LMD konf. darbai, 198–203
1997, Vilnius

Adaptyvus glodinimo pločio parinkimas pasiskirstymo tankio statistiniame įvertyje

M. Kavaliauskas (KTU, VDU)

Paskutiniu metu, vystantis kompiuterinei technikai, suintensyvėjo sudėtingesnių statistinės analizės metodų plėtojimas. Statistinės analizės paketais pradeda naudotis ne tik matematikai, bet ir kitų sričių specialistai: medikai, sociologai, ekonomistai, etc., todėl atsiranda automatizuotų procedūrų poreikis. Statistinės procedūros dirbančios be žmogaus pagalbos (įvairių parametrų nustatymo ir pan.) yra ypač reikalingos, jei jos yra sudėtingesnio statistinės analizės paketo sudėtine dalimi. Šis darbas yra bandymas gauti pakankamai tikslų ir automatizuotą neparametrinį pasiskirstymo tankio įvertį, bei ištirti jo savybes Monte-Karlo būdu.

Nagrinėsime branduolinį pasiskirstymo tankio įvertį:

$$\hat{f}(x) = \hat{f}_h(x) = \frac{1}{n}\sum_{j=1}^{n} W(x - X_j), \tag{1}$$

čia $W(x) = \frac{1}{h}\psi(\frac{x}{h})$, h – branduolio plotis, $\psi(x)$ – branduolio funkcija, tenkinanti sąlygą:

$$\int_{-\infty}^{\infty} \psi(x)\,dx = 1. \tag{2}$$

Šiame darbe modeliavimas buvo atliekamas naudojant Jepaničnikovo branduolį:

$$\psi(x) \begin{cases} \frac{3}{4}(1-x^2), & |x| \leqslant 1, \\ 0, & |x| > 1. \end{cases}$$

Vertinant pasiskirstymo tankio reikšmę taške x, natūralu parinkti glodinimo plotį atsižvelgiant ne tik į imties tūrį n, bet ir į tankio savybes taško x aplinkoje. Tuo būdu $h = h(x, n)$. Remsimės asimptotinėmis įverčių savybėmis, kai $n \to \infty$, $h = h(n) \to 0$. Nuostolių funkcija laikysime $E(f(x) - \hat{f}(x))^2$. Jei apsiribosime įverčiais su neneigiamais lyginiais branduoliais, kaip paprastai daroma praktikoje, tai įverčio poslinkis dukart tolygiai diferencijuojamo tankio atveju lygus:

$$b_h(x) = c_1 \frac{f''(x)h^2}{2} + o(h^2), \tag{4}$$

kur

$$c_1 = \int y^2 \psi(y)\,dy. \tag{5}$$

Taip pat gerai žinoma asimptotinė įverčio dispersijos išraiška:

$$\sigma_h^2 = \frac{c_2 f(x) - h f^2(x)}{nh} + O\left(\frac{h^2}{n}\right) = \frac{c_2 f(x)}{nh} + O\left(\frac{1}{n}\right). \tag{6}$$

Čia

$$c_2 = \int \psi^2(y)\, dy. \tag{7}$$

Iš šių formulių seka, kad jei taške x $f''(x) \neq 0$, tai asimptotiškai optimalus branduolio plotis $h(x, n)$ yra lygus:

$$h(x, n) = \left(\frac{c_2 f(x)}{c_1^2 (f''(x))^2 n}\right)^{1/5}. \tag{8}$$

Betarpiškai pasinaudoti (8) negalima, nes $f(x)$ ir jos išvestinės nėra žinomos. Be to reikia atsižvelgti į galimą situaciją, kai $f''(x)$ artima nuliui. Todėl konstruosime įvertį tokiu būdu:

$$h = h(x, n)\colon\ \hat{b}_h^2(x) + \hat{\sigma}_h^2(x) \longrightarrow min. \tag{9}$$

Pastaba. Žymėjimų paprastumo sumetimais, tiek optimalų branduolio plotį, tiek jo įvertį žymėsime h tikėdamiesi, kad tai neįves painiavos.

