Liet. matem. rink., T. 40, spec. nr., 290–295

Branduolinis pasiskirstymo tankio įvertinimas taikant kryžminį patikrinimą

Mindaugas KAVALIAUSKAS (MII)
el. paštas: snaiperiui@takas.lt

Įvadas

Šis straipsnis yra darbo [3] tęsinys. Jame nagrinėjamas neparametrinis branduolinis pasiskirstymo tankio įvertinimo metodas. Siekiama sukurti automatizuotą pasiskirstymo tankio įvertinimo procedūrą, kuri dirbtų be vartotojo pagalbos (parametrų parinkimo ir pan.). Nagrinėjama įverčio parametrų parinkimo procedūra, kuri remiasi kryžminio patikrinimo metodika. Aptariami modeliavimo būdu gauti rezultatai. Modeliavimui naudojami Gauso skirstinių mišiniai.

Tiesioginio pakeitimo metodas

Šiame skyriuje trumpai aprašomas tirtas branduolinio pasiskirstymo tankio įvertinimo metodas. Detalesnį metodo aprašymą galite rasti straipsnyje [3].

Tegul turime nepriklausomų vienodai pasiskirsčiusių su pasiskirstymo funkcija $f(x)$ atsitiktinių dydžių imtį $\mathbf{X} = (X_1, \ldots, X_n)$. Tuomet branduolinį pasiskirstymo tankio įvertį $\widehat{f}_h$ apibrėšime

$$\widehat{f}_h(x) = n^{-1} \sum_{t=1}^{n} K_h(X_t - x), \tag{1}$$

kur $K_h(y) = \frac{1}{h} K\left(\frac{y}{h}\right)$. Čia $h = h(x, \mathbf{X})$ – branduolio pločio funkcija, $K(y)$ – branduolio funkcija tenkinanti sąlygą $\int K(y)\,\mathrm{d}y = 1$. Šiame darbe buvo naudojama Jepaničnikovo branduolio funkcija.

Pagrindinis sunkumas, atsirandantis vertinant pasiskirstymo tankį, yra branduolio pločio h parinkimas. Kaip parodė darbo [3] rezultatai, tikslinga branduolio plotį skaičiuoti tiesioginio pakeitimo (angl. *plug-in*) metodu.

Paprastai branduolio plotį h siekiama parinkti taip, kad minimizuotų vidutinę kvadratinę paklaidą, t.y.

$$h_{opt}(x) = \arg\min_h \mathbf{E}\big(\widehat{f}_h(x) - f(x)\big)^2 = \arg\min_h \big(b_h^2(x) + \sigma_h^2(x)\big). \tag{2}$$

Čia $b_h(x)$ – įverčio poslinkis, o $\sigma_h^2(x)$ įverčio dispersija. Kadangi tikrosios įverčio poslinkio ir dispresijos reikšmės nėra žinomos, jos keičiamos įverčiais.

Darbe buvo naudojamas dispersijos įvertis,

$$\widehat{\sigma}_h^2(x) = \frac{c\widehat{f}_h(x)}{nh}, \quad c = \int K^2(y)\,\mathrm{d}y, \tag{3}$$

kuris gaunamas iš asimptotinių pasiskirstymo tankio įverčio savybių (žr. [3]).

Branduolinio pasiskirstymo tankio įverčio (1) poslinkis yra

$$b_h(x) = \mathbf{E}\widehat{f}_h(x) - f(x) = \int \left[f(x+hy) - f(x)\right]K(y)\,\mathrm{d}y. \tag{4}$$

Norint gauti tankio įverčio poslinkio įvertį, tiesiog nežinoma pasiskirstymo tankio funkcijos reikšmė išraiškoje (4) keičiame jos įverčiu, ir pointegraliniame reiškinyje panaudojamas modulis (motyvaciją žr. [3]):

$$\widehat{b}_{h,\Delta}(x) = \int \left|\widehat{f}_\Delta(x+hy) - \widehat{f}_\Delta(x)\right|K(y)\,\mathrm{d}y, \quad \Delta = \Delta(h). \tag{5}$$

Remiantis (2), (3), (4) apibrėžkime tiesioginio pakeitimo metodo branduolio pločio įvertį

$$h_{PI} = \arg\min_h \left[\widehat{\sigma}_h^2(x) + \widehat{b}_{h,\Delta(h)}^2(x)\right]. \tag{6}$$

Įverčio kokybei pagerinti buvo papildomai naudojamos glodinimo ir multiplikatyvaus poslinkio sumažinimo procedūros, kurios šiame straipsnyje nebus aptariamos.

