Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
54 t., 2013, 6–11

ISSN 0132-2818
www.mii.lt/LMR/

Neparametrinių tankio įvertinių lyginamoji analizė Monte Karlo metodu

Indrė Drulytė, Tomas Ruzgas

Kauno technologijos universitetas, Fundamentaliųjų mokslų fakultetas
K. Donelaičio g. 73, LT-44029 Kaunas

E. paštas: drulyte.indre@inbox.lt, tomas.ruzgas@ktu.lt

Santrauka. Darbas skirtas pasiskirstymo tankio neparametriniam statistiniam įvertinimui. Monte Karlo metodu tiriamas branduolio funkcijos poveikis daugiamodalinio tankio įvertinimo tikslumui. Parodoma, kad pasiūlyta nauja branduolio funkcija efektyvi asimetriškiems sunkių uodegų skirstiniams.

Raktiniai žodžiai: neparametrinis vertinimas, tankio vertinimas, branduolio funkcija.

Įvadas

Neparametrinis vertinimas yra vienas iš statistinių metodų, kuris leidžia lengvai apdoroti duomenis, netaikant jokių parametrinių prielaidų funkcijai. Šis metodas plačiai išnagrinėtas Silverman veikale [14].

Tegul $X(1), \ldots, X(n)$ yra stebimi nepriklausomi atsitiktiniai dydžiai su nežinomu pasiskirstymo tankiu $f(x)$. Jei funkcija $f(x)$ nėra parametrizuota, jai įvertinti taikomi neparametriniai metodai. Nors šiuolaikinėje duomenų analizėje žinoma gausybė pasiskirstymo tankio įvertinimo metodų, praktikoje nėra lengva parinkti efektyvią įvertinimo procedūrą, jei duomenų pasiskirstymo tankis daugiamodalinis, o imties tūris nėra didelis. Minėtas atvejis dažnai pastebimas praktiniuose tyrimuose ir remiasi Gauso skirstinių mišinio modeliu, kurio taikymai yra gana populiarūs ir sutinkami įvairiose mokslo kryptyse, nagrinėjant aktualias medicinos, gamtos mokslų, sociologijos problemas [1, 7, 5, 18]. Tarp neparametrinių tankio vertinimo metodų ypač plačiai paplitę branduoliniai įvertiniai. Šio įvertinio optimalaus glodinimo parametro parinkimas priklauso nuo nežinomo pasiskirstymo tankio $f(x)$ glodumo taško $X(t)$ aplinkoje, o tai nustatyti yra beveik neįmanoma jei imtis nėra didelė. Jei glodinimo parametras yra pastovus, tada mažėja vertinimo tikslumas, ypač daugiamodaliniu atveju.

Šio darbo tikslas – Monte Karlo metodu atlikti branduolinio tankio įvertinio lyginamąją analizę naudojant skirtingas branduolio funkcijas.

Šio straipsnio struktūra: 1 dalyje aprašyti neparametrinių tankių vertinimo metodai; 2 dalyje pateikiamos naudojamos branduolio funkcijos; 3 dalyje aprašytos pradinės sąlygos bei matas, skirtas įvertinti tankių tikslumą; 4 dalyje pateikiami eksperimentinio tyrimo rezultatai; 5 dalyje pateiktos tiriamojo darbo išvados.

1 Neparametriniai tankių vertinimo metodai

Praktikoje yra sutinkama daug neparametrinių tankio vertinimo metodų, tokių, kaip histograma, neparametrinė arba pusiau parametrinė regresija taip pat duomenų apgaubimo analizė (angl. data envelopment analysis). Pavyzdžiui, histograma – vienas paprasčiausių ir seniausių tankio įvertinių. Kaip žinoma, duomenys histogramos pavidalu (be grafinio vaizdavimo) pirmą kartą buvo pateikti 1661 metais, nustatant mirtingumo tikimybes skirtingose amžiaus grupėse [12]. Nuo 1891 metų histogramos terminą pirmasis pradėjo vartoti Karl Pearson [8]. Šis įvertinys, naudojamas kaip duomenų pateikimo priemonė, yra lengvai suprantamas ir patogus. Tačiau, norint atlikti papildomus skaičiavimus, kuriuose reikia žinoti tankio įverčio išvestines, susiduriama su didele problema: kadangi funkcija nėra tolydi, todėl laiptuotoje diagramoje neišvengiamai atsiranda trūkio taškai, kuriuose funkcija nediferencijuojama. Ši problema lengvai išsprendžiama vietoj histogramos naudojant branduolinį tankio įvertinį. Ši funkcija gaunama sumuojant atskirų stebinių glodžiuosius iškilimus – branduolius, tokiu būdu gaunant tolydžią funkciją. Tuomet vienmatis fiksuoto pločio branduolinis tankio įvertinys FK su branduolio funkcija K ir fiksuotu (globaliu) branduolio pločio parametru h, kuriuo galima įvertinti vienamačių duomenų $X \in R$ tankį $f(x)$, apibrėžiamas taip:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{1}$$