Poslinkio ir dispersijos įverčius nusakykime lygybėmis:

$$\hat{b}_n(x) = \frac{c_1 \lambda(x, h) h^2}{2}, \tag{10}$$

kur $\lambda(x, h) = \hat{f}_h''(x)$,

$$\hat{\sigma}_h^2(x) = \frac{c_2 \hat{f}_h(x)}{nh}. \tag{11}$$

Į poslinkio įvertį įeinantis pasiskirstymo tankio antrosios išvestinės įvertis $\lambda(x, h)$ randamas tokiu būdu. Pažymėkime:

$$Q(y) = F(x + y) - F(x - y) - \frac{y}{h}\big(F(x + h) - F(x - h)\big). \tag{12}$$

Skleisdami pasiskirstymo funkciją Teiloro eilute taško x aplinkoje gauname:

$$Q(y) = \frac{f''(x)(y^3 - yh^2)}{3} + o(yh^3). \tag{13}$$

Todėl $\lambda(x, h)$ apibrėžkime taip:

$$\lambda = \lambda(x, h)\colon \max_{0 \leqslant y \leqslant h} \left|\hat{Q}(y) - \frac{\lambda(y^3 - yh^2)}{3}\right| \longrightarrow \min, \tag{14}$$

čia $\hat{Q}(y)$ apibrėžiamas (12) lygybe, keičiant $F(x)$ į empirinę pasiskirstymo funkciją $\hat{F}(x)$.

Tokiu būdu gauto pasiskirstymo tankio įverčio testavimas parodė, kad jį reikia pataisyti. Norint įvertinti pasiskirstymo tankio antrąją išvestinę, reikia imti platesnę taško x aplinką, negu vertinant pačią tankio funkciją. Todėl įvertis pataisomas įvedant papildomą koeficientą $c > 1$:

$$\hat{b}_h(x) = \frac{c_1 \lambda(x, ch)h^2}{2}. \tag{15}$$

Pasiskirstymo tankio įvertis apibrėžiamas (1), (9), (11), (15) lygybėmis buvo tiriamas Monte-Karlo būdu. Tyrimui naudojami Gauso skirstinių mišiniai, nes būtent mišinių atveju, kai tankio glodumo savybės stipriai skiriasi įvairioms x reikšmėms, efektyvu naudoti branduolio plotį priklausomą nuo x. Be to, neparametrinis Gauso skirstinių tankio įvertis reikalingas netgi norint gauti geros kokybės parametrinius įverčius (žr. [2]).

Aprašytas įvertis buvo lygintas su:

I. Asimptotiškai optimaliu pseudoįverčiu, kurio parametras h gaunamas analogiškai kaip ir (8) formulėje, bet jis nepriklauso nuo x:

$$h = \left(\frac{c_2}{c_1^2 n \int (f''(x))^2 dx} \right)^{1/5}. \tag{16}$$

II. Statistiniu įverčiu, kurio h nepriklauso nuo x, bet parenkamas pagal imtį, minimizuojant nuostolių $E\|\hat{f} - f\|_2^2$ įvertį.

III. Adaptyviu pseudoįverčiu, kurio h optimaliai parenkamas pagal $f(x)$ lokalias savybes taško x aplinkoje.

Trumpai aprašysime **II** ir **III** statistikų konstrukcijas.

II *įvertis.* Šis įvertis aprašytas [1].

