Liet. matem. rink., T. 43, spec. nr., 475-479

Pasiskirstymo „uodegos“ indekso pasikeitimo nustatymas

Kęstutis GADEIKIS (VU)
el. paštas: gadeikis@ldr.lt

Šiame straipsnyje nagrinėjamas prof. V. Paulausko pasiūlytas naujas įvertis pasiskirstymo „uodegos“ indeksui nustatyti, tiriant hipotezę apie galimą jo pasikeitimą iš anksto nežinomoje vietoje. Šalia nesudėtingų teorinių skaičiavimų pateikiami modeliavimo rezultatai, leidžiantys tikėtis sėkmingų naujojo įverčio praktinių taikymų.

Šalia eksponentiniu gesimu pasižyminčių pasiskirstymų pastaruoju metu ypač daug dėmesio skiriama sunkiomis „uodegomis“ pasižymintiems atsitiktiniams dydžiams – jų pasiskirstymo funkcija F gęsta laipsniškai ir didelėms argumento x reikšmėms tenkina sąryšį

$$1 - F(x) = x^{-\alpha} L(x),$$

čia koeficientas $\alpha > 0$ yra vadinamas „uodegos“ indeksu, o funkcija L begalybės aplinkoje yra lėtai kintanti:

$$\lim_{x \to \infty} \frac{L(tx)}{L(x)} = 1 \quad \forall t > 0.$$

Vienas iš labiausiai žinomų ir dažniausiai praktikoje taikomų „uodegos“ indekso įverčių yra B.M. Hill pasiūlytas parametro $\gamma = 1/\alpha$ įvertis:

$$\gamma_{n,k} = \frac{1}{k} \sum_{i=0}^{k-1} \log X_{n,n-i} - \log X_{n,n-k},$$

čia $X_{n,1} \leqslant X_{n,2} \leqslant \ldots \leqslant X_{n,n}$ yra atsitiktinių dydžių $X_1, \ldots, X_n$ variacinė seka. Beje, parametro $1 \leqslant k \leqslant n$ parinkimas yra gana sunkus uždavinys, priklausanti nuo nežinomų pasiskirstymo funkcijos parametrų, ypač nežinant tikrosios „uodegos“ indekso reikšmės.

Prof. V. Paulausko pasiūlytas naujas įvertis yra kur kas paprastesnis ir lengviau realizuojamas. Padalinkime seką $X_1, \ldots, X_N$ į n lygių grupių $V_1, \ldots, V_n$, turinčių po m atsitiktinių dydžių, paprastumo dėlei laikydami, jog $N = n \cdot m$. Tegu

$$M_{ni}^{(1)} = \max\{X_j \colon X_j \in V_i\},$$

o $M_{ni}^{(2)}$ pažymėkime antrąjį pagal dydį tos pačios grupės V_i elementą. Tada apibrėžkime

$$\kappa_{ni} = \frac{M_{ni}^{(2)}}{M_{ni}^{(1)}}, \quad S_n = \sum_{i=0}^{n} \kappa_{ni}, \quad Z_n = n^{-1} S_n.$$

Kaip matome, κ_{ni} yra i-tosios grupelės dviejų didžiausių elementų santykis, o Z_n – tokių santykių vidutinė reikšmė.

[2] straipsnyje įrodoma, kad kai $N \to \infty$, $m = m(N) \to \infty$ ir $m/N \to 0$,

$$Z_n \xrightarrow{b.t.} \frac{\alpha}{\alpha + 1}.$$

Beje, įrodyme pateikiamas κ_{ni} konvergavimas pagal pasiskirstymą $\forall i$, kai $m \to \infty$:

$$\kappa_{ni} \xrightarrow{D} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^{1/\alpha},$$

