VILNIAUS UNIVERSITETAS

Gintautas Jakimauskas

DUOMENŲ TYRYBOS EMPIRINIŲ BAJESO METODŲ TYRIMAS IR TAIKYMAS

Daktaro disertacija

Fiziniai mokslai, Informatika (09 P)

Vilnius, 2014

Disertacija rengta 2013–2014 metais Vilniaus universiteto Matematikos ir informatikos institute.
Disertacija ginama eksternu.

Mokslinis konsultantas prof. habil. dr. Leonidas Sakalauskas (Vilniaus universitetas, fiziniai mokslai, informatika – 09P).

PADĖKA

Disertacijos autorius dėkoja moksliniam konsultantui prof. habil. dr. Leonidui Sakalauskui už vertingas mokslines konsultacijas ir nuolatinį skatinimą, ilgamečiams bendradarbiams habil. dr. Rimantui Rudzkiui, dr. Marijui Radavičiui ir dr. Jurgiui Sušinskui bei visiems pateikusiems pastabas ir pasiūlymus.

Gintautas Jakimauskas

TURINYS

ŽYMĖJIMAI IR SANTRUMPOS

n.v.p. – nepriklausomi vienodai pasiskirstę

a.d. – atsitiktinis dydis (atsitiktiniai dydžiai)

a.v. – atsitiktinis vektorius (atsitiktiniai vektoriai)

p.f. – pasiskirstymo funkcija

$\mathbf{R}^d$ – d-matė Euklidinė erdvė

N – imties dydis

F – pasiskirstymo funkcija

P – skirstinys

$\mathcal{N}(a, \sigma^2)$ – Gauso skirstinys su vidurkiu a ir dispersija σ^2

$\mathcal{N}(\mu, R)$ – daugiamatis Gauso skirstinys su vidurkių vektoriumi μ ir kovariacine matrica R

$\mathcal{B}(\lambda, n)$ – binominis skirstinys su įvykio tikimybe λ ir bandymų skaičiumi n

$\mathbf{X} = (X(1),...,X(N))$ – n.v.p. atsitiktinio vektoriaus X su pasiskirstymo funkcija F erdvėje $\mathbf{R}^d$ stebėjimų imtis

$(X_1(j), X_2(j), \dots, X_d(j))^{\mathrm{T}}$ – imties $\mathbf{X}$ j-tojo stebėjimo komponentės

X^N – n.v.p. atsitiktinio vektoriaus X su pasiskirstymo funkcija F erdvėje $\mathbf{R}^d$ stebėjimų imtis (kitas pažymėjimas)

$q_j, j = 1, 2, \dots, N$ – imties stebėjimų pasikartojimų skaičius, t. y. stebėjimas $X(j)$ pasikartoja q_j kartų

$NGrp$ – grupuotos imties dydis

X_{Grp}^{NGrp} – grupuota imtis

S – stačiakampis gretasienis erdvėje $\mathbf{R}^d$

$Z_k(j), j = 1, 2, \dots, k$ – grupavimo taškai po k-tojo žingsnio

$Z_k(j), j = 1, 2, \dots, k$ – grupavimo taškai po k-tojo žingsnio

E_k^2 – grupavimo paklaida po k-tojo žingsnio

$\Delta^2(k+1, m, v)$ – būsimas grupavimo paklaidos sumažėjimas po k+1-ojo žingsnio m-tajam stačiakampiui gretasieniui pagal koordinatę v

$Q_{\max}(k)$ – k-tojo procedūros žingsnio viršutinė operacijų skaičiaus riba

$1_{a \in A}$ – elemento a priklausymo aibei A indikatorius

Ω – populiacija

$\pi(Y)$ – populiacijos Ω sąlyginė tikimybė su sąlyga Y

T_k – testinė χ^2-tipo statistika

p – mišinio parametras, $p \in (0, 1)$

$\mathcal{F}_H$ ir $\mathcal{F}_A$ – dvi nepersikertančios d-mačių skirstinių klasės

H – neparametrinė hipotezė $H : F \in \mathcal{F}_H$ prieš $A : F \in \mathcal{F}_A$

X_1 ir X_2 – a.v. komponentės, $X_1 \in R^{d_1}$ ir $X_2 \in R^{d_2}$, $d_1 + d_2 = d$

G – skirstiniai iš klasės $\mathcal{F}_H$

G_1 ir G_2 – marginaliniai G skirstiniai, atitinkantys komponentes X_1 ir X_2

$L(F; F_0)$ – nuostolių funkcija tarp p.f. F ir F_0

$\mathbf{X}^{(H)} = (X^{(H)}(1), \ldots, X^{(H)}(M))$ – n.v.p. a.v. Ω_H imtis, nepriklausoma nuo $\mathbf{X}$

$\mathbf{Y} = \mathbf{X} \,||\, \mathbf{X}^{(H)} = (X(1), \ldots, X(N), X^{(H)}(1), \ldots, X^{(H)}(M))$ – apjungta imtis

$\hat{F}$ – empirinis imties $\mathbf{Y}$ pasiskirstymas

$\hat{\mathbf{E}}$ – vidurkis pagal empirinį imties $\mathbf{Y}$ pasiskirstymą $\hat{F}$

$\mathcal{P} = \{P_k, k = 0, 1, \ldots, K\}, P_0 = \mathbf{R}^d, P_{k-1} \subset P_k, k = 1, 2, \ldots, K,$ – erdvės $\mathbf{R}^d$ padalijimų seka

$\{\mathcal{A}_k, k = 0, 1, \ldots, K\}$ – σ-algebrų seka, generuota pagal seką $\mathcal{P}$

W_k – $\mathcal{A}_k$-išmatuojama svorio funkcija

c_α – kritinė reikšmė

$C_\alpha = \{T > c_\alpha\}$ – statistikos T kritinė sritis

$\mathbf{Y}^\tau$ – randomizuota imtis

T_k^τ – randomizuota testinė statistika

NML – neparametrinis didžiausio tikėtinumo (įvertinys) (*nonparametric maximum likelihood* (*estimator*))

NMLE – neparametrinis didžiausio tikėtinumo įvertinys (*nonparametric maximum likelihood estimator*)

EB – empirinis Bajeso (įvertinys) (*empirical Bayes* (*estimator*))

NEB – neparametrinis empirinis Bajeso (įvertinys) (*nonparametric empirical Bayes* (*estimator*))

$\Lambda = (A_1, A_2, ..., A_K)$ – K populiacijų aibė

N_j – individų skaičius *j*-toje populiacijoje

P_j – nežinomos tikimybės *j*-toje populiacijoje

Y_j , $j = \overline{1, K}$ – stebimas įvykių skaičius populiacijose

α_j , $j = \overline{1, K}$ – logitai (*logits*), $\alpha_j = \ln \frac{P_j}{1 - P_j}$

λ_j – nežinomos tikimybės *j*-toje populiacijoje (kitas pažymėjimas)

$\{\overline{\lambda}_j^{MRR}\} \equiv \overline{\lambda}^{MRR}$, $j = 1, 2, ..., K.$ – vidutinės santykinės rizikos (*mean relative risk* (MRR)) įvertis

$\{\overline{\lambda}_j^{RR}\} = \overline{\lambda}_j^{RR}$, $j = 1, 2, \ldots, K$, – santykinės rizikos (*relative risk* (RR)) įvertis

PAVEIKSLŲ SĄRAŠAS .....

LENTELIŲ SĄRAŠAS .....

1. BENDROJI DARBO CHARAKTERISTIKA

1.1. TYRIMŲ SRITIS

Darbo tyrimų sritis yra didelio matavimo didelių populiacijų duomenų tyrybos (*data mining*) metodai ir algoritmai.

Efektyvus informacijos, slypinčios duomenyse, atskleidimas ir panaudojimas yra svarbiausias konkurencingumo didinimo veiksnys šiuolaikinėje dinamiškoje tyrimų ir verslo aplinkoje. Duomenų tyryba (DT) yra šiuolaikinė informacijos analizės sritis, atsiradusi duomenų bazių technologijų, dirbtinio intelekto ir statistinės duomenų analizės sankirtoje. DT yra labai plati sritis, apimanti daug metodų, algoritmų bei taikomųjų programinių sistemų. Jei įprasti duomenų analizės metodai padeda atskleisti tiriamų kintamųjų priklausomumą, tai DT unikali tuo, kad analizės rezultatas yra naujų priklausomybių, kurios buvo, ar net nebuvo, įtariamos egzistuojant, radimas. Šiuolaikinės DT technologijos pagrindas yra šablonų (*pattern*), atvaizduojančių daugiabriaunius duomenų tarpusavio santykius, koncepcija.

Galima išskirti pagrindinius veiksnius, į kuriuos atsižvelgiama sprendžiant DT uždavinius:

• tenka apdoroti didelio kiekio įvairialytę informaciją,

• analizės rezultatai gali būti teikiami plačiam vartotojų, turinčių skirtingų poreikių, ratui.

DT metodams, pritaikomiems prognozavimui ir sprendimų priėmimui, būdingos dvi fazės: pirmojoje, pasinaudojant sukauptų duomenų imtimi, atskleidžiamos duomenų struktūrų savybės ir parengiamos taisyklės, kuriomis pasinaudojama antrojoje fazėje prognozuojant ir priimant sprendimus. Didelė metodų ir algoritmų įvairovė atskleidžia duomenų tyrybos sudėtingumą ir leidžia jos technologijas pritaikyti įvairiose nagrinėjamose situacijose. Dažnai DT uždaviniui išspręsti taikomi keli metodai iš eilės ar net sudėtingi jų deriniai. Uždavinių bei metodų įvairovę papildo grupė duomenų tyrybos algoritmų. Nė vienas iš jų nėra universalus ar nepriekaištingas. Parenkant

algoritmus atsižvelgiama į jų operacinį ir loginį sudėtingumą, sugaištamą analizei kompiuterio laiką bei atmintį, analizės patikimumą.

Žinios, atskleidžiamos DT metodais bei technologijomis, formaliai pateikiamos prielaidų (hipotezių) apie duomenų modelius arba modelių parametrų įverčių pavidalą: pvz., reikia įvertinti, ar duomenys yra tendencijų susidaryti panašių objektų klasteriams arba grupėms, ar tam tikri požymiai yra tarpusavyje susiję, ar duomenis galima pateikti sutrauktu pavidalu, neprarandant esminės informacijos ir pan. Sprendžiant DT uždavinius, dažniausiai pritaikomi šie metodai: klasterizavimas, daugiamatis skaliavimas, klasifikavimas, atraminių vektorių mašinos (*support vector machines*), regresinė analizė ir krikingas, pagrindinių komponenčių metodas (*Principal Component Analysis*), ir kt. (žr., pvz., Nisbet *et al.* (2009), Ye (2003)).

Bajeso sprendimų teorija yra plačiai taikoma duomenų tyryboje, kuomet informacija apie parametrus gali būti nusakoma tam tikru tikimybiniu skirstiniu. Šia teorija besiremiantys metodai pasižymi daugeliu privalumų (žr., pvz., DeGroot (1970), Carlin, Louis (1996), Diuk, Samoilenko (2002), Rossi *et al.* (2003), Diaconis (2009), Press *et al.* (2007), Richey (2010)), lyginant su klasikiniais (*„frequentist“*) metodais. Vis dėlto, šie metodai pradėti plačiau taikyti tik pastaraisiais dešimtmečiais daugiausiai dėl dviejų priežasčių. Viena iš dažnai nurodomų priežasčių yra tai, kad jie gali nebūti objektyvūs, t. y., jei statistiniai skaičiavimai nebus atlikti kruopščiai patikrinus ir įvertinus iš anksto numanomas nagrinėjamojo objekto savybes, tai gali padaryti subjektyvią įtaką rezultatams. Kita svarbi prielaida Bajeso metodams vystyti yra sparti kompiuterinės technikos plėtotė maždaug nuo 1980 metų, nes Bajeso metodams net ir gana paprastais atvejais reikia nemažų skaičiavimų.

1.2. PROBLEMOS AKTUALUMAS

Didelio duomenų rinkinio apdorojimas, sprendžiant pakankamai tiksliai ir pakankamai greitai įvairius uždavinius, yra vienas iš pagrindinių duomenų tyrybos uždavinių. Vienas iš sprendimo būdų yra naudoti našesnę aparatinę ar

programinę įrangą, tačiau šis sprendimo būdas praktikoje ne visada prieinamas. Kitas sprendimo būdas – pakeisti pradinį duomenų rinkinį mažesniu, pagal galimybes išlaikant pradines duomenų rinkinio savybes.

Didelių populiacijų didelio matavimo duomenų tyrybos uždaviniai dažnai iškyla biometrikoje, medicinoje, draudime, tinklų analizėje ir pan. Pvz., retų įvykių didelėse populiacijose (pvz., tam tikros ligos tikimybių, mirčių, savižudybių ir t. t.) vertinimo problema yra aktuali statistinėje epidemiologijoje, draudimo srityje adekvatus draudiminių įvykių tikimybių vertinimas gali duoti žymų praktinį efektą ir leisti tiksliau planuoti sąnaudas.

Pastaruoju metu nėra visuotinai pripažintos metodologijos daugiamačių neparametrinių hipotezių testavimui bei didelių populiacijų duomenų modelių parametrų vertinimui. Tradiciniai metodai testuojant daugiamates neparametrines hipotezes remiasi empirine charakteringąja funkcija, neparametriniais pasiskirstymo tankio įverčiais ir glodinimu, daugiamačiais neparametriniais Monte-Carlo testais, ir klasikinėmis vienmatėmis neparametrinėmis statistikomis, naudojamoms duomenims, projektuotiems į kryptis, rastas tikslinio projektavimo (*projection pursuit*) metodu. Sudėtingesni metodai remiasi Vapnik-Chervonenkis teorija, tolydžia funkcionaline centrine ribine teorema ir didelių nuokrypių tikimybių nelygybėmis (žr. Marcoulides, Hershberger (1997), Hirukawa (2012)). Pastaruoju metu pradėti plačiai taikyti, ypač praktiniuose uždaviniuose, Bajeso metodai ir Markovo grandinių Monte-Carlo metodai (žr. Andrieu *et al.* (2003), Berg (2004), Asmussen, Glynn (2007), Sakalauskas, Vaičiulytė (2012), Vaičiulytė, Sakalauskas (2011)).

Todėl empirinio Bajeso metodų ir algoritmų tyrimas bei taikymas neparametrinių hipotezių testavimui didelio matavimo duomenims bei didelių populiacijų statistinių modelių parametrų vertinimui yra aktuali teorinė ir praktinė duomenų tyrybos problema.

1.3. TYRIMŲ OBJEKTAS

Darbo tyrimų objektas yra duomenų tyrybos empiriniai Bajeso metodai ir algoritmai, taikomi didelio matavimų skaičiaus didelių populiacijų duomenų analizei.

1.4. TYRIMŲ TIKSLAS IR UŽDAVINIAI

Darbo tyrimų tikslas yra sudaryti metodus ir algoritmus didelių populiacijų neparametrinių hipotezių tikrinimui ir duomenų modelių parametrų vertinimui.

Šiam tikslui pasiekti yra sprendžiami tokie uždaviniai:

1. Sudaryti didelio matavimo duomenų skaidymo algoritmą.
2. Pritaikyti didelio matavimo duomenų skaidymo algoritmą neparametrinėms hipotezėms tikrinti.
3. Pritaikyti empirinį Bajeso metodą daugiamačių duomenų komponenčių nepriklausomumo hipotezei tikrinti su skirtingais matematiniais modeliais, nustatant optimalų modelį ir atitinkamą empirinį Bajeso įvertinį.
4. Sudaryti didelių populiacijų retų įvykių dažnių vertinimo algoritmą panaudojant empirinį Bajeso metodą palyginant Puasono-gama ir Puasono-Gauso matematinius modelius.
5. Sudaryti retų įvykių logistinės regresijos algoritmą panaudojant empirinį Bajeso metodą.

1.5. MOKSLINIS NAUJUMAS

Darbo metu gauti šie nauji rezultatai:

1. Didelio matavimo duomenų binarinio skaidymo metodas, paremtas erdvės skaidymu, naudojamu duomenų klasifikavime, kuris įgalina atlikti didelio matavimo duomenų skaidymą, pritaikomą duomenų grupavimui bei neparametrinių hipotezių tikrinimui.
2. Naujas metodas didelio matavimo nekoreliuotų duomenų pasirinktų komponenčių nepriklausomumo tikrinimui.

3. Naudojant skirtingus matematinius modelius parengtas naujas metodas parenkant didelių populiacijų retų įvykių optimalų modelį ir atitinkamą empirinį Bajeso įvertinį. Pateikta nesinguliarumo sąlyga Puasono-gama modelio atveju.

1.6. PRAKTINĖ DARBO REIKŠMĖ

Darbo metu gauti šie praktiniai rezultatai:

1. Sudarytas ir ištirtas didelio matavimo duomenų skaidymo algoritmas, pritaikytas MII sukurtoje programinėje įrangoje daugiamačių Gauso mišinių klasifikavimui.

2. Sudarytas skaitmeninis algoritmas didelio matavimo nekoreliuotų duomenų pasirinktų komponenčių nepriklausomumo tikrinimui, kurio galia yra didesnė, palyginti su klasikiniu komponenčių nepriklausomumo tikrinimo algoritmu.

3. Darbe sudaryti skaitmeniniai maksimalaus tikėtinumo algoritmai vertinti kelių Puasono-gama ir Puasono-Gauso empirinių Bajeso modelių parametrus, kurių buvo pritaikyti medicininių ir sociologinių duomenų analizei. Pateikta ir ištirta nesinguliarumo sąlyga Puasono-gama modelio atveju.

1.7. DARBO REZULTATŲ APROBAVIMAS

Tyrimų rezultatai buvo pristatyti ir aptarti šiose nacionalinėse ir tarptautinėse konferencijose:

1. Jakimauskas, Gintautas. Gamma and logit models in empirical bayesian estimation of probabilities of rare events // STOPROG 2012: Stochastic programming for implementation and advanced applications: international workshop, July 3-6, 2012, Neringa, Lithuania.
2. Jakimauskas, Gintautas; Sakalauskas, Leonidas. Empirical Bayesian estimation for Poisson-gamma model // 24th Mini EURO conference on

continuous optimization and information-based technologies in the financial sector (MEC EurOPT 2010), Vilnius

3. Gurevičius, Romualdas; Jakimauskas, Gintautas; Sakalauskas, Leonidas. Empirical Bayesian estimation of small mortality rates // 5th international Vilnius conference [and] EURO-mini conference "Knowledge-based technologies and OR methodologies for decisions of sustainable development" (KORSD-2009): September 30 – October 3, 2009, Vilnius, Lithuania.
4. Sakalauskas, Leonidas; Jakimauskas, Gintautas; Sušinskas, Jurgis. Analysis of medical data by empirical Bayes method // Computer data analysis and modeling: complex stochastic data and systems: Ninth international conference: Minsk, September 7-11, 2010.
5. Radavičius, Marijus; Jakimauskas, Gintautas; Sušinskas, Jurgis. Empirical Bayes testing goodness-of-fit for high-dimensional data // Computer data analysis and modeling: complex stochastic data and systems: Ninth international conference: Minsk, September 7-11, 2010.
6. Radavičius, Marijus; Jakimauskas, Gintautas; Sušinskas, Jurgis. Testing of independency for high-dimensional data // Computer data analysis and modeling: complex stochastic data and systems: Eighth international conference: Minsk, September 11-15, 2007.
7. Radavičius, Marijus; Jakimauskas, Gintautas; Sušinskas, Jurgis. Clustering and Testing in High-Dimensional Data, The 8th Tartu Conference on Multivariate Statistics, Tartu, 26-29 June 2007.
8. Jakimauskas, Gintautas; Sušinskas, Jurgis. Testing independence of high-dimensional random vectors, Nordic Conference on Mathematical Statistics 2008, Vilnius, 16-19 June 2008.

1.8. DARBO REZULTATŲ PUBLIKAVIMAS

Tyrimų rezultatai publikuoti šiuose moksliniuose leidiniuose:

Recenzuojamuose Lietuvos ir užsienio leidiniuose:

1. G. Jakimauskas. Efficiency analysis of one estimation and clusterization procedure of one-dimensional Gaussian mixture // Informatica, ISSN 0868-4952, 8(3), 1997, p. 331-343.
2. G. Jakimauskas, R. Krikštolaitis. Influence of projection pursuit on classification errors: computer simulation results // Informatica, ISSN 0868-4952, 11(2), 2000, p. 115-124.
3. G. Jakimauskas, R. Krikštolaitis. Bootstrap methods in selection of the discriminant subspace // Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai, ISSN 0132-2818, T. 40, 2000, p. 281-286.
4. G. Jakimauskas. Procedure of the removal of the outliers from the sample satisfying the multidimensional Gaussian mixture model // Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai, ISBN 9986-680-16-6, T. 42, 2002, p. 523-528.
5. Jakimauskas, Gintautas; Radavičius, Marijus; Sušinskas, Jurgis. A simple method for testing independence of high-dimensional random vectors // Austrian journal of statistics, ISSN 1026-597X, Vol. 37, no. 1, 2008, p. 101-108.
6. Jakimauskas, Gintautas. Efficient algorithm for testing goodness-of-fit for classification of high dimensional data // Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai, ISSN 0132-2818, T. 50, 2009, p. 293-297.
7. Jakimauskas, Gintautas; Sušinskas, Jurgis. Application of the empirical Bayes approach to nonparametric testing for high-dimensional data //

Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai, ISSN 0132-2818, T. 51, 2010, p. 402-407.

8. Jakimauskas, Gintautas; Sakalauskas, Leonidas. Empirical Bayesian regression model for estimation of small rates // Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai, ISSN 0132-2818, T. 53, ser. A, 2012, p. 42-47.

Recenzuojamuose tarptautinių konferencijų darbuose:

1. Gurevičius, Romualdas; Jakimauskas, Gintautas; Sakalauskas, Leonidas. Empirical Bayesian estimation of small mortality rates // 5th international Vilnius conference [and] EURO-mini conference "Knowledge-based technologies and OR methodologies for decisions of sustainable development" (KORSD-2009): September 30 – October 3, 2009, Vilnius, Lithuania. Vilnius: Technika, ISBN 9789955284826, 2009, p. 290-295.
2. Jakimauskas, Gintautas; Sakalauskas, Leonidas. Empirical Bayesian estimation for Poisson-gamma model // 24th Mini EURO conference on continuous optimization and information-based technologies in the financial sector (MEC EurOPT 2010). Vilnius: Technika, ISBN 9789955285984, 2010, p. 254-257.
3. Jakimauskas, Gintautas. Gamma and logit models in empirical Bayesian estimation of probabi-lities of rare events // STOPROG 2012: Stochastic programming for implementation and advanced applications: proceedings of international workshop, July 3-6, 2012, Lithuania. Vilnius: Technika, ISBN 9786099524146, 2012, p. 43-48.
4. Radavičius, Marijus; Jakimauskas, Gintautas. Robust projection pursuit // Computer data analysis and modeling: complex stochastic data and systems: proceedings of the Seventh international conference: Minsk, September 6-10, 2004. Vol. 1. Minsk: Publishing centre BSU, ISBN 985-445-492-4, 2004, p. 114-117.

5. Radavičius, Marijus; Jakimauskas, Gintautas; Sušinskas, Jurgis. Testing of independency for high-dimensional data // Computer data analysis and modeling: complex stochastic data and systems: proceedings of the Eighth international conference: Minsk, September 11-15, 2007. Vol. 1. Minsk: Publishing centre BSU, ISBN 9789854765082, 2007, p. 174-177.
6. Radavičius, Marijus; Jakimauskas, Gintautas; Sušinskas, Jurgis. Empirical Bayes testing goodness-of-fit for high-dimensional data // Computer data analysis and modeling: complex stochastic data and systems: proceedings of the Ninth international conference: Minsk, September 7-11, 2010. Vol. 1. Minsk: Publishing center BSU, ISBN 9789854768472, 2010, p. 199-202.
7. Sakalauskas, Leonidas; Jakimauskas, Gintautas; Sušinskas, Jurgis. Analysis of medical data by empirical Bayes method // Computer data analysis and modeling: complex stochastic data and systems: proceedings of the Ninth international conference: Minsk, September 7-11, 2010. Vol. 1. Minsk: Publishing center BSU, ISBN 9789854768472, 2010, p. 203-206.

1.9. DISERTACIJOS STRUKTŪRA

Disertaciją sudaro 5 skyriai, literatūros sąrašas ir priedai.

1-asis skyrius yra įvadinis. Jame pateikiama disertacijos tyrimų sritis, problemos aktualumas, tyrimų objektas, tyrimų tikslas ir uždaviniai, mokslinis naujumas, praktinė darbo reikšmė bei darbo rezultatų aprobavimas ir publikavimas.

2-ajame skyriuje pateikiamas didelės apimties ir didelio matavimo duomenų skaidymo algoritmas ir juo paremtos procedūros pritaikymas klasifikavimo procedūros skaičiavimų laiko sumažinimui. Kiti šios procedūros taikymai yra pateikiami sekančiame skyriuje, o jos aprašymas patogumo dėlei išskirtas į atskirą skyrių, taip pat priede pateikiamas algoritminis procedūros aprašymas.

Šiuo algoritmu paremta viena iš procedūrų iš 1992–1995 metais MII sukurtos programinės įrangos daugiamačių Gauso mišinių klasifikavimui.

3-ajame skyriuje pateikiami 2-ajame skyriuje aprašyto didelės apimties ir didelio matavimo duomenų skaidymo algoritmo taikymai.

3-iojo skyriaus 1-ajame skyrelyje pateikiama procedūra duomenų modelio verifikavimui. Šio skyrelio pagrindiniai rezultatai buvo pateikti žurnale *Liet. mat. rink. LMD darbai* – Jakimauskas (2009).

3-iojo skyriaus 2-ajame skyrelyje pateikiama procedūra neparametriniam didelio matavimo atsitiktinių vektorių nepriklausomumo testavimui. Šio skyrelio pagrindiniai rezultatai buvo pateikti žurnale *Austrian Journal of Statistics* – Jakimauskas, Radavičius, Sušinskas (2008).

3-iojo skyriaus 3-ajame skyrelyje pateikiama procedūra naudojanti empirinį Bajeso metodą, kuri leidžia gauti efektyvesnius hipotezių tikrinimo kriterijus uždaviniams pateiktiems pirmuose dviejuose skyreliuose. Šio skyrelio pagrindiniai rezultatai buvo pateikti žurnale *Liet. mat. rink. LMD darbai* – Jakimauskas, Sušinskas (2010).

4-ajame skyriuje pateikiami empirinio Bajeso metodo taikymai retų dažnių populiacijose analizei.

4-ojo skyriaus 1-ajame skyrelyje nagrinėjamas retų įvykių modeliavimas naudojant empirinį Bajeso metodą. Šio skyrelio pagrindiniai rezultatai buvo pateikti Vilniuje vykusios KORSD-2009 tarptautinės konferencijos darbuose – Gurevičius, Jakimauskas, Sakalauskas (2009). Naudojami Lietuvos Higienos instituto pateikti duomenys.

