Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
54 t., 2013, 23–28

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

Ležandro polinomo eilės parinkimas tikslinio projektavimo tankio įvertinime

Mindaugas Kavaliauskas

Kauno technologijos universitetas, Fundamentaliųjų mokslų fakultetas
Studentų g. 50, LT-51368 Kaunas
E. paštas: kavaliauskas.mindaugas@gmail.com

Santrauka. Straipsnyje aptariamas tikslinio projektavimo algoritmas ir jo panaudojimas pasiskirstymo tankio vertinime. Analizuojamas J.H. Friedman pasiūlytas tankio įvertinys, kai duomenų projekcijų tankiai yra vertinami ortogonaliųjų Ležandro polinomų bazėje. Sprendžiamas Ležandro polinomo eilės parinkimo klausimas. Išvados daromos naudojant Monte Karlo modeliavimo tyrimą.

Raktiniai žodžiai: tikslinis projektavimas, pasiskirstymo tankis, neparametrinis vertinimas.

Įvadas

Pasiskirstymo tankis yra viena iš pagrindinių charakteristikų apibūdinančių atsitiktinį dydį. Parametriniai įvertiniai, branduoliniai įvertiniai ir histogramos dažniausiai naudojami tankiui įvertinti. Sprendžiant praktinius uždavinius paprastai naudojamas statistinis hipotezių tikrinimas, duomenų klasifikavimas, priklausomybės tarp kintamųjų tyrimas, prognozavimas. Gali susidaryti įspūdis, kad pasiskirstymo tankio įvertinimas yra vadovėlinis uždavinys. Tikriausiai taip ir būtų, jeigu tankio įvertis būtų galutinis statistinės analizės rezultatas, tačiau nereikia pamiršti, kad tankis yra naudojamas, kaip tarpinis rezultatas kitiems uždaviniams spręsti. Pavyzdžiui, diskriminantinės analizės uždaviniai remiasi aposteriorinėm klasifikavimo tikimybėmis, kurios apskaičiuojamos naudojant pasiskirstymo tankius. Tad tokių klasifikavimo uždavinių kaip ligos diagnozavimas ar rašto atpažinimas sprendimas tarpiniuose skaičiavimuose dažniausiai naudos tankio įverčius.

Jeigu duomenų dimensija yra didelė, neparametriškai vertinant daugiamatį tankį susiduriama su taip vadinamu „daugiamatiškumo prakeikimu“ (angl. curse of dimensionality). Šį reiškinį galima iliustruoti paprastu pavyzdžiu – tegu turime vienetinį 10-matį rutulį tolydžiai užpildytą taškais, tuomet į jo viduje esantį $(0.05)^{1/10} \approx 0.74$ spindulio rutulį pateks tik 5% pradinio rutulio taškų. Būtent tokia situacija (mažai imties taškų didelio spindulio aplinkoje) atsitinka bandant naudoti branduolinį tankio įvertinį daugiamačiu atveju.

Tikslinio projektavimo idėja pirmą kartą aprašyta [4]. Jos esmė – rasti „įdomias duomenų“ projekcijas, kurios atspindėtų daugiamačių duomenų struktūrą. Projekcijos turinčios Gauso skirstinį laikomos labiausiai „neįdomiomis“. Tai pagrindžiančius argumentus galima rasti aprašytus [7, 6]. Vėliau tikslinio projektavimo idėja buvo panaudota regresijos įvertinime [2], diskriminantinėje analizėje [6, 9]. Pasiūlyti net du skirtingi tankio įvertinimo būdai [3, 1]. Jie aptarti ir skirtumai tarp jų išryškinti [10].

1 Tankio įvertinimas naudojant tikslinį projektavimą

Šiame straipsnyje aprašomas tankio įvertinimo procedūros, pasiūlytos [1], tyrimas. Šis tiksliniu projektavimu paremtas tankio įvertinimas susideda iš žingsnių:

1. *Duomenų sferinimas.* Pradinę imtį, sudarytą iš atsitiktinio dydžio X stebinių, transformuojame taip, kad ji turėtų nulinį vidurkį ir vienetinę kovariacinę matricą (dėl šių savybių transformacija vadinama sferinimu). Taip transformuotą stebimą atsitiktinį dydį žymėsime Z. Tolimesni veiksmai ir tankio vertinimas bus atliekamas šiam atsitiktiniam dydžiui. Ši transformacija padaro procedūrą invariantiška poslinkio ir mastelio atžvilgiu. Šiuo požiūriu tikslinis projektavimas labai skiriasi nuo pagrindinių komponenčių metodo (PKM), nes PKM svarbiomis laikomos tos kryptys, kurių dispersija yra didžiausia. Tiksliniame projektavime krypties „įdomumą“ nulemia skirstinio tankio konfigūracija, o ne dispersija.

