Liet. matem. rink. **45**, *spec. nr.*, 2005, 369–374

Daugiamačių Gauso skirstinių mišinio modelio panaudojimas neparametrinių tankių vertinime

Tomas RUZGAS (MII), Mindaugas KAVALIAUSKAS (KTU)
el. paštas: tomas.ruzgas@ktu.lt, snaiperiui@takas.lt

1. Įvadas

Sprendžiant įvairius taikomuosius uždavinius, daugelis algoritmų veikia gerai, jei žinoma pasiskirstymo tankių šeima. Deja, tikrovėje šie tankiai paprastai nežinomi, ir tankių vertinimas tampa kritiškai reikšmingas. Skirtingai negu parametriniame tankio vertinime daromos prielaidos apie duomenų parametrinį pasiskirstymą, neparametriniame tankio vertinime prielaidos apie duomenų pasiskirstymą yra ne tokios griežtos.

Turint d-mačius stebėtus duomenis $\{X_i, i = 1, \ldots, n\}$, daugiamačio tankio vertinimo uždavinys yra rasti įvertinį $\hat{f}$, kuris „geriausiai“ aproksimuoja tikrąjį tankį f.

Tradiciškai pasiskirstymo tankio funkcija konstruojama naudojant Gauso branduolį su fiksuotu branduolio pločio parametru h:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big). \tag{1}$$

Paprastai, stebėti duomenys nėra vienodai išsisklaidę visomis kryptimis. Todėl labai pageidautina pakeisti duomenų mastelį panaikinant didžiausius išsibarstymo skirtumus skirtingose koordinačių kryptyse. Vienas patrauklus metodas yra pradinių duomenų standartizavimas tiesinės transformacijos pagalba, kad jų vidurkis taptų nulinis, o kovariacinė matrica vienetinė ($\mathbf{E}[Z] = \mathbf{0}$ bei $\mathbf{E}[ZZ^T] = \mathbf{I}$), ir skirtingo branduolio pločio naudojimas skirtingiems stebėjimams [4]. Šis metodas vadinamas adaptuotu branduoliniu tankio įvertinimu su jautrumo parametru γ.

$$\hat{f}(z) = \frac{1}{n} \sum_{i=1}^{n} h^{-d} \lambda_i^{-d} K\Big(\frac{z - Z_i}{h\lambda_i}\Big), \quad \lambda_i = \big(\tilde{f}(Z_i)/g\big)^{-\gamma}, \tag{2}$$

čia g yra $\tilde{f}(z)$ geometrinis vidurkis, kai $\tilde{f}(z)$ įvertintas (1) pagalba.

Tikslinio projektavimo tankio vertinimo esmė yra „įdomių“, mažo matavimo duomenų projekcijų, kurios parodo skirstinio struktūras, ieškojimas. Nors „įdomumą“ gali būti sunku išreikšti, Huberis pateikė euristinį pasiūlymą, kad Gauso skirstinys privalo būti laikomas mažiausiai įdomiu. Performuojant ši pasiūlymą, Fridmanas pasiūlė

algoritminę procedūrą [1], vadinamą tiriamuoju tiksliniu projektavimu daugiamačio neparametrinio tankio vertinimui.

$$\hat{f}(z) = \varphi(z^{(M)}) \prod_{m=1}^{M} \frac{\hat{f}_{\alpha_m}(\alpha_m^T z^{(m-1)})}{\varphi(\alpha_m^T z^{(m)})}, \tag{3}$$

čia $\hat{f}_{\alpha_m}$ – duomenų $Z^{(m-1)}$ struktūros išilgai m-tos projekcijos α_m tankio įvertis, o φ yra standartinis daugiamatis Gauso skirstinio tankis.

Pusiau parametrinio branduolinio tankio vertinimo atveju d-matis baigtinis atsitiktinis vektorius X gali būti atitinkamai padalintas į s ir $(d-s)$-mates dalis, $X = (Y, Z)$, o tankio funkcija tuomet atitinka

$$f_X(x) = f_{(Y,Z)}(y, z) = f_Y(y) f_{Z|Y=y}(z), \quad x = (y, z) \in \mathbf{R}^d. \tag{4}$$

Čia f_X ir f_Y yra atitinkamai X ir Y tankiai, o $f_{Z|Y=y}$ yra Z tankis prie sąlygos $Y = y$. Tarkime, jog sąlyginis tankis $f_{Z|Y=y}$ yra Gauso, t.y. daugiamatis normalusis, bet taip, kad tankis f_Y nepriklauso jokiai parametrinei šeimai. Tada paprastas f_X įvertis gaunamas vertinant f_Y neparametriškai (1) ir pritaikant daugiamatį normalųjį tankį kiekvienam $f_{Z|Y=y}$[3].