Apskaičiuokime įverčio vidutinę kvadratinę paklaidą:

$$\|\hat{f}(x) - f(x)\|_2^2 = \int \hat{f}^2(x)\,dx - 2\int \hat{f}(x) f(x)\,dx + \int f^2(x)\,dx. \tag{17}$$

Kadangi paskutinis narys nepriklauso nuo įverčio, tai apibrėžkime h tokiu būdu:

$$h: \int \hat{f}^2(x)\,dx - 2\int \hat{f}(x) f(x)\,dx \longrightarrow \min, \tag{18}$$

$$\begin{aligned} \int \hat{f}^2(x)\,dx &= \int \left(\frac{1}{nh} \sum_{i=1}^{n} \psi\left(\frac{x - X_i}{h} \right) \right)^2 dx = \frac{1}{n^2 h} \sum_{i,j=1}^{n} K\left(\frac{X_i - X_j}{h} \right) \\ &= \frac{2}{n^2 h} \sum_{i=1}^{n} \sum_{j=i+1}^{n} K\left(\frac{X_i - X_j}{h} \right) + \frac{K(0)}{nh}, \end{aligned} \tag{19}$$

čia $K(x)$ yra branduolio $\psi(x)$ sąsuka su pačiu savimi.

$$\int \hat{f}(x)f(x)\,dx \approx \int \hat{f}(x)\,dF_n(x) \approx \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1,j\neq i}\psi\left(\frac{X_i - X_j}{h}\right). \tag{20}$$

Pastebėsime, kad dėmenų su $j = i$ išmetimas sumažina poslinkį.

Taigi iš (19) ir (20) gauname:

$$h\colon \frac{2}{n^2h}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(K\left(\frac{X_i - X_j}{h}\right) - 2\psi\left(\frac{X_i - X_j}{h}\right)\right) + \frac{K(0)}{nh} \longrightarrow \min. \tag{21}$$

III *pseudoįvertis.* Natūralu pseudoįvertį apibrėžti naudojant tikrąjį poslinkį ir dispersiją

$$h = h(x, n)\colon\ b_h^2(x) + \sigma_h^2(x) \longrightarrow \min, \tag{22}$$

tačiau taip apibrėžtas pseudoįvertis priklauso ne tik nuo lokalių tankio savybių taško x aplinkoje, bet ir nuo tankio reikšmių daug nutolusių nuo x. Todėl pseudoįvertis modifikuotas tokiu būdu. Kadangi:

$$\begin{aligned} b_h(x) &= \int_{-1}^{1}\psi(y)\,dy\int_{0}^{hy}f'(x+z)\,dz = \int_{0}^{1}\psi(y)\,dy\int_{0}^{hy}dz\int_{-z}^{z}f''(x+u)\,du \\ &= \int_{-h}^{h}f''(x+u)\left(\int_{|u|}^{h}\big(z-|u|\big)W(z)\,dz\right)du \overset{\text{def}}{=} \int_{-h}^{h}f''(x+u)G_h(u)\,du, \end{aligned} \tag{23}$$

tai pseudoįvertį $h(x)$ apibrėžkime:

$$h(x)\colon\ \big(b_h^*(x)\big)^2 + \sigma_h^2(x) \longrightarrow \min, \tag{24}$$

kur

$$\big|b_h^*(x)\big| = \int_{-h}^{h}\big|f''(x+y)\big|G_h(y)\,dy. \tag{25}$$

Tyrimo schema. Tyrimas atliktas Monte-Karlo metodu. Buvo generuojami vienmačių Gauso skirstinių mišinių su žinomais parametrais imtys. Jų tankis yra:

$$f(x) = \sum_{i=1}^{k} p_i\varphi(x, m_i, \sigma_i). \tag{26}$$

Gautoms imtims skaičiuojami aukščiau minėti pasiskirstymo tankio funkcijos įverčiai ir pseudoįverčiai. Šie įverčiai lyginami su teorine pasiskirstymo tankio funkcija.

Įverčio tikslumas vertintas L_1, L_2 ir L_∞ metrikose:

$$\varepsilon_1 = \int |\hat{f}(x) - f(x)|\,dx,$$
$$\varepsilon_2 = \int (\hat{f}(x) - f(x))^2 dx, \tag{27}$$
$$\varepsilon_\infty = \max_x |\hat{f}(x) - f(x)|.$$

Eksperimentų rezultatų lentelė.