   Algoritmo žingsniai 2–4 cikliškai kartojami. Indeksas k žymės ciklo numerį. Pradedama turint imtį sudaryta iš stebinių $Z^{(0)} = Z$.

2. *„Įdomios“ krypties paieška.* Šiame žingsnyje ieškoma projektavimo krypties, kuri maksimizuotų tam tikrą funkciją vadinamą projektavimo indeksu. Maksimizavimui paprastai naudojamas skaitinis kvazi-Niutono metodas.

3. *Imties projekcijos tankio įvertinimas.* Vienamatės stebinių projekcijos $\alpha_k^T Z^{(k-1)}$ pasiskirstymo funkciją žymėkime $F_{\alpha_k}^{(k-1)}$, o tankį $f_{\alpha_k}^{(k-1)}$. Šiame žingsnyje randamas šio tankio įvertis. Jis vėliau bus naudojamas stebimo daugiamačio atsitiktinio dydžio tankiui įvertinti.

4. *Struktūros panaikinimas.* Šiame etape imtis transformuojama taip, kad duomenų $Z^{(k-1)}$ projekcija rastąja kryptimi α_k turėtų Gauso skirstinį, taigi taptų „neįdomi“, o kryptimis, ortogonaliomis α_k, duomenų projekcijų reikšmės nepakistų. Struktūros panaikinimo transformacija $\Theta^{(k)}$ yra nusakoma formule

$$\begin{aligned} Z^{(k)} &= \Theta^{(k)}\big(Z^{(k-1)}\big) \\ &= Z^{(k-1)} - \big(\alpha_k^T Z^{(k-1)}\big)\alpha_k + \Phi^{-1}\big(F_{\alpha_k}^{(k-1)}(\alpha_k^T Z^{(k-1)})\big)\alpha_k, \end{aligned} \tag{1}$$

   čia antrasis dešinės pusės dėmuo panaikina projekcijos α_k kryptimi reikšmę (ji tampa lygi nuliui), o trečiojo dėmens dėka šios projekcijos skirstinys tampa normaliuoju. Φ žymi standartinę Gauso skirstinio funkciją.

   2–4 žingsniai kartojami, kol transformuotų duomenų $Z^{(k)}$ skirstinys tampa artimas daugiamačiam standartiniam Gauso skirstiniui.

5. *Pasiskirstymo tankio įvertinio apskaičiavimas.* Kadangi $Z^{(k)}$ skirstinys artėja prie Gauso skirstinio, kai k didėja, tai esant pakankamai dideliam krypčių kiekių k (praktikoje, tai būna nuo kelių iki keliolikos krypčių), tankio įvertinime $Z^{(k)}$ skirstinį aproksimuojame Gauso skirstiniu. Panaudojus ankščiau įvertintus projekcijų pasiskirstymo tankių įvertinius, atsitiktinio dydžio Z tankio įvertis $\hat{f}$ apskaičiuojamas

$$\hat{f}(z) = \varphi_d\big(T^{(K)}(z)\big) \prod_{k=1}^{K} \frac{\hat{f}_{\alpha_k}^{(k-1)}(\alpha_k^T T^{(k-1)}(z))}{\varphi(\Phi^{-1}(\hat{F}_{\alpha_k}^{(k-1)}(\alpha_k^T T^{(k-1)}(z))))}, \tag{2}$$

čia φ ir φ_d – atitinkamai vienamačio ir d-mačio Gauso skirstinio tankio funkcijos, o $T^{(k)} = \Theta^{(k)} \cdots \Theta^{(2)} \cdot \Theta^{(1)}$ – struktūros panaikinimo žingsnio transformacijų kompozicija. Empirinė pasiskirstymo funkcija paprastai naudojama, kaip pasiskirtymo funkcijos įvertis $\hat{F}_{\alpha_k}^{(k-1)}$.