Tankį vertinant logsplainais [2], ieškoma

$$f(x; \beta, t) = \exp\big(\beta_1 B_1(x; t) + \cdots + \beta_J B_J(x; t) - C(\beta, t)\big), \tag{5}$$

pavidalo funkcija, kurią atitinka log-tikėtinumo funkcija

$$l(\beta, t) = \sum_i f(X_i; \beta, t) = \sum_i \sum_j \beta_j B_j(X_i; t) - nC(\beta, t), \tag{6}$$

čia

$$C(\beta, t) = \log\left(\int_L^U \exp\big(\beta_1 B_1(x; t) + \cdots + \beta_J B_J(x; t)\big)\, \mathrm{d}x \right) < \infty, \quad L < x < U.$$

Maksimalaus tikėtinumo įvertis $\hat{\beta}$ randamas iš (6) $\hat{\beta} = \arg\max_{\beta \in B} l(\beta, t)$, ir atitinkamai $\hat{f} = f(x; \hat{\beta}, t)$, $L < x < U$.

Pasinaudokime apvertimo formule [5]

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} \mathrm{e}^{-it^T x} \psi(t)\, \mathrm{d}t, \quad \psi(t) = \mathbf{E}\mathrm{e}^{-it^T X}. \tag{7}$$

Pažymėję $u = |t|$, $\tau = t/|t|$ ir pakeitę kintamuosius į sferinę koordinačių sistemą gauname

$$f(x) = \frac{1}{(2\pi)^d} \int_{\tau:|\tau|=1} \mathrm{d}s \int_0^\infty \mathrm{e}^{-iu\tau^T x} \psi(u\tau) u^{d-1}\, \mathrm{d}u. \tag{8}$$

Čia pirmasis integralas suprantamas kaip paviršinis integralas ant vienetinės sferos paviršiaus.

Pasirinkę projektavimo krypčių, tolygiai išsidėsčiusių ant sferos, aibę T ir charakteristinę funkciją keisdami jos įvertiniu, gauname įverčio skaičiavimo formulę

$$\hat{f}(x) = \frac{c(d)}{\|T\|} \sum_{\tau \in T} \int_0^\infty \mathrm{e}^{-iu\tau^T x} \hat{\psi}_\tau(u) u^{d-1} \mathrm{e}^{-hu^2} \,\mathrm{d}u. \tag{9}$$

Po integralo ženklu įvestas papildomas daugiklis e^{-hu^2} atlieka papildomą įverčio $\hat{f}(x)$ glodinimą su Gauso branduolio funkcija.

Formulė (9) gali būti naudojama esant įvairiems projektuotų duomenų charakteristinės funkcijos įvertiniams. Gauso mišinio atveju patogu naudoti parametrinį šios funkcijos įvertinį

$$\hat{\psi}_\tau(u) = \sum_{j=1}^{\hat{q}_\tau} \hat{p}_j(\tau) \mathrm{e}^{-iu\hat{m}_j(\tau) - u^2 \hat{\sigma}_j^2(\tau)/2}. \tag{10}$$

2. Apvertimo formulė su triukšmo klasteriu

Dabar pasiūlysime naują daugiamačių Gauso skirstinių mišinio modelio panaudojimo metodą. Pasinaudokime apvertimo formule (7). Tankio įverčio skaičiavimo formulėje (9), charakteristinės funkcijos įvertinį konstruokime kaip Gauso mišinio ir tolygaus skirstinių charakteristinių funkcijų sąjungą su atitinkamais svoriais:

$$\hat{\psi}_\tau(u) = \sum_{j=1}^{\hat{q}_\tau} \hat{p}_j(\tau) \mathrm{e}^{iu\hat{m}_j(\tau) - u^2\hat{\sigma}_j^2(\tau)/2} + \hat{p}_0(\tau) \frac{2}{(b-a)u} \sin\frac{(b-a)u}{2} \cdot \mathrm{e}^{\frac{iu(a+b)}{2}}, \tag{11}$$

čia antras narys aprašo tolygaus pasiskirstymo triukšmo klasterį, p_0 – triukšmo klasterio svoris, $a = a(\tau)$, $b = b(\tau)$.