Apibrėždami įvertį dar nepaminėjome, kokia turi būti funkcija $\Delta(h)$. [2] siūloma naudoti $\Delta(h) \geqslant h$. Iš asimptotinių pasiskirstymo tankio įverčio savybių seka, kad naudotina nelygybė $\Delta(h) \geqslant c_K h$, kur konstanta c_K priklauso nuo branduolio funkcijos K parinkimo (detaliau žr. [3]). Šiame darbe buvo naudojama tokia Δ išraiška

$$\Delta(h) = \alpha\, c_K h, \tag{7}$$

kur α – įverčio parametras, kurį reikia parinkti, o naudotam Jepaničnikovo branduoliui $c_K \approx 1.54$.

Parametrų parinkimas taikant kryžminį patikrinimą

Vienas iš populiarių parametrų parinkimo metodų naudojamų pasiskirstymo tankio įvertinime yra kryžminio patikrinimo (angl. *cross-validation*) metodas. Dažniausiai įverčio efektyvumui nusakyti naudojamos vidutinės kvadratinės paklaidos, todėl labiausiai paplito mažiausių kvadratų kryžminio patikrinimo metodas. Jis minimizuoja nuostolių funkciją $\varepsilon_2 = \mathbf{E}\|\widehat{f} - f\|_2^2$. Tačiau toks metodas turi trūkumų – ši nuostolių funkcija

priklauso nuo mastelio parametro. Tai ypač svarbu vertinant mišinių tankius, kai mišinius sudaro klasteriai su labai skirtingomis glodumo savybėmis. Šio trūkumo neturi metodai paremti nuostolių $\varepsilon_1 = \mathbf{E}\|\widehat{f} - f\|_1$ erdvėje L_1 įvertinimu, tačiau tokius nuostolius sunku įvertinti.

Todėl siūlome naudoti tokią nuostolių funkciją

$$\varepsilon_A = \mathbf{E}\left\|\frac{\widehat{f} - f}{\sqrt{f}}\right\|_2^2 - 1. \tag{8}$$

Ši nuostolių funkcija yra nepriklausoma nuo mastelio parametro ir gali būti nesunkiai įvertinta (žr. [3]). Modifikuoto kryžminio patikrinimo metodo parametro įvertį apibrėžkime:

$$\alpha^* = \arg\min_{\alpha} \widehat{\varepsilon}_A(\alpha) = \arg\min_{\alpha} \Big(\int_{-\infty}^{+\infty} \frac{\widehat{f}_\alpha^2(x)}{g(x)} dx - \frac{2}{n} \sum_{t=1}^{n} \frac{\widehat{f}_\alpha(X_t|t)}{g(X_t)} \Big). \tag{9}$$

Čia $g(x)$ – koks nors pasiskirstymo tankio įvertis, o $\widehat{f}_\alpha(x|t)$ pasiskirstymo tankio įvertis taške x, apskaičiuotas išmetus stebėjimą X_t. Tankio įverčio parametro α (žr. (7)) parinkimas, panaudojant šią kryžminio patikrinimo metodo modifikaciją, buvo tiriamas imitacinio modeliavimo būdu.

Branduolio pločio parinkimas taikant kryžminį patikrinimą

Ankstesniame skyriuje buvo aptartas kryžminio patikrinimo metodas skirtas parametrui parinkti. Tačiau šį metodą galima taikyti ir branduolio pločiui $h(x)$ parinkti. Tokia idėja yra aprašyta straipsnyje [1]. Siūlomas toks algoritmas:

1) fiksuojamas x ašies tinklelis $x_1, \ldots, x_p$;
2) funkcija $h(x)$ apibrėžiama kaip splainas einantis per taškus $(x_1, h_1), \ldots, (x_p, h_p)$, t.y. $h(x) = h(x, h_1, \ldots, h_p)$;
3) $h_1, \ldots, h_p$ parenkami taip, kad minimizuotų pasirinktą nuostolių funkciją.

Šiame algoritme $h_1, \ldots, h_p$ atlieka daugiamačio parametro vaidmenį. Pažymėję $\alpha = (h_1, \ldots, h_p)$, galime taikyti kryžminio patikrinimo metodą, nusakytą formule (9), pasiskirstymo tankio įverčio branduolio pločiui rasti.