2 Tyrimas

Norint įvertinti pasiskirstymo tankį aprašytuoju metodu, reikia papildomai pasirinkti projektavimo indeksą bei tankio įvertinį vienmatėms duomenų projekcijoms. Po tikslinio projektavimo metodo idėjos pasiūlymo, projektavimo indeksai ir jų savybės buvo gana plačiai analizuoti pačių idėjos autorių ir kitų tyrėjų. Populiariausių indeksų apžvalgą ir palyginimą galime rasti [5, 8].

Šiame tyrime naudojome projektavimo indeksą pasiūlytą [1]:

$$I(\alpha) = \int_{-1}^{1} \left(f_R(x) - \frac{1}{2} \right)^2 dx, \tag{3}$$

kur $R = 2\Phi(\alpha^T Z) - 1$. Jo reikšmė parodo imties projekcijos $\alpha^T Z$ transformacijos R tankio atstumą nuo tolygiojo intervale $[-1, 1]$ skirstinio tankio. Jei $\alpha^T Z$ pasiskirstęs pagal Gauso dėsnį (t. y., projekcija yra „neįdomi“), tai R skirstinys yra tolygusis intervale $[-1, 1]$ ir projektavimo indekso $I(\alpha)$ reikšmė lygi nuliui. Esant kitokiam projekcijos skirstiniui, $I(\alpha)$ reikšmė yra teigiama. Projektavimo indekso reikšmę siūloma vertinti tankį f_R skleidžiant ortogonaliųjų Ležandro polinomų bazėje ir iš imties įvertinant pirmuosius $4 \leqslant J \leqslant 8$ skleidinio koeficientus.

Esant tokiam projektavimo indekso pasirinkimui, natūralu ir patį projekcijos tankį vertinti Ležandro polinomų bazėje. Tuomet tie patys skleidinio koeficientai nusakys parametrinę tankio išraišką.

Asimptotiškai optimali J eilė, kai $n \to \infty$, ieškant pirmosios projektavimo krypties α_1 yra išvesta [5], tačiau praktinės parinkimo problemos išlieka, nes optimalus J parinkimas priklauso ne tik nuo imties dydžio n, bet ir duomenų dimensijos d, bei sunkiai matematiškai nusakomos priklausomybės nuo daugiamačių duomenų skirstinio. Be to, domina optimalus J parinkimas ne tik pirmosios krypties paieškoje.

Šiame darbe buvo siekiama suformuluoti praktiškai pritaikomą kriterijų Ležandro polinomo eilei J parinkti. Tyrimas buvo atliekamas naudojant kompiuterinio modeliavimo būdą.

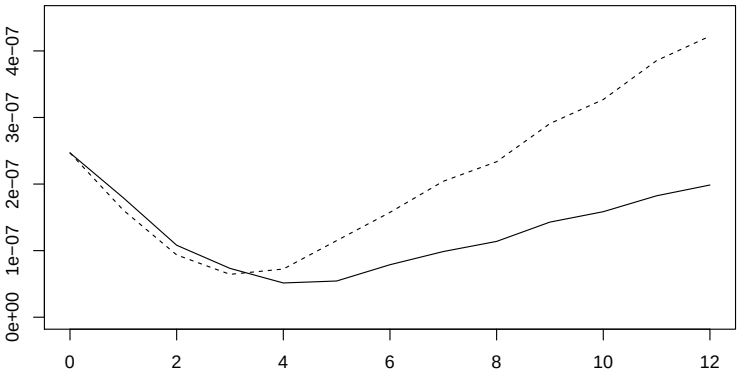
Tikslinio projektavimo metodiką verta taikyti, kai imties skirstinys yra sudėtingos struktūros – daugiamodalis. Tokio tipo skirstiniai yra gerai aproksimuojami skirstinių mišiniais, todėl tyrime imtis buvo generuojama naudojant Gauso skirstinių mišinio modelį. Tuomet stebimo atsitiktinio vektoriaus skirstinio tankis yra lygus

$$f(x) = \sum_{j=1}^{q} p_j \varphi(x, M_j, R_j), \quad \sum_{j=1}^{q} p_j = 1, \tag{4}$$

čia q – mišinio komponenčių kiekis, p_j, M_j ir R_j – mišinio komponenčių svoriai, vidurkiai ir kovariacinės matricos.

Atliekant tyrimą tikslumo matu buvo pasirinkta paklaida

$$\varepsilon = \sum_{i=1}^{m} \left(\hat{f}(\tilde{X}_i) - f(\tilde{X}_i) \right)^2 \approx \int \left(\hat{f}(x) - f(x) \right)^2 f(x)\, dx, \tag{5}$$



1 pav. Paklaidos priklausomybė nuo krypčių kiekio. Imties parametrai $d = 5$, $n = 500$, trys klasteriai. Ištisinė linija — $J = 4$, punktyrinė - - - $J = 8$.