4-ojo skyriaus 2-ajame skyrelyje nagrinėjami Puasono-Gauso ir Puasono-gama modeliai empiriniame Bajeso mažų tikimybių vertinime. Šio skyrelio pagrindiniai rezultatai buvo pateikti Neringoje vykusios STOPROG-2012 tarptautinės konferencijos darbuose –Jakimauskas (2012). Šiame skyrelyje naudojami duomenys iš Lietuvos Statistikos departamento duomenų bazės.

4-ojo skyriaus 3-ajame skyrelyje pateikiamas modifikuotas regresinis empirinis Bajeso įvertis mažų tikimybių vertinimui. Šio skyrelio pagrindiniai rezultatai buvo pateikti žurnale *Liet. mat. rink. LMD darbai* – Jakimauskas, Sakalauskas

(2012). Šiame skyrelyje taip pat naudojami duomenys iš Lietuvos Statistikos departamento duomenų bazės.

5-ajame skyriuje pateikiami rezultatai ir išvados.

Pabaigoje pateikiamas literatūros sąrašas ir priedai.

2. DUOMENŲ SUTRAUKIMAS DUOMENŲ TYRYBOJE

Duomenų tyryba (*data mining*) yra šiuolaikinė informacijos analizės sritis, atsiradusi duomenų bazių technologijų, dirbtinio intelekto ir statistinės duomenų analizės sankirtoje. Duomenų tyryba yra labai plati sritis, apimanti daug metodų, algoritmų bei taikomųjų programinių sistemų.

Didelio duomenų rinkinio apdorojimas, sprendžiant pakankamai tiksliai ir pakankamai greitai įvairius uždavinius, yra vienas iš pagrindinių duomenų tyrybos uždavinių, diegiant interaktyvias analitinio apdorojimo sistemas (OLAP – *on-line analytical processing*). Apibrėžiant terminą „didelis duomenų rinkinys“, reikia atsižvelgti į turimos aparatinės (*hardware*) ir programinės (*software*) įrangos našumą, taip pat į sprendžiamą uždavinį, nes sudėtingesniems duomenų modeliams realizuoti net ir palyginti nedidelė duomenų apimtis gali sudaryti rimtų problemų.

Vienas iš sprendimo būdų yra naudoti našesnę aparatinę ar programinę įrangą, tačiau šis sprendimo būdas praktikoje ne visada prieinamas. Kitas sprendimo būdas – pakeisti pradinį duomenų rinkinį mažesniu, pagal galimybes išlaikant pradines duomenų rinkinio savybes. Trivialus šio sprendimo būdo pavyzdys yra atsitiktinis mažesnės apimties duomenų imties išrinkimas. Tačiau šiuo atveju, sprendžiant analizės uždavinius, atitinkamai padidėja parametrų įverčių dispersija, į ką būtina atsižvelgti. Imčių metodai yra labai plačiai taikomi tais atvejais, kai duomenų rinkimo sąnaudos yra didelės (pvz., demografinėje statistikoje, ūkinėje statistikoje, sociologiniuose tyrimuose ir pan.).

Tarkime, kad yra pateiktas didelis duomenų rinkinys, ir reikia pakeisti šį duomenų rinkinį mažesniu, maksimaliai išlaikant pagrindines duomenų rinkinio savybes ir panaudojant informaciją iš visų duomenų rinkinio elementų. Pradinio duomenų rinkinio pakeitimui mažesniu yra naudojami įvairūs metodai, kuriuos galima suskirstyti į dvi grupes – skaidymo (*partitioning*) ir hierarchiniai (*hierarchical*) metodai (Zhou, Sander, 2003). Skaidymo algoritmai suskaido duomenų rinkinį į klasterius, hierarchiniai algoritmai

pateikia hierarchinę klasterinę struktūrą, tačiau neapibrėžia pačių klasterių išreikštiniu pavidalu.

Pastaruoju metu dažnai naudojamas DuMouchel *et al.* (1999) pasiūlytas duomenų sutraukimo (*data squashing*) metodas. Šis metodas siekia sutraukti (*squash*) duomenis tokiu būdu, kad statistinė analizė, atliekama naudojant sutrauktus duomenis, duotų kiek įmanoma panašesnius rezultatus, kaip ir naudojant visą duomenų rinkinį. Tokiu būdu duomenų analizę galima atlikti su sutrauktais duomenimis įprastais metodais ir gauti žymiai tikslesnius rezultatus, nei naudojant panašaus dydžio išrinktą atsitiktinę imtį. Straipsnyje Madigan *et al.* (2002) pateikiamas pavyzdys, kaip logistinės regresijos uždavinyje su 750000 stebėjimų sutrauktas (*squashed*) duomenų rinkinys su 8443 stebėjimų davė apie 500 kartų mažesnę vidutinę regresijos koeficientų paklaidą, palyginti su atsitiktinai išrinktu rinkiniu iš 7543 stebėjimų. Minėtame straipsnyje naudojamas tikėtinumo funkcija paremtas duomenų sutraukimas (likelihood-based data squashing), kuris skiriasi nuo DuMouchel (2002) pasiūlyto duomenų sutraukimo tuo, kad duomenys nebūtinai skaidomi pagal stačiakampį tinklelį (*rectangular grid*), ir suskaidymai gali būti nereguliarūs.

Vienas iš plačiai naudojamų metodų klasifikavime ir regresinėje analizėje yra CART (*classification and regression tree*) metodas (žr., pvz. Hastie *et al*. (2001), nuoroda į laisvai prieinamą paskutinės versijos (2013 m.) *.pdf failą: *http://www-stat.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf*).

Šiame metode nagrinėjama erdvės sritis (daugiamatis stačiakampis gretasienis) yra pažingsniui skaidoma į padalijimus, kol patenkinama tam tikra sustojimo sąlyga. Naudojant grafų teorijos sąvokas, CART metodo atliekamus žingsnius galima pavaizduoti kaip medį (*tree*), prasidedantį iš šakninės viršūnės (*root node*), iš kurio tam tikros viršūnės (*node*) konkrečiame žingsnyje išeina dvi ar daugiau kraštinių (*edges*), besibaigiančių atitinkamomis viršūnėmis. Jei kiekviename žingsnyje padalijimo elementas skaidomas į dvi dalis (*binary partitions*), tai turime binarinio medžio (*binary tree*) metodą. Šis metodas daugiausia taikomas (kaip galima spręsti iš jo pavadinimo)

klasifikavimui ir regresinei analizei. Kaip pavyzdį pateiksime iliustraciją iš aukščiau minėto *.pdf failo, p. 306 (žr. Pav. 2.1).

Panagrinėsime CART algoritmo modifikaciją, skirtą greitam didelės apimties daugiamačių duomenų grupavimui. Šis uždavinys atitinka algoritmo pritaikymą regresinei analizei, tik naudojamas didelis padalijimų skaičius, ir taikomas pats paprasčiausias skaidymas, kad maksimaliai būtų sutrumpintas skaičiavimų laikas. Tikslas yra pritaikyti grupuotus duomenis Gauso mišinių

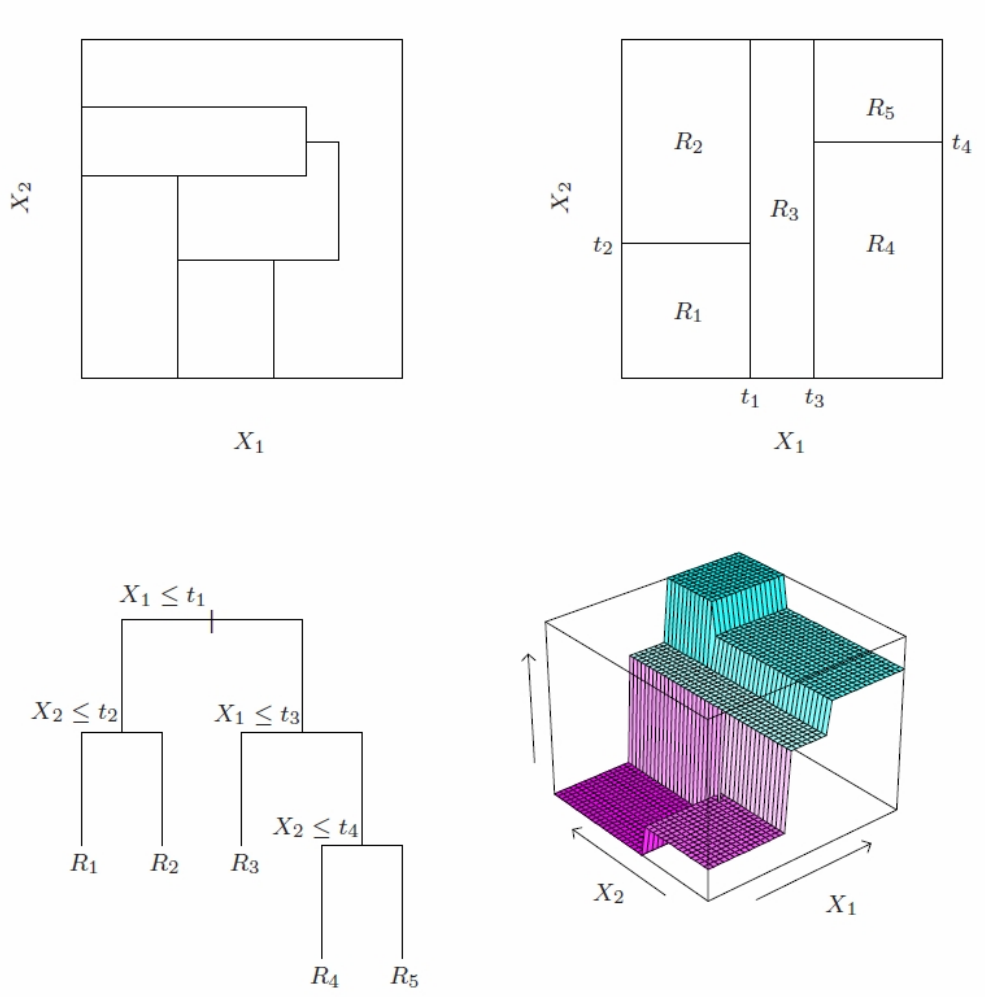


FIGURE 9.2. *Partitions and CART. Top right panel shows a partition of a two-dimensional feature space by recursive binary splitting, as used in CART, applied to some fake data. Top left panel shows a general partition that cannot be obtained from recursive binary splitting. Bottom left panel shows the tree corresponding to the partition in the top right panel, and a perspective plot of the prediction surface appears in the bottom right panel.*

Pav. 2.1. CART procedūros pavyzdys

klasifikavimo procedūros tam tikrame etape, kad būtų sumažintas bendras skaičiavimų laikas.

Tarkime, kad nagrinėjame imtį X^N, kurios ilgis yra N. Jei N yra pakankamai didelis, pvz. N=2000 (programinis apribojimas N=10000), nagrinėjamą imtį galima sugrupuoti ir gauti trumpesnę imtį $X_{Grp}{}^{NGrp}$, kur $NGrp$ gerokai mažesnis už N (pvz. $NGrp$ gali būti apie 150). Tuomet pirmieji klasifikavimo procedūros veiksmai atliekami su pradine imtimi X^N, toliau pagrindiniai skaičiavimai atliekami su grupuota imtimi $X_{Grp}{}^{NGrp}$ (atliekama daug ciklų pridedant naują klasterį, panaikinant tam tikrą klasterį, naudojant tam tikrą EM algoritmo iteracijų, tikrinant modelio adekvatumą ir t. t.). Kuomet šie skaičiavimai baigiami, grįžtama prie pradinės sekos X^N ir atliekami galutiniai skaičiavimai ir patikslinimai su visa seka. Žinoma, labai svarbu, kad grupavimo procedūra veiktų kaip galima greičiau, nes tik tada galima tikėtis bendro klasifikavimo procedūros sunaudojamo laiko sumažėjimo.

1991–1993 metais MII buvo sukurta programinė įranga daugiamačių Gauso mišinių klasifikavimui. Ši programinė įranga buvo sukurta užsakius Maskvos CEMI (Centralnyi Ekonomiko-Matematicheskyi Institut), kur ji buvo kruopščiai testuota ir kaip vienas iš punktų buvo įjungta į bendrovės Stat-Dialogue sukurtą programinę įrangą Class Master, kuri vėliau buvo platinama kaip komercinė programinė įranga. Vienas iš naudojamų originalių algoritmų MII pateiktoje programinėje įrangoje yra didelio matavimo binarinė duomenų skaidymo procedūra, šiuo atveju naudojama kaip duomenų grupavimo procedūra, siekiant sumažinti visos klasifikavimo procedūros naudojamą laiką.

2.1. MATEMATINIS ALGORITMO APRAŠYMAS

Tegul $\mathbf{X} = (X(1),...,X(N))$ tam tikra stebėjimų erdvėje $\mathbf{R}^d$ imtis, t. y.

$$X(j) = (X_1(j), X_2(j),..., X_d(j))^{\mathrm{T}} \in \mathbf{R}^d,\ j = 1,2,...,N.$$

Taip pat naudosime imties stebėjimų pasikartojimų skaičių (kas dažnai pasitaiko praktiniuose uždaviniuose) q_j, $j = 1, 2, \ldots, N$, t. y. stebėjimas $X(j)$ pasikartoja q_j kartų. Jei pasikartojimų nėra, tai tiesiog priskiriamos reikšmės $q_j = 1, j = 1, 2, \ldots, N$.

Tarsime, kad turime d-matį stačiakampį gretasienį S, nusakomą atitinkamais kiekvienos ašies intervalais $[a_i, b_i]$, $i = 1, 2, \ldots, d$, (jį galima parinkti laisvai, tačiau rekomenduojama jį parinkti kaip galima mažesnį, pvz., pagal minimalias ir maksimalias reikšmes kiekvienoje ašyje), tokį, kad visi stebėjimai telpa į šį stačiakampį gretasienį:

$$a_i \le X_i(j) \le b_i,\ i = 1,2,\ldots,d,\ j = 1,2,\ldots,N.$$

Procedūros tikslas yra skaidyti minėtąjį d-matį stačiakampį gretasienį į rinkinius d-mačių stačiakampių gretasienių (kurių kiekviename yra bent vienas imties $\mathbf{X}$ taškas) $S(k) = \{S_k(1), S_k(2),\ldots, S_k(k)\}$, $k = 1,2,\ldots,k_{\max}$, $k_{\max} \le N$ (pagal apibrėžimą $S(1) = S$), taip, kad kiekviename žingsnyje k minimizuotume imties $\mathbf{X}$ grupavimo paklaidą. Yra iš anksto nustatomas maksimalus skaičius M, $M > 1$, kuris yra dvejeto laipsnis , t. y. $M \in \{2, 4, 8, 16,\ldots\}$, kuris nusako, į kiek dalių maksimaliai galima suskaidyti kiekvieną ašį (pvz., $M = 256$).

Kiekvienam užbaigtam žingsniui k pažymėkime grupavimo taškus

$$Z_k(j) = \frac{\sum_{l=1}^{N} 1_{X(l)\in S_k(j)} \cdot q_l \cdot X(l)}{\sum_{l=1}^{N} 1_{X(l)\in S_k(j)} \cdot q_l},\ j = 1,2,\ldots,k,$$

ir bendrą imties grupavimo paklaidą po k-tojo žingsnio

$$E_k^2 = E_k^2(\mathbf{X}, S(k)) = \sum_{j=1}^{k} E_k^2(j, \mathbf{X}, S(k)) = \sum_{j=1}^{k}\sum_{l=1}^{N} 1_{X(l)\in S_k(j)} \cdot q_l \cdot |X(l) - Z_k(j)|^2 .$$

Be to, baigiant k žingsnį, iš anksto apskaičiuojami (k+1) žingsnio bendros imties grupavimo paklaidos sumažėjimai. Po vieną parenkamas stačiakampis gretasienis $S_k(m)$, $m = 1,2,...,k$, rinkinio tvarka yra pakeičiama taip, kad šis stačiakampis gretasienis būtų paskutinis rinkinyje. Pirmuosius k–1 stačiakampius gretasienius pažymėkime

$$\{S_{k+1}^m(1), S_{k+1}^m(2),..., S_{k+1}^m(k-1)\} = \{S_k^m(1), S_k^m(2),..., S_k^m(k-1)\},$$

kadangi jie tiek k žingsnyje, tiek k+1 žingsnyje lieka tokie patys. Paskutinis rinkinio stačiakampis gretasienis $S_k^m(k)$ suskaidomas į du – $\widetilde{S}_{k+1}^{m,v}(k)$ ir $\widetilde{S}_{k+1}^{m,v}(k+1)$ (dalijant atitinkamos ašies kraštinę į dvi lygias dalis) pagal kiekvieną ašį $v = 1,2,...,d$. Kiekvienas iš abiejų gautų stačiakampių gretasienių yra „apkarpomas“ pagal visas ašis, t.y., jei padalijus pusiau vienoje pusėje nėra nė vieno stebėjimo, ta pusė atmetama (jei reikia, atmetimas atliekamas keletą kartų). Tuo būdu, kiekvienam $m = 1,2,...,k$, gauname rinkinius

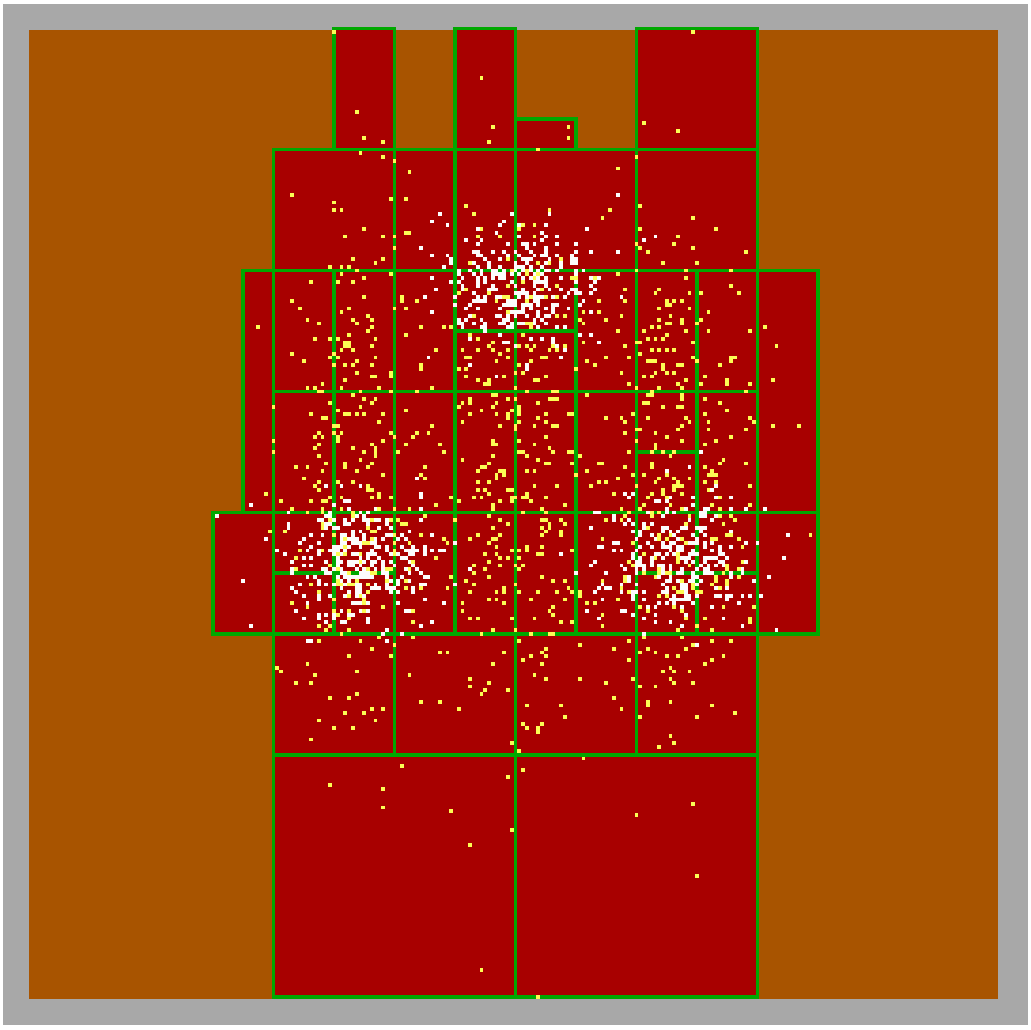
$$S^{m,v}(k+1) = \{S_{k+1}^m(1), S_{k+1}^m(2),..., S_{k+1}^m(k-1), \widetilde{S}_{k+1}^{m,v}(k), \widetilde{S}_{k+1}^{m,v}(k+1)\}, v = 1,2,...,d.$$

Atitinkamai kiekvienam rinkiniui apskaičiuojamas grupavimo paklaidos sumažėjimas, pasinaudojant tuo, kad reikia skaičiuoti tik $S_k^m(k)$ grupavimo paklaidos sumažėjimą. Tuo būdu, kiekvienam $m = 1,2,...,k$, gauname reikšmes

$$\Delta^2(k+1,m,v) =$$
$$= E_k^2(k,\mathbf{X},S_k^m(k)) - (E_{k+1}^2(k,\mathbf{X},\widetilde{S}_{k+1}^{m,v}(k) + E_{k+1}^2(k,\mathbf{X},\widetilde{S}_{k+1}^{m,v}(k+1))), v = 1,2,...,d,$$

ir apskaičiuojame

$$v^* = v^*(k+1,m) = \arg\max_{v=1,2,...,d} \Delta^2(k+1,m,v).$$



Pav. 2.1.1: Skaidymo procedūros pavyzdys

Tuo būdu, prieš atlikdami k+1 žingsnį, žinome kad k+1 žingsnyje skaidyti reikės (jei šis žingsnis bus atliekamas) rinkinio $S(k)$ stačiakampį gretasienį $S_k(m^*)$ pagal ašį v^*. Todėl paprastumo dėlei šį stačiakampį gretasienį perkeliame į paskutinę vietą, t. y. pakeistas rinkinys bus

.

$$S_k = \{S_k(1), S_k(2), ..., S_k(m^*-1), S_k(m^*+1), S_k(m^*+2), ..., S_k(k), S_k(m^*)\}$$

Po to sekančiu žingsniu apskaičiuojame

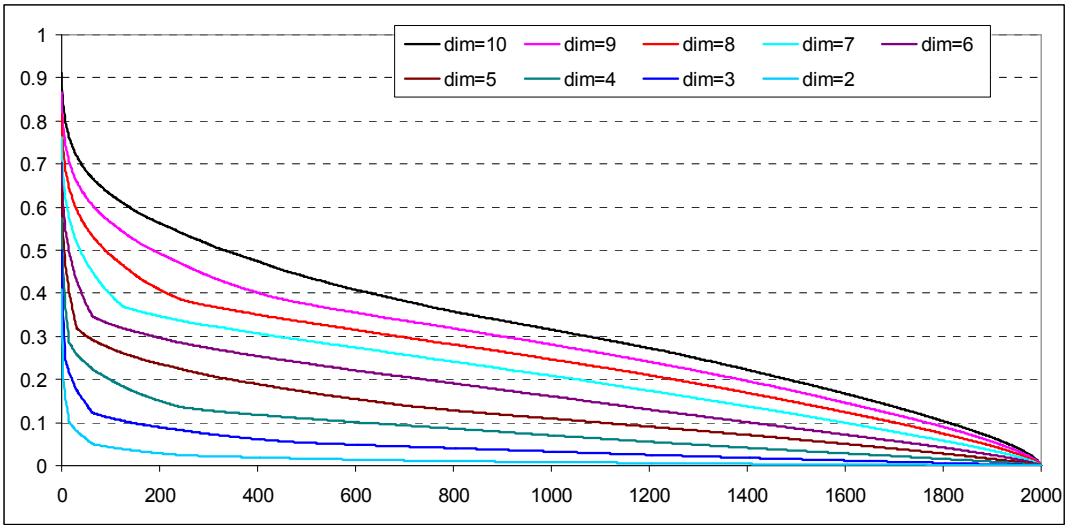
$$m^* = m^*(k+1) = \arg\max_{m=1,2,...,k} \Delta^2(k+1, m, v^*(k+1, m)) .$$

Kaip iliustracija pateiktas skaidymo procedūros pavyzdys (žr. Pav. 2.2.1). Šioje iliustracijoje parodyta skaidymo procedūra po tam tikro tarpinio žingsnio, kuomet duomenys yra dvi standartizuotos imtys (sujungtos į vieną), iš kurių pirmoji tenkina trijų komponenčių dvimatį Gauso mišinių modelį, o antrosios imties skirstinys yra dvimatis standartinis Gauso. o taip pat išsaugome (kadangi kitu žingsniu daugumos dydžių nereikės perskaičiuoti) atitinkamus jau apskaičiuotus dydžius $\{v^*(k+1,m), m=1,2,...,k\}$ ir $\{\Delta^2(k+1,m,v^*(k+1,m), m=1,2,...,k\}$.

Algoritmas baigiamas, kai visi stačiakampiai gretasieniai suskaidomi iki mažiausio dydžio, nusakomo parametru M. Pastebėsime, kad maksimalus žingsnių skaičius $k_{\max}$ gali būti mažesnis už N, priklausomai nuo imties, parinkto pradinio stačiakampio gretasienio S ir parametro M. Algoritmą galima pabaigti ir anksčiau, jei pasiekiamas norimas padalijimų skaičius k, arba pasiekiama norima normuota vidutinė imties grupavimo paklaida.