Eksperimentinis tyrimas

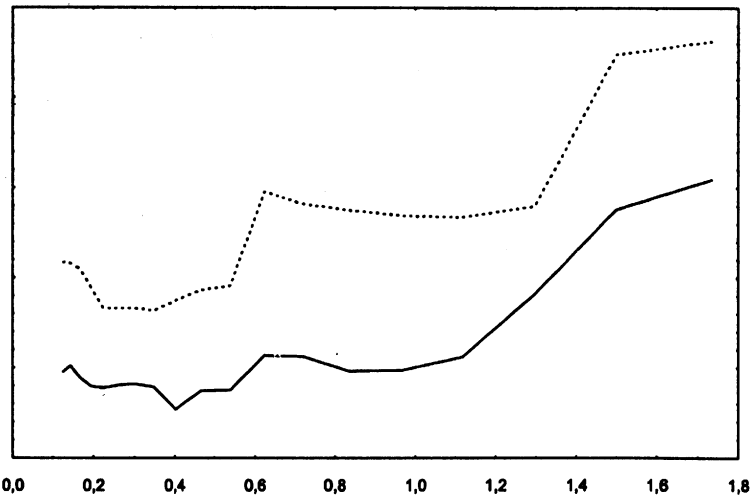
Ankstesniuose skyriuose aprašyti įverčiai buvo tirti Monte-Karlo būdu. Toks tyrimo metodas sudarė galimybes išmatuoti tikrąsias pasiskirstymo tankių įverčių paklaidas, nes generuojamų dydžių pasiskirstymo tankis buvo žinomas. Tyrimui buvo naudojami Gauso skirstinių mišiniai. Ypač akcentuojami rezultatai gauti naudojant mišinius, kurių komponenčių glodumas yra skirtingas, nes būtent šiuo atveju yra tikslinga naudoti adaptyvius

branduolinius pasiskirstymo tankio įverčius, t.y. įverčius, kurie gaunami naudojant skirtingą branduolio plotį h skirtingoms argumento x reikšmėms. Tiriamų įverčių tikslumas buvo matuojamas erdvių L_1 ir L_2 metrikose.

Parameto įvertinimo rezultatai

Šiame poskyryje aptariami rezultatai gauti bandant parametrą α (žr. formulę (7)) parinkti modifikuotu kryžminio patikrinimo būdu. Gautų įverčių paklaidos lyginamos su neblogus rezultatus duodančia fiksuota parametro reikšme $\alpha = 1$.

Nors [2] yra siūloma naudoti $\Delta(h) \geqslant h$, o iš asimptotinių tankio įverčio savybių seka, kad $\alpha \geqslant 1$ (žr. formulę (7)), tačiau modeliavimo rezultatai rodo, kad abi šios sąlygos nebūtinai turi būti tenkinamos. Beveik visais tirtais atvejais tankio įvertis buvo netikslus, kai $\alpha \geqslant 1.4$ ir paklaidos greitai augdavo didėjant α reikšmei. Tuo tarpu gana dažnai geriausi rezultatai buvo gaunami naudojant α reikšmes mažesnes už 1, o kartais paklaidas pavykdavo minimizuoti, net prie tokių mažų reikšmių, kaip pvz., $\alpha = 0.2$. Vienas iš tokių atvejų pateiktas grafike (1 pav.).



1 pav. Pasiskirstymo tankio įverčio paklaidos erdvės L_1 metrikos prasme (ištisinė linija) ir nepriklausančios nuo mastelio paklaidos ε_A įverčio (punktyrinė linija) priklausomybė nuo parametro α.

Apibendrinant galima teigti, kad:

– 80% atvejų α parinkimas modifikuotu kryžminio patikrinimo metodu sumažino paklaidas vidutiniškai 10%, kartais iki 30%;

– 20% atvejų papildoma α parinkimo procedūra padidino paklaidas 20-25%, kartais iki 60%.

Matome, kad parametro parinkimo procedūra nėra stabili, kaip neretai atsitinka su sudėtingomis statistinėmis procedūromis.