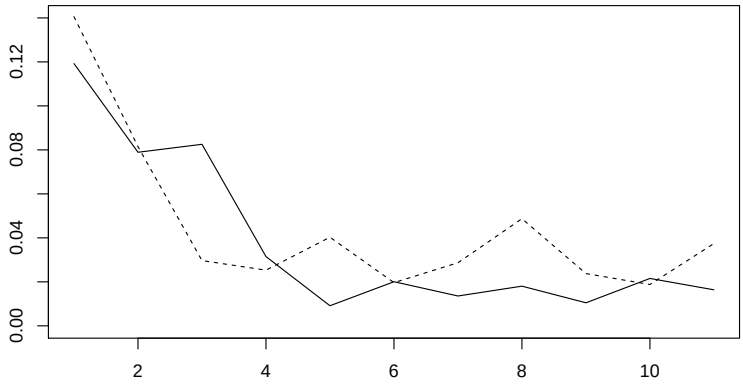
čia $(\tilde{X}_1, \ldots, \tilde{X}_m)$ – tikslumo matavimo imtis, $\hat{f}$ – tikslinio projektavimo metodu gautas tankio įvertis, f – tikrasis skirstinio tankis. Nors ši paklaida vertina ne vidutinę integruotą kvadratinę paklaidą, o vidutinė *svorinę* integruotą kvadratinę paklaidą, tačiau tokios paklaidos parinkimas yra įprastas neparametriniame pasiskirstymo tankio įvertinime, nes čia svarbu ne tankio įvertinio savybės skirstinio „uodegose“, o pagrindinės tankio dalies įvertinimo tikslumas. Šis įvertis kaip tik suteikia mažą svorį kvadratinei paklaidai srityse, kur tankio reikšmės yra mažos. Tikslumo įvertinimui buvo generuota imtis skirtinga nuo metodo tyrime naudotos pagrindinės imties, kuri buvo sudaryta iš taškų turinčių tą patį skirstinį, kaip ir tiriama imtis. Toks atskiros, tik tikslumo įvertinimui skirtos, imties generavimas turi du privalumus – ji objektyviau nusako įvertinio tikslumą, nes įvertinys nėra prisitaikęs prie šios imties; leidžia tiksliau įvertinti paklaidas, kai pagrindinės imties tūris yra mažas.

Tyrime naudoti tokie parametrai: duomenų dimensija $d = 2$, 5 arba 10; naudojamas mišinio komponenčių kiekis $q = 3$, o mišinio parametrai p_j, M_j ir R_j buvo parinkti dviem būdais – kad komponentės būtų persidengiančios, arba, kad nepersidengiančios; imties tūris $n = 100$, 200, 500 arba 1000; Ležandro polinomo eilė $J = 4$, 6, arba 8; tikslumo įvertinimo imties tūris $m = 1000$.

Monte Karlo tyrimo rezultatus pristatysime keliais tipiniais paklaidų grafikais. Tankio įvertinio paklaida mažėja iki pasiekiamas optimalus projektavimo krypčių kiekis (dažniausiai kelios kryptis). Jeigu tikslinio projektavimo algoritmas tęsiamas, tankio įverčio paklaida didėja. Dėl šios priežasties optimalus algoritmo stabdymas yra svarbus, tačiau šiame darbe nagrinėjamas tik polinomo eilės J parinkimo klausimas.

Viena iš aiškių tendencijų pastebėta visoms imtims yra ta, kad esant didesnei Ležandro polinomo eilei, paklaidos didėjimas esant perteklinėms projektavimo kryptims yra greitesnis nei esant mažai polinomo eilei. Ši tendencija aiškiai matoma 1 pav. palyginant punktyrine ir taškine linija pavaizduotus grafikus.

