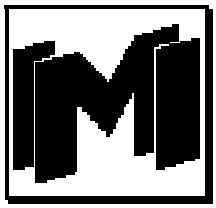


MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Tomas Ruzgas

Daugiamačio pasiskirstymo tankio neparametrinis įvertinimas naudojant stebėjimų klasterizavimą

Daktaro disertacija

Fiziniai mokslai, matematika (01 P)

Vilnius, 2007

Disertacija rengta 2002–2006 metais Matematikos ir informatikos institute.

Darbo mokslinis vadovas

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

Turinys

Įvadas

Tiriamoji problema ir darbo aktualumas. Nagrinėjamoji problema yra glaudžiai susijusi su daugiamačių stebinių pasiskirstymo analize – viena iš esminių duomenų analizės šakų, kuria remiasi daugelio kitų uždavinių sprendimas (diskriminantinė analizė, vaizdų atpažinimas ir pan.). Pasiskirstymo tankių vertinimo metodologija sulaukia vis daugiau dėmesio dėl naujai atsiradusių taikymo sričių: genetinės informacijos apdorojimo, astronomijos tyrimų objektų analizės, kompiuterinės technikos bei jos periferijos duomenų tyrimo ir t. t. Nors yra daugybė duomenų pasiskirstymo tankių vertinimo metodų, įvairūs autoriai siūlo vis naujas idėjas [10, 19, 61, 108, 132, 168, 170, 171, 206].

Gana plačiai paplitę modeliai, kai skirstinio šeima yra žinoma. Ypač dažnai pasitaiko Gauso skirstinio modelis. Sprendžiant realius pasiskirstymo uždavinius, paprastai būna nežinoma skirstinio šeima (pvz., prielaida dėl modelio gausiškumo neretai būna klaidinga). Pagal tai, ar žinoma stebimo atsitiktinio dydžio skirstinio šeima, tankių vertinimas skirstomas į parametrinį ir neparametrinį. Parametrinio vertinimo atveju daroma prielaida, kad tankio funkcija $f(x)$, apibūdinanti d-matį atsitiktinį vektorių X, priklauso tam tikrai gana siaurai funkcijų šeimai $f(\cdot;\theta)$, sąlygojamai nedidelio skaičiaus parametrų $\theta = (\theta_1, \theta_2, ..., \theta_k)$. Pavyzdžiu galėtų būti daugiamačių normalinių skirstinių šeima, kurią nusako vidurkio vektorius ir kovariacinė matrica. Tankis, kuris apskaičiuojamas pagal parametrinį įvertinį, gaunamas iš pradžių apskaičiavus parametro θ įvertį $\hat{\theta}$ ir $\hat{f} = f(\cdot;\hat{\theta})$. Toks traktavimas statistiniu požiūriu yra labai efektyvus, tačiau jeigu nė vienas šeimos $f(\cdot;\theta)$ narys nėra artimas funkcijai f, gauti rezultatai gali būti labai netikslūs.

Praktikoje parametrinis modelis dažnai taikomas ir tuo atveju, kai tikrasis tankis tik apytiksliai aproksimuojamas parametriniu modeliu. Ypač tai būdinga tyrimams, kai imtys nėra didelės. Pavyzdžiui, parametrų įverčiams rasti galima taikyti *maksimalaus tikėtinumo metodą* (angl. *maximum likelihood method*; toliau – *MTM*). Deja, tokių įverčių apskaičiavimas yra rimta praktinė problema daugiamačiu atveju; ypač tai liečia skirstinių mišinių modelius, kurie taikomi daugiamodaliniam tankiui aproksimuoti. Didėjant duomenų matavimo ir mišinio komponentų skaičiui, parametrų dimensija greitai auga, o tikėtinumo funkcija turi daug lokaliųjų ekstremumų. Klasikinių optimizavimo metodų taikymas neduoda gerų rezultatų, todėl, analizuojant mišinius, parametrams vertinti dažniausiai naudojamas rekurentinis *tikėtinumo maksimizavimo* (angl. *expectation maximization*; toliau – *EM*) algoritmas. EM algoritmo savybės yra gerai ištirtos ir aprašytos, pavyzdžiui, [11, 24, 231]. EM algoritmas maksimizuoja tikėtinumo funkciją, todėl gali būti naudojamas MTM įvertiniui apskaičiuoti.

Neparametriniam tankio vertinimui jokios parametrinės prielaidos dėl $f(x)$ nėra reikalingos, tačiau vietoj to daromos kitos prielaidos, tarkime, dėl funkcijos $f(x)$ tolydumo (pvz.,

kad funkcija turi antrosios eilės tolydžiąją išvestinę). Turint iš nepriklausomų X kopijų sudarytą didelę imtį $\mathbf{X} = (X(1), \ldots, X(n))$, tankis $f(x)$ gali būti apskaičiuotas pakankamai tiksliai.

Tačiau ir naudojant neparametrinius metodus daugiamačio tankio vertinimas kelia problemų, ypač tuo atveju, kai nagrinėjamasis tankis yra daugiamodalinis. Ši situacija būdinga tyrimams, susijusiems su klasifikavimu, vaizdų analize ir pan.

Šiuolaikinėje duomenų analizėje naudojama daugybė neparametrinių daugiamačių atsitiktinių dydžių pasiskirstymo tankio statistinio vertinimo metodų. Ypač plačiai paplitę branduoliniai įvertiniai [113, 190]. Gana populiarūs ir splaininiai [139, 211] bei pusiau parametriniai [43, 78, 80, 81, 87, 103–107, 108, 134, 161, 193] algoritmai. Tačiau stokojama išsamių esamų populiarių įvertinių efektyvumo palyginimų daugiamodalinio tankio atveju. Taikant daugumą populiarių neparametrinio vertinimo procedūrų, praktikoje susiduriama su jų parametrų optimalaus parinkimo problema. Branduolinių įvertinių konstrukcijos svarbiausias elementas yra glodinimo plotis, nelengva parinkti splaininių įvertinių mazgus ir t. t. Nors yra sukurta nemažai adaptyvių minėtų parametrų parinkimo procedūrų [17, 30–36, 85, 114, 119–129, 190, 206, 217, 227], tačiau jos nėra pakankamai efektyvios, kai imties tūris nėra didelis, ypač jei stebinių dimensija yra didoka. Pastaruoju atveju tikslinga taikyti duomenų projektavimą [4, 72, 73, 132, 147, 167, 173], nes parametrų parinkimo uždavinys tuo sunkesnis, kuo didesnė stebimų atsitiktinių vektorių dimensija. Disertaciniame darbe tiriama galimybė duomenų projektavimą panaudoti daugiamačiam pasiskirstymo tankiui nustatyti. Pasiūlytos ir ištirtos procedūros, leidžiančios gana tiksliai nustatyti daugiamačių skirstinių mišinio pasiskirstymo tankį pagal jo vienamačių projekcijų įverčius, gaunamus aproksimuojant Gauso skirstinių mišinio modeliu, o daugiamatį tankį apskaičiuoti pagal gautų projekcijų skirstinių įverčius taikant apvertimo formulę.

Nagrinėjami algoritmai, pagrįsti stebinių projektavimu į didelį skaičių laisvai pasirinktų vienamačių krypčių. Turint pakankamai didelio skaičiaus projekcijų pasiskirstymo tankių įverčius, daugiamačio pasiskirstymo tankio įvertį galima gauti iš apvertimo formulės [132]. Taikant šią metodologiją skaičiavimai kompiuteriu užima daug laiko, todėl ja susidomėta tik labai išaugus kompiuterių spartai. Kitokia yra J. H. Friedman idėja [72], leidžianti išvengti daugelio su minėtos apvertimo formulės taikymu susijusių sunkumų (keblu parinkti glodinimo parametrus, reikalingi didelio skaičiaus projekcijų pasiskirstymo tankių įverčiai ir t. t.).

Vienas iš būdų mėginti padidinti neparametrinių įverčių tikslumą – daugiamodalinio tankio analizę traktuoti kaip vienamodalinių tankių vertinimą, o tiriamąjį tankį – kaip vienamodalinių tankių mišinį. Pagrindinė darbo dalis skirta šiai idėjai įgyvendinti. Siūloma pirmame tyrimų etape imtį klasterizuoti, o po to kiekvieną klasterį atitinkančius skirstinių mišinio komponentus įvertinti atskirai. Pradinio klasterizavimo idėja nėra nauja, tačiau ji taikyta

tik kartu su populiariu branduoliniu tankių vertinimo metodu, klasterizavimui naudojant geometrinius ir hierarchinius algoritmus [52, 117].

Šiame darbe Monte Karlo [95, 96, 110, 145] metodu buvo siekiama atlikti įvairių neparametrinių įvertinių tikslumo lyginamąją analizę tuo atveju, kai stebinių skirstinio tankis yra daugiamodalinis, o imties tūris nėra didelis, ir nustatyti, ar tikslinga, vertinant tokio tipo tankius, imtį preliminariai suskaidyti į klasterius ir tuos klasterius patikslinti. Tačiau neužtenka nustatyti, kad pirminio imties klasterizavimo procedūras taikyti yra tikslinga – reikia parinkti efektyvius imties klasterizavimo algoritmus. Nemažai vietos skiriama įvairių klasterizavimo metodų palyginimui.

Tikslas ir uždaviniai. *Darbo tikslas* – sukurti ir ištirti daugiamačio pasiskirstymo tankio neparametrinio vertinimo algoritmus, kurie būtų efektyvūs daugiamodališkumo atveju.

Darbo uždaviniai:

- atlikti populiarių neparametrinių tankių statistinių įverčių tikslumo lyginamąją analizę daugiamodaliniu atveju;
- ištirti pirminio duomenų klasterizavimo poveikį daugiamodalinio tankio statistinio vertinimo tikslumui;
- palyginti įvairių klasterizavimo procedūrų taikymo pasiskirstymo tankiams vertinti efektyvumą.

Mokslinis naujumas

1. Atlikta skirtingų tipų statistikų, skirtų daugiamačių tankių vertinimui, lyginamoji analizė daugiamodaliniu atveju. Tyrimui atrenkant konkrečias pasiskirstymo tankių vertinimo procedūras, buvo siekiama, kad jos atstovautų populiarių įverčių klasėms ir jau būtų eksperimentiškai tirtos kitų tyrėjų. Pateikiamame darbe ypač daug dėmesio skiriama projektavimo metodams, kurių efektyvumas daugiamodaliniu atveju dar menkai tirtas.
2. Ištirtas neparametrinės daugiamodalinio tankio aproksimacijos Gauso mišinių modeliais tikslingumas esant nuosaikiam imties dydžiui.
3. Pasiūlyta originali daugiamodalinio tankio analizės metodika, paremta tiriamo tankio traktavimu vienamodalinių tankių mišiniu ir imties projektavimu. Ištirtas pirminio imties klasterizavimo efektyvumas vertinant tankį.
4. Palygintos įvairios pirminio imties klasterizavimo procedūros ir parodyta, kad, vertinant neparametrinį daugiamodalinį tankį, negriežtas klasterizavimas yra pranašesnis už griežtą.

Ginamieji teiginiai

1. Atliekant populiarių neparametrinių tankio įverčių tikslumo lyginamąją analizę daugiamodališkumo atveju, parodyta, kad vertinimo rezultatai labai pagerėja, jei stebiniai pirmiausia klasterizuojami traktuojant jų daugiamodalinį tankį kaip vienamodalinių tankių mišinį, o tankių vertinimo metodai yra taikomi kiekvienam klasteriui atskirai.
2. Parodyta, kad imties skaidymas į klasterius, taikant Gauso skirstinių mišinio modelį ir EM algoritmą, yra akivaizdžiai pranašesnis už populiarius geometrinius klasterizavimo metodus, o negriežtas klasterizavimas yra pranašesnis nei griežtas, kai klasterizavimo rezultatai taikomi tankių mišiniams statistiškai vertinti.
3. Pasiūlyta klasterių skaičiaus nustatymo taisyklė ir ištirtas jos efektyvumas.
4. Atlikta lyginamoji kelių populiarių neparametrinių įvertinių tikslumo analizė parodė, kad daugiamodalinio neparametrinio tankio vertinimo algoritmas gautas sujungiant pirminę imties klasterizaciją (taikant Gauso skirstinių mišinio modelį ir EM algoritmą) su J. H. Friedman procedūra, naudojančia duomenų projektavimą, yra efektyvesnis nei tirti kiti populiarūs vertinimo būdai.

Raktiniai žodžiai: daugiamatis pasiskirstymo tankis, neparametrinis vertinimas, imties klasterizavimas, tikslinis projektavimas, Monte Karlo metodas.

Rezultatų aprobavimas. Disertacinio darbo tematika yra išspausdintas 1 straipsnis leidinyje, įtrauktame į Mokslinės informacijos instituto duomenų bazę, 2 straipsniai – Lietuvos mokslo tarybos patvirtinto sąrašo tarptautinėse duomenų bazėse referuojamuose leidiniuose, 3 straipsniai – kituose recenzuojamuose mokslo leidiniuose. Straipsnių sąrašas pateikiamas disertacinio darbo pabaigoje.

Disertacinio darbo tematikai yra skirta 10 pranešimų Lietuvos ir tarptautinėse mokslinėse konferencijose. Konferencijų pranešimų sąrašas pateiktas disertacijos pabaigoje. Taip pat skaityti pranešimai Matematikos ir informatikos instituto Taikomosios statistikos skyriaus seminaruose, Matematikos ir informatikos instituto Matematinės statistikos, Tikimybių teorijos, Taikomosios statistikos skyrių bei Vilniaus universiteto Matematikos-informatikos fakulteto Matematinės analizės, Ekonometrinės analizės ir Matematinės informatikos katedrų jungtiniame seminare bei Vilniaus Gedimino technikos universiteto Matematinės statistikos katedros seminare.

Disertacinio darbo struktūra. Disertaciją sudaro įvadas, trys skyriai, išvados, literatūros sąrašas, publikacijų sąrašas ir priedai.

Pirmas skyrius skirtas daugiamačio pasiskirstymo tankio neparametriniam vertinimui ir šių įvertinių lyginamajai analizei. Antrame skyriuje apžvelgiami duomenų klasterizavimo

algoritmai, tiriamas pirminio imties klasterizavimo poveikis pasiskirstymo tankio neparametrinio vertinimo tikslumui. Trečiame skyriuje lyginamos populiarios geometrinio klasterizavimo procedūros bei tiriamas griežto ir negriežto pirminio imties klasterizavimo poveikis daugiamodalinio tankio vertinimo tikslumui ir šio tikslumo priklausomybė nuo pasirinkto klasterių skaičiaus. Darbo pabaigoje pateikiami straipsnių ir konferencijų pranešimų disertacijos tematika sąrašai.

1. Pasiskirstymo tankių neparametrinių įvertinių palyginimas

Sakysime, kad atsitiktinis vektorius $X \in \mathbf{R}^d$ tenkina skirstinių mišinio modelį, jeigu jo pasiskirstymo tankis $f(x)$ tenkina lygybę

$$f(x) = \sum_{k=1}^{q} p_k f_k(x) = f(x, \theta) . \tag{1.1}$$

Parametras q vadinamas mišinio klasterių (komponentų, klasių) skaičiumi, o p_k – *apriorinėmis* tikimybėmis. Jos tenkina sąlygas:

$$p_k > 0, \ \sum_{k=1}^{q} p_k = 1 . \tag{1.2}$$

(1.1) formulėje funkcija $f_k(x)$ yra pasiskirstymo tankio funkcija, o θ – daugiamatis modelio parametras.

Tarkime, kad X yra d-matis atsitiktinis vektorius, kurio pasiskirstymo tankis $f(x)$, ir turime iš nepriklausomų X kopijų sudarytą imtį $\mathbf{X} = (X(1), \ldots, X(n))$. Sakysime, kad imtis tenkina mišinio modelį, jeigu $X(t)$ tenkina (1.1). Dydį n vadinsime imties dydžiu (tūriu).

Vienas iš statistinių uždavinių yra stebimo atsitiktinio dydžio tankio vertinimas. Jei žinomas turimos imties pasiskirstymo tipas (normalinis, Puasono ir pan.), tuomet duomenų pasiskirstymo tankį galima įvertinti tiesiog naudojant vidurkio ir kovariacinės matricos įverčius, juos pritaikant apibrėžtam skirstiniui [12, 166, 186, 192, 210, 226]. Taigi, standartinis parametrinis metodas taikomas, kai prielaidos dėl tankio pavidalo yra tenkinamos. Vertinant tankį parametriniu būdu, reikia rasti skirstinio daugiamačio parametro θ reikšmę, o tai nėra paprasta, nes, didėjant dimensijai d, parametrų skaičius greitai auga. Pavyzdžiui, Gauso skirstinių mišinio atveju

$$\dim \theta = ½ \, qd(d + 1) + qd + q - 1, \tag{1.3}$$

ir net esant nedidelei dimensijai $d = q = 5$, modelį sudarys $\dim(\theta) = 104$ parametrai, o ieškant parametrų įverčių, gali tekti spręsti optimizavimo uždavinį 104-matėje erdvėje. Praktikoje klasterių skaičius q taip pat gali būti nežinomas, tuomet jį taip pat reikia įvertinti. Parametrinis metodas nėra naudingas, kai atsitiktinio dydžio skirstinys yra nežinomas. Tokiu atveju tam

tikroms tankio įverčių formoms nustatyti taikomi neparametriniai metodai [46, 66, 75, 153, 155, 165, 220].

Histograma – vienas paprasčiausių ir seniausių tankio įvertinių. Kiek žinoma, duomenys histogramos pavidalu (be grafinio vaizdavimo) pirmą kartą buvo pateikti 1661 metais nustatant mirtingumo tikimybes skirtingose amžiaus grupėse [212]. Patį terminą pirmasis pradėjo vartoti Karl Pearson [190] nuo 1891 metų. Aproksimuojant tankį $f(x)$ srityje Ω, skaičiuojamas stebinių $X(t)$ skaičius, patenkantis į Ω, ir dalijamas iš n bei srities Ω tūrio. Konstruojant įvertinį, pirmiausia randama erdvės sritis, į kurią patenka visi stebiniai, t. y. randami visų X projekcijų į ašis $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$ svyravimo intervalai. Stebinių svyravimo intervalai padalijami į l dalinių intervalų ir jais apribotuose hiperkubuose Ω_j ($j = 1, \ldots, r$) apskaičiuojamas tankio įvertis:

$$\hat{f}(x) = \frac{n(\Omega_j)}{n \cdot h_1 \cdot h_2 \ldots \cdot h_d}, \tag{1.4}$$

čia $n(\Omega_j)$ – į hiperkubą Ω_j patenkančių stebinių skaičius, o $h_j, j = 1, \ldots, d$ yra hiperkubo kraštinės [104, 128, 198]. Hiperkubų skaičių rekomenduojama (žr. [44, 187, 191] ir juose esančias nuorodas) parinkti $r \cong 1 + 3{,}32 \log n$, be to $l = \sqrt[d]{r}$ turi būti sveikasis skaičius, tai r yra parenkamas taip $\left\lceil \sqrt[d]{1 + 3{,}32 \log n} \right\rceil^d$, čia $\lceil \cdot \rceil$ žymi sveiką, suapvalintą į didesnę pusę, skaičių.

Histograma yra viena iš paprasčiausių duomenų pateikimo priemonių, kuri lengvai suprantama ir patogi. Šis įvertis yra funkcija, įgyjanti neneigiamas reikšmes, o jos integralas lygus vienetui. Tačiau jis nėra tolydus. Dėl to kyla problemų, kai svarbu žinoti tankio įverčio išvestines ir ypač, kai tankio vertinimas naudojamas tarpiniuose kitų metodų žingsniuose, pavyzdžiui, klasterizavimui taikant gradientinį algoritmą ar formuojant didelio matavimo duomenų lygio linijų grafiką.

1.1. Branduoliniai ir splaininiai įvertiniai

Anksčiau minėtos problemos nesunkiai galima išvengti bei pagerinti įvertinio tikslumą. Formuojant histogramą, reikia kiekvieną $X(t)$ įsivaizduoti kaip atskirą stulpelį, kurio aukštis $1/n$. Tada logiška stulpelio centrą pakeisti pačiu $X(t)$ ir gauti šią funkciją:

$$\hat{f}(x) = \frac{1}{n \cdot h_1 \cdot h_2 \ldots \cdot h_d} \sum_{t=1}^{n} I_{C_h(X(t))}(x), \tag{1.5}$$

čia C_h yra hiperkubas, kurio centras $X(t)$, o jo kraštinių ilgiai – $h_1, \ldots, h_d$. Ši statistika dažnai vadinama paprastuoju įvertiniu. Apibendrinant vietoj indikatorinės funkcijos kiekviename stebėtame taške galima naudoti glodųjį „iškilumą“ – branduolio funkciją. Tuomet daugiamatis fiksuoto pločio branduolinis tankio įvertinys (FKDE) su branduolio funkcija K ir fiksuotu (globaliu) branduolio pločio parametru h, kuriuo galima įvertinti daugiamačių duomenų $X \in \mathbf{R}^d$ tankį $\hat{f}(x)$, apibrėžiamas taip:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{t=1}^{n} K\left(\frac{x - X(t)}{h} \right). \tag{1.6}$$

Šiuolaikinėje duomenų analizėje tai vieni plačiausiai paplitusių pasiskirstymo tankio neparametrinių įvertinių [55, 63, 94, 144, 113, 118, 190]. Branduolio funkcija parenkama tokia, kad tenkintų sąlygą:

$$\int_{\mathbf{R}^d} K(x)dx = 1. \tag{1.7}$$

Kaip branduolys dažnai naudojama standartinio normalinio skirstinio tankio funkcija φ [79, 150]:

$$\varphi(x) = (2\pi)^{-d/2} \exp(-\tfrac{1}{2} x'x). \tag{1.8}$$

Dažnai stebiniai nebūna vienodai pasiskirstę visomis kryptimis. Todėl pageidautina pakeisti duomenų mastelį panaikinant didžiausius sklaidos skirtumus skirtingose koordinačių kryptyse. Vienas tam tinkamų metodų [76] yra duomenų standartizavimas, t. y. imties paveikimas tokia tiesine transformacija, kad transformuotų duomenų vidurkis būtų lygus nuliui, kovariacinė matrica būtų vienetinė, o (1.6) lygybę taikyti jau standartizuotiems duomenims. Tarkim, Z yra standartizuotas atsitiktinis vektorius:

$$Z = \mathbf{S}^{-1/2}(X - \overline{X}), \tag{1.9}$$

čia $\overline{X}$ yra empirinis imties vidurkis, o $\mathbf{S} \in \mathbf{R}^{d \times d}$ – empirinė kovariacinė matrica.

FKDE pagrindu sukonstruotas jau sudėtingesnis standartizuotų duomenų tankio įvertinys:

$$\hat{f}_Z(z) = \frac{1}{nh^d} \sum_{t=1}^{n} K\left(\frac{z - Z(t)}{h} \right), \tag{1.10}$$

$$\hat{f}(x) = \frac{(\det \mathbf{S})^{-1/2}}{nh^d} \sum_{t=1}^{n} K\left(\mathbf{S}^{-1/2} \frac{x - X(t)}{h} \right). \tag{1.11}$$

Optimalus FKDE branduolio plotis h^* nustatomas minimizuojant vidutinę integralinę kvadratinę paklaidą (MISE) [197]. Pavyzdžiui, kai stebinių skirstinys yra normalinis su vienetine kovariacine matrica, Gauso branduolių atveju h^* nustatyti buvo pasiūlyta [197] išraiška

$$h^* = An^{-\frac{1}{d+4}}, \text{ čia } A = [4/(2d+1)]^{\frac{1}{d+4}}. \tag{1.12}$$

Sudėtingesni branduolio pločio parinkimo metodai (tokie kaip mažiausių kvadratų kryžminio patikrinimo metodas) gaunami sudėtingesniais ir ilgesniais skaičiavimais [26, 65, 90, 100, 149, 162, 188, 203, 204, 221–224].

Praktiniuose tyrimuose branduolio plotis neretai parenkamas bandymų būdu. Jeigu h reikšmė yra maža, tankio funkcijos įvertis turi daugiau modų, kurios atitinka stebėtų duomenų išsidėstymą. Didesnė h reikšmė reiškia didesnį įverčio glodinimą.

Nors FKDE yra plačiai naudojami neparametriniam tankiui vertinti, jie dažnai turi ir tam tikrų praktinių trūkumų [197], pavyzdžiui, negali užtikrinti skirstinio galų vientisumo kartu per daug neglodinant pagrindinės tankio dalies.

Adaptuotas branduolinis tankio įvertinys. Geras FKDE patobulinimas yra adaptuotas branduolinis tankio įvertinys (AKDE) [197]. AKDE konstruojamas panašiai kaip FKDE: tankį aprašant branduoliu kiekviename stebėtame taške, tačiau šiuo atveju jau atsižvelgiama į branduolio plotį pereinant nuo vieno stebinio prie kito. Skirtingo glodumo srityse tikslinga imti skirtingus branduolių pločius. Šį metodą sudaro dvi pakopos: adaptuoto branduolio pločio bei tankio vertinimas branduoliniu metodu, naudojantis pirmame etape gauta informacija. Algoritmas gali būti apibendrintas taip:

1 žingsnis: Imties $\mathbf{X} = (X(1), \ldots, X(n))$ elementai standartizuojami į $\mathbf{Z} = (Z(1), \ldots, Z(n))$ taip, kad būtų $\hat{\mathbf{E}}[Z] = \mathbf{0}$ bei $\hat{\mathbf{E}}[ZZ'] = \mathbf{I}$.

2 žingsnis. Randami FKDE (1.10) įverčiai $\tilde{f}_Z(z)$, tenkinantys sąlygą $\tilde{f}_Z(Z(t)) > 0$, $\forall t$.

3 žingsnis. Nustatomas lokalus pločio parametras $\lambda_t = \left(\tilde{f}_Z(Z(t)) / g \right)^{-\gamma}$, čia g yra $\tilde{f}_Z(z)$ geometrinis vidurkis, t. y. $\log g = \frac{1}{n} \sum_{i=1}^{n} \log \tilde{f}_Z(Z(t))$, o γ – jautrumo parametras: $0 \le \gamma \le 1$.

4 žingsnis. Sudaromas adaptuotas branduolinis įvertinys $\hat{f}_Z(z)$ su kintamo pločio branduoliais:

$$\hat{f}_Z(z) = \frac{1}{n}\sum_{t=1}^{n} h^{-d}\lambda_t^{-d} K\left(\frac{z - Z(t)}{h\lambda_t}\right), \tag{1.13}$$

čia h – tas pats globalusis glodumo parametras kaip ir (1.6) lygybėje. Kuo didesnis γ, tuo jautresnė bus tankio atranka. Gana dažnai parenkamas $\gamma = \frac{1}{2}$ [1 , 197].

Pusiau parametrinis branduolinis tankio įvertinys. Kai duomenų yra nedaug, praktikoje parametriniai įverčiai dažnai yra taikomi ir tuo atveju, kai nežinomas tankis nėra parametrizuotas. Todėl vertas dėmesio yra parametrinio ir neparametrinio įverčių derinys. Pateiksime vieną tokį F. Hoti ir L. Holmström išnagrinėtą pusiau parametrinio branduolinio pasiskirstymo tankio įvertinio aprašymą, kuris tiriamąjį atsitiktinį vektorių suskaido į du subvektorius ir vieno iš jų pasiskirstymo tankį įvertina branduoliniu metodu, o kito sąlyginį tankį aproksimuoja normaliniu pasiskirstymo tankiu [108]. Tarkim, d ir s yra teigiami sveikieji skaičiai $d \geq 2$, $1 \leq s \leq d-1$. Taikant šį metodą, d-matis vektorius $X \in \mathbf{R}^d$ suskaidomas į du s ir $(d-s)$-mačius subvektorius $X = \begin{pmatrix} Y \\ Z \end{pmatrix}$, atitinkamai suskaidoma imtis: $\mathbf{X} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix}$, čia $Y \in \mathbf{R}^s$, $Z \in \mathbf{R}^{d-s}$. Vertinama tankio funkcija išreiškiama kaip atsitiktinio vektoriaus Y pasiskirstymo tankio ir atsitiktinio vektoriaus Z sąlyginio pasiskirstymo tankio sandauga:

$$f_X(x) = f_{(Y,Z)}(y,z) = f_Y(y) f_{Z|Y=y}(z \mid y), \; x = \begin{pmatrix} y \\ z \end{pmatrix} \in \mathbf{R}^d. \tag{1.14}$$

Čia f_X ir f_Y yra atitinkamai X ir Y tankiai, o $f_{Z|Y=y}$ yra Z tankis, kai $Y = y$. Tarkime, jog sąlyginis tankis $f_{Z|Y=y}$ yra Gauso, t. y. daugiamatis normalinis, bet tankis f_Y nepriklauso jokiai parametrinių funkcijų šeimai. Tada tankio f_X įvertinys gaunamas vertinant f_Y neparametriniu būdu ir taikant daugiamatį normalinį tankį kiekvienam $f_{Z|Y=y}$. Tankio funkcija $f_Y(y)$, kaip ir (1.11) yra vertinama branduoliniu metodu [197]. Kadangi imties elementai nėra standartizuoti, todėl glodumo parametras visomis kryptimis nėra vienodas ir, taikant branduolinį metodą, jis keičiamas s-mate matrica H:

$$\hat{f}(y) = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\det(H)} K\left(H^{-1}(y - Y(t))\right). \tag{1.15}$$

Dažniausiai H forma parenkama diagonali – $H = \mathrm{diag}(h_1, \ldots, h_s)$, o glodumo parametrai

$$h_j = \left(\frac{4}{s+2}\right)^{1/(s+4)} n^{-1/(s+4)} \sigma_j . \tag{1.16}$$

Pažymėtina, kad šią formą, kai s=1, pasiūlė B. W. Silverman [197]. Komponento Y_j standartinį nuokrypį σ_j pakeitus jo įverčiu ir pastebėjus, kad pirmasis daugiklis visada yra tarp 0,924 ir 1,059, gaunama D. W. Scott [190] taisyklė:

$$\hat{h}_j = n^{-1/(s+4)} \hat{\sigma}_j . \tag{1.17}$$

Šią Scott taisyklę nesunku apibendrinti glodumo matricai H:

$$\hat{H} = n^{-1/(s+4)} \hat{\Sigma}^{1/2}, \tag{1.18}$$

čia $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, ..., \hat{\sigma}_s^2)$ yra Y empirinių dispersijų diagonalioji matrica.

Sąlyginis tankis $f_{Z|Y}(\cdot \mid y)$ aproksimuojamas Gauso skirstiniu $\mathcal{N}(m(y), C(y))$, čia $m(y)$, $C(y)$ žymi vektoriaus Y sąlyginį vidurkį ir sąlyginę kovariacinę matricą:

$$m(y) = \mathbf{E}(Z \mid Y = y), y \in \mathbf{R}^s, \tag{1.19}$$

$$C(y) = \mathbf{E}\left[(Z - m(y))'(Z - m(y)) \mid Y = y\right], y \in \mathbf{R}^s. \tag{1.20}$$

Vertinant $m(y)$ ir $C(y)$, siūloma taikyti branduolinį glodinimą:

$$\hat{m}(y) = \frac{\sum_{t=1}^{n} K_{H_2}(y - Y(t)) Z(t)}{\sum_{j=1}^{n} K_{H_2}(y - Y(j))} = \sum_{t=1}^{n} W_{H_2}(y - Y(t)) Z(t), y \in \mathbf{R}^s, \tag{1.21}$$

čia svoriai

$$W_{H_2}(y - Y(t)) = \frac{K_{H_2}(y - Y(t))}{\sum_{j=1}^{n} K_{H_2}(y - Y(j))}, \tag{1.22}$$

kurių suma lygi vienetui. (1.21) formulė gali būti suprantama kaip Nadaraya ir Watson [158], [225] sąlyginio vidurkio funkcijos m regresinis įvertinys. Panašiai galima vertinti sąlyginę kovariacinę matricą:

$$\hat{C}(y) = \sum_{t=1}^{n} W_{H_3}(y - Y(t))(Z(t) - \hat{m}(y))'(Z(t) - \hat{m}(y)), \; y \in \mathbf{R}^s. \tag{1.23}$$

Parametrinis sąlyginio tankio $f_{Z|Y=y}$ įvertinys atrodo taip:

$$\hat{f}_{Z|Y=y}(z) = [(2\pi)^{d-s} \det \hat{C}(y)]^{-1/2} \exp\left\{-\tfrac{1}{2}(z - \hat{m}(y))\hat{C}(y)^{-1}(z - \hat{m}(y))'\right\}, \; z \in \mathbf{R}^{d\text{-}s}. \tag{1.24}$$

Tuomet X pasiskirstymo tankio f_X įvertis yra:

$$\hat{f}_X(x) = \hat{f}_{(Y,Z)}(y,z) = \hat{f}_Y(y)\hat{f}_{Z|Y=y}(z), \; x=(y,z) \in \mathbf{R}^d. \tag{1.25}$$

Anksčiau aprašyta procedūra vadinama pusiau parametriniu branduoliniu tankio įvertiniu, sutrumpintai – SKDE. Praktikoje net jeigu sąlyginė kelių atsitiktinio vektoriaus komponenčių normalumo prielaida tenkinama, tankio vertinimo rezultatų tikslumui turi įtakos ir suskaidymo dimensijos s, taip pat ir pačių koordinačių parinkimas. Vienas iš būdų jiems parinkti yra darbe [108] rekomenduojamas naudoti mažiausių kvadratų metodu ar maksimaliu tikėtinumu pagrįstas kryžminio patikrinimo metodas. Parametrus H_2 ir H_3 [108] autoriai siūlo parinkti lygius $2H$.

