MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Mindaugas Kavaliauskas

# Daugiamačių Gauso skirstinių mišinio statistinė analizė, taikant duomenų projektavimą

Daktaro disertacija

Fiziniai mokslai, matematika (01 P)

Vilnius, 2005

MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Mindaugas Kavaliauskas

# Daugiamačių Gauso skirstinių mišinio statistinė analizė, taikant duomenų projektavimą

Daktaro disertacija

Fiziniai mokslai, matematika (01 P)

Vilnius, 2005

Disertacija rengta 1999–2004 metais Matematikos ir informatikos institute.

Mokslinis vadovas
prof. habil. dr. Rimantas RUDZKIS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika, 01 P).

# Padėka

*Dėkoju darbo vadovui prof. habil. dr. Rimantui Rudzkiui už vadovavimą disertaciniam darbui, skirtą laiką ir energiją, doc. dr. Marijui Radavičiui už išsamius atsakymus į visus klausimus, recenzentei doc. dr. Birutei Kryžienei už pastabas, padėjusias patobulinti disertaciją, bendrovei "DB Topas" už suteiktą kompiuterinę techniką ir galimybę dirbti netrukdant mokslams, Dovilei už įvairią pagalbą rašant disertaciją. Taip pat dėkoju visiems kitiems, kurie tiesiogiai ar netiesiogiai prisidėjo prie šio darbo.*

# Turinys

# Įvadas

**Tiriamoji problema ir darbo aktualumas.** Nagrinėjama problema yra glaudžiai susijusi su daugiamačių stebėjimų klasterizavimu — viena iš svarbių duomenų analizės šakų, kuria remiasi daugelis vaizdų atpažinimo (angl. *pattern recognition*) uždavinių. Klasterizavimo metodologija susilaukia vis didesnio dėmesio dėl naujai atsiradusių taikymo sričių: prekių-pirkėjų grupavimo didžiulėse prekybos duomenų bazėse, su internetu susijusių duomenų analizės, genetinės informacijos apdorojimo ir t.t. Nežiūrint į daugybę egzistuojančių duomenų klasterizavimo metodų, įvairūs autoriai vis siūlo naujas idėjas (pvz., [13], [37], [52], [83], [112], [172]) ir algoritmus: MULTICLUS [46], CLARANS [57], MCLUST [63], EMMIX [111], SMEM [171] ir kitus.

Praktikoje daugelis duomenų yra klasterizuojama naudojant euristines, bet intuityviais argumentais pagrįstas procedūras, ir daugelis statistinės programinės įrangos yra šio tipo. Viena iš tokių klasifikavimo procedūrų grupių yra hierarchinio klasifikavimo procedūros, aprašytos [175]. Kita metodų grupė iteratyviai perskirstyto stebėjimus tarp grupių taip, kad minimizuotų kokią nors tikslo funkciją. Dažniausiai ši funkcija priklauso nuo atstumų tarp stebėjimų, todėl šie metodai paprastai vadinami *geometriniais klasterizavimo metodais*. Labiausiai žinomas ir paplitęs yra $k$-vidurkių metodas [73].

Tikimybiniai metodai yra kita populiari klasterizavimo metodų grupė, kuri atsirado nuo pat klasterinės analizės atsiradimo. Jie skiriasi nuo anksčiau minėtų tuo, kad nusako duomenų tikimybinį modelį ir įgalina geriau panaudoti turimą apriorinę informaciją. Neretai išaiškėja, kad žinomas euristinis metodas yra dalinis tikimybinio metodo atvejis. Pavyzdžiui, Gauso skirstinių mišinius su komponentėmis, kurių kovariacinės matricos yra vienodos ir proporcingos vienetinei matricai, optimaliai klasifikuoja $k$-vidurkių metodas.

Tikimybiniai klasterizavimo metodai paprastai yra pagrįsti skirstinių mišinio statistine analize. Tai lengva paaiškinti: jeigu stebimas atsitiktinis dydis priklauso vienai iš kelių klasių, kurios pasirenkamos atsitiktinai, tai atsitiktinio dydžio skirstinys bus minėtas klases atitinkančių skirstinių mišinys.

Gauso atsitiktiniai dydžiai ypač dažnai sutinkami praktikoje, nes jeigu atsitiktinis dydis priklauso nuo daugelio besisumuojančių atsitiktinių faktorių, tai, pagal centrinę ribinę teoremą (patenkinus tam tikras sąlygas), suma apytikriai tenkins Gauso skirstinio modelį. Taigi, tais atvejais, kai stebimas Gauso atsitiktinis dydis priklauso vienai iš kelių klasių,

stebėsime dydį, tenkinantį Gauso skirstinių mišinio modelį. Gauso skirstinių mišiniai aktualūs įvairiose srityse: biologijoje, medicinoje [5], astronomijoje [113], karyboje [176] ar net tekstilėje [26]. Daug klasterinės analizės uždavinių, tarp jų ir naudojančių šį mišinio modelį, galima rasti tam skirtoje monografijoje [7] ir knygoje [109]. Dėl dažno Gauso skirstinių mišinio modelio panaudojimo taikomuosiuose klasterizavimo tyrimuose, šio modelio statistinio identifikavimo uždavinys yra ypač aktualus, tačiau iki šiol nėra iki galo išspręstas.

Parametrų įverčiams rasti galima taikyti *maksimalaus tikėtinumo metodą* (angl. *maximum likelihood method*, toliau *MTM*), deja, tokių įverčių apskaičiavimas yra rimta praktinė problema daugiamačiu atveju. Didėjant duomenų matavimų skaičiui ir klasterių kiekiui, parametrų dimensija greitai didėja, o tikėtinumo funkcija turi daug lokalių ekstremumų. Klasikinių optimizavimo metodų taikymas neduoda gerų rezultatų. Todėl dažniausiai mišinių analizėje parametrų vertinimui ir imties klasterizavimui naudojamas rekurentinis *tikėtinumo maksimizavimo* (angl. *expectation maximization*, toliau *EM*), algoritmas. EM algoritmo savybės yra gerai ištirtos ir aprašytos, pavyzdžiui, [22], [180]. EM algoritmas maksimizuoja tikėtinumo funkciją, todėl gali būti naudojamas MTM įvertiniui apskaičiuoti. Deja, tikėtinumo funkcija turi daug lokalių ekstremumų, o EM algoritmas konverguoja lokaliai [136]. Taigi MTM įvertinį galima pasiekti tik parinkus pradinę vertinamo parametro reikšmę pakankamai arti tikrosios mišinio parametro reikšmės. Būtent šios pradinės reikšmės parinkimas vis dar yra aktuali, galutinai neišspręsta problema, kuri ir bus nagrinėjama disertaciniame darbe.

MTM įvertinys Gauso skirstinių mišinio modelio atveju yra asimptotiškai efektyvus. Taigi, turėdami pakankamai didelę imtį ir gerai parinkę pradinę parametro reikšmę, Gauso skirstinių mišinio modelio parametrų įvertinimo ir imties klasifikavimo uždavinius galėtume laikyti išspręstais. Tačiau didėjant duomenų dimensijai ir augant mišinio parametrų kiekiui asimptotiškai efektyvus įvertis gali būti pastebimai blogesnis (mažiau tikslus) už kai kuriuos paslinktus įverčius.

Imties stebėjimų projekcijos į vienamatę erdvę bus pasiskirsčiusios pagal vienamatį Gauso skirstinių mišinio modelį su daug mažesniu parametrų kiekiu, todėl MTM įvertinys gerai vertins šio vienamačio mišinio parametrus. Be to, vienamačiu atveju yra išspręstos ir kai kurios kitos mišinio analizės problemos. Pavyzdžiui, [142] yra pasiūlyta konstruktyvi procedūra pradinėms parametrų reikšmėms parinkti ir MTM įverčiams apskaičiuoti. Tiesa, šios procedūros tikslumas ir stabilumas didele dalimi priklauso nuo naudojamo neparametrinio tankio įverčių savybių. Todėl disertaciniame darbe yra pasiūlytas ir ištirtas nepara-

metrinis įvertinys, pritaikytas daugiamodalinių tankių su skirtingu lokaliu glodumu (šiomis savybėmis pasižymi Gauso skirstinių mišinių tankiai) vertinimui. Tad galime manyti, kad vienamačiu atveju tikslus MTM įvertis yra konstruktyviai apskaičiuojamas. Disertaciniame darbe yra tiriama duomenų projektavimo panaudojimo galimybė daugiamačio mišinio parametrams vertinti — pasiūlytos ir ištirtos procedūros, gerai vertinančios daugiamačio mišinio parametrus pagal vienamačių projekcijų parametrų MTM įverčius.

Projektavimo panaudojimo idėjomis disertacinis darbas išsiskiria iš kitų autorių darbų, skirtų mišinių modelių analizei. Paprastai projektavimas naudojamas ieškant mažesnio matavimo poerdvių išlaikančių klasterinę struktūra nusakančią informacija, pavyzdžiui, [132], [145]. Kitų autorių darbuose dažniausiai naudojami *projekcijų paieškos* (angl. *projection pursuit*) algoritmai ieškantys tam tikras savybes turinčių kelių projektavimo krypčių. Disertaciniame darbe nagrinėjami algoritmai pagrįsti projektavimu į didelį kiekį laisvai pasirinktų vienamačių krypčių. Ši metodologija reikalauja daug skaičiavimo kompiuteriu laiko ir tapo populiaria tik labai išaugus kompiuterių greitaeigiškumui. Artimiausios disertacijos idėjoms yra Friedman'o pasiskirstymo tankio vertinimo procedūros, naudojančios tikslinį projektavimą, aprašytos [68], [66], tačiau Friedman'o siūlomų metodų negalime tiesiogiai pritaikyti duomenims klasterizuoti, nes jo siūloma rekurentinė procedūra tankį vertina neparametriškai.

**Tikslas ir uždaviniai.** *Darbo tikslas* — konstruktyvių algoritmų, skirtų daugiamačių duomenų, tenkinančių Gauso skirstinių mišinio modelį, pasiskirstymo analizei ir klasifikavimui, sukūrimas.

*Darbo uždaviniai:* ištirti vienamačių duomenų projekcijų panaudojimo galimybę daugiamačių Gauso skirstinių mišinio pasiskirstymo tankiui vertinti bei duomenims klasterizuoti; sukurti procedūrą, pakankamai tiksliai parenkančią pradinę parametro reikšmę ir užtikrinančią rekurentinio EM algoritmo konvergavimą į MTM įvertį; ištirti, ar, naudojant duomenų projektavimą, mažų imčių atveju galima gauti tikslesnius įvertinius Gauso skirstinių mišinio parametrams vertinti, nei daugiamatis maksimalaus tikėtinumo metodo įvertinys.

**Naujumas.** Disertaciniame darbe pasiūlyti daugiamačio Gauso skirstinių mišinio parametrų identifikavimo ir duomenų klasterizavimo metodai, išsiskiria iš kitų autorių siūlomų metodų, naudojančių duomenų projektavimą.

1. Pasiūlytas ir ištirtas naujas neparametrinio vienamačio pasiskirstymo tankio įvertinimo metodas pritaikytas daugiamodaliniams tankiams vertinti.
2. Pasiūlyta ir ištirta nauja Gauso skirstinių mišinio pasiskirstymo tankio vertinimo pro-

cedūra, paremta duomenų projektavimu ir apvertimo formulės taikymu.

3. Pasiūlytos ir ištirtos naujos Gauso skirstinių mišinio parametrų identifikavimo procedūros, paremtos duomenų projektavimu ir mažiausių kvadratų metodu.
4. Pasiūlyta nauja geometrinė duomenų klasterizavimo procedūra paremta pseudoatstumais tarp imties taškų, kurie randami naudojantis projektuotų duomenų analize. Šios procedūros ir EM algoritmo derinys efektyviau identifikuoja Gauso skirstinių mišinius nei žinomos konstruktyvios MTM procedūros.
5. Pasiūlyta ir ištirta nauja daugiamačių Gauso skirstinių mišinio identifikavimo metodika, paremta duomenų projektavimo, pasiūlyto geometrinio klasterizavimo bei EM algoritmų ir mažiausių kvadratų metodo idėjų derinimu. Mažų imčių atveju kai kurioms parametrų reikšmėms įverčiai, gauti taikant šią metodiką, yra tikslesni (tirtų paklaidų prasme) už MTM įverčius.

**Pagrindiniai rezultatai.**

1. Pasiūlyta ir išnagrinėta neparametrinė vienamačio pasiskirstymo tankio statistinio įvertinimo procedūra, naudojanti branduolinį įvertį su adaptyviai parenkamu kintamu branduolio pločiu bei atliekanti įverčio poslinkio multiplikatyvią korekciją, tiksliau vertina Gauso skirstinių mišinio tankį, nei žinomuose statistiniuose paketuose naudojamos neparametrinės vertinimo procedūros, jei tankio lokalaus glodumo charakteristikos labai priklauso nuo argumento.
2. Maksimalaus tikėtinumo metodas daugelyje tirtų atvejų geriau nei kiti metodai tinka vertinti klasifikavimo tikimybėms, tačiau tirtais atvejais geometrinė klasterizavimo procedūra ($G$) dažnai geriau vertino Gauso skirstinių mišinio pasiskirstymo tankį esant pasirinktiems imties tūriams, jei tikslumo matu laikėme klaidos normą erdvėje $L_2$.
3. Sudėtinė procedūra ($G-EM$), pirmame etape naudojanti stebėjimų vienamačių projekcijų skirstinio parametrų statistinį įvertinimą, antrame etape pagal šiuos įverčius apibrėžianti pseudo-atstumą daugiamatėje erdvėje ir atliekanti geometrinį stebėjimų klasterizavimą, o trečiame etape taikanti $EM$ algoritmą, yra konstruktyvus ir patikimas būdas maksimalaus tikėtinumo metodo įverčiams apskaičiuoti.
4. Mažų imčių atveju, 3-iame punkte aprašyta procedūra, patikslinta mažiausių kvadratų metodu projektuotiems tankiams ($LSD$), tam tikroms parametrų reikšmėms duodavo geresnius rezultatus (visų matuotų paklaidų atžvilgiu), nei maksimalaus tikėtinumo metodas.

**Raktiniai žodžiai:** Gauso skirstinių mišinys, klasterizavimas, projektavimas, EM algoritmas, branduolinis tankio įvertis.

**Rezultatų aprobacija.** Diseracinio darbo tematika yra skaityta 15 pranešimų Lietuvos ir tarptautinėse mokslinėse konferencijose. Konferencijų pranešimų sąrašas pateiktas disertacijos pabaigoje. Taip pat skaityti pranešimai Kauno technologijos universiteto Taikomosios matematikos katedros, bei Matematikos ir informatikos instituto Taikomosios statistikos skyriaus seminaruose.

Disertacinio darbo tematika atspausdintas 1 straipsnis leidinyje, įtrauktame į Mokslinės informacijos instituto duomenų bazę, 4 straipsniai leidiniuose įrašytuose į Lietuvos mokslo ir studijų departamento patvirtintą sąrašą, 4 straipsniai kituose tarptautiniuose ir užsienio recenzuojamuose leidiniuose ir 1 straipsnis kituose leidiniuose. Straipsnių sąrašas pateikiamas disertacinio darbo pabaigoje.

**Disertacinio darbo struktūra.** Disertacija susideda iš įvado, keturių skyrių, išvadų, literatūros sąrašo.

Pirmajame skyriuje detalizuojamas tiriamas duomenų modelis, apžvelgiami kitų autorių darbai. Antras skyrius skirtas vienamačio pasiskirstymo tankio adaptyviam neparametriniam įvertinimui. Šiame skyriuje pasiūlyti metodai toliau naudojami daugiamačių duomenų analizėje, paremtoje projektavimu į vienamatę erdvę. Trečiame skyriuje pateikti siūlomi daugiamačių Gauso skirstinių mišinio analizės metodai. Ketvirtas skyrius skirtas eksperimentinių tyrimų metodikai bei tyrimų rezultatams apžvelgti. Darbo gale pateikiami straipsnių ir konferencijų pranešimų disertacijos tematika sąrašai.

# 1 Skyrius. Tyrimų apžvalga

## 1.1 Gauso skirstinių mišinio modelis ir jo identifikavimas

Sakysime, kad atsitiktinis vektorius $X \in \mathbb{R}^d$ tenkina Gauso (normaliųjų) skirstinių mišinio (toliau, trumpumo dėlei *Gauso mišinio*) modelį, jeigu jo pasiskirstymo tankis $f(x)$ tenkina lygybę

$$f(x) = \sum_{j=1}^{q} p_j \varphi_j(x) \stackrel{def}{=} f(x, \theta). \tag{1.1}$$

Parametrą $q$ vadinsime mišinio klasterių (klasių, komponenčių) kiekiu, o $p_j$ — apriorinėmis tikimybėmis. Jos tenkina sąlygas

$$p_j > 0, \quad \sum_{j=1}^{q} p_j = 1. \tag{1.2}$$

Formulėje (1.1) funkcija $\varphi_j(x)$ yra Gauso *pasiskirstymo tankio funkcija* (toliau, trumpumo dėlei *tankis*) su parametrais vidurkiu $M_j$ ir kovariacijų matrica $R_j$

$$\varphi_j(x) = \varphi(x; M_j, R_j) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|R_j|}} \, e^{-\frac{(x-M_j)' R_j^{-1} (x-M_j)}{2}}, \tag{1.3}$$

o $\theta = (p_j, M_j, R_j, j = 1, \ldots, q)$ — daugiamatis modelio parametras.

Tarkime, turime atsitiktinę nepriklausomų vienodai pasiskirsčiusių atsitiktinių vektorių imtį

$$\mathbb{X} = \{X_1, X_2, \ldots, X_n\}. \tag{1.4}$$

Sakysime, kad imtis tenkina Gauso mišinio modelį, jeigu $X_i$ tenkina Gauso mišinio modelį. Dydį $n$ vadinsime imties dydžiu (tūriu).

Vienas iš statistinių uždavinių yra stebimo atsitiktinio dydžio tankio įvertinimas. Gauso tankis gali būti įvairaus glodumo priklausomai nuo jo parametro - kovariacinės matricos. Kadangi Gauso mišinio komponenčių kovariacinės matricos gali būti skirtingos, jo tan-

kis gali turėti labai skirtingas glodumo savybes skirtingose srityse. Todėl neparametriškai vertinant tankį reikia naudoti adaptyvius įvertinius — sugebančius prisitaikyti prie tankio lokalių glodumo savybių. Parametriškai vertinant tankį, reikia įvertinti skirstinio daugiamačio parametro $\theta$ reikšmę, o tai nėra paprastas uždavinys, nes, didėjant dimensijai $d$, parametrų kiekis greitai auga

$$\dim\theta = q\frac{d(d+1)}{2} + qd + d - 1, \tag{1.5}$$

pavyzdžiui, net esant nedidelei dimensijai $d = q = 5$, modelį nusakys $\dim\theta = 104$ parametrai, ir, ieškant parametrų įverčių, gali tekti spręsti optimizavimo uždavinį 104-matėje erdvėje. Praktikoje klasterių kiekis $q$ taip pat gali būti nežinomas, tuomet jį taip pat reikės įvertinti. Buvo mėginta mišinio parametrus vertinti momentų metodu [44], bet pasirodė, kad toks parametrų įvertis nėra geras. Sprendžiant uždavinį momentų metodu dažniausiai apsiribojama atveju, kai mišinį sudaro du klasteriai su vienodomis kovariacinėmis matricomis. Vėliau mišinio parametrams vertinti pradėtas naudoti EM algoritmas.

Gauso mišinio modelis dažnai naudojamas klasterizavimo uždaviniuose. Praktikoje klasės, kuriai priklauso objektas, nustatymas yra dažnesnis ir svarbesnis uždavinys, negu tankio ar parametrų įvertinimas. Praplėsime Gauso mišinio modelio apibrėžimą ir parodysime glaudų parametrų identifikavimo ir klasterizavimo uždavinių ryšį.

Tegu $O(1), \ldots, O(n)$ stebimi objektai. Juos atitinka atsitiktinės poros $(X, \nu)$, kur $X$ — stebimas atsitiktinis požymių vektorius, o $\nu$ nestebimas klasės, kuriai priklauso objektas, numeris. Tegu kiekvienos klasės viduje požymių vektorius $X$ tenkina Gauso skirstinio modelį, tuomet sakysime, kad $X$ tenkina Gauso mišinio modelį. $\varphi_j(x)$ bus sąlyginis $X$ tankis prie sąlygos $\nu = j$, o $f(x)$ apibrėžtas (1.1) bus nesąlyginis $X$ tankis. *Klasterizavimo* (klasifikavimo be apmokymo imčių) uždavinys siekia įvertinti stebėto objekto klasės numerį, t.y. rasti įvertį $\widehat{\nu}(x) = \widehat{\nu}(x, \mathbb{X}) \in \{1, \ldots, q\}$. Toks klasterizavimas, priskiriantis stebėjimą vienai klasei, vadinamas *griežtu*. Klasterizavimas vadinamas *negriežtu*, jeigu jis parodo stebėjimo priklausymo klasteriams tikimybes, t.y. turime rasti aposteriorinių tikimybių

$$\pi(j, x) = \mathbb{P}\{\nu(x) = j | X = x\}, \quad j = \overline{1, q} \tag{1.6}$$

įverčius. Gauso mišinio modelio atveju iš Bajeso formulės seka, kad

$$\pi(j, x) = \frac{p_j\varphi_j(x)}{f(x, \theta)} \stackrel{def}{=} \pi_\theta(j, x). \tag{1.7}$$

Šio atveju klasės tikimybė

$$\pi_j = \mathbb{P}\{\nu(X) = j\}. \tag{1.8}$$

Formulė (1.7) sieja aposteriorines klasifikavimo tikimybes ir modelio parametrus, taigi kartu parodo glaudų parametrų identifikavimo ir klasterizavimo uždavinių ryšį. Turėdami negriežtą klasterizavimo taisyklę galime lengvai sukonstruoti griežto klasterizavimo taisyklę

$$\widehat{\nu}(x) = \arg\max_{j=\overline{1,q}} \widehat{\pi}(j, x). \tag{1.9}$$

Negriežto klasifikavimo tikimybių vertinimui galima taikyti *tiesioginio pakeitimo* (angl. *plug-in*) principą, formulėje (1.7) parametrą $\theta$ keičiant į jo įvertį $\widehat{\theta}$

$$\widehat{\pi}(j, x) = \pi_{\widehat{\theta}}(j, x) = \frac{\widehat{p}_j \widehat{\varphi}_j(x)}{f(x, \widehat{\theta})}. \tag{1.10}$$

Natūralu būtų taikyti maksimalaus tikėtinumo metodą (juo gautas rezultatas yra asimptotiškai efektyvus Gauso mišinio modelio atveju) šiam įverčiui gauti

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} l(\theta), \tag{1.11}$$

kur

$$l(\theta) \stackrel{def}{=} \sum_{k=1}^{n} \ln f(X(k), \theta) \tag{1.12}$$

yra logaritmuota tikėtinumo reikšmė.

Jei duomenų dimensija $d$ yra didelė, praktinis $\widehat{\theta}_{MLE}$ įverčio radimas yra sudėtingas.

**EM algoritmas.** Paprastai MTM įvertiniui rasti naudojamas EM algoritmas. Šis algoritmas Gauso mišinio analizei buvo nepriklausomai pasiūlytas kelių autorių: Hasselbland 1966 [79], Behboodian 1970 [9]. Vėliau jo savybės buvo gerai išnagrinėtos [180], [22] ir kituose darbuose. EM algoritmui skirta daug dėmesio įvairiuose apžvalginiuose straipsniuose ir monografijose, pvz., [45], [136], [59], [168], [110], [181]. EM algoritmas rekurentiškai

perskaičiuoja $\widehat{\pi}$ ir $\widehat{\theta}$ įverčius, naudodamas formulę (1.10) ir

$$\begin{aligned} \widehat{p}_j &= \sum_{x\in\mathbb{X}} \frac{\widehat{\pi}(j,x)}{n}, \\ \widehat{M}_j &= \sum_{x\in\mathbb{X}} \frac{\widehat{\pi}(j,x)}{\widehat{p}_j n} x, \\ \widehat{R}_j &= \sum_{x\in\mathbb{X}} \frac{\widehat{\pi}(j,x)}{\widehat{p}_j n} xx' - \widehat{M}_j \widehat{M}_j'. \end{aligned} \tag{1.13}$$

Deja, EM algoritmas konverguoja lokaliai ir jis konverguoja į statistiką (1.11) tik jeigu pradinė $\theta$ reikšmė yra arti $\widehat{\theta}_{MLE}$. Taigi, pradinės reikšmės pasirinkimo problema yra esminė.

Pradinių reikšmių parinkimas nėra galutinai išspręsta problema. Galimus sprendimo būdus galime suskaidyti į tokias grupes:

- **Atsitiktinis pradinės reikšmės parinkimas.** Tai vienas iš pirmųjų pasiūlytų būdų EM algoritmui inicializuoti. Deja, toks parinkimas negarantuoja patenkinamų rezultatų, ir tikimybė parinkti tinkamą pradinį tašką greit mažėja augant dimensijai $d$ [44]. Naudojant šį būdą paprastai algoritmas paleidžiamas keletą kartų iš įvairių atsitiktinių reikšmių, o geriausias rezultatas išsirenkamas naudojant tikėtinumo funkciją (1.12).
- **$k$-vidurkių metodas, kiti euristiniai klasterizavimo metodai.** Tai vienas iš dažnai naudojamų parinkimo būdų. Kadangi šie pradinio taško parinkimo metodai nėra tiesiogiai skirti Gauso mišiniams jie gerai veikia tik Gauso mišiniui tenkinant papildomas sąlygas, pvz., $k$-vidurkių metodas tinka EM algoritmui inicializuoti, kai klasterių kovariacinės matricos skiriasi nereikšmingai.
- **Nuoseklus klasterių išskyrimas.** Ši metodika skiriasi tuo, kad ji ne vykdo EM algoritmą nuo tam tikro pradinio taško, bet palaipsniui išskiria klasterius iš mišinio, ir, kiekvieną kartą, išskyrus naują klasterį, parametrai yra tikslinami EM algoritmu. Prie tokių metodų galima priskirti [142] aprašytą procedūrą, išskiriančią klasterius, kaip sritis, turinčias didesnę neparametrinio tankio įverčio reikšmę. Kitas nuoseklaus klasterių išskyrimo metodas paskelbtas darbe [172].
- **Hierarchinis Gauso mišinio klasifikavimas.** Hierarchiniai klasifikavimo metodai yra gerai žinomi, bet paprastai jie nebūna pritaikyti Gauso mišinių modelį tenkinantiems duomenimis. Pirmąkart toks klasifikavimo algoritmas, išlaikantis Gauso mišinio modelį, sudarant klasterių medį buvo aprašytas Fraley 1999 metais [61]. Taigi,

tai naujas ir gerai veikiantis algoritmas. Bene vienintelis iki šiol pastebėtas svarbus jo trūkumas yra tas, kad jis reikalauja daug skaičiavimo laiko ir atminties, jeigu duomenų kiekis yra didelis.

**Klasterizavimo programinė įranga.** Nežiūrint didelio publikuotų klasterizavimo procedūrų ir algoritmų kiekio, didesnioji jų dalis - duomenų kapstymosi (angl. *data mining*) tipo algoritmai, ir jie nėra pritaikyti Gauso mišinio modeliui. Daugelyje žinomų paketų realizuotas $k$-vidurkių, hierarchinio klasifikavimo (artimiausio kaimyno, tolimiausio kaimyno, vidutinio atstumo) ir kiti metodai, bet procedūrų, skirtų Gauso mišinio modelį tenkinantiems duomenims, taip pat pasigesime. Yra vos keletas programinių paketų, skirtų Gauso mišiniams klasterizuoti. Visi jie naudoja EM algoritmą mišinio parametrams rasti ir skiriasi tik pradinių reikšmių parinkimo būdais, klasterių kiekio nustatymo algoritmu:

- **NORMIX.** Šis paketas sukurtas dar 1967 metais [178]. Pradines parametrų reikšmes parenka atsitiktinai.
- **EMMIX.** 1999 m. sukurtas paketas, jis aprašytas [111]. Pradines parametrų reikšmes parenka atsitiktinai arba naudoja $k$-vidurkių metodą joms rasti. $k$-vidurkių algoritmas taip pat inicializuojamas pradedant nuo atsitiktinių centrų.
- **MCLUST.** Tai papildoma biblioteka prie paketo R (arba komercinės jo versijos S-Plus). Biblioteka naudoja hierarchinį Gauso mišinio klasifikavimą pradinėms EM algoritmo reikšmėms rasti. Papildomai galima apriboti klasterių kovariacinę struktūrą, pasirenkant sferinį, diagonalų, elipsoidinį modelį; modelį su vienodom ar skirtingom visų klasterių kovariacinėm matricom [61]. Klasterių kiekis parenkamas naudojant Bajeso informacinį kriterijų.

## 1.2 Pasiskirstymo analizės metodai, naudojantys duomenų projektavimą

Didėjant duomenų dimensijai, modelio parametrų kiekis sparčiai auga, mažiau pasireiškia maksimalaus tikėtinumo įvertinio asimptotinės savybės, sunkiau surasti pradines parametrų reikšmes EM algoritmui inicializuoti. Praktikoje galimos ir tokios situacijos, kai modelis tampa neidentifikuojamu, nes parametrų skaičius yra didesnis už imties dydį. Tokiu atveju, norint taikyti Gauso mišinio modelį, būtina mažinti modelio dimensiją. Vienas iš galimų dimensijos mažinimo būdų yra *diskriminantinės erdvės* panaudojimas.