2 pav. pavaizduota projektavimo indekso I priklausomybė nuo projektavimo krypties numerio. Matome, kad pirmose kryptyse stebimas gana greitas indekso reikšmės mažėjimas, tačiau veliau indeksas gali įgauti svyruojantį pobūdį (ypač didesnėms J reikšmėms), arba turėti reikšmes artimas nuliui (dažniau mažoms J reikšmėms). Mažiausia tankio įverčio paklaida pasiekiama esant panašiam krypčių kiekiui, kaip ir mažiausia projektavimo indekso reikšmė, tačiau naudoti vien tik indekso reikšmę,



2 pav. Projektavimo indekso priklausomybė nuo projektavimo krypties numerio. Imties parametrai $d = 5$, $n = 200$, trys klasteriai. Ištisinė linija — $J = 4$, punktyrinė - - - $J = 8$.

kaip algoritmo stabdymo sąlygą nėra pakankamai tikslu, nes paklaidų reikšmės gali pakankamai daug skirtis net tada, jei suklystama tik 1 ar 2 kryptimis, lyginant su optimaliu krypčių kiekiu.

Projektavimo indekso absoliuti reikšmė priklauso nuo duomenų dimensijos ir tankio formos, todėl sunku apibrėžti slenkstį indekso reikšmėms. Todėl buvo bandyta panaudoti statistiką, kurios absoliuti reikšmė butu labiau nepriklausoma nuo šių faktorių. Tam panaudotas Shapiro–Wilk normalumo testas duomenų projekcijai. Šio testo kritinis reikšmingumo lygis nepriklausomai nuo dimensijos ir tankio formos parodo projekcijos panašumą į Gauso skirstinį.

Pastebėta, kad esant sudėtingai projekcijos tankio formai (pvz., daugiamodaliniam tankiui), tankis tiksliau vertinamas aukštesnės eilės Ležandro polinomu, o jei tankio forma nesudėtinga – tikslesnius rezultatus duoda žemesnės eilės polinomai. Tad bandyta polinomo eilę J parinkti priklausomai nuo Shapiro–Wilk testo kritinės reikšmingumo reikšmės. Bandymai parodė, kad neblogi rezultatai gaunami jei naudojama $J = 8$, kai $p < 0{,}07$ ir $J = 4$, kai $p > 0{,}07$. Tokiu atveju pavyksta suderinti tikslų projekcijos tankio funkcijos įvertinimą ir gauti nedidelį paklaidos didėjimą esant perteklinėms projektavimo kryptims. Vienas iš tokių atvejų pavaizduotas 1 pav. ištisine linija.

3 Išvados

Atlikus tyrimą gautos šios išvados:

- esant pertekliniam krypčių kiekiui tankio įvertinimo paklaida didėja. Paklaida didėja sparčiau esant aukštesnei Ležandro polinomo eilei;
- Ležandro polinomo eilę projekcijos tankio įvertinime verta parinkti naudojant Shapiro–Wilk normalumo testą;
- optimalus krypčių kiekio parinkimas svarbus įvertinio tikslumui. Jį reiktų tirti papildomai.

Literatūra

[1] J.H. Friedman. Exploratory projection pursuit. *J. Am. Stat. Assoc.*, **82**(397):249–266, 1987.

[2] J.H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Am. Stat. Assoc.*, **76**(376):817–823, 1981.

[3] J.H. Friedman, W. Stuetzle and A. Schroeder. Projection pursuit density estimation. *J. Am. Stat. Assoc.*, **79**(387):599–607, 1984.

[4] J.H. Friedman and J.W. Turkey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, **23**(9):881–890, 1974.

[5] P. Hall. On polynomial-based projection indices for exploratory projection pursuit. *Ann. Stat.*, **17**(2):589–605, 1989.

[6] P.J. Huber. Projection pursuit. *Ann. Stat.*, **13**(2):435–475, 1985.

[7] M.C. Jones. *The Projection Pursuit Algorithm for Exploratory Data Analysis*. PhD thesis, University of Bath, 1983.

[8] G.P. Nason. *Design and Choice of Projection Indices*. PhD thesis, University of Bath, 1992.

[9] J. Polzehl. Projection pursuit discriminant analysis. *Comput. Stat. Data Anal.*, **20**(2):141–157, 1995.

[10] M. Zhu. On the forward and backward algorithms of projection pursuit. *Ann. Stat.*, **32**(1):233–244, 2004.

SUMMARY

Legendre polynomial order selection in projection pursuit density estimation
M. Kavaliauskas

Projection pursuit method and its application to probability density estimation is discussed. Method proposed by J.H. Friedman, based on projection density estimation using orthogonal Legendre polynomials, is analysed. Problem of Legendre polynomial order selection is solved. Conclusions are based on Monte Carlo simulation.

Keywords: projection pursuit, probability density, nonparametric estimation.