Histosplaininis tankio įvertinys. Teoriniai splainų glodinimo pagrindai yra aprašyti Duchon [50, 51], Meinguet [154] bei kituose darbuose, pavyzdžiui, [22, 37, 40, 48], ir išplėtoti Delicado bei del Rio [44] ir kitų autorių [57, 205]. Pasiskirstymo tankio vertinimas histosplainu susideda iš dviejų etapų. Pirmame etape sudaroma d-matė histograma. Antrame etape, jau turint histograminį netikslų tankio įvertį, ieškoma daugiamačio splaino regresinė priklausomybė, leidžianti patikslinti pirmame etape gautą tankio įvertį.

Histogramos (1.4) įverčius aproksimuojant d-mačiu splainu, pirmiausia d-matis baigtinis histogramos vidurio taškų tinklelis $\mathbf{x}$ padalijamas į s ir $(d–s)$-mačius subtinklelius $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ taip, kad $\mathbf{z}$ apibūdina regresijos tiesinę dalį, t. y. priklausomybė tarp $\mathbf{z}$ ir tankio įverčių histogramos hiperkubų centruose $\mathbf{w}$ yra tiesinė, o $\mathbf{y}$ apibūdina regresijos netiesinę dalį. Tada regresijos lygtis atrodo taip:

$$w(k) = g(y(k)) + z(k)\beta + e(k), \tag{1.26}$$

čia g yra nežinoma tolydžioji funkcija, β – nežinomas (d–s)-matis parametrų vektorius, o $e(k)$, $k = 1, \ldots, r$ yra nepriklausomos atsitiktinės paklaidos, kurių vidurkis lygus nuliui.

Aproksimavimo d-mačiu splainu procedūra sudaryta iš dviejų etapų – tiesinio ir netiesinio modelių vertinimo. $z(k)\beta$ yra tiesinė parametrinė modelio dalis, o $z(k)$ – tos regresijos nepriklausomi kintamieji. $g(y(k))$ yra netiesinė modelio dalis.

Regresijos koeficientai randami mažiausių kvadratų metodu:

$$\frac{1}{r}\sum_{k=1}^{r}(w(k) - g(y(k)) - z(k)\beta)^2 , \tag{1.27}$$

Netiesinė regresijos dalies funkcija $g(y(k))$ apibrėžiama [50] taip:

$$g(y(k)) = \theta_0 + \sum_{i=1}^{s}\theta_i y_i(k) + \sum_{j=1}^{r}\delta_j E_2(y(k) - y(j)), \tag{1.28}$$

čia $E_2(y(k) - y(j)) = \frac{1}{2^3\pi}\|y(k) - y(j)\|^2 \log(\|y(k) - y(j)\|)$.

Taigi, sprendžiant regresijos lygtį randami parametrai (β, δ, θ). Pažymėjus $\mathbf{K} = (K)_{kj} = E_2(y(k) - y(j))$ ir $\mathbf{T} = (T)_{kj} = (y_j(k))$, koeficientai (β, δ, θ) randami mažiausių kvadratų metodu minimizuojant funkciją $S(\beta, \delta, \theta)$:

$$S(\beta, \delta, \theta) = \tfrac{1}{r}\|\mathbf{y} - \mathbf{T}\theta - \mathbf{K}\delta - \mathbf{z}\beta\|^2 . \tag{1.29}$$

Tuomet histosplaininis tankio įvertinys (HSDE) yra

$$\hat{f}(x) = w(x; \hat{\beta}, \hat{\delta}, \hat{\theta}) . \tag{1.30}$$

Iš pateikto algoritmo aprašymo matyti, kad tankio vertinimo rezultatų tikslumui įtakos turi tinklelio $\mathbf{x}$ koordinačių suskirstymas į subtinklelius $(\mathbf{y}, \mathbf{z})$. Vienas iš tokio suskirstymo parinkimo būdų – naudoti Fišerio kriterijų hipotezei apie modelio netiesiškumą tikrinti (pvz., kai reikšmingumo lygmuo $\alpha = 0{,}05$).

Logsplaininis tankio įvertinys. Vienamačiais polinominiais splainais vadinami tam tikro laipsnio daliniai daugianariai. Lūžio taškai, kuriuose pereinama iš vieno daugianario į kitą,

vadinami mazgais. Tarkim, vektorius $\mathbf{t} = (t_1, ..., t_K) \in \mathbf{R}^K$ apibrėžia tokių K taškų rinkinį. Paprastai splainas nusako glodžius ryšius, rodančius, kaip sujungiamos skirtingos sritys, padalytos mazgais [185]. Šie apribojimai tiksliai nusakomi išreiškiant dalinius daugianarius tolydžiųjų išvestinių skaičiumi s. Turimos omenyje, pavyzdžiui, iš dalies tiesinės kreivės. Jei nėra jokių apribojimų, šių funkcijų mazguose leidžiami lūžio taškai. Iškėlus sąlygą, kad funkcijos yra globaliai tolydžios, reikalaujama, kad atskiros tiesinės dalys kiekviename mazge sueitų. Jei reikalingas didesnis glodumas (tolydžiųjų pirmosios eilės išvestinių), tuomet prarandamas mazgų lankstumas, o kreivės laikomos paprastomis tiesinėmis funkcijomis. Aproksimavimo teorijos literatūroje terminas „tiesinis splainas“ taikomas tolydžiajai dalinei tiesinei funkcijai. Atitinkamai terminas „kubinis splainas“ priskirtas tolydžiosioms kubinėms funkcijoms, turinčioms antrosios eilės tolydžiąsias išvestines, o mazguose leidžiančioms trečiosios eilės išvestinių šuolius. Apskritai įprasta naudoti splainus, glodžiausius ta prasme, kad visame daugianaryje būtų įtraukta kuo daugiau kokių nors tolydumo sąlygų.

Jei daugianario laipsnis yra b, o mazgų vektorius $\mathbf{t}$, tai daugianarių splainų, kurie turi s tolydžiųjų išvestinių, rinkinys sudaro tiesinę erdvę. Pavyzdžiui, tiesinių splainų rinkinys su mazgais sekoje $\mathbf{t}$ nusakomas funkcijomis

$$1, x, (x - t_1)_+, ..., (x - t_K)_+, \tag{1.31}$$

čia $(\cdot)_+ = \max(\cdot, 0)$. Šia aibe remsimės kaip erdvės baze. Apibendrinant reikia pasakyti, kad b laipsnio ir s glodumo splaino erdvės bazė yra sudaryta iš vienanarių, kurių pavidalas $(x - t_k)_+^{s+j}$, čia $1 \le j \le b - s$. Naudojantis šia formule, klasikinių kubinių splainų atveju $b = 3$ ir $s = 2$, taigi bazę sudaro elementai

$$1, x, x^2, x^3, (x - t_1)^3_+, ..., (x - t_K)^3_+. \tag{1.32}$$

Modelio požiūriu ši bazė yra patogi, nes atskiros funkcijos mazgų vietose susijungia. (1.31) ir (1.32) išraiškose kiekviena funkcija kaip tik siejasi su vienu iš mazgų, o šios funkcijos pašalinimas iš esmės atitinka paties mazgo pašalinimą.

Yra žinoma, kad funkcijų (1.31) ir (1.32) skaitinės savybės yra menkos. Pavyzdžiui, tiesinės regresijos uždaviniuose sprendimų matrica blogėja taip greitai, kaip greitai mažėja mazgų skaičius. Reikšminga to alternatyva yra vadinamoji *B-splaino bazė* [23, 136]. Šios funkcijos sudaromos taip, kad būtų palaikomos keliuose gretimuose intervaluose, kurie apibrėžiami mazgais (glodžiausiems splainams naudojami b+1 gretimi intervalai). Tarkime,

galima rasti bazę $B_1(x; \mathbf{t})$, ..., $B_J(x; \mathbf{t})$ b laipsnio erdvės splainams, kurių glodumas s, o mazgų seka $\mathbf{t}$, kad bet kuri funkcija erdvėje galėtų būti užrašyta kaip

$$g(x;\boldsymbol{\beta},\mathbf{t}) = \beta_1 B_1(x;\mathbf{t}) + ... + \beta_J B_J(x;\mathbf{t}), \tag{1.33}$$

kur atitinkamas koeficientų vektorius $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)'$. Jei naudojamos didžiausio glodumo splainų erdvės, tai $J = K+b+1$, kaip matyti iš (1.31) ir (1.32).

Pagal poskyrio pavadinimą šios analizės objektas yra logaritminis tankis. Tarkim, X yra atsitiktinis vektorius, įgyjantis reikšmes iš intervalo (L, U). Atskiru atveju L ir U gali būti $\pm\infty$. Kooperberg ir Stone [99, 137, 138, 206] metodas, žinomas logsplaino pavadinimu (LSDE), yra realizuotas su kubiniu splainu. Kubinis splainas aprašomas (1.32); šios funkcijos taip pat yra dukart tolydžiai diferencijuojamos, o daliniai daugianariai atitinkamai apibrėžiami mazgų sekoje $\mathbf{t} = (t_1, ..., t_K)$. Kiekviename intervale $[t_1, t_2]$, ..., $[t_{K-1}, t_K]$ kubiniai splainai yra kubiniai daugianariai, tačiau kraštuose $(L, t_1]$ ir $[t_K, U)$ jie yra tiesinės funkcijos. Mažiausias mazgų skaičius yra $K \geq 3$ (priešingu atveju turima tiesinė funkcija arba konstanta). Parenkama bazė, kurios pavidalas yra 1, $B_1(x; \mathbf{t})$, ..., $B_J(x; \mathbf{t})$, čia $J = K - 1$.

Sakoma, vektorius $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)' \in \mathbf{R}^J$ egzistuoja, jei:

$$C(\boldsymbol{\beta},\mathbf{t}) = \log\left(\int_L^U \exp(\beta_1 B_1(x;\mathbf{t}) + ... + \beta_J B_J(x;\mathbf{t}))dx\right) < \infty. \tag{1.34}$$

Tarkim, $\mathcal{B}$ žymi tokių galimų vektorių rinkinį. Parinkus $\boldsymbol{\beta} \in \mathcal{B}$, apibrėžiama teigiamųjų tankio funkcijų šeima intervale (L, U). Jos pavidalas:

$$g(x;\boldsymbol{\beta},\mathbf{t}) = \exp(\beta_1 B_1(x;\mathbf{t}) + ... + \beta_J B_J(x;\mathbf{t}) - C(\boldsymbol{\beta},\mathbf{t})),\ L<x<U. \tag{1.35}$$

Dabar, turint n didumo atsitiktinę imtį $X(1)$, ..., $X(n)$ iš intervalo (L, U) su nežinoma tankio funkcija f, logtikėtinumo funkcija, atitinkanti logsplainų (1.35) modelį, yra:

$$l(\boldsymbol{\beta},\mathbf{t}) = \sum_i g(X_i;\boldsymbol{\beta},\mathbf{t}) = \sum_i \sum_j \beta_j B_j(X_i;\mathbf{t}) - nC(\boldsymbol{\beta},\mathbf{t}),\ \boldsymbol{\beta} \in \mathcal{B}. \tag{1.36}$$

Maksimalaus tikėtinumo įvertis $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathcal{B}} l(\boldsymbol{\beta},\mathbf{t})$, atitinkamai tankio įvertinys $\hat{f} = g(x;\hat{\boldsymbol{\beta}},\mathbf{t})$, $L < x < U$.

Sakykime, pažingsninio nustatymo procedūros metu modelių seka žymima v, v-asis modelis turi J_v bazinių funkcijų. Geriausiam modeliui parinkti naudojamas apibendrintas Akaike informacijos kriterijus (AIC) [3]. Tarkim, kad $\hat{l}_v$ nusako logtikėtinumo funkcijos (1.36) įvertį v-ajam modeliui ir kad $\mathrm{AIC}_{a,v}(\mathbf{t}) = -2\hat{l}_v(\mathbf{t}) + aJ_v$ lygtimi apibrėžiamas Akaike informacijos kriterijus, kurio modelio nuostolių parametras a. Iš daugelio modelių parenkamas tas, kurio v reikšmė minimizuoja $\mathrm{AIC}_{a,v}$. Stone [206] rekomenduoja naudoti $a = \log n$.

1.2. Duomenų projektavimu paremti tankių įvertiniai

Apvertimo formulės tankio įvertinys. Nagrinėjant parametrinių metodų aproksimacijas, reikia pabrėžti, jog, didėjant duomenų dimensijai, modelio parametrų skaičius sparčiai auga, todėl sunkiau rasti tikslius parametrų įverčius. Vienamačių duomenų projekcijų

$$X_\tau = \tau' X \tag{1.37}$$

tankį f_τ rasti daug lengviau negu daugiamačių duomenų tankį f. Kadangi egzistuoja abipus vienareikšmė atitiktis

$$f \leftrightarrow \{f_\tau,\ \tau \in \mathbf{R}^d\}, \tag{1.38}$$

tai visai natūralu yra bandyti rasti daugiamatį tankį f naudojant vienamačių stebinių projekcijų tankių įverčius $\hat{f}_\tau$ [132].

Pažymėtina, kad mišinio (1.1) atveju, kai skirstiniai yra Gauso, stebinių projekcijos (1.37) taip pat pasiskirsčiusios pagal (vienamatį) Gauso skirstinių mišinio modelį:

$$f_\tau(x) = \sum_{k=1}^{q} p_k(\tau)\varphi_{k,\tau}(x) = f_\tau(x, \theta(\tau)), \tag{1.39}$$

čia $\varphi_{k,\tau}(x) = \varphi(x; m_k(\tau), \sigma_k^2(\tau))$ – vienamatis Gauso tankis. Daugiamačio mišinio parametrą θ ir duomenų projekcijų pasiskirstymo parametrus $\theta(\tau) = (p_k(\tau), m_k(\tau), \sigma_k^2(\tau))$, $k = 1, ..., q$ sieja lygybės:

$$
\begin{aligned}
p_j(\tau) &= p_j, \\
m_j(\tau) &= \tau' M_j, \\
\sigma_j^2(\tau) &= \tau' R_j \tau.
\end{aligned}
\tag{1.40}
$$

Pasinaudojus apvertimo formule

$$
f(x) = \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} e^{-it'x} \psi(t) dt, \tag{1.41}
$$

čia

$$
\psi(t) = \mathbf{E} e^{-it'X} \tag{1.42}
$$

žymi atsitiktinio dydžio X charakteristinę funkciją. Pažymėjus $u = |t|$, $\tau = t/|t|$ ir pakeitus kintamuosius į sferinę koordinačių sistemą, gaunama:

$$
f(x) = \frac{1}{(2\pi)^d} \int_{\tau:|\tau|=1} ds \int_0^\infty e^{-iu\tau'x} \psi(u\tau) u^{d-1} du. \tag{1.43}
$$

Čia pirmasis integralas suprantamas kaip vienetinės sferos paviršinis integralas.

Pažymėjus stebimo atsitiktinio dydžio projekcijos charakteristinę funkciją

$$
\psi_\tau(u) = \mathbf{E} e^{-iu\tau'X}, \tag{1.44}
$$

galioja

$$
\psi(u\tau) = \psi_\tau(u). \tag{1.45}
$$

Pasirinkus projektavimo krypčių, tolygiai išsidėsčiusių ant sferos, aibę T ir charakteristinę funkciją keičiant jos įvertiniu, gaunama formulė įverčiui apskaičiuoti [132, 133, 2A]:

$$
\hat{f}(x) = \frac{c(d)}{\#T} \sum_{\tau \in T} \int_0^\infty e^{-iu\tau'x} \hat{\psi}_\tau(u) u^{d-1} e^{-hu^2} du, \tag{1.46}
$$

čia ir toliau # žymi aibės elementų skaičių. Pasinaudojus *d*-mačio rutulio tūrio formule

$$V_d(R) = \frac{\pi^{\frac{d}{2}} R^d}{\Gamma(\frac{d}{2}+1)} = \begin{cases} \frac{\pi^{\frac{d}{2}} R^d}{(\frac{d}{2})!}, & \text{kai } d \bmod 2 \equiv 0 \\ \frac{2^{\frac{d+1}{2}} \pi^{\frac{d-1}{2}} R^d}{d!!}, & \text{kai } d \bmod 2 \equiv 1 \end{cases} \tag{1.47}$$

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

$$c(d) = \frac{(V_d(1))'_R}{(2\pi)^d} = \frac{d 2^{-d} \pi^{-\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}. \tag{1.48}$$

Kompiuterinio modeliavimo tyrimai parodė, jog naudojant apvertimo formulę gauti tankio įverčiai yra neglodūs. Todėl (1.46) formulėje po integralo ženklu naudojamas papildomas daugiklis e^{-hu^2}. Šis daugiklis įvertį $\hat{f}(x)$ papildomai glodina su Gauso branduolio funkcija. Kaip matome toliau, tokia daugiklio forma leidžia analitiškai apskaičiuoti integralo reikšmę, o Monte Karlo tyrimai parodė, jog jį naudojant gerokai sumažėja įverčių paklaidos.

(1.46) formulė gali būti naudojama esant įvairiems projektuotų duomenų charakteristinės funkcijos įvertiniams. Aptarsime du būdus, kurie naudojami šiame darbe. Vienas jų remiasi tankio aproksimacija Gauso skirstinių mišinio modeliu (IFDE). Nagrinėjamu atveju naudojamas parametrinis charakteristinės funkcijos įvertinys:

$$\hat{\varphi}_\tau(u) = \sum_{k=1}^{\hat{K}_\tau} \hat{p}_k(\tau) e^{iu\hat{m}_k(\tau) - u^2\hat{\sigma}_k^2(\tau)/2}. \tag{1.49}$$

Į (1.46) įrašius (1.49), gaunama:

$$\begin{aligned} \hat{f}(x) &= \frac{c(d)}{\#T} \sum_{\tau\in T} \sum_{k=1}^{\hat{K}_\tau} \hat{p}_k(\tau) \int_0^\infty e^{iu(\hat{m}_k(\tau)-\tau' x) - u^2(h+\hat{\sigma}_k^2(\tau)/2)} u^{d-1} du \\ &= \frac{c(d)}{\#T} \sum_{\tau\in T} \sum_{k=1}^{\hat{K}_\tau} \hat{p}_k(\tau) I_{d-1}\left(\frac{\hat{m}_k(\tau) - \tau' x}{\sqrt{\hat{\sigma}_k^2(\tau) + 2h}} \right) \left(\sqrt{\hat{\sigma}_k^2(\tau) + 2h} \right)^{-d}, \end{aligned} \tag{1.50}$$

čia

$$I_j(y) = \mathrm{Re}\left[\int_0^\infty e^{iyz - z^2/2} z^j dz\right]. \tag{1.51}$$

Pažymėtina, jog čia galima nagrinėti tik realią išraiškos dalį (menamųjų dalių suma turi būti lygi nuliui), nes tankio įvertis $\hat{f}(x)$ gali įgyti tik realias reikšmes. Pasirinkta glodinimo daugiklio forma e^{-hu^2} leidžia susieti glodinimo parametrą h su projekcijų klasterių dispersijomis – skaičiavimuose dispersijos tiesiog padidinamos dydžiu $2h$.

Apskaičiuojama (1.51) išraiška. Pažymėjus

$$K_j(y) = \int_0^\infty \cos yz \cdot e^{-z^2/2} \cdot z^j dz\,, \tag{1.52}$$

$$S_j(y) = \int_0^\infty \sin yz \cdot e^{-z^2/2} \cdot z^j dz\,, \tag{1.53}$$

galioja lygtis

$$\int_0^\infty e^{-iyz - z^2/2} z^j dz = K_j(y) + iS_j(y)\,. \tag{1.54}$$

Integruojant dalimis, gaunama:

$$\begin{aligned} K_j(y) &= e^{-z^2/2} z^{j-1} \cos yz\Big|_0^\infty + \int_0^\infty e^{-z^2/2}((j-1)z^{j-2}\cos yz - yz^{j-1}\sin yz)dz = \\ &= \mathbf{1}_{\{j=1\}} + (j-1)K_{j-2}(y) - yS_{j-1}(y),\ j \geq 1. \end{aligned} \tag{1.55}$$

Analogiškai išreiškus $S_j(y)$ bei atsižvelgus į j indekso apribojimus, gaunamos rekurentinės lygtys:

$$K_j(y) = (j-1)K_{j-2}(y) - yS_{j-1}(y)\,, j \geq 2, \tag{1.56}$$

$$K_1(y) = 1 - yS_0(y)\,, \tag{1.57}$$

$$S_j(y) = (j-1)S_{j-2}(y) - yK_{j-1}(y)\,, j \geq 2, \tag{1.58}$$

$$S_1(y) = yK_0(y)\,. \tag{1.59}$$

Funkcijoms $K_0(y)$ bei $S_0(y)$ apskaičiuoti pasinaudojama tuo, kad

$$(S_0(y))'_y = \int_0^\infty z\cos yz \cdot e^{-z^2/2} dz = K_1(y). \tag{1.60}$$

Iš (1.57) ir (1.60) gaunama, kad S_0 tenkina diferencialinę lygtį

$$S_0'(y) = 1 - yS_0(y),\ S_0(0) = 0. \tag{1.61}$$

Ši lygtis sprendžiama S_0 skleidžiant Teiloro eilute:

$$S_0'(y) = \sum_{l=0}^{\infty} c_{l+1}(l+1)y^{l+1} = 1 - \sum_{l=2}^{\infty} c_{l-1}y^l\ . \tag{1.62}$$

Sulyginus koeficientus prie panašių narių, randamos jų reikšmės:

$$\begin{gathered} c_0 = 0,\ c_1 = 1, \\ c_l = -c_{l-2}/l,\ l \geq 2. \end{gathered} \tag{1.63}$$

Taigi,

$$S_0(y) = \sum_{l=0}^{\infty} \frac{(-1)^l y^{2l+1}}{(2l+1)!!} = y - \frac{y^3}{3!!} + \frac{y^5}{5!!} - \frac{y^7}{7!!} + \dots . \tag{1.64}$$

K_0 randama iš (1.52) išraiškos:

$$\begin{aligned} K_0(y) &= \int_0^\infty \cos yz \cdot e^{-z^2/2} dz = \frac{1}{2}\int_{-\infty}^{\infty} \cos yz \cdot e^{-z^2/2} dz \\ &= \frac{1}{2}\int_{-\infty}^{\infty} (\cos yz - i\sin yz) \cdot e^{-z^2/2} dz = \sqrt{\frac{\pi}{2}} e^{-y^2/2}. \end{aligned} \tag{1.65}$$

Ieškomo integralo (1.46) reikšmė

$$I_j(y) = K_j(y). \tag{1.66}$$

Kitas naudotas (1.46) formulės taikymo būdas vertinant pasiskirstymo tankį yra ne aproksimuoti jį Gauso skirstinių mišiniu, o vertinti neparametriškai branduoliniu AKDE metodu kaip branduolio funkciją naudojant Gauso pasiskirstymo tankį φ ir taikyti apvertimo formulę (sutrumpintai vadinama IKDE). Šis būdas įgalina (1.46) dešinėje pusėje esantį integralą apskaičiuoti analitiškai [3A]. Kiekvienam $\tau \in T_0$

$$\hat{f}_\tau(v) = \frac{1}{n}\sum_{t=1}^{n} \varphi\left(\frac{v - \tau' X(t)}{h_t}\right) \Big/ h_t \,,\; h_t = h_t(\tau) \tag{1.67}$$

ir

$$\hat{\psi}_\tau(u) = \frac{1}{n}\sum_{t=1}^{n} \exp\left\{iu\,\tau' X(t) - h_t^2 u^2 / 2\right\}. \tag{1.68}$$

Vartojant tuos pačius žymėjimus kaip ir AKDE algoritmo aprašyme, turima $h_t(\tau) = h\lambda_t(\tau)$, čia h randamas naudojant (1.12), o λ_t yra lokalusis glodumo parametras.

Į (1.46) įrašius (1.73), gaunama:

$$\begin{aligned} \hat{f}(x) &= \frac{c(d)}{\#T}\sum_{\tau\in T}\sum_{t=1}^{n}\frac{1}{n}\int_0^\infty e^{iu(X(t)-\tau' x)-u^2(h+h_t^2(\tau)/2)} u^{d-1}\,du \\ &= \frac{c(d)}{\#T}\sum_{\tau\in T}\sum_{t=1}^{n}\frac{1}{n} I_{d-1}\left(\frac{\tau' X(t) - \tau' x}{\sqrt{h_t^2(\tau) + 2h}}\right)\left(\sqrt{h_t^2(\tau) + 2h}\right)^{-d}. \end{aligned} \tag{1.69}$$

Elgiantis kaip (1.50) formulės atveju randama $I_j(y)$ išraiška (1.66).

Pasinaudojus lygybe

$$f^{(k)}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iu}\psi(u)(iu)^k\,du\,, \tag{1.70}$$

įverčio (1.69) išraišką galima užrašyti taip:

$$\hat{f}(x) = \frac{c(d)}{\#T}\sum_{\tau\in T}\sum_{t=1}^{n}\frac{1}{n}\varphi^{(d-1)}\left(\frac{\tau' X(t) - \tau' x}{\sqrt{h_t^2(\tau) + 2h}}\right)\left(\sqrt{h_t^2(\tau) + 2h}\right)^{-d}, \tag{1.71}$$

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

Apvertimo formulės modifikuotas tankio įvertinys. Vienas iš apvertimo formulės metodo, apibrėžto (1.46), trūkumų yra tas, kad šiuo įvertiniu aprašomas Gauso skirstinių mišinio modelis (1.1) (kai $f_k = \varphi_k$) gerai vertina tik jam artimo pasiskirstymo stebinių tankį. Aproksimuojant tiriamą tankį Gauso skirstinių mišiniu dažnai IFDE tampa sudėtingas dėl didelio komponentų su mažomis *apriorinėmis* tikimybėmis skaičiaus. Jų skaičių galima sumažinti įvedant triukšmo klasterį.

Panagrinėkime modifikuotą algoritmą, sukurtą remiantis daugiamačių Gauso skirstinių mišinio modelio naudojimu. Pasinaudokime apvertimo formule (1.41). Apibrėžkime parametrinį tolygaus skirstinio tankio charakteristinės funkcijos parametrinį įvertinį:

$$\hat{\varphi}(u) = \frac{2}{(b-a)u} \sin\frac{(b-a)t}{2} u \cdot e^{\frac{iu(a+b)}{2}}. \tag{1.72}$$

Tankio įverčio skaičiavimo formulėje (1.46) charakteristinės funkcijos įvertinį konstruokime kaip Gauso skirstinių mišinio ir tolygaus skirstinio charakteristinių funkcijų sąjungą su atitinkamomis *apriorinėmis* tikimybėmis:

$$\hat{\varphi}_\tau(u) = \sum_{k=1}^{K_\tau} \hat{p}_k(\tau) e^{iu\hat{\mu}_k(\tau) - u^2\hat{\sigma}_k^2(\tau)/2} + \hat{p}_0(\tau) \frac{2}{(b-a)u} \sin\frac{(b-a)u}{2} \cdot e^{\frac{iu(a+b)}{2}}, \tag{1.73}$$

čia antrasis narys aprašo tolygaus pasiskirstymo triukšmo klasterį, $\hat{p}_0$ – triukšmo klasterio svoris, $a = a(\tau)$, $b = b(\tau)$. Remdamiesi nustatytais tolygaus skirstinio parametrų įverčiais ir projektuotais duomenimis, galime užrašyti:

$$a = (\tau' x)_{\min} - \frac{(\tau' x)_{\max} - (\tau' x)_{\min}}{2(n-1)} \tag{1.74}$$

ir

$$b = (\tau' x)_{\max} + \frac{(\tau' x)_{\max} - (\tau' x)_{\min}}{2(n-1)}. \tag{1.75}$$

Įrašę (1.73) į (1.46), gauname:

$$\hat{f}(x) = \frac{c(d)}{\#T} \sum_{\tau \in T} \left[\sum_{k=1}^{k_\tau} p_k(\tau) \int_0^\infty e^{iu(m_k(\tau) - \tau' x) - u^2 (h + \sigma_k^2(\tau)/2)} u^{d-1} du \right.$$
$$\left. + \frac{2 p_0(\tau)}{b-a} \int_0^\infty e^{iu\left(\frac{a+b}{2} - \tau' x\right) - u^2 h} \cdot \sin \frac{b-a}{2} u \cdot u^{d-2} du \right]. \tag{1.76}$$

Naudodami žymėjimus kaip ir (1.50), galime užrašyti:

$$\hat{f}(x) = \frac{c(d)}{\#T} \sum_{\tau \in T} \left[\sum_{k=1}^{k_\tau} p_k(\tau) I_{d-1}\left(\frac{m_k(\tau) - \tau' x}{\sqrt{\sigma_k^2(\tau) + 2h}} \right) \cdot \left(\sigma_k^2(\tau) + 2h\right)^{-\frac{d}{2}} \right.$$
$$\left. + \frac{2 p_0(\tau)}{b-a} J_{d-2}\left(\frac{a+b-2\tau' x}{2\sqrt{2h}}, \frac{b-a}{2\sqrt{2h}} \right) \cdot (2h)^{-\frac{d-1}{2}} \right], \tag{1.77}$$

čia $I_j(y)$ išraiška yra kaip ir (1.51), o jo reikšmė – (1.66) bei

$$J_j(y,z) = \mathrm{Re}\left[\int_0^\infty e^{iyu - u^2/2} \cdot \sin zu \cdot u^j du \right]. \tag{1.78}$$

Integruodami gauname:

$$\int_0^\infty e^{iyu - u^2/2} \cdot \sin zu \cdot u^j du = \int_0^\infty (\cos yu + i \sin yu) \cdot \sin zu \cdot e^{-u^2/2} \cdot u^j du$$
$$= \int_0^\infty \left(\frac{\sin(y+z)u + \sin(z-y)u}{2} + i \frac{\cos(y-z)u - \cos(y+z)u}{2} \right) \cdot e^{-u^2/2} u^j du \cdot \tag{1.79}$$
$$= \frac{1}{2} S_j(y+z) + \frac{1}{2} S_j(z-y) + i \frac{1}{2} K_j(y-z) - i \frac{1}{2} K_j(y+z),$$

čia $S_j(y)$ ir $K_j(y)$ apibrėžti (1.52) ir (1.53).

Ieškomo integralo (1.78) reikšmė

$$J_j(y,z) = ½ S_j(y+z) + ½ S_j(z-y). \tag{1.80}$$

Anksčiau minėta procedūra vadinama modifikuotu apvertimo formulės tankio įvertiniu (sutrumpintai – MIDE) [5A].

Tikslinio projektavimo tankio įvertinys. Subtilesnė yra J. H. Friedman ir bendraautorių idėja [71, 72, 73], leidžianti išvengti daugelio su minėtos apvertimo formulės taikymu susijusių

sunkumų (keblu parinkti glodinimo parametrus, reikalingi didelio skaičiaus projekcijų pasiskirstymo tankio įverčiai ir t. t.). Tikslinio projektavimo tankio įvertinio (PPDE) sudarymo esmė yra ieškojimas „įdomių“, mažo matavimo duomenų projekcijų, kurios parodo skirstinio struktūras, t. y. kur projekcijos turi skirstinius, labai besiskiriančius (kokio nors projektavimo indekso prasme) nuo Gauso. Nors „įdomybių“ mintį gali būti ir sunku išreikšti, P. J. Huber [111] pateikė euristinį pasiūlymą Gauso skirstinį laikyti neįdomiausiu. Tai yra grindžiama tuo, kad:

- daugiamatis Gauso skirstinys yra visiškai apibrėžtas savo tiesinės struktūros (vidurkio ir kovariacijų matricos), o norima apčiuopti duomenų struktūrą, kuri nepriklausytų nuo koreliacinės duomenų struktūros ir tiesinių transformacijų, pvz., mastelio parametro;
- visos daugiamačio Gauso skirstinio projekcijos taip pat yra 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, daugelis projekcijų turės skirstinį, artimą normaliniam. Šis teiginys išplaukia iš centrinės ribinės teoremos;
- esant pastoviai dispersijai, Gauso skirstinys laikomas neinformatyviausiu.

Friedman išplėtojo Huber mintį ir pasiūlė algoritmą, vadinamą tiriamuoju tiksliniu projektavimu, skirtu daugiamačiam neparametriniam tankiui vertinti. PPDE procedūrą sudaro penkios pakopos:

1) duomenų standartizavimas: supaprastina išsidėstymą, mastelį bei koreliacines struktūras;
2) projektavimo indeksas: nustatomi įvairių krypčių „įdomumo“ laipsniai;
3) optimizavimo strategija: ieškoma tokia kryptis, kurioje projektavimo indeksas yra pats didžiausias;
4) duomenų transformavimas: pasirinktoje kryptyje apskaičiuojamas vienmatis tankis ir duomenys gausianizuojami;
5) tankio formavimas: daugiamatis tankis formuojamas iš apskaičiuotų vienmačių tankių, t. y. daugiamatis tankis yra vienmačių tankių funkcionalas.