Įstatę (11) į (9), gauname

$$\hat{f} = \frac{c(d)}{\|T\|} \sum_{\tau \in T} \Bigg[\sum_{j=1}^{\hat{q}_\tau} \hat{p}_j(\tau) \int_0^\infty \mathrm{e}^{iu(\hat{m}_j(\tau) - \tau^T x) - u^2(h + \hat{\sigma}_j^2(\tau)/2)} u^{d-1} \,\mathrm{d}u$$
$$+ \frac{2\hat{p}_0(\tau)}{b-a} \int_0^\infty \mathrm{e}^{iu(\frac{a+b}{2} - \tau^T x) - u^2 h} \cdot \sin\frac{b-a}{2} u \cdot u^{d-2} \,\mathrm{d}u \Bigg], \tag{12}$$

Pastebėsime, kad čia galime nagrinėti tik realiąją išraiškos dalį (menamųjų dalių suma turi būti lygi nuliui), nes tankio įvertis $\hat{f}(x)$ gali įgyti tik realias reikšmes. Pasirinkta glodinimo daugiklio forma e^{-hu^2} leidžia susieti glodinimo parametrą h su projekcijų klasterių dispersijomis.

3. Eksperimentinis tyrimas

Ankstesniuose skyriuose aprašytų tankių vertinimo algoritmų efektyvumo tyrimas atliktas Monte-Karlo metodu. Toks algoritmų palyginimo būdas sudaro galimybę išmatuoti tikrąsias stebėjimų tankio reikšmes ir tuo būdu įvertinti algoritmų efektyvumą.

Tyrimui buvo naudojami penkiamačiai ($d = 5$) ir dvimačiai ($d = 2$) Koši skirstinių mišiniai

$$\sum_{j=1}^{q} p_j C(x, m_j, u_j), \quad C(x, m_j, u_j) = \prod_{k=1}^{d} \frac{u_{jk}}{\pi[u_{jk}^2 + (x_k - m_{jk})^2]}.$$

Tankių vertinimo tikslumui išreikšti skaičiuojamos vidutinė absoliutinė (13a) ir vidutinė absoliutinė santykinė (13b) paklaidos

$$\delta_1 = \frac{1}{n} \sum_{t=1}^{n} \left| f(x(t)) - \hat{f}(x(t)) \right| \cong \int \left| f(x) - \hat{f}(x) \right| f(x)\, \mathrm{d}x, \tag{13a}$$

$$\delta_2 = \frac{2}{n} \sum_{t=1}^{n} \left| \frac{f(x(t)) - \hat{f}(x(t))}{f(x(t)) + \hat{f}(x(t))} \right| \cong \int \left| f(x) - \hat{f}(x) \right| \mathrm{d}x. \tag{13b}$$

Skaičiavimai atlikti su imties dydžiais $n = 50, 100, 200, 400, 800$, keičiant mišinį sudarančių skirstinių kiekį, jų svorius bei atstumą tarp centrų. Kiekvienu atveju generuota po 100 imčių.

$q = 2$, $p_1 = (1 - p_2)$, $p_2 = 0.1, 0.2, 0.3, 0.4, 0.5$, $m_1 = (0, 0, 0, 0, 0)$, $m_2 = (0.5i, 0.5i, 0.5i, 0.5i, 0.5i)$, $i = 1, 2, \ldots, 6$.

$q = 3$, $p_1 = p_2 = (1 - p_3)/2$, $p_3 = 0.1, 0.2, 1/3, 0.4, 0.6, 0.8$, $m_1 = (0, 0, 0, 0, 0)$, $m_2 = (0.5i, 0.5i, 0.5i, 0.5i, 0.5i)$, $m_3 = (0.5i, 0.5i, 0, 0, 0)$, $i = 1, 2, \ldots, 6$.

$q = 4$, $p_1 = p_2 = p_3 = (1 - p_4)/3$, $p_4 = 0.1, 0.16, 0.25, 0.4, 0.7$, $m_1 = (0, 0, 0, 0, 0)$, $m_2 = (0.5i, 0.5i, 0.5i, 0.5i, 0.5i)$, $m_3 = (0.5i, 0.5i, 0, 0, 0)$, $m_4 = (0, 0, 0.5i, 0.5i, 0.5i)$, $i = 1, 2, \ldots, 6$.

$q = 2$, $p_1 = (1 - p_2)$, $p_2 = 0.1, 0.3, 0.5$, $m_1 = (0, 0)$, $m_2 = (0.5i, 0.5i)$, $i = 1, 2, \ldots, 6$.

$q = 3$, $p_1 = p_2 = (1 - p_3)/2$, $p_3 = 0.1, 1/3, 0.8$, $m_1 = (0, 0)$, $m_2 = (0.5i, 0.5i)$, $m_3 = (0.5i, 0)$, $i = 1, 2, \ldots, 6$.