2.2. MODELIAVIMO REZULTATAI

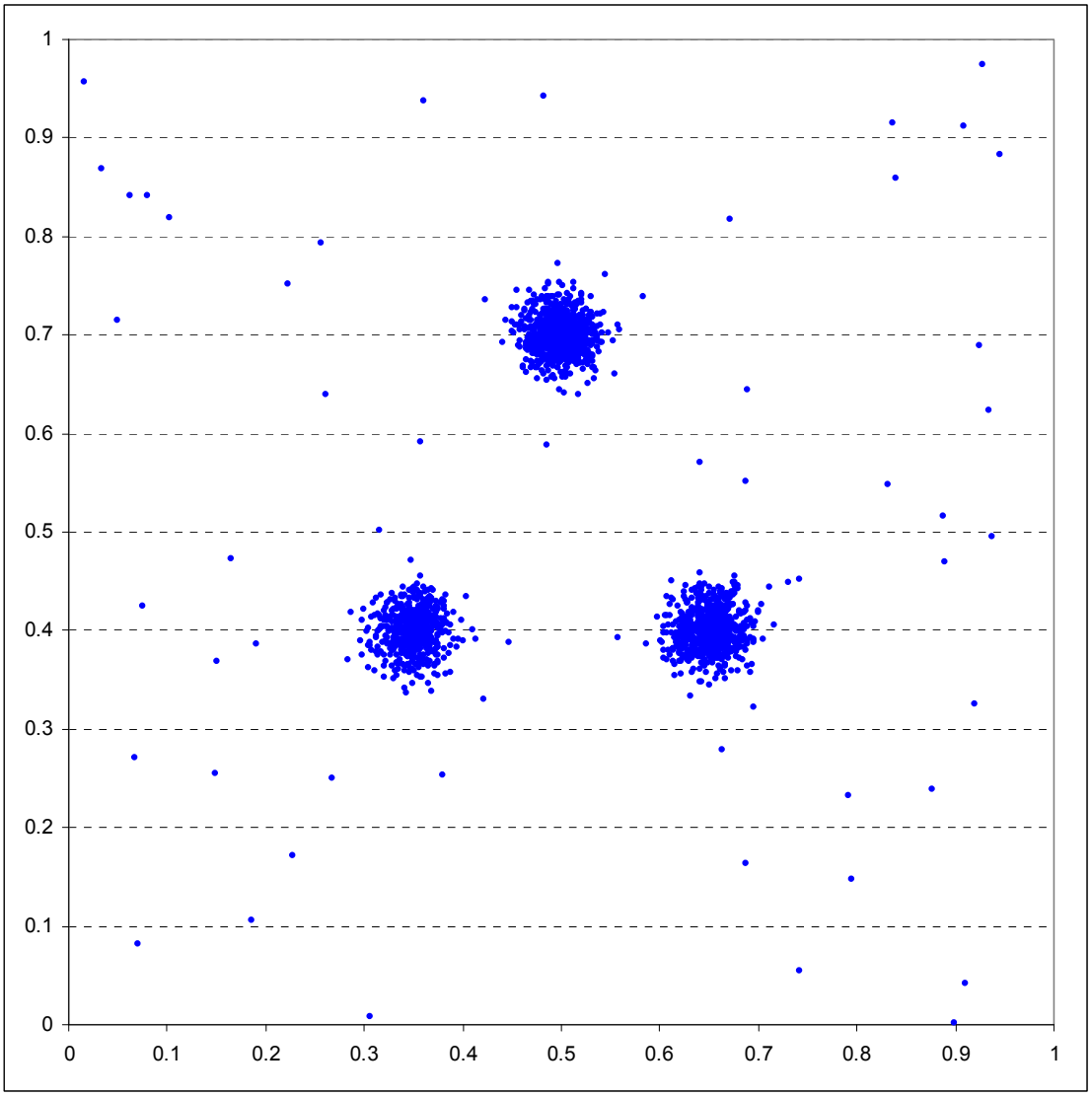
Pirmiausia panagrinėsime atvejį, kuomet taškai yra tolygiai pasiskirstę vienetiniame d-mačiame kube.



Pav. 2.2.1. Normuotos vidutinės paklaidos kitimas priklausomai nuo matavimo tolygaus skirstinio atveju

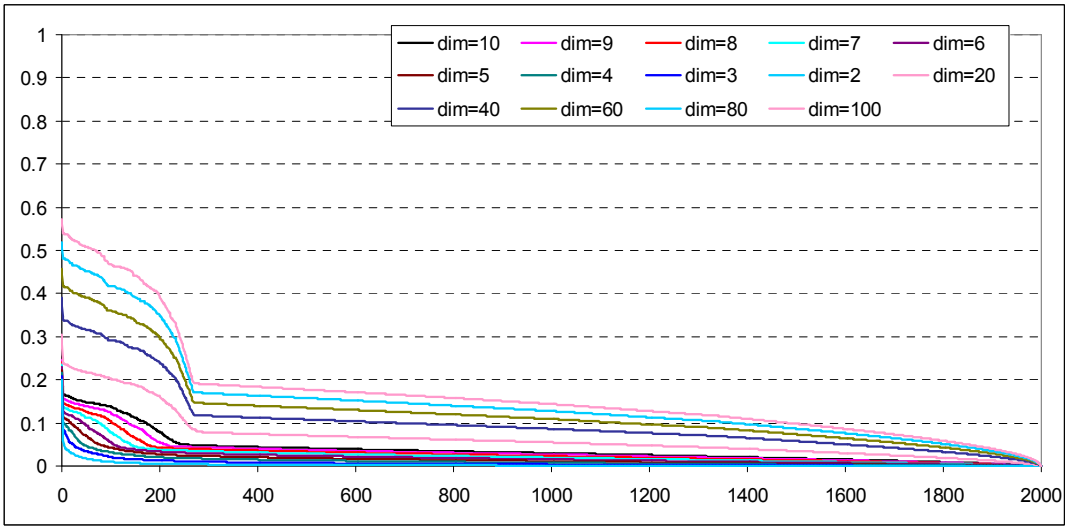
Palyginsime atskirų realizacijų normuotos vidutinės paklaidos (pagal aukščiau pateiktus pažymėjimus $(E_k^2/N)^{1/2}$) kitimą priklausomai nuo grupuotos imties ilgio $k = 1,2,...,N$, (nagrinėtos fiksuoto ilgio $N = 2000$ atsitiktinės sekos) esant skirtingiems erdvės matavimams (žr. Pav. 2.2.1).

Toliau panagrinėsime atvejį, kuris yra kaip tik tinkamas nagrinėjamos procedūros pritaikymui (žr. Pav. 2.2.2). Nagrinėsime taškus vienetiniame *d*-mačiame kube (imties dydis $N = 2000$), kurių nedidelė dalis (nagrinėjamu atveju 3 proc.) yra tolygiai pasiskirstę vienetiniame *d*-mačiame kube, o kiti taškai gaunami modeliavimui taikant *d*-matį Gauso mišinių modelį iš trijų

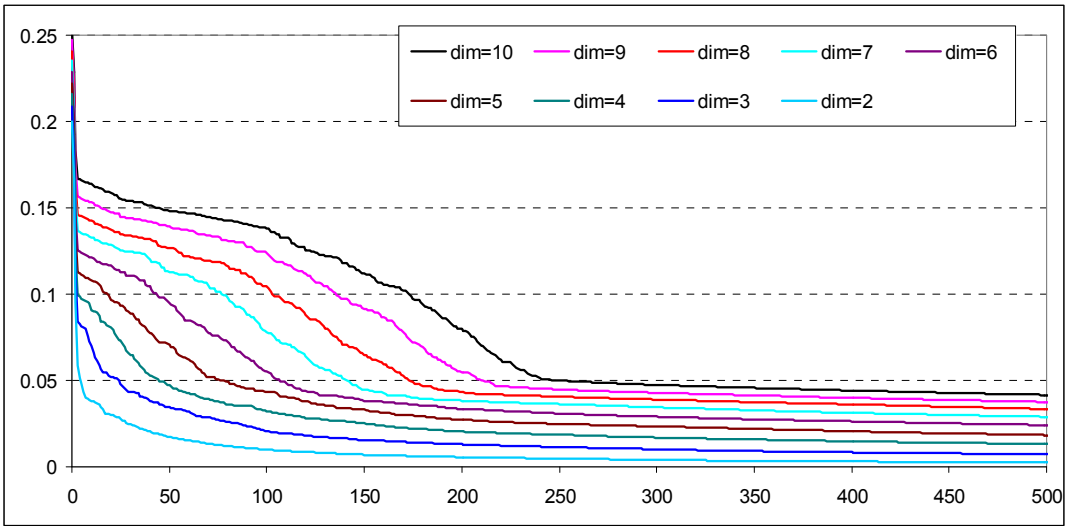


Pav. 2.2.2. Gauso mišinio iš trijų komponenčių pavyzdys, kai $d = 2$.

komponenčių (lygiomis tikimybėmis) su vienetinėmis kovariacinėmis matricomis, padaugintomis iš konstantos $c = 0.02$, o komponenčių vidurkiai yra (0.5–0.15, 0.5–0.1, 0.5, 0.5, ... , 0.5), (0.5, 0.5+0.2, 0.5, 0.5, ... , 0.5), (0.5+0.15, 0.5–0.1, 0.5, 0.5, ... , 0.5).



Pav. 2.2.3. Normuotos vidutinės paklaidos kitimas priklausomai nuo matavimo Gauso mišinio iš trijų komponenčių atveju



Pav. 2.2.4. Normuotos vidutinės paklaidos kitimas (kai $k = 1,2,...,500$) priklausomai nuo matavimo Gauso mišinio iš trijų komponenčių atveju

Šiuo atveju taip pat pateiksime atskirų realizacijų normuotos vidutinės paklaidos $(E_k^2 / N)^{1/2}$ kitimą priklausomai nuo grupuotos imties ilgio $k = 1,2,...,N$, $N = 2000$, esant skirtingiems erdvės matavimams (žr. Pav. 2.2.3–2.2.4).

Iš gautų rezultatų matome, kad nagrinėjamu atveju normuota vidutinė paklaida pirmiausia gerokai sumažėja po pirmųjų padalijimų (tai lemia skirtingų didelių klasterių atskyrimas, žr. Pav. 2.2.3), po to paklaidos mažėjimą lemia atskirų taškų atskyrimas, kuris baigiasi maždaug intervale nuo $k = 150$ iki $k = 250$ (žr. Pav. 2.2.4), o po to yra tik nedidelis daugiau mažiau tolygus normuotos vidutinės paklaidos mažėjimas (žr. Pav. 2.2.3).

2.3. IŠVADOS

Sutraukiant didelius duomenų rinkinius bei siekiant išsaugoti pagrindines duomenų rinkinio savybes ir panaudojant informaciją iš visų duomenų rinkinio elementų, dažnai naudojami skaidymo algoritmai. Iš darbe gautų tyrimo rezultatų galima padaryti išvadą, kad sudaryta skaidymo procedūra labiausiai tinka duomenims su išreikšta klasterine struktūra. Atlikus palyginti nedidelį žingsnių skaičių galima žymiai sumažinti pradinę normuotos vidutinės paklaidos reikšmę iki tokio lygmens, kad, viena vertus, gauname gerokai trumpesnę grupuotą duomenų seką, antra vertus, ši normuotos vidutinės paklaidos reikšmė leidžia pakankamai tiksliai atlikti skaičiavimus su grupuota seka (vietoje pradinės sekos), taip sumažinant skaičiavimų laiką.

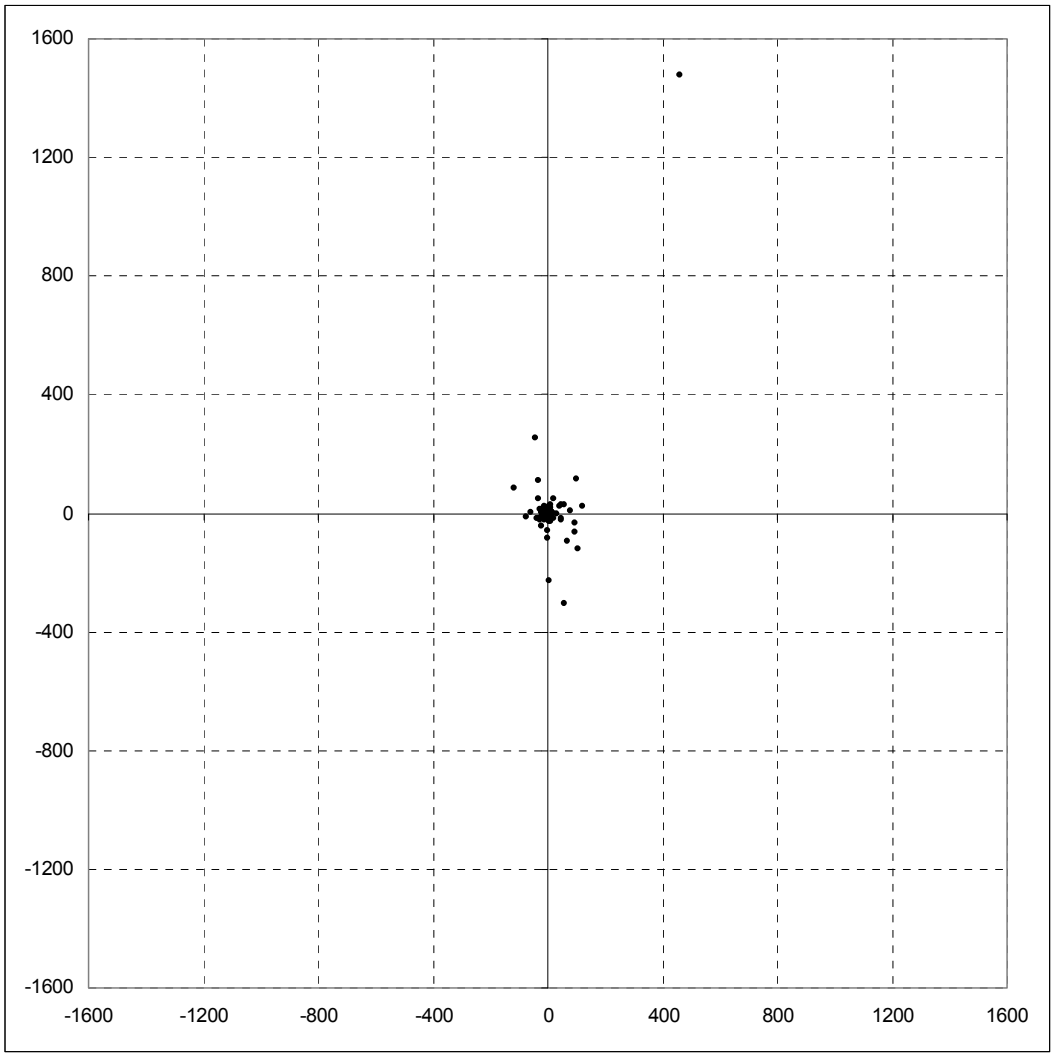
3. SKAIDYMO PROCEDŪRA DUOMENŲ MODELIO VERIFIKAVIMUI

Šiame skyriuje nagrinėjamas ankstesniame skyriuje sudarytos skaidymo procedūros taikymas duomenų modelio verifikavimui, požymių nepriklausomumo tikrinimui ir hipotezių apie duomenų poslinkį tikslinimui empiriniu Bajeso metodu.

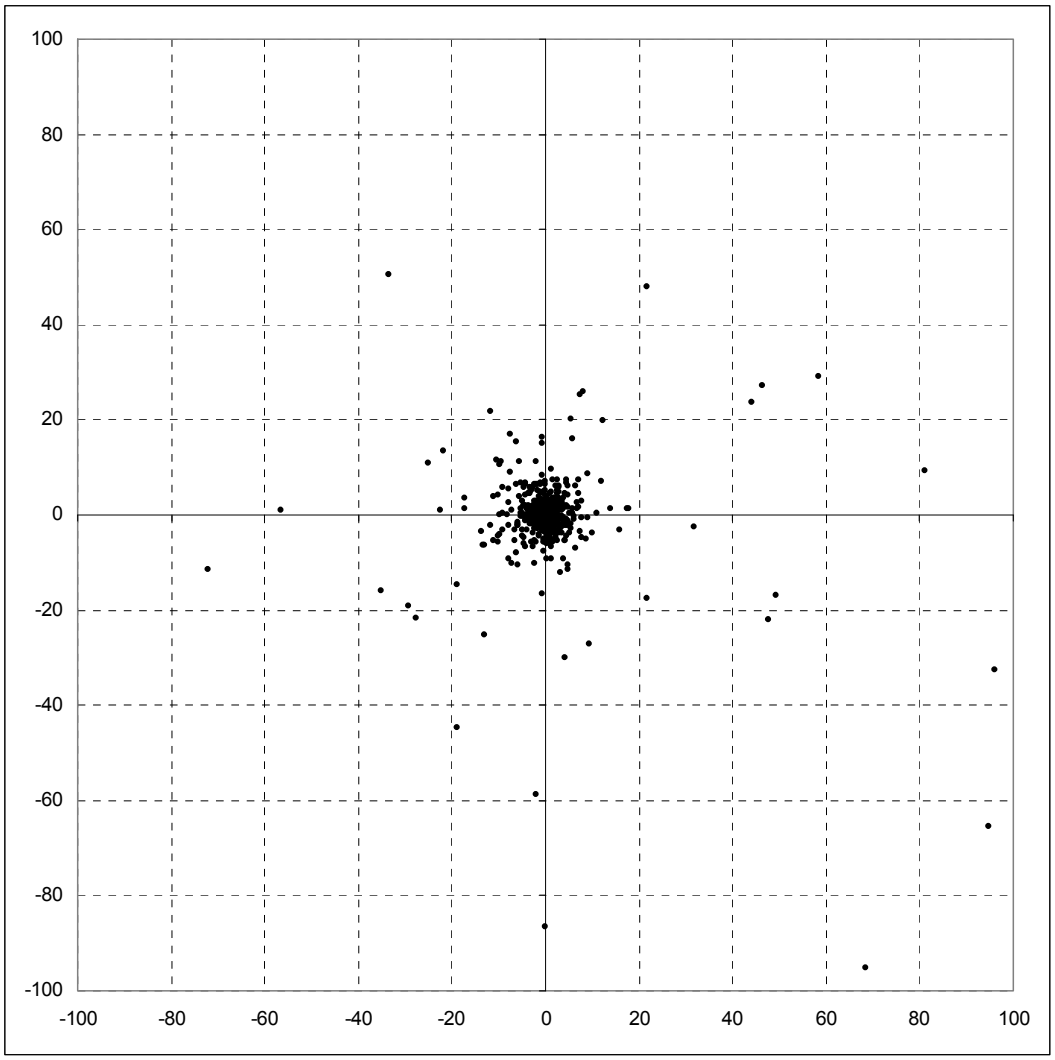
Sprendžiant duomenų tyrybos uždavinius su didelio matavimo duomenimis tiesiogiai taikyti daugumą statistinių metodų yra praktiškai neįmanoma dėl labai didelės skaičiavimų apimties. Pvz., naudoti įprastą Gauso mišinių modelį duomenų klasifikavimui esant dideliems matavimams nėra galimybių dėl didelio parametrų skaičiaus ir įverčių paklaidų, kurios sparčiai auga, didėjant matavimo skaičiui. Tačiau yra galimybė taikyti šį modelį duomenims, suprojektuotiems į nedidelio matavimo poerdvį, jei projektuojant į šį poerdvį neprarandama informacija apie aposteriorines tikimybes. Tikrinant požymių nepriklausomumą didelio matavimo duomenims, klasikinių įverčių paklaidos sparčiai didėja augant duomenų matavimui, todėl praktikoje dažniausiai apsiribojama žymiai paprastesniais koreliacinės analizės metodais. Vis dėlto, jei duomenys yra nekoreliuoti, tenka ieškoti metodų, kurie galėtų pakankamai efektyviai išspręsti požymių nepriklausomumo uždavinį tokio tipo duomenims.

Nagrinėjant duomenų modelio verifikavimą (*goodness-of-fit*) galima panaudoti tam tikrą tiesioginį metodą patikrinti modelio adekvatumą. Tarkime, kad d-matėje erdvėje turime nepersikertančias aibes A_1, A_2. ... , A_k, ir žinomos tikimybės d-mačiam atsitiktiniam dydžiui su žinomu skirstiniu (atitinkančiu tikrinamą modelį) patekti į šias aibes, t. y., $p_1 = \mathbf{P}(A_1)$, $p_2 = \mathbf{P}(A_2)$, ... , $p_k = \mathbf{P}(A_k)$. Žinoma, svarbu paimti aibes taip, kad tikimybių suma būtų artima vienetui. Tuomet nagrinėjant duomenų modelio verifikavimą d-mačiams duomenims $X^N = (X_1, X_2, \ldots, X_N)$, kurių ilgis yra N, galima lyginti tikimybes p_1, p_2, ... , p_k su atitinkamomis empirinėmis tikimybėmis q_1, q_2, ... , q_k, kur q_j yra imties X^N taškų, patekusių aibę A_j, skaičius, padalintas iš N , $j = 1, 2, \ldots k$.

Pateiksime pavyzdį, kuris gerai atspindi pasirinktą metodą. Nagrinėsime atvejį, kai matavimas $d = 2$. Tarkime, kad turime Cauchy skirstinį (jo komponentės yra nekoreliuotos, tačiau priklausomos) ir jo realizaciją, $N = 1000$ (žr. Pav. 3.1). Kai kurie šios realizacijos elementai įgyja dideles absoliutines reikšmes. Pav. 3.2 pateikta dalis realizacijos, kai reikšmės kiekvienoje ašyje moduliu neviršija 100, o Pav. 3.3 palyginimui pateikta ta pati realizacija, kurios antrosios komponentės indeksai yra sumaišyti atsitiktine tvarka (kas neturėtų pakeisti skirstinio, jei komponentės būtų nepriklausomos). Kaip matome, sumaišius antrosios komponentės indeksus, pasikeičia skirstinys (atsiranda charakteringas „kryžius“) ir atsiranda galimybė tikrinti komponenčių nepriklausomumo hipotezę lyginant abiejų realizacijų skirstinius.

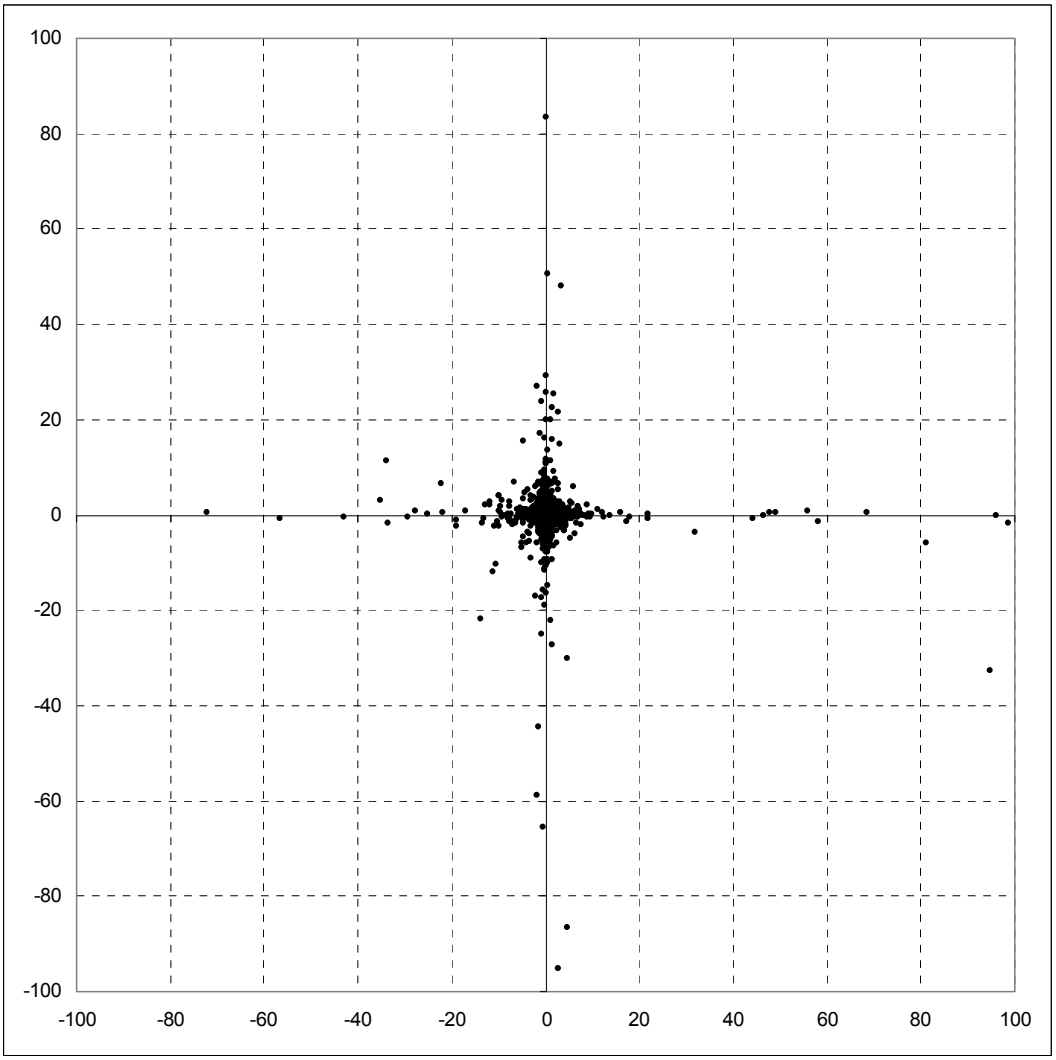


Pav. 3.1. Cauchy skirstinio pavyzdys, kai $d = 2$.



Pav. 3.2. Cauchy skirstinio pavyzdys, kai $d = 2$ (parodytos tik reikšmės, moduliu neviršijančios 100 kiekvienoje ašyje).

Pastaruoju metu nėra visuotinai pripažintos metodologijos didelio matavimo daugiamačių neparametrinių hipotezių testavimui. Tradiciniai metodai testuojant daugiamates neparametrines hipotezes remiasi empirine charakteringąja funkcija (Baringhaus, Henze, 1988), neparametriniais pasiskirstymo tankio įverčiais ir glodinimu Bowman, Foster (1993), Huang (1997), daugiamačiais neparametriniais Monte-Carlo testais, Zhu, Neuhaus (2000), Bajeso metodu ir Monte-Carlo Markovo grandinių metodu, Verdinelli, Wasserman (1998), ir klasikinėmis vienmatėmis neparametrinėmis statistikomis, naudojamomis duomenims, projektuotiems į kryptis, rastas duomenų tyryboje naudojamu tikslinio projektavimo (*projection pursuit*)



Pav. 3.3. Cauchy skirstinio realizacija, atsitiktine tvarka sumaišius antrosios komponentės indeksus, kai $d = 2$ (parodytos tik reikšmės, moduliu neviršijančios 100 kiekvienoje ašyje).

metodu Zhu *et al.* (1997), Szekely, Rizzo (2005). Sudėtingesni metodai remiasi Vapnik-Chervonenkis teorija, tolydžia funkcionaline centrine ribine teorema ir didelių nuokrypių tikimybių nelygybėmis Vapnik (1988), Bousquet *et al.* (2004).

Straipsnyje Jakimauskas *et al.* (2008) buvo pasiūlyta efektyvi, paremta duomenimis ir skaičiavimų prasme procedūra neparametriniam didelio matavimo atsitiktinių vektorių nepriklausomumo testavimui. Ši procedūra paremta randomizacija ir saviranka (*bootstrap*), specialia ankstesniame

skyriuje aprašyta pažingsnine duomenų skaidymo procedūra ir χ^2-tipo statistika. Straipsnyje Jakimauskas (2009) ši procedūra buvo pritaikyta testuojant tikslinio projektavimo metode tikrinamą hipotezę, teigiančią, kad papildomoje erdvėje $\mathbf{R}^{d-k}$ yra standartinis Gauso skirstinys, o erdvėje $\mathbf{R}^k$ yra daugiamatis Gauso mišinys tam tikram skaičiui k. Straipsnyje Jakimauskas, Sušinskas (2010) buvo pasiūlytos statistikos, kurios turi didesnę galią palyginti su χ^2-tipo testinėmis statistikomis, ir kurios paremtos neparametriniais didžiausio tikėtinumo ir empiriniais Bajeso įverčiais pagalbiniame mišinių modelyje.

3.1. MODELIO ADEKVATUMO TESTAVIMO ALGORITMAS DIDELIO MATAVIMO DUOMENŲ KLASIFIKAVIMUI

3.1.1. Tikslinio projektavimo metodas

Tikslinis projektavimas (*projection pursuit*) naudojamas duomenų tyryboje tiriamų požymių skaičiui sumažinti (žr., pvz. Friedman, Tukey (1974), Aivazyan *et al.* (1983)).

Tegul $X = X^N$ yra dydžio N imtis, tenkinanti matavimo d Gauso mišinių modelį (tegul matavimas d yra didelis) su pasiskirstymo funkcija (p.f.) F.

