Liet. mat. rink. LMD darbai, **50**, 2009, 298–303

Pasiskirstymo tankio įvertinimas naudojant duomenų projektavimą

Mindaugas KAVALIAUSKAS
Kauno technologijos universitetas
Studentų g. 50, LT-51368 Kaunas
el. paštas: kavaliauskas.mindaugas@gmail.com

Santrauka. Straipsnyje nagrinėjama neparametrinio daugiamačio pasiskirstymo tankio įvertinimo problema. Tiriamas J.H. Friedman pasiūlytas pasiskirstymo tankio įvertinys pagrįstas tikslinio projektavimo algoritmu. Šis algoritmas taikytinas, kai duomenų dimensija yra didelė, o tankis turi daugiamodalinę struktūra. Pasiskirstymo tankio įvertinys pagrįstas tiksliniu projektavimo yra sudėtinis algoritmas, kuris naudoja kitas statistines procedūras: projektavimo indeksą, vienamačių pasiskirstymo tankio projekcijų įvertinį. J.H. Friedman siūlė naudoti projekcinį tankio įvertinį Ležandro polinomų bazėje vienamačių duomenų projekcijų tankiams vertinti, bei projektavimo indeksą pagrįstą tankio išraiška Ležandro polinomų bazėje. Straipsnio autorius bando keisti originaliai pasiūlytas procedūras, klasikinėmis statistinėmis procedūromis. Bandoma naudoti branduolinį pasiskirstymo tankio įvertinį duomenų projekcijų tankiams vertinti, bei projektavimo indeksą paremtą Komogorovo-Smirnovo statistika. Tiriami ir palyginami modifikuoto ir originalaus algoritmų tikslumai. Palyginamoji analizė atliekama kompiuterinio modeliavimo būdu.

Raktiniai žodžiai: tikslinis projektavimas, pasiskirstymo tankio funkcija, neparametrinis vertinimas, projektavimo indeksas.

1. Įvadas

Pasiskirstymo tankis – viena iš pagrindinių funkcijų apibūdinančių atsitiktinį dydį, todėl suprantama, kad jo įvertinimas išlieka svarbus tiek savaime (pvz., vizualiai pateikiant statistinius duomenis), tiek kaip sudėtinė kitų algoritmų dalis. Didėjant stebimo dydžio matavimui daugelis statistinių procedūrų susiduria su „daugiamatiškumo prakeikimo“ problema (angl. *curse of dimensionality*). Šią problemą lengva iliustruotu tokiu pavyzdžiu – jei turime 10-matę vienetinę sferą tolydžiai užpildytą atsitiktiniais taškais, tai tam, kad apimtume 5% šios vienetinės sferos taškų (pvz., konstruodami branduolinį tankio įvertinį), turime imti sferą kurios spindulys lygus $(0.05)^{1/10} = 0.74$. Todėl sukonstruoti branduolinis tankio įvertis negali gerai įvertinti tankio formos 10-matėje erdvėje, jei duomenų kiekis nėra milžiniškas [4]. Vienas iš metodų padedančių išvengti šios problemos yra tikslinis projektavimas (angl. *projection pursuit*). J.H. Friedman tikslinio projektavimo idėją pirmą kart pasiūlė [2]. Vėliau ši idėja buvo pritaikyta daugiamačiam pasiskirstymo tankiui vertinti. Detalų jos aprašymą rasime [1]. Nepaisant to, kad nuo metodo aprašymo praėjo daugiau nei du dešimtmečiai, atsiranda vis naujų darbų, kuriuose taikoma tikslinio projektavimo idėja. Šiame straipsnyje tirsime originalią tankio vertinimo procedūra pasiūlytą J.H. Friedman. Bandysime atlikti procedūros modifikacija, kompiuterinio modeliavimo būdu įvertinti jų įtaką metodo tikslumui.

2. Originalus algoritmas

Trumpai aprašysime originalų pasiskirstymo tankio įvertinį pagrįstą tiksliniu projektavimu. Tai yra iteracinis algoritmas pagrįstas vienamačių duomenų projekcijų kiek galima labiau besiskiriančių nuo Gauso skirstinio paieška ir duomenų transformavimu taip, kad šios projekcijos įgautų Gauso pasiskirstymą.