**Diskriminantinės erdvė panaudojimas.** Diskriminantinės erdvės panaudojimas Gauso mišinio modelį tenkinantiems duomenims klasterizuoti yra aprašytas straipsniuose [143], [144], [145] bei disertacijoje [99]. Apibrėžkime diskriminantinę erdvę. Nemažindami bendrumo laikysime, kad $\mathbb{E}X = 0$ ir $\mathrm{cov}(X, X) = \mathbf{I}_d$, jei taip nėra, duomenis standartizuosime. Vektoriaus $u \in \mathbb{R}^d$ projekciją į tiesinį poerdvį $H \subset \mathbb{R}^d$ žymėsime $u_H$. *Diskriminantine erdve* $H$ vadinsime tiesinį poerdvį $H \subset \mathbb{R}^d$, tenkinantį sąlygą $\mathbb{P}\{\nu = i|X = x\} = \mathbb{P}\{\nu = i|X_H = x_H\}$, $i = 1, \ldots, q$, $x \in \mathbb{R}^d$ ir turintį mažiausią dimensiją.

Tegu $k \stackrel{def}{=} \dim H$, o vektoriai $u_1, u_2, \ldots, u_k$ sudaro ortonormuotą diskriminantinės erdvės bazę. Pažymėkime $U = (u_1, \ldots, u_k)$, tada $\pi(i, x) = \mathbb{P}\{\nu = i|U'X = U'x\}$, $\forall x \in \mathbb{R}^d$, $i = 1, \ldots, q$. Taigi $(U'X_1, \ldots, U'X_n)$ yra pakankama statistika $\pi(\cdot, \cdot)$ įvertinimui, t.y. projektuodami duomenis į diskriminantinę erdvę neprarandame informacijos apie klasterinę duomenų struktūrą. Tuomet atsitiktinio dydžio $UX$ skirstinys bus Gauso mišinys su tankiu

$$f^H(z) = \sum_{j=1}^{q} p_j \varphi_j^H(z) = f^H(z, \theta^H), \quad z \in \mathbb{R}^k, \tag{1.14}$$

čia $\varphi_j^H(z) = \varphi(z, M_j^H, R_j^H)$ yra $k$-mačio Gauso skirstinio tankis su vidurkiu $M_j^H = U'M_j$ ir kovariacine matrica $R_j^H = U'R_jU$, $\theta^H = (p_j, M_j^H, R_j^H)$ — daugiamatis parametras.

Tegul $\mathcal{F}$ žymi vienamačių standartizuotų Gauso mišinių pasiskirstymo funkcijų aibę, o $\Phi$ — standartinę normalinę pasiskirstymo funkciją. Tegul $\rho$ žymi funkcionalą su apibrėžimo sritimi $\mathcal{F} \times \mathcal{F}$ ir $\forall G, \Psi \in \mathcal{F}$, tenkinantį sąlygas:

$$\rho(G, \Psi) > 0, \quad \text{jei } G \neq \Psi, \tag{1.15}$$

$$\rho(G, G) = 0, \tag{1.16}$$

$$\rho(G * \Phi, \Psi * \Phi) < \rho(G, \Psi), \quad \text{jei } G \neq \Psi, \tag{1.17}$$

$$\rho\left(G\left(\frac{\cdot + \lambda}{c}\right), \Psi\left(\frac{\cdot + \lambda}{c}\right)\right) = \rho(G, \Psi), \quad \forall c > 0, \lambda \in \mathbb{R}, \tag{1.18}$$

čia $*$ žymi sąsūką. Šias sąlygas tenkinančių funkcionalų pavyzdžiais yra Kolmogorovo ir Helingerio atstumai.

Tegu $F_u$ žymi atsitiktinio dydžio $u'X$ pasiskirstymo funkciją. Tuomet teorema (žr. [144]) teigia, kad jei funkcionalas $\rho$ tenkina savybes (1.15)-(1.18), tai vektoriai $u_1, \ldots, u_k$

tenkinantys sąlygas

$$u_1 = \arg\max_{|\tau|=1} Q_1(\tau), \tag{1.19}$$

$$u_i = \arg\max_{\substack{\tau\perp \mathrm{span}(u_1,\ldots,u_{i-1}) \\ |\tau|=1}} Q_i(\tau), \quad i=2,\ldots,k, \tag{1.20}$$

sudaro ortonormuotą erdvės $H$ bazę, kur

$$Q_i(\tau) = \max_{v\in\{0,\pm u_1,\ldots,\pm u_{i-1}\}} \rho(F_{\tau+v}, \Phi * F_v). \tag{1.21}$$

Jei papildomai apribosime kovariacines matricas $R_1 = R_2 = \ldots = R_q$, tuomet $u_1,\ldots,u_k$ sudarys $H$ bazę prie sąlygos

$$u_i = \arg\max_{\substack{\tau\perp \mathrm{span}(u_1,\ldots,u_{i-1}) \\ |\tau|=1}} Q(\tau), \quad i=1,\ldots,k, \tag{1.22}$$

kur

$$Q(\tau) = \rho(F_\tau, \Phi). \tag{1.23}$$

Diskriminantinė erdvė paprastai randama naudojant *tikslinį projektavimą* (angl. *projection pursuit*). Tikslinio projektavimo metodai ieško "įdomių" krypčių daugiamatėje erdvėje. Krypčių "įdomumą" apibrėžia tikslo funkcija, kuri vadinama *projektavimo indeksu*. Mūsų nagrinėjamu atveju projektavimo indeksas yra atstumas $\rho(\cdot,\Phi)$, taigi ieškosime krypčių į kurias suprojektuotų duomenų pasiskirstymo funkcija labiausiai skiriasi nuo standartinės Gauso pasiskirstymo funkcijos. Taigi, diskriminantinės erdvės įvertį apibrėšime

$$\widehat{H} = \mathrm{span}(\widehat{u}_1,\ldots,\widehat{u}_k), \tag{1.24}$$

kur

$$\widehat{u}_i = \arg\max_{\substack{\tau\perp \mathrm{span}(\widehat{u}_1,\ldots,\widehat{u}_{i-1}) \\ |\tau|=1}} \widehat{Q}(\tau), \quad i=1,\ldots,k, \tag{1.25}$$

$$\widehat{Q}(\tau) = \rho(\widehat{F}_\tau, \Phi). \tag{1.26}$$

Jeigu diskriminantinės erdvės dimensija $k$ nežinoma, ji vertinama pavyzdžiui

$$\widehat{k} = \min\{i : \widehat{Q}(\widehat{u}_i) < \alpha\}. \tag{1.27}$$

**Tankio vertinimas naudojant tikslinį projektavimą.** Tikslinio projektavimo metodiką pirmą kartą pasiūlė J.B. Kruskal 1969 metais [100]. Vėliau šią metodiką išplėtojo J.H. Friedman ir bendraautoriai darbuose [69], [68], [66], beje paskutiniai du darbai yra skirti daugiamačio tankio vertinimui, panaudojant tikslinį projektavimą. Šių darbų tematika yra artimiausia disertacinio darbo tematikai. Trumpai aprašysime J.H. Friedman siūlomą tankio įvertinimo metodą.

Šis metodas, kaip kitų autorių siūlomi metodai, naudoja projektavimo indeksą, kuris lygina projektuotų duomenų skirstinį su Gauso skirstiniu. Projekcijos, turinčios Gauso skirstinį, laikomos mažiausiai "įdomiomis". Tai yra grindžiama:

- Daugiamatis Gauso skirstinys yra pilnai apibrėžtas savo tiesinės struktūros (vidurkio ir kovariacijų matricos), o mes norime apčiuopti duomenų struktūrą, kuri nepriklausytų nuo koreliacinės duomenų struktūros ir tiesinių transformacijų, pvz., mastelio parametro.
- Visos daugiamačio Gauso skirstinio projekcijos yra taip pat Gauso skirstiniai. Taigi jeigu tikslinio projektavimo būdu rasta projekcija nereikšmingai skirsis nuo Gauso Skirstinio, tai rodys, kad ir daugiamatis duomenų skirstinys yra artimas Gauso skirstiniui.
- Daugiamačių duomenų, turinčių struktūrą keliose projektavimo kryptyse, t.y. kur projekcijos turi skirstinius labai besiskiriančius (kokio nors projektavimo indekso prasme) nuo Gauso, daugelis projekcijų turės skirstinį artimą normaliajam. Šis teiginys seka iš centrinės ribinės teoremos.
- Esant fiksuotai dispersijai, Gauso skirstinys neša mažiausiai informacijos.

Autoriai siūlo tokią projektavimo indekso konstrukciją. Tarkime, kad stebimo atsitiktinio dydžio $X$ projekcija kryptimi $\tau$ turi Gauso skirstinį. Tuomet atsitiktinis dydis

$$R = 2\Phi(\tau' X) - 1 \tag{1.28}$$

bus tolygiai pasiskirstęs intervale [-1;1], ir jo tankio reikšmė šiame intervale bus lygi $\frac{1}{2}$. $R$ netolygumo matu laikykime integruotą kvadratinę paklaidą

$$\int_{-1}^{1} \left( f_R(r) - \frac{1}{2} \right)^2 dr = \int_{-1}^{1} f_R^2(r)\, dr - \frac{1}{2}, \tag{1.29}$$

čia $f_R$ yra $R$ tankio funkcija. Siūlomas projektavimo indeksas yra išraiškos (1.29) aproksi-

macija. Išskleiskime $f_R$ Ležandro polinomais

$$\int_{-1}^{1} f_R^2(r)\,dr - \frac{1}{2} = \int_{-1}^{1} \left( \sum_{j=0}^{\infty} a_j L_j(r) \right) f_R(r)\,dr - \frac{1}{2}, \tag{1.30}$$

kur Ležandro polinomai apibrėžiami

$$\begin{aligned} &L_0(r) = 1, \quad L_1(r) = r, \\ &L_j(r) = \frac{(2j-1)rL_{j-1}(r) - (j-1)L_{j-2}(r)}{j}, \quad j \geq 2, \end{aligned} \tag{1.31}$$

o koeficientai

$$a_j = \frac{2j+1}{2} \int_{-1}^{1} L_j(r) f_R(r)\,dr = \frac{2j+1}{2}\,\mathbb{E}L_j(r). \tag{1.32}$$

Taigi,

$$\int_{-1}^{1} f_R^2(r)\,dr - \frac{1}{2} = \sum_{j=1}^{\infty} \frac{2j+1}{2}\,\mathbb{E}^2 L_j(r). \tag{1.33}$$

Projektavimo indeksas yra apibrėžiamas

$$I(\tau) = \sum_{j=1}^{J} \frac{2j+1}{2}\,\mathbb{E}^2 L_j(2\Phi(\tau' X) - 1), \tag{1.34}$$

o iš imties jis vertinamas

$$\widehat{I}(\tau) = \sum_{j=1}^{J} \frac{2j+1}{2} \left( \frac{1}{n} \sum_{i=1}^{n} L_j(2\Phi(\tau' X(i)) - 1) \right)^2. \tag{1.35}$$

Pastebėsime, kad begalinę suma pakeitėme baigtine. Toks pakeitimas turi privalumų: suma tapo greičiau skaičiuojama, be to tai suteikia projektavimo indeksui robastiškumo, nes sumuojant tik baigtinį skaičių narių, lėtai gęstančios projekcijų skirstinių "uodegos" mažiau įtakos projektavimo indekso reikšmę. Si ūloma naudoti $4 \leq J \leq 8$.

Naudodami projektavimo indeksą (1.35) ieškosime "įdomių" duomenų projekcijų. Tačiau dažniausiai nepakanka rasti vienos projekcijos, kad pakankamai tiksliai įvertintume daugiamatį tankį. Ieškodami diskriminantinės erdvės, kiekvienos sekančios krypties ieškojome ortogonalios ankščiau rastoms, tačiau bendru atveju "įdomios" kryptys neb ūtinai turi

būti ortogonalios ir gali tekti naudoti didesnį projektavimo krypčių kiekį, nei duomenų dimensija. Todėl vertinant tankį tikslinio projektavimo būdu yra naudojamas taip vadinamas duomenų struktūros panaikinimas. Jis atlieka netiesinę mastelio pakeitimo transformaciją rasta projektavimo kryptimi taip, kad transformuotų duomenų skirstinys šia kryptimi tampa normalusis, t.y. "neįdomus". Tai užtikrina, kad ieškodami sekančios projektavimo krypties nerasime ką tik rastos. Duomenų struktūros panaikinimas remiasi tuo, kad jeigu vienamatė duomenų projekcija $\tau' X$ turi pasiskirstymo funkciją $F_\tau$, tai tuomet atsitiktinis dydis $\Phi^{-1}(F_\tau(\tau' X))$ turės Gauso skirstinį. Čia $\Phi^{-1}$ žymi atvirkštinę Gauso pasiskirstymo funkciją. Tiksliai nusakysime daugiamačių duomenų transformavimo procedūrą. Tegu $U$ yra ortonormuota $d \times d$ matrica su projektavimo krypties vektoriumi $\tau$ pirmoje eilutėje. Tuomet $UX$ bus posūkio transformacija, kurią atlikus pirmoji koordinatė bus $\tau' X$. Tegul $T(V)$ yra vektorinė transformacija, kuri pirmąją koordinatę transformuoja taip, kad skirstinys tampų normaliuoju, o kitų vektoriaus koordinačių nekeičia, t.y.

$$T(V) = \begin{pmatrix} \Phi^{-1}(F_\tau(V_1)) \\ V_2 \\ \vdots \\ V_d \end{pmatrix}. \tag{1.36}$$

Tuomet transformacija

$$X' = U' T(UX) \tag{1.37}$$

transformuos daugiamačius duomenis taip, kad jų projekcija į kryptį $\tau$ turės Gauso skirstinį, o skirstiniai ortogonaliomis kryptimis išliks nepakitę.

Pastebėsime, kad jei stebėjimų tankis yra $f$, tai, radę pirmą kryptį su projekcijos skirstiniu mažiausiai panašiu į Gauso skirstinį ir atlikę

aukščiau aprašytą duomenų transformavimą, turėsime modifikuotus stebėjimus, kurių tankio funkcija yra

$$f_1(x^{(1)}) = f(x)\varphi(\tau_1')/f_{\tau_1}(\tau_1' x), \tag{1.38}$$

čia $\varphi$ žymi standartinio Gauso skirstinio, o $f_{\tau_1}$ — duomenų projekcijos $\tau_1' x$ tankius, viršutiniai indeksai skliausteliuose vektoriams rodo, kiek kartų jie buvo transformuoti naudojant transformaciją (1.37). Atlikę du žingsnius turėsime modifikuotus stebėjimus, kurių tankio

funkcija yra

$$f_2(x^{(2)}) = f_1(x^{(1)})\varphi(\tau_2' x^{(1)})/f_{\tau_2}^{(1)}(\tau_2' x^{(1)}) = f(x)\frac{\varphi(\tau_1' x)\,\varphi(\tau_2' x^{(1)})}{f_{\tau_1}(\tau_1' x)\,f_{\tau_2}^{(1)}(\tau_2' x^{(1)})}, \tag{1.39}$$

čia tankio funkcijų viršutiniai indeksai rodo kiek kartų transformuotų stebėjimų projekcijas jos atitinka, pvz., $f_{\tau_3}^{(2)}$ yra atsitiktinio dydžio $\tau_3' X^{(2)}$ tankis, kur $X^{(3)}$ rekurentiškai apibrėžiamas naudojantis lygybe

$$X^{(k)} = U_{k-1}' T_{k-1}(U_{k-1} X^{(k-1)}). \tag{1.40}$$

Atlikę $k$ žingsnių turėsime duomenis su tankiu

$$f_k(x^{(k)}) = f(x)\prod_{i=1}^{k}\frac{\varphi(\tau_i' x^{(i-1)})}{f_{\tau_i}^{(i-1)}(\tau_i' x^{(i-1)})}, \tag{1.41}$$

Jei, po $k$ žingsnių, tikslinio projektavimo procedūra neranda krypties, kurioje transformuota projekcija reikšmingai skiriasi nuo Gauso skirstinio, tai transformuotų stebėjimų skirstinys artimas daugiamačiam Gauso skirstiniui. Tuomet $f_k$ keisdami Gauso tankiu, išsireiškę $f$ ir nežinomus tankius keisdami jų įverčiais, gausime

$$\widehat{f}(x) = \varphi(x^{(k)})\prod_{i=1}^{k}\frac{\widehat{f}_{\tau_i}^{(i-1)}(\tau_i' x^{(i-1)})}{\varphi(\tau_i' x^{(i-1)})}. \tag{1.42}$$

# 2 Skyrius. Adaptyvus vienamačio tankio neparametrinis statistinis įvertinimas

## 2.1 Branduoliniai įvertiniai ir jų savybės

Tegu $X$ yra atsitiktinis dydis su nežinomu tankiu $f(x)$, o $\mathbb{X} = (X_1, \ldots, X_n)$ žymi $n$ dydžio imtį, sudarytą iš nepriklausomų $X$ kopijų. *Branduolinis tankio įvertis* $\widehat{f}_h(x)$ apibrėžiamas

$$\widehat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \tag{2.1}$$

kur

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right). \tag{2.2}$$

Čia $K(h)$ yra *branduolio formos funkcija*, tenkinanti sąlygą

$$\int K(x)\, dx = 1, \tag{2.3}$$

o $h$ — *branduolio plotis*.

**Pastaba:** bendruoju atveju sąlyga (2.3) turi būti tenkinama integruojant visoje realiųjų skaičių aibėje, tačiau dažniausiai branduolys $K(x)$ įgyja reikšmes nelygias nuliui tik baigtiniame intervale. Todėl galime taip transformuoti branduolio funkciją, kad ji būtų nelygi nuliui tik intervale $(-1; 1)$. Tuomet (2.3) sąlygoje pakanka integruoti branduolį tik šiame intervale. Todėl, čia ir toliau, jei integravimo intervalas nenurodytas, tarsime, kad jis yra $(-1; 1)$.

Paprastai naudojami branduoliai tenkina ne tik (2.3) sąlygą, bet ir yra neneigiami, lyginiai. Jeigu norime, kad tankio įvertis būtų diferencijuojamas, turime parinkti diferencijuojamą branduolio funkciją, nes, kaip matome iš (2.1), tankio įverčiui būdingos branduolio

savybės. Kaip branduolio pavyzdį, galime pateikti Epaničnikovo branduolį:

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2), & \text{kai } |x| < 1 \\ 0, & \text{kai } |x| \geq 1 \end{cases} \tag{2.4}$$

Parinkti branduolio funkciją nėra sudėtinga. Yra keletas optimalių pagal įvairius kriterijus branduolio funkcijų. Todėl tankio įverčio kokybė labiausiai priklauso nuo branduolio pločio $h$, kurio parinkimas yra sudėtingesnis uždavinys. Jeigu parinksime per mažą branduolio plotį, tankio įvertis nebus glodus ir pasižymės santykinai didele dispersija. Jei parinksime per didelį branduolio plotį, tankio įvertis blogai atspindės tikrąją tankio formą, nes ją pernelyg suglodins. Be to, toks įvertis pasižymės dideliu poslinkiu. Vertindami tankį, galime naudoti tą patį branduolio plotį $h$ visoms $x$ reikšmėms arba parinkti branduolio plotį priklausomą nuo argumento $x$. Pirmuoju atveju branduolio plotį $h$ vadinsime *branduolio pločio parametru* arba tiesiog *parametru* $h$. Antruoju atveju, kai $h = h(x)$, vartosime terminą *branduolio pločio funkcija* arba *funkcija* $h$. Parametrą $h$ galime apskaičiuoti naudodami kryžminio patikrinimo metodą, tačiau jeigu tankio glodumas, esant įvairioms argumento reikšmėms, yra labai skirtingas, reikia naudoti branduolio pločio funkciją, nes taip galėsime prisitaikyti (adaptuotis) prie lokalių tankio funkcijos savybių. Dėl šios priežasties tankio įverčiai, kuriuose naudojamos branduolio pločio funkcijos, vadinami adaptyviaisiais arba įverčiais su lokaliai parenkamu glodinimo pločiu.

**Asimptotinės tankio įverčių savybės.** Išanalizuokime tankio įverčio, apibrėžto (2.1), asimptotines savybes. Tegu tenkinamos sąlygos

$$n \longrightarrow \infty, \quad h = h(x) \longrightarrow 0, \quad nh \longrightarrow \infty. \tag{2.5}$$

Sąlygos (2.5) yra pakankamos, kad įvertis b$\bar{\text{u}}$tų suderintas.

Imkime branduolio funkcijas, kurios yra lyginės ir tenkina šias sąlygas:

$$\begin{aligned} &\int x^j K(x)\,dx = 0, \quad \text{kai } j = 1, \ldots, k-1 \\ &\int x^k K(x)\,dx = c_1 \neq 0. \end{aligned} \tag{2.6}$$

ir panagrinėkime tankio įverčio poslinkio asimptotinį elgesį, kai tankis $f$ turi tolydinę $k$

eilės išvestinę.

$$
\begin{aligned}
b_h(x) &= \mathbb{E}\widehat{f}_h(x) - f(x) = \mathbb{E}K_h(x-X) - f(x) \\
&= \frac{1}{h}\int\limits_{-\infty}^{+\infty} K\left(\frac{x-y}{h}\right) f(y)\,dy - f(x) = \int K(y) f(x+hy)\,dy - f(x).
\end{aligned}
\tag{2.7}
$$

Skleisdami $f(x+hy)$ Teiloro eilute ir pasinaudodami (2.6), gauname

$$
\begin{aligned}
b_h(x) &= \int K(y)\left(f(x) + \frac{f'(x)h}{1!}y + \cdots + \frac{f^{(k)}(x)h^k}{k!}y^k + o(h^k)\right)dy - f(x) \\
&= \frac{f^{(k)}(x)h^k}{k!}\int K(y)y^k dy + o(h^k) = \frac{c_1 f^{(k)}(x)h^k}{k!} + o(h^k).
\end{aligned}
\tag{2.8}
$$

Taigi, įverčio poslinkis priklauso nuo branduolio pločio $h$ parinkimo. Tam, kad tankio įvertis būtų suderintas, branduolio plotis, turi tenkinti sąlygą $h = h(n) \overset{n\to\infty}{\longrightarrow} 0$. Esant šiai sąlygai tankio įverčio poslinkis tampa netiesiogiai priklausomas nuo imties turio $n$. Branduolio, apibrėžto (2.4), $k = 2$, o poslinkis yra lygus

$$
b_h(x) = \frac{c_1 f''(x)h^2}{2} + o(h^2), \quad c_1 = 0.2. \tag{2.9}
$$

Galima parinkti sudėtingesnius branduolius, kuriems (2.6) galiotų kai $k > 2$, ir tokiu būdu gauti poslinkius su aukštesnės eilės, nei $o(h^2)$, nykstančiais dydžiais, tačiau to neįmanoma padaryti su neneigiamomis branduolio funkcijomis. Naudojant branduolius, kurie įgyja neigiamas reikšmes, gaunami tankio įverčiai, kurie taip pat įgyja neigiamas reikšmes ir dėl šios priežasties nepriklauso pasiskirstymo tankio fukcijų klasei. Tokiu atveju reikia naudoti papildomas tankio įverčio transformacijas, kad būtų išvengta neigiamumo. Be to, tokių įverčių dispersija didesnė. Šiame darbe bus naudojamos tik neneigiamos branduolio funkcijos, kurios paprastai taikomos praktiniuose tyrimuose.

Analogiškai galime nustatyti tankio įverčio dispersijos asimptotines savybes.

$$
\begin{aligned}
\sigma_h^2(x) &= n^{-1}\mathbb{D}K_h(x-X) = n^{-1}\left(\mathbb{E}K_h^2(x-X) - \mathbb{E}^2 K_h(x-X)\right) \\
&= (nh)^{-1}\int K^2(y) f(x+hy)dy - n^{-1}\mathbb{E}^2\widehat{f}_h(x).
\end{aligned}
\tag{2.10}
$$

Po integralo ženklu skleisdami tankį Teiloro eilute gauname

$$\begin{aligned}\sigma_h^2(x) &= (nh)^{-1}\int K^2(y)\left(f(x)+O(h)\right)dy - n^{-1}\left(f(x)+O(h^2)\right)^2 \\ &= \frac{c_2 f(x)+O(h)}{nh} - \frac{f^2(x)+O(h^2)}{n} = \frac{c_2 f(x)}{nh} + O\left(\frac{1}{n}\right),\end{aligned} \tag{2.11}$$

kur

$$c_2 = \int K^2(x)dx. \tag{2.12}$$

Turėdami asimptotines tankio įverčio poslinkio ir dispersijos formules, galime rasti asimptotiškai optimalų branduolio plotį $h$. Paprastai stengiamasi parinkti $h$ taip, kad vidutinė kvadratinė paklaida būtų minimali. Todėl pažymėkime

$$h_{opt}(x) = \arg\min_h \mathbb{E}\left(\widehat{f}_h(x) - f(x)\right)^2. \tag{2.13}$$

Pasinaudodami (2.9) ir (2.11), galime apskaičiuoti asimptotinę vidutinę kvadratinę paklaidą

$$\mathbb{E}\left(\widehat{f}_h(x) - f(x)\right)^2 = b_h^2(x) + \sigma_h^2(x) \approx \frac{(c_1 f''(x))^2 h^4}{4} + \frac{c_2 f(x)}{nh} \tag{2.14}$$

Prilyginę (2.14) nuliui ir diferencijuodami pagal $h$, galime apskaičiuoti asimptotiškai optimalią branduolio pločio funkciją

$$h_{AS}(x) = \left(\frac{c_2 f(x)}{(c_1 f''(x))^2 n}\right)^{1/5}. \tag{2.15}$$

Norėdami rasti optimalų parametrą $h$, naudosime asimptotinės vidutinės kvadratinės paklaidos išraišką

$$\begin{aligned}\int_{-\infty}^{+\infty}\mathbb{E}\left(\widehat{f}_h(x) - f(x)\right)^2 &= \int_{-\infty}^{+\infty}\left(b_h^2(x)+\sigma_h^2(x)\right)dx \\ &\approx \frac{c_1^2 h^4 \int\limits_{-\infty}^{+\infty}(f''(x))^2 dx}{4} + \frac{c_2}{nh}.\end{aligned} \tag{2.16}$$

Tuomet asimptotiškai optimalus branduolio pločio parametras bus lygus

$$h_{AS} = \left( \frac{c_2 f(x)}{c_1^2 \|f''\|_2^2 n} \right)^{\frac{1}{5}}, \tag{2.17}$$

čia ir toliau $\|\cdot\|_p$ žymėsime normą funkcijų erdvėje $L_p$, $0 < p < \infty$, t.y.

$$\|f\|_p = \left( \int_{-\infty}^{+\infty} |f(x)|^p \, dx \right)^{1/p}. \tag{2.18}$$

## 2.2 Statistiniai metodai naudojami neparametriniame tankio įvertinime

**Parametrų parinkimas kryžminio patikrinimo b̄udu.** *Kryžminio patikrinimo metodas* (angl. *cross-validation*) pirmą kartą buvo paskelbtas Rudemo (1982) ir Bowmano (1984). Jis pagrįstas idėja, kad statistika, apskaičiuota naudojantis vienais imties elementais,

yra tikrinama, naudojant kitus imties elementus. Vienas iš populiariausių kryžminio patikrinimo metodų, vertinančių tankį, yra mažiausių kvadratų kryžminio patikrinimo metodas. Įverčio parametras ieškomas toks, kad minimizuotų integruotą kvadratinę paklaidą:

$$\begin{aligned} \alpha &= \arg\min_{\alpha} \int_{-\infty}^{+\infty} \left( \widehat{f}_\alpha(x) - f(x) \right)^2 dx \\ &= \arg\min_{\alpha} \left( \|\widehat{f}_\alpha\|_2^2 - 2 \int_{-\infty}^{+\infty} \widehat{f}_\alpha(x) f(x) \, dx + \|f\|_2^2 \right) \\ &= \arg\min_{\alpha} \left( \|\widehat{f}_\alpha\|_2^2 - 2 \int_{-\infty}^{+\infty} \widehat{f}_\alpha(x) \, dF(x) \right), \end{aligned} \tag{2.19}$$

čia $\alpha$ vertinamas parametras, o $F(x)$ stebimo atsitiktinio dydžio pasiskirstymo funkcija. Keisdami nežinomą pasiskirstymo funkciją į empirinę pasiskirstymo funkciją gausime parametro įverčio išraišką

$$\begin{aligned} \widehat{\alpha} &= \arg\min_{\alpha} \left( \|\widehat{f}_\alpha\|_2^2 - 2 \int_{-\infty}^{+\infty} \widehat{f}_\alpha(x) \, d\widehat{F}(x) \right) \\ &= \arg\min_{\alpha} \left( \|\widehat{f}_\alpha\|_2^2 - 2n^{-1} \sum_{i=1}^{n} \widehat{f}_\alpha(X_i) \right). \end{aligned} \tag{2.20}$$

Pastebėsime, kad vietoje (2.20) geriau naudoti formulę

$$\widehat{\alpha} = \arg\min_{\alpha} \left( \|\widehat{f}_\alpha\|_2^2 - 2n^{-1} \sum_{i=1}^{n} \widehat{f}_\alpha(X_i|i) \right), \tag{2.21}$$

čia $\widehat{f}_\alpha(x|i)$ įverčio reikšmę taške $x$, kuri apskaičiuojama pašalinus iš stebėjimų reikšmę $X_i$. Be to, empiriniai tyrimai rodo, kad kryžminio patikrinimo metode geriau ieškoti ne globalaus minimumo, o didžiausio lokalaus minimumo taško (žr. [95]).