Kadangi erdvės matavimas yra didelis, natūralu projektuoti imtį X į matavimo k (k = 1, 2, …) tiesinius poerdvius naudojant tikslinio projektavimo metodą (žr., pvz., Friedman (1987)) Aivazyan (1996). Jei standartizuotos projektuotos imties papildomoje erdvėje yra standartinis Gauso, šis tiesinis poerdvis H vadinamas diskriminantiniu poerdviu. Pvz., jei turime q Gauso mišinio komponenčių su lygiomis kovariacinėmis matricomis, tuomet diskriminantinio poerdvio matavimas yra q–1.

Turint diskriminantinio poerdvio įvertį, žymiai lengviau atlikti klasifikavimą naudojant suprojektuotą imtį (žr. pvz., Jakimauskas, Krikštolaitis (2000a, 2000b)), .

Pažingsninė procedūra, kuri taikoma standartizuotai imčiai, yra tokia (k = 1, 2,…, kol diskriminantinio poerdvio hipotezė yra patenkinta tam tikram k):

1. Geriausio matavimo k tiesinio poerdvio radimas tikslinio projektavimo metodu (žr., pvz., Behboodian (1970), Rudzkis, Radavičius (1997), (1999), Radavičius, Jakimauskas (2004)).

2. Gauso mišinių modelio įvertinimas iš imties, suprojektuotos į matavimo k tiesinį poerdvį (žr., pvz., Hasselblad (1966), Behboodian (1970), Everitt, Hand (1981), Rudzkis, Radavičius (1995), Jakimauskas (1997), Jakimauskas (2002)).

3. Įvertinti modelio adekvatumą (*goodness-of-fit*) įvertinto matavimo d modelio, laikant, kad pasiskirstymas papildomoje erdvėje yra standartinis Gauso. Jei testas nepatenkinamas, padidinamas k ir einama į žingsnį 1.

Problemos, susijusios su pirmu ir antru žingsniais, nagrinėjamos minėtuose straipsniuose. Jei naudojame plačiai priimtus metodus trečiame žingsnyje, problema yra tam tikro neparametrinio tankio įverčio lyginimas su tam tikru parametriniu tankio įverčiu didelio matavimo erdvėje. Problemos, susijusios su didelio matavimo duomenimis, dažnai vadinamos didelio matavimo prakeiksmu (*curse of dimensionality*), žr., pvz., Hastie *et al.* (2001)). Kaip alternatyvų metodą naudosime Monte-Carlo metodą ir specialią duomenų skaidymo procedūrą. Tiksliau, iš naujo generuosime turimą imtį laikydami, kad papildomoje erdvėje skirstinys yra standartinis Gauso. Testinei statistikai naudosime apjungtą imtį ir kiekviename padalijimo elemente suskaičiuosime taškų (t. y. imties elementų) skaičių, priklausančių pradinei imčiai ir priklausančių generuotai imčiai. Testinė statistika parenkama taip, kad jei hipotezė yra teisinga, testinės statistikos skirstinys silpnai priklausytų nuo matavimo d ir nuo pasiskirstymo tiesiniame poerdvyje. Testinis kriterijus gaunamas modeliuojant pakankamai didelį skaičių (pvz., 100 ar 1000) nepriklausomų generuotų imčių kurioms hipotezė yra teisinga, ir palyginant testinio kriterijaus reikšmę su iš anksto nustatytu lygiu.

Kriterijaus galia remiasi silpna testinio kriterijaus priklausomybe nuo matavimo d ir nuo skirstinio tiesiniame poerdvyje. Efektyvumas skaičiavimų prasme remiasi labai efektyvia binarine duomenų skaidymo procedūra ir paprastu testinės statistikos skaičiavimu.

Pateiksime kai kuriuos modeliavimo rezultatus. Naudojami metodai gali būti pritaikomi ir kitose situacijose, pvz., testuojant didelio matavimo atsitiktinių vektorių nepriklausomumą.

3.1.2. Projektuotų duomenų adekvatumo testavimas

Tarkime, kad turime standartizuotą matavimo *d* Gauso mišinių modelį su p.f. *F*. Pažymėkime F_H matavimo *d* Gauso mišinių modelio, kurio pasiskirstymas papildomame matavimo *d–k* poerdvyje yra standartinis Gauso, p.f. (priminsime, kad *k* yra tiesinio poerdvio, gauto tikslinio projektavimo metodu, matavimas). Nagrinėsime mišinio modelį

$$F_{(p)} = (1-p)F_H + pF, \quad p \in (0,1),$$

dviejų populiacijų Ω_H ir Ω su p.f. F_H ir *F*, atitinkamai. Fiksuokime *p* ir tegul *Y* yra atsitiktinis vektorius su p.f. $F_{(p)}$. Tegul $\pi(Y)$ yra populiacijos Ω sąlyginė tikimybė su sąlyga *Y*, t. y.

$$\pi(Y) = \mathbf{P}\{\Omega \mid Y\} = \frac{pf(Y)}{pf(Y) + (1-p)f_H(Y)}.$$

Čia *f* ir f_H yra atitinkamai p.f. *F* ir F_H pasiskirstymo tankiai.

Tegul X_H yra dydžio *M* n.v.p. atsitiktinių vektorių iš Ω_H, nepriklausančių nuo *X*, imtis. Apjungta imtis žymima *Y*, o *Z*(*j*), *j* = 1, 2, … , *N*+*M*, yra atitinkama populiacijos Ω indikatorių seka. Tegul $P = \{P_k, k = 0, 1, \dots, K\}$, $P_0 = \mathbf{R}^d$, yra erdvės $\mathbf{R}^d$ padalijimų seka, priklausoma nuo *Y* ir tegul $\{A_k, k = 0, 1, \dots, K\}$ yra atitinkamų σ-algebrų, generuotų šių padalijimų, seka. Skaičiavimų prasme efektyvus *P* pasirinkimas binarinis pakoordinatinis padalijimas minimizuojant kiekviename žingsnyje vidutinę kvadratinę paklaidą padalijimo elementuose. Natūralus testinės statistikos pasirinkimas yra χ^2-tipo statistika

$$T_k = \hat{\mathbf{E}}(Z_k - p)^2, \quad p = N/(N+M),$$

kur $\hat{\mathbf{E}}$ yra vidurkis pagal *Y* empirinį pasiskirstymą $\hat{F}$, o $Z_k = \hat{\mathbf{E}}(Z \mid A_k),\ k \in \{1, 2, \dots, K\}$.

3.1.3. Kompiuterinio modeliavimo rezultatai

Kompiuteriniam modeliavimui buvo parinkta $M = N$, ir kriterijaus statistika

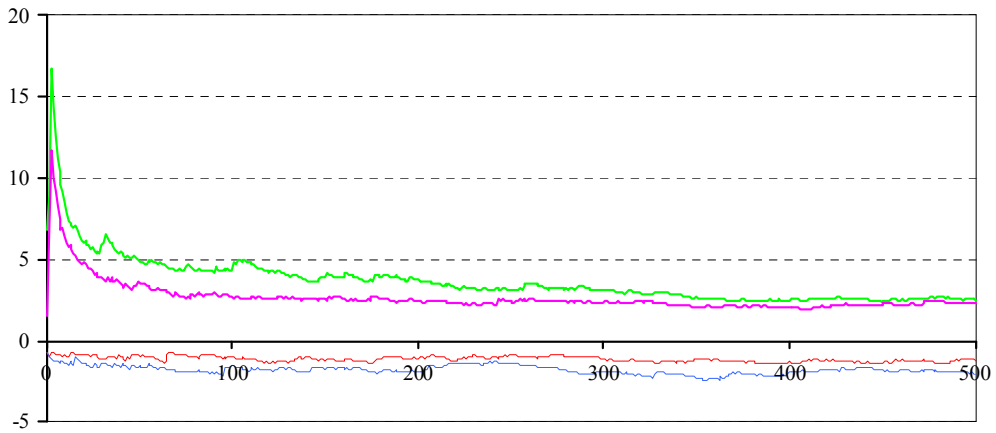
$$T_k = \frac{S_k - (k-1)}{\sqrt{(2k-1)}}, k = 1, 2, \ldots, K,$$

kur

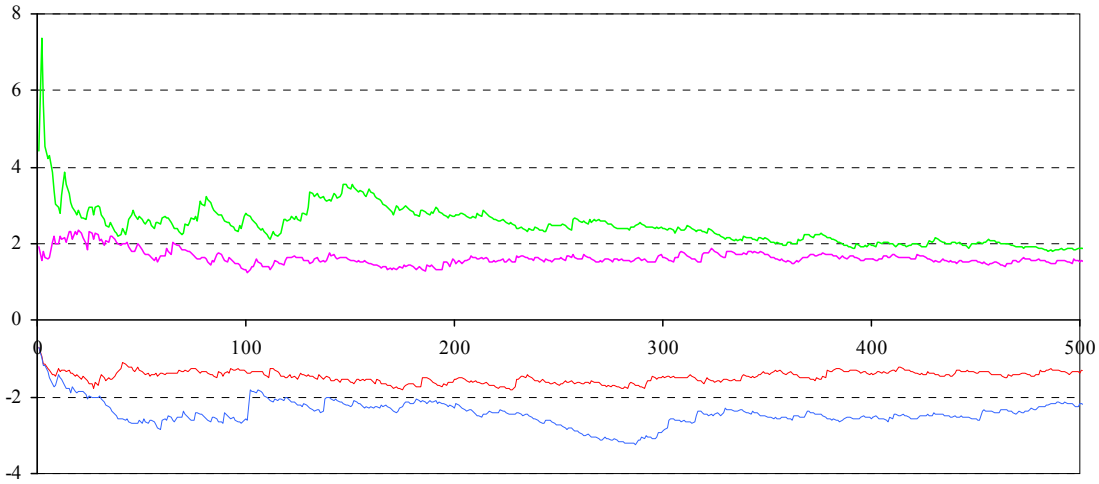
$$S_k = \frac{1}{2N} \sum_{j=1}^{k} (n^{j,k} - m^{j,k})^2, k = 1, 2, \ldots, K,$$

$n^{j,k}$ ir, atitinkamai, $m^{j,k}$, yra imties elementų skaičius (atitinkamai, imties X_H elementų skaičius) j-tajame padalijimo P_k elemente.

Buvo padaryta prielaida, kad diskriminantinė erdvė žinoma tiksliai (nėra paklaidų randant geriausią tiesinį poerdvį). Buvo generuojamos 100 nepriklausomų realizacijų. Buvo gautos testinės statististikos minimalios ir maksimalios reikšmės atitinkamoms apjungtoms realizacijoms. Taip pat buvo gautos testinės statististikos minimalios ir maksimalios reikšmės atmetus 5 proc. didžiausių ir 5 proc. mažiausių reikšmių. Matavimas buvo parenkamas iki 100, dažniausiai 10. Diskriminantinio poerdvio matavimas buvo parenkamas intervale 1–4 (šis matavimas priklauso nuo mišinio komponenčių skaičiaus ir jo parametrų), ir buvo naudojamas atitinkamas tiesinių poerdvių matavimų intervalas.



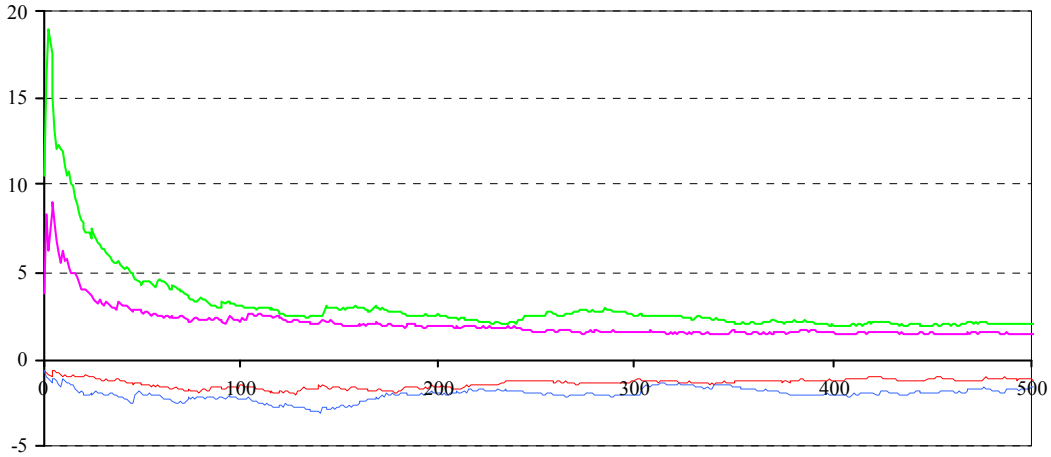
Pav. 3.1.1. Statistikos T_k minimumų ir maksimumų elgesys (modelis 1, projekcija į $k = 1$ matavimo poerdvį).



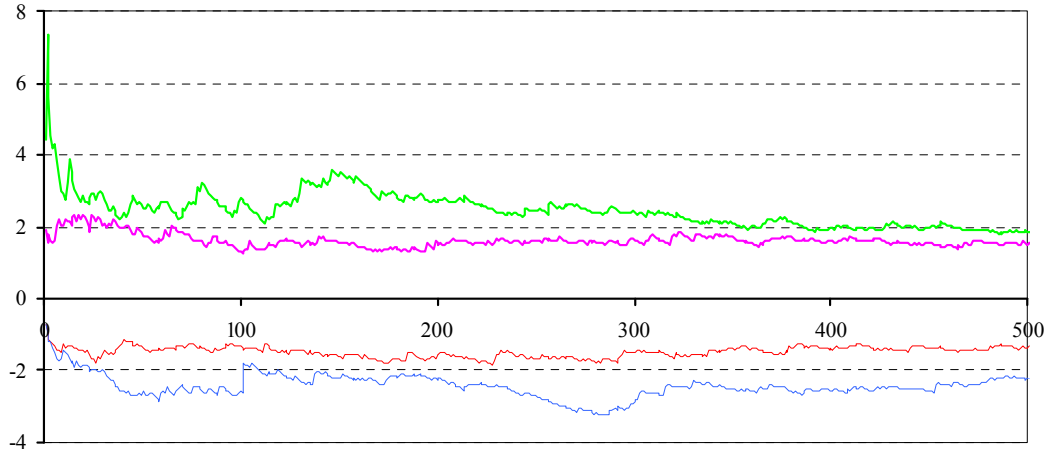
Pav. 3.1.2. Statistikos T_k minimumų ir maksimumų elgesys (modelis 1, projekcija į $k = 2$ matavimo poerdvį).

Modelis 1. Pav. 3.1.1 ir Pav. 3.1.2 pateikiamos minimalios ir maksimalios testinės statistikos reikšmės (taip pat reikšmės atmetus 5 proc. didžiausių ir 5 proc. mažiausių reikšmių) pavyzdžiui su 3 Gauso mišinio komponentėmis dešimtmatėje erdvėje su komponenčių vidurkiais (–4, –1, 0, … , 0), (0, 2, 0, … , 0), (4, –1, 0, … , 0) ir vienetinėmis kovariacinėmis matricomis (modelis 1). Žinoma, šiuo atveju diskriminantinio poerdvio matavimas lygus 2. Pav. 3.1.1 (atitinkamai, Pav. 3.1.2) buvo projektuojami dešimtmačiai duomenys į vienmatį (atitinkamai, dvimatį) poerdvį.

Modelis 2. Pav. 3.1.3 ir Pav. 3.1.4 pateikiamos minimalios ir maksimalios testinės statistikos reikšmės (taip pat reikšmės atmetus 5 proc. didžiausių ir 5 proc. mažiausių reikšmių) pavyzdžiui su 2 komponenčių Gauso mišiniu dešimtmatėje erdvėje su nuliniais komponenčių vidurkiais ir diagonalinėmis kovariacinėmis matricomis su diagonaliniais elementais (10, 1, 1, … , 1) ir (1, 10, 1, … , 1), atitinkamai (modelis 2). Šiuo atveju diskriminantinio poerdvio matavimas lygus 2. Pav. 3.1.3 (atitinkamai, Pav. 3.1.4) buvo projektuojami dešimtmačiai duomenys į vienmatį (atitinkamai, dvimatį) poerdvį.



Pav. 3.1.3. Statistikos T_k minimumų ir maksimumų elgesys (modelis 2, projekcija į $k = 1$ matavimo poerdvį).



Pav. 3.1.4. Statistikos T_k minimumų ir maksimumų elgesys (modelis 2, projekcija į $k = 2$ matavimo poerdvį).

Rezultatai parodė, kad yra labai silpna priklausomybė nuo parinkto mišinio modelio ir erdvės matavimo. Testinės statistikos, atmetus 5 proc. didžiausių ir 5 proc. mažiausių reikšmių, maksimumas yra tinkamas kriterijus priimti ar atmesti nagrinėjamą hipotezę.

3.2. DIDELIO MATAVIMO DUOMENŲ KOMPONENČIŲ NEPRIKLAUSOMUMO TESTAVIMAS

3.2.1. Duomenų komponenčių nepriklausomumo testavimas

Mūsų tikslas yra pateikti santykinai paprastą, priklausomą nuo duomenų ir skaičiavimų prasme efektyvią procedūrą komponenčių nepriklausomumo testavimui, kuomet erdvės X matavimas d yra didelis.

Tegul $\mathbf{X} = (X(1),...,X(N))$ yra n.v.p. atsitiktinio vektoriaus X su pasiskirstymo funkcija (p.f.) F erdvėje $\mathbf{R}^d$ stebėjimų imtis. Nagrinėsime tam tikrų F savybių testavimą. Tegul $\mathcal{F}_H$ ir $\mathcal{F}_A$ yra dvi nepersikertančios d-mačių skirstinių klasės.

Nagrinėkime neparametrinės hipotezės testavimo problemą:

$$H : F \in \mathcal{F}_H \quad \text{prieš} \quad A : F \in \mathcal{F}_A \tag{1}$$

Dviejų komponenčių $X_1 \in R^{d_1}$ ir $X_2 \in R^{d_2}$, $d_1 + d_2 = d$; $X = (X_1', X_2')$, nepriklausomumo testavimą atitinka

$$\mathcal{F}_H = \left\{ G : G(x) = G_1(x_1) \cdot G_2(x_2), x = (x_1, x_2), x_1 \in \mathbf{R}^{d_1}, x_2 \in \mathbf{R}^{d_2} \right\}, \tag{2}$$

kur G_1 ir G_2 yra marginaliniai G skirstiniai, atitinkantys komponentes X_1 ir X_2.

Procedūra remiasi Vapnik ir Chervonenkis idėja vertinant nuokrypį tarp empirinio ir tikrojo skirstinio panaudoti nuokrypį tarp dviejų nepriklausomų empirinių skirstinių (Vapnik, Chervonenkis (1981)) ir gerai žinoma modelio adekvatumo testavimo problema kaip klasifikacijos problemos interpretacija (žr., pvz., Hastie *et al.* (2001)), specialia pažingsnine duomenų padalijimo procedūra, randomizacija ir saviranka (*bootstrap*), nuoseklaus testavimo elementais. Vertinant procedūros efektyvumą naudojamas Monte-Carlo metodas.

Šiuo metu nėra visuotinai pripažintos metodologijos didelio matavimo daugiamačių neparametrinių hipotezių testavimui. Tradiciniai neparametrinių

hipotezių testavimo metodai remiasi empirine charakteringąja funkcija (Baringhaus and Henze (1988)), neparametriniais tankio įverčiais ir glodinimu (Bowman, Foster (1993), Huang (1997)) bei klasikinėmis vienmatėmis neparametrinėmis statistikomis duomenims projektuotiems į kryptis gautas tikslinio projektavimo (*projection pursuit*) metodu (Zhu *et al.*, (1997), Szekely, Rizzo (2005)).

Sudėtingesnė technika remiasi Vapnik-Chervonenkis teorija, tolydžia funkcionaline centrine ribine teorema ir didelių nuokrypių tikimybių nelygybėmis (Vapnik (1998), Bousquet *et al*. (2004)). Pastaruoju metu, ypač taikymuose, plačiai naudojami Bajeso metodas ir Markovo grandinių Monte-Carlo methodas (žr., pvz. Verdinelli, Wasserman (1998) ir ten esančias nuorodas). Daugiamatės kopulos (*copulas*) taip pat yra tinkamas būdas atspindėti statistinę priklausomybę tarp atsitiktinių vektorių.

Nepriklausomumo testavimo kriterijų asimptotinės savybės ir jų galingumas buvo detaliai studijuojami (žr., pvz., Genest, Remillard (2004)). Tačiau šie rezultatai tiesiogiai nepritaikomi mūsų atveju, kadangi komponentės X_1 ir X_2 yra didelio matavimo.

Didelio matavimo duomenų priklausomumo-nepriklausomumo struktūros nustatymui naudojama nepriklausomų komponenčių analizė (*independent component analysis* (ICA)), pastaruoju metu išvystytas pagrindinių komponenčių metodo ir projektavimo metodo papildymas, žr. Hyvarinen *et al.* (2001), kur pateikiamas efektyvus metodas sąlyginio nepriklausomumo testavimui. Susijusi tematika taip pat yra Szekely, Rizzo (2006), Polonik (1999), L.-X. Zhu, Neuhaus (2000) darbuose.

Skyrelyje 3.2.2 pateikiama neparametrinės hipotezės testavimo procedūra (žr. Radavičius, Jakimauskas, Sušinskas (2007)). Kitame skyrelyje pateikiami Monte-Carlo modeliavimo rezultatai ir baigiamosios pastabos (žr. Jakimauskas *et al.* (2008)).

3.2.2. Statistinis testas

Testinė statistika. Tegul $\mathcal{F} = \mathcal{F}_H \cup \mathcal{F}_A$. Tarkime, kad atvaizdavimas $\Psi : \mathcal{F} \to \mathcal{F}_H$ yra toks, kad $\mathcal{F}_H = \{G \in \mathcal{F} : \Psi(G) = G\}$. Duotam $F \in \mathcal{F}$, pažymėkime $F_H = \Psi(F)$. Nepriklausomumo hipotezei $F_H = F_1 \cdot F_2$.

Nagrinėkime dviejų populiacijų Ω_H ir Ω mišinių modelį

$$F_p = (1–p)F_H + pF, \ 0 < p < 1,$$

su p.f. F_H ir F, atitinkamai. Fiksuokime p ir tegul $Y = Y_{(p)}$ yra atsitiktinis vektorius (a.v.) su mišinio skirstiniu. Tegul $\pi(Y)$ yra populiacijos Ω sąlyginė tikimybė su sąlyga Y, t. y.

$$\pi(Y) = \mathbf{P}[\Omega \mid Y] = \frac{pf(Y)}{pf(Y) + (1-p)F_H(Y)}.$$

Čia f ir f_H yra pasiskirstymo tankiai (σ-baigtinio mato μ atžvilgiu) atitinkamai F ir F_H.

Įveskime nuostolių funkciją $L(F; F_0) = \mathbf{E}(\pi(Y) – p)^2$. Aišku, kad

$$L(F, F_H) = 0 \text{ tada ir tik tada kai } F=F_H,$$

kadangi sąlyginė tikimybė $\pi(Y)$ lygi tikimybei p tada ir tik tada kai $F = F_H$.

Tegul $\mathbf{X}^{(H)} = (X^{(H)}(1), \ldots, X^{(H)}(M))$ yra n.v.p. a.v. Ω_H imtis, nepriklausoma nuo $\mathbf{X}$. Apjungta imtis žymima

$$\mathbf{Y} = \mathbf{X} \parallel \mathbf{X}^{(H)} = (X(1), \ldots, X(N), X^{(H)}(1), \ldots, X^{(H)}(M))$$

o $Z(t) = \mathbf{1}\{t \le N\}, t = 1,\ldots N + M,$ yra atitinkama populiacijos Ω indikatorių seka. Tegul $\mathcal{P} = \{P_k, k = 0, 1, \ldots, K\}, P_0 = \mathbf{R}^d, P_{k-1} \subset P_k, k = 1, 2, \ldots, K,$ yra erdvės $\mathbf{R}^d$

padalijimų seka, priklausanti nuo **Y**, ir tegul $\{\mathcal{A}_k,\ k = 0, 1, \ldots, K\}$ yra atitinkama σ-algebrų seka, generuota šių padalijimų.

Skaičiavimų prasme efektyvus $\mathcal{P}$ pasirinkimas yra pažingsninis binarinis pakoordinatinis padalijimas, kiekviename žingsnyje minimizuojant vidutinę kvadratinę paklaidą imties **Y** elementams padalijimų aibėse. Kaip alternatyva galėtų būti padalijimas į aibes su maždaug lygiu imties **Y** elementų kiekiu.

Taip apibrėžus nuostolių funkciją $L(F;\ F_0)$, natūralus testo statistikos pasirinkimas yra χ^2-tipo statistika

$$T_k = \hat{\mathbf{E}}(Z_k - p)^2,\ \ p = \frac{N}{N+M}, \tag{3}$$

kur $\hat{\mathbf{E}}$ yra vidurkis pagal empirinį imties **Y** pasiskirstymą $\hat{F}$, o

$$Z_k = \hat{\mathbf{E}}(Z \mid \mathcal{A}_k),$$

kur $k \in \{1, 2, \ldots, K\}$. Skaičius k gali būti laikomas glodinimo parametru. Jis charakterizuoja, kokio smulkumo yra padalijimas. Taip pat nagrinėsime (3) versiją su svoriais

$$T_k = \hat{\mathbf{E}}(Z_k - p)^2 W_k, \tag{4}$$

kur W_k yra tam tikra $\mathcal{A}_k$-išmatuojama svorio funkcija. Pasirinkus $W_k = |S \cap \mathbf{Y}|/(p(1-p))$ ant padalijimo aibės $S \in P_k$, gauname L_2 atstumą tarp stebėtų ir tikėtinų dažnumų teisingai hipotezei H.

Kadangi optimali k reikšmė nežinoma, nagrinėsime tokį testo statistikos apibrėžimą:

$$T = \max_{k_0 \le k \le K} (T_k - a_k)/b_k, \tag{5}$$

kur $k_0 \geq 1$, a_k ir b_k yra centravimo ir mastelio parametrai, kurie yra atskirai apibrėžiami.

Pastaba. Kadangi kriterijaus kritinė sritis turi formą $C_\alpha = \{T > c_\alpha\}$, kur c_α yra kritinė reikšmė, atitinkanti reikšmingumo lygį α, natūralu išreikšti C_α kaip pažingsninę testavimo procedūrą:

Žingsnis 1: Priskiriame $k = k_0 - 1$.

Žingsnis 2: $k+1 \to k$; jei $k > K$, tada STOP, kitu atveju skaičiuoti T_k.

Žingsnis 3: Jei $T_k > a_k + c_\alpha b_k$, atmetame hipotezę H_0 ir STOP, kitu atveju einame prie Žingsnio 2.

Nulinės testinės statistikos pasiskirstymas. Tegul $\tau : I \to I$ yra atsitiktinis $I = \{1; 2, \ldots ; N + M\}$ perstatymas su lygiomis tikimybėmis, o $\mathbf{Y}^\tau$ yra atitinkamas **Y** perstatymas. Bet kuriai statistikai ξ pažymėkime ξ^τ, jei ši statistika skaičiuojama iš randomizuotos imties $\mathbf{Y}^\tau$. Atskiru atveju, $\mathbf{X}_\tau = (Y^\tau(1); Y^\tau(2), \ldots ; Y^\tau(N))$. Jei hipotezė H yra teisinga, $\mathbf{Y}^\tau = \mathbf{Y}$ pagal pasiskirstymą. Todėl galima nagrinėti sąlyginį randomizuotos testinės statistikos T_k^τ pasiskirstymą, kai duota imtis **Y**, tam, kad įvertintume pradinės testinės statistikos T_k savybes.