Šiuo atveju branduolio funkcija $K(u)$ turi tenkinti savybes:

$$K(u) \geqslant 0, \qquad \int K(u)\,du = 1,$$

$$K(-u) = K(u) \text{ – visoms } u \text{ reikšmėms}, \qquad \int u^2 K(u)\,du < \infty.$$

Atliekant visapusę modeliavimo analizę lyginamajame tyrime, kaip atskaitinius tankius naudinga išnagrinėti Gauso skirstinių mišinius, kuriuos 1992 m. savo darbe aprašė Marron ir Wand [8]. Be to, šiuos mišinius savo darbuose naudoja daugelis tyrėjų, tokių kaip Chen et al., 2001 [2], Torres-Carrasquillo et al., 2004 [15], Wasito et al., 2007 [17] bei daugelis kitų [9, 3].

Mokslininkai Fryer (1976 m.) ir Deheuvels (1977 m.) neparametrinių tankių vertinimo tikslumui išreikšti, pirmą kartą įrodė, kad vidutinė integruojama kvadratinė paklaida (MISE) gali būti apskaičiuojama tiksliai, kuomet ir tankio, ir branduolio funkcijos priklauso Gauso kreivių šeimai. Iš to matyti, kad šis vertinimo tikslumo matas gali būti apskaičiuojamas remiantis funkcijų sąsūkomis, kurių viena imama kaip fiksuotas filtras, dar vadinamas branduoliu.

Vidutinė integruojama kvadratinė paklaida randama iš:

$$\mathsf{MISE}(h) = E\int \left(\hat{f}_h(x) - f(x)\right)^2 dx,$$

čia $f(x)$ yra nežinomas tankis, o $\hat{f}_h(x)$ – tankio įvertis.

Dėl šios mišinių grupės „lankstumo“ galima atlikti nesudėtingą, tačiau tikslią analizę, nagrinėjant įvairių formų tankio funkcijas. Pagrindinis tokios analizės tikslas yra

gauti kuo tikslesnius modeliavimo rezultatus ir palengvinti skaičiavimus lengvai atliekamais metodais.

2 Branduolys ir jo funkcijos

Aukščiau aprašytame skyriuje aptarėme neparametrinių tankių vertinimo metodų privalumus. Šiame darbe neparametrinį tankių vertinimui pasirinktas branduolinis tankio įvertinys. Trumpai aprašysime jį.

Tegul $X(1), \ldots, X(n)$ yra stebimi nepriklausomi atsitiktiniai dydžiai su nežinomu pasiskirstymo tankiu $f(x)$, tuomet funkcijos $f(x)$ vienmatis fiksuoto pločio branduolinis tankio įvertinys FK aprašomas formule (1). Maža h reikšmė lemia didelę įvertinio dispersiją ir mažas paklaidas. Tačiau, parinkus didelę h gauname nedidelį standartinį išaugusių paklaidų nuokrypį. Optimalus Gauso branduolio funkcijos plotis h, nustatomas mažinant vidutinę integralinę kvadratinę paklaidą (MISE) ir apskaičiuojamas remiantis formule:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1,06\hat{\sigma}n^{-\frac{1}{5}},$$

čia $\hat{\sigma}$ – imties standartinis nuokrypis [11, 10, 13, 12, 16].

Taikant branduolinį tankio vertinimo metodą, buvo atrinktos keturios branduolio funkcijos pagal geriausius rezultatus:

1. Gauso (Gaussian) branduolio funkcija – viena populiariausių branduolio funkcijų, naudojamų neparametriniuose tankių vertinimo metoduose:

$$K(u) = \varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right).$$

2. Epanechnikovo branduolio funkcija yra optimali dispersijos minimumo atveju [4] ir aprašoma:

$$K(u) = \frac{3}{4}\left(1-u^2\right)\mathbb{1}_{\{|u|\leqslant 1\}}.$$

3. Trisvorė (Triweight) branduolio funkcija:

$$K(u) = \frac{35}{32}\left(1-u^2\right)^3\mathbb{1}_{\{|u|\leqslant 1\}}.$$

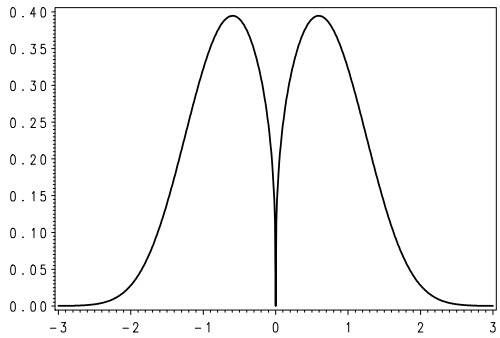
Visos aukščiau išvardintos branduolio funkcijos priklauso specialiai polinomų šeimos klasei ir yra išreiškiamos formule:

$$k_s(u) = \frac{(2s+1)!!}{2^{s+1}s!}(1-u^2)^s\mathbb{1}_{\{|u|\leqslant 1\}}.$$

Čia dvigubas faktorialas yra $(2s+1)!! = (2s+1)(2s-1)\ldots 5\cdot 3\cdot 1$.

Gauso branduolys yra gaunamas po dimensijų pakeitimo, kuomet skaičiuojama riba, kai $s \to \infty$.

Branduolių, kurių išraiškose yra aukštesni s laipsniai, funkcijos pavidalas yra glotnesnis. Tokiu atveju gaunami įverčiai, kurie taip pat yra glotnesni ir turintys daugiau išvestinių. Ieškant įverčių, naudojant Gauso branduolius, egzistuoja visų laipsnių išvestines [6].



1 pav. Siūlomos branduolio funkcijos forma, kai $\alpha = 0{,}75$.

4. Siūloma branduolio funkcija (tarpinė tarp Koši ir Gauso branduolio funkcijų):

$$K(u) = \varphi\big(|u|^{\frac{1}{\alpha}}\big)\frac{1}{\alpha}\big(|u|^{\frac{1}{\alpha}}\big)^{1-\alpha}.$$

Šios branduolio funkcijos pavidalas priklauso nuo pasirenkamo parametro α. Atlikto tyrimo parametro α reikšmės pasirinktos atsitiktinai ir lygios 0,25, 0,5 ir 0,75 (žiūrėti 1 pav.).

Siūlomos branduolio funkcijos forma yra netipinė, lyginant su aukščiau pateiktomis branduolio funkcijų išraiškomis. Taip pat ji mažiau jautri vertinamo taško labai artimai aplinkai, o tai atskirais pasiskirstymo tankių atvejais leidžia išvengti per didelio įverčio glodinimo.

3 Pradinės sąlygos ir vertinimo tikslumas

Tyrime buvo taikomas plačiai naudojamas Monte Karlo metodas. Šio metodo pagrindinis pritaikomumo principas statistinėje analizėje yra tai, jog duomenys parenkami atsitiktinai. Dėl to modeliavimui naudotos mažos ir vidutinio didumo imtys (16, 32, 64, 128, 256, 512, 1024, 2048). Kiekvienu atveju generuota po 30 nepriklausomų imčių.

Tankių vertinimo tikslumui išreikšti buvo naudojama vidutinė procentinė absoliutinė paklaida MAPE, kuri apskaičiuojama remiantis formule:

$$\mathsf{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f(X_i) - \hat{f}_h(X_i)}{f(X_i)}\right| \cong \int \big|f(x) - \hat{f}_h(x)\big|\,dx$$

čia $f(x)$ – tankio tikroji reiškmė, o $\hat{f}_h(x)$ – tankio įvertis.

4 Eksperimento rezultatai

Neparametrinis tankių vertinimas buvo atliktas naudojant branduolinius įvertinius su atitinkamomis branduolio funkcijomis, aptartomis antrajame skyriuje. Remiantis lengvai pritaikomu Monte Karlo metodu tankių generavimui, buvo panaudoti Marron ir Wand [8] pasiūlyti 15 Gauso skirstinių mišinių.

Šiuo atveju pagrindinis tyrimo tikslas buvo surasti geriausias branduolio funkcijas su kuriomis būtų tiksliausia įvertinti pasiskirstymo tankius. Dėl to, gautų rezultatų interpretacija atlikta išryškinant branduolio funkciją ir imties tūrį, kuriam esant gaunamas mažiausias tankių vertinimo poslinkis.