Gauso mišinių sudėtis:

1) $n = 500$; $p_1 = 0.7$, $p_2 = 0.3$; $m_1 = 0$, $m_2 = 3$; $\sigma_1 = 0.4$, $\sigma_2 = 1$;
2) $n = 2000$; $p_1 = 0.7$, $p_2 = 0.3$; $m_1 = 0$, $m_2 = 3$; $\sigma_1 = 0.4$, $\sigma_2 = 1$;
3) $n = 2000$; $p_1 = 0.7$, $p_2 = 0.3$; $m_1 = 0$, $m_2 = 6$; $\sigma_1 = 0.4$, $\sigma_2 = 2$.

Gauso mišiniai	Metrika	Siūlomas įvertis	Asimptotinis pseudoįvertis (I)	Fiksuoto pločio įvertis (II)	Optimalus pseudoįvertis (III)
	L_1	0.130268	0.132292	0.093799	0.094259
1	L_2	0.065963	0.075425	0.048306	0.049532
	L_∞	0.095950	0.086234	0.074546	0.051814
	L_1	0.065705	0.077963	0.059230	0.055269
2	L_2	0.033966	0.041728	0.030135	0.029183
	L_∞	0.037409	0.084537	0.037334	0.035365
	L_1	0.064688	0.100593	0.081787	0.056844
3	L_2	0.032094	0.042985	0.032597	0.028600
	L_∞	0.038153	0.081336	0.040551	0.036456

Preliminarūs tyrimo rezultatai rodo, kad siūlomas įvertis duoda mažesnes paklaidas negu asimptotinis pseudoįvertis **I**, tačiau nevisuomet vertina geriau negu fiksuoto pločio įvertis **II**. Iš pateiktos lentelės matome, kad esant mažam imties kiekiui, geresnis yra įvertis **II**, didesnės imties atveju (antras mišinys) įverčio ir fiksuoto pločio įverčio **II** kokybės yra panašios, tačiau pakeitus mišinio formą taip, kad glodumas įvairiose tankio dalyse dar labiau skirtųsi (trečias mišinys), geresnius rezultatus duoda įvertis apibrėžiamas (1), (9), (11), (15).

Išvados.

Įverčių palyginamoji analizė parodė, kad įvertį galima tobulinti. Vienas iš neišspręstų klausimų yra konstantos c įeinančios į (15) išraišką parinkimas. Bandymai rodo, kad c turėtų būti x funkcija, nes esant mažiems h nepakanka dvigubo pločio antrajai pasiskirstymo tankio išvestiniai įvertinti. Taip pat pastebėtas minimumo apibrėžiamo (9) nestabilumas. Esant mažiems h funkcija $\hat{b}_h^2(x) + \hat{\sigma}_h^2(x)$ yra „šokinėjanti", o pernelyg dideliems h galima rasti kitų šios funkcijos minimumų, todėl minimumo reikia ieškoti optimaliai parinktame intervale $[h_{min}; h_{max}]$. Manoma, kad įvertis geriau vertins tankį, jei h įvertį lokaliai suglodinsime.

Įverčio savybių išsamesnio tyrimo modeliavimo būdu rezultatai bus pateikti kitame straipsnyje.

Dėkoju prof. R.Rudzkiui už pagalbą rašant šį straipsnį.

LITERATŪRA

[1] B. W. Silverman, *Density Estimation for Statistics Data Analysis*, Chapman and Hall, 1986. ISBN 0-412-24620-1.

[2] M. Radavičius and R. Rudzkis, Statistical estimation of a mixture of Gaussian distribution, *Acta Applicandae Mathematicae*, **38** (1995), 37–54.

An adaptive choice of smoothing width in statistical estimation of distribution density

M. Kavaliauskas

This article illustrates the adaptive kernel probability density function estimation method, shows encounted problems, gives solutions and suggestions. The simulation results are discussed.