Fiksuokime imtį **Y**. Padalijimui $P_k = \{S_{k,1}, S_{k,2}, \ldots, S_{k,J_k}\}$ apibrėžkime

$$n(k) = (n_1(k), \ldots, n_{J_k}(k)) = \left(|\, S_{k,j} \cap \mathbf{Y} \,|, j = 1, \ldots J_k\right),$$

$$\nu(k) = (\nu_1(k), \ldots, \nu_{J_k}(k)) = \left(|\, S_{k,j} \cap \mathbf{X} \,|, j = 1, \ldots J_k\right), k = 1, \ldots, K.$$

kaip J_k matavimo vektorius, t. y. stebėtų **Y** ir **X** elementų dažnių vektorius. Tuomet $Z_k^\tau = \nu_j^\tau(k) / n_j(k)$ padalijimo aibėje $S_{k;j}$. Kadangi diskretus a.v. $\nu^\tau(k)$ turi daugiamatį hipergeometrinį skirstinį su parametrais $N + M$, $n(k)$, N, sąlyginis T^τ skirstinys turint duotą imtį **Y** ir padalijimą P, priklauso nuo **Y** tik per padalijimų aibių $n(k)$, $k = 1, \ldots, K$, dydžius. Tai duoda pagrindą apibrėžti a_k ir b_k formulėje (5), analizuoti asimptotinį statistikos T pasiskirstymą ir

nagrinėti eksponentines didelių statistikos T nuokrypių nelygybes. Žemiau pateiksime modeliavimo eksperimentą.

3.2.3. Nepriklausomumo testavimas

Tam, kad generuotume imtį iš $F_H = F_1 \cdot F_2$, naudosime savirankos (*bootstrap*) metodą ir generuosime nepriklausomas realizacijas skirstinio $\hat{F}_H = \Psi(\hat{F}) = \hat{F}_1 \cdot \hat{F}_2$ kur $\hat{F}_i$ yra F_i, $i = 1; 2$, empirinis skirstinys.

Tegul **X** yra nepriklausomi stebėjimai su daugiamačiu Student'o skirstiniu su m laisvės laipsnių. Nors X komponentės yra nekoreliuotos, jos yra priklausomos. Kadangi X konverguoja pagal pasiskirstymą į standartinį Gauso vektorių, kai $m \to \infty$, komponenčių priklausomumas beveik išnyksta dideliems m. Statistikos T_k centravimo ir mastelio parametrai skaičiuojami naudojant Gauso skirstinio aproksimacijas. Tarkime, kad $M = N$ ir $J_k = k + 1$. Tuomet standartizuotos χ^2-tipo statistikos $\hat{T}_k$ (3) ir L_2 atstumas $\hat{T}_k^{(2)}$ (4) su svorio funkcija $W_k = |S \cap \mathbf{Y}|$, atitinkamai, turi pavidalą

$$\hat{T}_k = \frac{T_k - k}{\sqrt{2k}}, T_k = \sum_{j=0}^{k} \frac{(n_j(k) - 2\nu_j(k))^2}{n_j(k)}, , \tag{6}$$

ir

$$\hat{T}_k^{(2)} = \frac{T_k^{(2)} - 2N}{2\sqrt{N}}, T_k^{(2)} = \sum_{j=0}^{k} (n_j(k) - 2\nu_j(k))^2, , \tag{7}$$

Nagrinėsime testinę statistiką T (5), paremtą (6). Palyginus su (7), ji priskiria didesnius svorius padalijimo aibėms su mažesniu imties elementų skaičiumi. Pagal modeliavimo rezultatus galima pasiūlyti $k_0 = 10$ ir $K = (M + N)/10$ kaip tinkamą pasirinkimą T_k maksimizavimo intervalui $[k_0; K]$. Kritinė reikšmė c_α šiai testo statistikai parenkama taip, kad generuotų imčių dalis kurioms teisinga nulinė hipotezė yra atmetama, neviršytų duoto reikšmingumo lygmens α, pvz., $\alpha = 0.05$. Konstantos c_α radimui naudojamas Monte-Carlo

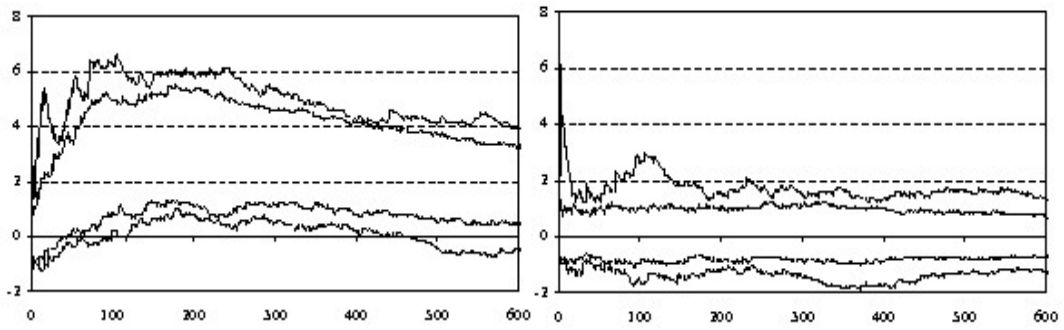
metodas. Rezultatai rodo, kad plačiame matavimų, imties dydžių, nulinio skirstinio intervale testinės statistikos elgesys yra labai panašus (žr. Pav. 3.3.1 ir Pav. 3.3.2, Pav 3.3.5). Taip pat procedūra buvo pritaikyta modelio adekvatumo testavimui, kuomet nagrinėjami daugiamačiai Gauso mišiniai. Pasirinkimas $c_{0.05} = 2.7$ yra daugeliu atvejų tinkamas.

Kompiuterinis modeliavimas buvo atliktas parinkus $d \leq 20$, $200 \leq N$, $M \leq 1000$, ir $m = 1; 2, \ldots ; 7; 25; 100$. Nepriklausomų komponenčių X_1 ir X_2 matavimai d_1 ir d_2, atitinkamai, buvo parinkti dviem būdais. Pirmu atveju $d_1 = d_2 = d/2$, o antru atveju $d_1 = 1$, $d_2 = d - 1$. Paprastai generavimų skaičius $R = 1000$. Žemiau pateikiami rezultatai kai $d = 2; 10$ ir $N = M = 1000$.

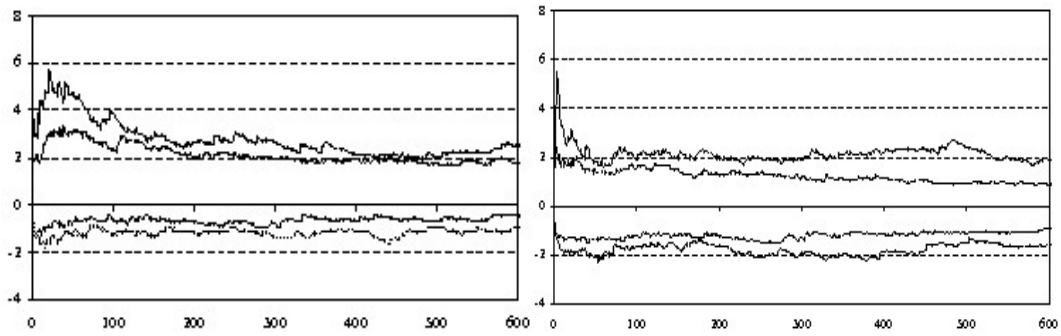
Trumpumo dėlei, testinė procedūra T (5) vadinama JRS testu. Procedūros galia buvo lyginama su klasikiniu Blum-Kiefer-Rosenblatt testu (trumpumo dėlei, BKR testu, žr. Blum *et al.* (1961)), kuris remiasi kriterijumi, paremtu Cramer-Von Mises tipo testine statistika nepriklausomumo testavimui:

$$\omega^2_{BKR} = N \int_{\mathbf{R}^{d_1}} \int_{\mathbf{R}^{d_2}} (\hat{F}(u,v) - \hat{F}_1(u)\hat{F}_2(v))^2 d\hat{F}(u,v). \tag{8}$$

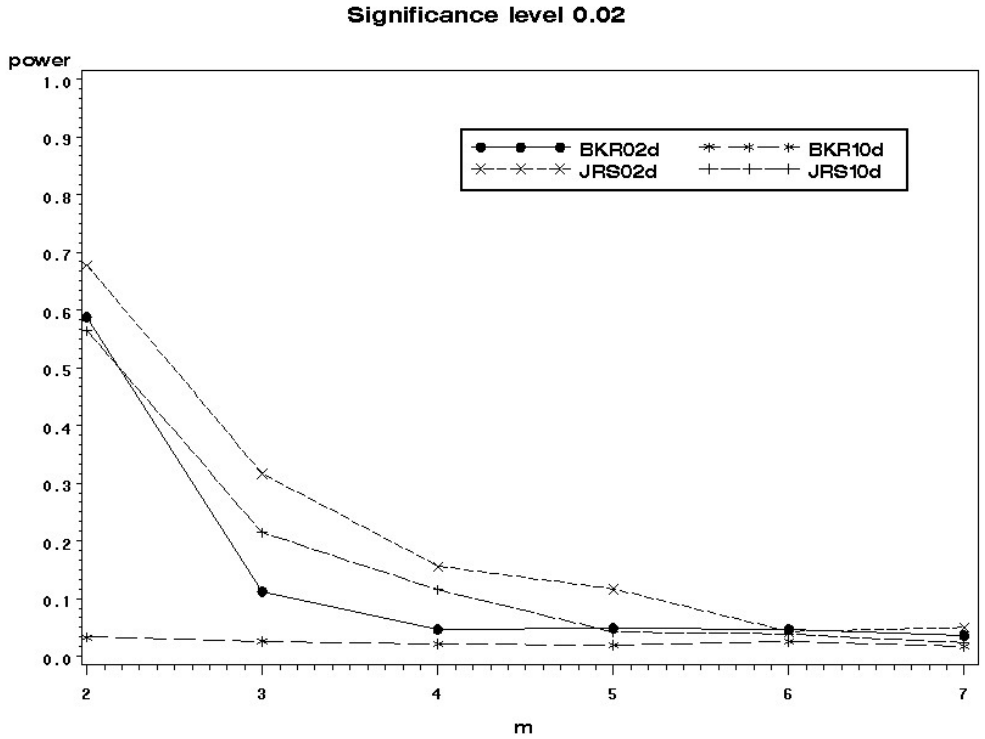
Čia $\hat{F}_i$ yra empirinė komponentės X_i, paremtos imtimi $\mathbf{X}$ ($i = 1; 2$), pasiskirstymo funkcija.



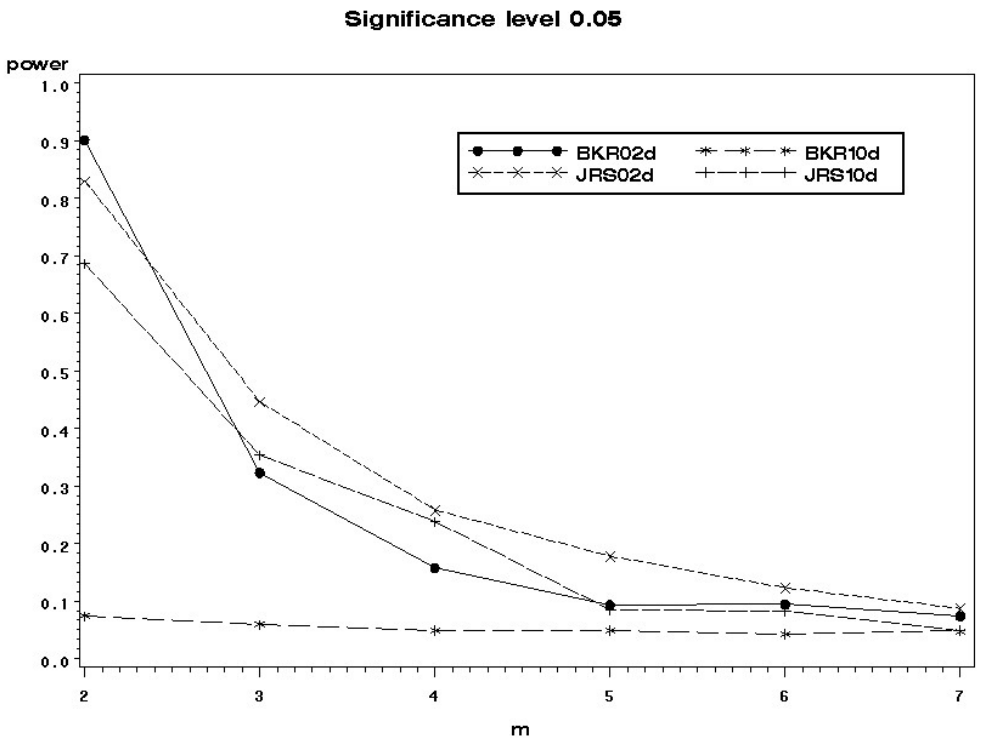
Pav. 3.2.1. Statistikos T_k maksimumas, minimumas ir dvipusiai 0.9 lygio pasikliovimo lygmenys imčiai iš Cauchy skirstinio ($m = 1$) ir atitinkamiems kontroliniams duomenims; $d = 20$, $d_1 = d_2 = 10$, $N = 1000$.



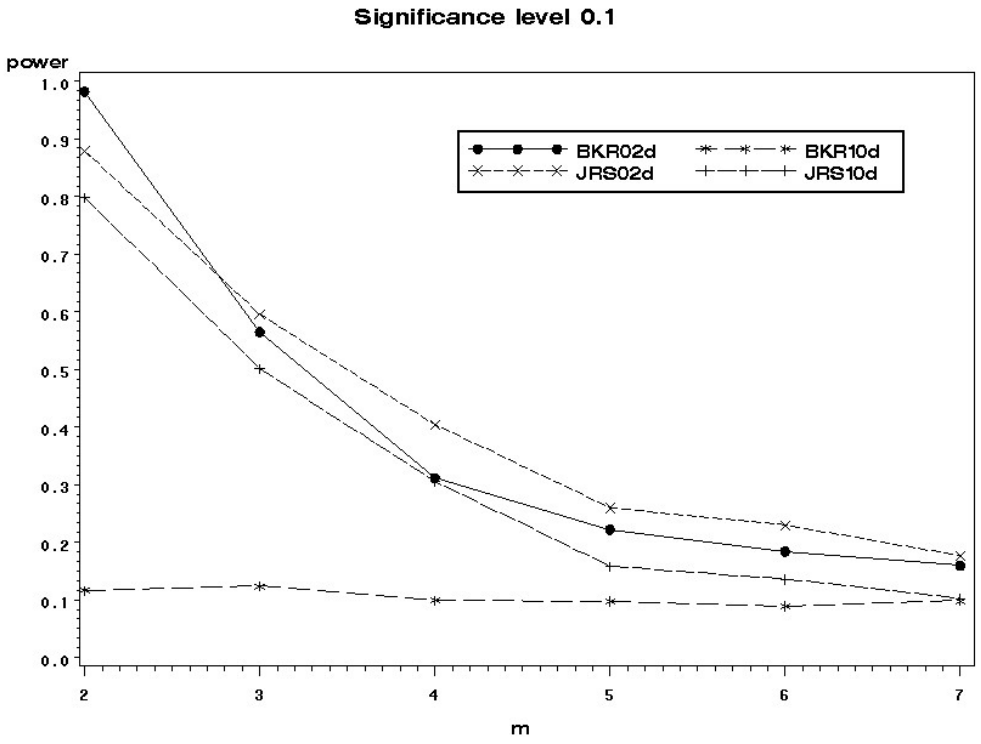
Pav. 3.2.2. Statistikos T_k maksimumas, minimumas ir dvipusiai 0.9 lygio pasikliovimo lygmenys imčiai iš Student'o skirstinio ($m = 3$) ir atitinkamiems kontroliniams duomenims; $d = 10$, $d_1 = 1$, $d_2 = 9$, $N = 1000$.



Pav 3.2.3. BKR testo galios funkcijos matavimams $d = 2$ ('BKR02d') ir $d = 10$ ('BKR10d') ir atitinkamos JRS testo galingumo funkcijos ('JRS02d' ir 'JRS10d'); reikšmingumo lygis $\alpha = 0.02$.



Pav 3.2.4. BKR ir JRS testų galios funkcijos, reikšmingumo lygis $\alpha = 0.05$.



Pav 3.2.5. BKR ir JRS testų galios funkcijos, reikšmingumo lygis $\alpha = 0.1$.

JRS testo galia lyginama su BKR testo galia. Tam, kad įvertintume nepriklausomumo testų galios funkcijas buvo atliktas Monte-Carlo modeliavimas su R = 1000 realizacijų. Rezultatai pateikti Pav. 3.2.3, 3.2.4 ir 3.2.5 reikšmingumo lygiams α = 0.02; 0.05; 0.1 ir matavimams d = 2 ir d = 10 su $d_1 = d_2 = d/2$. JRS testo galia nežymiai mažėja didėjant matavimui d, ir, kai d = 10, ji yra artima BKR testo galiai kai d = 2. BKR testo galia kai d = 10 yra labai maža.

Monte-Carlo modeliavimų rezultatai rodo, kad pasiūlyta procedūra yra gana perspektyvi. Ši procedūra pralenkia klasikinį Blum-Kiefer-Rosenblatt testą netgi mažo matavimo duomenims. Kritinė reikšmė c_α mažai priklauso nuo matavimo d ir padalijimo procedūros ir gali būti dar sumažinta pridėjus papildomus reikalavimus padalijimo procedūrai.

3.3. HIPOTEZIŲ APIE DIDELIO MATAVIMO ATSITIKTINIO VEKTORIAUS VIDURKĮ TIKRINIMAS TAIKANT EMPIRINĮ BAJESO METODĄ

3.3.1. Empirinio Bajeso metodo taikymas tikrinant hipotezes

Ankstesniuose skyreliuose naudojama χ^2-tipo statistika neatsižvelgia į sekos elementų, patekusių į atskiras padalijimo aibes, pasiskirstymą. Kai kuriais atvejais (pvz., kai skirstiniai didesnėje erdvės dalyje sutampa, o ryškiai skiriasi tik nedidelėje erdvės dalyje) didelis nereikšmingų nuokrypių skaičius gali užmaskuoti palyginti nedidelę dalį reikšmingų nuokrypių ir taip sumažinti hipotezės tikrinimo efektyvumą. Yra gana paprastas būdas – išmesti dalį (pvz., ketvirtadalį) mažiausių absoliutiniu dydžiu nuokrypių ir hipotezės tikrinimui naudoti tik didžiausius absoliutiniu dydžiu nuokrypius. Šis paprastas būdas, nepaisant privalumų, turi ir vieną trūkumą – statistikos pasiskirstymas esant teisingai hipotezei stipriai priklauso nuo tikrinamo skirstinio. Pateiksime (pasiremdami Jiang, Zhang (2009) darbu) metodą, kuris hipotezės tikrinimui naudoja empirinį Bajeso metodą.

Tegul $\mathbf{X} = (X(1), \ldots, X(N))$ yra dydžio N atsitiktinio vektoriaus (a.v.) X su skirstiniu P erdvėje $\mathbf{R}^d$ n.v.p. stebėjimų imtis. Nagrinėsime neparametrines P savybes tuo atveju, kai stebėjimų matavimas d yra didelis.

Straipsnyje Jakimauskas *et al.* (2008) buvo pateikta paprasta, paremta duomenimis ir skaičiavimų prasme efektyvi procedūra didelio matavimo duomenų neparametriniam testavimui. Ši procedūra remiasi randomizavimu ir saviranka (*bootstrap*), specialia pažingsnine duomenų skaidymo procedūra ir χ^2-tipo statistikomis. Šiame skyrelyje bus pateikta efektyvesnė už χ^2-tipo statistiką testinė statistika, paremta neparametriniu didžiausio tikėtinumo įverčiu (*nonparametric maximum likelihood* (NML)) ir empiriniu Bajeso (EB) įverčiu papildomame mišinių modelyje.

Tegul $\mathcal{P}_0$ ir $\mathcal{P}_1$ yra dvi nesusikertančios matavimo d skirstinių klasės, $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$. Nagrinėkime neparametrinės hipotezės testavimo problemą:

$$H_0 : P \in \mathcal{P}_0 \quad \text{prieš} \quad H_1 : P \in \mathcal{P}_1 \tag{1}$$

Tarkime, kad egzistuoja tolydus atvaizdavimas $\Psi : \mathcal{P} \to \mathcal{P}_0$ toks, kad $\mathcal{P}_0 = \{P \in \mathcal{P} : \Psi(P) = P\}$. Galima paimti, pvz., $\Psi(P) = \arg\min_{Q \in \mathcal{P}_0} \rho(Q, P)$, kur ρ yra atstumas erdvėje $\mathcal{P}$.

Tegul $\hat{\mathcal{P}}$ žymi empirinį skirstinį, paremtą imtimi **X** ir apibrėžkime $\hat{\mathcal{P}}_0 = \Psi(\hat{P})$. Jei patenkinta nulinė hipotezė, empiriniai skirstiniai $\hat{\mathcal{P}}$ ir $\hat{\mathcal{P}}_0$ dideliems N turėtų būti artimi, nes jie abu yra aproksimacijos to paties skirstinio P_0. Todėl bet kuris nukrypimo tarp $\hat{\mathcal{P}}$ ir $\hat{\mathcal{P}}_0$ matas gali būti paimtas kaip testinė statistika (1). Straipsnyje Jakimauskas *et al*. (2008) buvo nagrinėjamas žemiau pateikiamas nukrypimo matas T_0.

Generuokime dvi dydžio N nepriklausomas atsitiktines imtis $\mathbf{X}_P$ ir $\mathbf{X}_0$ su skirstiniais $\hat{\mathcal{P}}$ ir $\hat{\mathcal{P}}_0$, atitinkamai. Tegul $\mathbf{X}^*$ žymi apjungtą $\mathbf{X}_P$ ir $\mathbf{X}_0$, imtį,

$$\mathbf{X}^* = \mathbf{X}_P \,||\, \mathbf{X}_0 = \{\mathrm{X}_P(1), \ldots, \mathrm{X}_P(N), \mathrm{X}_0(1), \ldots, \mathrm{X}_0(N)\}.$$

Toliau, tegul $\mathcal{S} = \{S_k,\ k = 1, \ldots, K\}$, yra $\mathbf{X}^*$ padalijimų seka su $|\mathcal{S}_\mathrm{k}| = k$ elementų, gaunama iš tam tikro binarinio padalijimų algoritmo. Iš pradžių $\mathcal{S}_1 = \{\mathbf{X}^*\}$, o kai $k = 2, \ldots, K$, kitas padalijimas $\mathcal{S}_k$ gaunamas iš ankstesnio padalijimo $\mathcal{S}_{k-1}$ padalijus vieną iš padalijimo $\mathcal{S}_{k-1}$ aibių į du nesikertančius poaibius.

Fiksavus padalijimą $\mathcal{S}_k = \{S_1^k, \ldots, S_k^k\}$ ir $Q \in \{P, 0\}$, apibrėžkime

$$Y_Q = Y_Q(k) = (Y_Q(1), \ldots, Y_Q(k))^{\mathrm{T}} = (|\, S_j^k \cap \mathbf{X}_Q \,|,\ j = 1, \ldots, k)^{\mathrm{T}}. \tag{2}$$

Todėl Y_Q yra matavimo k vektorius su j-ąja komponente lygia $\mathbf{X}_Q$ elementų skaičiui aibėje S_j^k, $j = 1, \ldots, k$. Apibrėžkime

$$\eta_0 = (Y_P - Y_0)/\sqrt{Y_P + Y_0} \in \mathbf{R}^k, \tag{3}$$

čia veiksmai atliekami pakoordinačiui. Kuomet stebėjimų skaičius $Y_P(j) + Y_0(j)$ kiekvienoje aibėje S_j^k, $j = 1, \ldots, k$, yra didelis ir galioja nulinė hipotezė H_0, vektoriaus η_0 skirstinys gali būti aproksimuojamas $(k-1)$-mačiu standartiniu Gauso skirstiniu. Todėl natūralu naudoti χ^2 statistiką $|\eta_0|^2$ kaip nukrypimo matą tarp $\hat{P}$ ir $\hat{P}_0$ ir naudoti ją kaip testinę statistiką (1). Faktiškai, su statistika $|\eta_0|^2$, nulinė hipotezė

$$H_0^\eta : \mathbf{E}\eta_0 = 0_k \quad \text{prieš} \quad H_1^\eta : \mathbf{E}\eta_0 \neq 0_k$$

Yra tikrinama (čia 0_k žymi nulinį vektorių erdvėje $\mathbf{R}^k$). Statistikos T_0 aproksimacinė kovariacinė matrica tuo tarpu priklauso nuo alternatyvos H_P. Todėl naudojama nuokrypį stabilizuojanti transformacija, ir gaunamas naujas nuokrypio vektorius

$$\eta = \sqrt{Y_P + Y_0}\left(\arcsin\left(\sqrt{\frac{Y_P}{Y_P + Y_0}}\right) - \arcsin\left(\sqrt{\frac{Y_0}{Y_P + Y_0}}\right)\right). \tag{4}$$

Be to, χ^2 testas turi nedidelę galią, kuomet η matavimas n yra didelis, o kiekviena vidurkio $\theta = \mathbf{E}\eta$ komponentė tik nedaug skiriasi nuo 0_n, arba tik nedidelis skaičius θ komponenčių yra nenulinės. Kitame skyrelyje panaudosime neparametrinį maksimalaus tikėtinumo įvertį ir neparametrinį empirinį Bajeso metodą, kad sukonstruoti galingesnį kriterijų testuoti hipotezes H_0^η ir H_0.