Glodinimo pločio parametrą $h$ galima apskaičiuoti kryžminio patikrinimo metodu. Šiuo atveju $\alpha = h$. Apskaičiuokime minimizuojamos išraiškos (2.20) narius

$$\begin{aligned} \|\widehat{f}_h\|_2^2 &= \int_{-\infty}^{+\infty} \left( n^{-1} \sum_{i=1}^{n} K_h(x - X_i) \right)^2 dx \\ &= n^{-2}h^{-2} \int_{-\infty}^{+\infty} \left( \sum_{i,j=1}^{n} K\left( \frac{x - X_i}{h} \right) K\left( \frac{x - X_j}{h} \right) \right) dx \\ &= n^{-2}h^{-1} \sum_{i,j=1}^{n} \int_{-\infty}^{+\infty} K\left( \frac{X_i - X_j}{h} - x \right) K(x)\, dx \\ &= n^{-2}h^{-1} \sum_{i,j=1}^{n} K^*\left( \frac{X_i - X_j}{h} - x \right), \end{aligned} \tag{2.22}$$

kur $K^*(x)$ yra branduolio sąs$\bar{\text{u}}$ka su savimi

$$K^*(x) = \int K(x - s)K(s)\, ds. \tag{2.23}$$

Apskaičiuokime antrąjį (2.20) išraiškos narį:

$$\int_{-\infty}^{+\infty} \widehat{f}_\alpha(x)\, d\widehat{F}(x) = n^{-1} \sum_{i=1}^{n} \widehat{f}_\alpha(X_i) = (n^2 h)^{-1} \sum_{i,j=1}^{n} K\left( \frac{X_i - X_j}{h} \right) \tag{2.24}$$

Apskaičiuokime šio dydžio vidurkį:

$$\begin{aligned}\mathbb{E}\int_{-\infty}^{+\infty}\widehat{f}_\alpha(x)\,d\widehat{F}(x) &= \mathbb{E}(n^2h)^{-1}\sum_{i,j=1}^{n}K\left(\frac{X_i-X_j}{h}\right)\\ &= (hn)^{-1}(n-1)\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}K\left(\frac{x-y}{h}\right)f(x)f(y)\,dx\,dy+\frac{K(0)}{nh}\\ &= n^{-1}(n-1)\mathbb{E}\int_{-\infty}^{+\infty}\widehat{f}_\alpha(x)f(x)\,dx+\frac{K(0)}{nh}.\end{aligned} \tag{2.25}$$

Matome, kad šis įvertis yra paslinktas dydžiu $\frac{K(0)}{nh}\overset{h\to 0}{\longrightarrow}\infty$, todėl, skaičiuojant įvertį taške $X_i$, reikia išmesti elementą $X_i$ iš imties. Jeigu to nepadarytume, tai mažėjant $h$ reikšmei, antrasis narys artėtų į begalybę, o visa minimizuojama išraiška artėtų į $-\infty$. Taigi antrąjį formulės (2.20) narį skaičiuosime

$$\int_{-\infty}^{+\infty}\widehat{f}_\alpha(x)\,d\widehat{F}(x)\approx(n^2h)^{-1}\sum_{\substack{i,j=1\\ i\neq j}}^{n}K\left(\frac{X_i-X_j}{h}\right). \tag{2.26}$$

Kadangi branduolys $K(h)$ — lyginis, galime sumažinti sumuojamų narių kiekį ir išraišką (2.20) perrašome

$$\widehat{h}=\arg\min_h\left(\frac{2}{n^2h}\sum_{\substack{i,j=1\\ i<j}}^{n}\left(K^*\left(\frac{X_i-X_j}{h}\right)-2K\left(\frac{X_i-X_j}{h}\right)\right)+\frac{K^*(0)}{nh}\right). \tag{2.27}$$

Ieškodami $\widehat{h}(x)$ skaitiniu būdu mes randame funkcijos reikšmes tik tam tikroms argumento reikšmes, o norėdami rasti tarpinių taškų funkcijos reikšmes naudojame interpoliaciją. Todėl tankio įvertį, kuriam rasti naudojama branduolio pločio funkcija $\widehat{h}(x)$, galima laikyti tankio įverčiu su daugiamačiu parametru, sudarytu iš branduolio pločio funkcijos reikšmių. Tokiu atveju galėsime rasti adaptyvųjį tankio įvertį naudodamiesi kryžminio patikrinimo metodu. Kaip tik toks įvertis siūlomas [60]. Jis skaičiuojamas tokiu būdu:

1. fiksuojamos argumento reikšmės $x_1,\ldots,x_p$;
2. funkcija $h(x)$ apibrėžiama kaip trečios eilės splainas, einantis per taškus $(x_1,h_1),\ldots,(x_p,h_p)$;
3. aprašytu kryžminio patikrinimo būdu randame vektoriaus $(h_1,\ldots,h_p)$ įvertį.

**Artimiausių kaimynų metodas.** *Artimiausių kaimynų metodas* (angl. nearest neighbour) yra vienas iš metodų, kuris prisitaiko prie lokalių tankio savybių. Apibrėžkime $d_i(x)$ kaip atstumą nuo taško $x$ iki artimiausio $i$-tojo imties taško. Tuomet artimiausių kaimynų tankio įvertis apibrėžiamas

$$\widehat{f}(x) = \frac{k}{2nd_k(x)}. \tag{2.28}$$

Parametru $k$ galime keisti artimiausių kaimynų tankio įverčio glodumą. $k$ parenkamas žymiai mažesnis už imties dydį, paprastai naudojamas $k \approx n^{1/2}$. Norint tiksliau parinkti parametrą siūloma naudoti kryžminio patikrinimo metodą. Nors artimiausių kaimynų metodas yra adaptyvus ir sparčiai skaičiuojamas kompiuteriu, tačiau jis turi daug trūkumų. Šiuo būdu gautas tankio įvertis nepriklauso pasiskirstymo tankio fukcijų klasei, nes jo integralas nėra lygus 1. Imdami $x$ didesnį už didžiausią imties elementą, gausime $d_k(x) = x - X_{(n-k+1)}$. Čia $X_{(k)}$ žymime $k$-tąjį imties variacinės eilutės elementą. Todėl tokioms $x$ reikšmėms

$$\widehat{f}(x) = \frac{k}{2n(x - X_{(n-k+1)})}, \tag{2.29}$$

o $\int_{-\infty}^{+\infty} \widehat{f}(x) = \infty$, nes $\widehat{f}(x)$ "uodegos" gęsta $x^{-1}$ greičiu. Artimiausių kaimynų metodą galima apibendrinti

$$\widehat{f}(x) = \frac{1}{nd_k(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{d_k(x)}\right). \tag{2.30}$$

Taip apibrėžtas artimiausių kaimynų įvertis priklauso branduolinių įverčių klasei. Šiuo atveju $\widehat{f}(x)$ "uodegų" gesimo greitis, o kartu ir tankio įverčio integralo reikšmė, priklauso nuo branduolio funkcijos $K(x)$ parinkimo.

**Multiplikatyvus poslinkio koregavimas.** Branduoliniai tankio įverčiai, priklausomai nuo branduolio pločio, yra daugiau ar mažiau paslinkti. Jų paklaidas galima sumažinti taikant papildomus poslinkio sumažinimo metodus. Multiplikatyvus poslinkio koregavimas yra vienas iš tokių metodų. Tegu $\widehat{f}_h(x)$ yra branduolinis tankio įvertis apibrėžtas (2.1), o $K_h(x)$ — branduolio funkcija apibrėžta (2.2). Tuomet multiplikatyviai sumažinto poslin-

kio įvertis $\widetilde{f}(x)$ užrašomas formule

$$\widetilde{f}(x) = \frac{\widehat{f}_h(x)}{n} \sum_{i=1}^{n} \frac{K_h(x - X_i)}{\widehat{f}_h(X_i)}. \tag{2.31}$$

Trumpai paaiškinsime šią formulę. Tegu

$$\alpha(x) = \frac{f(x)}{g(x)}. \tag{2.32}$$

Imkime tokį $\alpha(x)$ įvertį

$$\widehat{\alpha}(x) = n^{-1} \sum_{i=1}^{n} \frac{K_h(x - X_i)}{g(X_i)}. \tag{2.33}$$

Kad $\widehat{\alpha}(x)$ yra $\alpha(x)$ įvertis, grindžiame taip:

$$\begin{aligned} \mathbb{E}\widehat{\alpha}(x) &= \int_{-\infty}^{+\infty} n^{-1} \sum_{i=1}^{n} \frac{K_h(x - y)}{g(y)}\, dy = \int_{-\infty}^{+\infty} K_h(x - y) \frac{f(y)}{g(y)}\, dy \\ &= \int K(z)\alpha(x + hz)\, dz \approx \alpha(x). \end{aligned} \tag{2.34}$$

Koreguoto tankio įverčio poslinkis, kai $h \overset{n\to\infty}{\longrightarrow} 0$, sumažėja nuo $O(h^2)$ iki $O(h^4)$. Išsamią šio metodo asimptotinių savybių analizę galima rasti [94].

## 2.3 Adaptyv̄us tankio įverčiai ir jų modifikacijos

**Tankio įvertis, pagrįstas antrosios išvestinės vertinimu.** Panagrinėkime įvertį, sudarytą remiantis vidutinės kvadratinės paklaidos minimumo ieškojimu. Imkime nuostolių funkciją

$$\mathbb{E}\left(\widehat{f}_h(x) - f(x)\right)^2 = b_h^2(x) + \sigma_h^2(x). \tag{2.35}$$

Kadangi, dydžiai $b_h^2(x)$ ir $\sigma_h^2(x)$ yra nežinomi, tai juos keiskime įverčiais. Įverčius apibrėšime remdamiesi asimptotinėmis formulėmis (2.9) ir (2.11), bei keisdami tankį ir jo

išvestines įverčiais

$$\widehat{b}_h(x) = \frac{c_1 \lambda(x,h) h^2}{2}, \quad \lambda(x,h) = \widehat{f}_h''(x) \tag{2.36}$$

$$\widehat{\sigma}_h^2(x) = \frac{c_2 \widehat{f}_h(x)}{nh}. \tag{2.37}$$

Branduolio pločio funkciją $h$ apibrėžkime

$$h(x) = \arg\min_h \left( \widehat{b}_h^2(x) + \widehat{\sigma}_h^2(x) \right). \tag{2.38}$$

Tankio funkcijos antrosios išvestinės įvertis $\lambda(x,h)$ randamas tokiu būdu. Pažymėkime

$$Q(y) = F(x+y) - F(x-y) - \frac{y}{h}(F(x+h) - F(x-h)). \tag{2.39}$$

Skleisdami pasiskirstymo funkciją Teiloro eilute taško $x$ aplinkoje, gauname, kad

$$\begin{aligned} Q(y) &= \left( 2f(x)y + \frac{f''(x)y^3}{3} + o(y^4) \right) - \frac{y}{h} \left( 2f(x)h + \frac{f''(x)h^3}{3} + o(h^4) \right) \\ &= \frac{f''(x)(y^3 - yh^2)}{3} + o(yh^3), \end{aligned} \tag{2.40}$$

todėl $\lambda(x,h)$ apibrėžkime

$$\lambda(x,h) = \arg\min_\lambda \max_{0 \le y \le h} \left| \widehat{Q}(y) - \frac{\lambda(y^3 - yh^2)}{3} \right|, \tag{2.41}$$

čia

$$\widehat{Q}(y) = \widehat{F}(x+y) - \widehat{F}(x-y) - \frac{y}{h} \left( \widehat{F}(x+h) - \widehat{F}(x-h) \right). \tag{2.42}$$

Pastebėsime, kad taip apibrėžtą $\lambda(x,h)$ galima tiksliai apskaičiuoti per baigtinį operacijų skaičių, nes dėl funkcijos $\widehat{Q}(y)$ pobūdžio pakanka patikrinti tik baigtinį tr ukio taškų kiekį. Todėl (2.41) galime užrašyti

$$\lambda(x,h) = \arg\min_\lambda \max_i |a_i + \lambda b_i|, \tag{2.43}$$

kur

$$a_i = \widehat{Q}(y_i), \tag{2.44}$$

$$b_i = -\frac{y_i^3 - y_i h^2}{3}, \tag{2.45}$$

o $y_i$ yra empirinės pasiskirstymo funkcijos $\widehat{F}(x)$ tr ukio taškai intervale $[x-h; x+h]$.

Tyrimai parodė, kad taip apibrėžtą tankio įvertį reikia pakoreguoti. Tankio funkcijos antrajai išvestinei įvertinti reikia platesnės taško $x$ aplinkos, negu vertinant pačią tankio funkciją. Atsižvelgiant į tai, įvertis buvo koreguotas pavartojant papildomą koeficientą $c \geq 1$ ir (2.36) keičiant į

$$\widehat{b}_h(x) = \frac{c_1 \lambda(x, ch) h^2}{2}. \tag{2.46}$$

Neblogi rezultatai gaunami imant $c = 2$. Tiksliau šio parametro reikšmę galima parinkti kryžminio patikrinimo b udu.

**Tiesioginio pakeitimo metodas.** *Tiesioginio pakeitimo metodas* (angl. *plug-in*) pagrįstas tuo, kad nežinomi dydžiai išraiškose keičiami jų statistiniais įverčiais. Šis metodas tapo populiarus pradėjus naudoti kompiuterinę techniką. Jis dažnai b una gana paprastas ir nesiremia sudėtinga matematine analize. Aukščiau aprašytas metodas, paremtas antrosios tankio funkcijos išvestinės vertinimu, taip pat remiasi ir asimptotinėmis tankio poslinkio, ir dispersijos išraiškomis. Pažvelgę į (2.15) matome, kad asimptotinės formulės netinka branduolio pločio funkcijai $h$ vertinti, kai $f''(0) \approx 0$, nes tokiu atveju $h \to \infty$. Todėl mes si ulome vertinant tankio įverčio poslinkį remtis ne asimptotinėmis formulėmis, o pasinaudoti (2.7) formule

$$b_h(x) = \int K(y) f(x+hy)\, dy - f(x) = \int (f(x+hy) - f(x)) K(y)\, dy \tag{2.47}$$

ir šioje formulėje tankį pakeisti jo įverčiu. Taigi gauname

$$\widehat{b}_{h,\Delta}(x) = \int (\widehat{f}_\Delta(x+hy) - \widehat{f}_\Delta(x)) K(y)\, dy. \tag{2.48}$$

Pastebėsime, kad šioje išraiškoje tankio įverčiui skaičiuoti reikia naudoti branduolio plotį $\Delta \geq h$, nes esant didelėms $n$ reikšmėms $b_h(x)$ vertinimas yra ekvivalentus antrosios tankio funkcijos išvestinės vertinimui, o antrajai išvestinei vertinti si uloma naudoti didesnį glodinimo plotį (žr. [95]).

Dydžiui $\Delta$ parinkti pasinaudosime asimptotine analize. Imkime $\Delta = \Delta(x)$. Tegu tenkinamos sąlygos (2.5), tuomet $\widehat{b}_{h,\Delta}(x)$ poslinkis ir dispersija bus lygūs

$$\mathbb{E}\widehat{b}_{h,\Delta}(x) = b_h(x) + o(h^2), \quad \text{jei } \Delta \to 0 \tag{2.49}$$

$$\mathbb{D}\widehat{b}_{h,\Delta}(x) = \frac{f(x)c_1^2c_3}{4}\frac{h^4}{n\Delta^5} + o\left(\frac{h^4}{n\Delta^5}\right), \quad c_3 = \|K''\|_2^2. \tag{2.50}$$

Šių lygybių išvedimas analogiškas kaip ir (2.9) ir (2.11).

Statistiškai vertinant poslinkį būtina siekti, kad

$$\mathbb{D}\widehat{b}_{h,\Delta}(x) < b_h^2(x). \tag{2.51}$$

Kai branduolio plotis $h$ artimas asimptotiškai optimaliam branduolio pločiui išreikštam formule (2.15), iš (2.9), (2.50) ir (2.51) gauname sąlygą

$$\Delta(h) \geq c_4 h, \quad c_4 = \left(\frac{c_3c_1^2}{c_2}\right)^{1/5}. \tag{2.52}$$

Todėl siūlome naudoti

$$\Delta(h) = \alpha c_4 h. \tag{2.53}$$

Parametrą $\alpha$ galime parinkti kryžminio patikrinimo būdu arba paprastumo dėlei imti $\alpha = 1$.

Modeliavimo būdu tiriant tankio įvertį, gautą vartojant branduolio plotį

$$h(x) = \arg\min_h \left(\widehat{b}_{h,\Delta}^2(x) + \widehat{\sigma}_h^2(x)\right), \tag{2.54}$$

kur $\Delta$ apibrėžtas (2.53), $\widehat{b}_{h,\Delta}(x)$ — (2.48), o $\widehat{\sigma}_h(x)$ — (2.37), buvo pastebėta, kad įvertis yra nestabilus. Tankio funkcijos antrosios išvestinės vingio taškuose skaičiuojant Integralą, gaunamos reikšmės artimos nuliui, nes pointegralinė funkcija įgyja tiek teigiamų, tiek neigiamų reikšmių, kurios kompensuoja viena kitą. Šiuo atveju gaunamos pernelyg didelės branduolio pločio $h$ reikšmės. Todėl siūlome skaičiuojant tankio įverčio poslinkio įvertį remtis ne (2.48) išraiška, o

$$\left|\widehat{b}_{h,\Delta}(x)\right| = \int_0^1 \left|\widehat{f}_\Delta(x+hy) - \widehat{f}_\Delta(x-hy) + 2\widehat{f}_\Delta(x)\right| K(y)\,dy. \tag{2.55}$$

Šios išraiškos yra asimptotiškai ekvivalenčios, bet pastaroji yra stabilesnė.

**Įverčio modifikacijos.** Tiriant aprašytą įvertį, buvo naudojamos įvairios modifikacijos, kad įverčio kokybė būtų geresnė. Buvo pastebėta, kad branduolio plotis, gautas remiantis (2.54) išraiška, nėra pakankamai glodus. Todėl jis buvo papildomai glodinamas

$$\bar{h}(x) = \frac{\sum_{i=1}^{n} h(X_i) K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}. \tag{2.56}$$

Šioje išraiškoje vietoje branduolio pločio $h$ galima vartoti pačią funkcijos $h(X_i)$ reikšmę, tačiau mes siūlome vartoti funkcijos $h(x)$ reikšmes taškuose, gretimuose taškui $x$, ir tiesiškai interpoliuoti. Tokia modifikacija padaro statistiką $\bar{h}(x)$ stabilesne, nes jeigu $h(X_i)$ viename taške yra labai mažas dėl skaičiavimo metodo nestabilumo, tai naudodami tą patį pernelyg mažą glodinimo plotį taške $x$, glodinamos funkcijos nesuglodinsime.

Panaši glodinimo procedūra buvo taikoma ne tik branduolio pločiui $h(x)$, bet ir pačiam tankio įverčiui glodinti. Ją užrašykime

$$\bar{\widehat{f}}_h(x) = \frac{\sum_{i=1}^{n} \widehat{f}_h(X_i) K_{h^*}(x - X_i)}{\sum_{i=1}^{n} K_{h^*}(x - X_i)}, \tag{2.57}$$

čia

$$h^* = h^*(x) = \beta h(x). \tag{2.58}$$

Tokia glodinimui naudojamo branduolio pločio išraiška buvo pasirinkta tam, kad galėtume reguliuoti glodinimo operacijos "stiprumą". Imdami mažesnes parametro $\beta$ reikšmes, mažiau suglodinsime tankio įvertį. Konkrečiu atveju parametrą galima parinkti taikant kryžminio patikrinimo metodą arba vartojant $\beta = 1$.

**Modifikuotas kryžminio patikrinimo metodas.** Aukščiau aprašyto kryžminio patikrinimo metodo kokybės kriterijus yra minimali integruota kvadratinė paklaida. Ši paklaida yra labai dažnai vartojama, vertinant įverčių kokybę. Tačiau ji turi trūkumą, nes yra priklausoma nuo mastelio parametro. Tarkime, kad tam tikro atsitiktinio dydžio $X$ su tankiu $f(x)$, tankio įverčio $\widehat{f}(x)$ integruota vidutinė kvadratinė paklaida yra lygi $\delta$. Pakeiskime dydžio $X$ mastelį, daugindami jį iš $k \geq 1$. Tuomet $kX$ tankis bus $\frac{1}{k} f\left(\frac{x}{k}\right)$. $kX$ tankiui įvertinti vartokime tą patį tik pakeisto mastelio įvertį, t.y. $\frac{1}{k} \widehat{f}\left(\frac{x}{k}\right)$. Apskaičiuokime šio įverčio

integruotą vidutinę kvadratinę paklaidą

$$\int\limits_{-\infty}^{+\infty}\left(\frac{1}{k}\widehat{f}\left(\frac{x}{k}\right)-\frac{1}{k}f\left(\frac{x}{k}\right)\right)^2 dx = k^{-1}\int\limits_{-\infty}^{+\infty}\left(\widehat{f}(x)-f(x)\right)^2 dx = k^{-1}\delta. \tag{2.59}$$

Matome, kad atsitiktinį dydį padauginus iš $k>1$, tankio įverčio paklaida sumažėjo $k$ kartų. Tai rodo, kad vertinant tankį, sudarytą iš kelių skirtingo mastelio komponenčių, kryžminio patikrinimo metodas "kreips didžiausią dėmesį" į komponentę, kurios tankis yra labiausiai "suspaustas". Todėl integruota vidutinė kvadratinė paklaida nėra geras optimalumo kriterijus, o (2.19) apibrėžtas kryžminio patikrinimo metodas nėra tinkamas, kai vertiname mišinių tankius.

Šio trūkumo neturi kriterijus, pagrįstas vidutinių nuostolių erdvėje $L_1$ dydžiu $\mathbb{E}\|\widehat{f}-f\|_1$, tačiau šiuos nuostolius sunku statistiškai įvertinti. Todėl siūlome modifikuoti kryžminio patikrinimo metodą keičiant (2.19) į

$$\alpha = \arg\min_{\alpha}\int\limits_{-\infty}^{+\infty}\left(\frac{\widehat{f_\alpha}(x)-f(x)}{\sqrt{f(x)}}\right)^2 dx. \tag{2.60}$$

Taip gauto parametro $\alpha$ reikšmė nepriklauso nuo mastelio parametro. Be to, jei tankis užrašomas parametriškai $f(x)\in\{f_\alpha(x),\ \alpha\in\Theta\}$, tai parametro $\alpha$ pseudo-įvertis apibrėžtas (2.60) asimptotiškai sutampa su maksimalaus tikėtinumo metodo įverčiu (žr. [133]). Norėdami apskaičiuoti parametrus taikydami modifikuotą kryžminio patikrinimo metodą, turime fiksuoti kokį nors tankio įvertį $\widehat{f_0}(x)$ ir vartoti jį (2.60) išraiškoje vietoje tikrojo tankio $f(x)$. Tam, kad išvengtume dalybos iš nulio ir modifikuotą kryžminio patikrinimo metodą padarytume stabilesniu, imkime

$$g(x) = \widehat{f_0}(x) + \frac{\varepsilon}{\sqrt{n}\max\limits_{i,j=1,\dots,n}|X_i-X_j|}, \tag{2.61}$$

kur $\varepsilon>0$ - pasirinktas parametras. Taigi $\widehat{\alpha}$ skaičiuokime tokiu būdu:

$$\widehat{\alpha} = \arg\min_{\alpha}\left(\int\limits_{-\infty}^{+\infty}\frac{\widehat{f}_\alpha^2(x)}{g(x)}\,dx - \frac{2}{n}\sum_{i=1}^{n}\frac{\widehat{f_\alpha}(X_i|i)}{\sqrt{g(X_i)}}\right)dx. \tag{2.62}$$

Šį parametrų įvertinimo metodą toliau vadinsime *modifikuotu kryžminio patikrinimo metodu*.

**Modifikuotas artimiausių kaimynų metodas.** Vienas iš artimiausių kaimynų metodo, apibrėžto (2.28), trūkumų yra tas, kad šio įverčio "uodegos" gęsta $x^{-1}$ greičiu. To galima išvengti naudojant branduolinį įvertį su branduolio pločiu apskaičiuojamu pagal $k$-kaimynų principą. Siūlome naudoti ne atstumą iki artimiausio $k$-tojo kaimyno $d_k(x)$, o taško $x$ $k$-kaimynų plotį $\rho_k(x)$, kurį apibrėžkime kaip minimalų intervalo ilgį, į kurį telpa $k$ artimiausių taškui $x$ imties taškų. Taip apibrėžtas $\rho_k(x)$ taškams $x$, esantiems imties viduje, apytiksliai lygus $d_k(x)$. Skaičiuodami $\rho_k(x)$, kai $x$ didesnis už didžiausią imties elementą, gausime $\rho_k(x) = X_{(n)} - X_{(n-k+1)}$. Taigi, $\rho_k(x)$ yra aprėžta funkcija, kai $x \to \infty$, tankio įvertis

$$\widehat{f}(x) = \frac{1}{n\rho_k(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\rho_k(x)}\right) \tag{2.63}$$

neturės lėtai gęstančių uodegų ir tankio įverčio integralas bus baigtinis.

Tiriant artimiausių kaimynų tankio įvertį, buvo pastebėta, kad branduolio plotis $h(x) = \rho_k(x)$ nėra glodus. Todėl buvo taikyta papildoma $h(x)$ glodinimo procedūra, apibrėžta (2.56). Taip pat buvo bandoma taikyti multiplikatyvų poslinkio sumažinimo metodą ir papildomą tankio glodinimo operaciją, apibrėžtą (2.57). Parametrui $k$ parinkti siūlome naudoti kryžminio patikrinimo ar modifikuotą kryžminio patikrinimo metodą.

# 3 Skyrius. Daugiamačio Gauso mišinio statistinio identifikavimo metodai

Kaip jau minėjome ankščiau, didėjant duomenų dimensijai, modelio parametrų kiekis sparčiai auga, sunkiau surasti tikslius parametrų įverčius. Daug lengviau yra įvertinti vienamačių duomenų projekcijų

$$X_\tau = \tau' X \tag{3.1}$$

tankį $f_\tau$, negu daugiamačių duomenų tankį $f$. Kadangi egzistuoja abipus vienareikšmė atitinkamybė

$$f \leftrightarrow \{f_\tau,\, \tau \in \mathbb{R}^d\}, \tag{3.2}$$

tai yra natūralu bandyti įvertinti daugiamatį tankį $f$ naudojant vienamačių stebėjimų projekcijų tankių įverčius $\widehat{f}_\tau$. Būtent tokius metodus aptarsime šiame disertacijos skyriuje.

Pastebėsime, kad mūsų nagrinėjamu Gauso mišinio atveju (1.1) stebėjimų projekcijos (3.1) taip pat pasiskirsčiusio pagal (vienamatį) Gauso mišinio modelį

$$f_\tau(x) = \sum_{j=1}^{q} p_j(\tau)\varphi_{j,\tau}(x) \stackrel{def}{=} f_\tau(x, \theta(\tau)), \tag{3.3}$$

čia $\varphi_{j,\tau}(x) = \varphi(x; m_j(\tau), \sigma_j^2(\tau))$ vienamatis Gauso tankis. Daugiamačio mišinio parametrą $\theta$ ir duomenų projekcijų pasiskirstymo parametrus
$\theta(\tau) = (p_j(\tau), m_j(\tau), \sigma_j^2(\tau))$, $j = 1, \ldots, q$ sieja lygybės

$$\begin{aligned} p_j(\tau) &= p_j, \\ m_j(\tau) &= \tau' M_j, \\ \sigma_j^2(\tau) &= \tau' R_j \tau. \end{aligned} \tag{3.4}$$

## 3.1 Apvertimo formulės taikymas

Pasinaudokime apvertimo formule

$$f(x) = \frac{1}{(2\pi)^d} \int\limits_{\mathbb{R}^d} e^{-it'x} \psi(t)\, dt, \tag{3.5}$$

čia

$$\psi(t) = \mathbb{E} e^{it'X} \tag{3.6}$$

žymi atsitiktinio dydžio $X$ charakteristinę funkciją. Pažymėję $u = |t|$, $\tau = t/|t|$ ir pakeitę kintamuosius į sferinę koordinačių sistemą gausime

$$f(x) = \frac{1}{(2\pi)^d} \int\limits_{\tau:\, |\tau|=1} ds \int\limits_0^\infty e^{-iu\tau'x} \psi(u\tau) u^{d-1}\, du. \tag{3.7}$$

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

Pažymėkime stebimo atsitiktinio dydžio projekcijos charakteristinę funkciją

$$\psi_\tau(u) = \mathbb{E} e^{iu\tau' X}, \tag{3.8}$$

tuomet

$$\psi(u\tau) = \psi_\tau(u). \tag{3.9}$$

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

$$\widehat{f}(x) = \frac{c(d)}{\|T\|} \sum_{\tau \in T} \int\limits_0^\infty e^{-iu\tau'x} \widehat{\psi}_\tau(u) u^{d-1} e^{-hu^2}\, du, \tag{3.10}$$

čia ir toliau $\|\cdot\|$ žymi aibės elementų skaičių. Pasinaudoję $d$-mačio rutulio tūrio formule

$$V_d(R) = \frac{\pi^{\frac{d}{2}} R^d}{\Gamma(\frac{d}{2}+1)} = \begin{cases} \frac{\pi^{\frac{d}{2}} R^d}{(\frac{d}{2})!}, & \text{kai } d \equiv 0 \pmod 2 \\ \frac{2^{\frac{d+1}{2}}\, \pi^{\frac{d-1}{2}} R^d}{d!!}, & \text{kai } d \equiv 1 \pmod 2 \end{cases} \tag{3.11}$$

galime apskaičiuoti konstantą $c(d)$, priklausančią nuo duomenų dimensijos

$$c(d) = \frac{(V_d(1))'_R}{(2\pi)^d} = \frac{d\, 2^{-d}\, \pi^{-\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}. \tag{3.12}$$

Jau pirmieji kompiuterinio modeliavimo tyrimai parodė, kad naudojant apvertimo formulę gauti tankio įvertiniai yra neglodūs. Todėl formulėje (3.10) po integralo ženklu įvedėme papildomą daugiklį $e^{-hu^2}$. Šis daugiklis atlieka papildomą įverčio $\widehat{f}(x)$ glodinimą su Gauso branduolio funkcija. Kaip pamatysime vėliau, tokia daugiklio forma leidžia analitiškai suskaičiuoti integralo reikšmę, o Monte-Karlo tyrimai parodė, kad jį naudojant žymiai sumažėja įverčių paklaidos.