Buvo pasiūlyta tokia projektavimo indekso konstrukcija. Žinoma, kad daugiamačio Gauso skirstinio visos projekcijos yra vienmačiai Gauso skirstiniai, taigi, jeigu nors viena kryptimi skirstinys yra ne Gauso, tai ir daugiamatis skirstinys taip pat nėra Gauso. Vadinasi, projektavimo indeksas $I(\tau)$ parodo, kiek vienmatis tankis $f_\tau(y)$ kryptimi τ ($Y = \tau' Z$) yra nutolęs nuo Gauso skirstinio, kai Z yra standartizuotas dydis [89]:

$$\widetilde{I}(\tau) = \int_{-\infty}^{\infty} (f_\tau(y) - \varphi(y))^2 dy \text{, čia } \varphi(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}. \qquad (1.81)$$

Projektavimo kryptis τ, maksimizuojanti $\widetilde{I}(\tau)$, projektuojant skirstinį, parenkama taip, kad išryškėtų to skirstinio daugiamodalinė ar kita netiesinė struktūra. Jeigu transformuojame duomenis *y* pagal lygybę

$$R = 2\Phi(Y) - 1 = 2\Phi(\tau' Z) - 1\,,\ R \in [-1,\ 1], \tag{1.82}$$

kur $\Phi(u)$ yra standartinio normalinio skirstinio funkcija:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{\frac{-t^2}{2}}\, dt\,, \tag{1.83}$$

o transformuoto dydžio *R* pasiskirstymo tankis

$$f_R(r) = \frac{f_\tau(y)}{\left|\frac{\partial r}{\partial y}\right|} = \frac{f_\tau(y)}{2\varphi(y)}, \tag{1.84}$$

tai, (1.81) lygybę galima perrašyti kintamąjį *y* pakeičiant į *r*:

$$\widetilde{I}(\tau) = \int_{-1}^{1} 2\varphi(y)(f_R(r) - 1/2)^2\, dr = \int_{-1}^{1} 2\varphi(\Phi^{-1}(\tfrac{R+1}{2}))(f_R(r) - 1/2)^2\, dr\,. \tag{1.85}$$

Friedman [73] pasiūlė šiek tiek kitokią projektavimo indekso $I(\tau)$ formą, *R* netolygumo matu laikant integruotą kvadratinę paklaidą:

$$I(\tau) = \int_{-1}^{1} (f_R(r) - 1/2)^2\, dr = \int_{-1}^{1} f_R^2(r) dr - 1/2\,. \tag{1.86}$$

Pažymėtina, kad jeigu *Y* skirstinys yra Gauso, tai $f_R(r) = 1/2$, $\forall r$, o projektavimo indeksas $I(\tau)$ lygus nuliui. Kuo daugiau *Y* skirstinys skiriasi nuo normalinio, tuo didesnė indekso $I(\tau)$ reikšmė. Kadangi $R \in [-1, 1]$, tai $f_R(r)$ gali būti išskleistas ortogonaliaisiais Ležandro daugianariais $\{\psi_j\}_{j=0}^{\infty}$, t. y. $f_R(r) = \sum_{j=0}^{\infty} b_j \psi_j(r)$:

$$I(\tau) = \int_{-1}^{1} f_R^2(r) dr - 1/2 = \int_{-1}^{1} \left[\sum_{j=0}^{\infty} b_j \psi_j(r)\right] f_R(r) dr - 1/2\,. \tag{1.87}$$

Ortogonalieji Ležandro daugianariai apibrėžiami iteracine išraiška:

$$\psi_0(r) = 1, \quad \psi_1(r) = r,$$
$$\psi_j(r) = \frac{(2j-1)r\psi_{j-1}(r) - (j-1)\psi_{j-2}(r)}{j}, \text{ kai } j \geq 2. \tag{1.88}$$

Iš ortogonalumo savybės išplaukia, jog koeficientai $\{b_j\}$ gali būti apskaičiuojami taip:

$$b_j = \frac{2j+1}{2}\int_{-1}^{1}\psi_j(r)f_R(r)dr = \frac{2j+1}{2}\mathbf{E}_R[\psi_j(r)] = \frac{2j+1}{2}\frac{1}{n}\sum_{t=1}^{n}\psi_j(2\Phi(Y(t))-1), \tag{1.89}$$

čia $\int_{-1}^{1}\psi_j(r)f_R(r)dr = \mathbf{E}_R[\psi_j(r)]$ išraiška aproksimuojama imties vidurkiu. Taigi, (1.86) lygybę galima užrašyti taip:

$$I(\tau) = \int_{-1}^{1} f_R^2(r)dr - 1/2 = \sum_{j=1}^{s}\frac{2j+1}{2}\mathbf{E}_R^2[\psi_j(r)]. \tag{1.90}$$

Pažymėtina, kad begalinė suma pakeista baigtine. Toks pakeitimas turi pranašumų: suma yra greičiau skaičiuojama, be to, tai suteikia projektavimo indeksui robastiškumo, nes sumuojant tik baigtinį skaičių narių, lėtai gęstančios projekcijų skirstinių „uodegos“ mažiau turi įtakos projektavimo indekso reikšmei. Siūloma parinkti $4 \leq s \leq 7$.

„Įdomioms“ projekcijoms ieškoti siūloma daug metodų. Geri taikymo rezultatai gaunami atsitiktinio starto principu, kai vienetinėje hipersferoje generuojamas didelis skaičius krypčių, o projektavimo kryptys po kiekvienos *k*-osios transformacijos parenkamos nuosekliai perrenkant:

$$\tau = \arg\max_{\tau}\{I(\tau)\}, \ \tau'\tau = 1. \tag{1.91}$$

Kitas siūlomas metodas, ieškant „įdomiausios“ projektavimo krypties, yra *mišri optimizavimo strategija* [73, 77, 83]. Apibrėžus projektavimo indekso analitinę išraišką, jos gradientas projektavimo kryptimi τ gaunamas toks:

$$\frac{\partial I}{\partial \tau} = \frac{2}{\sqrt{2\pi}} \sum_{j=1}^{s} (2j+1)\mathbf{E}[\psi_j(r)]\mathbf{E}[\psi_j'(r)e^{-y^2/2}(z-\tau y)], \tag{1.92}$$

čia Ležandro daugianario išvestinė lengvai apskaičiuojama pagal iteracinę formulę:

$$\begin{aligned} \psi_1'(r) &= 1, \text{ o} \\ \psi_j'(r) &= r\psi_{j-1}'(r) + j\psi_{j-1}(r), \text{ kai } j \geq 1. \end{aligned} \tag{1.93}$$

Pradžioje randamas apytikris pažingsninis optimizatorius atliekant paiešką pagrindinių komponenčių bei jų kombinacijų kryptyse taip, kad galėtų būti greitai pasiektas pradinis konvergavimas į maksimumą. Apytikris pažingsninis optimizatorius (stačiausias pakilimas) greitai parenka projekcijas, reikalingas pakilti į (lokalųjį) projektavimo indekso maksimumą.

Naudojant projektavimo indeksą (1.90), ieškoma „įdomių“ duomenų projekcijų. Tačiau dažniausiai nepakanka rasti vienos projekcijos, kad gana tiksliai būtų įvertintas daugiamatis tankis. Bendruoju atveju „įdomios“ kryptys nebūtinai turi būti ortogonalios ir gali tekti naudoti daugiau projektavimo krypčių nei yra duomenų dimensija. Todėl, vertinant tankį tikslinio projektavimo būdu, taikomas vadinamasis duomenų struktūros panaikinimas. Jis atlieka netiesinę mastelio pakeitimo transformaciją, rastą projektavimo kryptimi taip, kad transformuotų duomenų skirstinys šia kryptimi tampa normalinis, t. y. „neįdomus“. Tai užtikrina, kad, ieškant kitos projektavimo krypties, nebūtų rasta ta pati kryptis kaip ir prieš tai.

Duomenų struktūros panaikinimas remiasi tuo, kad jeigu vienamatė duomenų projekcija $\tau' Z$ turi pasiskirstymo tankį $f_\tau(y)$ ir atitinkamą skirstinio funkciją F_τ, tuomet atsitiktinis dydis

$$\widetilde{Y} = \Phi^{-1}(F_\tau(Y)), \tag{1.94}$$

čia Φ^{-1} yra pateiktai (1.83) lygybei atvirkštinė standartinio normalinio skirstinio funkcija. Friedman [73] pasiūlė empirinį skirstinio funkcijos įvertį skaičiuoti taip: $\hat{F}_\tau(y) = rank(Y)/n - \frac{1}{2n}$, čia $rank(y)$ yra Y rangas visoje n dydžio imtyje. Deja, šis įvertis nėra tikslus ir dažnai gaunama labai netolygi tankio funkcija. Pažymėję $Z^{(0)}=Z$, aptarsime, kaip iš $Z^{(k-1)}$ gaunamas $Z^{(k)}$. Remiantis (1.94) lygybe apibrėžiama

$$Z^{(k)} = Z^{(k-1)} + [\Phi^{-1}(F_\tau(\tau' Z^{(k-1)})) - \tau' Z^{(k-1)}]\tau. \tag{1.95}$$

Tokia pat „įdomiausios“ projekcijos ieškojimo procedūra atliekama su $Z^{(k)}$ ir ieškoma nauja kryptis. Ši veiksmų seka kartojama tol, kol daugiamatis skirstinys tampa artimas Gauso skirstiniui visomis kryptimis. Buvo pastebėta [73], kad gausianizavimas viena krytimi sutrikdo normalumą anksčiau nagrinėtomis kryptimis, taigi jų projektavimo indeksas $I(\tau)$ jau nebebūna lygus nuliui. Tačiau tyrimai rodo [72], kad atsiradę pokyčiai yra labai nedideli.

Daugiamatis tankis apskaičiuojamas iš vienmačių tankių įverčių. Sąsaja tarp daugiamačių $Z^{(k)}$ ir $Z^{(k\text{-}1)}$ tankių (čia $Z^{(k)}$ yra nutolusių duomenų $Z^{(k\text{-}1)}$ struktūra išilgai k-osios projekcijos $\tau(k)$) yra:

$$f_{\tau(k)}(z^{(k)}) = \frac{f_{\tau(k-1)}(z^{(k-1)})}{|J_k(z^{(k-1)})|}$$
$$f_{\tau(k-1)}(z^{(k-1)}) = f_{\tau(k)}(z^{(k)})\,|J_k(z^{(k-1)})|, \tag{1.96}$$

čia jakobianas

$$J_k(z^{(k-1)}) = \frac{\partial z^{(k)}}{\partial z^{(k-1)}} = \frac{\partial (Uz^{(k)})}{\partial (Uz^{(k-1)})} = \frac{\partial y^{(k)}}{\partial y^{(k-1)}} = \frac{f_{\tau(k)}(y^{(k-1)})}{\varphi(y^{(k)})} = \frac{f_{\tau(k)}(\tau'(k)z^{(k-1)})}{\varphi(\tau'(k)z^{(k)})} \geq 0. \tag{1.97}$$

Pradedant nuo daugiamačių pradinių duomenų $Z^{(0)}$, gausianizavimo procedūra atliekama kiekvienai „įdomiai“ projekcijai, rastai pagal $I(\tau)$. Po tam tikro skaičiaus, tarkime, po M, projekcijų daugiamačiai duomenys $Z^{(M)}$ nedaug tesiskiria nuo normalinio skirstinio, t. y. $f_{\tau(M)}(z^{(M)}) \approx \varphi(z^{(M)})$. $Z^{(0)}$ tankis gali būti apskaičiuotas taip:

$$\begin{aligned} f(z^{(0)}) &= f_{\tau(1)}(z^{(1)})J_1(z^{(0)}) = f_{\tau(2)}(z^{(2)})J_2(z^{(1)})J_1(z^{(0)}) = f_{\tau(M)}(z^{(M)})\prod_{k=1}^{M} J_k(z^{(k-1)}) \\ &\approx \varphi(z^{(M)})\prod_{k=1}^{M} J_k(z^{(k-1)}) = \varphi(z^{(M)})\prod_{k=1}^{M} \frac{f_{\tau(k)}(\tau'(k)z^{(k-1)})}{\varphi(\tau'(k)z^{(k)})}. \end{aligned} \tag{1.98}$$

Vienmatis projektuotų duomenų tankis $f_{\tau(k)}(\tau'(k)z^{(k\text{-}1)})$ apskaičiuojamas pagal (1.84) lygybę, t. y. $f_\tau = 2\varphi(y)f_R(r)$, arba konkrečiau:

$$f_{\tau(k)}(\tau'(k)z^{(k-1)}) = \varphi(\tau'(k)z^{(k-1)})\sum_{j=0}^{s}\frac{2j+1}{n}\sum_{t=1}^{n}\psi_j(r_t^{(k-1)})\psi_j(r^{(k-1)}). \tag{1.99}$$

Tuomet (1.98) dešinėje pusėje nežinomus vienmačius pasiskirstymo tankius pakeitus jų statistiniais įverčiais, gaunama:

$$\hat{f}(z) = \varphi(z^{(M)}) \prod_{k=1}^{M} \frac{\hat{f}_{\tau(k)}(\tau'(k) z^{(k-1)})}{\varphi(\tau'(k) z^{(k)})}. \tag{1.100}$$

Dėl daugianarinės projektavimo indekso formos ir iteratyvaus sąryšio tarp daugianarių PPDE apskaičiuojamas gana greitai.

1.3. Daugiamačio pasiskirstymo tankio įvertinių tikslumo tyrimas

Įvertinių tyrimo metodika. Tankio įvertinius tyrėme Monte Karlo metodu. Kadangi tikrasis imties tankis buvo žinomas, tai skaičiavome įvertinių paklaidas ir, jas lygindami, formulavome išvadas apie įvertinių kokybę. Šiame skyriuje konkretizuosime tyrime vartotas atsitiktines imtis, tirtus įvertinius ir jų paklaidas.

Atsitiktinės imtys. Tyrimui naudojome generuotas imtis, pasiskirsčiusias pagal Gauso ir Koši skirstinių mišinių modelius. Norint įvairiapusiškai ištirti siūlomus metodus, buvo varijuojamas mišinių komponentų skaičius, jų tikimybės, atstumai tarp komponentų centrų. Imties parametrams nurodyti vartosime sutrumpintą žymėjimą, pvz., $n = 400$, $p_1 = 0{,}65$, $m_1 = [0{,}0;\ 0{,}0]'$, $u_1 = [0{,}42;\ 0{,}51]'$, $p_2 = 0{,}35$, $m_2 = [2{,}0;\ 2{,}0]'$, $u_2 = [0{,}33;\ 0{,}46]'$ reikš imtį, sudarytą iš 400 stebinių, kurių tankis yra tam tikrų skirstinių mišinio ($q = 2$) tankis su atitinkamais komponentų svorio, padėties ir mastelio arba sklaidos parametrais.

Paklaidos. Tiriamųjų įvertinių tikslumą matavome naudodami kelias paklaidas, pasižyminčias skirtingomis savybėmis. Daugiausia dėmesio skyrėme paklaidoms

$$\delta_1 = \frac{1}{n} \sum_{t=1}^{n} \left| f(X(t)) - \hat{f}(X(t)) \right| \cong \int \left| f(x) - \hat{f}(x) \right| f(x) dx \tag{1.101}$$

ir

$$\delta_2 = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{f(X(t)) - \hat{f}(X(t))}{f(X(t)) + \hat{f}(X(t))} \right| \cong \frac{1}{2} \int \left| f(x) - \hat{f}(x) \right| dx\,. \tag{1.102}$$

Buvo skaičiuojami šių paklaidų matematinių vidurkių empiriniai analogai. Kiekvienu atveju apskaičiuoti paklaidų δ_i aritmetiniai vidurkiai $\bar{\delta}_i$ bei standartiniai nuokrypiai, gauti sugeneravus 100 nepriklausomų imčių. Juos ir pateiksime lentelėse bei grafikuose.

Tankių vertinimo parametrų parinkimo metodai. Taikant daugumą populiarių neparametrinio vertinimo procedūrų, praktikoje susiduriama su jų parametrų optimalaus parinkimo problema. Branduolinių įvertinių konstrukcijos svarbiausias elementas yra glodinimo plotis, splaininiams įverčiams nelengva parinkti mazgus ir t. t.

Parametrų parinkimas mažiausių kvadratų kryžminio patikrinimo būdu. Kryžminio patikrinimo metodas (angl. *cross-validation*) pirmą kartą buvo pasiūlytas Kurtz (1948) ir tikslintas Mosier (1951) bei plėtotas Krus ir Fuller (1982) [141, 157, 140]. Metodas pagrįstas idėja, jog statistika, apskaičiuota naudojantis vienais imties elementais, yra tikrinama naudojant kitus imties elementus [122, 175]. Vienas iš populiariausių kryžminio parinkimo metodų, vertinant tankį, yra mažiausių kvadratų kryžminio patikrinimo metodas (paskelbtas Breiman (1984) ir Burman (1989) [27, 29], kurio idėjos nagrinėtos ankstesniuose Stone darbuose (1974 ir 1977) [207, 208]). Įverčio parametro ieškoma tokio, kad šis minimizuotų integruotą kvadratinę paklaidą:

$$\begin{aligned}\theta &= \arg\min_{\theta} \int_{-\infty}^{\infty} \left(\hat{f}_{\theta}(x) - f(x)\right)^2 dx \\ &= \arg\min_{\theta} \left(\left\| \hat{f}_{\theta}(x) \right\|_2^2 - 2\int_{-\infty}^{\infty} \hat{f}_{\theta}(x) f(x) dx + \left\| f(x) \right\|_2^2 \right) \\ &= \arg\min_{\theta} \left(\left\| \hat{f}_{\theta}(x) \right\|_2^2 - 2\int_{-\infty}^{\infty} \hat{f}_{\theta}(x) dF \right),\end{aligned} \tag{1.103}$$

čia θ – vertinamas parametras, o $F(x)$ – stebimo atsitiktinio dydžio skirstinio funkcija. Keičiant nežinomą skirstinio funkciją į empirinę skirstinio funkciją, gaunama parametro įverčio išraiška:

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta} \left(\left\| \hat{f}_{\theta}(x) \right\|_2^2 - 2\int_{-\infty}^{\infty} \hat{f}_{\theta}(x) d\hat{F} \right) \\ &= \arg\min_{\theta} \left(\left\| \hat{f}_{\theta}(x) \right\|_2^2 - \frac{2}{n}\sum_{t=1}^{n} \hat{f}_{\theta}(X(t)) \right).\end{aligned} \tag{1.104}$$

Pažymėtina, kad vietoj (1.104) patogiau naudoti formulę

$$\hat{\theta} = \arg\min_{\theta} \left(\left\| \hat{f}_{\theta}(x) \right\|_2^2 - \frac{2}{n}\sum_{t=1}^{n} \hat{f}_{\theta}(X(t) \mid t) \right), \tag{1.105}$$

čia $\hat{f}_\theta(x \mid t)$ yra įverčio reikšmė taške x, kuri apskaičiuojama pašalinus iš stebinių reikšmę $X(t)$. Be to, empiriniai tyrimai rodo, kad taikant kryžminio patikrinimo metodą, geriau ieškoti ne globaliojo minimumo, o didžiausio lokaliojo minimumo taško [39, 53, 88, 126].

Parametrų parinkimas tikėtinumo kryžminio patikrinimo būdu. Nauji tikėtinumo kryžminio patikrinimo metodų tyrimai skelbiami Smyth (2000) bei Pavlic ir van der Laan (2003) straipsniuose [201] ir [164]. Formalizuojant tikėtinumo kryžminio patikrinimo metodą, naudojamas binarinis atsitiktinis vektorius $S_n \in \{0,1\}^n$, nepriklausantis nuo $\hat{F}$. S_n apibrėžia savotišką n stebinių imties padalijimą į mokomąją imtį $\{t \in \{1,\ldots,n\}: S_n(t)=0\}$ ir tikrinamąją imtį $\{t \in \{1, \ldots, n\}: \ S_n(t)=1\}$. Tarkim, $\hat{F}^1_{S_n}$, $\hat{F}^0_{S_n}$ yra empirinės atitinkamai tikrinamosios ir mokomosios imčių skirstinių funkcijos, o stebinių santykinis skaičius $p = \sum_{t=1}^{n} S_n(t)/n \in (0,1)$ tikrinamojoje imtyje yra pastovus. Tikėtinumo kryžminio patikrinimo kriterijus apibrėžiamas taip:

$$\hat{\theta}_{n(1-p)}(k) = -\mathbf{E}_{S_n} \int \log\left(\hat{f}_k(x \mid \hat{F}^0_{S_n})\right) d\hat{F}^1_{S_n}(x). \tag{1.106}$$

Pagal šį kriterijų randamas optimalus $\hat{k}$:

$$\hat{k} = \arg \min_{k \in \{1,\ldots,K(n)\}} \hat{\theta}_{n(1-p)}(k). \tag{1.107}$$

Pažymėtina, kad skirtingai parenkant atsitiktinius dydžius $S_n(t)$ aprėpiama daug kryžminio patikrinimo tipų: tokius kaip V sulenkimo (angl. *V-fold*) kryžminis patikrinimas, Monte Karlo (kartojami atsitiktiniai padalijimai) kryžminis patikrinimas ir imčių kartojimo (*bootstrap*) kryžminis patikrinimas. Pastarasis atitinka n didumo gražinamųjų imčių pakartojimus iš generalinės duomenų aibės ir vektorių $S_{n,t}$, kuris lygus t-ojo stebinio paėmimo kartų skaičiui. Šiuo atveju $\hat{F}^0_{S_n}$, $\hat{F}^1_{S_n}$ žymi atitinkamai stebinių pakartotų imčių ir jose išskirtų stebinių empirinius skirstinius.

Pasirinkto $\hat{k}$ atskaitos taškas apibrėžiamas lygtimi

$$\tilde{\theta}_{n(1-p)}(k) = -\mathbf{E}_{S_n} \int \log\left(\hat{f}_k(x \mid \hat{F}^0_{S_n})\right) dF(x) \tag{1.108}$$

ir yra minimizuojamas:

$$\widetilde{k} = \arg \min_{k \in \{1,\dots,K(n)\}} \widetilde{\theta}_{n(1-p)}(k). \tag{1.109}$$

Pažymėtina, kad $\widetilde{k}$ yra parinktas optimaliai, kai atitinka tokį *k*, kuriam esant Kulbako ir Leiblerio atstumo tarp vertinamo tankio *f*(*x*) ir mokomąja imtimi paremto jo įverčio $\hat{f}_k(\cdot \mid \hat{F}^0_{S_n})$ vidurkis pagal S_n yra mažiausias:

$$k \to \mathbf{E}_{S_n} \int \log\left(\frac{f(x)}{\hat{f}_k(x \mid \hat{F}^0_{S_n})} \right) dF(x). \tag{1.110}$$

Asimtotiškai kryžminio patikrinimo $\hat{k}$ parenkamas taip pat gerai kaip ir $\widetilde{k}$ ta prasme, kad sąlyginio Kulbako ir Leiblerio atstumo vidurkių santykis $(\mathbf{E}\widetilde{\theta}_{n(1-p)}(\hat{k}) - \theta^*)/(\mathbf{E}\widetilde{\theta}_{n(1-p)}(\widetilde{k}) - \theta^*)$ konverguoja į vienetą. Helingerio atstumas tarp funkcijų $f_{\hat{k}}(\cdot \mid \hat{F}_{n(1-p)})$ ir *f* kartu su atitinkamu Kulbako ir Leiblerio atstumo apribojimu yra:

$$\int \left(\sqrt{f(x)} - \sqrt{\hat{f}(x)} \right)^2 d\mu(x) \le \int \log\left(\frac{f(x)}{\hat{f}(x)} \right) f(x) d\mu(x). \tag{1.111}$$

Galiausiai tarp visų $\hat{f}$, tokių, kad $\hat{f} \to -\int \log(\hat{f}(x)) dF(x)$, apibrėžiamas minimumas:

$$\theta^* = -\int \log(f(x)) dF(x). \tag{1.112}$$

Pažymėsime, kad $\widetilde{\theta}_{n(1-p)}(\hat{k}) \ge \widetilde{\theta}_{n(1-p)}(\widetilde{k}) \ge \theta^*$.

Įvairias kryžminio patikrinimo būdų kompiuterines realizacijas galima rasti populiariuose statistiniuose paketuose, pavyzdžiui, [233, 234].

Parametrų įverčių „įdėties“ principas. Parametrų įverčių „įdėties“ (angl. *plug-in*) principas pagrįstas tuo, kad nežinomi dydžiai išraiškose keičiami jų statistiniais įverčiais [223]. Šis principas išpopuliarėjo pradėjus naudoti kompiuterinę įrangą. Dažnai jis būna paprastas, nes nereikia sudėtingos matematinės analizės.

Parametrų parinkimas pasiskirstymo tankių procedūrose. Šiame tyrime Monte Karlo metodu buvo siekiama atlikti anksčiau 1.1 ir 1.2 skyriuose aprašytų pasiskirstymo tankio neparametrinių įvertinių (adaptuoto branduolinio, pusiau parametrinio branduolinio,

histosplaininio, logsplaininio, apvertimo formulės taikymo, kuris remiasi tankio aproksimacija Gauso skirstinių mišinio modeliu, apvertimo formulės modifikuoto ir tikslinio projektavimo) tikslumo lyginamąją analizę. Taikant adaptuotą branduolinį metodą naudojamo jautrumo parametro reikšmę autoriai [113] siūlo rinkti iš aibės {0,2; 0,4; 0,6; 0,8}, o konkreti parametro reikšmė nustatoma tikėtinumo kryžminio patikrinimo būdu [142, 143]. Taikant pusiau parametrinį branduolinį metodą, buvo perrenkamos visos galimos subvektoriaus Y dimensijos s reikšmės ir jas atitinkančios koordinatės, o rezultatams palyginti su kitais tirtais metodais naudotos tiksliausios paklaidos. Histosplaininiu metodu parenkant histogramos vidurio taškų tinklelio suskaidymą į subtinklelius galima naudoti konstruktyvią programinės įrangos SAS procedūrą, kuri remiasi neparametrinės regresijos modeliu [183]; ji ir buvo naudojama darbe. Taikant logsplaininį metodą, parenkant bazinio splaino taškų skaičių, minimizuojamas Akaike informacijos kriterijus [137]. Šio įverčio apskaičiavimo kompiuterinė programa pateikta [232], ja ir buvo naudotasi tyrime. Apvertimo formule pagrįstas metodas ir jo modifikacija turi glodinimo parametrą h. Modeliavimo tyrimai parodė, kad šis metodas yra jautrus parametro parinkimui: parinkus per mažą h reikšmę, įvertinys tampa labai neglodus ir turi dideles paklaidas. Per daug suglodinus tankio įvertinį, labai nenukenčia jo kokybė. Atliekant tyrimus, buvo pastebėta, kad įvertinys tampa neglodus dėl to, kad kai kuriose kryptyse projektuotų stebinių reikšmės yra panašios, todėl išskiriami mažo svorio komponentai su nedidelėmis dispersijomis. Glodinimo parametras, taip pat triukšmo klasterio svorio (tikimybės) konkreti reikšmė iš aibės {0,05; 0,1; 0,15; 0,2; 0,3; 0,4} parenkami mažiausių kvadratų kryžminio patikrinimo būdu [197]. Taikant tikslinio projektavimo metodą ir remiantis straipsnio [113] rekomendacija, skleidinio (1.99) eilė $4 \le s \le 6$, o projektavimo kryptys buvo parenkamos maksimizuojant J. H. Friedman rekomenduoto projektavimo indekso (1.90) įverčio reikšmę.

Pastaba. Čia aprašyti parametrų parinkimai pasiskirstymo tankių procedūrose buvo naudoti ir vėlesniuose skyriuose pateiktiems sudėtiniams algoritmams tirti.

Tirti modeliai. Daugiausia dėmesio buvo skiriama atvejui, kai nepriklausomų d-mačių stebinių skirstinys yra Koši skirstinių mišinys. Daugiamatis Koši skirstinio tankis buvo aprašomas kaip sandauga vienamačių Koši skirstinio tankių su standartizuotais mastelio parametrais. Taigi,

$$f(x) = \sum_{i=1}^{q} p_i f_C(x, m_i),$$

čia

$$f_C(x, m_i) = \prod_{j=1}^{d} \frac{1}{\pi(1 + (x_j - m_{ij})^2)}.$$

Palyginimui buvo generuotos ir imtys, atitinkančios Gauso skirstinį su vienetine kovariacine mišinių matrica. Šiuo atveju

$$f(x) = \sum_{i=1}^{q} p_i f_N(x, m_i),$$

kur

$$f_N(x, m_i) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x_j - m_{ij})^2 / 2\right\}.$$

Nagrinėti šie modeliai:

A1: $q = 2$, $p_1 = (1 - p_2)$, $p_2 = 0{,}1, 0{,}3, 0{,}5$, $m_1 = (0; 0; 0; 0; 0)'$; $m_2 = (0{,}5k; 0{,}5k; 0{,}5k; 0{,}5k; 0{,}5k)'$ (čia ir toliau k perbėga reikšmes 1, 2, ..., 6).

A2: $q = 3$, $p_1 = p_2 = (1 - p_3)/2$, $p_3 = 0{,}1, 1/3, 0{,}8$, $m_1 = (0; 0; 0; 0; 0)'$, $m_2 = (0{,}5k; 0{,}5k; 0{,}5k; 0{,}5k; 0{,}5k)'$, $m_3 = (0{,}5k; 0{,}5k; 0; 0; 0)'$.

A3: $q = 4$, $p_1 = p_2 = p_3 = (1 - p_4)/3$, $p_4 = 0{,}1, 0{,}25, 0{,}7$, $m_1 = (0; 0; 0; 0; 0)'$, $m_2 = (0{,}5k; 0{,}5k; 0{,}5k; 0{,}5k; 0{,}5k)'$, $m_3 = (0{,}5k; 0{,}5k; 0; 0; 0)'$, $m_4 = (0; 0; 0{,}5k; 0{,}5k; 0{,}5k)'$.

A4: $q = 2$, $p_1 = (1 - p_2)$, $p_2 = 0{,}1, 0{,}3, 0{,}5$, $m_1 = (0; 0)'$, $m_2 = (0{,}5k; 0{,}5k)'$.

A5: $q = 3$, $p_1 = p_2 = (1 - p_3)/2$, $p_3 = 0{,}1, 1/3, 0{,}8$, $m_1 = (0; 0)'$, $m_2 = (0{,}5k; 0{,}5k)'$, $m_3 = (0{,}5k; 0)'$.

A6: $q = 4$, $p_1 = p_2 = p_3 = (1 - p_4)/3$, $p_4 = 0{,}1, 0{,}25, 0{,}7$, $m_1 = (0; 0)'$, $m_2 = (0{,}5k; 0{,}5k)'$, $m_3 = (0{,}5k; 0)'$, $m_4 = (0; 0{,}5k)'$.

Kiekvieno tipo duomenims generuotos įvairaus didumo imtys (50, 100, 200, 400, 800).

1.4. Pasiskirstymo tankio įvertinių tikslumo tyrimo rezultatai

Tyrimo rezultatai parodė, jog imtyje esant labai didelėms išskirtims logsplaininiu metodu įvertintas tankis duoda dideles paklaidas, ir tokiais atvejais šį metodą naudoti nėra rekomenduojama. Histosplaininis tankių vertinimo metodas gali konkuruoti su kitais tankių vertinimo metodais, kai duomenų dimensija yra nedidelė. Kai duomenų dimensija yra didesnė, šiuo metodu rezultatai gaunami prastesni nei tiriant kitais metodais [5A, 6A].

Skaičiuojant δ_1 paklaidą penkiamačiu atveju:

- kai $n = 50$, geriausi rezultatai gauti adaptuotu branduoliniu, apvertimo formulės modifikuotu ir apvertimo formulės taikymo metodais;
- kai $n = 100$, geriausi rezultatai gauti apvertimo formulės modifikuotu, adaptuotu branduoliniu ir tikslinio projektavimo metodais;
- kai $q = 2$, $n \geq 200$, geriausi rezultatai gauti pusiau parametriniu branduoliniu ir apvertimo formulės modifikuotu metodais;
- kai $q \geq 3$, $n = 200$, labai persidengiančių skirstinių atvejais ($k = 1, 2$) geriausi rezultatai gauti pusiau parametriniu branduoliniu, o labiau atsiskyrusių ($k \geq 3$) – apvertimo formulės modifikuotu metodu;
- kai $q = 3$, $n \geq 400$, geriausi rezultatai gauti pusiau parametriniu branduoliniu ir apvertimo formulės modifikuotu metodais;
- kai $q = 4$, $n = 400$, labai persidengiančių skirstinių atvejais ($k \leq 3$) geriausi rezultatai gauti pusiau parametriniu branduoliniu, o labiau atsiskyrusių ($k \geq 4$) – apvertimo formulės modifikuotu metodu;
- kai $q = 4$, $n \geq 400$, geriausi rezultatai gauti pusiau parametriniu branduoliniu ir apvertimo formulės modifikuotu metodais.

Skaičiuojant penkiamačių stebinių δ_2 paklaidą, labai ir mažai persidengiančių skirstinių atvejais ($k \leq 4$) geriausi rezultatai gauti pusiau parametriniu branduoliniu, o atsiskyrusių skirstinių – apvertimo formulės modifikuotu metodu (1.1 lentelė, skliausteliuose pateiktos paklaidų standartinių nuokrypių reikšmės).

Skaičiuojant δ_1 ir δ_2 paklaidas dvimačiu, labai ir mažai persidengiančių skirstinių ($k \leq 4$) atveju geriausi rezultatai gauti pusiau parametriniu branduoliniu, o atsiskyrusių skirstinių atveju – adaptuotu branduoliniu metodu.

δ_1 paklaidos atveju, kai mišinius sudaro labai persidengiantys skirstiniai, gauti rezultatai yra blogesni nei tada, kai skirstiniai yra atsiskyrę.