Tegu Z yra standartizuotas atsitiktinis dydis, t.y. dydis turintis nulinį vidurkį ir vienetinę kovariacinę matricą. Jei mūsų stebimas atsitiktinis dydis X netenkina šios savybės, tai Z gausime normuodami X. Pažymėkime $Z^{(0)} = Z$, tuomet $Z^{(k)}$, $k \geqslant 1$ yra gaunamas po šios procedūros. Tegu $g_k(u)$, $u \in \mathbb{R}$ yra vienamatės duomenų projekcijos $\tau' Z^{(k-1)}$ (kryptimi τ) pasiskirstymo tankis, o G_k tos pačios duomenų projekcijos pasiskirstymo funkcija. Tuomet

$$Z^{(k)} \overset{def}{=} Q_k(Z) = Z^{(k-1)} - \bigl(\tau' Z^{(k-1)}\bigr)\tau + \Phi^{-1}\bigl(G_k\bigl(\tau' Z^{(k-1)}\bigr)\bigr)\tau, \tag{1}$$

čia Φ – standartinė Gauso pasiskirstymo funkcija, o $\tau = \tau_k$ – projektavimo kryptis pasirinkta naudojant projektavimo indeksą.

Taigi atsitiktinis vektorius $Z^{(k)}$ yra gaunamas iš $Z^{(k-1)}$ taip, kad $Z^{(k)}$ projekcija kryptimi τ įgautų normalųjį pasiskirstymą, o projekcijos kitomis $d-1$ kryptimis, ortogonaliomis krypčiai τ, liktų nepakitę. Yra įrodyta [3], kad $Z^{(k)}$ konverguoja į standartinį normalinį atsitiktinį dydį, kai $k \to \infty$. Todėl pakankamai dideliam M gauname

$$f(z) \approx \varphi\bigl(z^{(M)}\bigr) \prod_{k=1}^{M} \frac{g_k(\tau_k' z^{(k-1)})}{\varphi(\tau_k' z^{(k)})}, \tag{2}$$

čia $z^{(k)} = Q_k(x)$, o φ – Gauso pasiskirstymo tankio funkcija. Nežinomus dydžius pakeitę statistiniais įverčiais gauname pasiskirstymo tankio įvertinį pagrįstą tiksliniu projektavimu.

2.1. *Duomenų projekcijų tankio įvertinimas*

[1] yra siūloma duomenų projekcijų tankius g_k vertinti naudojant projekcinius įverčius Ležandro polinomų bazėje. T.y.

$$\widehat{g}_k(u) = \varphi(u) \sum_{j=1}^{J} \frac{2j+1}{n} \sum_{t=1}^{n} \psi_j(\eta_t)\psi_j(u), \tag{3}$$

čia ψ_j – ortogonalūs Ležandro polinomai, J – skleidinio eilė ([1] siūloma naudoti $4 \leqslant J \leqslant 8$), $\eta_t = 2\Phi(\tau_k' Z_t^{(k)}) - 1$, Z_t – atsitiktinės imties elementas.

2.2. *Projektavimo krypties pasirinkimas*

Projektavimo kryptys τ_k turi būti parenkamos taip, kad duomenų projekcija pasirinkta kryptimi turėtų skirstinį, kiek galima mažiau panašų į Gauso. Šiam skirstinų palyginimui naudojama funkcija $I(\tau)$ vadinama projektavimo indeksu

$$\tau_k = \arg\max_{\tau} \widehat{I}(\tau). \tag{4}$$

[1] siūloma projektavimo indekso konstrukcija pagrįsta skleidiniu Ležandro polinomų bazėje

$$\widehat{I}(\tau) = \sum_{j=1}^{J} \frac{2j+1}{2n^2} \left(\sum_{t=1}^{n} \psi_j(\eta_t) \right)^2. \tag{5}$$

3. Algoritmo modifikacijos

3.1. Duomenų projekcijų tankio įvertinimas

Vietoje originalioje procedūroje siūlomo vienamačių duomenų projekcijų pasiskirstymo tankio projekcinio įvertinio Ležandro polinomų bazėje buvo bandytas taikyti branduolinis pasiskirstymo tankio įvertis

$$\widehat{g}_k(u) = \frac{1}{nh} \sum_{t=1}^{n} K\left(\frac{u - \tau_k' Z_t^{(k)}}{h} \right), \tag{6}$$

čia K – Gauso branduolio funkcija, h – branduolio plotis. Branduolio pločio radimui buvo naudota stabilius rezultatus duodanti formulė [7]

$$h = 0,9 \min\left(\widehat{STD}, \widehat{IQR}/1,34\right) n^{-1/5}, \tag{7}$$

čia $\widehat{STD}$ ir $\widehat{IQR}$ yra projektuotų duomenų standartinio nuokrypio ir tarpkvartilinio atstumo įverčiai.