Formulė (3.10) gali būti naudojama esant įvairiems projektuotų duomenų charakteristinės funkcijos įvertiniams. Mūsų nagrinėjamu Gauso mišinio atveju, patogu naudoti parametrinį šios funkcijos įvertinį

$$\widehat{\psi}_\tau(u) = \sum_{j=1}^{\widehat{q}_\tau} \widehat{p}_j(\tau)\, e^{iu\widehat{m}_j(\tau) - u^2\widehat{\sigma}_j^2(\tau)/2}. \tag{3.13}$$

Įstatę (3.13) į (3.10) gauname

$$\begin{aligned} \widehat{f}(x) &= \frac{c(d)}{\|T\|} \sum_{\tau\in T} \sum_{j=1}^{\widehat{q}_\tau} \widehat{p}_j(\tau) \int_0^\infty e^{iu(\widehat{m}_j(\tau) - \tau' x) - u^2(h + \widehat{\sigma}_j^2(\tau)/2)} u^{d-1}\, du \\ &= \frac{c(d)}{\|T\|} \sum_{\tau\in T} \sum_{j=1}^{\widehat{q}_\tau} \widehat{p}_j(\tau)\, I_{d-1}\left(\frac{\widehat{m}_j(\tau) - \tau' x}{\sqrt{\widehat{\sigma}_j^2(\tau) + 2h}}\right) \left(\sqrt{\widehat{\sigma}_j^2(\tau) + 2h}\right)^{-d}, \end{aligned} \tag{3.14}$$

kur

$$I_j(y) = \mathrm{Re}\left[\int_0^\infty e^{iyz - z^2/2} z^j\, dz\right]. \tag{3.15}$$

Pastebėsime, kad čia galime nagrinėti tik realią išraiškos dalį (menamųjų dalių suma turi

b$\bar{\text{u}}$ti lygi nuliui), nes tankio įvertis $\widehat{f}(x)$ gali įgyti tik realias reikšmes. Pasirinkta glodinimo daugiklio forma $e^{-hu^2}$ leidžia susieti glodinimo parametrą $h$ su projekcijų klasterių dispersijomis — tiesiog skaičiavimuose dispersijas padidinsime dydžiu $2h$.

Apskaičiuokime išraišką (3.15). Pažymėkime

$$K_j(y) = \int_0^\infty \cos yz \cdot e^{-z^2/2} \cdot z^j \, dz, \tag{3.16}$$

$$S_j(y) = \int_0^\infty \sin yz \cdot e^{-z^2/2} \cdot z^j \, dz, \tag{3.17}$$

tuomet

$$\int_0^\infty e^{iyz - z^2/2} z^j \, dz = K_j(y) + iS_j(y). \tag{3.18}$$

Integruodami dalimis gausime

$$\begin{aligned} K_j(y) = & -e^{-z^2/2} z^{j-1} \cos yz \Big|_0^\infty \\ & + \int_0^\infty e^{-z^2/2} \left( (j-1) z^{j-2} \cos yz - y z^{j-1} \sin yz \right) dz = \\ = & 1_{\{j=1\}} + (j-1) K_{j-2}(y) - y S_{j-1}(y), \quad j \geq 1. \end{aligned} \tag{3.19}$$

Analogiškai išreiškę $S_j(y)$ bei atsižvelgę į $j$ indekso apribojimus gausime rekurentines lygtis

$$K_j(y) = (j-1) K_{j-2}(y) - y S_{j-1}(y), \quad j \geq 2, \tag{3.20}$$

$$K_1(y) = 1 - y S_0(y), \tag{3.21}$$

$$S_j(y) = (j-1) S_{j-2}(y) + y K_{j-1}(y), \quad j \geq 2, \tag{3.22}$$

$$S_1(y) = y K_0(y). \tag{3.23}$$

Funkcijų $K_0(y)$ bei $S_0(y)$ apskaičiavimui pasinaudosime tuo, kad

$$(S_0(y))'_y = \int_0^\infty z \cos yz \cdot e^{-z^2/2} \, dz = K_1(y). \tag{3.24}$$

Iš (3.21) ir (3.24) gauname, kad $S_0$ tenkina diferencialinę lygtį

$$S_0'(y) = 1 - yS_0(y), \quad S_0(0) = 0. \tag{3.25}$$

Išspręskime šią lygtį, skleisdami $S_0$ Teiloro eilute

$$S_0'(y) = \sum_{k=0}^{\infty} c_{k+1}(k+1)y^{k+1} = 1 - \sum_{k=2}^{\infty} c_{k-1}y^k. \tag{3.26}$$

Sulyginę koeficientus prie panašių narių, rasime jų reikšmes

$$\begin{aligned} &c_0 = 0, \quad c_1 = 1, \\ &c_k = -c_{k-2}/k, \quad k \geq 2. \end{aligned} \tag{3.27}$$

Taigi

$$S_0(y) = \sum_{k=0}^{\infty} \frac{(-1)^k y^{2k+1}}{(2k+1)!!} = y - \frac{y^3}{3!!} + \frac{y^5}{5!!} - \frac{y^7}{7!!} + \dots. \tag{3.28}$$

$K_0$ rasime iš (3.16) išraiškos

$$\begin{aligned} K_0(y) &= \int_0^{\infty} \cos yz \cdot e^{-z^2/2}\, dz = \frac{1}{2}\int_{-\infty}^{\infty} \cos yz \cdot e^{-z^2/2}\, dz \\ &= \frac{1}{2}\int_{-\infty}^{\infty} (\cos yz - i \sin yz) \cdot e^{-z^2/2}\, dz = \sqrt{\frac{\pi}{2}}\, e^{-y^2/2}. \end{aligned} \tag{3.29}$$

Mūsų ieškomo integralo (3.10) reikšmė

$$I_j(y) = K_j(y). \tag{3.30}$$

Apvertimo formulę galima taikyti ne tik neparametriniams tankio įverčiams rasti, bet ir stebėjimams klasterizuoti. Tam būtina turėti suderintus tarp krypčių projektuotų duomenų pasiskirstymo parametrų įverčius, t.y. klasterių kiekis $\widehat{q}_t au$ ir projektuotų duomenų klasterių numeracija turi būti vienoda visoms projektavimo kryptims. Tokiu atveju, stebėjimų

priklausymo klasėms tikimybes rasime pagal formulę

$$\widehat{\pi}(j,x) = \frac{\widehat{f}_j(x)}{\sum_{j=1}^{q} \widehat{f}_j(x)}, \tag{3.31}$$

kur

$$\widehat{f}_j(x) = \frac{c(d)}{\|T\|} \sum_{\tau \in T} \widehat{p}_j(\tau)\, I_{d-1}\left( \frac{\widehat{m}_j(\tau) - \tau' x}{\sqrt{\widehat{\sigma}_j^2(\tau) + 2h}} \right) \left( \sqrt{\widehat{\sigma}_j^2(\tau) + 2h} \right)^{-d}. \tag{3.32}$$

Jeigu apsiribojame tankio $f(x)$ įvertinimo uždaviniu, apvertimo formule pagrįstas metodas gali naudotis projektuotų duomenų parametrų įveričiai, kurie turi skirtingą klasterių kiekį įvairiose kryptyse.

**Pastaba.** Kadangi

$$\begin{aligned}
&\int_0^{\infty} e^{-iu\tau' x} \widehat{\psi}_\tau(u) u^{d-1} e^{-hu^2}\, du \Bigg|_{\tau=-\tau} = \int_0^{\infty} e^{-iu(-\tau)' x} \widehat{\psi}_{-\tau}(u) u^{d-1} e^{-hu^2}\, du \\
&= \int_0^{-\infty} e^{-i(-u)(-\tau)' x} \widehat{\psi}_{-\tau}(-u)(-u)^{d-1} e^{-h(-u)^2}\, d(-u) \\
&= \int_{-\infty}^{0} e^{-iu\tau' x} \widehat{\psi}_\tau(u)(-u)^{d-1} e^{-hu^2}\, du,
\end{aligned} \tag{3.33}$$

tai prie sąlygos $d \equiv 1 \pmod 2$

$$\sum_{\tau \in \{\tau, -\tau\}} \int_0^{\infty} e^{-iu\tau' x} \widehat{\psi}_\tau(u) u^{d-1} e^{-hu^2}\, du = \int_{-\infty}^{\infty} e^{-iu\tau' x} \widehat{\psi}_\tau(u) u^{d-1} e^{-hu^2}\, du. \tag{3.34}$$

Tokiu atveju analitiškai suintegruoti (3.10) išraiškos integralą galime paprasčiau. Tegu $T' \stackrel{def}{=} \{-\tau,\ \tau \in T\}$.

$$\begin{aligned}
\widehat{f}(x) &= \frac{c(d)}{\|T\| + \|T'\|} \sum_{\tau \in T \cup T'} \int_0^{\infty} e^{-iu\tau' x} \widehat{\psi}_\tau(u) u^{d-1} e^{-hu^2}\, du \\
&= \frac{c(d)}{2\,\|T\|} \sum_{\tau \in T} \int_{-\infty}^{\infty} e^{-iu\tau' x} \widehat{\psi}_\tau(u) u^{d-1} e^{-hu^2}\, du
\end{aligned} \tag{3.35}$$

Pasinaudoję lygybe

$$f^{(k)}(x) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} e^{-iu} \psi(u) (iu)^k \, du, \tag{3.36}$$

randame įverčio išraišką

$$\widehat{f}(x) = \frac{\pi (-1)^{\frac{d-1}{2}} c(d)}{\|T\|} \sum_{\tau \in T} \sum_{j=1}^{\widehat{q}_\tau} \widehat{p}_j(\tau) \, \varphi^{(d-1)} \left( \frac{\widehat{m}_j(\tau) - \tau^{'} x}{\sqrt{\widehat{\sigma}_j^2(\tau) + 2h}} \right) \left( \sqrt{\widehat{\sigma}_j^2(\tau) + 2h} \right)^{-d}, \tag{3.37}$$

čia $\varphi^{(d-1)}$ žymi standartinio normaliojo tankio $d-1$ eilės išvestinę.

## 3.2 Mažiausių kvadratų metodo taikymas

**Mažiausių kvadratų metodas projektuotų duomenų parametrams.** Turėdami daugiamačio Gauso mišinio vienamačių projekcijų parametrų $\theta(\tau)$ įverčius ir panaudoję mažiausių kvadratų metodą, galime apskaičiuoti daugiamačio mišinio parametrą $\theta$. Kadangi tarp mišinio parametrų ir mišinio projekcijų parametrų galioja lygybės (3.4), daugiamačio parametro įvertį apibrėšime

$$\begin{aligned} \widehat{p}_j &: \sum_{\tau} \left( p_j - \widehat{p}_j(\tau) \right)^2 \longrightarrow min, \\ \widehat{M_j} &: \sum_{\tau} \left( \tau' M_j - \widehat{m}_j(\tau) \right)^2 \longrightarrow min, \\ \widehat{R}_j &: \sum_{\tau} \left( \tau' R_j \tau - \widehat{\sigma}_j^2(\tau) \right)^2 \longrightarrow min. \end{aligned} \tag{3.38}$$

Įverčius $\widehat{p}_j$, $\widehat{M_j}$ ir $\widehat{R}_j$ rasime naudodami įprastas mažiausių kvadratų metodo formules, tačiau taip rastas įvertis $\widehat{R}_j$ neb̄ūtinai yra teigiamai apibrėžta matrica. Dažniausiai ši sąlyga nėra patenkinta esant santykinai mažai imčiai ir dideliems parametrų kiekiams. Kovariacinės matricos įverčio teigiamam apibrėžtumui užtikrinti si ū lome naudoti matricos išdėstymą tikrinėmis reikšmėmis, ir neigiamas tikrines reikšmes keisti duomenų, projektuotų atitinkamo tikrinio vektoriaus kryptimi, dispersijų įverčiais. Pastebėsime, kad analogiška kovariacinės matricos įvertinimo metodika si ū loma [6], [19], norint gauti robastiškus įverčius, kai duomenų dimensija yra didelė.

**Mažiausių kvadratų metodas projektuotų duomenų tankiams.** Bandydami taikyti mažiausių kvadratų metodą projektuotų duomenų mišinio parametrams susidūrėme su klasterių suderinamumo problema: klasterių kiekis turi sutapti, išskirti klasteriai turi atspindėti artimas imties taškų aibes daugiamatėje erdvėje, klasteriai turi būti vienodai numeruoti visose kryptyse. Be to, panašūs tankiai gali turėti pakankamai skirtingas parametrų reikšmes. Todėl bandymas minimizuoti ne kvadratinių atstumų sumą tarp parametrų, o tarp projektuotų duomenų tankių, leistų išvengti šių trūkumų.

Taigi, apibrėžkime parametro įvertį, kuris minimizuoja atstumą tarp tankių

$$\widehat{\theta}:\ \sum_{\tau}\left\|f_\tau(\cdot,\theta)-\widehat{f_\tau}(\cdot)\right\|_2^2\longrightarrow min, \tag{3.39}$$

čia $f_\tau(\cdot,\theta)$ parametrinė tankio išraiška, apibrėžta (3.3), o $\widehat{f_\tau}(\cdot)$ koks nors atsitiktinio vektoriaus projekcijos tankio įvertis.

Ši lygybė gali būti perrašyta

$$\widehat{\theta}:\ \sum_{\tau}\int\limits_{\mathbb{R}}\left(f_\tau^2(x,\theta)-2f_\tau(x,\theta)\widehat{f_\tau}(x)\right)\,dx\longrightarrow min, \tag{3.40}$$

kur paskutinis skirtumo kvadrato narys $\widehat{f_\tau^2}(x)$ yra praleistas, nes jis nepriklauso nuo minimizavimo argumento $\theta$.

Mūsų tiriamu atveju galime naudoti parametrinį projekcijų tankio įvertį $\widehat{f_\tau}(\cdot)$, nes kiekvienoje projekcijoje galime įvertinti vienamačio mišinio parametrus, arba galime antros komponentės integralą pakeisti suma pagal stebėjimų projekcijas. Tyrimams pasirinkome pastarąjį atvejį, nes jis nenaudoja papildomos vienamačių mišinių parametrų vertinimo procedūros, taigi yra nepriklausomas nuo galimų jos trūkumų. Be to, tai nepadaro metodo sudėtingesniu. Taigi, pasirinktu atveju (3.40) perrašysime detalizuodami minimizuojamos funkcijos išraišką

$$\widehat{\theta}:\ Q(\theta)=\sum_{\tau}\left(\int\limits_{\mathbb{R}}f_\tau^2(x,\theta)\,dx-\frac{2}{n}\sum_{j=1}^{n}f_\tau(\tau' X_j,\theta)\right)\longrightarrow min. \tag{3.41}$$

Visų nežinomų modelio parametrų vektorių $\theta$ galime padalinti į tris dalis

$$\theta = \begin{pmatrix} \mathbf{p} \\ \mathbf{u} \\ \mathbf{v} \end{pmatrix}, \tag{3.42}$$

čia $\mathbf{p}$ yra klasterių tikimybių vektorius, $\mathbf{u}$ — visų klasterių vidurkio elementų vektorius, o $\mathbf{v}$ — visų klasterių kovariacinių matricų elementų vektorius. Kadangi Gauso tankių sandaugos integralas yra lygus

$$\int_{\mathbb{R}} \varphi(x; m_1, \sigma_1^2)\, \varphi(x; m_2, \sigma_2^2)\, dx = \varphi(m_1 - m_2; 0, \sigma_1^2 + \sigma_2^2), \tag{3.43}$$

nuostolių funkcija (3.41) turi formą

$$Q(\theta) = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) \varphi_{\mathbf{i}}(\mathbf{u}, \mathbf{v}) \longrightarrow min, \tag{3.44}$$

čia $\mathbf{i}$ yra vektorinis indeksas, sumuojantis pagal projekcijas $\tau$, klasterius $i, j = 1, ..., q$ ir imties taškus $l = 1, \ldots, n$, t.y. $\mathbf{i} \in T \otimes \{1, \ldots, q\} \otimes \{1, \ldots, q\} \otimes \{1, \ldots, n\}$. Pastebėsime, kad

$$\varphi_{\mathbf{i}} = \frac{1}{\sqrt{\beta_{\mathbf{i}}}} exp\left(-\frac{\alpha_{\mathbf{i}}}{2\beta_{\mathbf{i}}}\right), \text{ čia } \alpha_{\mathbf{i}} = \left((\mathbf{u} - a_{\mathbf{i}})^T c_{\mathbf{i}}\right)^2, \ \beta_{\mathbf{i}} = b_{\mathbf{i}}^T \mathbf{v}. \tag{3.45}$$

Šioje formulėje $a_{\mathbf{i}}$, $b_{\mathbf{i}}$ ir $c_{\mathbf{i}}$ yra vektoriai, priklausantys tik nuo imties ir projektavimo krypčių, t.y. turėdami fiksuotą imtį ir pasirinkę projektavimo kryptis galime šiuos vektorius laikyti konstantomis.

Apskaičiuokime nuostolių funkcijos $Q(\theta)$ išvestines.

$$\frac{d\varphi_{\mathbf{i}}}{d\mathbf{u}} = -\frac{\varphi_{\mathbf{i}}}{\beta_{\mathbf{i}}} \alpha_{\mathbf{i}}' = A_{\mathbf{i}}(\mathbf{u} - a_{\mathbf{i}}), \tag{3.46}$$

$$\text{čia } A_{\mathbf{i}} = -\frac{\varphi_{\mathbf{i}}}{\beta_{\mathbf{i}}} \left[c_{\mathbf{i}} c_{\mathbf{i}}^T\right],$$

$$\frac{d\varphi_{\mathbf{i}}}{d\mathbf{v}} = \varphi_{\mathbf{i}} \left(\frac{\alpha_{\mathbf{i}}}{2\beta_{\mathbf{i}}^2} - \frac{1}{2\beta_{\mathbf{i}}}\right) \beta_{\mathbf{i}}' = d_{\mathbf{i}} - B_{\mathbf{i}} \mathbf{v}, \tag{3.47}$$

$$\text{čia } \ d_{\mathbf{i}} = \frac{\varphi_{\mathbf{i}} \alpha_{\mathbf{i}}}{2\beta_{\mathbf{i}}^2} b_{\mathbf{i}}, \ B_{\mathbf{i}} = \frac{\varphi_{\mathbf{i}}}{2\beta_{\mathbf{i}}} \left[b_{\mathbf{i}} b_{\mathbf{i}}^T\right].$$

Taigi,

$$\frac{dQ}{d\mathbf{u}} = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) A_{\mathbf{i}}(\mathbf{u} - a_{\mathbf{i}}) = C\mathbf{u} - e, \tag{3.48}$$

$$\text{čia } C = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) A_{\mathbf{i}},\ e = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) A_{\mathbf{i}} a_{\mathbf{i}},$$

$$\frac{dQ}{d\mathbf{v}} = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p})(d_{\mathbf{i}} - B_{\mathbf{i}}\mathbf{v}) = g - D\mathbf{v}, \tag{3.49}$$

$$\text{čia } D = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) B_{\mathbf{i}},\ g = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) d_{\mathbf{i}}.$$

Čia $a_{\mathbf{i}}$, $b_{\mathbf{i}}$, $c_{\mathbf{i}}$, $d_{\mathbf{i}}$, $e$ ir $g$ yra vektoriai, o $A_{\mathbf{i}}$, $B_{\mathbf{i}}$, $C$ ir $D$ — matricos. Kadangi

$$\widehat{\theta}:\ Q(\theta) \longrightarrow \min \quad \Rightarrow \quad \frac{dQ}{d\mathbf{p}}(\widehat{\theta}) = 0,\ \frac{dQ}{d\mathbf{u}}(\widehat{\theta}) = 0,\ \frac{dQ}{d\mathbf{v}}(\widehat{\theta}) = 0, \tag{3.50}$$

galime pateikti rekurentinį algoritmą $\widehat{\theta}$ apskaičiavimui

$$\mathbf{u}^{(k+1)} = C^{-1}(\theta^{(k)})\, e(\theta^{(k)}),\ \mathbf{v}^{(k+1)} = D^{-1}(\theta^{(k)})\, g(\theta^{(k)}). \tag{3.51}$$

**Klasterių tikimybių apskaičiavimas.** Apskaičiuokime klasterių tikimybių vektorių $\mathbf{p}$. Detalizuokime (3.44) išraiką:

$$Q(\theta) = \sum_{i,j=1}^{q} p_i p_j S_1(i,j) - \sum_{i=1}^{q} p_i S_2(i), \tag{3.52}$$

kur

$$S_1(i,j) = \sum_{\tau \in T} \frac{exp\left(-\frac{(\tau'(M_i - M_j))^2}{2\tau'(R_i + R_j)\tau}\right)}{\sqrt{2\pi}\sqrt{\tau'(R_i + R_j)\tau}}, \quad S_2(i) = \frac{2}{n}\sum_{j=1}^{n}\sum_{\tau \in T} \frac{exp\left(-\frac{(\tau'(M_i - X_j))^2}{2\tau' R_i \tau}\right)}{\sqrt{2\pi}\sqrt{\tau' R_i \tau}}. \tag{3.53}$$

Tuomet

$$\frac{dQ}{dp_i} = 2\sum_{j=1}^{q} p_j S_1(i,j) - S_2(i) \tag{3.54}$$

ir vektorių $\mathbf{p}$ rasime iš lygties

$$\mathbf{p} = A^{-1}B, \quad \text{kur} \quad a_{i,j} = 2S_1(i,j), \quad b_i = S_2(i). \tag{3.55}$$

Pastebėsime, kad taip rastų klasterių tikimybių suma nebūtinai lygi 1, todėl pertvarkysime lygtis, kad užtikrintume sąlygos $p_1 + ... + p_q = 1$ galiojimą (paprastumo dėlei galima normuoti klasterių tikimybių vektorių kiekvienos iteracijos metu). Išreikškime

$$p_q = 1 - \sum_{i=1}^{q-1} p_i \tag{3.56}$$

ir įstatykime šią išraišką į (3.52). Tuomet sugrupavę panašius narius ir pasinaudoję funkcijos $S_1(i,j)$ simetriškumu argumentų atžvilgiu gausime

$$\begin{aligned} Q(\theta) =& S_1(q,q) + \sum_{i=1}^{q-1} p_i 2[S_1(i,q) - S_1(q,q)] \\ &+ \sum_{i,j=1}^{q} p_i p_j [S_1(i,j) - S_1(i,q) - S_1(q,j) + S_1(q,q)] \\ &- S_2(q) - \sum_{i=1}^{q-1} p_i [S_2(i) - S_2(q)], \end{aligned} \tag{3.57}$$

o

$$\begin{aligned} \frac{dQ}{dp_i} =& \sum_{j=1}^{q-1} p_j \, 2[S_1(i,j) - S_1(i,q) - S_1(q,j) + S_1(q,q)] \\ &- [S_2(i) - S_2(q) - 2S_1(i,q) + 2S_1(q,q)]. \end{aligned} \tag{3.58}$$

Prilyginę dalines išvestines nuliui tikimybių vektorių randame iš lygties

$$(p_1, ...p_{q-1})^T = A^{-1}B, \tag{3.59}$$

kur matrica $A_{q-1\times q-1}$ ir vektorius $B_{q-1\times 1}$ sudaryti atitinkamai iš

$$a_{i,j} = 2[S_1(i,j) - S_1(i,q) - S_1(q,j) + S_1(q,q)] \quad \text{ir} \tag{3.60}$$

$$b_i = S_2(i) - S_2(q) - 2S_1(i,q) + 2S_1(q,q), \tag{3.61}$$

o $p_q$ randame iš (3.56).

## 3.3 Geometrinis klasterizavimas

Taikant mažiausių kvadratų metodą projektuotiems parametrams, reikia pradinio suskaldymo tam, kad galėtume gauti suderintus klasterių įverčius projekcijose į įvairias kryptis. Taikant mažiausių kvadratų metodą projektuotiems tankiams, pradinė parametro reikšmė buvo reikalinga, kad galėtume, pradedant ja, minimizuoti norimą funkciją. Iteraciniam EM algoritmui taip pat reikalingas pradinis imties suskaldymas arba mišinio parametro reikšmė. Taigi, pradinio imties suskaldymo problema išlieka ir yra kertinė sprendžiant klasifikavimo ar parametrinio tankio įvertinimo uždavinius.

Aprašysime naują griežto klasterizavimo procedūrą, kuri padės išspręsti pradinio imties suskaldymo uždavinį. Sakykime, visiems $x, y \in \mathbb{R}^d$ apibrėžta neneigiama funkcija $\rho(x, y)$, kurią vadinsime pseudoatstumu. Bendru atveju ji gali netenkinti trikampio taisyklės ir netgi pseudoatstumas nuo taško iki jo pačio $\rho(x, x)$ nebūtinai bus lygus nuliui. Stebėjimų numerių aibę $N$ suskaldysime į nesusikertančius poaibius $\widehat{K}_1, ..., \widehat{K}_q$ taip, kad funkcionalas

$$Q(K_1, ..., K_q) = \sum_{j=1}^{q} \frac{1}{\|K_j\|} \sum_{s,r \in K_j} \rho(X(s), X(r)) \tag{3.62}$$

įgytų mažiausią reikšmę. Formule (3.62) nusakius imties taškų suskaldymą belieka apibrėžti pseudoatstumo funkciją $\rho$. Ši funkcija turi atspindėti taškų priklausomybę klasteriams, o ne geometrinį taškų išsidėstymą Euklido erdvėje, t.y. pseudoatstumas tarp taškų priklausančių tam pačiam klasteriui turi būti mažas, o skirtingiems — didelis. Funkciją $\rho$ apibrėšime naudodamiesi duomenų projektavimu. Pirmiausiai apibrėžkime pseudoatstumą tarp imties taškų projekcijų, o tuomet, naudodamiesi juo, nusakykime pseudoatstumą tarp imties taškų.

**Pseudoatstumo tarp imties taškų projekcijų parinkimas.** Kadangi pseudoatstumo funkcija turi atspindėti taškų priklausomybę klasteriams, remsimės imties taškų projekcijų klasifikavimo tikimybių

$$\pi_\tau(j, x) = \mathbb{P}\{\nu = j | \tau' X = \tau' x\} \tag{3.63}$$

įverčiais. Juos iš projektuotų duomenų skirstinio parametrų įverčių rasime naudodami formulę

$$\widehat{\pi}_\tau(j, x) = \frac{\widehat{p}_j(\tau)\widehat{\varphi}_{j,\tau}(x)}{f_\tau(x, \widehat{\theta}(\tau))} \tag{3.64}$$

Tyrimams pasirinkome dvi pseudoatstumo tarp imties taškų projekcijų funkcijas: funkciją, pagrįstą skirtumu tarp klasifikavimo tikimybių

$$\rho_\tau(x,y) = \sum_{j=1}^{\widehat{q}_\tau} |\widehat{\pi}_\tau(j,x) - \widehat{\pi}_\tau(j,y)|, \tag{3.65}$$

ir funkciją, įvertinančią nepriklausymo tai pačiai klasei tikimybę

$$\rho_\tau(x,y) = 1 - \sum_{j=1}^{\widehat{q}_\tau} \widehat{\pi}_\tau(j,x)\widehat{\pi}_\tau(j,y), \tag{3.66}$$

nes

$$1 - \sum_{j=1}^{q} \pi_\tau(j,x)\pi_\tau(j,y) = \mathbb{P}\{\nu(X) \neq \nu(Y) | X = x, Y = y\}. \tag{3.67}$$

**Pseudoatstumo tarp imties taškų parinkimas.** Pseudoatstumas tarp imties taškų bus konstruojamas remiantis pseudoatstumais tarp imties taškų projekcijų įvairiose kryptyse. Kadangi atskirose projekcijose duomenų klasės persidengs, tai faktas, kad taškų projekcijos priklauso vienai klasei (pseudoatstumas tarp taškų projekcijų yra mažas) kryptyje nerodo, kad šie taškai yra iš vienos klasės. Bet jei yra krypčių, kuriose taškų projekcijos aiškiai priklauso skirtingoms klasėms, tai ir patys taškai priklausys skirtingoms klasėms. Remiantis tokiais samprotavimais turėtume apibrėžti pseudoatstumą tarp taškų taip

$$\rho(x,y) = \max_\tau \rho_\tau(\tau' x, \tau' y). \tag{3.68}$$

Tačiau toks apibrėžimas gali neduoti stabilaus įverčio, nes esant netgi vienai krypčiai, kurioje buvo aiškiai blogai įvertinti imties projekcijos parametrai ir to pasekoje vieno klasterio taškai buvo priskirti skirtingiems klasteriams, pseudoatstumas tarp taškų $\rho(\cdot,\cdot)$ blogai atspindės tikrą taškų priklausomybę tam pačiam klasteriui. Todėl stabilesnius rezultatus turėtų duoti analogiškai apibrėžtas įvertinys, tik kai maksimumas randamas atmetus nedidelę $\alpha$ dalį krypčių, kur pseudoatstumas tarp tų taškų projekcijų yra didžiausias.