1.1 lentelė. δ_2 paklaidos priklausomybė nuo atstumo tarp komponentų centrų

Vertinimo metodai	Tankiai					
	$d = 5; p_1 = p_2 = p_3 = 1/3; n = 100$					
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
AKDE	0,8268	0,8257	0,8197	0,8128	0,8066	0,8075
	(0,076)	(0,0814)	(0,0848)	(0,0827)	(0,0788)	(0,0731)
PPDE	0,9243	0,9319	0,9303	0,93	0,9284	0,925
	(0,0501)	(0,0364)	(0,0375)	(0,0387)	(0,041)	(0,0433)
LSDE	0,8043	0,8162	0,8583	0,8611	0,8613	0,8711
	(0,0534)	(0,054)	(0,0491)	(0,0349)	(0,0434)	(0,0577)
SKDE	0,7158	0,7144	0,7088	0,7071	0,7179	0,7227
	(0,026)	(0,0905)	(0,0905)	(0,083)	(0,0631)	(0,0498)
IFDE	0,9459	0,8886	0,7857	0,8463	0,8761	0,8312
	(0,0362)	(0,1318)	(0,0706)	(0,038)	(0,111)	(0,0538)
MIDE	0,7389	0,7332	0,7235	0,7149	0,7121	0,7219
	(0,0279)	(0,0221)	(0,0338)	(0,0195)	(0,0208)	(0,0203)

δ_2 paklaidos atveju geresni rezultatai gauti tada, kai skirstinių centrai yra nutolę vidutiniškai ($k = 3, 4, 5$).

2. Duomenų pirminio klasterizavimo poveikis daugiamodalinių tankių statistinio vertinimo tikslumui

2.1. Klasterizavimo metodai

Klasterinė analizė – tai matematinių metodų visuma, skirta objektų ar reikšmių aibėms suskaidyti į tam tikra prasme vienalytes, homogenines grupes (klases, klasterius) taip, kad tos pačios grupės elementai būtų „artimi“ vienas kitam, o elementai iš skirtingų grupių – „tolimi“ vienas kitam. Toks uždavinys iškyla sprendžiant įvairias taikomąsias problemas: įvairių preparatų poveikio efektų tyrimas medicinoje, ekonomikos subjektų įvairių plėtros tendencijų išskyrimas, anketų ir respondentų tipologinė analizė sociologijoje, dangaus kūnų segmentų išskyrimas astronomijoje, pavyzdžiui [15, 200, 216].

Dauguma klasterinės analizės (klasterizavimo be mokymo) darbų pasirodė per pastaruosius keturis dešimtmečius, nors pirmieji darbai, kuriuose buvo paminėti klasterinės analizės metodai, pasirodė gana seniai. 1911 metais lenkų antropologas J. Czekanowicki iškėlė „struktūrinės klasifikacijos“ idėją, kuri realizuoja pagrindinį klasterinės analizės principą – išskirti objektų kompaktines grupes. Terminą „klasterinė analizė“ (angl. *cluster analysis*) 1939 metais pirmasis pasiūlė anglų mokslininkas R. C. Tryon. Žodis „cluster“ iš anglų kalbos gali būti verčiamas kaip „kekė“, „sankaupa“. Labai didelę įtaką klasterinei analizei padarė dviejų biologų R. R. Sokal ir P. H. Sneath knyga, kurioje tvirtinama, jog objektų suskaidymo į grupes struktūros nustatymas padeda atkurti šių struktūrų susikūrimo procesą, o skirtingų grupių (klasterių) objektų skirtumai ir panašumai gali būti evoliucinio proceso mechanizmo suvokimo pagrindas.

Klasterių sudarymo metodų yra daug. Jie skirstomi pagal tai, kokia yra skirstymo į klasterius strategija, kaip parenkami panašumo matai bei atstumo tarp klasterių nustatymo kriterijai. Pagal naudojamą metodiką klasterizavimo metodus sąlygiškai galima suskirstyti į „geometrinius“ ir „tikimybinius“, pagal klasterių svarbumą – į hierarchinius ir nehierarchinius [21, 25, 54, 177, 228, 4A].

Imties klasterizavimas naudojantis EM algoritmu. Jei atsitiktinio vektoriaus X pasiskirstymo tankis turi q maksimumų, tai jį galima bandyti aproksimuoti q vienamodalinių pasiskirstymo tankių mišiniu:

$$f(x) = \sum_{k=1}^{q} p_k f_k(x). \qquad (2.1)$$

Tarkime, kad X skirstinys priklauso nuo atsitiktinio dydžio v, kuris įgyja reikšmes $1, \ldots, q$ su atitinkamomis tikimybėmis $p_1, \ldots, p_q$. Klasifikavimo teorijoje v yra interpretuojamas kaip klasės, kuriai priklauso stebimas objektas, numeris. Taigi, stebinius $X(t)$ atitiktų $v(t)$, $t = 1, \ldots, n$. Funkcijos f_k traktuojamos kaip X sąlyginis pasiskirstymo tankis, esant sąlygai $v = k$. Remiantis tokiu požiūriu, imties negriežtas klasterizavimas suprantamas kaip *aposteriorinių* tikimybių

$$\pi_k(x) = \mathbf{P}\{v = k | X = x\} \tag{2.2}$$

vertinimas, kai visi $x \in \{X(1), \ldots, X(n)\}$. Griežtas imties klasterizavimas būtų atsitiktinių dydžių $v(1), \ldots, v(n)$ vertinimas, t. y. imtis suskaidoma į poaibius remiantis lygybe

$$\theta(t) = \arg\max_{k=1,\ldots,q} \hat{\pi}_k(X(t)). \tag{2.3}$$

Įverčiai $\hat{\pi}_k$ gaunami aproksimuojant nežinomus pasiskirstymo tankio komponentus φ_k normaliniais pasiskirstymo tankiais ir naudojant EM algoritmą. Jį trumpai aprašysime. Tarkime, galioja (2.1) lygybė ir f_k yra skirstinių $\mathcal{N}(M(k), R(k))$ tankio funkcijos, $k = 1, \ldots, q$. Šiuo atveju (2.1) lygybės dešiniąją pusę pažymėkime $f(x,\theta)$, čia $\theta = (p_k, M(k), R(k), k = 1, \ldots, q)$.

Galioja lygybės:

$$\pi_k(x) = \frac{p_k f_k(x)}{f(x,\theta)} \text{ ir } k = \overline{1,q}. \tag{2.4}$$

Turint θ įvertį, tikimybių π_k įverčiai gaunami iš (2.4) „*įdėties*“ metodu, t. y. keičiant dešinėje pusėje nežinomus parametrus jų statistiniais įverčiais. EM algoritmas yra rekurentinė procedūra, skirta θ maksimalaus tikėtinumo įverčiui

$$\theta^* = \arg\max_{\theta} L(\theta),\ L(\theta) = \prod_{t=1}^{n} f(X(t),\theta) \tag{2.5}$$

ir jį atitinkantiems įverčiams $\hat{\pi}_k$ apskaičiuoti. Šis algoritmas Gauso mišinio analizei buvo nepriklausomai pasiūlytas kelių autorių: Hasselbland (1966) [102], Behboodian (1970) [18]. Vėliau jo savybės buvo gerai išnagrinėtos [24, 41, 42, 214, 231] ir kituose darbuose. EM algoritmui skirta daug dėmesio įvairiuose apžvalginiuose straipsniuose ir monografijose [45, 58,

130, 152, 169, 213]. Tarkim, po r ciklų gavome įverčius $\hat{\pi}_k = \hat{\pi}_k^{(r)}$. Tada naujasis įvertis $\hat{\theta} = \hat{\theta}^{(r+1)}$ apibrėžiamas lygybėmis:

$$\hat{p}_k = \frac{1}{n}\sum_{t=1}^{n} \hat{\pi}_k(X(t)),$$

$$\hat{M}(k) = \frac{1}{n\hat{p}_k}\sum_{t=1}^{n} \hat{\pi}_k(X(t)) \cdot X(t),$$

$$\hat{R}(k) = \frac{1}{n\hat{p}_k}\sum_{t=1}^{n} \hat{\pi}_k(X(t))\left[X(t) - \hat{M}(k)\right]\cdot\left[X(t) - \hat{M}(k)\right]',$$

čia $k = 1, \ldots, q$. Įrašius $\hat{\theta}^{(r+1)}$ į (2.4) dešiniąją pusę, randami $\hat{\pi}^{(r+1)}(X(t))$, $k = \overline{1, q}$, $t = \overline{1, n}$. Šios rekurentinės procedūros išdavoje gauname nemažėjančią seką $L(\hat{\theta}^{(r)})$, tačiau ar ji konverguoja į globaliojo maksimumo tašką, labai priklauso nuo pradinio įverčio $\hat{\theta}^{(0)}$ (arba $\hat{\pi}^{(0)}$). Pradinės reikšmės parinkimas nėra galutinai išspręsta problema. Galimus sprendimo būdus galima suskirstyti į tokias grupes:

- Paprasčiausias pradinės reikšmės parinkimo problemos sprendimo būdas – taikyti *atsitiktinio starto* principą: EM algoritmas kartojamas daug kartų kiekvieną kartą pradinius įverčius $\hat{\pi}^{(0)}$ parenkant atsitiktinai. Galutinis rezultatas parenkamas taip, kad jį atitinkanti tikėtinumo funkcijos $L(\hat{\theta})$ reikšmė būtų maksimali.
- Geri rezultatai taip pat gauti taikant *nuoseklaus klasterių išskyrimo* procedūrą. Ši metodika skiriasi tuo, kad EM algoritmas nėra vykdomas nuo tam tikro pradinio taško, bet klasteriai palaipsniui išskiriami iš mišinio ir kiekvieną kartą, išskyrus naują klasterį, parametrai yra tikslinami EM algoritmu. Prie tokių metodų galima priskirti [172] aprašytą procedūrą, išskiriančią klasterius kaip sritis, turinčias didesnę neparametrinio tankio įverčio reikšmę. Kitas nuoseklaus klasterių išskyrimo metodas paskelbtas darbe [219].
- Gerai yra žinomi *hierarchiniai klasifikavimo metodai*, bet paprastai jie nėra pritaikyti Gauso mišinių modelį tenkinantiems duomenimis. Pirmąkart toks klasifikavimo algoritmas, išlaikantis Gauso mišinio modelį sudarant klasterių medį, buvo aprašytas Fraley 1998 metais [69]. Taigi, tai naujas ir gerai veikiantis algoritmas. Bene vienintelis iki šiol pastebėtas didelis jo trūkumas yra tas, kad ilgai trunka kompiuterio skaičiavimai ir reikia didelės kompiuterio atminties, jeigu duomenų yra daug.
- *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 tada, kai Gauso mišinys tenkina papildomas sąlygas, pvz., *k* vidurkių metodas tinka EM algoritmui inicializuoti kai klasterių kovariacinės matricos skiriasi nedaug.

Hierarchinis klasterizavimas. Nedideliam objektų (stebinių) skaičiui klasterizuoti dažniausiai naudojami hierarchiniai algoritmai. Hierarchija gali būti skaidančioji, kai jos formavimo procese pradedama nuo objektų visumos kaip vienos klasės, o baigiama tuo, kad kiekvienas objektas sudaro atskirą klasę. Atvirkštinis procesas, kai iš pradžių kiekvienas objektas sudaro atskirą klasę, o paskui sujungti galutiniame etape visi objektai sudaro vieną klasę, vadinamas jungiamąja hierarchija. Pažymėkime S_i – *i*-ąją objektų klasę (klasterį), n_i – objektų skaičių klasėje, $\rho(S_l, S_m)$ – atstumą tarp klasių S_l ir S_m. Stebinių klasifikacijai naudojant jungiamosios hierarchijos algoritmą pradinis skaidinys yra $S^{(0)} = \left(S_1^{(0)}, ..., S_n^{(0)}\right)$, čia $S_t^0 = \{X(t)\}$, *k*–lygio skaidinys $S^{(k)} = \left(S_1^{(k)}, ..., S_{n-k}^{(k)}\right)$, gaunamas iš $S^{(k-1)}$ skaidinio apjungus klasių porą (S_l^*, S_m^*):

$$(S_l^*, S_m^*) = \arg \min_{\substack{S_l \neq S_m \\ S_l, S_m \in S^{(k-1)}}} \rho(S_l, S_m). \tag{2.6}$$

Galutinę hierarchiją sudaro įkeltų skaidinių sistema $S^{(0)} \subset S^{(1)} \subset ... \subset S^{(n-1)} \equiv \mathbf{X}$, kurią galima vaizduoti grafiškai medžio pavidalo diagrama, vadinama dendrograma. Hierarchijos formavimo būdai plačiai nagrinėti Anderberg [6], Sneath ir Sokal [202], Hartigan [101], Everitt [59].

Klasterių artumo matai. Klasterinėje analizėje svarbu apibrėžti objektų homogeniškumo sąvoką. Tai sunkus uždavinys, nes jam spręsti nėra formalių metodų. Paprastai artumų matricoje elementai $\rho(S_l, S_m)$ apibūdina arba atstumą tarp objektų, arba įvertina jų panašumo (skirtumo) laipsnį. Atstumo tarp objektų metrikos $d(X(i), X(j))$ parinkimas yra vienas iš pagrindinių klasterinės analizės uždavinių. Nuo jos parinkimo priklauso ir galutinis objektų grupavimo į klasterius variantas. Praktikoje dažniausiai naudojamas Euklido atstumas [131]. Taip pat dažnai naudojamas Euklido kvadratinis atstumas, svertinis Euklido atstumas, Čebyševo atstumas, Manhateno atstumas, Minkovskio atstumas, absoliutusis laipsninis atstumas [116].

Populiarios ir dažnai naudojamos šios klasterių artumo metrikos [116, 131, 135, 156, 202]:

- vidutinio atstumo tarp visų galimų dviejų klasterių objektų porų:

$$\rho_{vid}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{X(i)\in S_l} \sum_{X(j)\in S_m} d(X(i), X(j));$$

- artimiausio kaimyno (angl. *nearest neighbor*):

$$\rho_{\min}(S_l, S_m) = \min_{\substack{X(i)\in S_l \\ X(j)\in S_m}} d(X(i), X(j));$$

- tolimiausio kaimyno (angl. *furthest neighbor*):

$$\rho_{\max}(S_l, S_m) = \max_{\substack{X(i)\in S_l \\ X(j)\in S_m}} d(X(i), X(j));$$

- centroidų (angl. *centroid*) – sunkio centrų metodas (matuojamas atstumas tarp klasterių „sunkio centrų“):

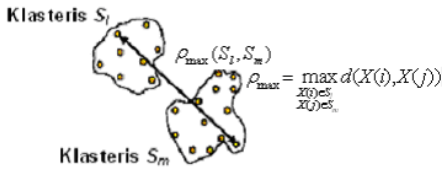
$$\rho_C(S_l, S_m) = d(\bar{S}_l, \bar{S}_m);$$

- Ward's:

$$\rho_W(S_l, S_m) = \frac{d(\bar{S}_l, \bar{S}_m)}{1/n_l + 1/n_m}.$$

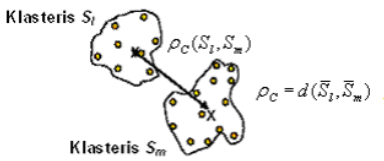
Tyrime hierarchinio klasterizavimo artumui tarp klasterių apibrėžti naudojame tris metrikas:

1. Tolimiausio kaimyno – atstumas tarp klasterių apibrėžiamas kaip atstumas tarp klasterių tolimiausių kaimynų (objektų).



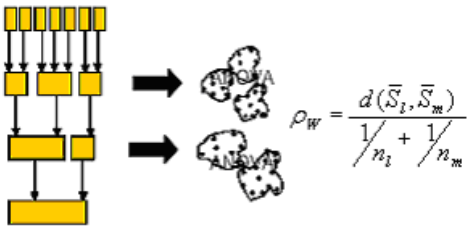
2.1 pav. Tolimiausio kaimyno atstumo tarp klasterių nustatymas

2. Centroidų – atstumas tarp klasterių apibrėžiamas kaip Euklido atstumo tarp klasterių centrų kvadratas.



2.2 pav. Centroidų atstumo tarp klasterių nustatymas

3. Taikant Ward‘s metodą, atstumas tarp dviejų klasterių apibrėžiamas kaip Euklido atstumų tarp visų įmanomų klasterius sudarančių objektų porų kvadratų suma. Tiesą sakant, kiekviename hierarchinio klasterizavimo etape atliekama dispersinė analizė.



2.3 pav. Ward’s atstumo tarp klasterių nustatymas

***k* vidurkių metodas.** Šis algoritmas klasterizuoja objektus, suskaidydamas juos į *k* klasterių. Jo idėją pasiūlė MacQueen [148] ir Hartigan [101]. Tai vienas iš EM algoritmo variantų, kurio tikslas – apibrėžti *k* vidurkių stebiniams, kurių skirstinys yra Gauso. Daroma prielaida, kad stebiniai yra daugiamačiai. Ieškoma suskaidymo, kuris minimizuoja dispersijas klasterių viduje arba funkciją

$$\sum_{i=1}^{k}\sum_{X\in S_i}\left\|X - m(S_i)\right\|^2, \qquad (2.7)$$

čia S_i, $i = 1, 2, \ldots, k$ yra klasteriai, o $m(S_i)$ yra klasterio S_i, sudaryto iš stebinių $X \in S_i$, „svorio centras“.

Algoritmas pradedamas suskaidant stebinius į k pradinių klasterių. Tuomet apskaičiuojamas kiekvieno klasterio vidurkis ar randamas jo centras. Atliekamas naujas suskaidymas stebinius priskiriant artimiausiems centrams. Vėliau perskaičiuojami naujų klasterių centrai ir šie du žingsniai kartojami tol, kol stebiniai nebekeičia klasterių, t. y. centrai stabilizuojasi.

Dėl greito savo konvergavimo šis algoritmas ypač populiarus praktikoje [13]. Ne vieną kartą pastebėta, jog algoritmo iteracijų skaičius yra daug mažesnis nei klasterizuojamų objektų skaičius.

***k* artimiausių kaimynų metodas.** Yra keletas *k* artimiausių kaimynų klasterizavimo metodo atmainų. Pateiksime vieną jų, aprašytą Gitman [82] ir Huizinga [112]. Klasterizavimas pradedamas nuo to, jog kiekvienas stebinys priskiriamas atskiram klasteriui. Toliau du klasteriai yra jungiami į vieną:

1) sudaromos visos įmanomos poros iš dviejų elementų;
2) skaičiuojamas kiekvienos poros tankis:

$$f_i(x) = \frac{n_i(x)}{n \cdot V(x)}, \tag{2.8}$$

   čia $n_i(x)$ – i-ojo klasterio kaimynų skaičius kartu priskaičiuojant ir patį stebinį x, $V(x)$ – $n_i(x)$ sudarančių stebinių užimamos erdvės hipertūris;
3) sujungiami du klasteriai, kurių bendras tankis yra didžiausias.

Toliau nagrinėjamas kiekvienas stebinys kartu su, atitinkamu vienam ar keliems kaimynams padedant, įvertintu tankiu (2.8). Tiriamasis stebinys gali priklausyti bet kuriam klasteriui ir tuo būdu randami jo *k* artimiausi kaimynai. Taigi, šis stebinys priskiriamas tam klasteriui, su kuriuo jį sujungus tankis (2.8) būna didžiausias ir ne mažesnis nei sujungus su bet kuriuo kaimynu.

Toks klasterių tikslinimas baigiamas, kai klasteriai nusistovi ir jų struktūra nebekinta [74, 229].

2.2. Klasterių skaičiaus nustatymo algoritmai

Viena iš problemų, su kuria dažnai susiduriama klasterinėje analizėje, – tai klasterių skaičiaus nustatymas arba modelio adekvatumo tyrimas [86]. Paprastai naudojami parametriniai kriterijai: tikėtinumo funkcijos ar kiti kriterijai, taip pat paremti tikėtinumo funkcija, pavyzdžiui, Akaike informacijos kriterijus (AIC). Tačiau taikant šiuos kriterijus susiduriama ir su tam tikromis problemomis. Visų pirma reikia apskaičiuoti funkcijos globalųjį maksimumą kaip lokaliųjų maksimumų didžiausią reikšmę, tačiau kartais tai atliekama ir su išimtimis. Tad tokiu atveju nė vienos procedūros taikymas negali garantuoti, kad toks globalusis maksimumas bus rastas. Antra vertus, taikant šiuos kriterijus daroma prielaida, kad vienas iš lyginamų parametrinių metodų yra teisingas. Ši prielaida kriterijų daro nestabilų. Pateikti argumentai verčia kelti klausimą: gal verta naudoti neparametrinius kriterijus, kad būtų patikrintas skirstinių mišinio modelio adekvatumas.

ω^2 tipo kriterijai. Šiuos kriterijus taikyti gana nesudėtinga [172]. Tarkim, $f^*(x) = f(\theta^*, x)$, čia θ^* – įvertis, gautas maksimalaus tikėtinumo metodu. Pažymėkime skirstinio funkciją F^*, o empirinę skirstinio funkciją – $\widetilde{F}(x) = n^{-1}\sum_{t=1}^{n}\prod_{j=1}^{d}\mathbf{1}_{\{X_j(t)<x_j\}}$. Apibrėždami

$$\psi_1 = \int_{\mathbf{R}^d}\left(F^*(x) - \widetilde{F}(x)\right)^2 d\widetilde{F}(x) = n^{-1}\sum_{t=1}^{n}\left[F^*(X(t)) - \widetilde{F}(X(t))\right]^2, \tag{2.9}$$

neatmesime nulinės modelio (2.1) adekvatiškumo hipotezės, jei tenkinama sąlyga

$$\psi_1 < \varepsilon_1. \tag{2.10}$$

Reikšmingumo lygmeniui α lygmuo ε_1 parenkamas kaip galima arčiau kvantilio u_α: $P\{\psi_1 > u_\alpha\} = \alpha$. Tam galima naudoti *bootstrap* metodą [2, 28, 91, 146, 230]. Jei G_1 pažymėsime statistikos ψ_1 skirstinio funkciją, tuomet, esant fiksuotam n, apibrėžiama atsitiktinio vektoriaus X skirstinio funkcija F, t. y.

$$P\{\psi_1 < u\} = G_1(u, F), \text{ visiems } u. \tag{2.11}$$

Kai yra duotas α, kintamasis ε_1 apibrėžiamas lygybe

$$G_1(\psi_1, F^*) = 1 - \alpha. \tag{2.12}$$

Funkciją $G_1(\cdot, F^*)$ galima gauti *bootstrap* metodu.

Tačiau nors ω^2 tipo kriterijus yra labai populiarus, reikia pripažinti, kad praktikoje kriterijus, paremtas pasiskirstymo tankių įverčių nukrypimais, o ne skirstinio funkcijomis, yra jautresnis. Tarkim, $\hat{f}(x)$ yra (1.6) tipo neparametrinis tankio $f(x)$ įvertis. Statistiką $\psi_2 = \| \hat{f}(x) - f^*(x) \|_2^2$, kaip modelio adekvatumo tyrimo kriterijų, rekomenduojama taikyti tikrinant sąlygą $\psi_2 < \varepsilon_2$.

Pažymėtina, kad jei hipotezė apie (2.1) modelį yra teisinga, kai $n \to \infty$, tai $\mathbf{E}\psi_1 = \mathrm{O}(n^{-1})$ ir $\mathbf{E}\psi_2 = \mathrm{O}(n^{-\gamma})$, čia $\gamma = 4/(4+d)$. Statistika ψ_2 konverguoja į nulį lėčiau nei ψ_1, o jos skirstinys iš esmės priklauso nuo θ parametro. Atskirai nuo ψ_2 kriterijaus kriterijus

$$\psi_3 = \int \frac{\hat{f}(x)}{f^*(x)} d\tilde{F}(x) - 1 = n^{-1} \sum_{t=1}^{n} \hat{f}(X(t)) \big/ f^*(X(t)) - 1 \tag{2.13}$$

vertas didesnio dėmesio. Kai ψ_2 yra įprastas nuostolių $\| \hat{f}(x) - f^*(x) \|_2^2$ neparametrinis įvertis, statistika ψ_3 yra nuostolių

$$\left\| (\hat{f}(x) - f^*(x)) \big/ \sqrt{f^*(x)} \right\|_2^2 = \int (\hat{f}(x) \big/ f^*(x)) dF(x) - 1$$

įvertis. Jei teisinga (2.1) hipotezė, kai $n \to \infty$, statistika ψ_3 konverguoja į nulį greičiau nei ψ_2, o $\mathbf{E}\psi_3 = \mathrm{O}(n^{-(1+\gamma)/2} \log n)$. Taip pat matyti, kad ψ_3 skirstinys nedaug priklauso nuo θ reikšmės. Taikant ψ_3 kriterijų, tikrinama sąlyga $\psi_3 < \varepsilon_3$.

Lygmenys ε_3 ir ε_2 nustatomi taip pat kaip (2.11) ir (2.12) naudojant parametrinį *bootstrap* metodą.

Tikėtinumo funkcijos prieaugio kriterijai. Pradedant nuo reikšmės $q = 1$, parametras q toliau nuosekliai didinamas tol, kol atmetama hipotezė apie (2.1) modelio adekvatumą (pvz., su reikšmingumo lygmeniu $\alpha = 0{,}05,\ 0{,}1$). Šiai hipotezei tikrinti galima taikyti kriterijų, paremtą tikėtinumo funkcijos prieaugio naudojimu. Tarkim, $\hat{\theta}(q)$ yra θ įvertis, gautas naudojantis EM algoritmu, kai klasterių skaičius yra q,

$$\psi = L(\hat{\theta}(q+1)) - L(\hat{\theta}(q)) . \tag{2.14}$$

Tarkim, $G(u)$ žymi ψ skirstinio įvertį, gautą parametriniu *bootstrap* metodu darant prielaidą, kad Gauso skirstinių mišinio modelis (2.1) yra teisingas. Tada hipotezė apie modelio (2.1) adekvatumą neatmetama, kai

$$1 - G(\psi) \geq \alpha. \tag{2.15}$$

Šitaip parinktą klasterių skaičių žymėsime q^* [1A].

Praktiniuose tyrimuose neretai susiduriama su problema, kai dėl didelio stebinių skaičiaus bei mažų tankio funkcijos reikšmių juose kompiuteriu apskaičiuota tikėtinumo funkcijos reikšmė lygi nuliui. Šiuo atveju galima naudoti funkciją, ekvivalenčią tikėtinumo funkcijai:

$$\widetilde{L}(\theta) = \prod_{t=1}^{n} \left(\aleph + f(X(t), \theta)\right), \tag{2.16}$$

čia $\aleph = \max\{u : \sum_{t=1}^{n} \mathbf{1}_{\{f(X(t),\theta)\leq u\}} \leq 0{,}05n\}$. Tuomet (2.14) lygybę atitinkamai galima perrašyti taip:

$$\widetilde{\psi} = \widetilde{L}(\hat{\theta}(q+1)) - \widetilde{L}(\hat{\theta}(q)). \tag{2.17}$$

Taikant $\widetilde{\psi}$ kriterijų, tikrinama sąlyga $1 - G(\widetilde{\psi}) \geq \alpha$.

Sarle kubinis klasterizavimo kriterijus. Klasteriams atskirti plačiai naudojamas optimizavimo kriterijus, žinomas kaip kvadratų sumos tarp klasterių stebinių vertinimas [180]. Paprastai naudojama statistika nusakanti klasterizuotų stebinių sklaidą:

$$R^2 = 1 - \frac{\mathrm{Tr}(\mathbf{W})}{\mathrm{Tr}(\mathbf{T})},$$

čia Tr(**W**) ($\mathbf{W} = \mathbf{X}'\mathbf{X} - \overline{\mathbf{X}}'\mathbf{V}'\mathbf{V}\overline{\mathbf{X}}$, čia **V** – klasterių indikatorinė matrica, kurios elementai $v_{tk} = 1$, jei *t*-asis stebinys yra iš *k*-ojo klasterio, kitu atveju $v_{tk} = 0$, $\overline{\mathbf{X}}$ – empirinis imties vidurkis) yra klasterius sudarančių stebinių kvadratų suma, Tr(**T**) yra visos imties elementų kvadratų suma, t. y. matricos **T**, apibrėžtos lygybe **T** = **X′X**, pėdsakas.

Taikant Sarle kubinį klasterizavimo kriterijų (CCC), tikrinamos tokios hipotezės:

H_0: stebinių skirstinys daugiamatis tolygusis.

H_1: stebinių skirstinys yra daugiamačių Gauso skirstinių mišinys.

Teigiamų CCC reikšmių atveju H_0 atmetama.

Norint apskaičiuoti CCC, pirmiausia įvertinama dispersija:

$$E(R^2) \cong 1 - \left[\sum_{j=1}^{d^*} \frac{1}{n+u_j} + \sum_{j=d^*+1}^{d} \frac{u_j^2}{n+u_j} \right] \left[\frac{(n-q)^2}{n} \right] \left[1 + \frac{4}{n} \right] \Big/ \sum_{j=1}^{d} u_j^2 \,. \tag{2.18}$$

čia s_j yra hiperkubo j-osios kraštinės ilgis, $u_j = s_j/c$, kai $c = (v/q)^{\frac{1}{d}}$ ir $v = \prod_{i=1}^{d} s_i$, d^* – didžiausias sveikasis skaičius mažesnis už q, bet toks, kad u_{d^*} būtų ne mažesnis už vienetą.

CCC, priklausantis nuo R^2, apskaičiuojamas taip:

$$CCC = \log\left[\frac{1 - E(R^2)}{1 - R^2} \right] \frac{\sqrt{nd^*/2}}{(0{,}001 + E(R^2))^{1.2}}, \tag{2.19}$$

čia $R^2 = 1 - \left[d^* + \sum_{j=d^*+1}^{d} u_j^2 \right] \Big/ \sum_{j=1}^{d} u_j^2$.

Ieškoma CCC reikšmės lokaliųjų maksimumų. Jei CCC reikšmės yra tarp 0 ir 2, tai taip parinkta klasterių struktūra yra galima, tačiau interpretuoti ją reikia atsargiai. Jei reikšmės yra didesnės už 2, tai sakoma, kad klasterių skaičius parinktas tinkamai.

Pseudo-*F* kriterijus. Matuojant klasterių atsiskyrimą, pseudo-*F* (PSF) kriterijumi taip pat mėginama nustatyti klasterių skaičių imties pagrindu. PSF kriterijus gali būti naudojamas kaip klasterių skaičiaus indikatorius, tačiau jo pavadinimas iš dalies yra klaidinantis. Ši statistika nėra pasiskirsčiusi kaip Fišerio atsitiktinis dydis, t. y. ji pasiskirsčiusi kaip Fišerio atsitiktinis dydis su $d(q-1)$ ir $d(n-q)$ laisvės laipsniais, jei tenkinamos ir dvi prielaidos, kurios beveik niekada nepasitvirtina:

1) klasterizavimo metodas stebinius klasteriams priskiria atsitiktinai,
2) stebiniai yra tarpusavyje nepriklausomi, pasiskirstę pagal daugiamatį normalinį skirstinį.

PSF apibrėžiamas taip:

$$PSF = \frac{\left(\sum_{t=1}^{n} \left\| X(t) - \overline{X} \right\|^2 \right) / (q-1)}{\left(\sum_{k=1}^{q} \sum_{i \in C_k} \left\| X(i) - \overline{X}_k \right\|^2 \right) / (n-q)}, \tag{2.20}$$

čia $\overline{X}$ yra imties vidurkis, o $\overline{X}_k$ – *k*-ojo klasterio vidurkis. PSF gali būti išreikštas statistikos R^2 nariais:

$$PSF = \frac{R^2/(q-1)}{(1-R^2)/(n-q)}. \tag{2.21}$$

PSF interpretuojamas panašiai kaip ir CCC: ieškoma PSF reikšmės lokaliųjų maksimumų ir sprendžiama apie klasterių pasirinkimą.

Pseudo-T^2 kriterijus. Pseudo-T^2 kriterijaus statistika (PST2) yra Hotelling T^2 kriterijaus [109] variantas, kuriuo lyginami dviejų daugiamačių aibių vidurkiai. PST2 naudojamas priimant sprendimą, kurie du klasteriai turi būti sujungti arba ne. Paprastai kriterijumi PST2 matuojamas atskyrimas tarp dviejų vėliausiai sujungtų klasterių. Jei PST2 reikšmė yra didelė, tai klasterių vidurkiai skiriasi reikšmingai, o klasteriai nesujungiami, ir atvirkščiai – jei PST2 reikšmė yra maža, tai klasterius galima sujungti užtikrintai.

Pseudo-T^2 statistikos skirstinys yra Fišerio skirstinys su d ir $d(n_k+n_l-2)$ laisvės laipsniais. PST2 apskaičiuojama sujungus klasterius C_k ir C_l į klasterį C_m:

$$PST2 = \frac{w_m - w_k - w_l}{(w_k + w_l)/(n_k + n_l - 2)}, \tag{2.22}$$

čia n_k yra *k*-ojo klasterio stebinių skaičius, o $w_k = \sum_{i \in C_k} \left\| X(i) - \overline{X}_k \right\|$.

Pagrindinė pseudo-T^2 statistikos (PST2) interpretavimo taisyklė – ieškoti klasterių skaičiaus mažėjimo kryptimi pirmos gerokai didesnės reikšmės, nei prieš tai buvusi, ir parinkti klasterių skaičių, kuris atitinka prieš tai esančią (mažesnę) statistikos reikšmę.