3.2. Projektavimo indekso pasirinkimas

Vietoje originalioje procedūroje siūlomo projektavimo indekso pagrįsto skleidiniu Ležandro polinomų bazėje (5) buvo bandyta naudoti projektavimo indeksą pagrįstą klasikine tikimybine Kolmogorovo–Smirnovo statistika

$$\widehat{I}(\tau) = \sup_u \left| \widehat{G}_k(u) - \Phi(u) \right|, \tag{8}$$

čia $\widehat{G}_k$ yra $\tau' Z^{(k-1)}$ empirinė pasiskirstymo funkcija, o Φ – standartinė Gauso pasiskirstymo funkcija.

4. Palyginamoji analizė

4.1. Tyrimų metodika

Metodai buvo tirti kompiuterinio modeliavimo būdu. Tai leido apskaičiuoti tikrąsias metodų paklaidas ir jas palyginti. Tyrimams buvo pasirinkti keli daugiamačiai vienmodaliniai ir daugiamodaliniai duomenų skirstiniai, turintys skirtingas simetriškumo, glodumo ir „uodegų sunkumo“ charakteristikas:

- vienmodalinis Gauso mišinys (stipriai persidengiantys klasteriai);
- dvimodalinis Gauso mišinys (mažai persidengiantys klasteriai);
- daugiamatis χ^2 skirstinys (koordinatės nepriklausomos pasiskirsčiusios pagal χ^2_4 skirstinį);

- daugiamatis Pareto skirstinys (koordinatės nepriklausomos pasiskirsčiusios pagal Pareto skirstinį su formos parametru $k = 1$).

Modeliuotų duomenų dimensija $d = 4$, imties dydis $n = 100, 200, 500, 1000$.

Metodų paklaidos buvo vertinamas naudojant šiuos tikslumo matus

$$\delta_1 = \frac{1}{n}\sum_{i=1}^{n}\left|f(X_t) - \widehat{f}(X_t)\right| \approx \int \left|f(x) - \widehat{f}(x)\right| f(x)\,\mathrm{d}x, \tag{9}$$

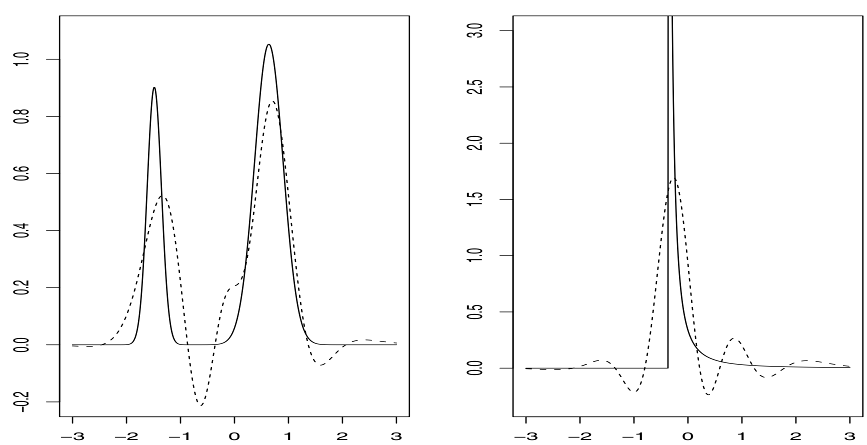
$$\delta_2 = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f(X_t) - \widehat{f}(X_t)}{f(X_t) + \widehat{f}(X_t)}\right| \approx \frac{1}{2}\int \left|f(x) - \widehat{f}(x)\right| \mathrm{d}x. \tag{10}$$

Toks tikslumo matų pasirinkimas leido rezultatus palyginti su kitais darbais, pvz., [5,6].