3.3.2. Pagalbinė testavimo problema ir empirinis Bajeso metodas

Nagrinėkime pagalbinę testavimo problemą:

$$H_0^{\eta}: \mathbf{E}\eta_0 = 0_n \quad \text{prieš} \quad H_1^{\eta}: \mathbf{E}\eta_0 \neq 0_n . \tag{5}$$

kur $\eta \sim \mathcal{N}(\theta, I_n)$ ir $\theta \in \mathbf{R}^n$ yra nežinomas vidurkių vektorius. Taikant (empirinį) Bajeso metodą, nežinomas parametras θ laikomas atsitiktiniu. Taigi, nagrinėsime neparametrinį Gauso mišinių modelį su mišinio skirstiniu G (žr. Radavičius, Jakimauskas, Sušinskas (2007)):

$$\eta = \theta + z\,, \quad \theta \text{ ir } z \text{ yra nepriklausomi,} \tag{6}$$

$$z \sim \mathcal{N}(0_n, I_n), \tag{7}$$

$$\theta_i \sim G, \quad \{\theta_i, i = 1, 2, \ldots, n\} \text{ yra n.v.p. a.d.} \tag{8}$$

Bet kuriam $\nu > 0$, žymėsime $\mu_\nu(y \mid G)$ posteriorinį θ_1 ν-momentą su sąlyga $\eta_1 = y$:

$$\mu_\nu(y \mid G) = \frac{\varphi_\nu(y \mid G)}{\varphi_0(y \mid G)}, \tag{9}$$

$$\varphi_l(y \mid G) = \int_{\mathbf{R}} u^l \varphi(y-u) dG(u), \quad l \geq 0. \tag{10}$$

Čia φ žymi standartinį Gauso tankį. Homogeniškumo hipotezė (5) tvirtina, kad faktiškai nėra jokio mišinio, o G yra išsigimęs skirstinys taške 0. Kadangi $\mathbf{E}\,|\eta|^2 = n\mathbf{E}\theta_1^2 + n$, nulinės hipotezės H_0^{η} testavimo kriterijus gali būti paremtas funkcionalo įverčiu:

$$\mu_2 = \mu_2(G) = \int_{\mathbf{R}} u^2 dG(u) = \theta_1^2. \tag{11}$$

Alternatyvos tiesioginiam įverčiui $(\hat{\mu}_2)_{\chi^2} = n^{-1} |\eta|^2 - 1$ yra neparametrinis didžiausio tikėtinumo įvertinys (*nonparametric maximum likelihood estimator* (NMLE))

$$(\hat{\mu}_2)_{ML} = \mu_2(\hat{G}_{ML}), \tag{12}$$

ir neparametrinis empirinis Bajeso įvertinys (*nonparametric empirical Bayes* (NEB))

$$(\hat{\mu}_2)_{EB} = \frac{1}{n}\sum_{j=1}^{n} \mu_2(\eta_j \mid \hat{G}_{ML}). \tag{13}$$

Čia $\hat{G} = \hat{G}_{ML}$ yra NMLE mišinio skirstinio *G*. Gauso mišiniams jis egzistuoja ir yra stipriai pagrįstas (*strongly consistent*), žr., pvz., van de Geer (2003)). Taip pat nagrinėsime NEB statistiką

$$(\hat{\mu}_1^2)_{EB} = \frac{1}{n}\sum_{j=1}^{n} \mu_1^2(\eta_j \mid \hat{G}_{ML}), \tag{14}$$

kuri yra paslinktas link 0 funkcionalo μ_2 įvertis.

Jiang, Zhang (2009) parodė, kad parametro θ NEB įvertis

$$\hat{\theta} = (\mu_1(\eta_j \mid \hat{G}_{ML}), \;\; j = 1, 2, \ldots, n)$$

asimptotiškai pasiekia vidutinės kvadratinės paklaidos R_n^* minimumą separabilių statistikų klasėje, jei

$$(\log n)^{9/2} \min(\sqrt{\log n}, \| \theta \|_\infty) = o(nR_n^*).$$

Taip pat jie parodė modeliavimo būdu, kad tam tikrais atvejais $\hat{\theta}$ gerokai lenkia kitus žinomus įverčius, įskaitant ir James–Stein įvertį. Kadangi $\hat{\theta}$ yra invariantiškas poslinkio atžvilgiu, tai rodo, kad kriterijus (5) testavimui paremtas statistika $(\hat{\mu}_1^2)_{EB}$ gali būti galingesnis, ypač artimoms alternatyvoms.

Asimptotinės $(\hat{\mu}_1^2)_{EB}$ savybės gali būti išvestos iš $|\hat{\theta} - \theta|^2$ savybių. Kitame skyrelyje nagrinėsime statistikos $(\hat{\mu}_1^2)_{EB}$ savybes baigtinėms imtims ir pateiksime kai kuriuos modeliavimo rezultatus tam tikroms natūralioms alternatyvoms.

3.3.4. Modeliavimo eksperimentas

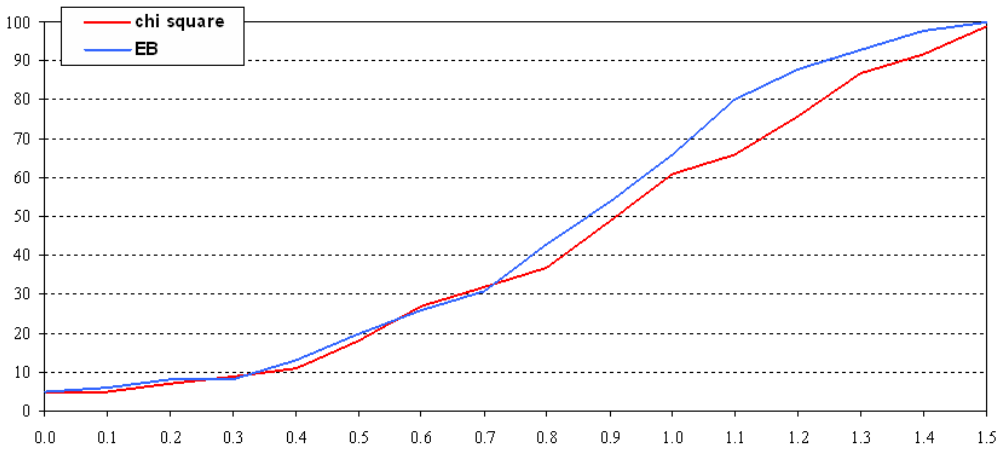
Nagrinėsime tris skirstinio θ_i alternatyvas:

(a1) $\theta_i = a u_i,\ u_i \sim \mathcal{N}(0, 1)$,

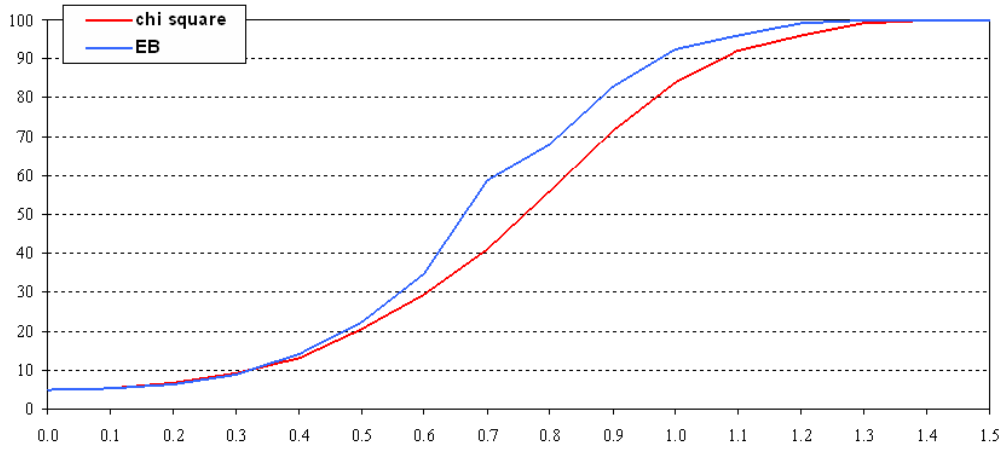
(a2) $\theta_i = a(2z_i - 1),\ z_i \sim \mathcal{B}(1/2, 1)$,

(a3) $\theta_i = a(-1)^i \cdot \mathbf{1}\{i \le m\},\ 1 < m < n$.

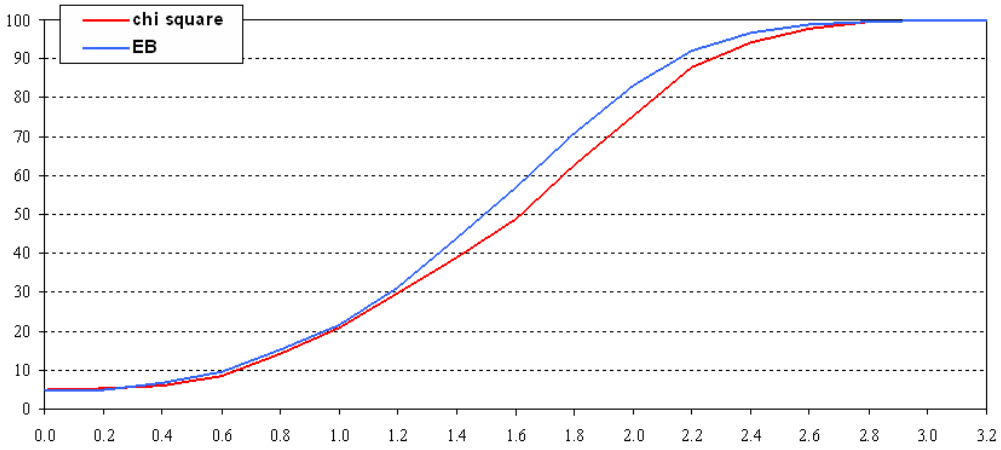
Įvairioms parametrų a, n ir m kombinacijoms buvo atliktas modeliavimas su 1000 nepriklausomų realizacijų (žr. Jakimauskas, Sušinskas (2010)). Parametras $a > 0$ nusako testavimo problemos sudėtingumą. Modeliavimo rezultatai rodo tam tikrą NEB testo galios padidėjimą, palyginti su χ^2 testo galia. Pav. 3.3.1–3.3.3 pateikiami tipiniai rezultatai. Čia pateikiami galios grafikai testinei statistikai $(\hat{\mu}_1^2)_{EB}$ ir χ^2 testinei statistikai priklausomai nuo parametro a, pateiktas atvejis, kai $n = 50$ ir $m = 8$.



Pav. 3.3.1. Testų galia alternatyvai (a1).



Pav. 3.3.2. Testų galia alternatyvai (a2).



Pav. 3.3.3. Testų galia alternatyvai (a3).

NML įvertis $\hat{G}_{ML}$ skaičiuojamas naudojant EM algoritmą Gauso mišiniui su iš anksto nurodytais ir fiksuotais mišinio komponenčių centrais (žr., pvz., Jiang, Zhang (2009)). Komponenčių skaičius $m = 15$. Tai reiškia, kad faktiškai NMLE su apribojimais pakeičia $\hat{G}_{ML}$.

3.4. IŠVADOS

Pateiktas duomenų skaidymo procedūros pritaikymas duomenų modelio verifikavimui leidžia panaudoti metodą patikrinti modelio adekvatumą, kuris mažai priklauso nuo duomenų dimensijos ir kurio efektyvumas remiasi subalansuotų daugiamatės erdvės srities suskaidymo rinkinių seka. Svarbu tai, kad šių subalansuotų rinkinių seka (priklausanti nuo turimų duomenų ir nuo tikrinamo duomenų modelio) gaunama paprastu būdu, be papildomai nurodomų parametrų.

Nagrinėjant modelio adekvatumo testavimo algoritmą didelio matavimo duomenims, tenkinantiems daugiamatį Gauso mišinių modelį, rezultatai parodė, kad yra labai silpna priklausomybė nuo parinkto mišinio modelio ir erdvės matavimo. Testinės statistikos, atmetus 5 proc. didžiausių ir 5 proc. mažiausių reikšmių, maksimumas yra tinkamas kriterijus priimti ar atmesti nagrinėjamą hipotezę.

Nagrinėjant didelio matavimo duomenų komponenčių nepriklausomumo testavimo algoritmą, Monte-Carlo modeliavimų rezultatai rodo, kad pasiūlyta procedūra yra pakankamai efektyvi. Ši procedūra pralenkia klasikinį Blum-Kiefer-Rosenblatt testą didesnio matavimo duomenims. Kritinė reikšmė c_α mažai priklauso nuo matavimo d ir padalijimo procedūros ir gali būti dar sumažinta pridėjus papildomus reikalavimus padalijimo procedūrai.

Taikant empirinį Bajeso metodą testuojant didelio matavimo duomenų komponenčių nepriklausomumą, pradinė didelio matavimo duomenų testavimo problema suvedama į papildomą testavimo problemą. Nulinė hipotezė H_0^η gali būti performuluota kaip $G = \delta_0$, kur G yra apriorinis nežinomų parametrų θ_i, $i = 1, \ldots, n$, skirstinys, o δ_0 yra išsigimęs skirstinys taške 0. Todėl bet kuris

nukrypimo matas tarp δ_0 ir skirstinio G NMLE įverčio $\hat{G}_{ML}$ gali būti naudojamas testavimui, pvz., χ^2 testas arba neparametrinis tikėtinumo santykio kriterijus. Modeliavimo būdu buvo nagrinėjamos baigtinių imčių savybės testui, paremtam NEB statistika $\hat{\mu}_1^2$. Modeliavimo rezultatai rodo NEB testo privalumus, palyginus su χ^2 testu. Kadangi NMLE įverčio $\hat{G}_{ML}$ skaičiavimas yra iteracinis ir užima daug laiko, rezultatai gali priklausyti nuo skaičiavimo metodo iteracijų skaičiaus.

4. RETŲ DAŽNIŲ ANALIZĖ NAUDOJANT EMPIRINĮ BAJESO METODĄ

Nagrinėkime retų įvykių didelėse populiacijose (pvz., tam tikros ligos tikimybių, mirčių, savižudybių ir t. t., tikimybių) vertinimo problemą. Atitinkamų įvykių skaičius priklauso nuo populiacijos dydžio ir nuo atskiro įvykio tikimybės. Klasikinis dažnių įvertis dažnai yra netinkamas, nes turi per dideles paklaidas. Padarysime prielaidą, kad įvykių skaičius populiacijoje turi Puasono skirstinį su tam tikrais parametrais. Pastebėsime, kad tokia aproksimacija yra pakankamai tiksli didelėms populiacijoms ir mažoms, bet ne pernelyg mažoms, tikimybėms.

Empiriniame Bajeso vertinimo metode daroma prielaida, kad įvykių tikimybės populiacijose yra atsitiktinės ir turi tam tikrą skirstinį. Gerai žinoma (žr., pvz., Clayton, Caldor (1987), Meza, (2003), Sakalauskas, Vaičiulytė 2012), kad nežinomų tikimybių Bajeso įverčiai turi gerokai mažesnę vidutinę kvadratinę paklaidą palyginti su paprastais santykinės rizikos įverčiais.

Bajeso statistinių sprendimų teorija yra taikoma tais atvejais, kuomet informacija apie parametrus gali būti nusakoma tam tikru tikimybiniu skirstiniu. Šią teoriją naudojantys metodai turi visą eilę privalumų (žr., pvz., DeGroot (1970), Carlin, Louis (1996)), lyginant su klasikiniais (*„frequentist“*) metodais. Vis dėlto, šie metodai pradėti plačiau taikyti tik prieš kelis dešimtmečius. Viena iš dažnai nurodomų priežasčių yra tai, kad jie yra gali nebūti objektyvūs – visi statistiniai skaičiavimai turėtų būti atliekami kruopščiai patikrinus ir įvertinus iš anksto numanomas nagrinėjamojo objekto savybes, o tai gali padaryti rezultatus tokius, kokius juos norėtų matyti konkrečią problemą nagrinėjantis statistikas. Bet viena iš svarbiausių priežasčių, vis dėlto, yra staigus kompiuterinės technikos vystymasis maždaug nuo 1980 metų. Bajeso metodams net ir atrodytų paprastais atvejais reikia nemažų skaičiavimų, pvz. sudėtingų integralų skaičiavimų ir pan.

Naudojant klasikinius metodus (žr., Carlin, Louis (1996)), tam tikros statistikos *d* rizika vertinant nežinomą parametrą θ yra funkcija nuo nežinomo parametro (čia *l* yra nuostolių funkcija, o *f* – imties skirstinys)

$$R(\theta,d) = \mathbf{E}_{X|\theta}[l(\theta,d(X))] = \int l(\theta,d(X)) f(X \mid \theta) dX .$$

Naudojant Bajeso metodą, tam tikros statistikos *d* rizika vertinant nežinomą parametrą θ yra vienas skaičius (čia π yra apriorinis skirstinys)

$$\rho(\pi,d(X)) = \mathbf{E}_{\theta|X}[l(\theta,d(X))] = \int l(\theta,d(X)) p(\theta \mid X) d\theta ,$$

kur

$$p(\theta \mid X) = \frac{f(X \mid \theta)\pi(\theta)}{m(X)},$$

$$m(X) = \int f(X \mid u)\pi(u) du .$$

Bajeso rizika, laikant kad statistika *d* yra iš tam tikros klasės *D*, apibrėžiama kaip rizikų apatinis rėžis

$$\rho^*(\pi) = \inf_{d \in D} \rho(\pi,d(X)) .$$

Kiekviena statistika d^*, kurios rizika lygi Bajeso rizikai, vadinama Bajeso įverčiu skirstiniui π.

Jei nuostolių funkcija yra kvadratinė, gausime vidutinį kvadratinį nuokrypį. Klasikiniu atveju vidutinis kvadratinis nuokrypis duotai parametro reikšmei θ yra

$$\mathbf{E}_{X|\theta}[l(\theta, d(X))] = \mathbf{E}_{X|\theta}[(\theta, d(X))^2] \overset{\text{def.}}{=} MSE_d(\theta).$$

Bajeso metodo atveju rizika duotam aprioriniam skirstiniui π yra

$$\mathbf{E}_{\theta|X}[l(\theta, d(X))] = \mathbf{E}_{\theta|X}[(\theta, d(X))^2] \overset{\text{def.}}{=} MSE_{d,\pi}(X).$$

Taip pat naudojama preposteriorinė rizika

$$\mathbf{E}_{\theta,X}[l(\theta, d(X))] = \mathbf{E}_{\theta,X}[(\theta, d(X))^2] \overset{\text{def.}}{=} MSE_{d,\pi},$$

kuriai galioja lygybės

$$MSE_{d,\pi} = \mathbf{E}_{\theta}[MSE_d(\theta)],$$

$$MSE_{d,\pi} = \mathbf{E}_{X}[MSE_{d,\pi}(X)].$$

Vienas iš Bajeso metodo privalumų yra tai, kad rizika yra vienas skaičius (priklausantis nuo imties), o tuomet, kaip taisyklė egzistuoja optimali statistika. Klasikiniu atveju nagrinėjamų statistikų klasei tenka įvesti tam tikrus apribojimus, kad nagrinėjamoje statistikų klasėje egzistuotų optimaliu procedūra.

Bajeso metodas taip pat turi ir trūkumą, kadangi tenka daryti prielaidą apie konkretaus apriorinio skirstinio egzistavimą ir, be to, laikyti jį yra žinomu. Šią problemą iš dalies galima išspręsti naudojant empirinį Bajeso (EB) metodą. Empirinio Bajeso metodo atveju stebėti duomenys naudojami įvertinti apriorinio skirstinio parametrus (parametrinis EB metodas), arba įvertinti apriorinio skirstinio formą (neparametrinis EB metodas). Yra bendra nuomonė, kad jei nagrinėjamame uždavinyje (daugelyje praktinių uždavinių)

parametrams galima priskirti tam tikrą tikimybinį skirstinį, tokie metodai yra visiškai priimtini ir naudingi.

4.1. RETŲ ĮVYKIŲ MODELIAVIMAS NAUDOJANT EMPIRINĮ BAJESO METODĄ

Pagrindinis atvaizdavimo žemėlapyje tikslas yra aprašyti geografinius skirtumus tarp ligos ar mirties tikimybių ir parodyti, kad tam tikri įvykiai gali būti iš dalies nulemti rizikos faktorių, kurie turi erdvinę struktūrą. Jei neatsižvelgsime į skirtingą populiacijų dydį, netiksliai gautas santykinės rizikos (*relative risk* (RR)) įvertis, paremtas tik keliais atvejais, gali lemti didelius nuokrypius žemėlapyje ir dominuoti jo vaizdą. Todėl erdvinės informacijos tinkamai analizei plačiai naudojamas Bajeso metodas, kadangi jis gali įvertinti atskirų įvykių tikimybes ne tik iš nagrinėjamos populiacijos duomenų, bet ir iš kitų populiacijų duomenų (žr. Tsutakava *et al.* (1985), Knorr-Held, Rasser (1999), Bradley *et al.* (2000), Leite *et al.* (2000), Quigley *et al.* (2007), ir t. t.). Kaip yra parodyta, empiriniai Bajeso įverčiai turi gerokai mažesnį vidutinį kvadratinį nuokrypį palyginti su RR įverčiais (žr. Clayton, Kaldor (1987), Meza J.L. (2003), ir t. t.). Paprastai stebimi dydžiai rizikos atvaizdavimuose yra dydžiai su Puasono skirstiniu, priklausančiu nuo įvykio tikimybės ir stebėjimų intervalo kiekvienoje populiacijoje. Empiriniam Bajeso mažų populiacijos tikimybių vertinimui reikalingas apriorinis populiacijos rizikos skirstinys. Buvo naudojami keletas apriorinių skirstinių, tarp jų gama, lognormalusis skirstinys, taip pat neparametriniai skirstiniai. Kol kas apriorinio skirstinio pasirinkimas buvo daugiausiai paremtas loginiais samprotavimais (žr. Yasui *et al.* (2000), Vaurioa, Jankala (2006) , ir t. t.).

Nagrinėsime empirinio Bajeso vertinimo techniką Puasono-Gauso modeliui, kuomet apriorinis logitų skirstinys yra Gauso su parametrais, vertinamais didžiausio tikėtinumo (*maximal likelihood* (ML)) metodu (žr. Tsutakava *et al.* (1985), Sakalauskas (1995)) Pateiksime apriorinių parametrų vertinimo nesinguliarumo sąlygas ir nagrinėsime paprastą iteracinį algoritmą apriorinio skirstinio vertinimui (žr. Gurevičius, Jakimauskas, Sakalauskas

(2009)). Kadangi empirinis Bajeso metodas naudojant Puasono-Gauso modelį atskiria skirtingas įvykių populiacijose tikimybes, nors įvykių skaičius skiriasi nedaug, aprašysime klasterizavimo algoritmą, panaudojantį šią savybę. Naudosime Lietuvos 2003–2004 metų mirtingumo duomenis (žr. Sakalauskas, Jakimauskas, Sušinskas (2010)), kad įvertintume nežinomas mirtingumo tikimybes ir parodytume galimybę pritaikyti pateikiamą metodą.

4.1.1. Puasono-Gauso modelis

Nagrinėkime aibę $\Lambda = (A_1, A_2, ..., A_K)$ iš K populiacijų, kur kiekviena populiacija A_j susideda iš N_j individų. Tarkime, kad tam tikri įvykiai (pvz., mirtis ar liga ar kiti įvykiai) gali atsitikti stebimoje populiacijoje. Mūsų tikslas yra įvertinti nežinomas įvykių P_j tikimybes , kuomet yra stebimas įvykių skaičius Y_j , $j = \overline{1,K}$, populiacijose. Kadangi paprastai santykinės rizikos įvertis $\frac{Y_j}{N_j}$ daugeliu atvejų negali būti naudojamas dėl didelių skirtumų tarp populiacijų dydžių N_j, taikysime empirinį Bajeso metodą.

Paprastai daroma prielaida, (žr. Bradley *et al.* (2000), Tsutakava (1985), Clayton, Kaldor (1987), ir t. t.), kad įvykių skaičius Y_j turi Puasono skirstinį su parametrais $\lambda_j = N_j \cdot P_j$:

$$f(Y_j, \lambda_j) = e^{-\lambda_j} \frac{(\lambda_j)^{Y_j}}{(Y_j)!}, \quad j = 1,..., K. \tag{1}$$

Nagrinėsime modelį, kuriame logitai

$$\alpha_j = \ln \frac{P_j}{1 - P_j} \tag{2}$$

yra Gauso atsitiktiniai dydžiai su parametrais μ, σ. Todėl logitų skirstinio tankis yra

$$g(\alpha_j,\mu,\sigma)=\frac{\exp\left(-\frac{(\alpha_j-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}. \tag{3}$$

Tuomet tikimybės P_j vertinamos kaip sąlyginės tikimybės su sąlyga, kad μ,σ įgyja duotas reikšmes:

$$P_j=\frac{\int_{-\infty}^{\infty}\frac{1}{1+e^{-\alpha}}f\left(Y_j,\frac{N_j}{1+e^{-\alpha}}\right)g(\alpha,\mu,\sigma)d\alpha}{D_j(\mu,\sigma)}, \tag{4}$$

kur

$$D_j(\mu,\sigma)=\int_{-\infty}^{\infty}f\left(Y_j,\frac{N_j}{1+e^{-\alpha}}\right)g(\alpha,\mu,\sigma)d\alpha. \tag{5}$$

yra *j*-tosios populiacijos, $j=\overline{1,K}$, įvykių skaičius.

Naudojant empirinį Bajeso metodą nežinomi parametrai μ,σ vertinami didžiausio tikėtinumo metodu (žr. Tsutakava *et al.* (1985), Bradley *et al.* (2000)). Po tam tikrų pertvarkymų gauname šio pavidalo logaritminę didžiausio tikėtinumo (*maximum likelihood* (ML)) funkciją:

$$L(\mu,\sigma)=-\sum_{j=1}^{K}\ln\left(\int_{-\infty}^{\infty}f\left(Y_j,\frac{N_j}{1+e^{-\alpha}}\right)g(\alpha,\mu,\sigma)d\alpha\right)=-\sum_{j=1}^{K}\ln\left(D_j(\mu,\sigma)\right), \tag{6}$$

Kuri yra minimizuojama, kad gauti parametrų μ, σ įverčius

$$L(\mu,\sigma)\to\min_{\mu,\sigma}.$$

4.1.2. Didžiausio tikėtinumo funkcijos išvestinės ir fiksuoto taško lygtis

ML funkcija (6) yra diferencijuojama parametrų atžvilgiu ir jos dalinės išvestinės turi pavidalą:

$$\frac{\partial L(\mu,\sigma)}{\partial \mu} = \sum_{j=1}^{K} \frac{\int_{-\infty}^{\infty} \frac{\alpha-\mu}{\sigma^2} \cdot f\left(Y_j, \frac{N_j}{1+e^{-\alpha}}\right) g(\alpha,\mu,\sigma) d\alpha}{D_j(\mu,\sigma)}, \tag{7}$$

$$\frac{\partial L(\mu,\sigma)}{\partial \sigma} = -\frac{1}{\sigma} \cdot \sum_{j=1}^{K} \frac{\int_{-\infty}^{\infty} \left(1-\frac{(\alpha-\mu)^2}{\sigma^2}\right) \cdot f\left(Y_j, \frac{N_j}{1+e^{-\alpha}}\right) g(\alpha,\mu,\sigma) d\alpha}{D_j(\mu,\sigma)}. \tag{8}$$

Prilyginus nuliui pirmąsias išvestines

$$\frac{\partial L(\mu,\sigma)}{\partial \mu} = 0, \tag{9}$$

$$\frac{\partial L(\mu,\sigma)}{\partial \sigma} = 0 \tag{10}$$

iš (7), (8) gauname "fiksuoto taško" lygtis μ ir σ ML įverčių skaičiavimui:

$$\mu = \frac{1}{K} \sum_{j=1}^{K} \frac{\int_{-\infty}^{\infty} \alpha f\left(Y_j, \frac{N_j}{1+e^{-\alpha}}\right) g(\alpha,\mu,\sigma) d\alpha}{D_j(\mu,\sigma)}, \tag{11}$$

$$\sigma^2 = \frac{1}{K} \sum_{j=1}^{K} \frac{\int_{-\infty}^{\infty} (\alpha-\mu)^2 f\left(Y_j, \frac{N_j}{1+e^{-\alpha}}\right) g(\alpha,\mu,\sigma) d\alpha}{D_j(\mu,\sigma)}. \tag{12}$$

Pastebėsime, kad šių lygčių sprendinys egzistuoja, tik kai patenkintos tam tikros sąlygos. Parametro σ ML įverčio nesinguliarumo sąlyga (t. y., $\sigma^2 > 0$) po tam tikros (6), (7), (8) analizės gali būti nusakoma šia teorema.