Beale *F* kriterijus. Iš daugelio formalių metodų, kurie naudojami klasterių skaičiui nustatyti, Beale *F* kriterijus laikomas vienu geriausių [84]. Jam apskaičiuoti reikalinga atstumų tarp stebinių kvadratų suma. Tam klasterių viduje skaičiuojama liekanų kvadratų sumos reikšmė:

$$w_1 = \sum_{k=1}^{q_1} \sum_{j=1}^{n_k} (X_k(j) - \overline{X}_k)'(X_k(j) - \overline{X}_k). \tag{2.23}$$

Tas pats daroma ir su kitu klasterių suskaidymu:

$$w_2 = \sum_{k=1}^{q_2} \sum_{j=1}^{n_k} (X_k(j) - \overline{X}_k)'(X_k(j) - \overline{X}_k). \quad (2.24)$$

Jei klasterių viduje skaičiuojamos liekanų kvadratų sumų reikšmės yra priskirtos taip, kad klasterizavimas su didesniu klasterių skaičiumi yra q_1, o atitinkama kvadratų suma w_1, tai:

1) jei w_1 yra daug mažesnis už w_2, tai pasirinkti didesnį q_1 yra geriau nei mažesnį q_2;
2) jei w_1 yra daug didesnis už w_2, tai geriau yra pasirinkti nedidelį klasterių skaičių q_2.

Žinoma, kyla klausimas, koks didelis skirtumas turi būti.

Tikrinant, ar skirtumas tarp dydžių w_1 ir w_2 yra statistiškai reikšmingas, skaičiuojama ši statistika:

$$F^* = \frac{w_2 - w_1}{w_1} \cdot \frac{c_1(n - q_1)}{c_2(n - q_2) - c_1(n - q_1)}, \quad (2.25)$$

čia q_1 yra didesnis klasterių skaičius nei q_2, w_1 ir w_2 – atitinkamos atstumų, esančių klasterių viduje, kvadratų sumos, o $c_1 = q_1^{(-2/d)}$ ir $c_2 = q_2^{(-2/d)}$.

Galiausiai, F^* reikšmė lyginama su Fišerio skirstinio kritine reikšme (pavyzdžiui, kai $\alpha = 0{,}05$), kai skaitiklio laisvės laipsnių skaičius

$$c_2 (n - q_2) - c_1 (n - q_1),$$

o vardiklio laisvės laipsnių skaičius

$$c_1 (n - q_1).$$

Jei Beale F statistika yra didesnė už Fišerio skirstinio kritinę reikšmę, tai klasterizavimas su daugiau klasterių yra geresnis. Kitu atveju pasirenkamas sprendimas su mažesniu klasterių skaičiumi.

2.3. Pirminio duomenų klasterizavimo poveikio pasiskirstymo tankio neparametrinio vertinimo tikslumui tyrimas

Vienas iš būdų mėginti padidinti tankių vertinimo tikslumą – pereiti nuo daugiamodalinio tankio analizės prie vienamodalinių tankių vertinimo traktuojant tiriamąjį tankį kaip vienamodalinių tankių mišinį. Siūlome pirmame tyrimų etape imtį klasterizuoti, o paskui kiekvieną klasterį atitinkančius skirstinių mišinio komponentus įvertinti atskirai. Be to, klasterizavimui galima panaudoti rekurentinę procedūrą, kuri remiasi stebinių skirstinio aproksimacija Gauso skirstinių mišiniu ir EM algoritmo taikymu, aprašytu 2.1 skyriuje. Kaip jau buvo anksčiau minėta, taikant šį algoritmą susiduriama su pradinės reikšmės parinkimo problema. Paprasčiausias jos sprendimo būdas – taikyti atsitiktinio starto principą [133]: EM algoritmas kartojamas daug kartų, kiekvieną kartą pradinius įverčius $\hat{\theta}^{(0)}$ parenkant atsitiktinai. Galutinis rezultatas parenkamas taip, kad jį atitinkanti tikėtinumo funkcijos $L(\hat{\theta})$ reikšmė būtų maksimali. Klasterių skaičius q parenkamas kryžminio patikrinimo būdu [143]. Gerus rezultatus taip pat galima gauti naudojant Lietuvos Matematikos ir informatikos institute sukurtą konstruktyvią procedūrą, kurioje klasterių skaičius parenkamas taikant ω^2 tipo kriterijų [172].

Šiame tyrime Monte Karlo metodu buvo siekiama atlikti anksčiau 1.1 ir 1.2 skyriuose aprašytų pasiskirstymo tankio neparametrinių įvertinių (kitaip nei 1 dalyje tirtame apvertimo formulės taikyme, čia naudotas ne IFDE, o IKDE metodas) tikslumo lyginamąją analizę tuo atveju, kai stebinių pasiskirstymo tankis yra daugiamodalinis, ir nustatyti, ar tikslinga, vertinant tokio tipo tankius imtį preliminariai suskaidyti į klasterius. Buvo generuojamos atsitiktinės imtys su nepriklausomais stebiniais, pasiskirsčiusiais pagal Gauso ir Koši skirstinių mišinius. Norint įvairiapusiškai ištirti siūlomus metodus, buvo keičiamas imties dydis, mišinių komponentų sanklotos laipsnis.

Paklaida. Norint palyginti gautus rezultatus su gautais [113] darbe, algoritmams vertinti naudota paklaida, apibrėžianti atstumą tarp dviejų funkcijų:

$$\delta = \mathbf{E}(g(X) - f(X))^2 / \mathbf{D}f(X). \tag{2.26}$$

Pažymėjus tankio funkciją f, o jos įvertį $\hat{f}$ – g ir paėmus jų empirinius analogus, paklaida

$$\delta = Err / Var, \tag{2.27}$$

čia $Err = \frac{1}{n}\sum_{t=1}^{n}(\hat{f}_t - f_t)^2$ reiškia vidutinę kvadratinę įvertinto tankio $\hat{f}_t$ ir tikrojo tankio $f_t = f(X(t))$ paklaidą, o $Var = \frac{1}{n}\sum_{t=1}^{n}(f_t - \bar{f})^2$ reiškia imties sklaidą, čia $\bar{f}$ nusako $f(X(1)), \ldots, f(X(n))$ vidurkį.

Kiekvienu atveju apskaičiuotas paklaidos δ aritmetinis vidurkis $\bar{\delta}$ bei standartinis nuokrypis gauti sugeneravus 100 nepriklausomų imčių. Juos ir pateiksime lentelėse bei grafikuose.

Tirti modeliai. Buvo naudojami tie patys duomenys, kuriuos savo darbe [113] jau buvo naudoję J. N. Hwang, S. R. Lay ir A. Lippman. Naudojami trijų tipų daugiamačiai ($d = \overline{2,5}$) Gauso ir Koši skirstinių su nepriklausomomis komponentėmis mišiniai. Koši skirstiniai turi sunkias uodegas, kokių neturi Gauso skirstiniai. Duomenų skirstinių tankių mišiniai aprašomi taip:

$$f(x) = \sum_{i=1}^{q} p_i f_N(x, m_i, \sigma_i) \text{ (Gauso skirstinių mišinys),}$$

$$f(x) = \sum_{i=1}^{q} p_i f_C(x, m_i, u_i) \text{ (Koši skirstinių mišinys)}$$

su apribojimais $\sum_{i=1}^{q} p_i = 1,\ p_i \ge 0,\ i = \overline{1,q}$. Čia

$$f_N(x, m_i, \sigma_i) = \frac{1}{\prod_{j=1}^{d}\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{d}\frac{(x_j - m_{ij})^2}{\sigma_{ij}^2}\right\},$$

$$f_C(x, m_i, u_i) = \prod_{j=1}^{d}\frac{u_{ij}}{\pi[u_{ij}^2 + (x_j - m_{ij})^2]}.$$

B1: *Vienos modos skirstinys*. Pirmasis duomenų tipas yra vienos modos skirstinys su tokiais parametrais (pažymėtina tai, jog $d = \overline{2,4}$ atvejais mišinio parametrai apibrėžiami imant $d = 5$ atvejo pirmus parametrų elementus): Gauso atveju,

p=1, m=(0,0; 0,0; 0,0; 0,0; 0,0)′, σ^2=(0,84; 1,02; 0,70; 1,20; 0,96)′;

Koši atveju

p=1, m=(0,0; 0,0; 0,0; 0,0; 0,0)′, u=(0,84; 1,02; 0,70; 1,20; 0,96)′.

B2: *Mažai persidengiantis dviejų modų skirstinys*. Antrasis duomenų tipas yra mažai persidengiantis dviejų modų skirstinys su tokiais parametrais: Gauso atveju,

p_1=0,65, m_1=(0,0; 0,0; 0,0; 0,0; 0,0)′, σ_1^2=(0,42; 0,51; 0,35; 0,60; 0,48)′,

p_2=0,35, m_2=(2,0; 2,0; 2,0; 2,0; 2,0)′, σ_2^2=(0,33; 0,46; 0,53; 0,43; 0,45)′;

Koši atveju

p_1=0,65, m_1=(0,0; 0,0; 0,0; 0,0; 0,0)′, u_1=(0,42; 0,51; 0,35; 0,60; 0,48)′,

p_2=0,35, m_2=(2,0; 2,0; 2,0; 2,0; 2,0)′, u_2=(0,33; 0,46; 0,53; 0,43; 0,45)′.

B3: *Labai persidengiantis dviejų modų skirstinys*. Trečiasis duomenų tipas yra labai persidengiantis dviejų modų skirstinys su tokiais parametrais: Gauso atveju,

p_1=0,65, m_1=(0,0; 0,0; 0,0; 0,0; 0,0)′, σ_1^2=(0,84; 1,02; 0,70; 1,20; 0,96)′,

p_2=0,35, m_2=(2,0; 2,0; 2,0; 2,0; 2,0)′, σ_2^2=(0,66; 0,92; 1,06; 0,86; 0,90)′;

Koši atveju

p_1=0,65, m_1=(0,0; 0,0; 0,0; 0,0; 0,0)′, u_1=(0,84; 1,02; 0,70; 1,20; 0,96)′.

p_2=0,35, m_2=(2,0; 2,0; 2,0; 2,0; 2,0)′, u_2=(0,66; 0,92; 1,06; 0,86; 0,90)′.

Kiekvieno tipo duomenims (tiek Gauso, tiek Koši skirstinių mišiniams) skirtinguose matavimuose ($d = \overline{2,5}$) buvo generuotos įvairaus didumo imtys (200, 400, 800, 1600, 3200).

2.4. Pirminio klasterizavimo poveikio pasiskirstymo tankio vertinimo tikslumui tyrimo rezultatai

Adaptuotu branduoliniu ir tikslinio projektavimo metodais gauti rezultatai yra panašūs kaip ir gauti J. N. Hwang, S. R. Lay bei A. Lippman: turimos mažesnio matavimo sunkių uodegų (Koši) imtys geriau aprašomos branduoline struktūra. Esant didesniam matavimų skaičiui ir didesnėms imtims (nuo 400 stebinių) ar Gauso skirstinių atveju, geresni rezultatai gauti tikslinio projektavimo tankio įvertiniu. Penkiamačių Gauso mišinių atveju geri rezultatai gauti naudojantis apvertimo formulės tankio įvertiniu. Vertinant vienamodalinius Gauso bei Koši skirstinius, paklaidos gautos net iki 4,6 karto mažesnės; tai ypač gerai matoma, kai imtys yra mažos. Pradinis duomenų sujungimas į homogenines grupes taikant Gauso skirstinių mišinio modelį ir EM algoritmą leido gauti geresnius rezultatus: esant mažoms imtims, paklaidos sumažėjo net 2 ar 3, o kai kuriais atvejais net 5 kartus, esant didesnėms imtims (n = 1600, 3200), šis santykis yra mažesnis ir siekia 1,05–2 kartus. *Šiame tyrime vienareikšmiškai geriausiu galima laikyti tikslinio projektavimo įvertinį* [3A]. Vienamodaliniai Gauso skirstiniai gerai vertinami taikant pusiau parametrinį branduolinį metodą, o penkiamačiai jų mišiniai – ir apvertimo formulės taikymo metodą. Buvo pastebėta, jog pavienių imčių atvejais logsplaininiu metodu buvo gauti labai tikslūs rezultatai, tačiau jei pasitaikydavo smarkiai išsiskiriančių stebinių, bendras vidutinis rezultatas iš esmės pablogėdavo. Apvertimo formulės taikymo algoritmas yra

labai lėtas ir skaičiavimams reikia kur kas dagiau CPU laiko, palyginti su kitais tirtais algoritmais.

Šio modeliavimo tyrimo tipiniai rezultatų pavyzdžiai pateikti 2.1, 2.2 ir 2.3 lentelėse. Skliausteliuose vaizduojamos paklaidų standartinių nuokrypių reikšmės.

2.1 lentelė. Vienamodaliniai keturmačiai skirstiniai

Metodas	Gauso skirstinys				Koši skirstinys			
	$n = 400$		$n = 1600$		$n = 400$		$n = 1600$	
	be klast.	su klast.	be klast.	su klast.	be klast.	su klast.	be klast.	su klast.
AKDE	0,2055	0,1178	0,1455	0,0845	0,1994	0,1787	0,1283	0,1052
	(0,0289)	(0,0121)	(0,0169)	(0,0169)	(0,0056)	(0,0324)	(0,0014)	(0,0051)
PPDE	0,126	0,0668	0,0457	0,0243	0,1804	0,1115	0,0445	0,0323
	(0,0088)	(0,0141)	(0,0061)	(0,0061)	(0,0034)	(0,0109)	(0,0073)	(0,0031)
IKDE	0,1764	0,1661	0,126	0,1159	0,2099	0,19	0,0777	0,0719
	(0,0063)	(0,0178)	(0,0037)	(0,0059)	(0,0087)	(0,0278)	(0,0094)	(0,0118)
LSDE	*****	0,1208	*****	0,1099	*****	0,1729	*****	0,0729
	*****	(0,0066)	*****	(0,0126)	*****	(0,0107)	*****	(0,0045)
SKDE	*****	0,0993	*****	0,0541	*****	0,1908	*****	0,0647
	*****	(0,0124)	*****	(0,0015)	*****	(0,0032)	*****	(0,0071)

2.2 lentelė. Dvimodaliniai mažai persidengiantys keturmačiai skirstiniai

Metodas	Gauso mišinys				Koši mišinys			
	$n = 400$		$n = 1600$		$n = 400$		$n = 1600$	
	be klast.	su klast.	be klast.	su klast.	be klast.	su klast.	be klast.	su klast.
AKDE	0,2963	0,2531	0,2495	0,1882	0,2173	0,1755	0,1706	0,1257
	(0,055)	(0,0166)	(0,0706)	(0,0178)	(0,1644)	(0,0466)	(0,0861)	(0,0214)
PPDE	0,2219	0,0928	0,059	0,0328	0,2106	0,2027	0,1834	0,1057
	(0,0242)	(0,0137)	(0,0229)	(0,0148)	(0,0804)	(0,0598)	(0,0266)	(0,0224)
IKDE	0,253	0,2531	0,1841	0,1766	0,227	0,2124	0,1851	0,1732
	(0,0621)	(0,0017)	(0,0316)	(0,0017)	(0,0685)	(0,0037)	(0,0847)	(0,0214)
LSDE	*****	0,1281	*****	0,0824	*****	0,213	*****	0,1378
	*****	(0,0148)	*****	(0,0136)	*****	(0,0283)	*****	(0,0077)
SKDE	*****	0,1393	*****	0,0759	*****	0,2011	*****	0,1418
	*****	(0,0107)	*****	(0,0147)	*****	(0,0313)	*****	(0,0201)

2.3 lentelė. Dvimodaliniai labai persidengiantys keturmačiai skirstiniai

Metodas	Gauso mišinys				Koši mišinys			
	$n = 400$		$n = 1600$		$n = 400$		$n = 1600$	
	be klast.	su klast.	be klast.	su klast.	be klast.	su klast.	be klast.	su klast.
AKDE	0,2526	0,1049	0,2039	0,0629	0,2478	0,1946	0,1416	0,1341
	(0,0471)	(0,0058)	(0,0729)	(0,0094)	(0,0889)	(0,0123)	(0,0434)	(0,0109)
PPDE	0,1684	0,0512	0,0591	0,0412	0,1879	0,1628	0,1403	0,0912
	(0,0278)	(0,0106)	(0,005)	(0,0063)	(0,0078)	(0,0429)	(0,0165)	(0,0021)
IKDE	0,2563	0,2321	0,1808	0,1644	0,2496	0,2239	0,1455	0,1427
	(0,0122)	(0,0018)	(0,005)	(0,0052)	(0,0258)	(0,0518)	(0,0373)	(0,0213)
LSDE	*****	0,1772	*****	0,1213	*****	0,2184	*****	0,1352
	*****	(0,0061)	*****	(0,0055)	*****	(0,0299)	*****	(0,0106)
SKDE	*****	0,0801	*****	0,0809	*****	0,2193	*****	0,1245
	*****	(0,0078)	*****	(0,0097)	*****	(0,0047)	*****	(0,0056)

1 priede grafiškai pavaizduoti tirtų tankių vertinimo metodų tikslumo rezultatai priklausomai nuo imties didumo, kai taikomas pradinis imties klasterizavimas, taip pat adaptuoto branduolinio ir tikslinio projektavimo metodų palyginimas naudojant imties klasterizavimą ir jo nenaudojant.

3. Įvairių klasterizavimo procedūrų taikymo efektyvumas pasiskirstymo tankiams vertinti

3.1. Paplitusių geometrinio klasterizavimo procedūrų taikymas pasiskirstymo tankiui statistiškai vertinti

Geometrinis duomenų klasterizavimas – vienas iš metodų, kuriam nereikia suderintų pasiskirstymo parametrų įverčių, todėl jis gali būti naudojamas kaip pradinių daugiamačių duomenų suskaidymo metodas. Vėliau geometriniu klasterizavimu gauti klasteriai gali būti tikslinami (perklasterizuojami) kitais metodais. Tyrime perklasterizuojama neparametriškai įvertinus tankį ir taikant Bajeso principą bei *aposteriorinių* tikimybių $\pi_k(x) = \mathbf{P}\{v = k | X = x\}$ įverčius $\hat{\pi}_k(x)$, $k = \overline{1,q}$, o imtis suskaidoma į klasterius remiantis (2.3) lygybėmis. Populiarios geometrinio klasterizavimo procedūros palygintos tiriant jų taikymą pasiskirstymo tankiui statistiškai vertinti. Nagrinėtų klasterių skaičiaus parinkimo kriterijų (Sarle, pseodo-F, pseudo-T^2, Beale F) geometrinio klasterizavimo atveju atlikta literatūros [179, 183] analizė parodė, jog šiame tyrime rekomenduojama taikyti Sarle kubinį klasterizavimo kriterijų bei Beale F kriterijų. Atlikus preliminarų kriterijų, aprašytų 2.2 skyriuje ir skirtų geometriniam klasterizavimui, palyginimo tyrimą buvo pasirinktas Sarle kubinis klasterizavimo kriterijus [7K]. Klasterizavimui atlikti bei klasterių skaičiui nustatyti galima taikyti statistinę programinę įrangą SAS (jos aprašymą žiūrėti [16, 182, 183, 184]); darbe ji ir buvo naudota. Iš visų 1.1 ir 1.2 skyriuose aprašytų pasiskirstymo tankio neparametrinių įvertinių šiame tyrime buvo naudoti adaptuotas branduolinis, pusiau parametrinis, histosplaininis bei Friedman tikslinio projektavimo įvertiniai [6A].

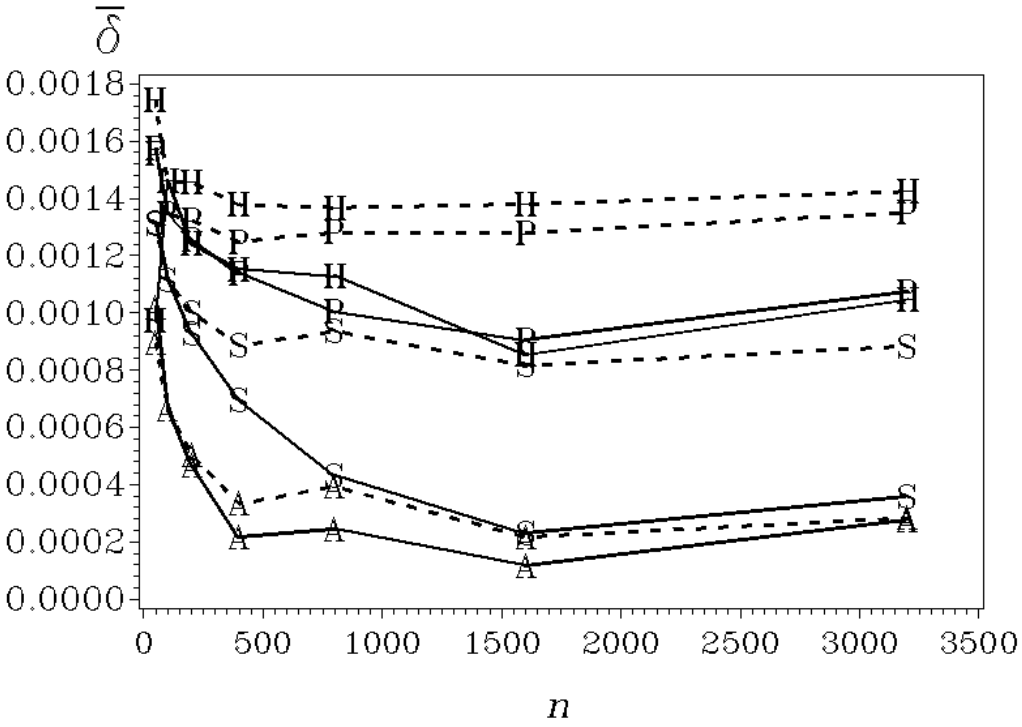
Paklaida. Praktikoje gali būti sprendžiamas ne tik klasifikavimo uždavinys, bet ir skirstinių mišinio parametrų arba tankio vertinimo uždavinys. Kadangi panašius tankius galima nustatyti visai skirtingais parametrais, tai tikslumui įvertinti naudosime L_2 atstumą tarp tankio ir jo įverčio:

$$\delta = \frac{1}{n}\sum_{t=1}^{n}\left(f(X(t)) - \hat{f}(X(t))\right)^2 \cong \int\left(f(x) - \hat{f}(x)\right)^2 f(x)dx\,. \tag{3.1}$$

Palyginimas. Tyrimui buvo naudoti tie patys duomenų modeliai B1, B2 ir B3, kuriuos naudojome 2.3 skyriuje. Kiekvieno tipo duomenims (tiek Gauso, tiek Koši skirstinių mišiniams) skirtinguose matavimuose ($d = 2$ ir $d = 5$) generuotos įvairaus didumo imtys (50, 100, 200, 400, 800, 1600, 3200).

Duomenų klasterizavimas įvairiais metodais daugeliu atvejų labiausiai pagerino mažų imčių rezultatus, kadangi mažos imtys dažniau buvo skirstomos į didesnį klasterių skaičių nei didesnės imtys. Koši skirstinių rezultatai pagerėjo labiau nei Gauso skirstinių. Tankių įverčių tikslumui daugiausia įtakos turėjo *k* artimiausių kaimynų klasterizavimo procedūra.

3.1 paveiksle pateiktos dvimačių Koši vienamodalinių tankių vertinimo paklaidos, kai duomenys pradžioje nebuvo klasterizuojami ir buvo klasterizuojami *k* artimiausių kaimynų procedūra. Čia simboliais A, P, S ir H atitinkamai žymimi AKDE, PPDE, SKDE ir HSDE algoritmai; punktyrine linija žymimos pasiskirstymo tankio įverčių paklaidos neatlikus pirminio duomenų klasterizavimo, o ištisine – atlikus pirminį duomenų klasterizavimą. Histosplaininio metodo taikymo kartu su pirminiu duomenų klasterizavimu rezultatai buvo gerokai blogesni nei gauti kitais metodais, todėl jo grafike nevaizduojame. Visų metodų rezultatai įvairaus tūrio imtims buvo pagerinti panaudojus šią duomenų klasterizavimo procedūrą. Koši vienamodalinis skirstinys (B1 modelis) yra vienas tų, kurio klasterizavimas skirtingomis procedūromis pagerino įvairaus didumo imčių tankio vertinimo rezultatus.

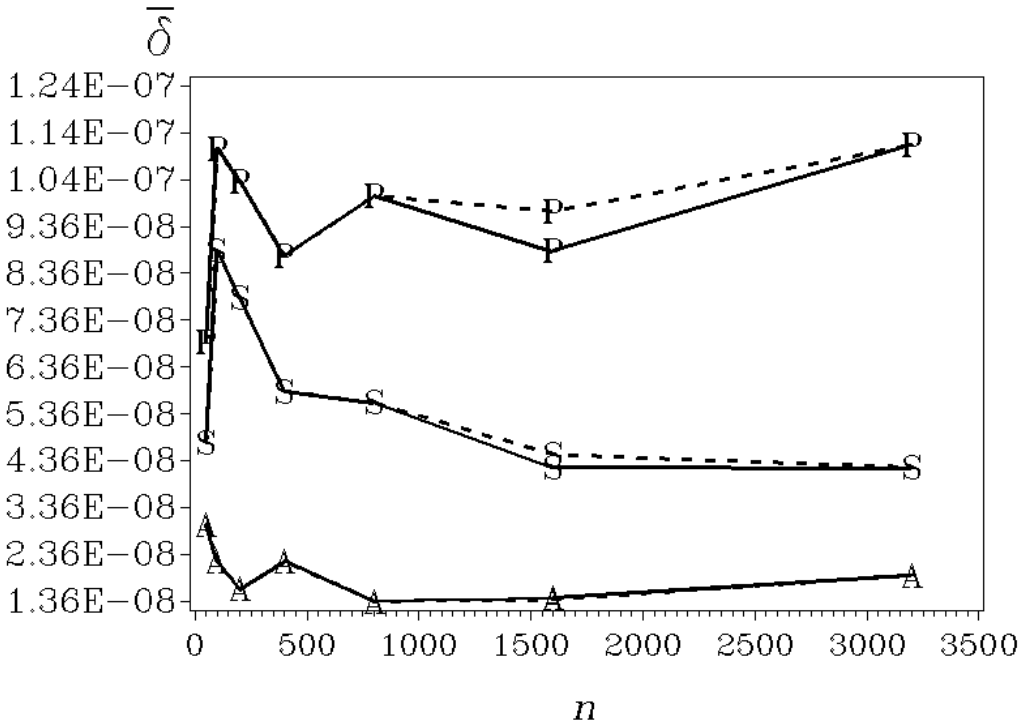


3.1 pav. Koši dvimačių vienmodalinių tankių paklaidų palyginimas duomenų neklasterizuojant ir juos klasterizuojant

Penkiamačiu atveju pradinis klasterizavimas geometrinėmis procedūromis pasiskirstymo tankio įverčius pagerino nedaug. Didžiausia buvo *k* artimiausių kaimynų klasterizavimo procedūros įtaka.

3.2 paveiksle pateiktos penkiamačių Koši vienamodalinių tankių vertinimo paklaidos, kai duomenys pradžioje nebuvo klasterizuojami ir kai buvo klasterizuojami *k* artimiausių kaimynų procedūra. Čia simboliais A, P ir S atitinkamai žymimi AKDE, PPDE ir SKDE algoritmai; punktyrine linija žymimos paklaidos neatlikus pirminio duomenų klasterizavimo, o ištisine – atlikus pirminį duomenų klasterizavimą. Penkiamačiu atveju paklaidos yra daug mažesnės, nes ir pačių tankių įverčių reikšmės yra mažesnės nei dvimačiu atveju. Kaip jau minėta, dvimačiu atveju duomenų klasterizavimas didžiausią įtaką turėjo Koši tipo skirstiniams, tačiau penkiamačiu atveju pirminis klasterizavimas pasiskirstymo tankio įverčių tikslumą pagerina mažiau nei dvimačiu atveju.

Duomenų perklasterizavimas dvimačiu atveju pagerino tankių vertinimo rezultatus, gautus kintamo pločio branduoliniu metodu, taip pat tikslinio projektavimo ir pusiau parametriniu branduoliniu metodu. Penkiamačiu atveju perklasterizavimas buvo naudingas tik taikant tikslinio projektavimo įvertinį, o taikant branduolinius įvertinius rezultatai neretai buvo blogesni.



3.2 pav. Koši penkiamačių vienmodalinių tankių paklaidų palyginimas duomenų neklasterizuojant ir juos klasterizuojant

Kaip parodė grafinė paklaidų analizė, naudojant geometrinį klasterizavimą gauti geresni rezultatai nei jo nenaudojant. Tačiau tolimesni tyrimai, kuriuose populiarios geometrinio klasterizavimo procedūros palygintos su tikimybinio klasterizavimo procedūra (EM algoritmu),

parodė, jog vertinant pasiskirstymo tankius tikimybiniai klasterizavimo metodai yra akivaizdžiai pranašesni už populiarius tirtus geometrinio klasterizavimo metodus. 1 priede pateikti grafinės analizės pavyzdžiai klasterizavimo (*k* artimiausių kaimynų ir Gauso skirstinių mišinio modelio bei EM algoritmo) metodams palyginti. Toliau aptarsime tikimybinių metodų taikymą, kai imtis yra klasterizuojama negriežtai.

3.2. Negriežto imties klasterizavimo naudojimas neparametriškai vertinant pasiskirstymo tankį

2 dalyje buvo ištirtas imties klasterizavimo, naudojant Gauso skirstinių mišinio modelį ir EM algoritmą, tikslingumas. Šiame skyriuje aptarsime tankių vertinimo metodų taikymą sąlyginiam pasiskirstymo tankiui vertinti, kai imtis yra klasterizuota negriežtai naudojant tą pačia procedūrą. Negriežto klasterizavimo atveju klasteriai suprantami kaip aibės

$$\{(X(1), \pi_i(1)), ..., (X(n), \pi_i(n))\},\ i = \overline{1, q}.$$

čia $\pi_i(t)$ rodo su kokiu svoriu (*aposteriorine* tikimybe) stebinys $X(t)$ priskiriamas i-ajai klasei.

Daroma prielaida, kad stebimas atsitiktinis vektorius X priklauso nuo latentinio atsitiktinio dydžio v, įgyjančio reikšmes 1, …, q, ir kad X sąlyginis pasiskirstymo tankis $f_i(x)$ esant sąlygai $\{v = i\}$, yra vienamodalinis $i = \overline{1, q}$. Jei klasterizavus imtį galioja (2.1), tai pasiskirstymo tankiui $f(x)$ vertinti naudojama lygybė

$$\hat{f}(x) = \sum_{i=1}^{q} \hat{p}_i \hat{f}_i(x). \tag{3.2}$$

Pakeitus *aposteriorines* tikimybes $\pi_i(x) = \mathbf{P}\{v = i | X = x\}$ jų įverčiais $\hat{\pi}_i(x)$ ir naudojant išraišką $p_i = \mathbf{E}\pi_i(X)$, turima:

$$\hat{p}_i = \frac{1}{n} \sum_{t=1}^{n} \hat{\pi}_i(X(t)),\ i = 1, \ldots, q.$$

Vertinant mišinio (3.2) komponentą $f_i(x)$ algoritmu AKDE, stebiniai $Z(t)$, kaip ir $X(t)$, yra nagrinėjami su svoriais $\hat{\pi}_i(t) = \hat{\pi}_i(Z(t)) = \hat{\pi}_i(X(t))$, $t = 1, \ldots, n$. Todėl (1.13) formulė keičiama į

$$\hat{f}_i(z) = \frac{1}{\hat{p}_i n} \sum_{t=1}^{n} \frac{\hat{w}_i(t)}{h^d \lambda_t^d} K\left(\frac{z - Z(t)}{h\lambda_t} \right). \tag{3.3}$$

Atitinkamai $f_i(x)$ vertinti algoritmu SKDE svoriai (1.22), kurie naudojami vertinant sąlyginį vidurkį (1.21) ir sąlyginę kovariacinę matricą (1.23), keičiami į

$$\hat{w}_{i,H_2}(y - Y(t)) = \frac{\hat{w}_i(t) K_{H_2}(y - Y(t))}{\sum_{j=1}^{n} \hat{w}_i(j) K_{H_2}(y - Y(j))}. \tag{3.4}$$

Analogiškai, naudojant IFDE algoritmą $f_i(x)$ vertinti, kiekvienoje projekcijoje $\tau \in T_0$ (1.67) formulė keičiama į

$$\hat{f}_{i,\tau}(v) = \frac{1}{\hat{p}_i n} \sum_{t=1}^{n} \frac{\hat{w}_i(t)}{h_t} \varphi\left(\frac{v - \tau' X(t)}{h_t} \right), \; h_t = h_t(\tau) \tag{3.5}$$

su atitinkama charakteristine funkcija $\hat{w}_\tau$, kuri naudojama įverčiui (1.69) užrašyti.