Kitas būdas, padedantis sumažinti išskirčių įtaką, yra maksimumo keitimas p-laipsnio vidurkiu. Šiuo atveju

$$\rho^p(x,y) = \tfrac{1}{\|T\|} \sum_\tau \rho_\tau^p(\tau' x, \tau' y). \tag{3.69}$$

Tokiu atveju reiktų parinktu dydį $p$, kuris būtų metodo parametru.

**Funkcionalo minimizavimas.** Geometriniai klasifikavimo metodai yra pakankamai išvystyti, ir (3.62) tipo funkcionalų minimizavimą, kai pseudoatstumas $\rho$ yra Euklido atstumas tarp taškų, atlieka daugelis statistinės analizės paketų. Tačiau dėl netradicinės atstumo funkcijos naudojimo teko naudoti specialią procedūrą funkcionalui minimizuoti. Ji savo esme analogiška k-centrų metodui, tačiau mūsų siūlomas algoritmas naudoja tik pseudoatstumus tarp imties taškų, bet nenaudoja pseudoatstumo tarp taško ir imties centro, kas yra svarbu praktiškai realizuojant klasifikavimo algoritmą.

Turėdami pseudoatstumą tarp imties taškų $\rho(x, y)$ apibrėžkime pseudoatstumą nuo imties taško iki klasės

$$\rho(x, K) = \frac{1}{\|K\|} \sum_{t \in K} \rho(x, X(t)). \tag{3.70}$$

Iš imties išskirsime $q$ klasterių, o paskui, taikydami rekurentinį algoritmą, taškus pergrupuosime taip, kad funkcionalas (3.62) būtų minimizuojamas.

Iš dar nesuklasifikuotų imties taškų išskirkime tokį, kurio pseudoatstumas iki likusių yra didžiausias. Iš šio taško ir artimiausių jo kaimynų numerių sudarykime naują $[n/q]$ dydžio klasterį, čia $[\cdot]$ žymi sveikąją skaičiaus dalį.Kartodami šią klasterių išskyrimo procedūrą išskirkime $q - 1$ pradinių klasterių. Paskutinį klasterį sudarysime iš likusių nesuklasifikuotų taškų.

Taip sudarysime pradinius klasterius, tenkinančius savybes:

1. $\bigcup\limits_{i=1}^{q} K_i = N$.
2. $K_i = \{s_i\} \cup \{s : \rho(s_i, s) \leq \rho(s_i, t), \forall s \in K_i^*, t \in K_i^* \backslash K_i\}$, $i = \overline{1, q-1}$, kur $s_i = \arg\max\limits_{s \in K_i^*} \rho(x(s), K_i^*)$, $\|K_i\| = [n/q]$, o $K_i^* = N \backslash \bigcup\limits_{j<i} K_j$.
3. $K_q = N \backslash \bigcup\limits_{j=1}^{q-1} K_j$.

Išskirtus klasterius rekurentiškai tikslinsime stengdamiesi minimizuoti funkcionalą $Q$. Iš kiekvienos klasterio išskirkime dalį taškų, labiausiai nutolusių nuo savo klasės, t.y. tokių, kuriems reikšmės $\rho(x(s), K_i), s \in K_i$ yra didžiausios. Iš šių taškų numerių sudarykime naują klasę $K^*$. Dabar šiuos taškus perskirstykime priskirdami tai klasei, kuriai priskirdami gausime mažiausią funkcionalo $Q$ reikšmę. Patikslintas imties suskaldymo klases

$\widetilde{K}_1, \ldots, \widetilde{K}_q$ apibrėžkime

$$\widetilde{K}_i = K_i \backslash K_* \cup \{s : s \in K_*, \eta(s) = i\}, \tag{3.71}$$

čia

$$\eta(s) = \arg \min_{i=1,\ldots,q} Q(K_1 \backslash K_*, ..., K_{i-1} \backslash K_*, K_i \backslash K_* \cup \{s\}, K_{i+1} \backslash K_*, ..., K_q \backslash K_*). \tag{3.72}$$

Prilyginę naujai gautas klases klasėms $K_1, ..., K_q$, galėsime atlikti sekančią klasių patikslinimo iteraciją ir toliau minimizuoti funkciją $Q$.

Pastebėsime, kad jeigu kiekviename rekurentiniame klasių patikslinimo algoritmo žingsnyje aibę $K_*$ sudarytume tik iš vieno taško, t.y. išskirtume vieną tolimiausią tašką iš kurios nors klasės, tai gautume monotoniškai mažėjančią funkcijos $Q$ seką, t.y. kiekviename žingsnyje

$$Q(\widetilde{K}_1, ..., \widetilde{K}_q) \leq Q(K_1, ..., K_q), \tag{3.73}$$

tačiau tokiu atveju rekurentinis algoritmas konverguotų į lokalų funkcijos $Q$ minimumą. Todėl siūlome iš pradžių atlikti iteracijas, priskiriant $K_*$ klasei pusę kiekvienos iš klasių $K_1, ...K_q$ taškų. Kai rekurentinis algoritmas nustoja konvergavęs, surasti klasių rinkinį, kuriam funkcija $Q$ turėjo mažiausią reikšmę (kai priskiriame $K_*$ daugiau nei vieną tašką, $Q$ reikšmes nebūtinai monotoniškai mažėja) ir iš šio rinkinio atlikti iteracijas perklasifikuojamų taškų kiekį sumažinus iki 1..5%. Kai rekurentinis algoritmas vėl nustoja konverguoti, tuomet vėl radus klasių rinkinį, minimizuojantį $Q$, atlikti iteracijas, priskiriant $K_*$ vieną tašką. Šiame etape $Q$ seka bus monotoniškai mažėjanti.

Taigi, aprašytas geometrinio klasterizavimo algoritmas susideda iš tokių etapų:

1. pasirenkame, pavyzdžiui tolygiai pasiskirsčiusių ant sferos, projektavimo krypčių aibę $T = \{\tau\}$ erdvėje $\mathbb{R}^d$;
2. kiekvienoje kryptyje, naudodami idėjas, aprašytas [142], bei EM algoritmą, įvertiname projektuotos imties skirstinio parametrus $\widehat{\theta}_\tau$;
3. naudodamiesi projektuotos imties parametrų įverčiais apskaičiuojame klasifikavimo tikimybių įverčius $\widehat{\pi}_\tau(j, x)$, o iš jų randame pseudoatstumų tarp projektuotų imties taškų funkcijos $\rho_\tau(x, y)$ reikšmes, o iš pastarųjų apskaičiuojame pseudoatstumų tarp imties taškų matricą $[\rho(x(i), x(j)]_{i,j=\overline{1,n}}$.

4. išskiriame pradinius klasterius ir, naudodami pseudoatstumų matricą, rekurentiškai minimizuojame funkcionalą $Q$.

Taip gauname griežtą imties taškų klasifikavimą. Jis gali b$\bar{\text{u}}$ti naudojamas, kaip pradinis imties suskaldymas, kitiems duomenų klasterizavimo algoritmams, arba kaip atskiras duomenų klasterizavimo ar mišinio parametrų įvertinimo metodas. Pastaruoju atveju Gauso mišinio klasterių svorius įvertintume $\widehat{p}_i = \frac{\|K_i\|}{n}$, o vidurkius ir kovariacines matricas - tiesiog skaičiuodami kiekvienos klasės imties taškų vidurkį ir empirinę kovariacinę matricą. Negriežto klasifikavimo tikimybių įverčius apskaičiuotume naudodami formulę (1.10).

# 4 Skyrius. Nagrinėtų statistinės analizės metodų tikslumo tyrimas

## 4.1 Adaptyvių vienamačių neparametrinio tankio įverčių tikslumo tyrimas

**Įverčių tyrimo metodika.** Tankio įverčius tyrėme Monte-Karlo metodu. Kadangi tikrasis imties tankis buvo žinomas, tai skaičiavome įverčių paklaidas ir jas lygindami darėme išvadas apie įverčių kokybę. Šiame skyriuje konkretizuosime tyrime vartotas atsitiktines imtis, tirtus įverčius ir jų paklaidas.

**Atsitiktinės imtys.** Tyrimui naudojome generuotas imtis, pasiskirsčiusias pagal Gauso mišinio modelį. Norint įvairiapusiškai ištirti siūlomus metodus, buvo varijuojami: klasterių kiekis, jų tikimybės, vidurkiai ir dispersijos. Imties parametrams nurodyti vartosime sutrumpintą pažymėjimą, pvz., $n = 500$, $(m, \sigma, p) = (0, 1, 0.3), (20, 5, 0.7)$ reikš imtį sudaryta iš 500 stebėjimų, kurių tankis yra Gauso mišinio tankis su parametrais $q = 2$, $m_1 = 0$, $\sigma_1 = 0$, $p_1 = 0.3$, $m_2 = 20$, $\sigma_2 = 5$, $p_2 = 0.7$.

**Paklaidos.** Tiriamų įverčių tikslumą matavome naudodami kelias paklaidas, pasižyminčias skirtingomis savybėmis (žr. skyrių 2.3). Daugiausia dėmesio skyrėme paklaidoms, matuojamoms erdvių $L_1$ ir $L_2$ metrikose,

$$\epsilon_1(\widehat{f}) = \|\widehat{f} - f\|_1, \tag{4.1}$$

$$\epsilon_2(\widehat{f}) = \|\widehat{f} - f\|_2, \tag{4.2}$$

o taip pat paklaidai, nuo mastelio nepriklausančioje metrikoje,

$$\epsilon_A(\widehat{f}) = \left\| \frac{\widehat{f} - f}{\sqrt{f}} \right\|_2. \tag{4.3}$$

Buvo skaičiuojami šių paklaidų empiriniai analogai $\varepsilon_1$, $\varepsilon_2$ ir $\varepsilon_A$, kuriuos ir pateiksime lentelėse.

**Skaitinių metodų taikymo ypatybės.** Skaičiuojant įverčius skaitiniais metodais atsiranda šiems metodams būdingų sunkumų. Pavyzdžiui, minimizuojant reiškinį reikia pa-

rinkti argumento intervalą ir taškų kiekį, kuriuose bus ieškomas minimumas. Tokios pat problemos atsiranda skaičiuojant integralą.

Branduolio pločio funkcijas skaičiavome 25-iuose taškuose, išdėstytuose $x$ ašyje taip, kad tarp jų būtų vienodas imties elementų kiekis. Branduolio plotį kituose taškuose skaičiavome naudodami tiesinę interpoliaciją. Skaičiuodami branduolio pločio parametrą, minimumo ieškojome intervale $\left[\frac{X_{(n)}-X_{(1)}}{8\sqrt{n}}; \frac{X_{(n)}-X_{(1)}}{4}\right]$. Optimalių branduolio pločio funkcijų reikšmių ieškojome intervale $\left[\frac{1}{2}h_{NN}(x); 2h_{NN}(x)\right]$, kur $h_{NN}(x)$ — branduolio pločio funkcijos reikšmė gauta naudojant artimiausių kaimynų metodą. Minimizuodami ir integruodami reiškinius, taškų kiekį, kuriuose skaičiavome reiškinių reikšmes, stengėmės parinkti derindami skaičiavimo spartą ir tikslumą.

**Tirti įverčiai.** Skaičiuojant tankio įverčius vartojome Epaničnikovo branduolį, apibrėžtą (2.4). Glodinimo procedūrose, nusakytose (2.56) ir (2.57), vartojome Parzeno branduolį

$$K(x) = \begin{cases} 1 - 6x^2 + g|x|^3, & \text{kai } |x| \le 1/2, \\ 2(1-x^2), & \text{kai } 1/2 \le |x| \le 1, \\ 0, & \text{kai } |x| \ge 1. \end{cases} \tag{4.4}$$

Tyrėme šiuos tankio įverčius:

- pastovaus branduolio pločio kryžminio patikrinimo įvertį apibrėžtą (2.20);
- pastovaus branduolio pločio modifikuoto kryžminio patikrinimo įvertį, apibrėžtą (2.62). Vietoje funkcijos $\widehat{f}_0(x)$ (žr. (2.61)) imdavome artimiausių kaimynų įvertį;
- artimiausių kaimynų įvertį, apibrėžtą (2.30), kai $k = n^{0.7}$ (tokį parametro parinkimą motyvavo atlikti modeliavimo tyrimai);
- artimiausių kaimynų įvertį, apibrėžtą (2.30), kai parametras $k$ buvo parenkamas naudojant kryžminio patikrinimo metodą;
- įvertį, pagrįstą antrosios išvestinės vertinimu, kai branduolio plotis apibrėžtas (2.38), (2.46), (2.37), o $c = 2$;
- tiesioginio pakeitimo įvertį, kurio branduolio plotis apibrėžtas (2.54), (2.55), (2.37), o $\Delta$ apibrėžtas (2.53), kai parametras $\alpha = 1$ (tokį parametro parinkimą motyvavo atlikti modeliavimo tyrimai);
- teorinis pseudo-įvertis.

Teorinis pseudo-įvertis apibrėžiamas analogiškai kaip ir tiesioginio pakeitimo įvertis, tik išraiškose (2.54), (2.55), (2.37) tankio įverčiai keičiami tikrosiomis tankio funkcijomis. Teorinį pseudo-įvertį tyrėme norėdami palyginti įverčių paklaidas su teoriškai minimaliomis paklaidomis. Šis įvertis nėra tikrasis tankio įvertis, nes jis naudojasi informacija apie imties skirstinį.

Tyrėme ne tik minėtus tankio įverčius, bet ir jų modifikacijas. Įverčius modifikavome glodindami branduolio pločio funkciją, remdamiesi (2.56), taikydami multiplikatyvų poslinkio sumažinimą, apibrėžtą (2.31), bei tankio įverčio glodinimą, apibrėžtą (2.57). (2.58) išraiškoje naudojome parametrą $\beta = 0.5$, nes funkcijos b ̄udavo pernelyg suglodinamos, kai imdavome tokį patį glodinimo plotį, kaip ir branduolio plotis.

Aprašydami tyrimo rezultatus vartosime tokius trumpinius:

| | |
|---|---|
| CV | Kryžminio patikrinimo metodas, |
| MCV | Modifikuotas kryžminio patikrinimo metodas. |

Modifikacijas trumpumo dėlei taip pat sunumeruokime:

| | |
|---|---|
| 1 | branduolio pločio funkcijos glodinimas, |
| 2 | multiplikatyvus poslinkio koregavimas, |
| 3 | tankio įverčio glodinimas. |

Taigi, “artimiausių kaimynų MCV įvertis mod. 2, 3” reikš artimiausių kaimynų įvertį po multiplikatyvaus poslinkio koregavimo ir tankio įverčio glodinimo, kai parametras $k$ parenkamas naudojant modifikuotą kryžminio patikrinimo metodą.

**Pradiniai tyrimai.** Pradinius metodų tyrimus atlikome naudodami dvi bandomąsias imtis: $n = 500$ $(m, \sigma, p) = (0, 1, 1)$ ir $n = 500$ $(m, \sigma, p) = (0, 1, 0.5), (20, 5, 0.5)$. Atlikdami šiuos tyrimus pastebėjome, kad įvertis, pagrįstas antrosios išvestinės vertinimu, yra nestabilus. Skaičiuodami branduolio plotį, kai kurioms $x$ reikšmėms gaudavome apribojimų intervalo $\left[\frac{1}{2}h_{NN}(x); 4h_{NN}(x)\right]$ viršutinio galo reikšmę, nors optimali $h$ reikšmė b ̄udavo mažesnė. Įverčio paklaidas pavyko sumažinti ieškant minimumo intervale $\left[\frac{1}{2}h_{NN}(x); 3h_{NN}(x)\right]$, tačiau atliekant kitus tyrimus tekdavo vis siaurinti šį intervalą. Jeigu metodas ir duodavo stabilius rezultatus, tai jie b ̄udavo nežymiai geresni už paprastesnio artimiausių kaimynų metodo rezultatus. Dėl šių priežasčių toliau šiame skyriuje neaptarinėsime įverčio, pagrįsto antrosios išvestinės vertinimu.

**Tiesioginio pakeitimo metodo parametro parinkimas.** Skaičiuodami tankio įvertį tiesioginio pakeitimo metodu turime pasirinkti parametro $\alpha$, vartojamo išraiškoje (2.53), reikšmę. Tirdami įvertį ėmėme įvairias parametro reikšmes. Kai imties parametrai $n = 500$ $(m, \sigma, p) = (0, 1, 0.5), (20, 5, 0.5)$, metodo paklaidos pateikiamos 4.1 lentelėje. Iš len-

| Metodas | $\alpha$ | $\varepsilon_1$ | $\varepsilon_2$ | $\alpha$ | $\varepsilon_1$ | $\varepsilon_2$ |
|---|---|---|---|---|---|---|
| Nemodifikuotas | 0.4 | 0.12339 | 0.02875 | 1.1 | 0.12472 | 0.02894 |
| Mod. 1,2,3 | | 0.08606 | 0.02760 | | 0.08137 | 0.02017 |
| Nemodifikuotas | 0.5 | 0.13388 | 0.03436 | 1.2 | 0.13209 | 0.03018 |
| Mod. 1,2,3 | | 0.08387 | 0.02738 | | 0.08308 | 0.01997 |
| Nemodifikuotas | 0.6 | 0.13762 | 0.03511 | 1.4 | 0.15062 | 0.03509 |
| Mod. 1,2,3 | | 0.08898 | 0.02796 | | 0.08029 | 0.01821 |
| Nemodifikuotas | 0.7 | 0.13839 | 0.03579 | 1.6 | 0.15336 | 0.03633 |
| Mod. 1,2,3 | | 0.09104 | 0.02870 | | 0.08199 | 0.01904 |
| Nemodifikuotas | 0.8 | 0.13261 | 0.03272 | 1.8 | 0.15910 | 0.03905 |
| Mod. 1,2,3 | | 0.09567 | 0.03045 | | 0.08347 | 0.01978 |
| Nemodifikuotas | 0.9 | 0.12941 | 0.03078 | 2.0 | 0.17414 | 0.03968 |
| Mod. 1,2,3 | | 0.08857 | 0.02492 | | 0.08548 | 0.02054 |
| Nemodifikuotas | 1.0 | 0.12628 | 0.02962 | 2.2 | 0.20621 | 0.04408 |
| Mod. 1,2,3 | | 0.08440 | 0.02214 | | 0.09030 | 0.02100 |

4.1 lentelė: Tiesioginio pakeitimo metodo parametro $\alpha$ parinkimas.

telės duomenų matome, kad paklaidos kinta be aiškios priklausomybės nuo $\alpha$. Panašūs rezultatai buvo gauti naudojant imtis su kitokiais parametrais. Jei bandytume analizuoti šiuo metodu apskaičiuotas branduolio pločio funkcijas, tai pastebėtume, kad naudojant per daug mažas parametro reikšmes gaunamos neglodžios funkcijos su daugybe lokalių ekstremumų, nors šie ekstremumai nedaug nutolę nuo vidutinės funkcijos reikšmės. Per daug padidinus parametrą, apskaičiuotas branduolio plotis būna didesnis už optimalų, ir metodo paklaidos pradeda didėti. Iš pradžių per didelį tankio funkcijos suglodinimą kompensuoja multiplikatyvus poslinkio sumažinimas, todėl modifikuoto įverčio paklaidos didėja lėčiau. Parametrui $\alpha$ parinkti buvo bandomas taikyti MCV metodas, tačiau ir jis pastebimai nepagerino įverčio, o tik prailgino skaičiavimo laiką ir padarė įvertį ne tokiu stabiliu. Atsižvelgiant į tai, vėlesniuose tyrimuose naudojome $\alpha = 1$.

**Artimiausių kaimynų metodo parametro parinkimas.** Artimiausių kaimynų metodas priklauso nuo parametro $k$. Parametras $k$ dažnai išreiškiamas

$$k = n^l, \qquad 0 < l < 1. \tag{4.5}$$

Tuomet turime pasirinkti laipsnio rodiklį $l$. Tyrėme artimiausių kaimynų metodą, kuriame branduolio plotį parinkome naudojant atstumą iki $k$-tojo kaimyno $d_k(x)$, aprašytą 2.2 skyriuje, bei artimiausių kaimynų metodą, kuriame branduolio plotis buvo parenkamas naudojant $k$ kaimynų plotį $\rho_k(x)$, apibrėžtą 2.3 skyriuje.

| Metodas | $l$ | Artimiausių kaimynų metodas naudojantis $d_k(x)$ | | | Artimiausių kaimynų metodas naudojantis $\rho_k(x)$ | | |
|---|---|---|---|---|---|---|---|
| | | $\varepsilon_1$ | $\varepsilon_2$ | $\varepsilon_A$ | $\varepsilon_1$ | $\varepsilon_2$ | $\varepsilon_A$ |
| Nemodifikuotas | 0.5 | 0.21211 | 0.09664 | 0.31526 | 0.22392 | 0.10132 | 0.30994 |
| Mod. 1,2,3 | | 0.23153 | 0.11253 | 0.32768 | 0.25323 | 0.11730 | 0.35052 |
| Nemodifikuotas | 0.55 | 0.19518 | 0.08275 | 0.30334 | 0.19313 | 0.08480 | 0.26381 |
| Mod. 1,2,3 | | 0.20652 | 0.09758 | 0.28500 | 0.21495 | 0.09938 | 0.29591 |
| Nemodifikuotas | 0.6 | 0.17662 | 0.06365 | 0.31237 | 0.16565 | 0.06505 | 0.21930 |
| Mod. 1,2,3 | | 0.18225 | 0.07683 | 0.23899 | 0.18868 | 0.07885 | 0.24166 |
| Nemodifikuotas | 0.65 | 0.16446 | 0.05137 | 0.37091 | 0.14068 | 0.05048 | 0.19520 |
| Mod. 1,2,3 | | 0.15454 | 0.06035 | 0.20882 | 0.15807 | 0.06210 | 0.20169 |
| Nemodifikuotas | 0.7 | 0.15211 | 0.04033 | 0.46878 | 0.11401 | 0.03497 | 0.18624 |
| Mod. 1,2,3 | | 0.11908 | 0.04052 | 0.17299 | 0.11934 | 0.04303 | 0.16914 |
| Nemodifikuotas | 0.75 | 0.18230 | 0.04437 | 0.76377 | 0.10691 | 0.02649 | 0.21268 |
| Mod. 1,2,3 | | 0.09307 | 0.02797 | 0.18371 | 0.08669 | 0.02941 | 0.13917 |
| Nemodifikuotas | 0.8 | 0.22730 | 0.05487 | 1.10946 | 0.16442 | 0.03974 | 0.60241 |
| Mod. 1,2,3 | | 0.09842 | 0.02827 | 0.21519 | 0.08980 | 0.02770 | 0.17507 |
| Nemodifikuotas | 0.85 | 0.35156 | 0.07893 | 2.10072 | 0.25180 | 0.05663 | 1.18909 |
| Mod. 1,2,3 | | 0.11028 | 0.02417 | 0.27319 | 0.08976 | 0.02073 | 0.17512 |
| Nemodifikuotas | 0.9 | 0.84890 | 0.23351 | 5.09582 | 0.80400 | 0.21112 | 6.83276 |
| Mod. 1,2,3 | | 0.87233 | 0.23554 | 5.47386 | 0.73636 | 0.20570 | 6.02714 |

4.2 lentelė: Artimiausių kaimynų metodo įverčių paklaidų priklausomybę parametro $l$.

Tyrimai parodė, kad šis metodas nėra jautrus $k$ parametro parinkimui. Kai imties tūris $n = 500$, metodas duoda panašias paklaidas tiek su $k = 22$ ($l = 0.5$), tiek su $k = 144$ ($l = 0.8$), ir šios paklaidos skiriasi nuo optimalių 30%–50% (žr. 4.2 lentelę). Optimalios nemodifikuotų paklaidų reikšmės pasiekiamos kai $l = 0.7$, o modifikuotų — kai $l = 0.75$. Parametrą $k$ bandėme parinkti kryžminio patikrinimo ir modifikuotu kryžminio patikrinimo metodu. Modifikuoto kryžminio patikrinimo metodo rezultatai nebuvo geri. Tai galėjo nulemti kelios priežastys. Visų pirma, modifikuotam kryžminio patikrinimo metodui reikalingas pagalbinis įvertis $\widehat{f}_0(x)$. Kadangi šiuo įverčiu laikėme patį artimiausių kaimynų metodo įvertį, tai gaudavome blogus paklaidų įverčius. Be to, 4.2 lentelėje matome, kad paklaida $\varepsilon_A$ mažiausia, kai $l = 0.55$, tačiau kitos paklaidos minimizuolamos $l$ reikšmė yra didesnė. Grafiškai palyginus tankių įverčius buvo pastebėta, kad MCV būdu parinkus glodinimo parametrą $l$ buvo gaunami neglodūs, daug lokalių ekstremumų turintys tankio įverčiai. Taigi šis metodas nėra tinkamas šiam glodinimo parametrui parinkti.

Taikant kryžminio patikrinimo metodą bandėme rasti optimalią parametro $l$ reikšmę intervale $[0.5; 0.8]$, tačiau taip parinkdami parametrą dažniausiai gaudavome didesnes paklai-

| Metodas | Mišinys | Pastovaus pločio CV metodas | | | Pastovaus pločio MCV metodas | | |
|---|---|---|---|---|---|---|---|
| | | $h$ | $\varepsilon_1$ | $\varepsilon_2$ | $h$ | $\varepsilon_1$ | $\varepsilon_2$ |
| Nemodifik. | A1 | 0.87955 | 0.07646 | 0.03878 | 0.36236 | 0.08993 | 0.05255 |
| Mod. 1,2,3 | | | 0.05210 | 0.03043 | | 0.11241 | 0.06600 |
| Nemodifik. | A2 | 0.91632 | 0.12076 | 0.02676 | 1.10808 | 0.11448 | 0.02582 |
| Mod. 1,2,3 | | | 0.13055 | 0.03287 | | 0.11557 | 0.02673 |
| Nemodifik. | B5 | 0.34965 | 0.16660 | 0.03838 | 1.05659 | 0.11917 | 0.03008 |
| Mod. 1,2,3 | | | 0.19241 | 0.04472 | | 0.10573 | 0.02558 |
| Nemodifik. | D1 | 0.34013 | 0.21840 | 0.04013 | 1.10138 | 0.12035 | 0.02567 |
| Mod. 1,2,3 | | | 0.25878 | 0.04820 | | 0.12055 | 0.02297 |
| Nemodifik. | D2 | 0.72827 | 0.19290 | 0.03713 | 1.35658 | 0.16606 | 0.03772 |
| Mod. 1,2,3 | | | 0.20899 | 0.03979 | | 0.17042 | 0.03606 |

4.3 lentelė: Kryžminio patikrinimo metodo ir jo modifikacijos palyginimas.

das, negu naudodami $l = 0.7$. Todėl artimiausių kaimynų metodo su parametru, parinktu kryžminio patikrinimo būdu, detaliau neaptarinėsime.

**Kryžminio patikrinimo ir modifikuoto kryžminio patikrinimo metodų palyginimas.** Pastovaus branduolio pločio tankio įverčius skaičiavome dviem būdais: kryžminio patikrinimo ir modifikuotu kryžminio patikrinimo. Naudodamiesi šių įverčių paklaidomis galėsime palyginti šiuos parametrų parinkimo metodus, nes, kaip buvo minėta aukščiau, šie metodai nepagerino artimiausių kaimynų ir tiesioginio pakeitimo įverčių. Kai kurių mišinių tyrimo rezultatai pateikti 4.3 lentelėje.

Pastebėjome, kad taikydami tiek kryžminio patikrinimo metodą (žr. mišinius B5, D1), tiek modifikuotą kryžminio patikrinimo metodą (žr. mišinį A1) kartais gauname per mažas parametro $h$ reikšmes ir tai rečiau atsitinka naudojant MCV metodą. Kaip ir tikėjomės, nemodifikuotu kryžminio patikrinimo metodu rastas branduolio plotis yra šiek tiek mažesnis. Taip yra todėl, kad kryžminio patikrinimo metodas labiau prisitaiko prie komponentės, turinčios "siauresnę" tankio funkciją (atitinkančios klasterį su mažesne dispersija), o tokios komponentės tankiui įvertinti reikia naudoti siauresnį branduolį. Rezultatai yra truputį netikėti tuo, kad modifikuoto metodo paklaidos $\varepsilon_2$ yra taip pat šiek tiek mažesnės, negu nemodifikuoto, nors būtent nemodifikuotas metodas vertina ir minimizuoja $\varepsilon_2$ paklaidas, o modifikuotas minimizuoja $\varepsilon_A$ paklaidos įverčius.