čia λ_1 ir λ_2 yra standartiniai (vidurkis lygus vienetui) eksponentiniai atsitiktiniai dydžiai. Pertvarkę reiškinį

$$E\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^{1/\alpha} = E\left(1 + \frac{\lambda_2}{\lambda_1}\right)^{-1/\alpha} = E(1+z)^{-1/\alpha}$$

ir nesunkiai radę atsitiktinio dydžio z tankio funkciją, lygią $x/(x+1), x \geqslant 0$, lengvai paskaičiuojame ne tik lygų $\alpha/(\alpha+1)$ vidurkį , bet ir dispersiją, lygią

$$\frac{\alpha}{(\alpha+1)^2(\alpha+2)}.$$

Reikėtų pastebėti, jog optimalaus narių skaičiaus grupelėse m, susieto su grupelių skaičiumi n sąryšiu $N = m \cdot n$, parinkimas vis dar išlieka problema, ypač esant ne itin dideliam duomenų kiekiui, nes nors atsitiktinis dydis Z_n yra asimptotiškai nepaslinktas, tai dar neužtikrina gerų taikymo praktikoje rezultatų. Prie šio klausimo dar sugrįšime.

Pastebėjus, jog tirti atsitiktinius dydžius κ_{ni} yra kur kas paprasčiau, o pasikeitus indeksui pasikeičia ir vidutinė Z_n reikšmė, belieka pasirinkti tinkamą pasikeitimo fakto ir vietos nustatymo metodą. Taip pat pastebėkime, jog nors sprendžiame modifikuotą uždavinį (nagrinėjami ne atsitiktiniai dydžiai $X_1, X_2, \ldots, X_N$, o išvestiniai $\kappa_{n1}, \ldots, \kappa_{nn}$) ir dalinai prarandame tikslumą, tačiau išlaikant Z_n konvergavimui keliamą sąlygą $m/N \to 0$, kai $N \to \infty$, toks tikslumo praradimas tampa nereikšmingas.

Natūralu, kad tiriama hipotezė H_0, jog pasiskirstymo „uodegos“ indeksas nepasikeitė, esant alternatyvai, jog egzistuoja toks taškas k^*, kad atsitiktiniai dydžiai $\kappa_{n1}, \ldots, \kappa_{nk^*}$ ir $\kappa_{n,k^*+1}, \ldots, \kappa_{nn}$ pasižymi skirtingais indeksais, čia k^* yra indekso pasikeitimo taškas:

$$\alpha_1 = \ldots = \alpha_{k^*} \neq \alpha_{k^*+1} = \ldots = \alpha_n.$$

Iš pradžių buvo pasirinktas [4] straipsnyje išdėstytas Kolmogorovo–Smirnovo kriterijus. Deja, kriterijus yra jautrus ne tik vidurkio, bet ir kitokio pobūdžio pasiskirstymo

pasikeitimams, be to, pasiskirstymo funkcijų perskaičiavimai užima gana daug laiko ir reikalauja papildomo optimizavimo. Nenuostabu, jog sunkia „uodega“ pasižyminčių atsitiktinių dydžių sugeneruotai serijai praktiniai kriterijaus taikymo rezultatai nebuvo itin sėkmingi, tai vaizdžiai galima pailiustruoti pavyzdžiu, pateiktu 1 pav.

Šiame brėžinyje tikrasis pasikeitimo taškas yra ties 500-aja pozicija, maksimumas pasiekiamas ties 460-aja. Pastebėkime, jog vizualiai pakankamai sunku nustatyti tikslią pasikeitimo vietą, nors maksimali statistikos reikšmė 1,64 atitiktų net 99,5% pasikliovimo lygmenį: $P(T \leqslant z) \to 1 - \mathrm{e}^{-2z^2}$, čia T yra nagrinėjama Kolmogorovo–Smirnovo statistika. Šoktelėjimas ties 460-aja pozicija atrodo kiek atsitiktinis, o šiek tiek mažesnė antroji „kupra“ taip pat klaidina vizualiai.