Teorema (Sakalauskas, 1995). *Lygčių sistemos* (11)*,* (12) *sprendinys egzistuoja, jei*

$$\sum_{j=1}^{K}(Y_j - N_j \cdot P)^2 \geq \sum_{j=1}^{K} Y_j \,. \tag{13}$$

Priešingu atveju ML įvertis turi tokį pavidalą:

$$\mu = \ln\frac{P}{1-P}\,, \tag{14}$$

$$\sigma = 0\,, \tag{15}$$

kur

$$P = \sum_{j=1}^{K} Y_j \Big/ \sum_{j=1}^{K} N_j \,. \tag{16}$$

Kaip seka iš sąlygos (13), singuliarumas daugiausiai atsiranda mažose populiacijose. Taigi ši sąlyga gali būti naudojama parinkti populiacijų su mažomis tikimybėmis aibes.

Nesunku pastebėti, kad singuliarumo atveju (t.y. $\sigma = 0$) įvykių populiacijose tikimybės nekinta ir yra tos pačios visoje populiacijoje

$$P_j \equiv P\,. \tag{17}$$

Atitinkama ML funkcijos reikšmė turi pavidalą:

$$L(\mu^*,0)=\sum_{j=1}^{K}\left(N_j\cdot P-Y_j\cdot\ln(N_j\cdot P)\right)=$$

$$=\sum_{j=1}^{K}Y_j\cdot(1-\ln(N_j\cdot P))\,. \tag{18}$$

4.1.3. Gauso skirstinio apriorinių parametrų vertinimas „paprastų iteracijų“ metodu

ML funkcijos savybių tyrimas parinkus įvairius populiacijų dydžius ir įvairų įvykių skaičių leidžia padaryti išvadą, kad pakankamai didelėje minimumo taško aplinkoje ši funkcija yra unimodalinė su vienu minimumo tašku. Taigi, tarkime, kad patenkinta nesinguliarumo sąlyga (13). Tuomet lygčių sistemos (9), (10) ar (11), (12) sprendinys gali būti randamas skaitiniais metodais. Pvz., „paprastų iteracijų“ metodas (žr. Kantorovitch, Akilov (1982)) gali būti pritaikytas išspręsti šias lygtis, kad gauti parametrų μ ir σ ML įverčius:

$$\mu_{t+1}=\frac{1}{K}\sum_{j=1}^{K}\frac{\int_{-\infty}^{\infty}\alpha\, f\left(Y_j,\frac{N_j}{1+e^{-\alpha}}\right)g(\alpha,\mu_t,\sigma_t)d\alpha}{D_j(\mu,\sigma)}, \tag{19}$$

$$\sigma^2_{t+1}=\frac{1}{K}\sum_{j=1}^{K}\frac{\int_{-\infty}^{\infty}(\alpha-\mu_t)^2\, f\left(Y_j,\frac{N_j}{1+e^{-\alpha}}\right)g(\alpha,\mu_t,\sigma_t)d\alpha}{D_j(\mu,\sigma)}. \tag{20}$$

Startinis taškas naudojant (19) ir (20) gali būti parinktas kaip

$$\mu_0=\frac{1}{K}\sum_{j=1}^{K}\alpha_j^0\,, \tag{21}$$

$$\sigma_0^2 = \frac{1}{K}\sum_{j=1}^{K}(\alpha_j^0 - \mu^0)^2 , \qquad (22)$$

kur

$$\alpha_j^0 = \ln \frac{Y_j / N_j}{1 - Y_j / N_j}, \quad j = \overline{1, K} .$$

Parametrų μ, σ^2 ML įverčiai taip pat gali būti randami kintamos metrikos metodu (žr. Dennis and Schnabel (1996)), ML funkcijos gradientui vertinti naudojant pradinį tašką (21), (22) ir išraiškas (7), (8).

Pastebėsime, kad visi integralai formulėse (5), (6), (7), (8), (11), (12), (19), (20) gali būti skaičiuojami naudojant Hermit'o-Gauso kvadratūrines formules (žr. Abramovich, Stegun (1968)) arba naudojant kitas skaitinio integravimo formules.

Paprastai ML funkcijos integravimas ir minimizavimas gali būti atliekamas naudojantis atitinkamomis matematinių programų MATHCAD, MAPLE, ir t. t. funkcijomis.

4.1.4. Taikymai skirti duomenų analizei

Pateiktas metodas buvo panaudotas analizuoti 2003–2004 metų nužudymų ir savižudybių mirtingumo Lietuvoje duomenis (visi įvykiai populiacijoje, vyrai ir moterys atskirai). Skaitinis integravimas ir ML funkcijos minimizavimas buvo atliekamas su programa MATHCAD ir programavimo kalba Pascal. Šių duomenų analizės rezultatus naudojant Puasono-gama modelį žr. Jakimauskas, Sakalauskas (2010).

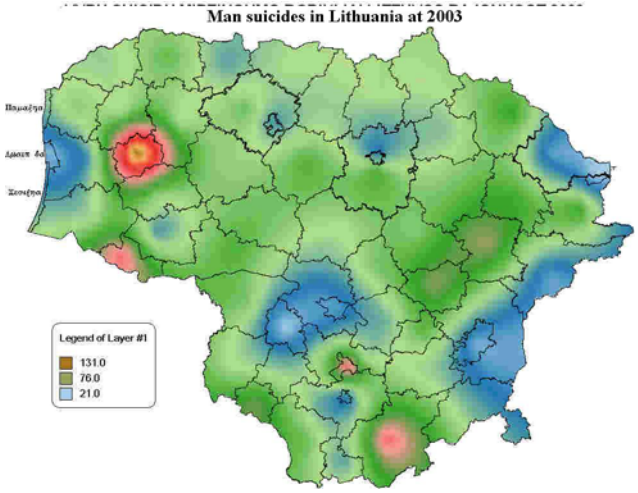
Nesinguliarumo sąlygos (13) tyrimo ir empirinio Bajeso įvykių tikimybių vertinimo rezultatai pateikti Lentelėse 4.1.1, 4.1.2.

	$\sum_{j=1}^{K}(Y_j - N_j \cdot P)^2 / \sum_{j=1}^{K} Y_j$	P^*10^5	μ	σ
AllSuic	14255.1/1434=9.941	41.656	-7.652	0.281
AllHom	453.3/325=1.395	9.440	-9.260	0.136
MenSuic	9484.2/1199=8.297	74.582	-7.707	0.288
MenHom	356.0/232=1.534	14.431	-8.840	0.159
WomenSuic	692.9/256=2.705	13.952	-8.826	0.371
WomenHom	76.8/100=0.768	5.450	-9.817	0.000

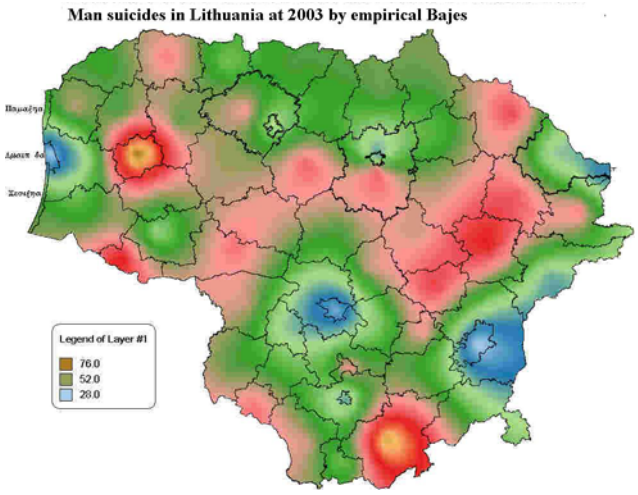
Lentelė 4.1.1. Savižudybių (*suicide*)/nužudymų (*homicide*) mirtingumo 2003 metais Lietuvoje tikimybių empirinis Bajeso vertinimas.

	$\sum_{j=1}^{K}(Y_j - N_j \cdot P)^2 / \sum_{j=1}^{K} Y_j$	P^*10^5	μ	σ
AllSuic	15421/1381	40.334	-7.652	0.278
AllHom	287/294	8.587	-9.260	0.000
MenSuic	11889/1124	70.337	-7.707	0.305
MenHom	313/200	12.515	-8.840	0.234
WomenSuic	1573.2/257	14.075	-8.826	0.257
WomenHom	84.1/93	5.093	-9.817	0.000

Lentelė 4.1.2. Savižudybių (*suicide*)/nužudymų (*homicide*) mirtingumo 2004 metais Lietuvoje tikimybių empirinis Bajeso vertinimas.



Pav. 4.1.1. Santykinės savižudybių tikimybės.



Pav. 4.1.2. Savižudybių tikimybių įverčiai, gauti empiriniu Bajeso metodu.

Tokiu būdu, apriorinio skirstinio singuliarumas, t. y. nulinė apriorinė dispersija buvo stebima tik su kai kuriais atskirais atvejais (moterų savižudybės 2003 m. ir 2004 m., ir visos savižudybės 2004 m.)

Pav. 4.1.1, 4.1.2 atvaizduotos santykinės nužudymų ir savižudybių tikimybės, įvertintos naudojant empirinį Bajeso metodą. Matome, kad empirinis Bajeso vertinimas įgalina pastebėti tam tikrus erdvinius nužudymų pasiskirstymo populiacijose efektus.

4.2. GAMA IR LOGIT MODELIAI EMPIRINIAME BAJESO MAŽŲ TIKIMYBIŲ VERTINIME

Nagrinėjama mažų tikimybių didelėse populiacijose (pvz., tam tikros ligos tikimybių, mirčių, savižudybių ir t. t., tikimybių) vertinimo problema. Nagrinėkime du nežinomų tikimybių pasiskirstymo modelius: tikimybės turi gama skirstinio modelį (modelis (A)), arba tikimybių logitai turi Gauso skirstinį (modelis (B)). Parinkome realius duomenis iš Lietuvos Statistikos departamento duomenų bazės (žr. *http://www.stat.gov.lt/*) – darbingo amžiaus asmenys, pirmą kartą pripažinti neįgaliaisiais pagal administracinę teritoriją (Lentelė M3140706), 2010 metai (teritorijų skaičius $K = 60$). Be to, naudojami vidutinio metinio gyventojų skaičiaus duomenys pagal administracinę teritoriją (Lentelė M3010211). Buvo gauti pradiniai parametrai naudojant paprastas iteracines procedūras modeliams (A) ir (B). Antrame etape buvo atlikti įvairūs testai naudojant Monte-Carlo modeliavimą (naudojant modelius (A) ir (B)) keičiant vieną pasirinktą parametrą ir skaičiuojant maksimalaus tikėtinumo (*maximum likelihood* (ML) įverčius atskirai imant modelį (A) ir modelį (B). Pagrindinis tikslas buvo parinkti tinkamiausią modelį ir duoti rekomendacijas naudoti gama ar logit modelį Bajeso vertinimui.

Rezultatai rodo, kad Monte-Carlo modeliavimas įgalina parinkti tinkamesnį vertinimo modelį.

Padarysime prielaidą, kad įvykių skaičius populiacijoje turi Puasono skirstinį su tam tikrais parametrais. Pastebėsime, kad tokia aproksimacija yra pakankamai tiksli didelėms populiacijoms ir mažoms, bet ne pernelyg mažoms, tikimybėms.

Empiriniame Bajeso vertinimo metode daroma prielaida, kad įvykių tikimybės populiacijose yra atsitiktinės ir turi tam tikrą skirstinį. Gerai žinoma (žr., pvz., (Clayton, Caldor, 1987), (Meza, 2003)), kad nežinomų tikimybių Bajeso įverčiai turi gerokai mažesnę vidutinę kvadratinę paklaidą palyginti su paprastais santykinės rizikos įverčiais.

Nagrinėkime du nežinomų tikimybių pasiskirstymo modelius: tikimybės turi gama skirstinį su formos (*shape*) parametru $\nu > 0$ ir mastelio (*scale*)

parametru $\alpha > 0$ (modelis (A)), arba tikimybių logitai turi Gauso skirstinį su vidurkiu μ ir dispersija σ^2 (modelis (B)).

Pradiniame etape buvo parinkti realūs duomenys iš Lietuvos Statistikos departamento duomenų bazės (žr. *http://www.stat.gov.lt/*) – darbingo amžiaus asmenys, pirmą kartą pripažinti neįgaliaisiais pagal administracinę teritoriją (Lentelė M3140706), 2010 metai (teritorijų skaičius K = 60). Be to, buvo naudojami vidutinio metinio gyventojų skaičiaus duomenys pagal administracinę teritoriją (Lentelė M3010211). Buvo gauti pradiniai parametrai (naudojant žemiau aprašytą paprastą iteracinę procedūrą) modeliams (A) ir (B). Antrojo etapo metu buvo atlikti įvairūs testai naudojant Monte-Carlo modeliavimą (naudojant modelius (A) ir (B)) keičiant vieną pasirinktą parametrą ir skaičiuojant maksimalaus tikėtinumo (*maximum likelihood* (ML) įverčius atskirai imant modelį (A) ir modelį (B). Pagrindinis tikslas buvo parinkti tinkamiausią modelį ir duoti rekomendacijas naudoti gama ar logit modelį Bajeso vertinimui.

4.2.1. Matematiniai modeliai

Tegul turime K populiacijų $A_1, A_2, \dots, A_K$, susidedančių, atitinkamai, iš N_j, ir tam tikri įvykiai (pvz., mirtis ar tam tikra liga) gali įvykti šiose populiacijose. Mes stebime įvykių skaičių $\{Y_j\} = Y_j, j = 1, 2, \dots, K$.

Padarykime prielaidą, kad įvykių skaičių lemia nežinomos tikimybės $\{\lambda_j\} = \lambda_j, j = 1, 2, \dots, K$, kurios yra lygios visiems individams iš tos pačios populiacijos. Tada įvykių skaičius $\{Y_j\}$ yra nepriklausomų atsitiktinių dydžių (n.a.d.) $\{\mathbf{Y}_j\} = \mathbf{Y}_j, j = 1, 2, \dots, K$, su binominiu skirstiniu (atitinkamai, su parametrais $(\lambda_j, N_j), j = 1, 2, \dots, K$), imtis. Aišku, kad,

$$\mathbf{E}\mathbf{Y}_j = \lambda_j N_j, j = 1, 2, \dots, K. \quad (1)$$

Paprastai daroma prielaida (žr., pvz., (Tsutakawa *et al.*, 1985), (Clayton, Caldor, 1987)), kad a.d. $\{\mathbf{Y}_j\}$ turi Puasono skirstinį su parametrais $\lambda_j N_j$, $j = 1, 2, \dots, K$,

$$\mathbf{P}\{\mathbf{Y}_j = m\} = h(m, \lambda_j N_j),\ m = 0, 1, \ldots;\ j = 1, 2, \ldots, K, \tag{2}$$

kur

$$h(m, z) = e^{-z}\frac{z^m}{m!},\ \ m = 0,\ 1,\ \ldots,\ \ z > 0, \tag{3}$$

Padarius tokią prielaidą, taip pat galioja (1).

Nagrinėsime matematinį modelį, darydami prielaidą, kad nežinomos tikimybės $\{\lambda_j\}$ yra nepriklausomi vienodai pasiskirstę (n.v.p.) a.d. su pasiskirstymo funkcija F iš tam tikros klasės $\mathcal{F}$. Mūsų tikslas yra gauti nežinomų tikimybių $\{\hat{\lambda}_j\}$ įverčius iš stebėtų įvykių skaičiaus $\{Y_j\}$, darant prielaidą, kad $F \in \mathcal{F}$.

Padarykime prielaidą, kad $\{\lambda_j\}$ yra n.v.p. gama a.d. su formos (*shape*) parametru $\nu > 0$ ir mastelio (*scale*) parametru $\alpha > 0$, t. y. pasiskirstymo funkcija (p.f.) F turi pasiskirstymo tankį

$$f(x) = f(x; \nu, \alpha) = \frac{\alpha \cdot (\alpha \cdot x)^{\nu-1}}{\Gamma(\nu)} e^{-\alpha x},\ \ 0 \le x < \infty. \tag{4}$$

Tuomet $\mathbf{E}\lambda_j = \nu\ /\ \alpha$, ir $\mathbf{D}\lambda_j = \nu\ /\ \alpha^2$. Be to,

$$\mathbf{E}(\lambda_j \mid \mathbf{Y}_j = Y_j) = \frac{Y_i + \nu}{N_j + \alpha},\ \ j = 1, 2, \ldots, K. \tag{5}$$

Pažymėkime šį modelį modeliu (A).

Nepriklausomai nuo $\{\lambda_j\}$ skirstinio, galime naudoti vidutinės santykinės rizikos (*mean relative risk* (MRR)) įvertį

$$\overline{\lambda}^{MRR} = \frac{\sum_{k=1}^{K} Y_k}{\sum_{k=1}^{K} N_k}, \tag{6}$$

Ir tokiu atveju laikome, kad $\{\overline{\lambda}_j^{MRR}\} \equiv \overline{\lambda}^{MRR}$, $j = 1, 2, \ldots, K$. Taip pat galime naudoti santykinės rizikos (*relative risk* (RR)) įvertį $\{\overline{\lambda}_j^{RR}\} = \overline{\lambda}_j^{RR}$, $j = 1, 2, \ldots, K$, kur

$$\overline{\lambda}_j^{RR} = \frac{Y_j}{N_j}, \quad j = 1, 2, \ldots, K. \tag{7}$$

Nagrinėsime empirinį Bajeso įvertį $\{\hat{\lambda}_j\}$, kuris yra tam tikras kompromisas tarp vidutinės santykinės rizikos įverčio $\{\overline{\lambda}_j^{MRR}\}$ ir santykinės rizikos įverčio $\{\overline{\lambda}_j^{RR}\}$. Šis įvertis gaunamas iš (5) naudojant parametrų $(\hat{\nu}, \hat{\alpha})$ įverčius.

Kaip alternatyvą nagrinėsime Bajeso įvertį $\{\tilde{\lambda}_j\}$, kuris gaunamas padarius prielaidą, kad nežinomos tikimybės yra n.v.p. a.d., tokie, kad jų logitai

$$\alpha_j = \ln \frac{\lambda_j}{1 - \lambda_j}, \; j = 1, 2, \ldots, K, \tag{8}$$

yra n.v.p. Gauso a.d. su vidurkiu μ ir dispersija σ^2. Pažymėkime šį modelį modeliu (B). Šiuo atveju sąlyginis $\{\lambda_j\}$ vidurkis turi tokią formą (žr. (Sakalauskas (1995), Gurevičius *et al.,* (2009)):

$$\mathbf{E}(\lambda_j \mid \mathbf{Y}_j = Y_j) = \frac{\int_{-\infty}^{\infty} \frac{1}{1+e^{-x}} h(Y_j, \frac{N_j}{1+e^{-x}}) \varphi(x; \mu, \sigma^2) dx}{D_j(\mu, \sigma^2)}, \tag{9}$$

$$D_j(\mu, \sigma^2) = \int_{-\infty}^{\infty} h(Y_j, \frac{N_j}{1+e^{-x}}) \varphi(x; \mu, \sigma^2) dx. \tag{10}$$

Nagrinėjant modelį (A), atitinkama maksimalaus tikėtinumo funkcija turi tokį pavidalą:

$$L_A(\nu,\alpha)=\sum_{j=1}^{K}\left(\ln\frac{\Gamma(Y_j+\nu)}{\Gamma(\nu)}+\nu\ln(\alpha)-(Y_j+\nu)\ln(N_j+\alpha)+Y_j\ln N_j\right) \tag{11}$$

Modelio (B) atveju, atitinkama maksimalaus tikėtinumo funkcija turi tokį pavidalą:

$$L_B(\mu,\sigma^2)=\sum_{j=1}^{K}\left(\ln D_j(\mu,\sigma^2)\right). \tag{12}$$

Maksimalaus tikėtinumo įverčiai $\{\hat{\lambda}_j^*\}$ ir $\{\tilde{\lambda}_j^*\}$ gaunami maksimizuojant (11), atitinkamai (12), ir pakeičiant parametrų reikšmes formulėje (5) ar (10) reikšmėmis (ν^*,α^*) ar $(\mu^*,(\sigma^*)^2)$. Praktikoje naudojami apytiksliai įverčiai $\{\hat{\lambda}_j\}$ ir $\{\tilde{\lambda}_j\}$, gaunami skaitiniais metodais (paprastai iteracinėmis procedūromis) apytikslių parametrų reikšmių $(\hat{\nu},\hat{\alpha})$, atitinkamai $(\tilde{\mu},\tilde{\sigma}^2)$, radimui.

4.2.2. Modeliavimo rezultatai

Kaip pradinius duomenis $\{Y_j\}$ modeliavimui parinkome realius duomenis iš Lietuvos Statistikos departamento duomenų bazės – darbingo amžiaus asmenys, pirmą kartą pripažinti neįgaliaisiais pagal administracinę teritoriją (Lentelė M3140706), 2010 metai (teritorijų skaičius K = 60), iš viso 9 duomenų rinkiniai:

1 – Iš viso (15432 atvejai)

2 – Tuberkuliozė (399)

3 – Piktybiniai navikai (2706)

4 – Psichikos ir elgesio sutrikimai (1303)

5 – Nervų sistemos ligos (1501)

6 – Kraujotakos sistemos ligos (3525)

7 – Jungiamojo audinio ir skeleto-raumenų sistemos ligos (2566)

8 – Traumos, apsinuodijimai (1116)

9 – Kitos priežastys (2316)

Taip pat buvo naudojamas vidutinis metinis gyventojų skaičius pagal administracinę teritoriją (Lentelė M3010211); iš viso gyventojų skaičius visose administracinėse teritorijose lygus 3286820.

Pradiniame etape realūs imties duomenys buvo įvertinti tiek modeliu (A), tiek modeliu (B), t. y. gauname startinius įverčius $\{\hat{\lambda}_j\}_0$ ir $\{\tilde{\lambda}_j\}_0$.

Modelio (A) atveju startinius įverčius gauname pasinaudoję iteracine procedūra (žr. (Clayton, Caldor, 1987)) naudodami apibrėžimą

$$\hat{\theta}_j = \frac{Y_j + \nu}{N_j + \alpha}, \quad j = 1, 2, \ldots, K, \tag{13}$$

ir dvi lygtis:

$$\frac{\nu}{\alpha} = \frac{1}{K}\sum_{j=1}^{K}\hat{\theta}_j\,, \tag{14}$$

ir

$$\frac{\nu}{\alpha^2} = \frac{1}{K-1}\sum_{j=1}^{K}(1+\frac{\alpha}{N_j})(\hat{\theta}_j - \frac{\nu}{\alpha})^2\,. \tag{15}$$

Iš (14) ir (15) gauname α kaip kvadratinės lygties šaknį ir po to gauname ν naudodami (15). Iteracinė procedūra startuoja nuo $\{\hat{\theta}_j\}_0 = \{\lambda_j^{RR}\}$, tada iš (14) ir (15) gauname $(\nu, \alpha)_0$, iš (13) gauname $\{\hat{\theta}_j\}_1$, ir t. t.

Čia pateikiama Puasono-gama modelio (modelio A) nesinguliarumo sąlyga, gaunama analogiškai, kaip ir Puasono-Gauso modelio (modelio B) nesinguliarumo sąlyga (Sakalauskas, 1995, žr. praėjusio skyrelio formulę (13)).

Jei ši sąlyga nepatenkinta, ieškant maksimalaus tikėtinumo funkcijos maksimumo, $\alpha,\nu \to \infty$,

$$\nu / \alpha \to \{\overline{\lambda}^{MRR}\} \overset{\text{def.}}{=} P, \tag{16}$$

nežinomų tikimybių empiriniai Bajeso įverčiai artėja prie konstantos $\{\hat{\lambda}_j\} = \{\hat{\theta}_j\} \to \{\overline{\lambda}^{MRR}\}$, o maksimalaus tikėtinumo funkcijos maksimumas nėra pasiekiamas.

Maksimizuojant maksimalaus tikėtinumo funkciją (11) reikia rasti tokias parametrų reikšmes (ν^*,α^*), kurioms maksimalaus tikėtinumo funkcijos išvestinė tiek pagal ν, tiek pagal α, lygi nuliui. Prilyginus išvestinę pagal α nuliui, gauname lygtį

$$\frac{\nu}{\alpha} = \frac{1}{K}\sum_{j=1}^{K}\frac{Y_j+\nu}{N_j+\alpha} \overset{\text{def.}}{=} \frac{1}{K}\sum_{j=1}^{K}\theta_j\,, \tag{17}$$

o prilyginę išvestinę pagal ν nuliui, gauname lygtį

$$\sum_{j=1}^{K}\sum_{s=0}^{Y_j-1}\frac{1}{\nu+s} + K\ln\alpha - \sum_{j=1}^{K}\ln(N_j+\alpha) = 0\,. \tag{18}$$

Padauginę (18) lygtį iš α, ir ją pertvarkius, gausime

$$\frac{\alpha}{\nu}\sum_{j=1}^{K}\sum_{s=0}^{Y_j-1}\frac{1}{1+s/\nu} - \alpha\sum_{j=1}^{K}\ln(1+N_j/\alpha) = 0\,. \tag{19}$$

Kuomet $\nu,\alpha \to \infty$, $\nu/\alpha = P\cdot(1+o(1/\nu))$, pasinaudoję formulėmis $1/(1+x) = 1-x+O(x^2)$ ir $\ln(1+x) = x-x^2/2+O(x^3)$, kai $x \to 0$, iš (19) gausime, kad

$$\lim_{\substack{\nu,\alpha\to\infty \\ \nu/\alpha=P\cdot(1+o(1/\nu))}} \alpha\frac{\partial L(\nu,\alpha)}{\partial\nu} = \lim_{\substack{\nu,\alpha\to\infty \\ \nu/\alpha=P\cdot(1+o(1/\nu))}} \left(\frac{\alpha}{\nu}\sum_{j=1}^{K}\sum_{s=0}^{Y_j-1}(1-s/\nu) - \alpha\sum_{j=1}^{K}\left(N_j/\alpha - N_j^2/2\alpha^2\right)\right), \quad (20)$$

arba, pasinaudojus aritmetinės progresijos formule, suprastinus narius, ir padauginus iš $2\alpha P^2$

$$\lim_{\substack{\nu,\alpha\to\infty \\ \nu/\alpha=P\cdot(1+o(1/\nu))}} 2\nu^2\frac{\partial L(\nu,\alpha)}{\partial\nu} = -\sum_{j=1}^{K}Y_j(Y_j-1) + \sum_{j=1}^{K}(N_j\cdot P)^2 = \sum_{j=1}^{K}Y_j - \sum_{j=1}^{K}(Y_j^2-(N_j\cdot P)^2). \quad (21)$$

Taigi, jei stebimas įvykių skaičius $\{Y_j\} = Y_j, j = 1, 2, \ldots, K$, populiacijose $A_1, A_2, \ldots, A_K$, susidedančių, atitinkamai, iš $N_j, j = 1, 2, \ldots, K$, individų, patenkina nesinguliarumo sąlygą

$$\sum_{j=1}^{K}Y_j - \sum_{j=1}^{K}(Y_j^2-(N_j\cdot P)^2) < 0, \quad (22)$$

tuomet egzistuoja maksimalaus tikėtinumo funkcijos (11) maksimumas tam tikriems baigtiniams α ir ν, kadangi, kaip nesunku įsitikinti, pakankamai mažoms α ir ν reikšmėms maksimalaus tikėtinumo funkcijos išvestinė pagal ν yra didesnė už 0.