$q = 4$, $p_1 = p_2 = p_3 = (1 - p_4)/3$, $p_4 = 0.1, 0.25, 0.7$, $m_1 = (0, 0)$, $m_2 = (0.5i, 0.5i)$, $m_3 = (0.5i, 0)$, $m_4 = (0, 0.5i)$, $i = 1, 2, \ldots, 6$.

Algoritmai: AKDE – adaptuotas branduolinis, PPDE – tikslinio projektavimo, LSDE – logsplainų, SKDE – pusiau parametrinis branduolinis, IFDE – apvertimo formulė, MIDE – apvertimo formulė su triukšmo klasteriu. IFDE ir MIDE metoduose naudojami mišinio parametrai apskaičiuoti Matematikos ir informatikos institute (Vilnius) sukurta programine įranga [6].

Reikšmių parinkimas. Adaptuoto branduolinio metodo naudotos jautrumo parametro reikšmės $\gamma = (0.2, 0.4, 0.6, 0.8)$. Tiksliniame projektavime modeliuota su trimis Lagranžo polinomų eilėmis: 4, 5 ir 6. Naudota tiek projektavimo krypčių, kad projektavimo indeksas $I(\alpha) = \int_{-1}^{1} f_r^2(r)\, \mathrm{d}r - 1/2 = \sum_{j=1}^{J} \frac{2j+1}{2} E_r^2[\psi_j(r)]$ į naują kryptį pasidarytų mažesnis už 0.001. Logsplainų metodas automatizuotai parenka bazinio splaino taškus minimizuojant Akaike informacijos kriterijų $\mathrm{AIC}(t) = -2l(t) + aJ(t)$, J – splaino laipsnis, $a = \log(n)$, l – tikėtinumo funkcija naudojama parenkant splaino koeficientus. Pusiau parametriniame branduoliniame metode X vektorius dalijamas į

$0 \leqslant s \leqslant d$ dalis, glodumo parametrai $h_1 = 0.2, 0.4, 0.6, h_2 = 2h_1$. Apvertimo formulės ir modifikuotos apvertimo formulės metoduose glodinimo parametras parenkamas „*cross-validation*“ būdu, triukšmo klasterio svoris $p_0 = 0.05, 0.1, 0.15, 0.2, 0.3, 0.4$.

Rezultatai. LSDE metodas esant labai didelėms išskirtims ($|x - m_j| > 100u_j$) jas grupuodamas su didesniu kiekiu reikšmių esančių arčiau skirstinio centro, ir apskaičiuotų splaino koeficientų pagalba, tankį išskirtyse įvertina 10^{100} eilės, kas yra nekorektiška, ir tokiais atvejais šio metodo naudojimas nėra rekomenduotinas.

Skaičiuojant vidutinę absoliutinę paklaidą penkiamačiu atveju:

- kai $n = 50$, geriausi rezultatai gaunami AKDE, MIDE ir IFDE metodais;
- kai $n = 100$, geriausi rezultatai gaunami MIDE, AKDE ir PPDE metodais;
- kai $q = 2, n \geqslant 200$, geriausi rezultatai gaunami SKDE ir MIDE metodais;
- kai $q \geqslant 3,\ n = 200$, smarkiai persidengiančių skirstinių atvejais ($i = 1, 2$) geriausi rezultatai gaunami SKDE, o labiau atsiskyrusių ($i \geqslant 3$) – MIDE metodu;
- kai $q = 3,\ n \geqslant 400$, geriausi rezultatai gaunami SKDE ir MIDE metodais;
- kai $q = 4,\ n = 400$, smarkiai persidengiančių skirstinių atvejais ($i \leqslant 3$) geriausi rezultatai gaunami SKDE, o labiau atsiskyrusių ($i \geqslant 4$) – MIDE metodu;
- kai $q = 4,\ n \geqslant 400$, geriausi rezultatai gaunami SKDE ir MIDE metodais.

Skaičiuojant vidutinę absoliutinę santykinę paklaidą penkiamačiu atveju smarkiai ir lengvai persidengiančių skirstinių atvejais ($i \leqslant 4$) geriausi rezultatai gaunami SKDE, o atsiskyrusių skirstinių – MIDE metodu (žr. 1 lentelę, skliaustuose vaizduojamos paklaidų standartinių nuokrypių reikšmės).

Dvimačiu atveju skaičiuojant abi paklaidas smarkiai ir lengvai persidengiantiems skirstiniams ($i \leqslant 4$) geriausi rezultatai gaunami SKDE, o atsiskyrusių skirstinių – AKDE metodais.