Įverčių tyrimas parodė, kad branduoliniams tankio įverčiams, kurių branduolio plotis fiksuotas, neverta taikyti multiplikatyvaus poslinkio sumažinimo ir tankio įverčio glodinimo procedūrų, nes taip modifikuoto įverčio paklaidos padidėja. 4.3 lentelėje paminėtų mišinių sudėtis:

| Metodas | Pak- | Imties tūris $n$ | | | | | |
|---|---|---|---|---|---|---|---|
| | laida | 50 | 100 | 200 | 500 | 1000 | 2000 |
| Pastovaus | $\varepsilon_1$ | 0.41806 | 0.25784 | 0.19738 | 0.11448 | 0.11917 | 0.08528 |
| pločio MCV | $\varepsilon_2$ | 0.09935 | 0.05948 | 0.04468 | 0.02582 | 0.03008 | 0.02106 |
| Artim. kaim. | $\varepsilon_1$ | **0.30593** | **0.18272** | 0.14370 | 0.11908 | 0.10868 | 0.09009 |
| mod. 1,2,3 | $\varepsilon_2$ | **0.09368** | **0.04790** | 0.03244 | 0.04052 | 0.02921 | 0.02492 |
| Ties. pakeit. | $\varepsilon_1$ | 0.34128 | 0.22402 | **0.11071** | **0.08440** | **0.07260** | **0.05081** |
| mod. 1,2,3 | $\varepsilon_2$ | 0.09687 | 0.07035 | **0.02794** | **0.02214** | **0.02090** | **0.01674** |

4.4 lentelė: Metodų tikslumo priklausomybė nuo imties tūrio.

- A1: $n = 500$, $(m, \sigma, p) = (0, 1, 1)$;
- A2: $n = 500$, $(m, \sigma, p) = (0, 1, 0.5), (20, 5, 0.5)$;
- B5: $n = 1000$, $(m, \sigma, p) = (0, 1, 0.5), (20, 5, 0.5)$;
- D1: $n = 1000$, $(m, \sigma, p) = (0, 1, 0.3333), (-10, 3, 0.3333), (25, 7, 0.3333)$;
- D2: $n = 1000$, $(m, \sigma, p) = (0, 1, 0.25), (-20, 4, 0.25), (15, 7, 0.25), (30, 2, 0.25)$.

**Įvertinimo kokybės priklausomybės nuo imties tūrio tyrimas.** Šio tyrimo metu tikrinome, kaip imties tūris įtakoja įverčių paklaidas. Tyrime vartojome mišinį, sudarytą iš dviejų skirtingo glodumo komponenčių: $(m, \sigma, p) = (0, 1, 0.5), (20, 5, 0.5)$. Tankio įvertinimo metodų paklaidos pateiktos 4.4 lentelėje. Tyrimai parodė, kad esant nedidelėms imtims geresni rezultatai gaunami naudojant modifikuotą artimiausių kaimynų metodą, o kai imtys didelės, patartina naudoti modifikuotą tiesioginio pakeitimo metodą. Tirdami imtis, sudarytas iš didesnio komponenčių kiekio, pastebėjome, kad tiesioginio pakeitimo metodo paklaidos mažesnės už kitų metodų paklaidas, kai komponentei tenka 150 ir daugiau imties taškų.

**Skirtingo glodumo tankių įvertinimo tyrimas.** Yra žinoma, kad pastovaus branduolio pločio tankio įverčiai neblogai vertina tankius, kurių glodumas įvairioms argumento $x$ reikšmėms nedaug skiriasi. Šio tyrimo tikslas — nustatyti kaip kinta įverčių tikslumas, kai tankio glodumo savybės kinta. Atsitiktinės imtys buvo sudaromos, keičiant mišinio komponenčių dispersijų santykį $\sigma_1/\sigma_2$ ir didinant atstumą tarp komponenčių vidurkių. Metodų paklaidos pateiktos 4.5 lentelėje. Mišinių parametrai yra tokie:

- A1: $n = 500$, $(m, \sigma, p) = (0, 1, 1)$;
- C2: $n = 500$, $(m, \sigma, p) = (0, 1, 0.5), (2, 3, 0.5)$;
- C3: $n = 500$, $(m, \sigma, p) = (0, 1, 0.5), (5, 3, 0.5)$;

| Metodas | Pastovaus pločio MCV metodas | | Artimiausių kaimynų metodas, mod. 1, 2, 3 | | Tiesioginio pakeitimo metodas, mod. 1, 2, 3 | |
|---|---|---|---|---|---|---|
| | $\varepsilon_1$ | $\varepsilon_2$ | $\varepsilon_1$ | $\varepsilon_2$ | $\varepsilon_1$ | $\varepsilon_2$ |
| A1 | 0.08993 | 0.05255 | 0.13957 | 0.10025 | 0.04976 | 0.03014 |
| C2 | 0.08411 | 0.02874 | 0.11552 | 0.05143 | 0.07199 | 0.02502 |
| C3 | 0.08311 | 0.02467 | 0.13175 | 0.04798 | 0.07665 | 0.02535 |
| C4 | 0.10056 | 0.02771 | 0.12138 | 0.04208 | 0.07480 | 0.02199 |
| C5 | 0.11783 | 0.02884 | 0.12061 | 0.03843 | 0.07988 | 0.02169 |
| A2 | 0.11448 | 0.02582 | 0.11908 | 0.04052 | 0.08440 | 0.02214 |
| C7 | 0.13606 | 0.03096 | 0.11858 | 0.03698 | 0.07777 | 0.02049 |
| C8 | 0.15521 | 0.03252 | 0.11915 | 0.03661 | 0.07861 | 0.01997 |

4.5 lentelė: Metodų tikslumo priklausomybė nuo tankio glodumo kitimo.

- C4: $n = 500$, $(m, \sigma, p) = (0, 1, 0.5), (10, 3, 0.5)$;
- C5: $n = 500$, $(m, \sigma, p) = (0, 1, 0.5), (16, 4, 0.5)$;
- A2: $n = 500$, $(m, \sigma, p) = (0, 1, 0.5), (20, 5, 0.5)$;
- C7: $n = 500$, $(m, \sigma, p) = (0, 1, 0.5), (30, 7, 0.5)$;
- C8: $n = 500$, $(m, \sigma, p) = (0, 1, 0.5), (40, 10, 0.5)$.

Lentelėje pateikti duomenys rodo, kad pastovaus branduolio pločio įvertis tikslesnis už adaptyvųjį artimiausių kaimynų įvertį paklaidos $\varepsilon_1$ prasme, kol tankio glodumo savybės įvairioms $x$ nesiskiria daug ($\frac{\sigma_2}{\sigma_1} \leq 5$). Didėjant skirtumui tarp glodumo savybių, pastovaus pločio MCV metodo paklaidos didėja, o artimiausių kaimynų metodo mažėja. Todėl, kai glodumo savybės pakankamai skirtingos (mišiniai C7, C8), tikslesni rezultatai gaunami naudojant artimiausių kaimynų tankio įvertinimo metodą.

Modifikuoto tiesioginio pakeitimo metodas yra tikslesnis už kitus metodus tiek $\varepsilon_1$, tiek $\varepsilon_2$ paklaidų prasme.

**Modifikavimo procedūrų efektyvumo analizė.** Tankio įverčių kokybę bandėme pagerinti taikant papildomas įverčių modifikavimo procedūras: branduolio pločio glodinimą, multiplikatyvų poslinkio koregavimą ir tankio įverčio glodinimą. Tam, kad įvertintume šių procedūrų efektyvumą, apskaičiuokime vidutinės paklaidos sumažėjimą po procedūros pritaikymo. 4.6 lentelėje pateiktos nemodifikuoto tiesioginio pakeitimo metodo ir visų jo modifikacijų paklaidų reikšmės, bei paklaidų reikšmių vidurkiai. Remdamiesi šiais dydžiais galime apskaičiuoti modifikavimo procedūrų efektyvumą. Pirma, antra ir trečia modifikacijos $\varepsilon_1$ paklaidas sumažino atitinkamai $1.058$, $1.524$ (!) ir $1.042$ karto, o $\varepsilon_2$ paklaidas

| Mišinys | Be modifikacijų | | Mod. 1 | | Mod. 1, 2 | | Mod. 1, 2, 3 | |
|---|---|---|---|---|---|---|---|---|
| | $\varepsilon_1$ | $\varepsilon_2$ | $\varepsilon_1$ | $\varepsilon_2$ | $\varepsilon_1$ | $\varepsilon_2$ | $\varepsilon_1$ | $\varepsilon_2$ |
| A1 | 0.06617 | 0.03337 | 0.07373 | 0.03433 | 0.05437 | 0.03188 | 0.04976 | 0.03014 |
| C2 | 0.10742 | 0.03973 | 0.10569 | 0.03608 | 0.07470 | 0.02635 | 0.07199 | 0.02502 |
| C3 | 0.12497 | 0.03775 | 0.10889 | 0.03117 | 0.07850 | 0.02761 | 0.07665 | 0.02535 |
| C4 | 0.13065 | 0.03568 | 0.11952 | 0.03239 | 0.07808 | 0.02403 | 0.07480 | 0.02199 |
| C5 | 0.14356 | 0.03652 | 0.13396 | 0.03292 | 0.08244 | 0.02391 | 0.07988 | 0.02169 |
| A2 | 0.12628 | 0.02962 | 0.12151 | 0.02725 | 0.08689 | 0.02459 | 0.08440 | 0.02214 |
| C7 | 0.14754 | 0.03551 | 0.14064 | 0.03208 | 0.08161 | 0.02308 | 0.07777 | 0.02049 |
| C8 | 0.15104 | 0.03565 | 0.13913 | 0.03103 | 0.08203 | 0.02267 | 0.07861 | 0.01997 |
| $\bar{\varepsilon}$ | 0.12470 | 0.03547 | 0.11788 | 0.03215 | 0.07732 | 0.02551 | 0.07423 | 0.02334 |

4.6 lentelė: Modifikacijų efektyvumo tyrimas.

atitinkamai $1.103$, $1.260$ ir $1.093$ karto. Atlikus visas modifikacijas $\varepsilon_1$ ir $\varepsilon_2$ paklaidos sumažėjo $1.680$ (!)ir $1.519$ (!) karto. Tai rodo, kad šios modifikacijos, o ypač multiplikatyvus poslinkio sumažinimas, žymiai pagerina įverčių kokybę.

Taikydami šias procedūras kitiems tankio įverčiams, detaliau ištyrėme jų savybes. Multiplikatyvaus poslinkio koregavimo procedūra sumažina tik pakankamai glodžių įverčių paklaidas. Ši procedūra sumažina tankio įverčio glodumą, todėl taikant ją neglodiems įverčiams, pastarieji yra dar labiau iškraipomi ir paklaidos padidėja. Tai lengva pastebėti tiriant artimiausių kaimynų įvertį, kurio glodumą keičiame keisdami parametrą $l$. Multiplikatyvaus poslinkio redukavimo procedūra nėra tinkama pastovaus branduolio pločio tankio įverčiams, nes kai kurioms argumento reikšmėms jų glodumas nėra pakankamas. Papildoma tankio glodinimo operacija reikalinga tam, kad atstatytume minėtą glodumo sumažėjimą.

Branduolio pločio glodinimas sumažina įverčių paklaidas beveik visais atvejais. Pastebėta, kad jis kartais nežymiai pablogina tik teorinio pseudo-įverčio savybes. Tai yra natūralu, nes šio pseudo-įverčio branduolio plotis yra teoriškai optimalus ir jo koregavimas turėtų padidinti įverčio paklaidas.

**Tiesioginio pakeitimo įverčio ir teorinio pseudo-įverčio palyginimas.** Tirdami teorinį pseudo-įvertį galime sužinoti kiek mūsų sukurtas įvertis "atsilieka" nuo teoriškai pasiekiamos ribos. Kai kuriems mišiniams gautos įverčių paklaidos pateiktos 4.7 lentelėje. Pastaba: mišinių sudėtį galite rasti prie 4.3 lentelės analizės. Iš lentelėje pateiktų duomenų galime apskaičiuoti, kad nemodifikuoto įverčio kokybė blogesnė už teoriškai pasiekiamą 1.288 karto pagal $\varepsilon_1$ ir 1.344 karto pagal $\varepsilon_1$ paklaidą. Palyginę modifikuotų metodų vidutines paklaidas pastebėsime, kad įvertis blogesnis už pseudo-įvertį tik 1.017 karto pagal $\varepsilon_1$

| Metodas | Mišinys | Tiesioginio pakeitimo įvertis | | Teorinis pseudo-įvertis | |
|---|---|---|---|---|---|
| | | $\varepsilon_1$ | $\varepsilon_2$ | $\varepsilon_1$ | $\varepsilon_2$ |
| Nemodifikuotas | A1 | 0.06617 | 0.03337 | 0.05721 | 0.03032 |
| Mod. 1, 2, 3 | | 0.04976 | 0.03014 | 0.05472 | 0.03176 |
| Nemodifikuotas | A2 | 0.12628 | 0.02962 | 0.09519 | 0.02307 |
| Mod. 1, 2, 3 | | 0.08440 | 0.02214 | 0.07314 | 0.02360 |
| Nemodifikuotas | B5 | 0.14638 | 0.04135 | 0.07834 | 0.01960 |
| Mod. 1, 2, 3 | | 0.07260 | 0.02090 | 0.06700 | 0.01836 |
| Nemodifikuotas | D1 | 0.14724 | 0.03073 | 0.11080 | 0.02107 |
| Mod. 1, 2, 3 | | 0.08703 | 0.01963 | 0.08481 | 0.01917 |
| Nemodifikuotas | D2 | 0.17055 | 0.03377 | 0.16840 | 0.03161 |
| Mod. 1, 2, 3 | | 0.12200 | 0.02884 | 0.12897 | 0.02890 |

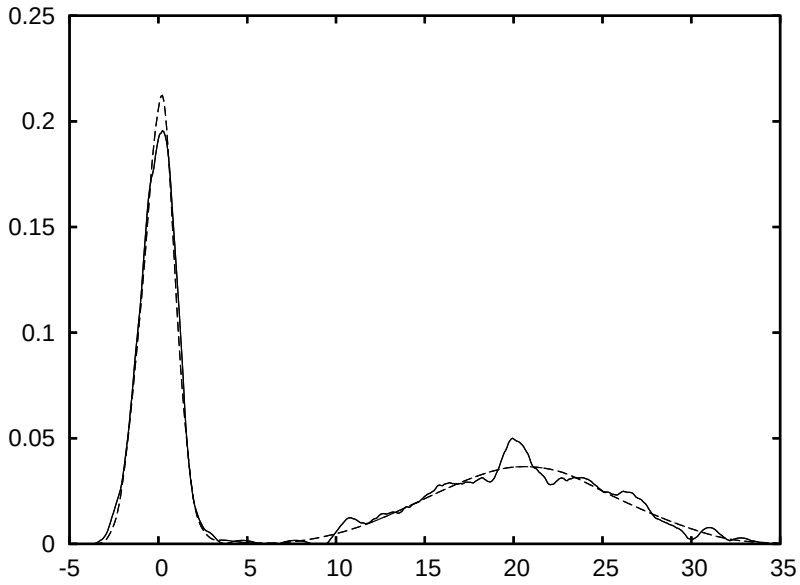
4.7 lentelė: Tiesioginio pakeitimo įverčio ir teorinio pseudo-įverčio palyginimas.

paklaidą, bet geresnis už jį pagal $\varepsilon_2$ paklaidą.

**Grafinis tankio įverčių palyginimas.** Tankio įverčiai turi būti ne tik optimalūs įvairių paklaidų prasme, bet ir gerai atitikti tikrojo tankio formą. Todėl įverčiai buvo tikrinami vizualiai. Tyrimai parodė, kad tiesioginio pakeitimo įverčiai gerai atitinka tikrąją tankio konfigūraciją, pastovaus pločio įverčiai blogai vizualiai atspindi mišinių komponentes su didele dispersija, o artimiausių kaimynų metodas iškraipo tankio formą mišinio komponenčių tankio viršūnėse. Pateiksime kai kurių tankio funkcijų įverčių grafikus. Imties parametrai yra: $n = 500$, $(m, \sigma, p) = (0, 1, 0.5), (20, 5, 0.5)$. Pastebėsime, kad 4.1 pav. parodytu atveju, tiesioginio pakeitimo metodo paklaidos $\varepsilon_1$ ir $\varepsilon_2$ yra šiek tiek didesnės, negu pastovaus pločio kryžminio patikrinimo metodo, tačiau tiesioginio pakeitimo metodo įvertis geriau atitinka tikrojo tankio formą.

## 4.2 Daugiamačio Gauso mišinio analizės procedūrų tikslumo tyrimas

Siūlomi daugiamačio Gauso mišinio identifikavimo ir duomenų klasterizavimo metodai buvo tiriami Monte-Karlo metodu. Buvo generuojamos atsitiktinės imtys (1.4) su nepriklausomais stebėjimas pasiskirsčiusiais pagal Gauso mišinio skirstinį. Norint įvairiapusiškai ištirti siūlomus metodus, buvo varijuojama imties dydžiu, klasterių kiekiu, jų tikimybėmis, vidurkiais ir kovariacinėmis matricomis, stengiantis aprėpti įvairias mišinio konfigūracijas: atsiskiriančius klasterius, persidengusius klasterius; panašius klasterių svorius, reikšmingai skirtingus klasterių svorius; mažas, dideles imtis ir pan.




4.1 pav.: Tankių įverčiai gauti tiesioginio pakeitimo metodu (taškinė linija) ir pastovaus pločio kryžminio patikrinimo metodu (ištisinė linija).

**Paklaidos.** Kadangi tikrasis imties skirstinys buvo žinomas, galėjome suskaičiuoti įverčių paklaidas ir jas lygindami darėme išvadas apie įverčių kokybę. Tiriamiems metodams palyginti naudojome kelias paklaidas. Stebėjimų klasterizavimo kokybei įvertinti naudojome vidutinę negriežto klasifikavimo paklaidą

$$\epsilon_\pi = \mathbb{E} \sum_{i=1}^{q} |\widehat{\pi}(j, X) - \pi(j, X)| \tag{4.6}$$

ir griežto klasifikavimo paklaidą

$$\epsilon_\nu = \mathbb{P}\{\widehat{\nu}(X) \neq \nu(X)\}. \tag{4.7}$$

Parinkus modelio parametrą $\theta$ ir imties tūrį $n$, generuojamos nepriklausomos atsitiktinės imtys $\mathbb{X}^{(s)} = (X^{(s)}(1), ..., X^{(s)}(n))$, $s = \overline{1, r}$ ir randamos nagrinėjamų įverčių realizacijos.

Tuomet skaičiuojami paklaidų (4.6) ir (4.7) empiriniai analogai

$$\varepsilon_\pi = \frac{1}{rn} \sum_{\substack{j=\overline{1,q} \\ t=\overline{1,n} \\ s=\overline{1,r}}} |\widehat{\pi}^{(s)}(j, X^{(s)}(t)) - \pi(j, X^{(s)}(t))|, \tag{4.8}$$

$$\varepsilon_\nu = \frac{1}{rn} \sum_{s=\overline{1,r}} \|\{t : \widehat{\nu}^{(s)}(X^{(s)}(t)) \neq \nu(X^{(s)}(t))\}\|. \tag{4.9}$$

Šiame darbe visuose eksperimentuose $r = 10$.

Praktikoje gali būti sutinkamas ne tik klasifikavimo uždavinys, bet ir Gauso mišinio parametrų arba tankio įvertinimo uždavinys. Kadangi panašius tankius galima nusakyti visai skirtingais parametrais, tai tikslumui įvertinti naudosime $L_2$ atstumą tarp tankio ir jo įverčio

$$\epsilon_{L2} = \mathbb{E} \int_{\mathbb{R}^d} \left(f_{\widehat{\theta}}(x) - f_\theta(x)\right)^2 dx. \tag{4.10}$$

Siekdami palyginti įverčių tikslumą srityse, kur tankio reikšmė yra didelė (tokių įverčių aktualumas pagrįstas [142]), naudosime paklaidą

$$\epsilon_{L2F} = \mathbb{E} \int_{\mathbb{R}^d} \left(f_{\widehat{\theta}}(x) - f_\theta(x)\right)^2 f(x)\, dx. \tag{4.11}$$

Šių paklaidų empiriniai analogai skaičiuojami naudojant formules

$$\varepsilon_{L2} = \frac{1}{r} \sum_{s=\overline{1,r}} \int_{\mathbb{R}^d} (f_{\widehat{\theta}}^{(s)}(x) - f_\theta(x))^2\, dx, \tag{4.12}$$

$$\varepsilon_{L2F} = \frac{1}{rn} \sum_{\substack{t=\overline{1,n} \\ s=\overline{1,r}}} \left(f_{\widehat{\theta}}^{(s)}(X^{(s)}(t)) - f_\theta(X^{(s)}(t))\right)^2. \tag{4.13}$$

Pastebėsime, kad pasinaudojus savybe

$$\int_{\mathbb{R}^d} \varphi(x; M_1, R_1)\, \varphi(x; M_2, R_2)\, dx = \varphi(M_1 - M_2; 0, R_1 + R_2), \tag{4.14}$$

išraiškoje (4.12) esantį integralą galime suskaičiuoti analitiškai

$$\begin{aligned}
&\int\limits_{\mathbb{R}^d} (f_{\widehat{\theta}}(x) - f_\theta(x))^2\,dx = \int\limits_{\mathbb{R}^d} \left( \sum_{i=1}^{\widehat{q}} \widehat{p}_i \widehat{\varphi}_i(x) - \sum_{i=1}^{q} p_i \varphi_i(x) \right)^2 dx = \\
&\quad = \sum_{i,j=1}^{\widehat{q}} \widehat{p}_i \widehat{p}_j \int\limits_{\mathbb{R}^d} \widehat{\varphi}_i(x) \widehat{\varphi}_j(x)\,dx - 2 \sum_{\substack{i=\overline{1,\widehat{q}} \\ j=\overline{1,q}}} \widehat{p}_i p_j \int\limits_{\mathbb{R}^d} \widehat{\varphi}_i(x) \varphi_j(x)\,dx \\
&\qquad + \sum_{i,j=1}^{q} p_i p_j \int\limits_{\mathbb{R}^d} \varphi_i(x) \varphi_j(x)\,dx \qquad (4.15) \\
&\quad = \sum_{i,j=1}^{\widehat{q}} \widehat{p}_i \widehat{p}_j\, \varphi(\widehat{M}_i - \widehat{M}_j; 0, \widehat{R}_i + \widehat{R}_j) - 2 \sum_{\substack{i=\overline{1,\widehat{q}} \\ j=\overline{1,q}}} \widehat{p}_i p_j\, \varphi(\widehat{M}_i - M_j; 0, \widehat{R}_i + R_j) \\
&\qquad + \sum_{i,j=1}^{q} p_i p_j\, \varphi(M_i - M_j; 0, \widehat{R}_i + R_j)
\end{aligned}$$

Jeigu metodas nevertina parametro $\theta$, o yra randamas neparametrinis tankio įvertis, tai išraiška (4.12) keičiama

$$\varepsilon_{L2} = \frac{1}{rn} \sum_{\substack{t=\overline{1,n} \\ s=\overline{1,r}}} \frac{\left( \widehat{f}^{(s)}(X^{(s)}(t)) - f_\theta(X^{(s)}(t)) \right)^2}{f_\theta(X^{(s)}(t))}, \qquad (4.16)$$

o išraiškoje (4.13) parametrinis tankio įvertis $f^{(s)}_{\widehat{\theta}}$ keičiamas neparametriniu $\widehat{f}^{(s)}$.

Modeliavimo tyrimai parodė, kad paklaidos $\varepsilon_\pi$ ir $\varepsilon_\nu$ elgiasi panašiai, t.y. metodai, tikslūs vienos paklaidos prasme, bus tikslūs ir kitos paklaidos prasme. Daugelyje atvejų panašiai elgiasi ir $\varepsilon_{L2}$ bei $\varepsilon_{L2F}$ paklaidos. Pastebėsime, kad maksimalaus tikėtinumo medodo įvertis paprastai būdavo tiksliausias $\varepsilon_\pi$ ir $\varepsilon_\nu$ prasme, tačiau kartais nusileisdavo kitiems įverčiams $\varepsilon_{L2}$ (ir dar dažniau $\varepsilon_{L2F}$) prasme.

**Tirti metodai.** Aprašyti metodai gali būti įvairiai kombinuojami siekiant pagerinti įverčio kokybę. Pažymėsime aukščiau aprašytus metodus, kad galėtume sutrumpintai užrašyti sudėtinio metodo schemą. Aprašyti metodai remiasi vienamačių duomenų klasterizavimo paketu, sukurtu Matematikos ir informatikos institute Taikomosios statistikos skyriuje. šis stebėjimų klasterizavimo paketas remiasi idėjomis, aprašytomis [142]. Jeigu tiriami metodai naudoja ne tikrus vienamačių duomenų projekcijų parametrų įverčius, gautus naudojant minėtą paketą, o pseudo-įverčius, gautus EM algoritmu, paleistu iš teorinių

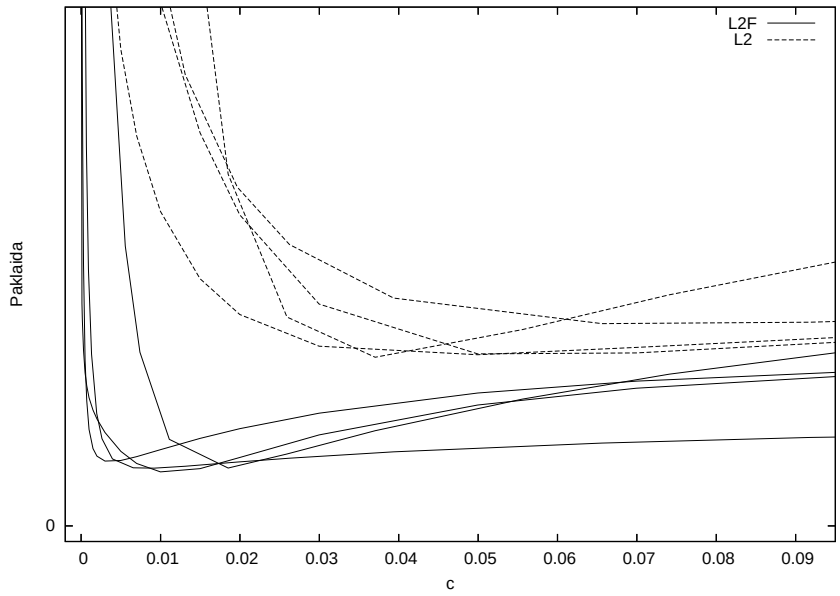
parametro reikšmių, virš procedūrų žymėsime ženklą "$\widetilde{\ }$", t.y. pvz., $\widetilde{LSP}$ reikš, kad naudojamas mažiausių kvadratų metodo įvertis, naudojantis duomenų vienamačių projekcijų pseudo-įverčius. Pastebėsime, kad minėtieji pseudo-įverčiai yra maksimalaus tikėtinumo metodo įverčiai. Naudosime šiuos pažymėjimus:

| Žymuo | Apibrėžimas |
|---|---|
| $EM$ | Rekurentinis EM algoritmas apibrėžtas (1.10) ir (1.13) |
| $I$ | Tankio įvertinimas, pagrįstas apvertimo formule ir apibrėžtas (3.37) |
| $LSP$ | Mažiausių kvadratų metodo įvertis, pagrįstas atstumų tarp projektuotų parametrų minimizavimu, apibrėžtas (3.38) |
| $LSD$ | Mažiausių kvadratų metodo įvertis, pagrįstas atstumų tarp projektuotų tankių minimizavimu, apibrėžtas (3.41) |
| $G$ | Geometrinis klasterizavimo metodas, aprašytas skyriuje 3.3 |

Siūlomų statistinių procedūrų kombinacijas žymėsime atskiras procedūras skirdami "–". Pvz., $G - EM$ reikš metodą, kai stebėjimų klasifikavimas randamas geometrinio klasterizavimo būdu, ir šis rezultatas naudojamas kaip pradinis taškas EM algoritmui.

Minėti duomenų analizės metodai buvo lyginami su maksimalaus tikėtinumo įvertiniu. Pastarasis buvo rastas naudojant EM algoritmą iš teorinių parametro reikšmių. Šį pseudo-įvertinį žymėsime $MLE$.