Atitinkamai, vertinant mišinio komponentą $f_i(x)$ algoritmu PPDE, stebiniai $Y = \tau' Z(t)$ nagrinėjami su svoriais $\hat{w}_i(t) = \hat{w}_i(Z(t))$, $t = 1, \ldots, n$. Todėl taikant (1.100) lygybę pasiskirstymo tankiui $f_i(x)$ vertinti, (1.99) formulė keičiama į

$$\hat{f}_{i,\tau(k)}(y) = \varphi(y) \sum_{j=0}^{s} \frac{2j+1}{\hat{p}_i n} \sum_{t=1}^{n} \hat{w}_i(t) \psi_j(r_t^{(k-1)}) \psi_j(r^{(k-1)}). \tag{3.6}$$

Taigi projektavimo indeksas (1.90) keičiamas į

$$I_i(\tau) = \sum_{j=1}^{s} \frac{2j+1}{2\hat{p}_i^2 n^2} \sum_{t=1}^{n} \left(\hat{w}_i(t) \psi_j(r_t) \right)^2, \tag{3.7}$$

o jo gradientas (1.92) į

$$\frac{\partial I_i}{\partial \tau} = \frac{2}{\sqrt{2\pi}} \sum_{j=1}^{s} \frac{2j+1}{\hat{p}_i^2 n^2} \sum_{t=1}^{n} \hat{w}_i(t) \psi_j(r_t) \sum_{l=1}^{n} \hat{w}_i(l) \psi_j'(r_l) e^{-y^2/2} (z - \tau y). \tag{3.8}$$

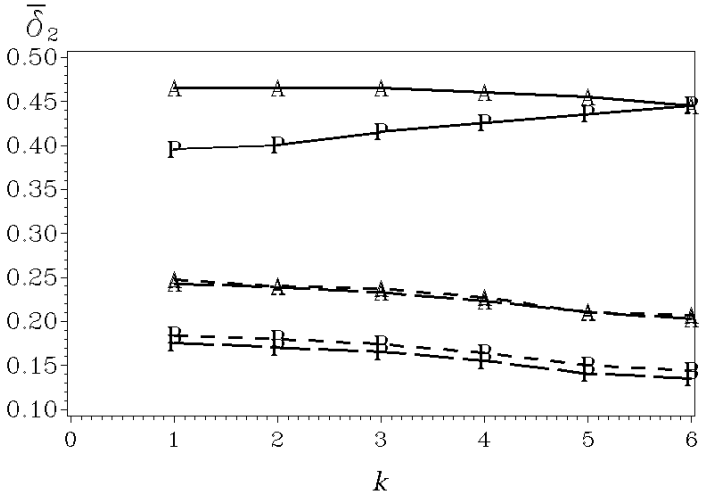
Įvertinių palyginimas. Siūlomų metodų tikslumas buvo lyginamas 1.3 skyriuje naudotų duomenų modelių A1–A6 Koši ir Gauso skirstinių mišinių pagrindu taikant (1.101) ir (1.102) paklaidas. Kiekvieno tipo duomenims generuotos įvairaus didumo imtys (50, 100, 200, 400, 800). Atlikta lyginamoji kelių populiarių neparametrinių įvertinių tikslumo analizė parodė, kad daugeliu nagrinėtų atvejų Friedman procedūra buvo efektyviausia ir tik kai imties tūris buvo mažas, kiek tikslesnis buvo branduolinis įvertinys. Todėl šiame skyriuje, atlikus imties klasterizavimą (taikant Gauso skirstinių mišinio modelį ir EM algoritmą), mišinio komponentai buvo vertinami dviem būdais – naudojant branduolinį įvertinį su adaptuotai parenkamu glodinimo pločiu bei Friedman pasiūlytą algoritmą, kuris remiasi tiksliniu projektavimu.

Atlikto kompiuterinio eksperimento rezultatai [1A] visiškai patvirtino 2.4 skyriaus išvadą [3A] apie klasterizavimo tikslingumą. Abiejų paklaidų δ_1 ir δ_2 priklausomybės nuo imties dydžio, klasterizavimo tipo ir atstumo tarp vertinamo tankio modų viršūnių buvo panašios (kokybiniu požiūriu).

Vertinant δ_1 paklaidą penkiamačiu Koši skirstinių mišinių atveju, kai imties didumas yra 50 ir 100, kai neatliekamas pirminis imties suskaidymas į klasterius, geriausi rezultatai gauti adaptuotu branduoliniu metodu, o didesnių imčių atvejais – tikslinio projektavimo metodu. Atlikus pirminį imties suskaidymą į klasterius, paklaidos gautos mažesnės ir visais atvejais išryškėjo tikslinio projektavimo metodo pranašumas. Taikant negriežtą klasterizavimą, paklaidos yra šiek tiek (~5 %) mažesnės nei taikant griežtą klasterizavimą, o atlikus perklasterizavimą ir patikslinus klasterius, tankių vertinimo rezultatai, nors ir neryškiai (~1–2 %), bet pagerėja. Didėjant atstumui tarp mišinio komponentų centrų, paklaidos mažėja, jei atliekamas pirminis klasterizavimas. Kai mišinį sudaro du klasteriai, paklaidos didėja, jei didinamas mišinio komponentų proporcijų (svorių) skirtumas. Didinant mišinį sudarančių komponentų skaičių, paklaidos didėja. Kai trijų klasterių mišinį sudaro labai persidengiantys komponentai ($k = 1$), geriausi rezultatai gauti, kai trečiasis komponentas yra mažiausias, blogesni – kai visi komponentai yra vienodų svorių, ir blogiausi – kai trečiasis komponentas yra didžiausias. Komponentus labiau atskiriant ($k \geq 2$), geriausi rezultatai gaunami, kai jie visi yra vienodo svorio. Kai mišinį sudaro keturi komponentai, mažiausios paklaidos gautos turint vienodų svorių klasterius, didžiausios – kai ketvirtasis komponentas yra didžiausias. Atlikus pirminį imties suskaidymą tiek trijų, tiek keturių mišinį sudarančių komponentų atveju, geriausi rezultatai gauti turint vienodų svorių komponentus, blogiausi – kai vienas komponentas yra didesnis už kitus.

Skaičiuojant δ_2 paklaidą penkiamačių labai ir mažai persidengiančių Koši skirstinių atvejais ($k \leq 4$), kai imtys ne didesnės kaip 200, geriausi rezultatai gauti adaptuotu branduoliniu tankio vertinimo metodu, o atsiskyrusių skirstinių – tikslinio projektavimo metodu. Kitos paklaidos kitimo tendencijos yra tokios pat kaip ir vertinant δ_1 paklaidą. 3.3 paveiksle pateiktas

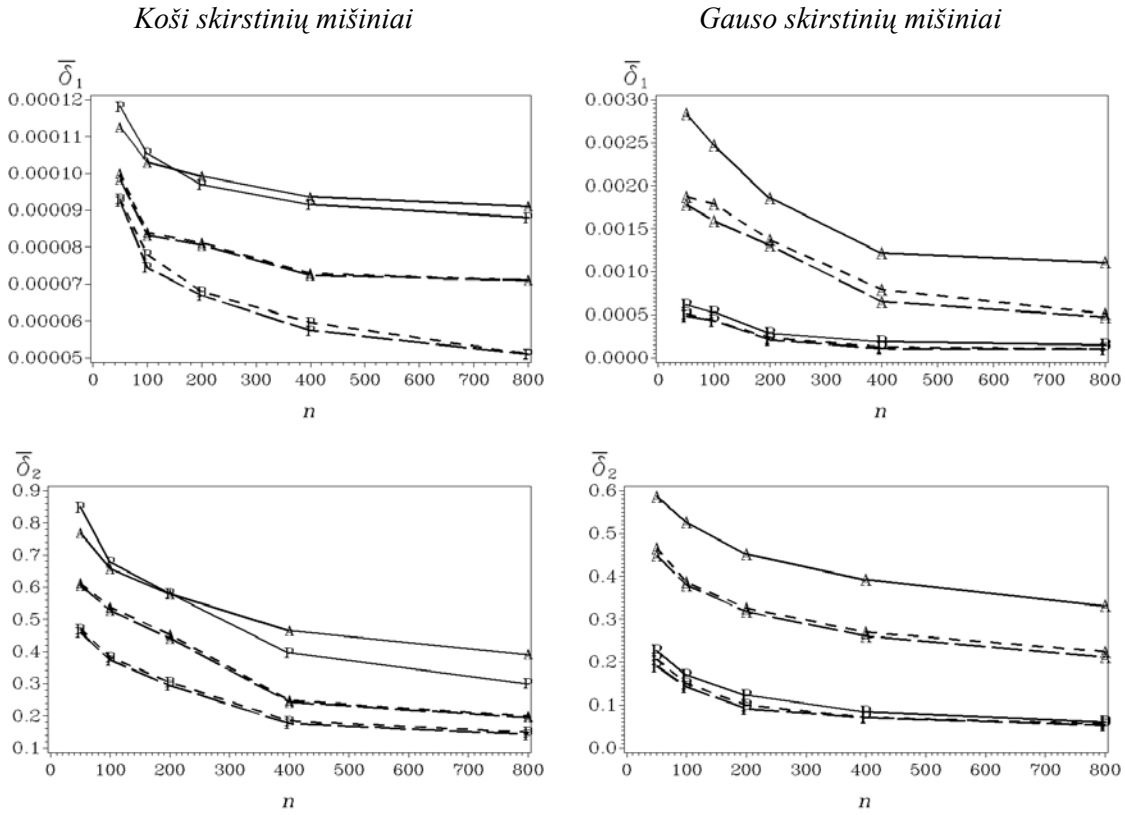
paklaidų kitimo priklausomai nuo atstumo tarp mišinio komponentų centrų pavyzdys (simboliais A ir P atitinkamai žymimi AKDE ir PPDE algoritmai; ištisine linija žymimos paklaidos neatlikus pirminio duomenų klasterizavimo, trumpa punktyrine – atlikus pirminį griežtą duomenų klasterizavimą, o ilga punktyrine – atlikus pirminį negriežtą duomenų klasterizavimą). Esant mažoms imtims ($n = 50$), tankių vertinimo δ_2 paklaidos siekia 0,7–0,9, atlikus pirminį imties suskaidymą į klasterius, jos sumažėja iki 0,35–0,7. Esant didelėms imtims ($n = 800$), δ_2 paklaidos yra 0,3–0,5, o atlikus pirminį imties suskaidymą į klasterius, jos sumažėja iki 0,1–0,3. Pirminiam imties suskaidymui, kai taikomas negriežtas klasterizavimas, mažoms imtims ($n = 50$, 100) gaunamos apie 2–5 % mažesnės paklaidos, didelėms ($n = 800$) apie 5–10 %, palyginti su paklaidomis, gautomis taikant griežtą klasterizavimo procedūrą, nors skirtumas tarp šių paklaidų kinta priešingai – mažoms imtims jis yra didesnis nei didelėms imtims. Perklasterizavimo įtaka buvo tokia pat esant tiek mažoms, tiek didelėms imtims, tačiau jį naudoti tikslingiau, kai mišinio komponentai yra ne vienodų, o skirtingų svorių – atitinkamai mažoms iki 1 % ir didelėms iki 0,6 %.



3.3 pav. Paklaidų priklausomybė nuo atstumo tarp mišinio komponentų centrų (Koši tankių mišinio modelis 2: $n = 400$; $d = 5$; $p_1 = p_2 = 0{,}45$, $p_3 = 0{,}1$)

Dvimačių Koši skirstinių mišinių atveju skaičiuojant abi δ_1 ir δ_2 paklaidas, kai imtys ne didesnės kaip 200, geriausi rezultatai gauti adaptuotu branduoliniu tankių vertinimo metodu. Kitos paklaidų kitimo tendencijos yra tokios pat kaip ir penkiamačiu atveju. Dvimačių mišinių atveju δ_2 paklaidos gautos mažesnės, palyginti su apskaičiuotomis vertinant penkiamačius mišinius, kartu mažėja ir pirminio imties suskaidymo į klasterius įtaka. Negriežto klasterizavimo pranašumas, palyginti su griežtu, esant mažoms imtims ($n = 50$), yra iki 3 %, esant didelėms imtims ($n = 800$), – iki 6 %. Perklasterizavimas negriežto klasterizavimo rezultatus pagerina tik iki 0,5 %

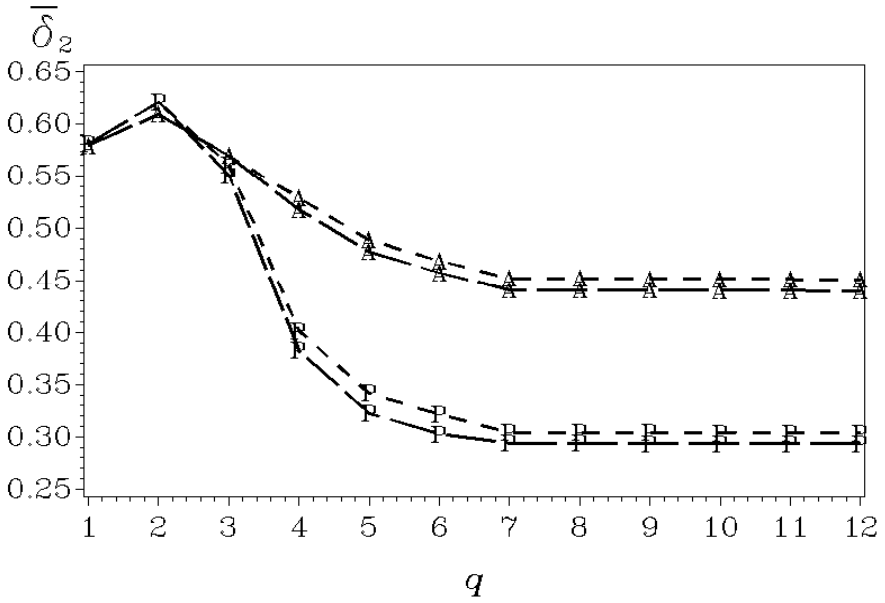
Vertinant dvimačių ir penkiamačių Gauso skirstinių mišinių tankius, geriausia taikyti tikslinio projektavimo metodą. Šiems mišiniams galioja analogiškos tendencijos kaip ir Koši skirstinių mišiniams, išskyrus tai, kad perklasterizavimas paklaidų nesumažino. Esant mažoms imtims ($n = 50$), tankių vertinimo δ_2 paklaidos siekia 0,4–0,7, po pirminio imties suskaidymo į klasterius jos sumažėja iki 0,2–0,5. Esant didelėms imtims ($n = 800$), δ_2 paklaidos yra 0,3–0,5, o po pirminio imties suskaidymo į klasterius jos sumažėja iki 0,05–0,3. 3.4 paveiksle pateikti trimodalinių Koši ir Gauso skirstinių mišinių δ_1 ir δ_2 paklaidų tipiniai pavyzdžiai (simboliais A ir P atitinkamai žymimi AKDE ir PPDE algoritmai; ištisine linija žymimos paklaidos neatlikus pirminio duomenų klasterizavimo, trumpa punktyrine – atlikus pirminį griežtą duomenų klasterizavimą, o ilga punktyrine – atlikus pirminį negriežtą duomenų klasterizavimą).



3.4 pav. Paklaidų priklausomybė nuo imties didumo (Koši ir Gauso skirstinių mišinių modelis A2: $d = 5$; $p_1 = p_2 = 0{,}45$, $p_3 = 0{,}1$; $k = 1$)

3.3. Pasiskirstymo tankio vertinimo tikslumo priklausomybė nuo pasirinkto klasterių skaičiaus

Klasterių skaičiaus parinkimas yra neatsiejama klasterizavimo procedūros dalis. Iš 2.2 skyriuje pasiūlytų algoritmų, skirtų tikimybinio klasterizavimo procedūrai, taikančiai EM algoritmą, pasiskirstymo tankio vertinimo priklausomybei nuo klasterių skaičiaus tirti buvo pasirinktas kriterijus, paremtas tikėtinumo funkcijos prieaugiu. Tyrimas atliktas 1.3 ir 3.2 skyriuose naudotų duomenų modelių A1–A6 Koši ir Gauso skirstinių mišinių pagrindu, naudojant (1.101) ir (1.102) paklaidas, o imtį klasterizavus, mišinio komponentai įvertinti naudojant branduolinį įvertinį su adaptuotai parenkamu glodinimo pločiu bei Friedman tikslinio projektavimo procedūrą. Naudojant kiekvieno tipo duomenis, generuotos įvairaus didumo imtys (50, 100, 200, 400, 800).



3.5 pav. Paklaidų priklausomybė nuo parinkto klasterių skaičiaus (Koši tankių mišinio modelis A2: $n = 200$; $d = 5$; $p_1 = p_2 = 0{,}45$, $p_3 = 0{,}1$; $k = 6$)

Analizuojant modeliavimo rezultatus priklausomybės nuo klasterių skaičiaus aspektu, pastebėta, jog Gauso skirstinių mišiniams klasterių skaičiaus įverčio reikšmė dažniausiai sutampa su tikruoju komponentų skaičiumi. Kai tarp q–1 vienodo didumo komponentų yra vienas nedidelio svorio komponentas, klasterių skaičiaus įverčio reikšmė buvo vienetu mažesnė nei tikrasis komponentų skaičius. Koši skirstinių mišiniams klasterių skaičiaus įverčio reikšmė buvo stebėta didesnė nei mišinio komponentų skaičius: esant mažoms imtims ($n = 50$), klasterių skaičius vienu–trimis, o esant didesnėms imtims ($n = 800$), – šešiais–aštuoniais viršijo tikrąjį mišinio komponentų skaičių. Tai galima paaiškinti tuo, kad stebinius mėginta aproksimuoti ne Koši, bet Gauso skirstinių mišiniais, taip pat kad tarp stebėtų duomenų pasitaikė daug išskirčių.

Tiriant, kaip paklaidos priklauso nuo klasterių skaičiaus, reikia paminėti, jog parinkus mažiau klasterių, nei yra komponentų, paklaidos padidėja: parinkus klasterių skaičių, mažesnį vienetu, paklaidos padidėja mažiau tuo atveju, kai vertinamas mišinys yra su skirtingų svorių Gauso skirstiniais, o šis padidėjimas siekia 20 % ar daugiau. Vienodų svorių Gauso skirstinių mišinių paklaidų padidėjimas siekia iki 2,5 karto, Koši skirstinių mišinių atveju klasterių skaičių sumažinus vienetu panaikinamas nedidelio svorio „išskirčių klasteris“, o kartu 5–20 % padidinamos tankių vertinimo paklaidos. Situacija pasikeičia, kai, mažinant klasterių skaičių, lieka du ar mažiau „išskirčių klasterių“, tuomet paklaidos išauga kaip ir Gauso skirstinių mišinių vertinimo atveju. Kai mišinį sudarančių ir tarpusavyje nutolusių Koši skirstinių yra daugiau nei du, o klasterius taip pat parinkus du, daugeliu atvejų paklaidos padidėjo palyginti su tuo atveju, kai nebuvo klasterizuojama ($q = 1$) (3.5 paveikslas; simboliais A ir P atitinkamai žymimi AKDE ir PPDE algoritmai; trumpa punktyrine linija žymimos paklaidos atlikus pirminį griežtą duomenų klasterizavimą, o ilga punktyrine – atlikus pirminį negriežtą duomenų klasterizavimą). Klasterių skaičių parinkus didesnį už komponentų skaičių, paklaidos šiek tiek sumažėjo (sukūrus tris papildomus klasterius, paklaidos sumažėdavo tik 0,05 %, palyginti su papildomais dviem klasteriais), tačiau papildomo klasterio svoris, palyginti su kitais, buvo nedidelis, o kiekvieno naujo klasterio atsiradimas prailgindavo papildomų kompiuterinių skaičiavimų trukmę [1A].

Darbo išvados

Disertacija skirta atsitiktinių vektorių daugiamodalinių pasiskirstymo tankių statistiniam vertinimui. Gauti tokie pagrindiniai rezultatai:

1. Daugiamodalinių pasiskirstymo tankių statistinio vertinimo rezultatai labai pagerėja, jei stebiniai pirmiausia klasterizuojami (traktuojant jų daugiamodalinį tankį kaip vienamodalinių tankių mišinį), o tankių vertinimo metodai yra taikomi kiekvienam klasteriui atskirai.

2. Daugeliu atvejų didžiausias tankių vertinimo efektyvumas buvo pasiekiamas, kai po pirminio imties suskaidymo kiekvieną klasterį atitinkantys tankio komponentai buvo įvertinti J. H. Friedman pasiūlyta rekurentine procedūra.

3. Parodyta, kad negriežtas imties klasterizavimas, kuris remiasi nagrinėjamo tankio aproksimacija Gauso pasiskirstymo tankių mišiniu ir EM algoritmu, yra pranašesnis nei griežtas ar kitos populiarios geometrinio klasterizavimo procedūros, kai klasterizavimo rezultatai taikomi daugiamodaliniams tankiams statistiškai vertinti.

4. *Bootstrap* metodu nustatomas klasterių skaičius yra artimas optimaliam.

Literatūros sąrašas

1. Abramson I., On Bandwidth Variation in Kernel Estimates – A Square Root Law, *The Annals of Statistics*, Vol. **10**, 1982, p. 1217–1223.
2. Aitkin I. S., Anderson D., Hinde J., Statistical modeling of data on teaching styles, *Journal of the Royal Statistical Society*, Series A, Vol. **144**, 1981, p. 419–461.
3. Akaike H., A new look at the statistical model identification. IEEE Trans. AC, Vol. **19**, 1974, p. 716–723.
4. Aladjem M., Projection Pursuit Mixture Density Estimation, *IEEE Transactions on Signal*, Vol. **53**, No. 11, 2005, p. 4376–4383.
5. Amemiya T., The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model, *Econometrica*, Vol. **45**, No. 4, 1977, p. 955–968.
6. Anderberg M. R., Cluster Analysis for Applications, *New York: Academic Press, Inc*., 1973.
7. Andrews D. W. K., Estimation when a parameter is on a boundary, *Econometrica*, Vol. **66**, 1999, p. 1341–1383.
8. Andrews D. W. K., The large sample correspondence between classical hypothesis tests and Bayesian posterior odds tests, *Econometrica*, Vol. **62**, No. 5, 1994, p. 1207–1232.
9. Andrews D. W. K., A stopping rule for the computation of generalized method of moments estimators, *Econometrica*, Vol. **65**, No. 4, 1997, p. 913–931.
10. Antoniadis A., Gregoire G., Nason G., Density and hazard rate estimation for right censored data using wavelet methods, *Journal of the Royal Statistical Society*, Series B, Vol. **61**, 1999, p. 63–84.
11. Archambeau C., Lee J. A. and Verleysen M., On the convergence problems of the EM algorithm for finite gaussian mixtures, *Artificial Neural Networks*, 2003, p. 99–106.
12. Archambeau C. and Verleysen M., Fully Nonparametric Probability Density Function Estimation with Finite Gaussian Mixture Models, *ICAPR*, 2003, p. 81–84.
13. Arthur D., Vassilvitskii S., How Slow is the k-means Method?, *Proceedings of the twenty-second annual symposium on Computational geometry*, 2006, p. 144–153.
14. Azzalini A. and Capitanio A., Statistical Applications of the Multivariate Skew Normal Distribution, *Journal of the Royal Statistical Society*, Series B, Vol. **61**, 1999, p. 579–602.
15. Basford K. E., and McLachlan G. J., Mixture Models: Inference and Applications to Clustering, *New York: Marcel Dekker*, 1988.

16. Base SAS 9.1.3 Procedures Guide, Second Edition, *Cary, NC, USA: SAS Institute Inc*., Volumes **1–4**, 2006, 1934 p.
17. Bashtannyk D. M. and Hyndman R. J., Bandwidth selection for kernel conditional density estimation, *Technical report, Department of Econometrics and Business Statistics, Monash University*, 1998.
18. Behboodian J., On a Mixture of Normal Distributions, *Biometrika*, Vol. **57**, 1970, p. 215–217.
19. Berger J. O., Bayesian Analysis: A Look at Today and Thoughts of Tomorrow, *New York: Chapman and Hall*, 2002, p. 275–290.
20. Berndt E., Hall B., Hall R., Hausman J., Estimation and inference in nonlinear structural models, *Annals of Economic and Social Measurement*, Vol. **3**, No. 4, 1974, p. 653–665.
21. Bock H. H., Probability models and hypotheses testing in partitioning cluster analysis, *Clustering and Classification*, 1996, p. 377–453.
22. Boneva L., Kendall D., Stafanov I., Spline transformations: Three new diagnostic aids for the statistical data analyst, *Journal of the Royal Statistical Society*, Vol. **33**, 1971, p. 1–70.
23. de Boor C., A practical guide to splines, *New York: Springer*, 1978.
24. Boyles R. A., On the convergence of the EM algorithm, *Journal of the Royal Statistical Society*, Series B, Vol. **45**, 1983, p. 47–50.
25. Bozdagan H., Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity, *Frontiers Statist. Modeling*, Vol. **2**, 1994, p. 69–113.
26. Breiman L., Meisel W., Purcel E., Variable kernel estimates of multivariate densities, *Technometrics*, Vol. **19**, 1977, p. 135–144.
27. Breiman L., Friedman J. H., Olshan R. A. and Stone C. J., Classification and Regression Trees, *Wadsworth & Brooks/Cole*, 1984.
28. Burke M. D. and Gombay E., The bootstrapped maximum likelihood estimator with an application, *Statistics and Probability Letters*, Vol. **12**, 1991, p. 421–427.
29. Burman P., A comparative study of ordinary cross-validation, v-fold cross-validati on and the repeated learning testing methods, *Biometrika*, Vol. **76**, 1989, p. 503–514.
30. Castellana J. V., Leadbetter M. R., On Smoothed Probability Density Estimation for Stacionary Processes*, Stochastic Processes and their Applications*, Vol. **21**, 1986, p. 179–193.
31. Chen S., Hong X. and Harris C. J., Sparse Kernel Density Construction Using Orthogonal Forward Regression With Leave-One-Out Test Score and Local Regularization, *IEEE Transactions on Systems*, Vol. **34**, No. 4, 2004, p. 1708–1717.
32. Cheng M. Y., Hall P., Turlach B. A., High-derivative parametric enhancements of nonparametric curve estimators, *Biometrika*, Vol. **86**, No. 2, 1999, p. 417–428.

33. Cheng M. Y., Choi E., Fan J., Hall P., Skewing-methods for two parameter locally-parametric density estimation, *Bernoulli*, Vol. **6**, 2000, p. 169–182.
34. Cheng M. Y., Hall P., Reducing variance in nonparametric surface estimation, *Journal of Multivariate Analysis*, Vol. **86**, 2003, p. 375–397.
35. Chernozhukov V., Hong H., An MCMC Approach to Classical Estimation. *Journal of Econometrics*, Vol. **115**. No. 2, 2003, p. 293–346.
36. Chiu S. T., Bandwidth selection for kernel density estimation, *The Annals of Statistics*, Vol. **19**, 1991, p. 1883–1905.
37. Ciesielski Z., Nonparametric polynomial density estimation, *Probability and Mathematical Statistics*, Vol. **9**, No. 1, 1988, p. 1–10.
38. Clayton D. G., A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrica*, Vol. **65**, 1978, p. 141–151.
39. Coppejans M. and Gallant A. R., Cross Validated SNP Density Estimates, *Manuscript*, 2000.
40. Cox D. D., Multivariate smoothing spline functions, *SIAM Journal of Numerical Analysis*, Vol. **21**, 1984, p. 789–813.
41. Cwik J., Koronacky J., Probability density estimation using a Gaussian clustering algorithm, *Neural Computing and Applications*, Vol. **4**, 1996, p. 149–160.
42. Cwik J., Koronacky J., Multivariate density estimation: A comparative study, *Neural Computing & Applications*, Vol. **6**, No. 3, 1997, p. 173–185.
43. Delgado M. A. and Robinson P. M., Nonparametric and semiparametric methods for economic research, *Journal of Economic Surveys*, Vol. **6**, No. 3, 1992, p. 201–249.
44. Delicato P., del Rio M., A generalization of histogram type estimators, *UPF Economics and Business Working Papers Series*, 422, 1999.
45. Dempster A. P., Laird N. M. and Rubin D. B., Maximum likelihood from incomplete data via the *EM* algorithm, *Journal of the Royal Statistical Society.* B, **39**, 1977, p. 1–38.
46. Devroye L., and Györfi L., Nonparametric Density Estimation : The L_1 View. *New York: Wiley*, 1985.
47. Diaconis P., Freedman D., On the consistency of bayes estimates, *Annals of Statistics*, Vol. **14**, 1986, p. 1–26.
48. Dias R., Density estimation via hybrid splines, *Journal of Statistical Computation and Simulation*, Vol. **60**, 1998, p. 277–294.
49. Dotan Y., Intrator N., Multimodality exploration in training an unsupervised projection pursuit neural network, *IEEE Neural Networks*, Vol. **9**, No. 3, 1996, p. 464–472.

50. Duchon J., Fonctions-Spline et Esperances Conditionnelles de Champs Gaussiens, *Ann. Sci. Univ. Clermont Ferrand II Math.*, Vol. **14**, 1976, p. 19–27.
51. Duchon J., Splines Minimizing Rotation-Invariant Semi-Norms in Sovolev Spaces, *Constructive Theory of Functions of Several Variables*, 1977, p. 85–100.
52. Duong T., Bandwidth matrices for multivariate kernel density estimation, *PhD thesis*, 2004, 161 p.
53. Duong T., and Hazelton M. L., Cross-validation bandwidth matrices for multivariate kernel density estimation, *Scandivanian Journal of Statistics*. Vol. **32,** 2005, p. 485–506.
54. Eilers P. H. C., Nonparametric density estimation with grouped observations, Statistica Neerlandica, Vol. **45**, 1991, p. 255–270.
55. Elgammal A., Duraiswami R., Harwood D., and Davis L. S., Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance, *IEEE*, Vol. **90**, No. 7, 2002.
56. Epanechnikov V. A., Nonparametric estimates of a multivariate probability density, *Theoretical Probability Applications*, Vol. **14**, 1969, p. 153–158.
57. Eubank R. L., Nonparametric Regression and Spline Smoothing, *New York: Marcel Dekker*, 1999.
58. Everitt B. S., Hand D. J., Finite Mixture Distributions, *New York: John Wiley*, 1981.
59. Everitt B. S., Landau S. and Leese M., Cluster analysis, *Oxford University Press, NY*, 2001.
60. Eğecioğlu Ö., Srinivasan A., A fast non-parametric density estimation algorithm, *Comunications in Numerical Methods and Engeenering*, Vol. **13**, 1997, p. 755–763.
61. Eğecioğlu Ö., Srinivasan A., Efficient nonparametric density estimation on the sphere with applications in fluid mechanics, *Siam j. sci. comput*., Vol. **22**, No. 1, 2000, p. 152–176.
62. Fadda D., Slezak E. and Bijaoui A., Density estimation with non-parametric methods, *Astron. Astrophys. Suppl. Ser*., Vol. **127**, 1998, p. 335–352.
63. Fan J., Heckman N. E. and Wand M. P., Local polynomial kernel regression for generalized linear models and quasi-likelihood functions, *Journal of the American Statistical Association*, Vol. **90**, 1995, p. 141–150.
64. Fan J. and Marron J. S., Best possible constant for bandwidth selection, *Annals of Statistics*, Vol. **20**, 1992, p. 2057–2070.
65. Fan J. and Gijbels I., Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation, *Journal of the Royal Statistical Society*, B, **57**, 1995, p. 371–394.
66. Fenstad G. U. and Hjort N. L., Two Hermite expansion density estimators, and a comparison with the kernel method, *Unpublished manuscript*, 1996.

67. Fenton V. M. and Gallant A. R., Qualitative and Asymptotic Performance of SNP Density Estimators, *Journal of Econometrics*, Vol. **74**, 1996a, p. 77–118.

68. Fenton V. M. and Gallant A. R., Convergence Rates of SNP Density Estimators, *Econometrica*, Vol. **64**, 1996b, p. 719–127.

69. Fraley C., Algorithms for Model-Based Gaussian Hierarchical Clustering, *SIAM Journal on Scientific Computing*, Vol. **20**, 1998, p. 270–281.

70. Freedman D., and Diaconis P., On the histogram as a density estimator: *L*2 theory. *Z. Wahrscheinlichkeit-stheorie verw.*, Gebeite, Vol. **57**, 1981, p. 453–476.

71. Friedman J. H., Turkey J. W., A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions on Computers*, Ser. C, **23**, 1974, p. 881–889.

72. Friedman J. H., Stuetzle W., Schroeder A., Projection pursuit density estimation, *Journal of the American Statistical Association*, Vol. **79**, 1984, p. 599–608.

73. Friedman J. H., Exploratory projection pursuit, *Journal of the American Statistical Association*, Vol. **82**, No**.** 397, 1987, p. 249–266.

74. Friedman J. H., Flexible metric nearest neighbor classification, *Tech. Report, Dept. of Statistics, Stanford University*, 1994.

75. Fryer M. J., A review of some non-parametric methods of density estimation, *J. Inst. Math. Applic*., Vol. **20**, 1977, p. 335–354.

76. Fukunaga K., Introduction to Statistical Pattern Recognition. *New York*: *Academic Press*, 1972.

77. Fukunaga K., and Hostetler L. D., The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, Vol. **21**, 1975, p. 32–40.

78. Gallant A. R., Nychka D. W., Semi-nonparametric Maximum Likelihood Estimation, *Econometrica*, Vol. **55**, No. 2, 1987, p. 363–390.

79. Gasser T., Müller, H. G., and Mammitzsch V., Kernels for nonparametric curve estimation, *Journal of the Royal Statistical Society*, B, **47**, 1985, p. 238–252.

80. Gill R. D., Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part I), *Scandinavian Journal of Statistics*, Vol. **16**, 1989, p. 97–128.

81. Gill R. D., and van der Vaart A. W., Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part II), *Scandinavian Journal of Statistics*, Vol. **20**, 1993, p. 271–288.

82. Gitman I., An Algorithm for Nonsupervised Pattern Classification, *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-**3**, 1973, p. 66–74.

83. Goffe W. L., Ferrier G. D., Rogers J., Global optimization of statistical functions with simulated annealing, *Journal of Econometrics*, Vol. **60**, 1994, p. 65–99.

84. Gordon A. D., Cluster validation, *Tokyo: Springer-Verlag*, 1998, p. 22–39.

85. Gu C. and Qiu C., Smoothing spline density estimation: theory, *Annals of Statistics*, Vol. **21**, 1993, p. 217–234.