Todėl buvo pereita prie vidurkių pasikeitimo tyrimo. Pastebėkime, kad jeigu $X_1, X_2, \ldots, X_n$ yra nepriklausomi vienodai pasiskirstę atsitiktiniai dydžiai su nuliniu vidurkiu ir vienetine dispersija, tai pasikeitimo tašku spėjant esant tašką $X_k, 1 < k < n$, pirmos ir antros dalies vidurkių skirtumo statistiką būtų galima užrašyti taip:

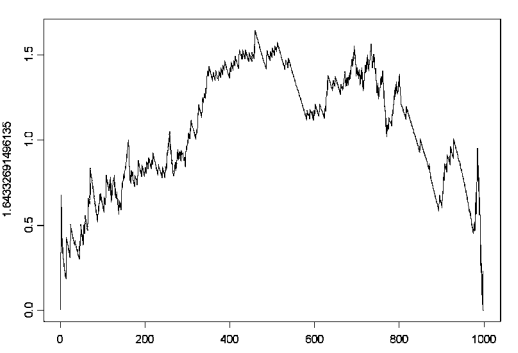
$$T_n(k) = \frac{1}{k}\sum_{i=1}^{k} X_i - \frac{1}{n-k}\sum_{i=k+1}^{n} X_i = \frac{1}{k}\sum_{i=1}^{k} X_i - \frac{1}{n-k}\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{k} X_i\right)$$
$$= \left(\frac{1}{k} + \frac{1}{n-k}\right)\sum_{i=1}^{k} X_i - \frac{1}{n-k}\sum_{i=1}^{n} X_i = \frac{n}{k(n-k)}\sum_{i=1}^{k} X_i - \frac{1}{n-k}\sum_{i=1}^{n} X_i.$$

Atitinkamai ją normavę, gauname

$$\widetilde{T}_n\,(k) = \frac{1}{\sqrt{n}}\,\frac{k(n-k)}{n}\,T_n(k) = \frac{1}{\sqrt{n}}\sum_{i=1}^{k} X_i - \frac{1}{\sqrt{n}}\,\frac{k}{n}\sum_{i=1}^{n} X_i.$$

Nesunku pastebėti, kad

$$\widetilde{T}_n\,(k) \xrightarrow{D} W(t) - tW(t) = W^0(t),$$



1 pav. Kolmogorovo–Smirnovo kriterijaus taikymas.

čia W(t) yra standartinis Vynerio procesas, o $0 < t < 1$ riboje atitinka k ir n santykį. Tokiu būdu gauname statistikos $\widetilde{T}_n\ (k)$ konvergavimą pagal pasiskirstymą į visiems gerai žinomą procesą – Brauno tiltą $W^0(t)$.

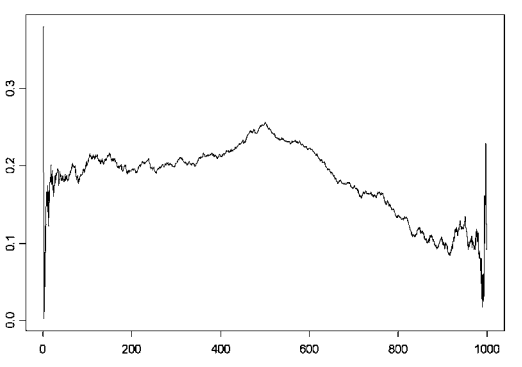
Kadangi mūsų atveju $E\kappa_{ni} \neq 0$ ir $DX\kappa_{ni} \neq 1$, nesunku įsitikinti, jog centravimas vidurkio nepakeičia, o normuoti galime dalindami ne tik iš vidutinio kvadratinio nuokrypio, kurio iš tikrųjų galime ir nežinoti, bet ir iš jo empirinės charakteristikos. Paėmę vidurkio pasikeitimą absoliutiniu dydžiu, gauname konvergavimą pagal pasiskirstymą į atitinkamą Brauno tilto funkcionalą.