$I$ **metodas.** Apvertimo formule pagrįstas metodas turi glodinimo parametrą $h$. Modeliavimo tyrimai parodė, kad šis metodas yra jautrus parametro parinkimui — parinkus per mažą $h$ reikšmę įvertis tampa labai neglodžiu ir turi dideles paklaidas. Pernelyg suglodinus tankio įvertį, jo kokybė labai nenukenčia. Atliekant tyrimus buvo pastebėta, kad įvertis tampa neglodžiu dėl to, kad kai kuriose kryptyse, stebėjimų projekcijų reikšmės yra panašios, ir klasterizavimo procedūra išskiria mažo svorio klasterius su mažomis dispersijomis. Todėl vienas iš būdų padaryti tankį glodesniu yra nenaudoti krypčių su dispersija ar svoriu, mažesniais už pasirinktą minimalią reikšmę. Tačiau taip išmetant kryptis galime išmesti visą sritį, todėl siūlome naudoti glodinimo parametrą $h$. Glodinimo plotį $h$ logiška parinkti proporcingą pradinių duomenų išsibarstymui (mastelio parametrui), todėl siūlome naudoti $h = c\sqrt{\lambda_{max}}$, kur $\lambda_{max}$ yra duomenų kovariacinės matricos didžiausia tikrinė reikšmė. Metodo paklaidų priklausomybės grafikai keliems iš tirtų pateikti 4.2 paveikslėlyje. Optimali $c$ reikšmė daugelyje iš tirtų atvejų priklausė intervalui $[0.002; 0.02]$, todėl galima naudoti pvz., $c = 0.015$. Pastebėsime, kad netgi suglodinus tankio įvertį, vis tiek išlieka svyruojančios šio įverčio "uodegos", kurios kai kurioms argumento reikšmėms įgyja nei-




4.2 pav.: Paklaidų $\varepsilon_{L2F}$ ir $\varepsilon_{L2}$ priklausomybė nuo glodinimo parametro $c$. Pateikiami grafikai gauti esant kelioms skirtingoms $\theta$ reikšmėms. Iš grafikų matome, kad $\varepsilon_{L2F}$ paklaida yra mažiausia, kai $c \approx 0.1$, o $\varepsilon_{L2}$ mažiausia, kai $c \approx 0.5$. Ordinačių ašies mastelis skirtingiems grafikams yra skirtingas.

giamas reikšmes. Dėl šios priežasties negalime tiksliau parinkti glodinimo parametro $c$ naudodamiesi tikėtinumo funkcija. Be to, esant didelėms santykinėms paklaidoms, įverčio "uodegos" žymiai padidina paklaidų $\varepsilon_{L2}$ dydį, nors $\varepsilon_{L2F}$ paklaidų prasme šis metodas gali konkuruoti su kitais. Todėl galime daryti išvadą, kad metodas $I$ tinkamas tankiui vertinti ten, kur tankio reikšmė yra didelė.

Taip pat modeliavimo tyrimai parodė, kad norint gauti pakankamai tikslius $I$ metodo įverčius, reikia naudoti didelį projektavimo krypčių kiekį, pvz., 10000 krypčių.

Panaudoti apvertimo formule pagrįsto metodo duomenims klasterizuoti nepavyko. Buvo testuotas $\widetilde{I}$ metodas, kuris naudojo suderintus tarpusavyje stebėjimų projekcijų pasiskirstymo parametrų maksimalaus tikėtinumo pseudo-įverčius, tačiau pakankamai tiksliai įvertinti atskirų klasterių tankio dedamųjų nepavyko — jų reikšmės buvo labai neglodžios ir paklaidos būdavo kelis ar net kelioliką kartų didesnės nei kitų tankio įvertinimo metodų.

$LSP$ **metodas.** Siekiant išsiaiškinti metodo galimybes buvo tiriamas pseudo-įvertis $\widetilde{LSP}$, kuris buvo paremtas vienamačiais maksimalaus tikėtinumo metodo įverčiais. Modeliavimo tyrimas parodė, kad šis metodas visais atvejais nusileido $MLE$ įverčiui, tačiau patikslinus jį $EM$ algoritmu, buvo gaunamas maksimalaus tikėtinumo metodo įvertis.

Tačiau vienas iš šio metodo trūkumų yra tas, kad jam reikalingi suderinti kryptyse duo-

menų projekcijų parametrų įverčiai, o tam reikia turėti pagalbinį pradinį duomenų suskaldymą. Todėl šis metodas gali būti tik naudojamas derinant jį su kitais metodais, pvz., geometrinio klasterizavimo algoritmu.

Pakankamas pasirinktų krypčių kiekis reikalingas šiam metodui yra mažesnis, negu $I$ metodui. Kai $dimX = 5$, pakakdavo naudoti 1000 projektavimo krypčių.

$LSD$ **metodas.** Šis metodas pranašesnis už $LSP$ metodą tuo, kad nereikalauja, kad projektuotų stebėjimų pasiskirstymo parametrų įverčiai būtų suderinti, tačiau tai yra iteracinis parametro įverčio patikslinimo metodas, ir jam reikalinga pradinė parametro reikšmė. Savo tyrimuose pradine parametro reikšme laikėme $MLE$ įvertinį (kuris praktikoje gali būti randamas pvz., derinant geometrini klasterizavimą ir $EM$ algoritmą). Tyrimas parodė, kad kai kuriems mišiniams metodas duoda tikslesnius parametrų įverčius (mūsų tirtų keturių paklaidų prasme), negu $MLE$ įverčio. Tačiau kitais atvejais metodas nebuvo toks tikslus. Viena iš galimų to priežasčių gali būti tai, kad $LSD$ metodu minimizuojami nuostoliai nėra $L2$ nuostoliai tarp tankio ir jos įverčio. Geresnių rezultatų būtų galima tikėtis minimizuojant $L2$ paklaidą

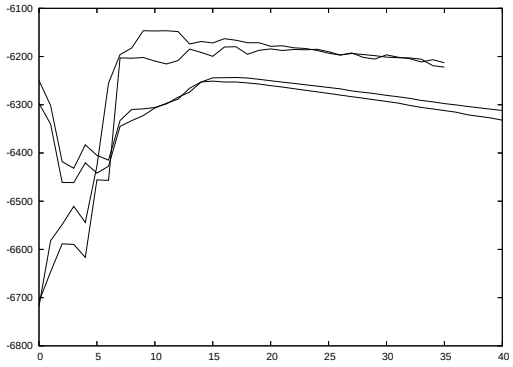
$$\begin{aligned}\int\limits_{\mathbb{R}^d} \Big(\widehat{f}(x) - f(x)\Big)^2\, dx &= \int\limits_{\tau:|\tau|=1} ds \int\limits_0^\infty \Big(\widehat{f}(t\tau) - f(t\tau)\Big)^2 u^{d-1}\, du \\ &\approx c(d) \sum_{\tau\in T} \int\limits_0^\infty \Big(\widehat{f}(t\tau) - f(t\tau)\Big)^2 u^{d-1}\, du\end{aligned} \tag{4.17}$$

Galime spėti, kad svorinės funkcijos naudojimas po integralo ženklu gali pagerinti $LSD$ metodą, tačiau tai reikalauja papildomų tyrimų. Tiesiogiai panaudotas daugiklis $u^{d-1}$ neduos rezultato, nes $f(t\tau) \neq f_\tau(t)$.

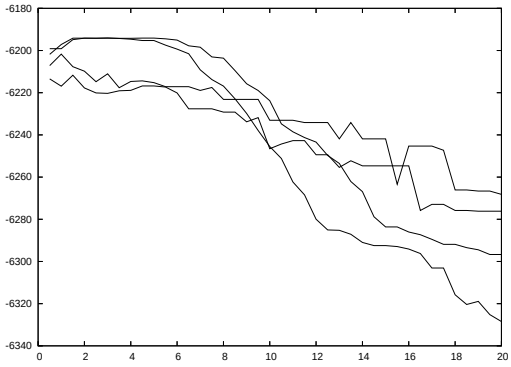
Šiam metodui, kaip ir $LSP$, kai $dimX = 5$ pakakdavo naudoti 1000 projektavimo krypčių.

$G$ **metodas.** Geometrinis duomenų klasterizavimas — vienas iš metodų, kuriam nereikia suderintų duomenų projekcijų pasiskirstymo parametro įverčių, todėl jis gali būti naudojamas kaip pradinis įvertis daugiamačiame duomenų suskaidyme. Vėliau $G$ metodu gauti įverčiai gali būti tikslinami kitais metodais.

Buvo pasiūlytos kelios $\rho_\tau$ ir $\rho$ pseudoatstumų funkcijų apibrėžimo alternatyvos. Atliekant modeliavimo tyrimą buvo stengiamasi išsiaiškinti, kurią šių funkcijų naudoti norint pasiekti tiksliausių rezultatų. Buvo pastebėta, kad naudojant išraiškas (3.65) ir (3.66) gaunami panašūs rezultatai.




4.3 pav.: Logaritmuotos tikėtinumo funkcijos reikšmės priklausomybę nuo geometrinio klasterizavimo algoritmo parametro $\alpha$.

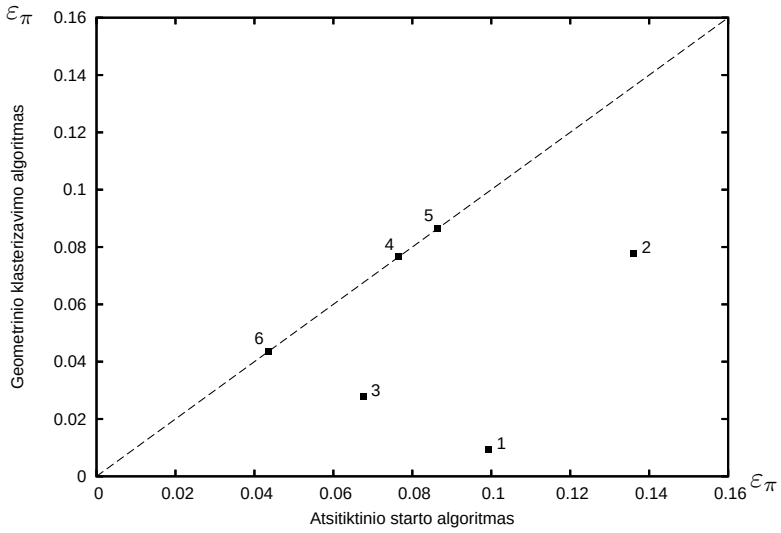



4.4 pav.: Logaritmuotos tikėtinumo funkcijos reikšmės priklausomybę nuo geometrinio klasterizavimo algoritmo parametro $p$.

Pseudoatstumą tarp daugiamačių stebėjimų galėjome apibrėžti naudodami (3.68) arba (3.69). Be to, abiem atvejais buvo sprendžiamas metodo parametrų parinkimo klausimas: reikėjo parinkti parametrą $\alpha$ arba $p$. Modeliavimo tyrimas parodė, kad optimalios šių parametrų reikšmės dažniausiai priklauso intervalams $\alpha \in [0, 0.25]$ ir $p \in [2, 8]$. Kadangi vienas iš pagrindinių disertacinio darbo uždavinių yra pasiūlyti algoritmą pradinei EM algoritmo parametro reikšmei parinkti taip, kad šis algoritmas konverguotų į MTM įvertį, o pasirinkimo kriterijumi galime naudoti tikėtinumo funkciją, tai siūlome šį metodą derinti su EM algoritmu, t.y. naudoti $G - EM$ metodą, o nežinomus parametrus parinkti iš nurodytų intervalų, remiantis tikėtinumo funkcijos reikšme. Paveikslėliuose 4.3 ir 4.4 vaizduojamos $G$ metodo logaritmuotos tikėtinumo funkcijos reikšmės priklausomai nuo parametrų $\alpha$ ir $p$ parinkimo.

Pastebėsime, kad daugelyje tirtų atvejų $G - EM$ metodas duodavo tokį pat rezultatą kaip ir $MLE$ įvertinys, todėl siūlome šį metodą naudoti konstruktyviam maksimalaus tikėtinumo metodo įverčiui skaičiuoti.

**Palyginimas su kitais algoritmais.** Siūlomų metodų tikslumas buvo lyginamas ne tik su MTM įvertiniu, bet ir su atsitiktinio starto procedūra. Šis procedūra naudojama EM algoritmui inicializuoti, ir ji yra ypač populiari statistine programinėje įrangoje. Tam, kad išvengti centrinės ribinės teoremos pasireiškimo, kai atsitiktinai parinkus pradinį stebėjimų suskaldymą, mišinio komponenčių parametrai tampa panašiais, buvo naudojama tik dalis imties taškų parametrui inicializuoti (angl. *subsampling*), taip kaip rekomenduojama [111]. Daugelyje tirtų atvejų EM algoritmui startuotant iš atsitiktinės mišinio parametro reikšmės, MTM įvertinys buvo randamas per 20–200 atsitiktinio starto bandymų. Tačiau,




4.5 pav.: Atsitiktinio starto ir geometrinio klasterizavimo procedūrų palyginimas. EM algoritmas buvo pradedamas nuo pradinės parametro reikšmes, kuri randama naudojant minėtas procedūras. Grafike atidėtos klasifikavimo tikslumą atspindinčios paklaidos $\varepsilon_\pi$.

jei apriorinės klasterių tikimybės skirdavosi reikšmingai, atsitiktinio starto procedūra nerasdavo klasterio, kurio tikimybė maža, net jei procedūra buvo kartojama labai daug kartų (buvo bandoma 4000 atsitiktinio starto taškų). Šio modeliavimo tyrimo rezultatai pateikti 4.5 paveikslėlyje[1]. Taškai pažymėti 1, 2, ir 3 atitinka mišinius su reikšmingai skirtingomis klasterių tikimybėmis. Esant tokioms parametrų reikšmėms, geometrinio klasterizavimo procedūra inicializuota EM algoritmas yra tikslesnis nei inicializuotas atsitiktinio starto procedūra. Kitais atvejais (taškai 4, 5 ir 6) abi procedūros užtikrino mišinio parametrų konvergavimą į MTM įvertinį, todėl abiejų metodų paklaidos yra lygios ir taškai išsidėstę ant pusiaukampinės.

[1]Tirtų mišinių klasterių tikimybės. Mišinys 1: $q = 2, p_1 = 0.92, p_2 = 0.08$. Mišinys 2: $q = 3, p_1 = p_3 = 0.05, p_2 = 0.9$. Mišinys 3: $q = 2, p_1 = 0.9, p_2 = 0.1$. Mišinys 4: $q = 4, p_1 = p_3 = 0.2, p_2 = p_4 = 0.3$. Mišinys 5: $q = 3, p_1 = p_3 = 0.35, p_2 = 0.3$. Mišinys 6: $q = 4, p_1 = p_2 = p_3 = p_4 = 0.25$.

| Metodas | $\varepsilon_{L2F} \times 10^3$ | $\varepsilon_{L2}$ | $\varepsilon_{\pi}$ | $\varepsilon_{\nu}$ | log-tikėtinumas |
|---|---|---|---|---|---|
| Mišinys 1 | | | | | |
| $MLE$ | 0.102915 | 0.0042922 | 0.02967 | 0.062 | -5082.6141 |
| $I$ | 0.147062 | 0.0142949 | - | - | - |
| $G$ | 0.094649 | 0.0040880 | 0.08669 | 0.084 | -5112.0377 |
| $G-EM$ | 0.102915 | 0.0042922 | 0.02967 | 0.062 | -5082.6141 |
| $\widetilde{LSP}$ | 0.131447 | 0.0066076 | 0.08317 | 0.100 | -6288.3265 |
| $\widetilde{LSP}-EM$ | 0.102915 | 0.0042922 | 0.02967 | 0.062 | -5082.6141 |
| $\widetilde{LSD}$ | 0.102894 | 0.0043588 | 0.02926 | 0.058 | -5122.4386 |
| $\widetilde{LSD}-EM$ | 0.102915 | 0.0042922 | 0.02967 | 0.062 | -5082.6141 |
| Mišinys 2 | | | | | |
| $MLE$ | 0.085176 | 0.0042035 | 0.10651 | 0.175 | -5361.9998 |
| $G$ | 0.211098 | 0.0049300 | 0.28909 | 0.324 | -5498.4503 |
| $G-EM$ | 0.129791 | 0.0050920 | 0.15941 | 0.211 | -5367.4009 |
| $\widetilde{LSP}$ | 0.164251 | 0.0047840 | 0.13045 | 0.209 | -5382.0408 |
| $\widetilde{LSP}-EM$ | 0.085176 | 0.0042035 | 0.10651 | 0.175 | -5361.9998 |
| $\widetilde{LSD}$ | 0.106742 | 0.0049146 | 0.12983 | 0.204 | -5380.1543 |
| $\widetilde{LSD}-EM$ | 0.085176 | 0.0042035 | 0.10651 | 0.175 | -5361.9998 |
| Mišinys 3 | | | | | |
| $MLE$ | 0.095754 | 0.0037712 | 0.04477 | 0.140 | -5790.7939 |
| $I$ | 0.124311 | 0.0299140 | - | - | - |
| $G$ | 0.078399 | 0.0035799 | 0.07827 | 0.149 | -5851.6277 |
| $G-EM$ | 0.095754 | 0.0037712 | 0.04477 | 0.140 | -5790.7939 |
| $\widetilde{LSP}$ | 0.106307 | 0.0042097 | 0.06947 | 0.148 | -5815.6094 |
| $\widetilde{LSP}-EM$ | 0.095754 | 0.0037712 | 0.04477 | 0.140 | -5790.7939 |
| $\widetilde{LSD}$ | 0.087631 | 0.0034004 | 0.04084 | 0.142 | -5797.1964 |
| $\widetilde{LSD}-EM$ | 0.095754 | 0.0037712 | 0.04477 | 0.140 | -5790.7939 |
| Mišinys 4 | | | | | |
| $MLE$ | 0.074533 | 0.0027717 | 0.00046 | 0.000 | -5303.8939 |
| $G$ | 0.074468 | 0.0028200 | 0.00047 | 0.000 | -5311.9212 |
| $G-EM$ | 0.074533 | 0.0027717 | 0.00046 | 0.000 | -5303.8939 |
| Mišinys 5, $n = 150$ | | | | | |
| $MLE$ | 0.017137 | 0.0021385 | 0.08650 | 0.113 | -1773.9857 |
| $I$ | 0.024355 | 0.0120174 | - | - | - |
| $G$ | 0.014452 | 0.0020510 | 0.12760 | 0.160 | -1825.7856 |
| $G-EM$ | 0.016509 | 0.0021073 | 0.09295 | 0.141 | -1781.9726 |
| $\widetilde{LSD}$ | 0.024882 | 0.0025698 | 0.09529 | 0.113 | -1805.9785 |
| $\widetilde{LSD}-EM$ | 0.017137 | 0.0021385 | 0.08650 | 0.113 | -1773.9857 |

4.8 lentelė: Paklaidų lentelė

| Metodas | $\varepsilon_{L2F} \times 10^3$ | $\varepsilon_{L2}$ | $\varepsilon_\pi$ | $\varepsilon_\nu$ | log-tikėtinumas |
|---|---|---|---|---|---|
| Mišinys 5, $n = 500$ | | | | | |
| $MLE$ | 0.011546 | 0.0014833 | 0.05609 | 0.110 | -6197.6356 |
| $I$ | 0.020124 | 0.0042214 | - | - | - |
| $G$ | 0.018089 | 0.0017667 | 0.12665 | 0.164 | -6227.7968 |
| $G - EM$ | 0.011546 | 0.0014833 | 0.05609 | 0.110 | -6197.6356 |
| $\widetilde{LSD}$ | 0.010702 | 0.0012891 | 0.05049 | 0.111 | -6210.1840 |
| $\widetilde{LSD} - EM$ | 0.011546 | 0.0014833 | 0.05609 | 0.110 | -6197.6356 |
| Mišinys 6 | | | | | |
| $MLE$ | 0.174804 | 0.0075923 | 0.07652 | 0.130 | -2084.7917 |
| $G$ | 0.162687 | 0.0065076 | 0.12903 | 0.185 | -2120.3574 |
| $G - EM$ | 0.174877 | 0.0076015 | 0.07686 | 0.130 | -2084.7918 |
| $\widetilde{LSD}$ | 0.185266 | 0.0078729 | 0.08052 | 0.125 | -2091.9379 |
| $\widetilde{LSD} - EM$ | 0.174804 | 0.0075923 | 0.07652 | 0.130 | -2084.7917 |

4.8 lentelė: Paklaidų lentelė (tęsinys)

| Paklaida | Imtis | $MLE$ | $\widetilde{G}$ | $\widetilde{G} - EM$ | $G$ | $G - EM$ |
|---|---|---|---|---|---|---|
| $\varepsilon_\pi$ | $n = 50$ | 0.01987 | 0.06111 | 0.07903 | 0.10053 | 0.11879 |
| | $n = 100$ | 0.01478 | 0.02578 | 0.01504 | 0.02773 | 0.01504 |
| | $n = 200$ | 0.01397 | 0.01814 | 0.01397 | 0.02315 | 0.01397 |
| | $n = 500$ | 0.00959 | 0.01748 | 0.00959 | 0.01753 | 0.00959 |
| $\varepsilon_{L2} \times 10^2$ | $n = 50$ | 0.41735 | 0.46795 | 0.47304 | 0.46489 | 0.48914 |
| | $n = 100$ | 0.26139 | 0.20742 | 0.23366 | 0.21509 | 0.23366 |
| | $n = 200$ | 0.14650 | 0.13488 | 0.14650 | 0.13682 | 0.14650 |
| | $n = 500$ | 0.07321 | 0.08112 | 0.07321 | 0.08265 | 0.07321 |

4.9 lentelė: Geometrinio klasifikavimo metodo paklaidų priklausomybė nuo imties dydžio. Naudotas mišinys nr. 7.

# Išvados

Remiantis tyrimų rezultatais, gynimui pateikiamos šios išvados:

1. Pasiūlyta neparametrinė vienamačio pasiskirstymo tankio statistinio įvertinimo procedūra tiksliau vertina Gauso skirstinių mišinio tankį nei žinomuose statistiniuose paketuose naudojamos neparametrinės vertinimo procedūros, jei tankio lokalaus glodumo charakteristikos labai priklauso nuo argumento.

2. Siūloma duomenų projektavimu paremta Gauso skirstinių mišinių identifikavimo metodika yra konstruktyvi ir apskaičiuoja MTM įvertį tiksliau negu kiti konstruktyvūs algoritmai.

3. Mažų imčių atveju, naudojant duomenų projektavimu pagrįstą procedūra sudarytą iš geometrinio klasterizavimo, EM algoritmo ir mažiausių kvadratų metodo, kai sumuojamos projektuotų duomenų pasiskirsymo tankių įverčių paklaidos $L_2$ metrikoje, galima gauti tikslesnius Gauso skirstinių mišinio parametrų įverčius negu maksimalaus tikėtinumo metodo įverčiai.

4. Sprendžiant daugiamačių Gauso skirstinių mišinio identifikavimo ir duomenų klasterizavimo uždavinius tikslinga naudoti duomenų projektavimą.

# Literatūra


[1] Adamson I.S. (1982), On Bandwidth Variation in Kernel Estimates - A Square Root Law, *The Annals of Statistics*, **10**, pp. 1217–1223.

[2] Aggarwal C.C., Procopiuc C., Wolf J.L., Yu P.S., Park J.S. (1999), A Framework for Finding Projected Clusters in High Dimensional Spaces, *Proc. of ACM SIGMOD International Conference on Management of Data*, Philadelphia, Pennsylvania, U.S., pp. 61–72.

[3] Agrawal R., Gehrke J., Gunopulos D., Raghavan P., Automatic Subspace Clustering of High Dimensional Data for Data Mining, http://citeseer.ist.psu.edu/

[4] Akaike H. (1974), A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, **19**, pp. 716–723.

[5] Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D., Levine A.J. (1999), Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, *Cell Biology*, **99**, pp. 6745–6750.

[6] Ammann L.P. (1993), Robust Singular Value Decompositions: A New Approach to Projection Pursuit, *Journal of the American Statistical Association*, **88** (422), pp. 505–514.

[7] Anderberg M.R. (1973), *Cluster Analysis for Applications*, New York: Academic Press.

[8] Banfield J.D., Raftery A.E. (1993), Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics*, **49**, pp. 803–821.

[9] Behboodian J. (1970), On a Mixture of Normal Distributions, *Biometrika*, **57**, pp. 215–217.

[10] Bensmail H., Celeux G. (1996), Regularized Gaussian Discriminant Analysis Through Eigenvalue Decomposition, *Journal of the American Statistical Association*, **91**, pp. 1743–1748.

[11] Bensmail H., Celeux G., Raftery A.E., Robert C.P. (1997), Inference in Model-Based Cluser Analysis, *Statistics and Computing*, **7**, pp. 1–10.

[12] Berkhin P. (2002), Survey Of Clustering Data Mining Techniques, *Technical Report*, Accrue Software, San Jose, U.S., http://citeseer.ist.psu.edu/berkhin02survey.html

[13] Biernacki C., Celeux G., Govaert G. (1999), An Improvement of the NEC Criterion for Assessing the Number of Clusters in Mixture Model, *Pattern Recognition Letters*, **20**, pp. 267–272.

[14] Biernacki C., Celeux G., Govaert G. (2000), Assessing a Mixture Model for Clustering With the Integrated Completed Likelihood, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, pp. 719–725.

[15] Biernacki C., Govaert G. (1999), Choosing Models in Model-Based Clustering and Discriminant Analysis, *Journal of Statistical Computation and Simulation*, **64**, pp. 49–71.

[16] Binder D.A. (1978), Bayesian Cluster Analysis, *Biometrika*, **65**, pp. 31–38.

[17] Bock H.H. (1996), Probabilistic Models in Cluster Analysis, *Computational Statistics and Data Analysis*, **23**, pp. 5–28.

[18] Bock H.H. (1998), Probabilistic Approaches in Cluster Analysis, *Bulletin of the International Statistical Institute*, **57**, pp. 603–606.

[19] Bolton R.J., Krzanowski W.J. (1999), A Characterization of Principal Components for Projection Pursuit, *The American Statistician*, **53** (2), pp. 108–109.

[20] Bowman A.W., Azzalini A. (1997), *Applied Smoothing Techniques for Data Analysis*, Oxford, U.K.: Clarendon Press.

[21] Bowman A.W., Foster P.J. (1993), Adaptive Smoothing and Density-Based Test of Multivariate Normality, *Journal of the American Statistical Association*, **88**, pp. 529–537.

[22] Boyles R.A. (1983), On the Convergence of the EM Algorithm, *Journal of the Royal Statistical Society*, Ser. B, **45**, pp. 47–50.

[23] Bradley P.S., Fayyad U., Reina C. (1998), Scaling EM (Expectation-Maximization) Clustering to Large Databases, *Technical Report MSR-TR-98-35*, Microsoft Research.

[24] Buhmann J.M. (1995), Learning and data clustering, in *Handbook of Brain Theory and Neural Networks* by Arbib M.

[25] Burman P., Nolan D. (1992), Location-Adaptive Density Estimation and Nearest-Neighbor Distance, *Journal of Multivariate Analysis*, **40**, pp. 132–157.

[26] Campbell J.G., Fraley C., Murtagh F., Raftery A.E. (1996), Linear Flow Detection in Woven Textiles Using Model-Based Clustering, *Pattern Recognition Letters*, **18**, pp. 1539–1548.

[27] Cao R., Cuevas A., Manteiga W.G. (1994), A Comparative Study of Several Smoothing Methods in Density Estimation, *Computational Statistics and Data Analysis*, **17**, pp. 153–176.

[28] Celeux G. (1998), Bayesian Inference for Mixtures: The Label-Switching Problem, In *COMPSTAT*, editors R. Payne and P. Green, Heidelberg and Vienna: Physica-Verlag, pp. 227–232.

[29] Celeux G., Chaveau D., Diebolt J. (1996), Stochastic Versions of the EM Algorithm: An Expiremental Study in the Mixture Case, *Journal of Statistical Computation and Simulation*, **55**, pp. 287–314.

[30] Celeux G., Govaert G. (1993), Comparison of the Mixture and the Classification Maximum Likelihood in Cluster Analysis, *Journal of Statistical Computation and Simulation*, **47**, pp. 127–146.

[31] Celeux G., Govaert G. (1995), Gaussian Parsimonious Clustering Methods, *Pattern Recognition*, **28**, pp. 781–793.

[32] Celeux G., Hurn M., Robert C. (2000), Computational and Inferential Difficulties With Mixture Posterior Distributions, *Journal of the American Statistical Assosiation*, **95**, pp. 957–970.

[33] Celeux G., Soromenho G. (1996), An Entropy Criterion for Assessing the Number of Clusters in a Mixture, *Journal of Classification*, **13**, pp. 195–212.

[34] Chang W.C. (1993), On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions, *Applied Statistics*, **32**, pp. 267–275.

[35] Chen J., Kalbfleisch J.D. (1996), Penalized Minimum-Distance Estimates in Finite Mixture Models, *Canadian Journal of Statistics*, **24**, pp. 167–175.

[36] Chen S.S., Gopinath R.A. (2000), Gaussianization, in book *NIPS*, pp. 423–429, http://citeseer.ist.psu.edu/456021.html

[37] Cheng R.C.H., Liu W.B. (2001), The Consistency of Estimators in Finite Mixture Models, *Scandinavian Journal of Statistics*, **28** (4), pp. 603–616.

[38] Cohen A.C. (1967), Estimation in Mixtures of Two Normal Distributions, *Technometrics*, **9**, pp. 15–28.

[39] Cook D., Buja A., Cabrera J. (1993), Projection Pursuit Indices Based on Expansions With Orthonormal Functions, *Journal of Computational and Graphical Statistics*, **2** (3), pp. 225–250.

[40] Cuevas A., Febrero M., Fraiman R. (2000), Estimating the Number of Clusters, *The Canadian Journal of Statistics*, **28** (2).

[41] Cutler A., Cordero-Brana O.I. (1996), Minimum Hellinger Distance Estimation for Finite Mixture Models, *Journal of the American Statistical Association*, **91**, pp. 1716–1723.

[42] Ćwik J., Koronacki J. (1996), Probability Density Estimation Using a Gaussian Clustering Algorithm, *Neural Computation Applications* **4**, pp. 149–160 .

[43] Ćwik J., Koronacki J. (1997), A Combined Adaptive-Mixtures Plug-in Estimator of Multivariate Probability Densities, *Computational Statistics and Data Analysis*.

[44] Day N.E. (1969), Estimating the Components of a Mixture of Normal Distributions, *Biometrika*, **56**, pp. 463–474.