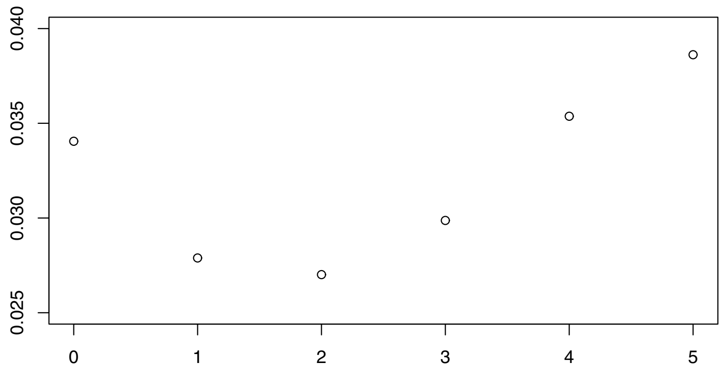
4.2. Tyrimų rezultatai

Pasiskirstymo tankio įvertinių gautų naudojant originalų projektavimo indeksą pagrįstą Ležandro polinomais bei Kolmogorovo–Smirnovo statistika pagrįstą projektavimo indeksą paklaidos buvo panašios ir įvairiems tirtiems skirstiniams skyrėsi mažiau nei 5%. Todėl galima daryti išvada, kad tiksliniu projektavimu pagrįsta tankio įvertinimo procedūra nėra labai jautri projektavimo indekso pasirinkimui. Kolmogorovo–Smirnovo statistikos panaudojimo privalumas yra tas, kad šio indekso reikšmė yra greičiau apskaičiuojama. Projektavimo krypčių radimas pagreitėja beveik 10 kartų (nors tai nedaug įtakoja suminį tankio įvertinimo algoritmo skaičiavimo laiką).

Atliekant modeliavimo tyrimus paaiškėjo, kad Ležandro polinomais pagrįstas projekcinis įvertis blogai vertina tankius skirstiniams labai besiskiriantiems nuo Gauso skirstinio. Tankio įvertis gali įgyti neigiamas reikšmes, o skirstiniams, kurių tankis



1 pav. Tikrasis pasiskirstymo tankis (ištisinė linija) ir jo projekcinis įvertis Ležandro polinomų bazėje (punktyrinė linija). Gauso skirstinių mišinio ir Pareto skirstinio atveju.



2 pav. Pasiskirstymo tankio įvertinio kokybė priklausomai nuo to, keliose pirmose projektavimo kryptyse naudotas branduolinis pasiskirstymo tankio įvertinys.

nėra glodus, įvertis įgija „svyruojančią“ formą. Šie atvejai pavaizduoti 1 pav. Tačiau vertinant tankius artimus Gauso tankiams Ležandro polinomais pagrįstas metodas pranašesnis. Be to, šis metodas duoda geresnius rezultatus esant mažoms imtims.

1 lentelėje pateikiamos δ_1 paklaidų vidutinės reikšmės (apskaičiuotos kartojant tyrimą 20 kartų). Matome, kad branduolinio metodo panaudojimas gali pagerinti daugiamačio pasiskirstymo tankio įvertinio kokybę. Tyrimai parodė, kad Pareto skirstinio atveju visais atvejais tikslinga taikyti branduolinį įvertinį, tačiau rezultatai nėra vienareikšmiai Gauso mišinių atveju – vienoms mišinio parametrų reikšmėms geresnius rezultatus duoda vienas metodas, kitoms kitas. Aiškus dėsningumas nuo mišinio tankio glodumo savybių nepastebėtas. Tačiau aišku, kad branduolinį metodą tikslinga taikyti didesnėms imtis. Naudojant mažo dydžio imtis ($n = 100$) Ležandro polinomų panaudojimas dažnai duoda geresniu rezultatus.