86. Guo P., Chen C. L. P., and Lyu M. R., Cluster Number Selection for a Small Set of Samples Using the Bayesian Ying–Yang Model, *IEEE Transactions On Neural Networks*, Vol. **13**, No. 3, 2002.

87. Hafner C. M., Rombouts J. V. K., Semiparametric multivariate GARCH models, *Forthcoming in Econometric Theory*, 2006.

88. Hall P., Large sample optimality of least squares cross-validation in density estimation, *Annals of Statistics*, Vol. **11**, 1983, p. 1156–1174.

89. Hall P., On polynomial-based projection indices for exploratory projection pursuit, *Annals of Statistics*, Vol. **17**, No. 2, 1989, p. 589–605.

90. Hall P., Sheather S. J., Jones M. C. and Marron J. S., On optimal data-based bandwidth selection in kernel density estimation, *Biometrika*, Vol. **78**, 1991, p. 263–269.

91. Hall P., The bootstrap and edgeworth expansion, *New York: Springer*, 1992.

92. Hall P., Huber C., Owen A., Coventry A., Asymptotically optimal balloon density estimates, *Journal of Multivariate Analysis*, Vol. **51**, 1994, p. 352–371.

93. Hall P., Presnel B., Density estimation under constraints, *Journal of Computational and Graphical Statistics*, Vol. **8**, No. 2, 1999, p. 259–281.

94. Hall P., Huang L. S., Nonparametric kernel regression subject to monotonicity constraints, *Akt*., Vol. **39**, 1999, p. 125–153.

95. Hammersley J. M. and Handscomb D. C., Monte Carlo Methods. *New York: Chapman and Hall*, 1964.

96. Hands S. and Everitt B. S., A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques, *Multivariate Behavioral Research*, Vol. **22**, 1987, p. 235–243.

97. Hansen L. P., Large sample properties of generalized method of moments estimators, *Econometrica*, Vol. **50**, No. 4, 1982, p. 1029–1054.

98. Hansen L. P., Heaton J., Yaron A., Finite-sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics*, Vol. **14**, 1996, p. 262–280.

99. Hansen M. H., Kooperberg C., Spline Adaptation in Extended Linear Models, *Statistical Science*, Vol. **17**, No. 1, 2002, p. 2–20.

100. Hart J. D., Vieu P., Data-Driven Bandwidth Choice for Density Estimation on Dependent data, *Annals of Statistics*, Vol. **18**, 1990, p. 873–890.

101. Hartigan J. A., Clustering Algorithms, *New York: John Wiley & Sons, Inc*., 1975.

102. Hasselbland V., Estimation of Parameters for a Mixture of Normal Distributions, *Technometrics*, Vol. **8**, 1966, p. 431–444.

103. Hyndman R. J. and Yao Q., Nonparametric estimation and symmetry tests for conditional density functions, *Working paper 17/98, Department of Econometrics and Business Statistics, Monash University*, 1998.

104. Hjort N. L., On frequency polygons and average shifted histograms in several dimensions, *Technical Report, Stanford University*, Vol. **22**, 1986.

105. Hjort N. L., Semiparametric estimation of parametric hazard rates, *Survival Analysis: State of the Art*, 1991, p. 211–236.

106. Hjort N. L. and Glad I. K., Nonparametric density estimation with a parametric start, *Annals of Statistics*, Vol. **23**, 1995, p. 882–904.

107. Hjort N. L., and Jones M. C., Locally Parametric Nonparametric Density Estimation, *The Annals of Statistics*, Vol. **24**, No. 4, 1996, p. 1619–1647.

108. Hoti F., Holmström L., Application of Semiparametric Density Estimation to Classification. *ICPR*, Vol. **3**, 2004, p. 371–374.

109. Hotelling H., Multivariate quality control. *In: Eisenhart C., Hastay M.W., Wallis W.A., editors. Techniques of statistical analysis. New York: McGraw-Hill*, 1947, p. 111–84.

110. Huber P. J., Robust regression: Asymptotics, conjectures, and Monte Carlo, *Annals of Statistics*, Vol. **1**, 1973, p. 799–821.

111. Huber P. J., Projection pursuit, The Annals of Statistics, Vol. **13**, No. 2, 1985, p. 435–475.

112. Huizinga D. H., A Natural or Mode Seeking Cluster Analysis Algorithm, *Technical Report 78–1, Behavioral Research Institute, 2305 Canyon Blvd., Boulder, Colorado 80302*, 1978.

113. Hwang J. N., Lay S. R. and Lippman A., Nonparametric Multivariate Density Estimation: A Comparative Study, *IEEE Transactions on Signal Processing*, Vol. **42**, No. 10, 1994, p. 2795–2810.

114. Härdle W. and Müller M., Multivariate and semiparametric kernel regression, *New York: Wiley*, 2000, p. 357–391.

115. Izenman A. J., Recent developments in nonparametric density estimation, *Journal of the American Statistical Association*, Vol. **86**, No. 413, 1991, p. 205–224.

116. Jain A. K. and Dubes R. C., *Algorithms for Clustering Data*. Prentice Hall, 1988.

117. Jeon B. and Landgrebe D. A., Fast Parzen Density Estimation Using Clustering-Based Branch and Bound, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. **16**, No. 9, 1994, p. 950–954.

118. Jones M. C., Discretized and interpolated kernel density estimates, *Journal of the American Statistical Association*, Vol. **84**, 1989, p. 733–741.

119. Jones M. C., Potential for automatic bandwidth choice in variations of kernel density estimation, *Statistics and Probability Letters*, Vol. **13**, 1992, p. 351–356.

120. Jones M. C., Kernel density estimation when the bandwidth is large, *Austral. J. Statist*., Vol. **35**, 1993, p. 319–326.

121. Jones M. C., Simple boundary correction for kernel density estimation, *Statistics and Computing*, Vol. **3**, 1993, p. 135–146.

122. Jones M. C. and Foster P. J., Generalized jackkning and higher order kernels, *Journal of Nonparametric Statistics*, Vol. **3**, 1993, p. 81–94.

123. Jones M. C., On kernel density derivative estimation, *Comm. Statist. Theory Methods*, Vol. **23**, 1994, p. 2133–2139.

124. Jones M. C., Davies S. J. and Park B. U., Versions of kernel-type regression estimators, *Journal of the American Statistical Association*, Vol. **89**, 1994, p. 825–832.

125. Jones M. C., On close relations of local likelihood density estimation, *Unpublished manuscript*, 1995.

126. Jones M. C., Marron J. S. and Sheather S. J., A Brief Survey of Bandwidth Selection for Density Estimation, *Journal of the American Statistical Association*, Vol. **91**, 1996, p. 401–407.

127. Jones M. C. and Foster P. J., A simple nonnegative boundary correction method for kernel density estimation, *Statistica Sinica*, Vol. **6**, 1996, p. 1005–1013.

128. Jones M. C., Samiuddin M., Al-Harbey A. H. and Maatouk T. A. H., The edge frequency polygon, *Biometrika*, Vol. **85**, 1998, p. 235–239.

129. Jones M. C., Signorini D. F., Hjort N. L., On multiplicative bias correction in kernel density estimation, *The Indian Journal of Statistics*, Vol. **61**, 1999, p. 422–430.

130. Jordan M. I., Xu L., Convergence results for the EM approach to mixtures of expert architectures, *Neural Networks*, Vol. **8**, 1995, p. 1409–1431.

131. Karypis G., Han E. H. and Kumar V., Chameleon: A hierarchical clustering algorithm using dynamic modeling, *IEEE Computer*, Vol. **32**, No. 8, 1999, p. 68–75.

132. Kavaliauskas M., Rudzkis R., Projection-based Estimation of Multivariate Distribution Density*, Lietuvos matematikos rinkinys*, **42**(spec. nr.), 2002, p. 529–536.

133. Kavaliauskas M., Daugiamačių Gauso skirstinių mišinio statistinė analizė, taikant duomenų projektavimą, *daktaro disertacija*, 2005, 88 p.

134. Khan, S., Powell, J.L., Two step estimation of semiparametric censored regression models, *Journal of Econometrics*, Vol. **103**, 2001, p. 73–110.

135. King B., Step-wise clustering procedures, *Journal of the American Statistical Association*, Vol. **69**, 1967, p. 86–101.

136. Koo J. Y., Bivariate B-splines for tensor logspline density estimation, *Computational Statistics and Data Analysis*, Vol. **21**, 1996, p. 31–42.

137. Kooperberg C. and Stone C. J., A Study of Logspline Density Estimation, *Computational Statistics and Data Analysis*, Vol. **12**, 1991, p. 327–347.

138. Kooperberg C. and Stone C. J., Logspline Density Estimation for Censored Data, *Journal of Computational and Graphical Statistics*, Vol. **1**, 1992, p. 301–328.

139. Kooperberg C., Bivariate density estimation with an application to survival analysis, *Journal of Computational and Graphical Statistics*, Vol. **7**, 1998, p. 322–341.

140. Krus D. J. and Fuller E. A., Computer-assisted multicross-validation in regression analysis, *Educational and Psychological Measurement*, Vol. **42**, 1982, p. 187–193.

141. Kurtz A. K., A research test of Rorschach test, *Personnel Psychology*, Vol. **1**, 1948, p. 41–53.

142. van der Laan M. J., Efficient and Inefficient Estimation in Semiparametric Models, *Technical Report, CWI Amsterdam*, 1996.

143. van der Laan M. J., Dudoit S. and Keles S., Asymptotic Optimality of Likelihood-Based Cross-Validation, *Statistical Applications in Genetics and Molecular Biology*: Vol. **3**: No. 1, Article 4, 2004.

144. Lambert C., Harrington S., Harvey C., and Glodjo A., Efficient on-line nonparametric kernel density estimation, *Algorithmica*, Vol. **25**, 1999, p. 37–57.

145. Law A., Kelton D., Simulation modelling and analysis, *New York: McGraw Hill*, 1991.

146. Leger Ch. and Politis D. N., Bootstrap technology and applications, *Technometrics*, Vol. **34**, 1992, p. 378–398.

147. Likhterov N. and Aladjem M., Two Dimensional Projection Pursuit Applied to Gaussian Mixture Model Fitting, *Journal of Systemics, Cybernetics and Informatics*, Vol. **1**, No. 4, 2003.

148. MacQueen J. B., Some Methods for Classification and Analysis of Multivariate Observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. **1**, 1967, p. 281–297.

149. Marron J. S., An asymptotically efficient solution to the bandwidth problem of kernel density estimation, *Annals of Statistics*, Vol. **13**, 1985, p. 1011–1023.

150. Marron J. S., Nolan D., Canonical kernels for density estimations, *Statistics and Probability Letters*, Vol. **7**, No. 3, 1988, p. 195–199.

151. Marron J. S., and Ruppert D., Transformations to reduce boundary bias in kernel density estimation, *Journal of the Royal Statistical Society*, B, **56**, 1994, p. 653–671.

152. McLachlan G. J., Krishnan T., The EM Algorithm and Extensions, *New York: John Wiley*, 1997.

153. McLachlan G. and Peel D., Finite Mixture Models, *New York: John Wiley*, 2000.

154. Meinguet J., Multivariate Interpolation at Arbitrary Points Made Simple, *J. Appl. Math. Phys. (ZAMP)*, Vol. **5**, 1979, p. 439–468.

155. Minnote M. C., Achieving higher-order convergence rates for density estimation with binned data, *Journal of the American Statistical Association*, Vol. **93**, 1998, p. 663–672.

156. Mojena R., Hierarchical grouping methods and stopping rules: An evaluation, *Computer Journal*, Vol. **20**, 1977, p. 359–363.

157. Mosier C. I., Problems and designs of cross-validation, *Educational and Psychological Measurement,* Vol. **11**, 1951, p. 5–11.

158. Nadaraya E. A., On estimating regression, *Theoretical Probability Applications*, Vol. **9**, 1964, p. 141–142.

159. Nadaraya E. A., On nonparametric estimates of density functions and regression curves, *Theoretical Probability Applications*, Vol. **10**, 1965, p. 186–190.

160. Nielsen J. P., Multiplicative bias correction in kernel hazard estimation, *Scandinavian Journal of Statistics*, Vol. **11**, 1998, p. 453–466.

161. Olkin I. and Spiegelman C. H., A semiparametric approach to density estimation, *Journal of the American Statistical Association*, Vol. **82**, 1987, p. 858–865.

162. Park B. U., Marron J. S., Comparison of Data-Driven Bandwidth Selectors, *Journal of the American Statistical Association*, Vol. **85**, 1990, p. 66–72.

163. Parzen E., On the estimation of probability density and mode. *Ann. Math. Statist*., **33**, 1962, p. 1065–1076.

164. Pavlic M. and van der Laan M. J., Fitting of mixtures with unspecified number of components using cross validation distance estimate, *Computational Statistics and Data Analysis*, Vol. **41**, 2003, p. 413–428.

165. Peters C. A., Valafar F., Comparison of Three Nonparametric Density Estimation Techniques Using Bayes' Classifiers Applied to Microarray Data Analysis, *METMBS*, 2003, p. 119–125.

166. Priebe C. E., Adaptive mixtures, *Journal of the American Statistical Association*, Vol. **89**, 1994, p. 796–806.

167. Radavicius M., Rudzkis R., Consistent Estimation of Discriminant Space in Mixture Model by Using Projection Pursuit, *Proceedings of 7th Vilnius Conference in Probability Theory and Mathematical Statistics*, 1998, p. 617–626.

168. Raftery A. E., Bayesian model selection in social research (with discussion), *Sociological Methodology*, Vol. **25**, 1995, p. 111–193.

169. Render R. A., Walker H. F., Mixture densities, maximum likelihood and the EM algorithm, *SIAM Rev*., Vol. **26**, 1984, p. 195–239.

170. Robinson P. M., Nonparametric Estimators for Time Series, *Journal of Time Series Analysis*, Vol. **4**, 1983, p. 185–207.

171. Rosenblatt M., Remarks on some nonparametric estimates of a density function, *Ann. Math. Statist.*, Vol. **27**, 1956, p. 832–837.

172. Rudzkis R., Radavicius M., Statistical Estimations of a Mixture of Gaussian Distributions. *Acta Applicandae Mathematicae*, Vol. **38***,* 1995, p. 37–54.

173. Rudzkis R., Radavicius M., Testing Hypotheses on Discriminant Space in the Mixture Model of Gaussian Distributions. *Acta Applicandae Mathematicae*, Vol. **79***,* 2003, p. 105–114.

174. Ruppert D., Cline D. B. H., Bias reduction in kernel density estimation by smoothed empirical transformations, *The Annals of Statistics*, Vol. **22**, 1994, p. 185–210.

175. Sain S. R., Baggerly K. A., Scott D. W., Cross-validation of multivariate densities, *Journal of the American Statistical Association*, Vol. **89**, No. 427, 1994, p. 807–817.

176. Sain R. R. and Scott D. W., On Locally Adaptive Density Estimation, *Journal of the American Statistical Association*, Vol. **91**, 1996, p. 1525–1534.

177. Salvador R., and Ayanz J. S. M., An extension of a nonparametric clustering algorithm to derive radiometrically homogeneous objects pointed by seeds, *Journal of the Remote Sensing*, Vol. **23**, No. 6, 2002, p. 1197–1205.

178. Samiuddin M. and El-Sayyad G. M., On nonparametric kernel density estimates, *Biometrika*, Vol. **77**, 1990, p. 865–874.

179. Sarle W. S., Cluster Analysis by Least Squares, *Proceedings of the Seventh Annual SAS Users Group International Conference*, 1982, p. 651–653.

180. Sarle W. S., The Cubic Clustering Criterion. *SAS technical report A-108, SAS Institute, Cary, NC*, 1983.

181. SAS 9.1 Macro Language: Reference, *Cary, NC, USA: SAS Institute Inc*., 2004, 348 p.

182. SAS/IML 9.1 User's Guide, *Cary, NC, USA: SAS Institute Inc*., Volumes **1** and **2**, 2004, 1040 p.

183. SAS/STAT 9.1 User's Guide, *Cary, NC, USA: SAS Institute Inc*., Volumes **1**–**7**, 2004, 5136 p.

184. SAS/GRAPH 9.1 Reference, *Cary, NC, USA: SAS Institute Inc*., Volumes **1**, **2**, and **3**, 2004, 1628 p.

185. Schumaker L. L., Spline Functions: Basic Theory, *New York: J. Wiley*, 1981.

186. Schuster E. and Yakowitz S., Parametric/nonparametric mixture density estimation with application to good-frequency analysis, *Water Resources Bulletin*, Vol. **21**, 1985, p. 797–804.

187. Scott D. W., On optimal and data-based histograms, *Biometrika*, Vol. **66**, 1979, p. 605–610.

188. Scott D. W. and Sheather S.J., Kernel density estimation with binned data, *Comm. Statist. Theory Meth.*, 14, 1985, p. 1353–1359.

189. Scott D. W., Wand M. P., Feasibility of multivariate density estimates, *Biometrika*, Vol. **78**, 1991, p. 197–205.

190. Scott D. W., Multivariate Density Estimation: Theory, Practice, and Visualization, *New York: John Wiley*, 1992.

191. Scott D. W., Averaged shifted histograms: effective nonparametric density estimators in several dimensions, *Annals of Statistics*, Vol. **13**, 1985, p. 1024–1040.

192. Scott D. W., Remarks on Fitting and Interpreting Mixture Models, *Computing Science*, Vol. **31**, 1999.

193. Severini T. A., Staniswalis J. G., Quasi likelihood estimation in semiparametric models, *Journal of the American Statistical Association*, Vol. **89**, 1994, p. 501–511.

194. Shapiro S. S., and Wilk M. B., An analysis of variance test for normality (complete samples), *Biometrika*, Vol. **52**, 1965, p. 591–611.

195. Sheather S. J. and Jones M. C., A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation, *Journal of the Royal Statistical Society*, B, **53**, 1991, p. 683–690.

196. Silverman B. W., Using Kernel Density Estimates to Investigate Multimodality, *Journal of the Royal Statistical Society*, Ser. B, **43**, 1981, p. 97–99.

197. Silverman B. W., Density Estimation for Statistics and Data Analysis, *London: Chapman and Hall*, 1986.

198. Simonoff J. S., The anchor position of histograms and frequency polygons: quantitative and qualitative smoothing, *Comm. Statist. Simul. Computat*., Vol. **24**, 1995, p. 691–710.

199. Simonoff J. S., Smoothing Methods in Statistics, *New York: Springer*, 1996.

200. Symons M., Clustering Criteria and Multivariate Normal Mixtures, *Biometrics*, Vol. **37**, 1981, p. 37–43.

201. Smyth P., Model selection of probabilistic clustering using cross-validated likelihood, *Statistics and Computing*, Vol. **10**, 2000, p. 63–72.

202. Sneath P. H. A. and Sokal R. R., Numerical Taxonomy, *San Francisco: Freeman*, 1973.

203. Steckley S. G., Henderson S. G., A Kernel Approach To Estimating The Density Of A Conditional Expectation. *Simulation Conference, 2003. Proceedings of the 2003 Winter*, Vol. **1**, 2003, p. 383– 391

204. Stone C. J., An asymptotically optimal window selection rule for kernel density estimates, *The Annals of Statistics*, Vol. **12**, 1984, p. 1285–1297.

205. Stone C. J., The use of polynomial splines and their tensor products in multivariate function estimation, *The Annals of Statistics*, Vol. **22**, 1994, p. 118–184.

206. Stone C. J., Hansen M., Kooperberg C. and Truong Y. K., Polynomial Splines and Their Tensor Products in Extended Linear Modeling, *The Annals of Statistics*, Vol. **25**, 1997, p. 1371–1470.

207. Stone M., Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society*, B, **36**, 1974, p. 111–147.

208. Stone M., Asymptotics for and against cross-validation, *Biometrika*, Vol. **64**, 1977, p. 29–35.

209. Sweeting T. J., Coverage probability bias, objective Bayes and the likelihood principle. *Biometrika*, Vol. **88**, 2001, p. 657–675.

210. Utsugi A., Density estimation by mixture models with smoothing priors, *Neural Computation*, Vol. **10**, No. 8, 1998, p. 2115–2135.

211. Takada T., Nonparametric density estimation: A comparative study. *Economics Bulletin*, Vol. **3**, No. 16, 2001, p. 1–10.

212. Tapia R. A. and Thompson J. R., Nonparametric Probability Density Estimation. *Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore and London*, 1990.

213. Titterington D. M., Smith A. F. M., Markov U. E., Statistical Analysis of Finite Mixture Distributions, *New York: John Wiley*, 1985.

214. Tsuda K., Akaho S., Asai K., The *em* algorithm for kernel matrix completion with auxiliary data, *The Journal of Machine Learning Research*, Vol. **4,** No. 1, 2004, p. 67–81.

215. Vincent P. and Bengio Y., Locally Weighted Full Covariance Gaussian Density Estimation, *CIRANO*, 2004.

216. Zhang T., Ramakrishnan R., Linvy M., BIRCH: An Efficient Data Clustering Method for very Large Databases, *SIGMOD*, 1996, p. 103–114.

217. Zhang X., King M. L., Hyndman R. J., Bandwidth Selection for Multivariate Kernel Density Estimation Using MCMC*, Computational Statistics and Data Analysis*, Vol. **50**, 2004, p. 3009–3031.

218. Zhao Y., Atkeson C. G., Implementing projection pursuit learning, *IEEE Transactions on Neural Networks*, Vol. **7**, No. 2, 1996, p. 362–373.

219. Vlassis N., Likas A., A Greedy EM algorithm for Gaussian Mixture Lear Learning, *Neural Processing Letters*, Vol. **15**, No. 1, 2000, p. 77–87.

220. Walter G. G. and Ghorai J. K., Advantages and disadvantages of density estimation with wavelets, *Comp. Sci. Stat.*, Vol. **24**, 1993, p. 234–243.

221. Wand M. P. and Jones M. C., Kernel Smoothing. *London: Chapman and Hall*, 1995.

222. Wand M. P. and Jones M. C., Comparison of smoothing parameterizations in bivariate kernel density estimation, *Journal of the American Statistical Association*, Vol. **88**, 1993, p. 520–528.

223. Wand M. P., Jones M. C., Multivariate plug-in bandwidth selection, *Computational Statistics*, Vol. **9**, 1994, p. 97–116.

224. Wand M. P., Fast computation of multivariate kernel estimators, *Journal of Computational and Graphical Statistics*, Vol. **3**, No. 4, 1994, p. 433–445.

225. Watson G., Smooth regression analysis, *Sankhya*, Ser. A, **26**, 1964, p. 359–372.

226. West M., Modelling With Mixtures, *Bayesian Statistics*, Vol. **4**, 1992, p. 503–524.

227. Whittle P., On the smoothing of probability density functions, *Journal of the Royal Statistical Society*, B, **20**, 1958, p. 334–343.

228. Wong M., A Hybrid Clustering Method for Identifying High-Density Clusters, *Journal of the American Statistical Association*, Vol. **77**, 1982, p. 841–847.

229. Wong M. and Lane T., A Kth-Nearest Neighbor Clustering Procedure, *Journal of the Royal Statistical Society*, Series B, Vol. **45**, 1983, p. 362–368.

230. Wong M., A bootstrap testing procedure for investigating the number of subpopulations, *Journal of Statistical Computation and Simulation*, Vol. **22**, 1985, p. 99–112.

231. Wu C. F. J., On convergence properties of the EM algorithm, *The Annals of Statistics*, Vol. **11**, 1983, p. 95–103.

232. http://bear.fhcrc.org/~clk/ (žiūrėta 2006-12-10).

233. http://support.sas.com/ (žiūrėta 2006-12-10).

234. http://www.r-project.org/ (žiūrėta 2006-12-10).

Autoriaus publikacijų sąrašas

Straipsniai Mokslinės informacijos instituto (ISI) duomenų bazėse referuojamuose leidiniuose

1A Rudzkis R., Ruzgas T., Clustering Effect on the Statistical Estimation Accuracy of Distribution Density, *Acta Applicandae Mathematicae*, Dordrecht, Kluwer Academic Publishers, 2007. ISSN 0167-8019. (Accepted)

Straipsniai Lietuvos mokslo tarybos patvirtinto sąrašo tarptautinėse duomenų bazėse referuojamuose leidiniuose

2A Kavaliauskas M., Rudzkis R., Ruzgas T., The Projection-based Multivariate Distribution Density Estimation, *Acta et Commentationes Universitatis Tartuensis de Mathematica*, Vol. **8**, Tartu, Tartu University Press, 2004, p. 135–141. ISSN 1406-2283. (Zentralblatt MATH)

3A Ruzgas T., Rudzkis R. and Kavaliauskas M., Application of Clustering in the Non-Parametric Estimation of Distribution Density, *Nonlinear Analysis*: *Modelling and Control*, Vol. **11**, No 4, Vilnius, 2006, p. 393–411. ISSN 1392-5113. (VINITI, Zentralblatt MATH)

Straipsniai kituose recenzuojamuose mokslo leidiniuose

4A Ruzgas T., Įvairių klasterizavimo algoritmų efektyvumo palyginimas, *Lietuvos matematikos rinkinys*, spec. nr., Vol. **42**, Vilnius, MII, 2002, p. 571–576. ISSN 0132-2818.

5A Ruzgas T., Kavaliauskas M., Daugiamačių Gauso skirstinių mišinio modelio panaudojimas neparametrinių tankių vertinime, *Lietuvos matematikos rinkinys*, spec. nr., Vol. **45**, Vilnius, MII, 2005, p. 369–374. ISSN 0132-2818.

6A Šmidtaitė R., Ruzgas T., Neparametrinių tankių vertinimo algoritmų tyrimas panaudojant klasterizavimo metodus, *Lietuvos matematikos rinkinys*, spec. nr., Vol. **46**, Vilnius, MII, 2006, p. 273–279. ISSN 0132-2818

Autoriaus pranešimų konferencijose sąrašas

1K Ruzgas T., Janilionis V., Hipersferos spindulio parinkimas klasifikuojant Gauso skirstinių mišinį, *Matematika ir matematinis modeliavimas*, KTU, Kaunas, 2002.

2K Ruzgas T., Įvairių klasterizavimo algoritmų efektyvumo palyginimas, *Lietuvos matematikų draugijos XLIII konferencija*, LKA, Vilnius, 2002.

3K Kavaliauskas M., Rudzkis R., Ruzgas T., Daugiamačių Gauso skirstinių mišinio parametrų įvertinimas taikant vienamates projekcijas, *Lietuvos matematikų draugijos XLIV konferencija*, VPU, Vilnius, 2003.

4K Kavaliauskas M., Rudzkis R., Ruzgas T., The Projection-based Multivariate Distribution Density Estimation, *The 7th Tartu Conference on Multivariate Statistics. Satellite meeting of ISI 54th session in Berlin*, University of Tartu, Tartu, 2003.

5K Ruzgas T., Kavaliauskas M., Gauso skirstinių mišinių klasterizavimas taikant neparametrinius metodus, *Matematika ir matematinis modeliavimas*, KTU, Kaunas, 2004.

6K Ruzgas T., Kavaliauskas M., Daugiamačių Gauso skirstinių mišinio modelio panaudojimas neparametrinių tankių vertinime, *Lietuvos matematikų draugijos XLVI konferencija*, VU, Vilnius, 2005.

7K Šmidtaitė R., Ruzgas T., Klasterizavimo panaudojimas neparametrinių tankių vertinime, *Taikomoji matematika*, KTU, Kaunas, 2006.

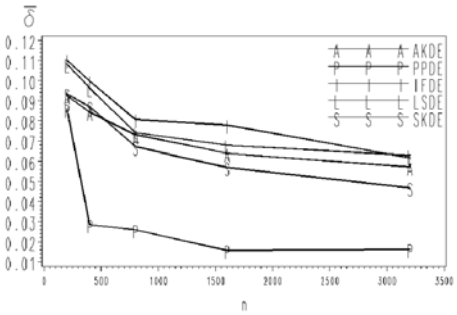
8K Ruzgas T., Rudzkis R., Kavaliauskas M., Negriežto klasterizavimo panaudojimas daugiamačių neparametrinių tankių vertinime, *Lietuvos matematikų draugijos XLVII konferencija*, KTU, Kaunas, 2006.

9K Ruzgas T., Šmidtaitė R., Neparametrinių tankių vertinimo algoritmų tyrimas panaudojant klasterizavimo metodus, *Lietuvos matematikų draugijos XLVII konferencija*, KTU, Kaunas, 2006.

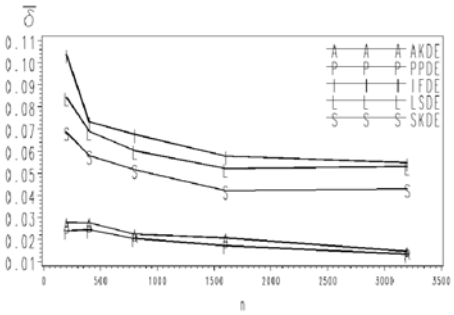
10K Rudzkis R, Ruzgas T., Nonparametric Estimation of Distribution Density Applying Clustering Procedures, *9th Vilnius Conference on Probability Theory and Mathematical Statistics*, VGTU, Vilnius, 2006.

1 priedas. Tankių vertinimo rezultatų diagramos

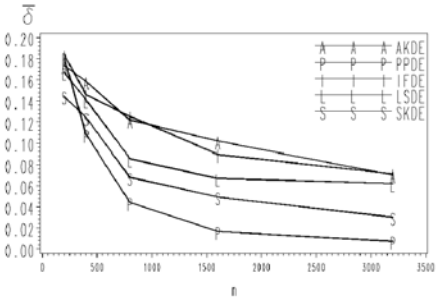
1P.1–1P.3 paveikslų grupėse pateikti duomenų B1, B2 ir B3 modelių, kurie aprašyti 2.3 skyriuje, tankių vertinimo rezultatai (šio priedo paveiksluose paklaida apibrėžta (2.27) formule), kai pradinis klasterizavimas atliktas naudojant EM algoritmą.



