm{x})/\mathrm{w} * \mathrm{w}(0)$ un l ir vienādi abām metodēm.
- Ja $\mathrm{a}(\mathrm{x}) = (1 - |\mathrm{x}|)1(|\mathrm{x}| \leq 1)$, tad DWB ir vienāds ar MBB.

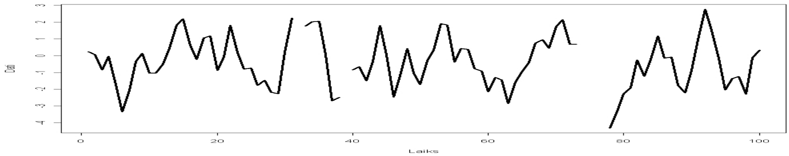
Vispārīgā gadījumā DBW ir lietojams lineārām statistikām.

- Dotai izlasei $X_1, \ldots, X_n$ ar sadalījumu F parametra $\theta = T(F)$ novērtējums ir $\hat{\theta} = T(\rho_n)$, kur $\rho_n = n^{-1} \sum_{t=1}^{n} \delta_{X_t}$.
- $T_n(\rho_n) = T(F) + n^{-1} \sum_{t=1}^{n} IF(X_t, F) + R_n$.
- Pie nosacījumiem $n\mathrm{var}(\hat{\theta}) = n^{-1}\mathrm{var}(\sum_{t=1}^{n} IF(X_t, \rho_n)) + o(1)$.
- Butstrapotais mērs $\rho_n^* = n^{-1} \sum_{i=1}^{n} (W_i + 1 - \bar{W}_n)\delta_{X_i}$, kur $\bar{W}_n = n^{-1} \sum_{t=1}^{n} W_t$ un $\{W_t\}_{t=1}^{n}$ ir gadījuma lielumi.

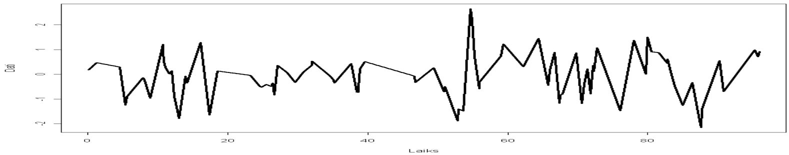
Ja $T(F) = \int x dF$, tad $T(\rho_n) = \bar{X}_n$ un $T(\rho_n^*) = n^{-1} \sum_{t=1}^{n} (W_t + 1 - \bar{W}_t) = \bar{X}_n + n^{-1} \sum_{t=1}^{n} W_t(X_t - \bar{X}_n)$.
Vispārīgākām statistikām var būt problēmas ar butstrapa izlases iegūšanu, jo ρ_n^* nav varbūtību mērs.

- Šobrīd nav iegūti rezultāti par second-order correctness, bet ir zināms, ja W_t tiek izvēlēts ar normālo sadalījumu, tad vispārīgā gadījumā second-order correctness netiks sasniegts.
- Ja X_t ir normālais sadalījums, tad, iespējams, DWB būs second-order correctness ar W_t normālo sadalījumu, gludas funkcijas modeļa gadījumā.
- Šobrīd nav rekomendācijas, kā izvēlēties W_t pie dotiem $X_1, \dots, X_n$, lai sasniegtu second-order correctness.
- No precizitātes viedokļa svarīgāk ir izvēlēties $a(\cdot)$ un l (first-order), kā W_t.

- Trūkstoši elementi (missing values).



- Laika intervāli ir gadījuma lielumi.



$X_t = 0.6\sin(X_{t-1}) + Z_t$, kur Z_t ir $N(0,1)$ un $n = 200$. $T_n = \bar{X}$ ar $a(x) = (1 - |x|)1(|x| \leq 1)$.