Atsižvelgus į tai, kad pirmosiomis projektavimo kryptimis gautos duomenų projekcijos turi skirstinį daug besiskiriantį nuo Gauso, o vėlesnėse iteracijose transformuotų duomenų skirstinys (o kartu ir jų projekcijų skirstinys) artėja prie Gauso skirstinio, buvo bandyta sukonstruoti algoritmą, kur pirmųjų projekcijų tankio vertinimui būtų naudojamas branduolinis įvertinys (geriau veikiantis kai duomenų skirstinys nepanašus į Gauso), o kitų projekcijų tankio vertinimui – Ležandro polinomais

1 lentelė. Paklaidos δ_1 reikšmių palyginimas Ležandro polinomais ir branduoliniu įvertiniu pagrįstiems metodams įvairiems duomenų skirstiniams. Imties dydis $n = 500$

Skirstinys	Ležandro	Branduolinis
Vienmodalinis Gauso mišinys	**0,01639**	0,01765
Dvimodalinis Gauso mišinys	0,02906	**0,02783**
Daugiamatis χ^2	0,02894	**0,02615**
Daugiamatis Pareto	1,10244	**0,43753**

pagrįstas įvertinys. 2 pav. matome pavaizduotus tyrimo rezultatus vienam iš daugiamodalinių Gauso mišinių, kai imties dydis $n = 500$. Daugiamačio tankio įvertinime buvo naudotos į projektavimo kryptys. Pirmose k krypčių naudotas branduolinis pasiskirstymo tanko įvertinys, o likusiose Ležandro polinomais pagrįstas įvertinys, t.y. kai $k = 0$, naudojamas vien tik Ležandro polinomais pagrįstas įvertinys, o kai $k = 5$ vien tik branduolinis įvertinys. Matome, kad branduolinį įvertinį tikslinga taikyti pirmųjų duomenų projekcijų tankiams vertinti. Optimalus projekcijų kiekis kurios naudotinas branduolinis įvertinys yra skirtingas įvairiems skirstiniams. Čia tikslinga tęsti tyrimus siekiant sukurti metodą parinkti optimalaus projekcijų kiekio parinkimui priklausomai nuo projektavimo indekso reikšmės.

5. Išvados

Iš atliktų tyrimų seka išvados:

1. Naudojant tiek Ležandro polinomais pagrįstą projektavimo indeksą, tiek Kolmogorovo–Smirnovo statistika pagrįstą projektavimo indeksą gaunami nereikšmingai skirtingos tankio įvertinio paklaidos. Tačiau Kolmogorovo–Smirnovo statistika pagrįstas projektavimo indeksas yra greičiau apskaičiuojamas.
2. Neglodžių skirstinių atveju tikslinga taikyti branduolinį pasiskirstymo tankio įvertinį duomenų projekcijų tankiams vertinti.
3. Esant mažoms imtims tikslinga taikyti Ležandro polinomais pagrįstą pasiskirstymo tankio įvertinį duomenų projekcijų tankiams vertinti.
4. Tikslinga naudoti kombinuotą algoritmą taikant tiek branduolinį, tiek Ležandro polinomais pagrįstą pasiskirstymo tankio įvertinį. Reikalingi papildomi tyrimai optimaliam šių įvertinių deriniui nustatyti.

Literatūra

1. J.H. Friedman. Exploratory projection pursuit. *Journal of American Statistical Association*, 82(397):249–266, 1987.
2. J.H. Friedman, J.W. Turkey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23(9):881–890, 1974.
3. J.H. Friedman, W. Stuetzle, A. Schroeder. Projection pursuit density estimation. *Journal of American Statistical Association*, 79(387):599–607, 1984.
4. P.J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.
5. J.-N. Hwang, S.-R. Lay, A. Lippman. Nonparametric multivariate density estimation: a comparative study. *IEEE Transactions on Signal Processing*, 42(10):2795–2810, 1994.
6. T. Ruzgas, R. Rudzkis, M. Kavaliauskas. Application of clustering in the non-parametric estimation of distribution density. *Nonlinear Analysis*: *Modeling and Control*, 11(4):393–411, 2006.
7. B.W. Silverman. *Density Estimation for Statistics Data Analysis*. Chapman and Hall, London, 1986.

SUMMARY

M. Kavaliauskas. Probability density estimation using data projection

Nonparametric estimation of multivariate multimodal probability density is analysed. The projection pursuit density estimator was proposed by J.H. Friedman. Author of this paper proposes the modifications of original Friedman algorithm: employing a kernel density estimator, and a projection index based on Kolmogorov–Smirnov statistic. The efficiency of proposed modifications is analysed using computer simulation technique.

Keywords: projection pursuit, probability density function, nonparametric estimation, projection index.