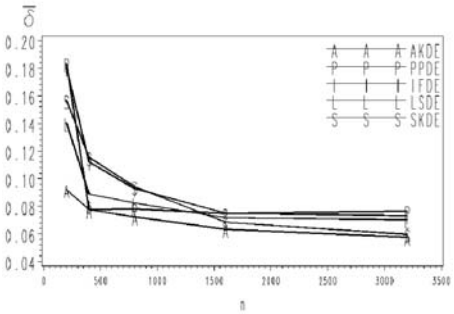
Gauso skirstinys d=2



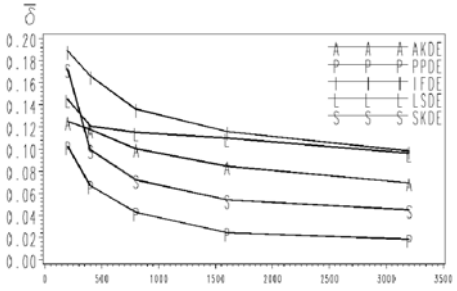
Koši skirstinys d=2



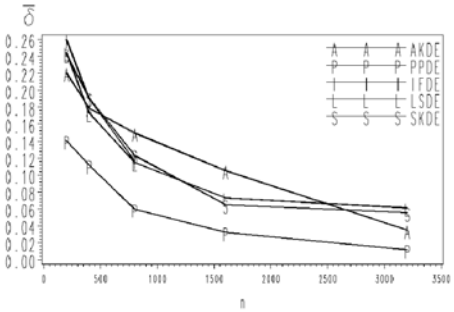
Gauso skirstinys d=3



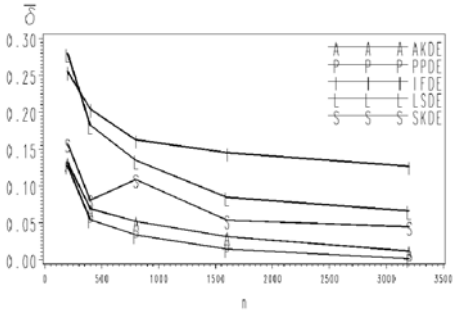
Koši skirstinys d=3



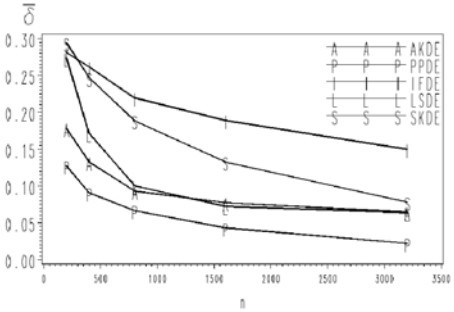
Gauso skirstinys d=4



Koši skirstinys d=4

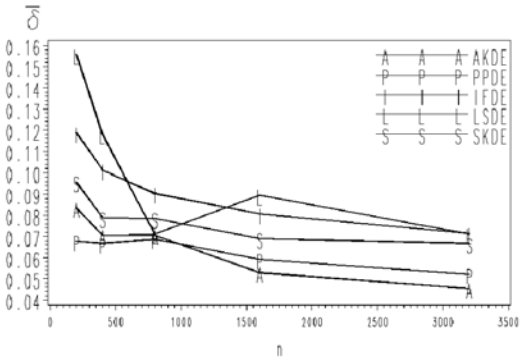


Gauso skirstinys d=5

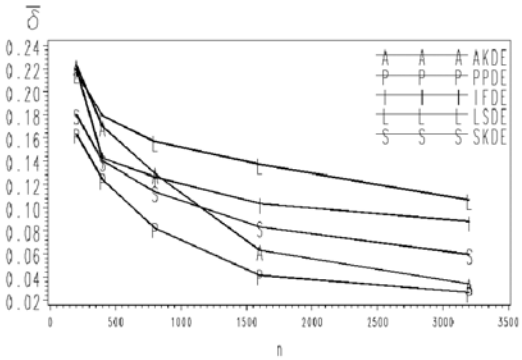


Koši skirstinys d=5

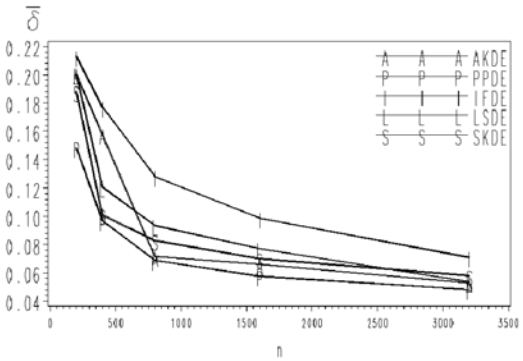
1P.1 pav. Vienos modos Gauso ir Koši skirstiniai



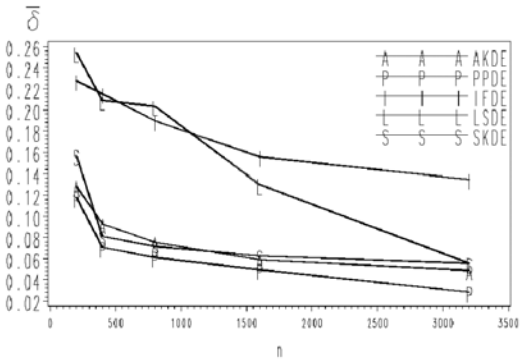
Gauso skirstinių mišinys d=2



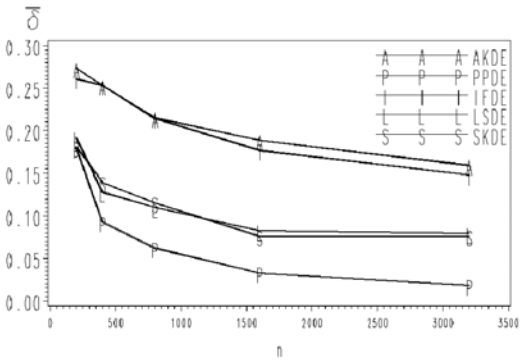
Koši skirstinių mišinys d=2



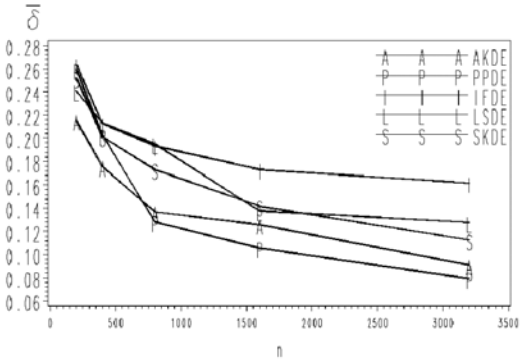
Gauso skirstinių mišinys d=3



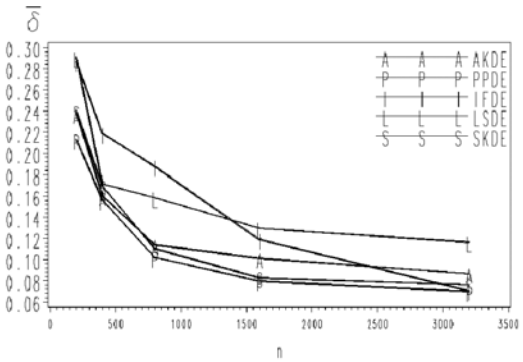
Koši skirstinių mišinys d=3



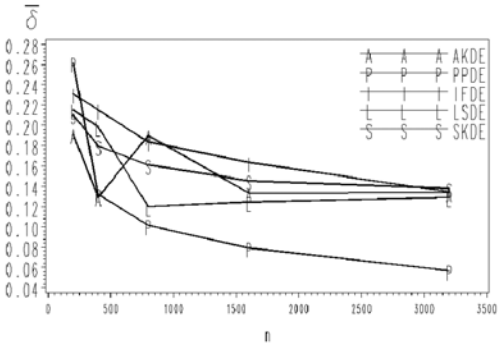
Gauso skirstinių mišinys d=4



Koši skirstinių mišinys d=4

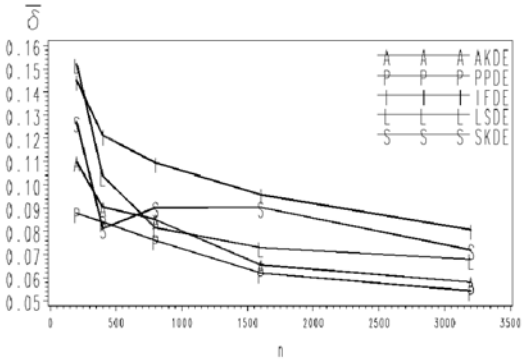


Gauso skirstinių mišinys d=5

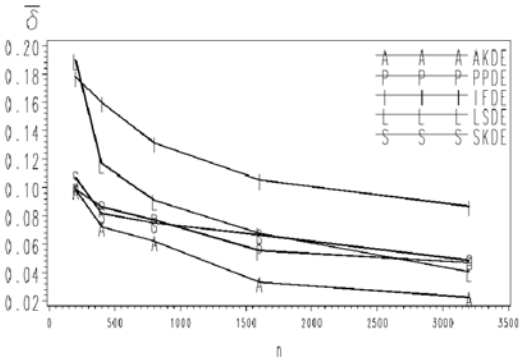


Koši skirstinių mišinys d=5

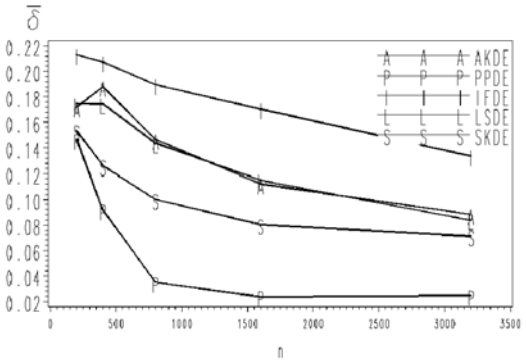
1P.2 pav. Mažai persidengiantys Gauso ir Koši skirstinių mišiniai



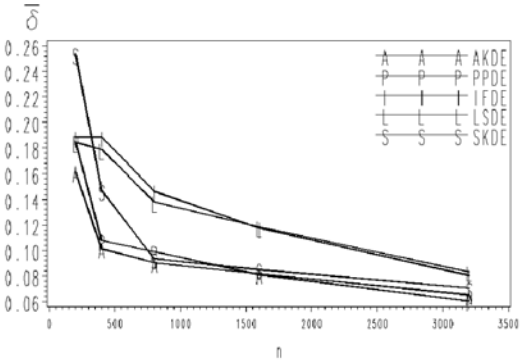
Gauso skirstinių mišinys d=2



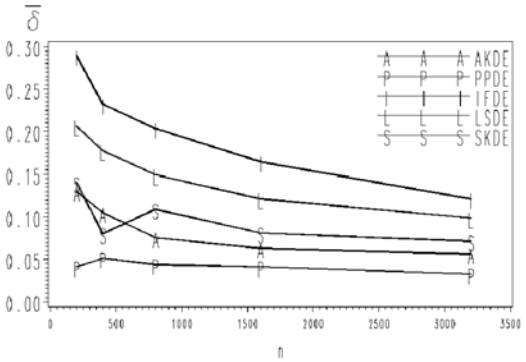
Koši skirstinių mišinys d=2



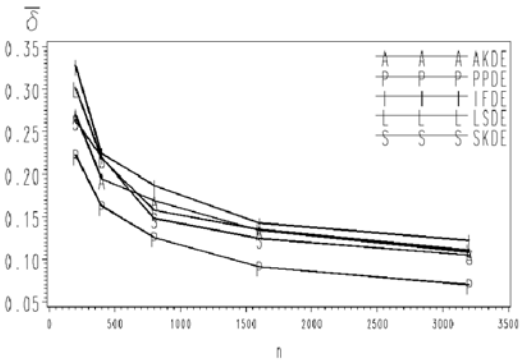
Gauso skirstinių mišinys d=3



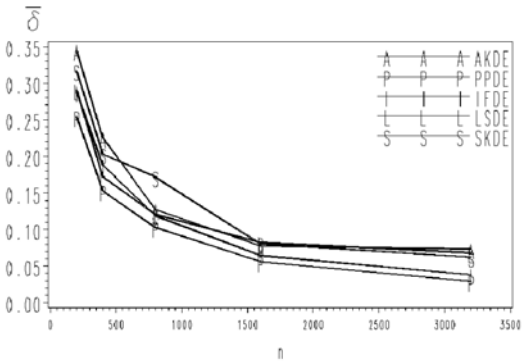
Koši skirstinių mišinys d=3



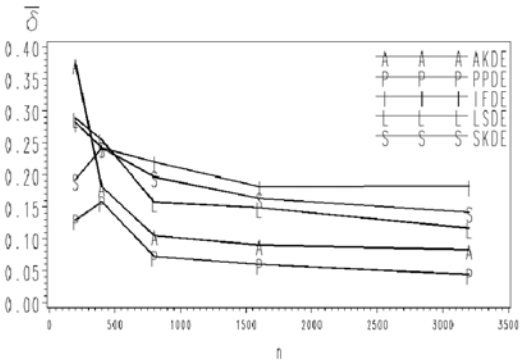
Gauso skirstinių mišinys d=4



Koši skirstinių mišinys d=4



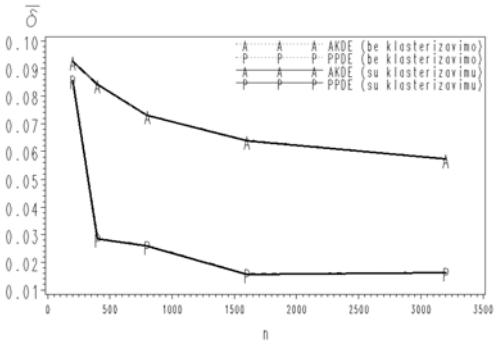
Gauso skirstinių mišinys d=5



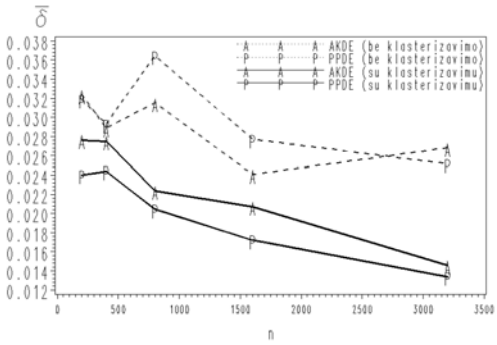
Koši skirstinių mišinys d=5

1P.3 pav. Labai persidengiantys Gauso ir Koši skirstinių mišiniai

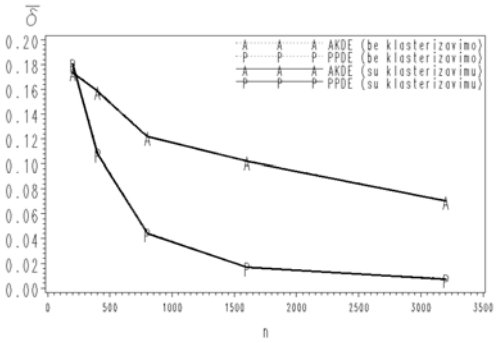
1P.4–1P.6 paveikslų grupėse pateikti duomenų B1, B2 ir B3 modelių tankių vertinimo rezultatai, gauti adaptuoto branduolinio ir tikslinio projektavimo metodais, palyginant pradinio klasterizavimo, atlikto naudojant EM algoritmą, poveikį su rezultatais gautais neatlikus klasterizavimo.



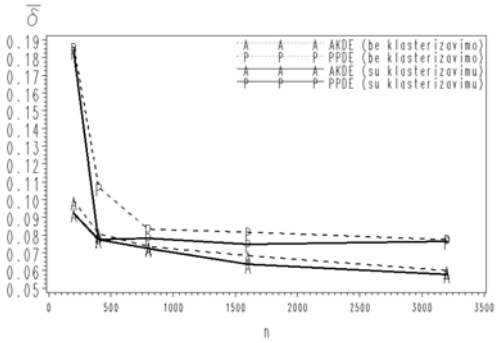
Gauso skirstinys d=2



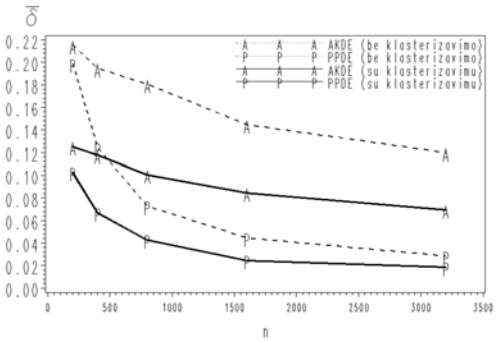
Koši skirstinys d=2



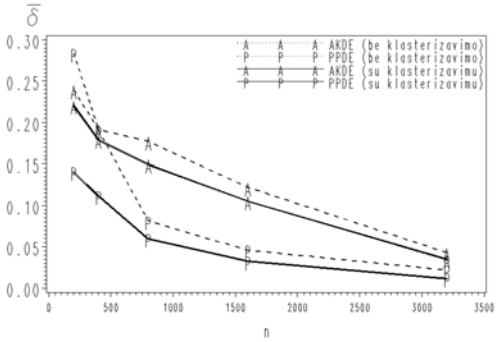
Gauso skirstinys d=3



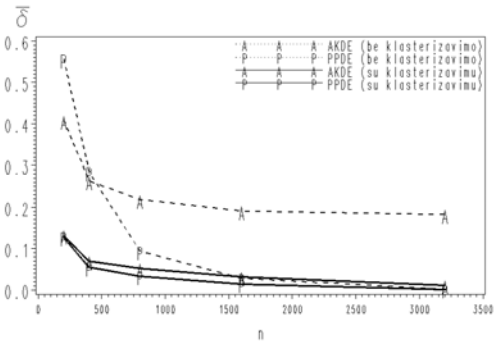
Koši skirstinys d=3



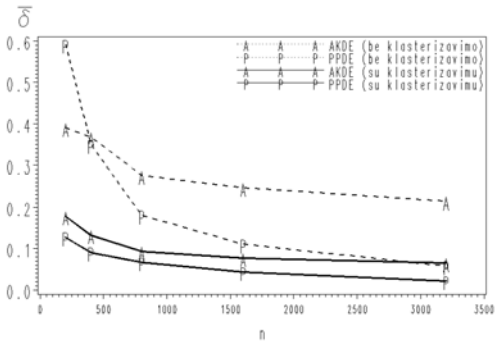
Gauso skirstinys d=4



Koši skirstinys d=4

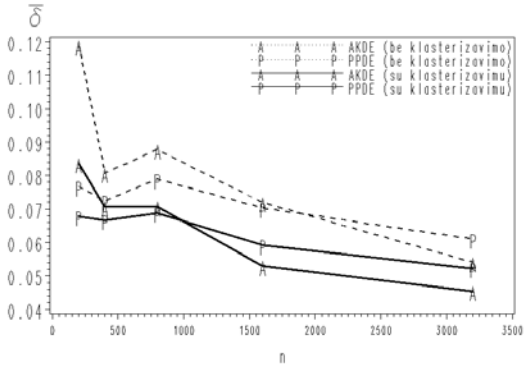


Gauso skirstinys d=5

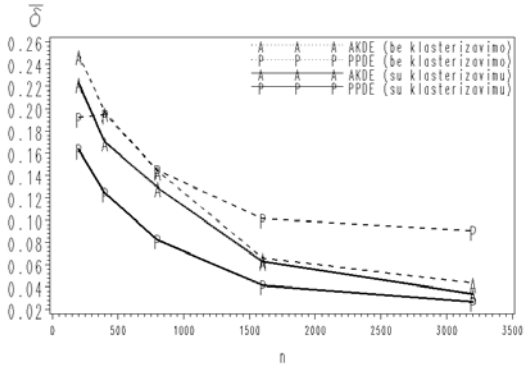


Koši skirstinys d=5

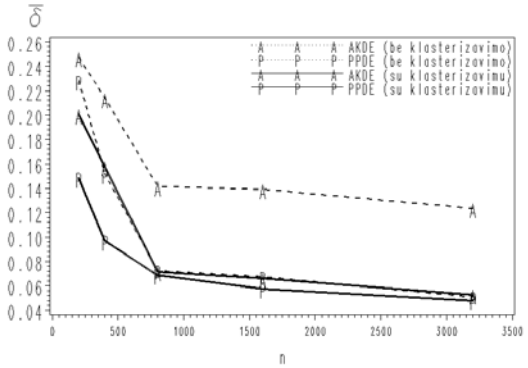
1P.4 pav. Vienos modos Gauso ir Koši skirstiniai



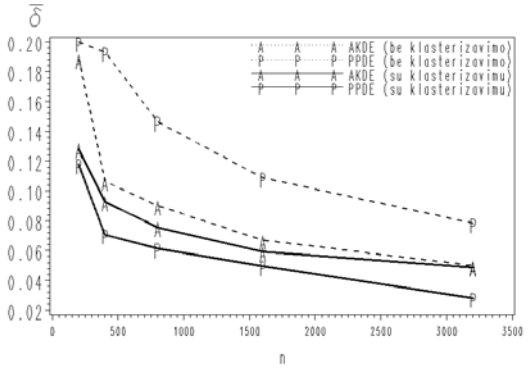
Gauso skirstinių mišinys *d*=2



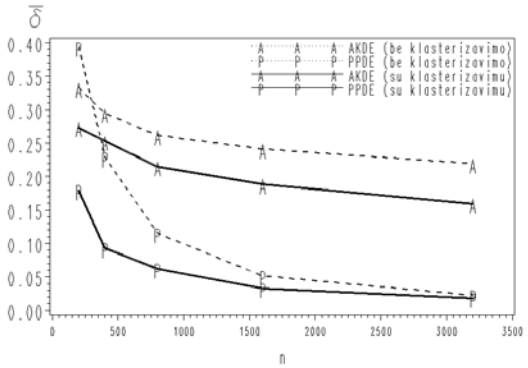
Koši skirstinių mišinys *d*=2



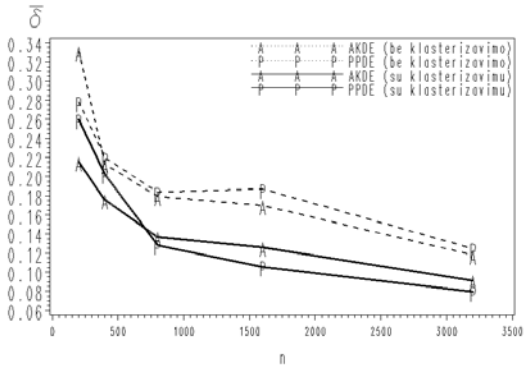
Gauso skirstinių mišinys *d*=3



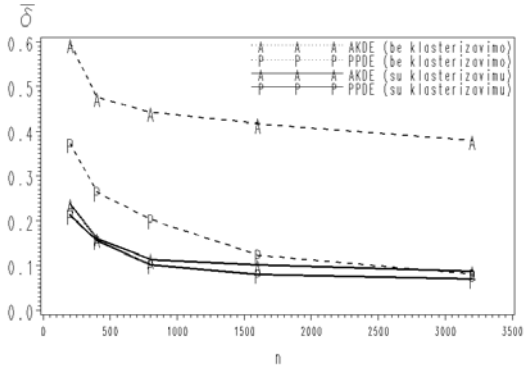
Koši skirstinių mišinys *d*=3



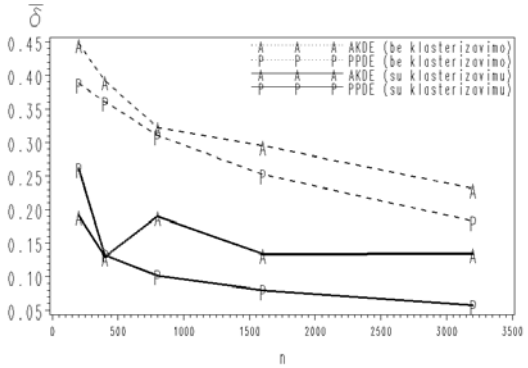
Gauso skirstinių mišinys *d*=4



Koši skirstinių mišinys *d*=4

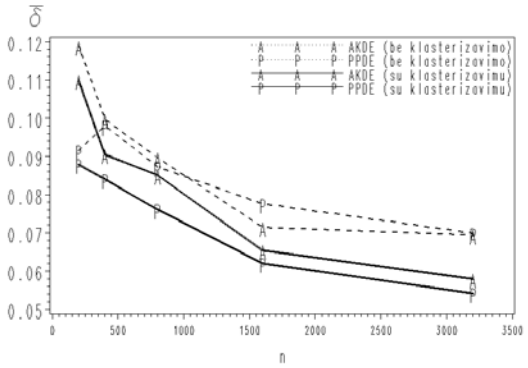


Gauso skirstinių mišinys *d*=5

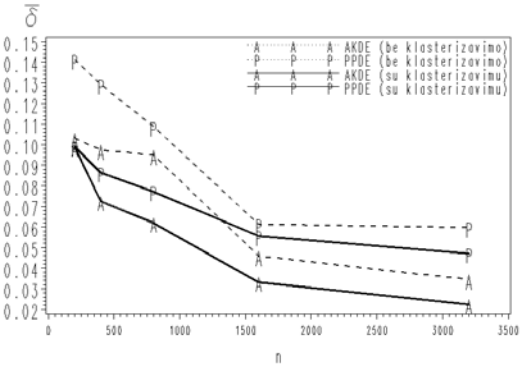


Koši skirstinių mišinys *d*=5

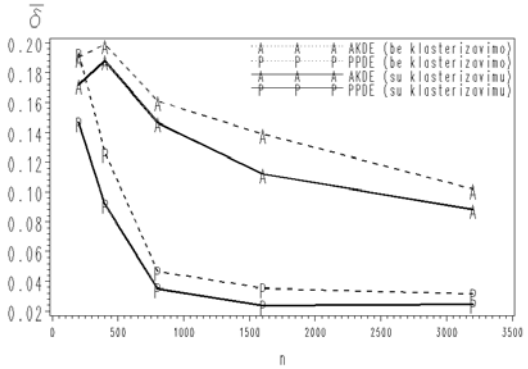
1P.5 pav. Mažai persidengiantys Gauso ir Koši skirstinių mišiniai



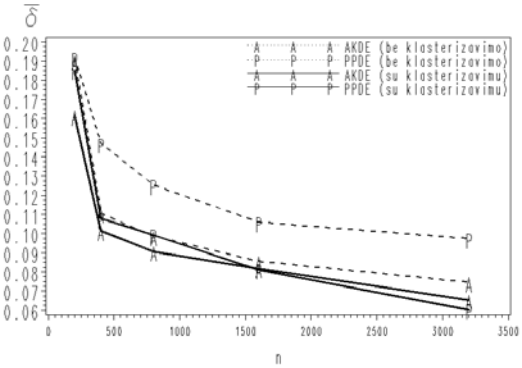
Gauso skirstinių mišinys *d*=2



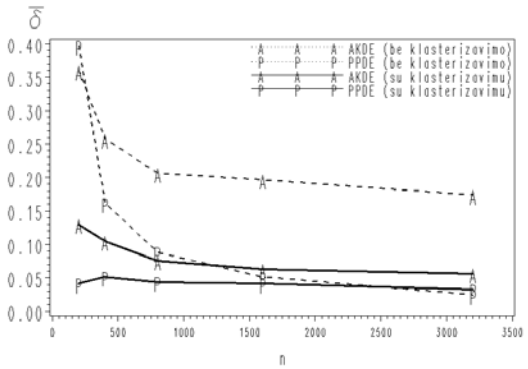
Koši skirstinių mišinys *d*=2



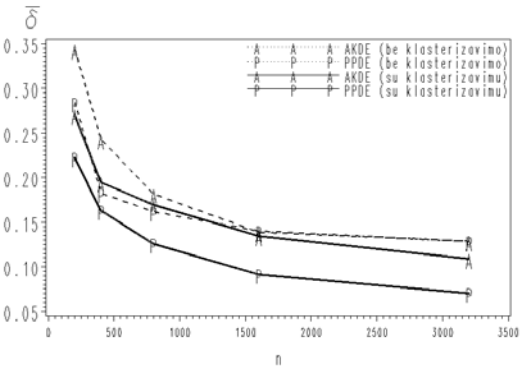
Gauso skirstinių mišinys *d*=3



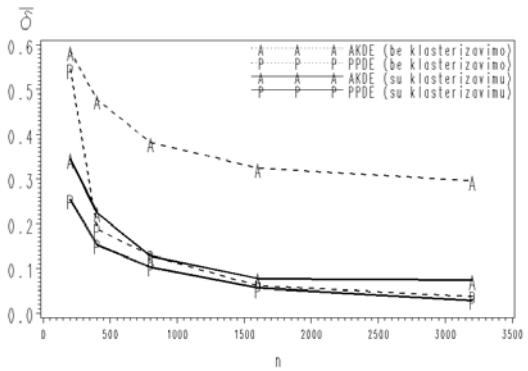
Koši skirstinių mišinys *d*=3



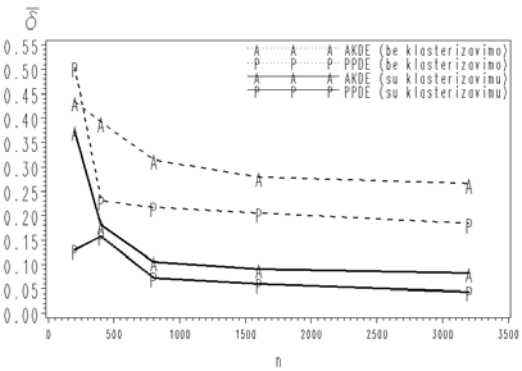
Gauso skirstinių mišinys *d*=4



Koši skirstinių mišinys *d*=4



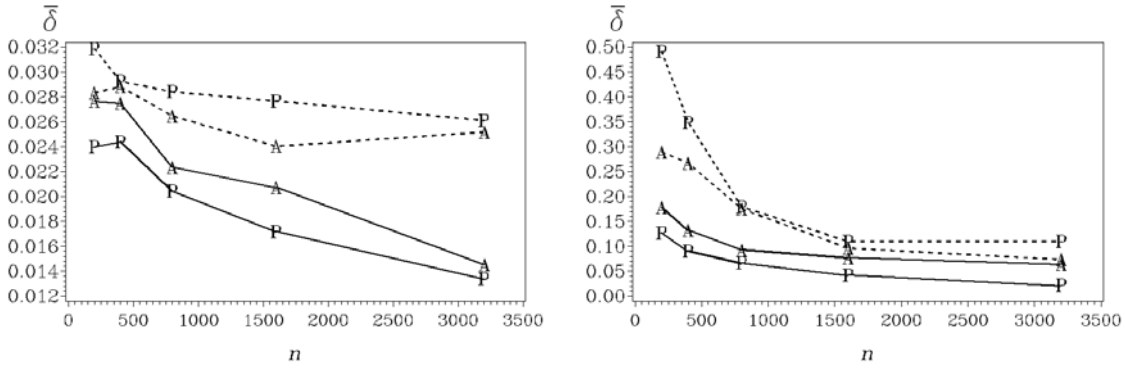
Gauso skirstinių mišinys *d*=5



Koši skirstinių mišinys *d*=5

1P.6 pav. Labai persidengiantys Gauso ir Koši skirstinių mišiniai

1P.7 paveiksle pateikti duomenų B1 tipo Koši tankio vertinimo rezultatai, gauti adaptuoto branduolinio ir tikslinio projektavimo metodais naudojant pirminį imties klasterizavimą *k* artimiausių kaimynų procedūra ir Gauso skirstinių mišinio modeliu bei EM algoritmu; simboliais A ir P atitinkamai žymimi AKDE ir PPDE algoritmai; ištisine linija žymimos paklaidos atlikus pirminį duomenų klasterizavimą naudojant Gauso skirstinių mišinio modelį bei EM algoritmą, punktyrine – *k* artimiausių kaimynų procedūra.



Koši skirstinys *d*=2 Koši skirstinys *d*=5

1P.7 pav. Vienos modos Koši skirstiniai

2 priedas. Kompiuterinio modeliavimo programos kodas

Modeliavimas atliktas naudojant SAS [233]ir R [234] programinę įrangą. Priede pateiktas pusiau parametrinio branduolinio pasiskirstymo tankio vertinimo programos kodo pavyzdys, užrašytas SAS programavimo kalba [16, 181, 182, 183, 184].

```
%macro SKDE(s, XC=X);
   proc iml;
      use X;  read all into X;  close X;
      use &XC;  read all into XC;  close &XC;
      %if %sysfunc(exist(svoriai)) %then
      %do;
         use svoriai;  read all into w;  close svoriai;
      %end;
      %else %do;
         w=j(nrow(XC),1,1);
      %end;
      if &s ^= 0 then
      do;
         YC=XC[,1:&s];
         Y=X[,1:&s];
         if &s ^= &d then
         do;
            ZC=XC[,&s+1:&d];
            Z=X[,&s+1:&d];
         end;
         if nrow(YC)=1 then std=j(1,ncol(YC),1);
         else std=sqrt(((YC-
repeat((YC#w)[+,]/w[+],nrow(YC),1))##2#w)[+,]/w[:]/(nrow(YC)-1));
         H=diag(nrow(YC)**(-1/(ncol(YC)+4))*std);
         start kdepoint(arg1,Z1,h1,h2,w);
            K=0;
            do i2=1 to nrow(Z1);
               arg2=(arg1-Z1[i2,])*inv(h1);
               K2=(2*constant('PI'))**(-ncol(Z1)/2)*exp(-1/2*arg2*arg2`);
               K=K+w[i2,]*K2;
               arg2=(arg1-Z1[i2,])*inv(h2);
               K2=(2*constant('PI'))**(-ncol(Z1)/2)*exp(-1/2*arg2*arg2`);
               if i2=1 then WW=w[i2,]*K2/det(h2);
               else WW=WW||(w[i2,]*K2/det(h2));
            end;
            f_w=(K/(w[+]*det(h1)))||WW;
            return(f_w);
         finish kdepoint;
         do i1=1 to nrow(Y);
            arg1=Y[i1,];
            f_tmp=kdepoint(arg1,YC,H,2*H,w);
            f=f_tmp[,1];
            WW=f_tmp[,2:(nrow(YC)+1)];
            if &s ^= &d then
            do;
               C=shape(0,ncol(ZC),ncol(ZC));
               if WW[+] ^= 0 then
               do;
                  m=(WW`#ZC)[+,]/WW[+];
                  do i3=1 to nrow(ZC);
                     C=C+WW[,i3]*(ZC[i3,]-m)`*(ZC[i3,]-m)/WW[+];
                  end;
                  if nrow(ZC)<=%eval(&d-&s) then C=I(%eval(&d-&s));
```

```
               end;
            end;
            if i1=1 then
            do;
               f_y=f;
               if &s ^= &d then
                  if abs(det(C))<=1E-100 then f_z=0;
                  else f_z=((2*constant('PI'))**ncol(C)*det(C))**(-1/2)*exp(-
1/2*(Z[i1,]-m)*inv(C)*(Z[i1,]-m)`);
            end;
            else
            do;
               f_y=f_y//f;
               if &s ^= &d then
                  if abs(det(C))<=1E-100 then f_z=f_z//0;
                  else f_z=f_z//(((2*constant('PI'))**ncol(C)*det(C))**(-
1/2)*exp(-1/2*(Z[i1,]-m)*inv(C)*(Z[i1,]-m)`));
            end;
         end;
      end;
      else
      do;
         ZC=XC;
         Z=X;
         m=(ZC#w)[+,]/w[+];
         C=shape(0,ncol(ZC),ncol(ZC));
         do i3=1 to nrow(ZC);
            C=C+w[i3,]*(ZC[i3,]-m)`*(ZC[i3,]-m)/w[+];
         end;
         if nrow(ZC)<=%eval(&d-&s) then C=I(%eval(&d-&s));
         do i1=1 to nrow(Z);
            if i1=1 then
            do;
               if abs(det(C))<=1E-100 then f_z=0;
               else f_z=((2*constant('PI'))**ncol(C)*det(C))**(-1/2)*exp(-
1/2*(Z[i1,]-m)*inv(C)*(Z[i1,]-m)`);
            end;
            else
            do;
               if abs(det(C))<=1E-100 then f_z=f_z//0;
               else f_z=f_z//((2*constant('PI'))**ncol(C)*det(C))**(-
1/2)*exp(-1/2*(Z[i1,]-m)*inv(C)*(Z[i1,]-m)`);
            end;
         end;
      end;
      if &s ^= 0 then
      do;
         if &s ^= &d then SKDE=f_y#f_z;
         else SKDE=f_y;
      end;
      else SKDE=f_z;
      create SKDE from SKDE [colname='f'];
      append from SKDE;
   quit;
%mend SKDE;
```

Padėka

Nuoširdžiai dėkoju darbo vadovui prof. habil. dr. Rimantui Rudzkiui už vadovavimą disertaciniam darbui, skirtą laiką ir energiją, doc. dr. Marijui Radavičiui už išsamias konsultacijas, Kauno technologijos universiteto fundamentaliųjų mokslų fakulteto dekanui doc. dr. Vytautui Janilioniui už suteiktą kompiuterinę techniką. Ačiū tėvams už jų visapusį palaikymą, Ingai už kantrybę ir įvairią pagalbą rašant disertaciją. Taip pat dėkoju visiems kitiems, kurie tiesiogiai ar netiesiogiai prisidėjo prie šio darbo.

Feci, quod potui, faciant meliora potentes