Naujai pasiūlyta branduolio funkcija geriausius rezultatus (12 iš 15 nagrinėtų skirstinių mišinių atvejų) parodė, esant mažoms imtims, t. y. kai imties tūris atitinkamai lygus 16, 32, 64, 128. Tuo tarpu trisvorė branduolio funkcija labiausiai tinkama turint vidutinio dydžio imtis, kurių tūriai yra 256, 512, 1024, 2048. Nagrinėjant gautus rezultatus su Epanechnikovo branduolio funkcija, akivaizdu, jog ši funkcija negali būti vienareikšmiškai naudojama tik atitinkamo dydžio imtims. Gauso branduolio funkciją yra aktualu nagrinėti, kai skaičiavimai atliekami su didesnėmis nei vidutinio dydžio imtimis. Šiuo atveju imties tūris turėtų būti didesnis nei 2048.

5 Išvados

Neparametrinio tankių vertinimo algoritmų lyginamasis tyrimas Monte Karlo metodu parodė, jog plačiai naudojamos branduolio funkcijos yra tinkamiausias sprendimas norint atlikti skaičiavimus su vidutinio dydžio bei didelėmis imtimis. Tuo tarpu, pasiūlyta nauja branduolio funkcija, esant mažoms imtims, daugeliu Gauso skirstinių mišinių atvejų, kai turimas asimetriškumas su sunkiomis uodegomis, ar esant sudėtingai modų struktūrai, buvo geresnė už kitas tirtas branduolio funkcijas.

Literatūra

[1] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *J. Clasif.*, **13**:195–212, 1996.

[2] T. Chen, Ch. Huang, E. Chang and J. Wang. Automatic accent identification using Gaussian mixture models. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition*, pp. 343–346, 2001.

[3] S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the IEEE Symposiumon Foundations of Computer Science*, pp. 633–644, New York, 1999.

[4] V.A. Epanechnikov. *Nonparametric Estimation of a Multivariate Probability Density.* 1969.

[5] H. Greenspan. *Constrained Gaussian Mixture Model Framework for Automatic Segmentation of MR Brain Images.* 2006.

[6] B.E. Hansen. *Lecture Notes on Nonparametrics.* 2009.

[7] Y. Huang. *A Gaussian Mixture Model Based Classification Scheme for Myoelectric Control of Powered Upper Limb Prostheses.* 2005.

[8] J.S. Marron and M.P. Wand. Exact mean integrated squared error. *Ann. Stat.*, **20**:712–736, 1992.

[9] R.J. Martis, C. Chakraborty and A.K. Ray. A two-stage mechanism for registration and classification of ECG using Gaussian mixture model. *Pattern Recognit. Lett.*, **42**:2979–2988, 2009.

[10] V.C. Raykar and R. Duraiswami. *Very Fast Optimal Bandwith Selection for Univariate Kernel Density Estimation.* Department of Computer Science, University of Maryland, Collegepark.

[11] V.C. Raykar and R. Duraiswami. Fast optimal bandwith selection for kernel density estimation. In *Proceedings of the Sixth SIAM International Conference on Data Mining*, pp. 524–528, Bethesda, 2006.

[12] D.W. Scott. On optimal and data-based histograms. *Biometrika*, **66**:605–610, 1979.

[13] S.J. Sheather and M.C. Jones. A reliable data-based bandwith selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B*, **53**:683–690, 1991.

[14] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Taylor & Francis, 1986.

[15] P.A. Torres-Carrasquillo and T.P. Gleason, and D.A. Reynolds. Dialect identification using Gaussian mixture models. In *Proc. Odyssey: The Speaker and Language Recognition Workshop in Toledo*, pp. 297–300, 2004.

[16] M.P. Wand and M.C. Jones. *Kernel Smoothing*. London, 1995.

[17] I. Wasito, S.Z.M. Hashim and S. Sukmaningrum. Iterative local gaussian clustering for expressed genes identification linked to malignancy of human colorectal carcinoma. *Bioinformation*, pp. 175–181, 2007.

[18] H. Zhang, C.L. Giles, H.C. Foley and J. Ye. Probabilistic community discovery using hierarchical latent Gaussian mixture model. In *Proceedings of the 22nd National Conference on Artificial Intelligence,* vol. 1, pp. 663–668, 2007.

SUMMARY

Research of comparative analysis of nonparametric density estimation by applying Monte Carlo method

I. Drulyte, T. Ruzgas

This paper presents nonparametric statistical estimation of distribution density. The Monte Carlo method is used to show the effects of kernel function for multimodal kernel density estimation. Here it is shown that the novel kernel function is effective for asymmetrical heavy tails distributions.

Keywords: non-parametric estimation, density evaluation, kernel function.