Visi straipsnyje aprašyti rezultatai buvo modeliuojami statistiniu paketu S-PLUS generuojant sunkia „uodega“ pasižyminčių atsitiktinių dydžių serijas, randant atitinkamas skaitines charakteristikas ir rezultatus pavaizduojant grafiškai. Atliekant praktinius skaičiavimus, buvo remtasi [5] straipsnyje pateikiamais patarimais. Deja, ribota straipsnio apimtis neleidžia pateikti visų gautų rezultatų, tuo labiau palyginamųjų grafikų.

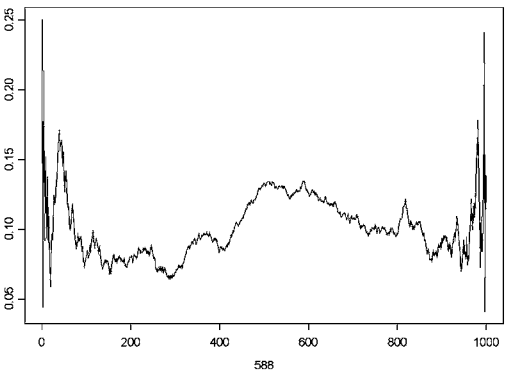
Tikrinant indekso, tiksliau, jį atitinkančios vidutinių $\kappa_{ni}, i = 1, \ldots, n$, reikšmių pasikeitimą, akivaizdu, jog pačioje sekos pradžioje ar pabaigoje vidurkių skirtumas gali būti didelis ne tiek dėl indekso pasikeitimo, kiek dėl nedidelio duomenų skaičiaus, todėl praktiniuose skaičiavimuose spėjamas „uodegos“ indekso pasikeitimo taškas ieškomas ne kaip $1 < k^* < n$, bet $\theta n \leqslant k^* \leqslant (1-\theta)n, 0 < \theta < 0,5$. Pasirenkant θ atsižvelgiama į turimų duomenų kiekį, paprastai imama $\theta = 0,1$, o esant daugiau duomenų galima pasirinkti ir $\theta = 0,05$. Tipinis vidurkių skirtumo modulio grafikas pavaizduotas 2 pav.

Generuojant šimtą 10 000 a.d. serijų su pasikeitimo tašku sekos viduryje ir indeksais $\alpha_1 = 0,7$ ir $\alpha_2 = 1,3$, imant narių skaičių grupelėje $m = 10$ tikėtiniausias κ_{ni} pasikeitimo taškas svyravo nuo 428 iki 588 su vidutine reikšme 499,2. Kaip matyti iš kraštutinio atvejo $k^* = 588$ grafiko, nustatant pasikeitimo tašką vizualiai atsitiktinio šuoliuko būtų galima išvengti (3 pav.).

Siekiant dar geresnių rezultatų, ypač esant mažesniam duomenų kiekiui, tiek optimalaus narių skaičiaus grupėleje parinkimo, tiek indekso pasikeitimo taško nustatymo metodai bus toliau tobulinami.



2 pav. Tipinis vidurkių skirtumo modulio grafikas.



3 pav. Kraštinių reikšmių ignoravimas.

Literatūra

[1] V. Paulauskas, A new estimator for tail index (to appear in *Acta Appl. Mathematicae*) (2003).
[2] Yu. Davydov, V. Paulauskas, A. Račkauskas, More on p-stable convex sets in Banach spaces, *J. of Theor. Probability*, **13**, 39–64 (2001).
[3] C. Quintos, Zh. Fan, P. Phillips, Structural change tests in tail behaviour and the Asian Crisis, *Review of Economic Studies*, **13**, 633–663 (2001).
[4] E. Carlstein, Nonparametric change-point estimation, *Ann. of Statistics*, **16**, 188–197 (1988).
[5] R. Davidson, J.G. MacKinnon, Graphical methods for investigating the size and power of hypothesis tests, *The Manchester School*, **66**, 1–26 (1998).

Determination of a change-point of tail index

K. Gadeikis

This article deals with a new estimator for tail index, introduced by prof. V. Paulauskas in [1]. Hypothesis of a change-point at beforehand-unknown place is examined. Besides simple theoretical calculations, simulation results are presented, which enable to expect successful practical applications of the estimator.