Pateiktame grafike matome vieną minimizuojamos funkcijos $\widehat{\varepsilon}_A$ pavyzdį ir tikrąsias L_1 paklaidų reikšmes. Grafikai yra pakankamai panašūs. Tai rodo, kad šiuo atveju paklai-

dos ε_A įvertis neblogai atspindi paklaidą L_1 metrikos prasme. Minimizuojama funkcija turi keletą lokalių minimumų, kas sukelia papildomų sunkumų ieškant minimumo taško.

Buvo pastebėta, kad įvertį galima padaryti šiek tiek stabilesniu, jeigu ieškosime ne globalaus minimumo, o didžiausio lokalaus minimumo srityje $\alpha \leqslant 1.5$, tačiau šiuo atveju šiek tiek nukenčia įverčio kokybė. Pastebėsime, kad [2] autorius, taikydamas mažiausių kvadratų kryžminio patikrinimo metodą, taip pat siūlo naudoti didžiausią lokalų minimumą. Be to, modeliavimo tyrimai parodė, kad dažnai didžiausias lokalaus minimumo taškas yra arti reikšmės $\alpha = 1$. Vienas tokių atvejų yra pavaizduotas grafike. Kadangi naudojant reikšmę $\alpha = 1$ gaunami pakankamai stabilūs rezultatai, tai natūralu, kad ir didžiausio lokalaus minimumo ieškojimas, kai $\alpha \leqslant 1.5$, padarys α parinkimo procedūrą stabilesne.

Praktiškai bandant panaudoti šį metodą reikia atsižvelgti ir į metodo skaičiavimo laiką, kuris išauga. Ypač jeigu bandoma minimizuoti pagal didelį α taškų kiekį.

Branduolio pločio parinkimo rezultatai

Bandymas parinkti branduolio plotį $h(x)$ tiesiogiai taikant modifikuotą kryžminio patikrinimo metodą nedavė gerų rezultatų. Gauto pasiskirstymo tankio įverčio paklaidos buvo apie 30% didesnės už tiesioginio pakeitimo įvertį apibrėžtą (1), (6). Toks branduolio pločio parinkimo metodas duodavo aiškiai per mažas funkcijos $h(x)$. Šios reikšmės labiau tinkamos naudoti, kaip apatinis apribojimas $h(x)$ funkcijų reikšmėms, negu kaip optimali $h(x)$ reikšmė. Bandymas minimizuojant ieškoti ne globalaus, o didžiausio lokalaus minimumo taško retai duodavo geresnius rezultatus, nes minimizuojant daugiamatį parametrą $h_1, \ldots, h_p$ buvo naudojama ciklinė procedūra, kuri kiekviename žingsnyje, dėl ir taip didžiulės skaičiavimų apimties, minimizuodavo tik pagal keletą reikšmių, o tokiu atveju sunku rasti didžiausią lokalų minimumą.

Išvados

Modifikuotas kryžminio patikrinimo metodas duoda blogesnius rezultatus negu tiesioginio pakeitimo metodas, bandant jį tiesiogiai naudoti vertinant branduolio plotį $h(x)$.

Modifikuotas kryžminio patikrinimo metodas yra tinkamas branduolio pločio parametrui α vertinti, nes jo naudojimas šiek tiek pagerina pasiskirstymo tankio įverčio kokybę. Siūlome taikyti šį metodą, jeigu įverčio skaičiavimo laikas nėra svarbus.

Branduolinio pasiskirstymo tankio įverčio paklaidos metrikoje L_1, prie $\alpha \leqslant 0.5$, reikalauja papildomo tyrimo, nes gauti modeliavimo rezultatai šiek tiek prieštarauja teorinėms išvadoms, paremtomis asimptotine analize.

Literatūra

[1] J. Fan, P. Hall, M.A. Martin and P. Patil, On local smoothing of nonparametric curve estimator, *Journal of the American Statistical Association*, **91**(433), 258–266 (1996).

[2] M.C. Jones, J.S. Maron and S.J. Sheather, A brief survey of bandwidth selection for density estimation, *Journal of the American Statistical Association*, **91**(433), 401–407, March (1996).
[3] R. Rudzkis and M. Kavaliauskas, On local bandwidth selection for density estimation, *Informatica*, **9**(4), 479–490 (1998).
[4] B.W. Silverman, *Density Estimation for Statistics Data Analysis*, Chapman and Hall, London (1986).

Kernel distribution density estimation based on cross-validation

M. Kavaliauskas

The kernel density estimation procedure is proposed. Parameter selection method based on cross-validation technique is analyzed. The results of investigation by simulation means are discussed.