[45] Dempster A.P., Laird N.M., Rubin D.B. (1977), Maximum Likelihood for Incomplete Data via the EM Algorithm (with discussion), *Journal of the Royal Statistical Society*, Ser. B, **39**, pp. 1–38.

[46] DeSarbo W.S., Howard D.J., Jedidi K. (1991), MULTICLUS: A New Method for Simultaneously Performing Multidimensional Scaling and Cluster Analysis, *Psychometrika*, **56**, pp. 121–136.

[47] Devlin S.J., Gnanadesikan R., Kattenring J.R. (1981), Robust Estimation of Dispersion Matrices and Principal Components, *Journal of the American Statistical Association*, **76**, pp. 354–362.

[48] Devroye L., Lugosi G. (1996), Acceptable Smoothing Factor for Kernel Density Estimates, *The Annals of Statistics*, **24** (6), pp. 2499–2512.

[49] Devroye L, Lugosi G. (1998), Variable Kernel Estimates: On the Impossibility of Tuning the Parameters, http://citeseer.ist.psu.edu/devroye98variable.html

[50] Dick N.P., Bowden D.C. (1973), Maximum Likelihood Estimation for Mixtures of Two Normal Distributions, *Biometrika*, **29**, pp. 781–790.

[51] Diday E., Lechevallier Y., Schader M., Bertrand P., Burtschy B. (1994), *New Approaches in Classification and Data Analysis*, New York: Springer-Verlag.

[52] Diebolt J., Robert C. (1994), Estimation of Finite Mixtures Through Bayesian Sampling, *Journal of the Royal Statistical Society*, Ser. B, **56**, pp. 363–375.

[53] Ding C., He X., Zha H., Simon H.D., Adaptive Dimension Reduction for Clustering High Dimensional Data, http://citeseer.ist.psu.edu/

[54] Duran B.S., Odell P.L. (1974), *Cluster Analysis*, New York: Springer-Verlag.

[55] Edwards A.W.F., Cavalli-Sforza L.L. (1965), A Method for Cluster Analysis, *Biometrics*, **21**, pp. 362–375.

[56] Escobar M.D., West M. (1995), Bayesian Density Estimation and Inference Using Mixtures, *Journal of the American Statistical Association*, **90**, pp. 1301–1312.

[57] Ester M., Kriegel H.P., Sander J., Xu X. (1996), A Density-Based Algorithms for Discovering Clusters in Large Spatial Databeses with Noise, *Proc. of the $2^{nd}$ International Conference on Knowledge Discovery and Data Mining (KDD-96)*.

[58] Everitt B.S. (1993), *Cluster Analysis*, London: Edward Arnold.

[59] Everitt B.S, Hand D.J. (1981), *Finite Mixture Distributions*, New York: Wiley.

[60] Fan J., Hall P., Martin M.A., Patil P. (1996), On Local Smoothing of Nonparametric Curve Estimators, *Journal of the American Statistical Association*, **91** (433), pp. 258–266.

[61] Fraley C. (1998), Algorithms for Model-Based Gaussian Hierarchical Clustering, *SIAM Journal on Scientific Computing*, **20**, pp. 270–281.

[62] Fraley C., Raftery A.E. (1998), How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis, *The Computer Journal*, **41**, pp. 578–588.

[63] Fraley C., Raftery A.E. (1999), MCLUST: Software for Model-Based Cluster Analysis, *Journal of Classification*, **16**, pp. 297–306.

[64] Fraley C., Raftery A.E. (2002), Model-Based Clustering, Discriminant Analysis and Density Estimation, *Journal of the American Statistical Association*, **97** (458), pp. 611–631.

[65] Friedman H.P., Rubin J. (1967), On Some Invariant Criteria for Grouping Data, *Journal of the American Statistical Association*, **62**, pp. 1159–1178.

[66] Friedman J.H. (1987), Exploratory Projection Pursuit, *Journal of the American Statistical Association*, **82** (397), pp. 249–266.

[67] Friedman J.H. (1989), Regularized Discriminant Analysis, *Journal of the American Statistical Association*, **84**, pp. 165–175.

[68] Friedman J.H., Stuetzle W., Schroeder A. (1984), Projection Pursuit Density Estimation, *Journal of the American Statistical Association*, **79** (387), pp. 599–608.

[69] Friedman J.H., Turkey J.W. (1974), A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions on Computers*, Ser. C, **23**, pp. 881–889.

[70] Fryer J.G., Robertson C.A. (1972), A Comparison of Some Methods of Estimating Mixed Normal Distributions, *Biometrika*, **59**, pp. 639–648.

[71] Green P.J. (1995), Reversible Jump MCMC Computation and Bayesian Model Determination, *Biometrika*, **82**, pp. 711–732.

[72] Grund B., Polzehl J. (1996), Bias Corrected Bootstrap Bandwidth Selection, *Technical Report 611*, School of Statistics, University of Minnesota, Minneapolis, U.S., http://citeseer.ist.psu.edu/grund96bias.html

[73] Hartigan J.A, Wong M.A. (1978), Algorithm AS 136: A $k$-Means Clustering Algorithm, *Applied Statistics*, **28**, pp. 100–108.

[74] Hall P. (1981), On the Non-Parametric Estimation of Mixture Proportions, *Journal of the Royal Statistical Society*, Ser. B, **43**, pp. 147–156.

[75] Hall P. (1989), On Polynomial-Based Projection Indices for Exploratory Projection Pursuit, *The Annals of Statistics*, **17** (2), pp. 589–605.

[76] Hall P. (1992), A Global Properties of Variable Bandwidth Density Estimates, *The Annals of Statistics*, **20**, pp. 762–776.

[77] Hall P., Sheather S.J., Jones M.C., Marron S.J. (1991), On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation, *Biometrika*, **78**, pp. 263–269.

[78] Hartigan J.A. (1975), *Clustering Algorithms*, New York: Wiley.

[79] Hasselbland V. (1966), Estimation of Parameters for a Mixture of Normal Distributions, *Technometrics*, **8**, pp. 431–444.

[80] Hastie T., Tibshirani R. (1996), Discriminant Analysis of by Gaussian Mixtures, *Journal of the Royal Statistical Society*, Ser. B, **58**, pp. 155–176.

[81] Hinneburg A., Aggarwal C.C., Keim D.A. (2000), What Is the Nearest Neighbor in High Dimensional Spaces?, *Proc. of the $26^{th}$ VLDB Conference*, Cairo, Egypt, pp. 506–515.

[82] Hjort N.L, Glad I.K. (1995), Nonparametric Density Estimation with a Parametric Start, *The Annals of Statistics*, **23** (3), pp. 882–904.

[83] Hollmen J. (1999), Penalized Likelihood Estimation in Gaussian Mixture Models, http://citeseer.ist.psu.edu

[84] Hosmer D.W. (1973), On MLE of the Parameters of a Mixture of Two Normal Distributions When Sample Size is Small, *Commun. Statist.*, **1**, pp. 217–227.

[85] Hosmer D.W. (1978), A Use of Mixtures of Two Normal Distributions in a Classification Problem, *Journal of Statistical Computation and Simulation*, **6**, pp. 281–294.

[86] Huber P.J. (1981), *Robust Statistics*, New York: Wiley.

[87] Huber P.J. (1985), Projection Pursuit (with discussion), *The Annals of Statistics*, **13** (2), pp. 435–475.

[88] Jain A.K., Dubes R.C. (1988), *Algorithms for Clustering Data*, New York: Prentice Hall, Englewood Cliffs.

[89] Jain A.K., Murty M.N., Flynn P.J. (1999), Data Clustering: A Review, *ACM Computing Surveys*, **31** (3), pp. 264–323.

[90] Jain A.K., Duin R.P.W., Mao J. (2000), Statistical Pattern Recognition: A Review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22** (1), pp. 4–37.

[91] Jakimauskas G. (1997), Efficiency Analysis of One Estimation and Clusterization Procedure of One-Dimensional Gaussian Mixture, *Informatica*, **8** (3), pp. 331–343.

[92] Jambu M., Lebeaux M.O. (1991), *Cluster Analysis and Data Analysis*, Elsevier Science.

[93] Jones M.C. (1983), *The Projection Pursuit Algorithm for Exploratory Data Analysis*, Ph.D. Thesis, University of Bath.

[94] Jones M.C., Linton O., Nielsen J.P. (1995), A Simple Bias Reduction Method for Density Estimation, *Biometrika*, **82** (2), pp. 327–338.

[95] Jones M.C., Marron J.S., Sheather S.J. (1996), A Brief Survey of Bandwidth Selection for Density Estimation, *Journal of the American Statistical Association*. **91** (433), pp. 401–407.

[96] Jones M.C., Sibson R. (1987), What is Projection Pursuit (with discussion)?, *Journal of the Royal Statistical Society*, Ser. A, **150**, pp. 1–36.

[97] Ka E.N.K., Fu A.W. (2002), Efficient Algorithm for Projected Clustering, *Proc. of the $18^{th}$ International Conference on Data Engineering*.

[98] Kaufman L., Rousseeuw P.J. (1990), *Finding Gruops in Data*, New York: Wiley.

[99] Krikštolaitis R. (2001), *Daugiamačių duomenų klasterizavimas panaudojant stebėjimų projektavimą*, Daktaro disertacija, VDU, Kaunas.

[100] Kruskal J.B. (1969), Toward Apractical Method Which Helps Uncover the Structure of a Set of Multivariate Observations by Finding the Linear Transformation Which Optimizes the New 'Index of Condensation', in *Statistical Computation*, ed. by R.C. Milton and J.A. Nelder, New York: Academic.

[101] Lavine M., West M. (1992), A Bayesian Method for Classification and Discrimination, *Canadian Journal of Statistics*, **20**, pp. 451–461.

[102] Leroux M. (1992), Consistent Estimation of a Mixing Distributions, *The Annals of Statistics*, **20**, pp. 1350–1360.

[103] Li X.Q., King I. (1999), Gaussian Mixture Distance for Information Retrieval, *Proc. of the International Conference on Neural Networks* pp. 2544–2549.

[104] Mack Y.P., Rosenblatt M. (1979), Multivariate $k$-Nearest Neighbor Density Estimates, *Journal of Multivariate Analysis*, **9**, pp. 1–15.

[105] Marron J.S. (1985), A Asymptotically Efficient Solution to the Bandwidth Problem of Kernel Density Estimation, *The Annals of Statistics*, **13**, pp. 1011–1023.

[106] Marron J.S., Wand M.P. (1992), Exact Mean Integrated Error, *The Annals of Statistics*, **20**, pp. 712–736.

[107] McLachlan G.J. (1987), On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture, *Applied Statistics*, **36**, pp. 318–324.

[108] McLachlan G.J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

[109] McLachlan G.J., Basford K.E. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

[110] McLachlan G.J., Krishnan T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

[111] McLachlan G.J., Peel D., Basford K.E., Adams P. (1999), The EMMIX Software for the Fitting of Mixtures of Normal and t-Components, *Journal of Statistical Software*. **4** (2), http://www.jstatsoft.org/

[112] Moore A. (1999), Very Fast EM-based Mixture Model Clustering Using Multiresolution $kd$-Trees, in *Advances in Neural Information Processing Systems* editors M. Kearns, S. Solla, D. Cohn, Vol. 11, pp. 543–549.

[113] Mukherjee S., Feigelson E.D., Badu G.J., Murtagh F., Fraley C., Raftery A.E. (1998), Three Types of Gamma Ray Bursts, *The Astrophysical Journal*, **508**, pp. 314–327.

[114] Müller P., Erkanli A., West M. (1996), Bayesian Curve Fitting Using Multivariate Normal Mixtures, *Biometrika*, **83**, pp. 67–80.

[115] Nason G.P. (1992), *Design and Choice of Projection Index*, Ph.D. Thesis, University of Bath.

[116] Nason G.P. (1995), Three-Dimensional Projection Pursuit, *Journal of the Royal Statistical Society*, Ser. C, **44**, pp. 411–430.

[117] Nason G.P. (1996), Robust Projection Indices, *Technical report*, University of Bristol, UK.

[118] Ng R.T., Han J. (1994), Efficient and Effective Clustering Methods for Spatial Data Mining, *Proc. of the $20^{th}$ International Conference on Very Large Data Bases*, Santiago, Chile, pp. 144–155.

[119] Oh M.S., Raftery A.E. (2001), Bayesian Multidimensional Scaling and Choice of Dimension, *Journal of the American Statistical Association*, **96**, pp. 1031–1044.

[120] Olkin I., Spiegelman C.H. (1987), A Semiparametric Approach to Density Estimation, *Journal of the American Statistical Association*, **82** (399), pp. 858–865.

[121] Ordonez C., Omiecinski E., FREM: Fast and Robust RM Clustering for Large Data Sets, http://citeseer.ist.psu.edu/536108.html

[122] Ormoneit D., Tresp V. (1998), Improved Gaussian Mixture Density Estimates Using Bayesian Penalty Terms and Network Averaging, in *Advances in Neural Information Processing Systems* editors David S. Touretzky D.S., Mozer M.C., Hasselmo M.E., Vol. 8, pp. 542–548.

[123] Ormoneit D., Tresp V. (1998), Averaging, Maximum Penalized Likelihood and Bayesian Estimation for Improving Gaussian Mixture Probability Density Estimates, *IEEE Transactions on Neural Networks*, **9**, pp. 639–649

[124] Paclik P., Novovicova J., Number of Components and Initialization in Gaussian Mixture Model for Pattern Recognition, http://citeseer.ist.psu.edu

[125] Park B.U., Maroon J.S. (1990), Comparison of Data-Driven Bandwidth Selectors, *Journal of the American Statistician Association*, **85** (409), pp. 66–72.

[126] Park B.U., Turlach B.A. (1992), Practical Performance of Several Data Driven Bandwidth Selectors, *Computational Statistics*, **7** (3), pp. 251–270.

[127] Parzen E. (1962), On Estimation of a Probability Density Function and Mode, *The Annals of Mathematical Statistics*, **33**, pp. 1215–1230.

[128] Posse C. (1990), An Effective Two-Dimensional Projection Pursuit Algorithm, *Comm. Statist. Simul. Comput.*, **19** (4), pp. 1143–1164.

[129] Posse C. (2001), Hierarchical Model-Based Clustering for Large Datasets, *Journal of Computational and Graphical Statistics*, **10**, pp. 464–486.

[130] Prakasa Rao B.L.S. (1983), *Nonparametric Functional Estimation*, New York: Academic Press.

[131] Procopiuc C.M., Jones M., Agarwal P.K., Marali T.M. (2002), A Monte Carlo Algorithm for Fast Projective Clustering, *Proc. of ACM SIGMOD International Conference on Management of Data*, Madison, USA.

[132] Radavičius M., Rudzkis R. (1998), Consisten Estimation of Discriminant Space in Mixture Model by Using Projection Pursuit, *Proceedings of the $7^{th}$ Vilnius Conference in Probability Theory and Mathematical Statistics*, pp. 617–626.

[133] Radavičius M., Rudzkis R. (1999), Locally Minimax Efficiency of Nonparametric Density Estimators for $\chi^2$ Type Losses, *Lietuvos matematikos rinkinys*, **39** (3), pp. 398-424.

[134] Raftery A.E. (1995), Bayesian Model Selection in Social Research (with disscusion), *Sociological Methodology*, **25**, pp. 111–193.

[135] Rasmussen C.E. (2000), The Infinite Gaussian Mixture Model, in *Advances in Neural Information Processing Systems* editors Solla S.A., Leen T.K., Muller K.R., Vol. 12, pp. 554–560.

[136] Render R.A., Walker H.F. (1984), Mixture Densities, Maximum Likelihood and the EM Algorith, *SIAM Rev.*, **26**, pp. 195–239.

[137] Richardson S., Green P.J. (1997), On Bayesian Analysis of Mixtures With an Unknown Number of Components (with discussion), *Journal of the Royal Statistical Society*, Ser. B, **59**, pp. 731–792.

[138] Ripley B.D. (1992), *Pattern Recognition and Neural Networks*, U.K.: Cambridge University Press.

[139] Roeder K., Wasserman L. (1997), Practical Bayesian Density Estimation Using Mixtures of Normals, *Journal of the American Statistical Association*, **92**, pp. 894–902.

[140] Romesberg H.C. (2000), *Cluster Analysis for Researchers*, New York: Wiley, John & Sons.

[141] Roosen C.B., Hastie T.J. (1994), Automatic Smoothing Spline Projection Pursuit, *Journal of Computational and Graphical Statistics*, **3**, pp. 235–248.

[142] Rudzkis R., Radavičius M. (1995), Statistical Estimation of a Mixture of Gaussian Distributions, *Acta Applicandae Mathematicae*, **38**, pp. 37–54.

[143] Rudzkis R., Radavičius M. (1997), Celenapravlenoe projecirovanie v medeliach smesi gausovskich raspredilenij, sochraniajushcheje infomaciju o klasternoj strukture, *Lietuvos matematikos rinkinys*, **37** (4), pp. 550–563, (rusų k.).

[144] Rudzkis R., Radavičius M. (1999), Characterization and Statistical Estimation of a Discriminant Space for Gaussian Mixtures, *Acta Applicandae Mathematicae*, **59**, pp. 279–290.

[145] Rudzkis R., Radavičius M. (2003), Testing Hypotheses on Discriminant Space in the Mixture Model of Gaussian Distributions, *Acta Applicandae Mathematicae*, **79**, pp. 105–114.

[146] Sain S.R., Baggerly K.A., Scott D.W. (1992), Cross-Validation of Multivariate Densities, *Journal of the American Statistical Association*, **89** (427), pp. 807–817.

[147] Sain S.R. (1994), *Adaptive Kernel Density Estimation*, Ph.D. Thesis, Rice University, Houston, U.S., http://citeseer.ist.psu.edu/sain94adaptive.html

[148] Sain S.R., Scott D.W. (1996), On Locally Adaptive Density Estimation, *Journal of the American Statistical Association*, **91** (436), pp. 1525–1534.

[149] Sain S.R. (2000), Bias Reduction and Elimination with Kernel Estimators, http://citeseer.ist.psu.edu/

[150] Samuiddin M., El-Sayyad G.M. (1990), On Nonparametric Kernel Density Estimates, *Biometrika*. **77**, pp. 865–874.

[151] Schwarz G. (1978), Estimating the Dimension of a Model, *The Annals of Statistics*, **6**, pp. 461–464.

[152] Scott A.J., Symons M.J. (1971), Clustering Methods Based on Likelihood Ratio Criteria, *Biometrics*. **27**, pp. 387–397.

[153] Scott D.W. (1985), Average Shifted Histograms: Effective Nonparametric Density Estimation in Several Dimensions, *The Annals of Statistics*, **13**, pp. 1024–1040.

[154] Scott D.W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, New York: Wiley.

[155] Scott D.W., Terrell G.R. (1987), Biased and Unbiased Cross-Validation in Density Estimation, *Journal of the American Statistical Association*, **82**, pp. 1131–1146.

[156] Sheather S.J., Jones M.C. (1991), A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation, *Journal of the Royal Statistical Society*, Ser. B, **53**, pp. 683–690.

[157] Silverman B.W. (1986), *Density Estimation for Statistics Data Analysis*, London: Chapman and Hall.

[158] Simonoff J.S. (1996), *Smoothing Methods in Statistics*, New York: Springer-Vergal.

[159] Singer Y., Warmuth M.K. (1998), A New Parameter Estimation Method for Gaussian Mixtures, http://citeseer.ist.psu.edu/singer98new.html

[160] Smyth P. (2000), Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood, *Statistics and Computing*, **10**, pp. 63–72.

[161] Stephens M. (1997), Discussion on 'On Bayesian Analysis of Mixtures With an Unknown Number of Components' (by S. Richardson S and P.J. Green), *Journal of the Royal Statistical Society*, Ser. B, **59**, pp. 768–769.

[162] Stephens M. (1997), *Bayesian Methods for Mixtures of Normal Distributions*, Ph.D. Thesis, University of Oxford.

[163] Stephens M. (2000), Bayesian Analysis of Mixture Models With an Unknown Number of Components — an Alternative to Reversible Jump Methods, *The Annals of Statistics*. **28**, pp. 40–74.

[164] Stephens M. (2000), Dealing With Label Switching in Mixture Models, *Journal of the Royal Statistical Society*, Ser. B, **62** (4), pp. 795–809.

[165] Sun J. (1991), Significance Level in Exploratory Projection Pursuit, *Biometrika*, **78** (4), pp. 759–769.

[166] Terrel G.R., Scott D.W. (1992), Variable Kernel Density Estimation, *The Annals of Statistics*, **20**, pp. 1236–1265.

[167] Tipping M.E. (1999), Deriving Cluster Analytic Distance Functions from Gaussian Mixture Models, *Proc. of $9^{th}$ Int. Conf. on Artificial Neural Networks*, Edinburgh, pp. 815–820.

[168] Titterington D.M., Smith A.F.M., Markov U.E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

[169] Tsuda K., Minoh M., A Nonparametric Density Model for Classification in a High Dimensional Space, http://citeseer.ist.psu.edu/78925.html

[170] Turlach B.A., Bandwidth Selection in Kernel Density Estimation: A Review, http://citeseer.ist.psu.edu/214125.html

[171] Ueda N., Nakano R., Ghahramani Z., Hinton G.E. (2000), SMEM Algorithm for Mixture Models, *Neural Computation*, **12** (9), pp. 2109–2128.

[172] Vlassis N., Likas A. (2000), A Greedy EM algorithm for Gaussian Mixture Learning, *Neural Processing Letters*, **15** (1), pp. 77–87.

[173] Wand M.P., Jones M.C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

[174] Wand M.P., Marron J.S., Ruppert D. (1991), Transformations in Density Estimation (with discussion), *Journal of the American Statistical Association*, **86**, pp. 343–361.

[175] Ward J.H. (1963), Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, **58**, pp. 234–244.

[176] Windham M.P., Cutler A. (1992), Information Ratios for Validating Mixture Analyses, *Journal of the American Statistical Association*, **87**, pp. 1188–1192.

[177] Witherspoon N.H., Holloway J.H., Davis K.S., Miller R.W., Dubey A.C. (1995), The Coastal Battlefield Reconnaissance and Analysis (COBRA) Program for Minefield Detection, *Proceedings of SPIE*, Orlando, pp. 500–508.

[178] Wolfe J.H. (1967), NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixture Distributions, *Technical Bulletin USNPRA SRM 68-2*, U.S. Naval Personnel Research activity, San Diego.

[179] Wolfe J.H. (1970), Pattern Clustering by Multivariate Mixture Analysis, *Multivariate Behavioral Research*, **5**, pp. 329–350.

[180] Wu C.F.J. (1983), On Convergence Properties of the EM Algorithm, *The Annals of Statistics*. **11**, pp. 95–103.

[181] Xu L., Jordan M.I. (1996), On Convergence Properties of the EM Algorithm for Gaussian Mixtures, *Neural Computation*, **8** (1), pp. 129–151.

[182] Yeung K.Y., Fraley C., Murua A., Raftery A.E., Ruzzo W.L. (2001), Model-Based Clustering and Data Transformations for Gene Expression Data, *Bioinformatics*, **17**, pp. 763–774.

# Publikacijų sąrašas

## Straipsniai recenzuojamuose leidiniuose, įtrauktuose į Mokslinės informacijos instituto duomenų bazę

1. Kavaliauskas M., Rudzkis R. (2005), Multivariate Data Clustering for the Gaussian Mixture Model, *Informatica*, **16** (1), Vilnius, MII. [ISSN 0868-4952]

## Straipsniai leidiniuose, įtrauktuose į Mokslo ir studijų departamento prie Švietimo ir mokslo ministerijos patvirtintą sąrašą

1. Kavaliauskas M. (1997), Adaptyvus glodinimo pločio parinkimas pasiskirstymo tankio statistiniame vertinime, *Lietuvos matematikų draugijos XXXVII konferencijos darbai. Specialus Lietuvos matematikos rinkinio priedas*, Vilnius, Technika, pp. 198–203. [ISBN 9986-05-350-1]

2. Rudzkis R., Kavaliauskas M. (1998), On Local Bandwidth Selection for Density Estimation, *Informatica*, **9** (4), Vilnius, MII, pp. 171–178. [ISSN 0868-4952]

3. Kavaliauskas M. (2000), Branduolinio pasiskirstymo tankio įvertinimas taikant kryžminį patikrinimą, *Lietuvos matematikos rinkinys, spec. nr.*, **40**, Vilnius, MII, pp. 290–295. [ISSN 0132-2818]

4. Kavaliauskas M., Rudzkis R. (2002), Projection-based Estimation of Multivariate Distribution Density, *Lietuvos matematikos rinkinys, spec. nr.*, **42**, Vilnius, MII, pp. 537–540. [ISSN 0132-2818]

## Straipsniai kituose recenzuojamuose tarptautiniuose ir užsienio leidiniuose

1. Kavaliauskas M. (1999), Adaptive Density Estimation Method, *Proc. of $11^{th}$ European Young Statisticians Meeting*, Marly-le-Roi, INRA, pp. 112–116.

2. Kavaliauskas M. (2001), An Adaptive Kernel Density Estimation Method, *Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods. Proc. of $6^{th}$ Intl. Conf.*, **1**, Minsk, BSU, pp. 171–178. [ISBN 985-445-490-8 (Vol.1), ISBN 985-445-489-4]

3. Kavaliauskas M., Rudzkis R. (2004), The Projection-based Estimation of the Gaussian Mixture Parameters, *Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods. Proc. of 7$^{th}$ Intl. Conf.*, **1**, Minsk, BSU, pp. 160–164. [ISSN 985-445-492-4]

4. Kavaliauskas M., Rudzkis R., Ruzgas T. (2004), The Projection-based Multivariate Distribution Density Estimation, *Acta et Commentationes Universitatis Tartuensis de Mathematica*, **8**, Tartu, Tartu University Press, pp. 1–7. [ISSN 1406-2283]

## Kiti straipsniai

1. Kavaliauskas M., Ruzgas T. (2004), Gauso skirstinių mišinių klasterizavimas taikant neparametrinius metodus, *Matematika ir matematinis modeliavimas. Konferencijos pranešimų medžiaga*, Technologija, Kaunas, pp. 14–19. [ISBN 9955-09-616-0]

# Konferencijų pranešimų sąrašas

1. Kavaliauskas M., Adaptyvus glodinimo pločio parinkimas pasiskirstymo tankio statistiniame vertinime, *Lietuvos matematikų draugijos XXXVIII konferencija*, VGTU, Vilnius, 1997.

2. Kavaliauskas M.,Rudzkis R., Adaptyv ̄us pasiskirstymo tankio įver ̆ciai, *Lietuvos matematikų draugijos XXXIX konferencija*, KMA, Kaunas, 1998.

3. Kavaliauskas M., Glodinimo pločio parinkimo b ̄udai pasiskirstymo tankio įvertinime, *Lietuvos matematikų draugijos XL konferencija*, MII, Vilnius, 1999.

4. Kavaliauskas M., Adaptive Density Estimation Method, *$11^{th}$ European Young Statisticians Meeting*, Marly-le-Roi, INRA, 1999.

5. Kavaliauskas M., Branduolinio pasiskirstymo tankio įvertinimas taikant kryžminį patikrinimą, *Lietuvos matematikų draugijos XLI konferencija*, ŠU, Šiauliai, 2000.

6. Kavaliauskas M., Daugiamačio tankio vertinimas taikant projektavimą, *Lietuvos matematikų draugijos XLII konferencija*, KU, Klaipėda, 2001.

7. Kavaliauskas M., An Adaptive Kernel Density Estimation Method, *Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods*, BSU, Minsk, BSU, 2001.

8. Kavaliauskas M., Rudzkis R., Projection-based Estimation of Multivariate Distribution Density, *Lietuvos matematikų draugijos XLIII konferencija*, LKA, Vilnius, 2002.

9. Rudzkis R, Kavaliauskas M., Projection-based Multivariate Distribution Density Estimation, *$8^{th}$ Vilnius Conference on Probability Theory*, VGTU, Vilnius, 2002.

10. Kavaliauskas M., Rudzkis R., Ruzgas T., Daugiamačių Gauso skirstinių mišinio parametrų įvertinimas taikant vienamates projekcijas, *Lietuvos matematikų draugijos XLIV konferencija*, VPU, Vilnius, 2003.

11. Kavaliauskas M., Rudzkis R., Ruzgas T., The Projection-based Multivariate Distribution Density Estimation, *The $7^{th}$ Tartu Conference on Multivariate Statistics. Satellite meeting of ISI $54^{th}$ session in Berlin*, University of Tartu, Tartu, 2003.

12. Kavaliauskas M., Ruzgas T., Gauso skirstinių mišinių klasterizavimas taikant neparametrinius metodus, *Matematika ir matematinis modeliavimas*, Kauno technologijos universitetas, 2004.

13. Kavaliauskas M., Rudzkis R., Vienamačių projekcijų panaudojimas daugiamačio Gauso mišinio modelio atveju, *Lietuvos matematikų draugijos XLV konferencija*, LŽŪU, Kaunas, 2004.

14. Rudzkis R., Kavaliauskas M., Ispolzovanie odnomernych proekcij v algoritmach klasifikacii, osnovannych na modeli smesi mnogomernych raspredilenij, *$6^{th}$ Intl. School-Workshop on Multidimensional Analysis and Econometrics*, Tsakhkadzor, Armėnija, 2004, (plenarinis pranešimas, rusų kalba).

15. Kavaliauskas M., Rudzkis R., The Projection-based Estimation of the Gaussian Mixture Parameters, *Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods*, BSU, Minsk, 2004.