Vidutinės absoliutinės paklaidos atveju, kuomet mišinius sudaro smarkiai persidengiantys skirstiniai, rezultatai gauti blogesni nei tada, kuomet skirstiniai yra atsiskyrę.

Vidutinės absoliutinės santykinės paklaidos pavyzdys

Vertinimo metodai	Tankiai					
	$d = 5;\ p_1 = p_2 = p_3 = 1/3;\ n = 100$					
	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	0.826808648 (0.07603562)	0.825704404 (0.08142691)	0.819752795 (0.08480668)	0.812811066 (0.08271640)	0.806631805 (0.07880315)	0.807497095 (0.07308145)
PPDE	0.924337085 (0.05008109)	0.931891639 (0.03642851)	0.930329934 (0.03748276)	0.930005809 (0.03867695)	0.928387201 (0.04104483)	0.925006274 (0.04333755)
LSDE	0.804288749 (0.0533969)	0.816248491 (0.05403439)	0.858309622 (0.04909458)	0.861129929 (0.03487466)	0.861347825 (0.04344762)	0.871065205 (0.05765906)
SKDE	0.715837322 (0.02599728)	0.714378646 (0.09050244)	0.708771976 (0.09054359)	0.707097612 (0.08299397)	0.717894995 (0.06305649)	0.722714169 (0.04982716)
IFDE	0.945929689 (0.0362443)	0.888569351 (0.13177255)	0.785702837 (0.07055977)	0.846303854 (0.03802011)	0.87606106 (0.11104464)	0.831186625 (0.05383089)
MIDE	0.738921095 (0.02797055)	0.733226436 (0.02215356)	0.723520668 (0.0337667)	0.714941374 (0.01954811)	0.712088441 (0.02078429)	0.721910599 (0.02027824)

Vidutinės absoliutinės santykinės paklaidos atveju geresni rezultatai gauti kuomet skirstinių centrai nutolę vidutiniškai ($i = 3, 4, 5$).

4. Išvados

1. Trumpai apžvelgti populiarūs ir dažnai praktikoje sutinkami neparametriniai tankių vertinimo algoritmai.
2. Ištirtas tankių vertinimo apvertimo formule algoritmas. Triukšmo klasterio įtraukimas ženkliai pagerino apvertimo formulės taikymo rezultatus, kas ypač buvo akivaizdu mišiniams su išsiskiriančiais stebėjimais.
3. Didesnės dimensijos atveju ($d \sim 5$) skaičiuojant vidutinę absoliutinę paklaidą ir esant labai mažoms imtims ($n \sim 50$) su išskirtimis rekomenduotina naudoti adaptuoto branduolio metodą, imtims $n \sim 100$ – modifikuotos apvertimo formulės metodą, didesnėms imtims su persidengiančiais skirstiniais – pusiau parametrinį branduolinį, o su labiau atsiskyrusiais – modifikuotos apvertimo formulės metodus; vidutinės absoliutinės santykinės paklaidos atveju imtims su persidengiančiais skirstiniais rekomenduotina pusiau parametrinis branduolinis, o imtims su atsiskyrusiais – modifikuotos apvertimo formulės metodai. Mažos dimensijos atveju ($d \sim 2$) imtims su persidengiančiais skirstiniais rekomenduotina pusiau parametrinis branduolinis, o imtims su atsiskyrusiais – adaptuoto branduolio metodai.

Literatūra

1. J.H. Friedman, Exploratory projection pursuit, *Journal of the American Statistical Association*, **82**(397), 249–266 (1987).
2. M.H. Hansen, C. Kooperberg, Spline adaptation in extended linear models, *Statistical Science*, **17**(1), 2–20 (2002).
3. L. Holmström, F. Hoti, Application of semiparametric density estimation to classification, *ICPR*, (3), 371–374 (2004).
4. 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).
5. M. Kavaliauskas, R. Rudzkis, Projection-based estimation of multivariate distribution density, *Liet. matem. rink.*, **42**(spec. nr.), 529–536 (2002).
6. R. Rudzkis, M. Radavicius, Statistical estimations of a mixture of Gaussian distributions, *Acta Applicandae Mathematicae*, **38**, 37–54 (1995).

SUMMARY

T. Ruzgas, M. Kavaliauskas. Nonparametric density estimation using a multidimensional mixture model of Gaussian distributions

This paper algorithmically and empirically studies five major types of nonparametric multivariate density estimation techniques, where no assumption is made about data being drawn from any of known parametric families of distribution. There is developed method of inversion formula where noise cluster is included to general Gaussian mixture model.

Keywords: nonparametric density estimation, inversion formula, characteristic function.