Modelio (B) atveju buvo naudojama iteracinė procedūra (žr. Gurevičius *et al.* (2009)), kuri naudoja ML įverčio išvestines. Šiuo atveju iteracinė procedūra startuoja nuo $(\mu, \sigma^2)_0 = (\mu_0, (\sigma^2)_0)$, kur

$$\mu_0 = \frac{1}{K}\sum_{j=1}^{K}\ln\frac{\overline{\lambda}_j^{RR}}{1-\overline{\lambda}_j^{RR}}, \quad (23)$$

$$(\sigma^2)_0 = \frac{1}{K}\sum_{j=1}^{K}\left(\ln\frac{\overline{\lambda}_j^{RR}}{1-\overline{\lambda}_j^{RR}} - \mu_0\right)^2. \quad (24)$$

Iš $(\mu, \sigma^2)_i = (\mu_i, (\sigma^2)_i)$, nauji parametrai $(\mu, \sigma^2)_{i+1} = (\mu_{i+1}, (\sigma^2)_{i+1})$ gaunami naudojant šias formules (skaičiavimams buvo naudojami skaitiniai metodai iš (Abramovich, Stegun (1968)):

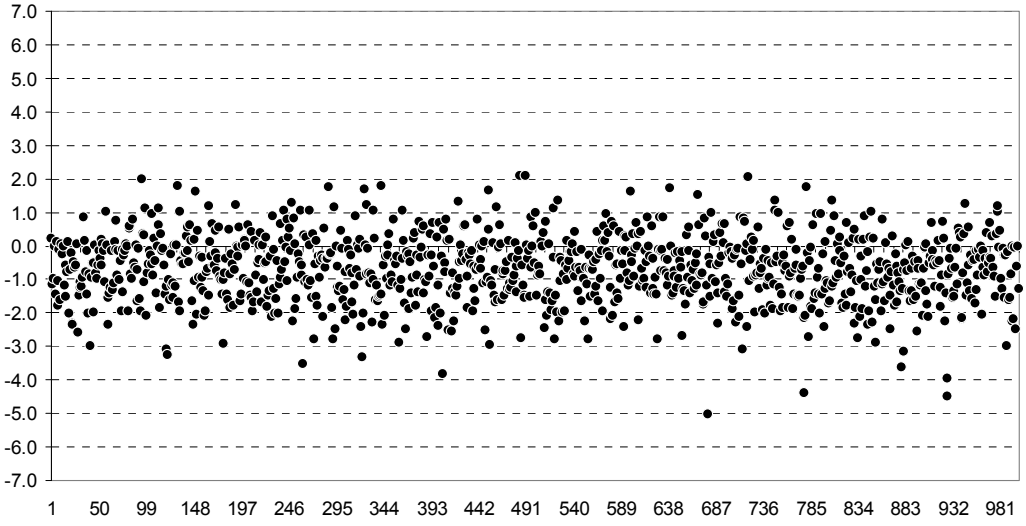
$$\mu_{i+1} = \frac{1}{K}\sum_{j=1}^{K} \frac{\int_{-\infty}^{\infty} xh(Y_j, \frac{N_j}{1+e^{-x}})\varphi(x; \mu_i, (\sigma^2)_i)dx}{D_j(\mu_i, (\sigma^2)_i)}, \tag{25}$$

$$(\sigma^2)_{i+1} = \frac{1}{K}\sum_{j=1}^{K} \frac{\int_{-\infty}^{\infty} (x-\mu_i)^2 h(Y_j, \frac{N_j}{1+e^{-x}})\varphi(x; \mu_i, (\sigma^2)_i)dx}{D_j(\mu_i, (\sigma^2)_i)}. \tag{26}$$

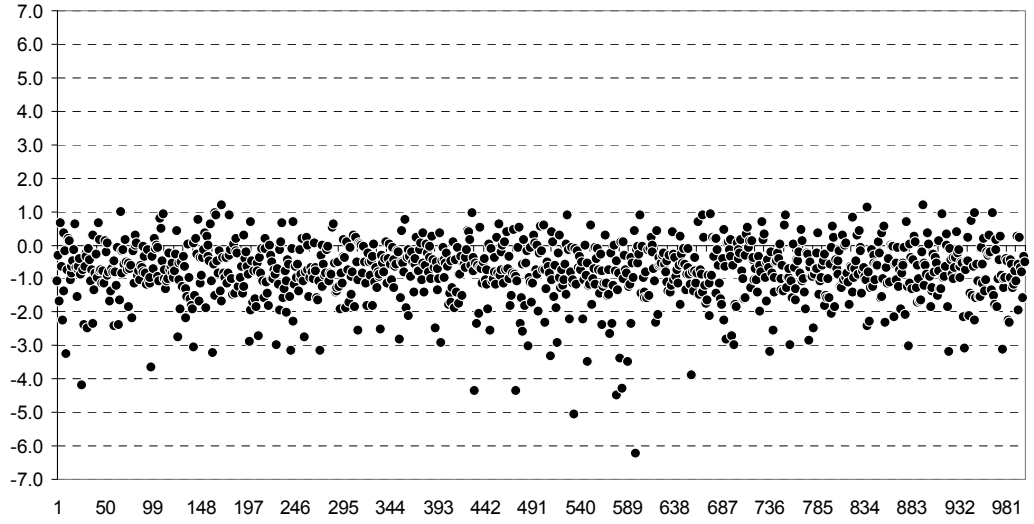
Turint startinį įvertį, buvo modeliuojamos atsitiktinės realizacijos naudojant Monte-Carlo metodą (paprastai 100 nepriklausomų realizacijų), naudojant modelį (A) ir modelį (B). Kiekvieno modelio atveju, generuotoms realizacijos buvo skaičiuojamas įvertis $\{\hat{\lambda}_j\}$, įvertis $\{\tilde{\lambda}_j\}$, po to – parametrų įverčiai $(\hat{\nu}, \hat{\alpha})$ ir $(\tilde{\mu}, \tilde{\sigma}^2)$, atitinkamai, ir atitinkamos ML funkcijos reikšmės $L_A(\hat{\nu}, \hat{\alpha})$ ir $L_B(\tilde{\mu}, \tilde{\sigma}^2)$.

Įverčiams $\{\hat{\lambda}_j\}$ ir $\{\tilde{\lambda}_j\}$ rasti buvo pritaikytas paprastas optimizavimo algoritmas. Tegul turime startinį įvertį, pvz., $\{\hat{\lambda}_j\}_0$, ir atitinkamus parametrus $(\nu, \alpha)_0$. Sukonstruojame dydžio m x m stačiakampę gardelę (buvo parinktas m = 5) su gardelės žingsniais h_1 ir h_2 ir centru taške $(\nu, \alpha)_0$. Modeliui (A) buvo parinkta h_1 = 0.1 parametrui ν, ir h_2 = 100.0 parametrui α. Atitinkamai, modeliui (B) buvo parinkta h_1 = 0.001 parametrui μ, ir h_2 = 0.001 parametrui σ. Iteracijos įverčiui $\{\hat{\lambda}_j\}$, atitinkamai įverčiui $\{\tilde{\lambda}_j\}$, rasti atliekamos skaičiuojant ML funkcijos reikšmes gardelės taškuose. Jei maksimali reikšmė nėra gardelės centre, paslenkame gardelę taip, kad maksimali reikšmė būtų naujos gardelės centre, ir suskaičiuojame ML funkcijos reikšmes naujos gardelės taškuose. Jei maksimali reikšmė yra gardelės centre, sumažiname

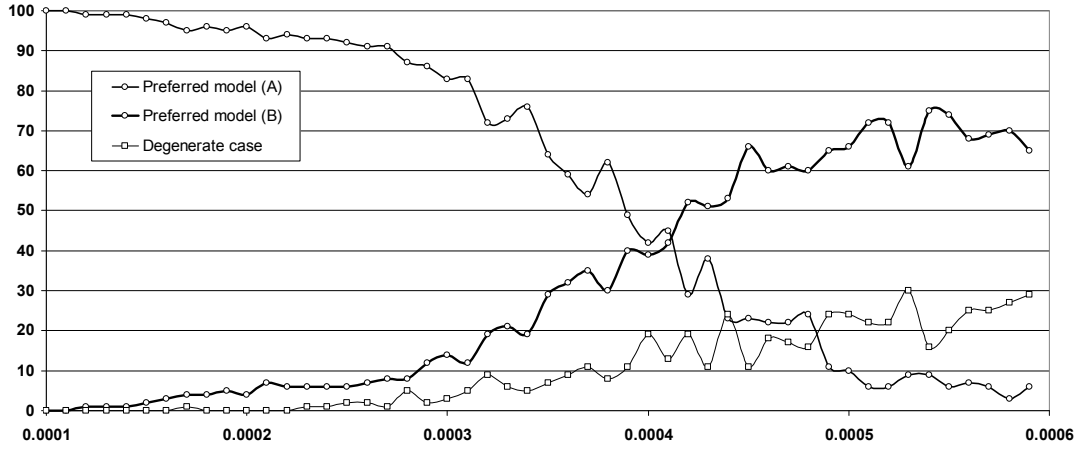
gardelės žingsnius (buvo parinkta $h_1/1.1$ ir $h_2/1.1$ modelio (A) atveju, ir $h_1/1.15$ ir $h_2/1.15$ modelio (B) atveju) ir pakartojame procedūrą naujoje gardelėje. Buvo parinktas toks sustojimo kriterijus: maksimalus iteracijų skaičius 100 arba abiejų parametrų išvestinės mažesnės kaip 0.02.



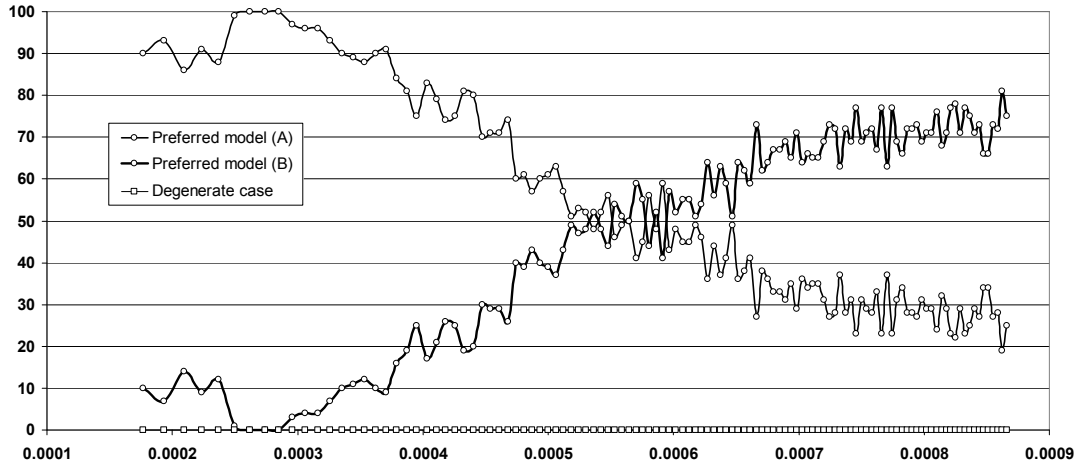
Pav. 4.2.1. ML funkcijų skirtumai $L_A - L_B$, generavimas su modeliu (A), duomenų rinkinys nr. 1.



Pav. 4.2.2. ML funkcijų skirtumai $L_A - L_B$, generavimas su modeliu (B), duomenų rinkinys nr. 1.



Pav. 4.2.3. Modelių (A) ir (B) efektyvumas, generavimas su modeliu (A), ant *x* ašies ν/α reikšmės, fiksuota reikšmė $\nu/\alpha^2 = 32/(64000)^2$.



Pav. 4.2.4. Modelių (A) ir (B) efektyvumas, generavimas su modeliu (A), ant *x* ašies ν/α reikšmės, fiksuota reikšmė $\nu/\alpha^2 = 4/(8000)^2$.

Rezultatai rodo, kad daugumai duomenų rinkinių 1–9 realizacijų (generuotų su modeliu (A), ir generuotų su modeliu (B)) paprastai $L_B(\tilde{\mu},\tilde{\sigma}^2) > L_A(\hat{\nu},\hat{\alpha})$.

Rezultatai duoda ML funkcijų palyginimą naudojant skirtingus generavimo modelius ir skirtingus vertinimo modelius (1000 nepriklausomų realizacijų). Čia pateiksime tik rezultatus duomenų rinkiniui nr. 1 (žr. Pav. 4.2.1–4.2.2). Šiam duomenų rinkiniui vertinimas su modeliu (B) yra labiau tinkamas (*preferable*), nei duomenims, generuotiems su modeliu (A). Kalbant apie kitus duomenų rinkinius, rezultatai rodo, kad mažesniems duomenų rinkinio dydžiams modelis (B) tampa dar labiau tinkamas.

Sprendžiant modelio (A) ar modelio (B) tinkamumą, daug kas priklauso nuo pradinių modelio parametrų. Tai buvo patikrinta generuojant nepriklausomas realizacijas su įvairiomis parametrų reikšmėmis. Pateiksime modeliavimo rezultatus (žr. Pav. 4.2.3) reikšmių aibei $\alpha = (1280, 14080, 15360, \ldots 39680)$ su vienodais intervalais ir fiksuota reikšme $\nu/\alpha^2 = 32/(64000)^2$. Kiekvienai parametrų ($\nu$, α)) porai buvo generuotos 100 nepriklausomų realizacijų. Pav. 4.2.3 ašyje *x* turime ν/α reikšmes. Pav. 4.2.4 pateikti modeliavimo rezultatai reikšmių aibei $\nu = (0.5\alpha_1, 0.6\alpha_2, 0.7\alpha_3, \ldots, 12.0\alpha_{116})$ su tokiais α_j, kad būtų fiksuota reikšmė $\nu/\alpha^2 = 4/(8000)^2$. Abiem atvejais atitinkamai *y* ašyje pateikta (procentais) tinkamiausio vertinimo modelio dalis ir išsigimusių atvejų skaičiaus dalis (žr. (Gurevičius *et al.* (2009)).

Rezultatai rodo, kad Monte-Carlo modeliavimo metodas leidžia nuspręsti, kuris vertinimo modelis yra tinkamesnis.

4.3. MODIFIKUOTAS REGRESINIS EMPIRINIS BAJESO ĮVERTIS MAŽŲ TIKIMYBIŲ VERTINIMUI

Nagrinėsime papildomo regresijos kintamojo pridėjimo prie logit modelio efektyvumą nagrinėjant mažų tikimybių didelėse populiacijose problemą. Nagrinėkime du nežinomų tikimybių pasiskirstymo modelius: tikimybės pasiskirsčiusios pagal gama skirstinį (modelis (A)), arba tikimybių logit'ai turi Gauso skirstinį (modelis (B)). Modifikuotame modelyje B naudosime papildomą regresijos kintamąjį Gauso skirstinio vidurkiui (modelis BR).

Naudosime realius duomenis iš Lietuvos Statistikos departamento duomenų bazės (žr. http://www.stat.gov.lt/)

Pagrindiniai duomenys yra – darbingo amžiaus asmenys, pirmą kartą pripažinti neįgaliaisiais pagal administracinę teritoriją (Lentelė M3140706), 2010 metai (teritorijų skaičius $K = 60$).

Be to, naudosime vidutinio metinio gyventojų skaičiaus duomenis pagal administracinę teritoriją (Lentelė M3010211).

Papildomas regresijos kintamasis paremtas šiais duomenimis – ligoninėse gydytų ligonių skaičius (Lentelė 3140312), 2010 metai.

Buvo gauti pradiniai parametrai naudojant paprastas iteracines procedūras modeliams (A), (B) ir (BR). Antrame etape buvo atlikti įvairūs testai naudojant Monte-Carlo modeliavimą (naudojant modelius (A), (B) ir (BR)). Pagrindinis tikslas buvo parinkti tinkamiausią modelį ir duoti rekomendacijas naudoti gama ar logit modelį (su ar be papildomu regresijos kintamuoju) Bajeso vertinimui. Rezultatai rodo, kad Monte-Carlo modeliavimas įgalina parinkti tinkamesnį vertinimo modelį.

4.3.1. Matematiniai modeliai

Tegul įvykių skaičius $\{Y_j\}$ yra nepriklausomų atsitiktinių dydžių (a.d.) $\{\mathbf{Y}_j\} = \mathbf{Y}_j$, $j = 1, 2, \ldots, K$, su binominiu skirstiniu (atitinkamai, su parametrais (λ_j, N_j), $j = 1, 2, \ldots, K$), imtis. Aišku, kad,

$$\mathbf{E}\mathbf{Y}_j = \lambda_j N_j, j = 1, 2, \dots, K. \quad (1)$$

Dažnai daroma prielaida (žr., pvz., (Tsutakawa *et al.*, 1985), (Clayton, Caldor, 1987)), kad a.d. $\{\mathbf{Y}_j\}$ turi Puasono skirstinį su parametrais $\lambda_j N_j$, $j = 1, 2,\dots, K$,

$$\mathbf{P}\{\mathbf{Y}_j = m\} = h(m, \lambda_j N_j), m = 0, 1, \dots; j = 1, 2, \dots, K,$$

kur

$$h(m, z) = e^{-z}\frac{z^m}{m!}, \ m = 0, 1, \dots, \ z > 0,$$

Padarius tokią prielaidą, taip pat galioja (1).

Nagrinėsime matematinį modelį, laikydami, kad nežinomos tikimybės $\{\lambda_j\}$ yra nepriklausomi vienodai pasiskirstę (n.v.p.) a.d su pasiskirstymo funkcija F iš tam tikros klasės $\mathcal{F}$. Mūsų uždavinys yra rasti nežinomų tikimybių įverčius $\{\hat{\lambda}_j\}$ iš stebimo įvykių skaičiaus $\{Y_j\}$, laikant, kad $F \in \mathcal{F}$.

Padarykime prielaidą, kad $\{\lambda_j\}$ yra n.v.p. gama a.d. su formos (*shape*) parametru $\nu > 0$ ir mastelio (*scale*) parametru $\alpha > 0$, t. y. pasiskirstymo funkcija (p.f.) F turi pasiskirstymo tankį

$$f(x) = f(x; \nu, \alpha) = \frac{\alpha \cdot (\alpha \cdot x)^{\nu-1}}{\Gamma(\nu)} e^{-\alpha x}, \ 0 \le x < \infty.$$

Tuomet $\mathbf{E}\lambda_j = \nu / \alpha$, ir $\mathbf{D}\lambda_j = \nu / \alpha^2$. Be, to,

$$\mathbf{E}(\lambda_j \mid \mathbf{Y}_j = Y_j) = \frac{Y_i + \nu}{N_j + \alpha}, \ j = 1, 2, \dots, K. \quad (2)$$

Pažymėkime šį modelį modeliu (A).

Empirinis Bajeso įvertis $\{\hat{\lambda}_j\}$, kuris yra tam tikras kompromisas tarp vidutinės santykinės rizikos įverčio (*mean relative risk estimate* (MRR)) $\{\overline{\lambda}_j^{MRR}\}$ ir santykinės rizikos įverčio (*relative risk estimate* (RR)) $\{\overline{\lambda}_j^{RR}\}$ gaunamas iš (2) naudojant parametrų įverčius $(\hat{\nu},\hat{\alpha})$.

Kaip alternatyvą nagrinėsime empirinį Bajeso įvertį $\{\widetilde{\lambda}_j\}$, kuris gaunamas padarius prielaidą, kad nežinomos tikimybės yra n.v.p. a.d., tokie, kad jų logitai

$$\alpha_j = \ln\frac{\lambda_j}{1-\lambda_j},\ j = 1, 2, \ldots, K,$$

yra n.v.p. Gauso a.d. su vidurkiu μ ir dispersija σ^2. Pažymėkime šį modelį modeliu (B).

Be to, įveskime papildomą regresijos kintamąjį Z_j, laikydami, kad $\mu = \mu(j) = \mu_0 + \mu_1 Z_j$, $j = 1, 2, \ldots, K$. Pažymėkime šį modelį modeliu (BR). Šis kintamasis laikomas neatsitiktiniu, todėl visos formulės modeliui (B) galioja taip pat ir modeliui (BR).

Modelio (B) ir modelio (BR) atveju sąlyginis $\{\lambda_j\}$ vidurkis turi tokią formą:

$$\mathbf{E}(\lambda_j \mid \mathbf{Y}_j = Y_j) = \frac{\int_{-\infty}^{\infty} \frac{1}{1+e^{-x}} h(Y_j, \frac{N_j}{1+e^{-x}})\varphi(x;\mu,\sigma^2)dx}{D_j(\mu,\sigma^2)}, \tag{3}$$

$$D_j(\mu,\sigma^2) = \int_{-\infty}^{\infty} h(Y_j, \frac{N_j}{1+e^{-x}})\varphi(x;\mu,\sigma^2)dx.$$

Nagrinėjant modelį (A), atitinkama ML funkcija turi tokį pavidalą:

$$L_A(\nu,\alpha) = \sum_{j=1}^{K}\left(\ln\frac{\Gamma(Y_j+\nu)}{\Gamma(\nu)} + \nu\ln(\alpha) - (Y_j+\nu)\ln(N_j+\alpha) + Y_j\ln N_j\right) \tag{4}$$

Modelio (B) ir modelio (BR) atveju, atitinkama ML funkcija turi tokį pavidalą:

$$L_B(\mu,\sigma^2)=\sum_{j=1}^{K}\left(\ln D_j(\mu,\sigma^2)\right). \tag{5}$$

ML įverčiai $\{\hat{\lambda}_j^*\}$ ir $\{\tilde{\lambda}_j^*\}$ gaunami maksimizuojant (4), atitinkamai (5), ir pakeičiant parametrų reikšmes formulėse (2) ar (3) reikšmėmis (ν^*,α^*) ar, atitinkamai $(\mu^*,(\sigma^*)^2)$. Praktikoje naudojami apytiksliai įverčiai $\{\hat{\lambda}_j\}$ ir $\{\tilde{\lambda}_j\}$ gaunami naudojant skaitinius metodus (paprastai iteracines procedūras) tam, kad rastume apytiksles parametrų $(\hat{\nu},\hat{\alpha})$, atitinkamai, $(\tilde{\mu},\tilde{\sigma}^2)$, reikšmes.

4.3.2. Kompiuterinio modeliavimo rezultatai

Kaip pradinius duomenis $\{Y_j\}$ modeliavimui parinkome realius duomenis iš Lietuvos Statistikos departamento duomenų bazės – darbingo amžiaus asmenys, pirmą kartą pripažinti neįgaliaisiais pagal administracinę teritoriją (Lentelė M3140706), 2010 metai (teritorijų skaičius K = 60), viso 9 duomenų rinkiniai:

1 – Iš viso (15432 atvejai)

2 – Tuberkuliozė (399)

3 – Piktybiniai navikai (2706)

4 – Psichikos ir elgesio sutrikimai (1303)

5 – Nervų sistemos ligos (1501)

6 – Kraujotakos sistemos ligos (3525)

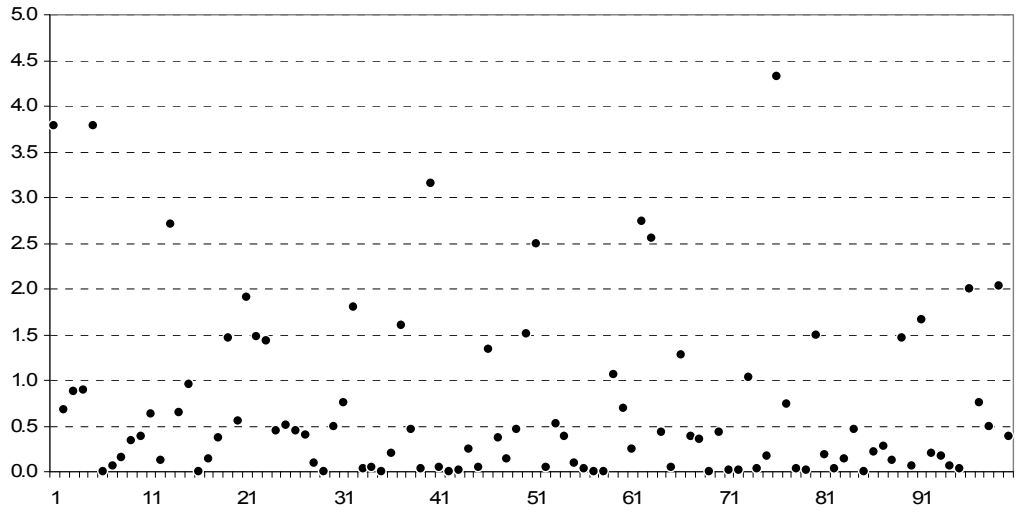
7 – Jungiamojo audinio ir skeleto-raumenų sistemos ligos (2566)

8 – Traumos, apsinuodijimai (1116)

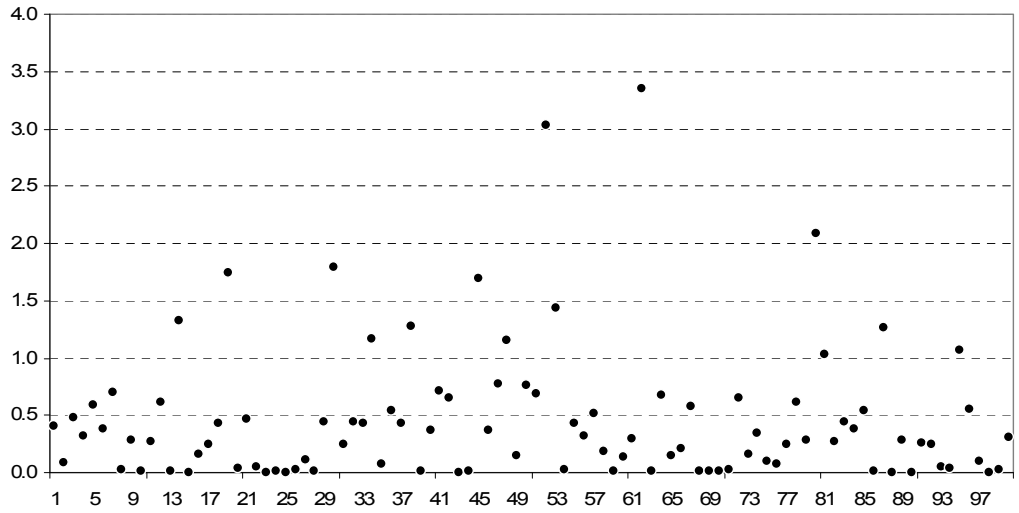
9 – Kitos priežastys (2316)

Taip pat buvo naudojamas vidutinis metinis gyventojų skaičius pagal administracinę teritoriją (Lentelė M3010211), iš viso gyventojų skaičius visose administracinėse teritorijose lygus 3286820.

Papildomas regresijos kintamasis buvo paremtas šiais duomenimis – ligoninėse gydytų ligonių skaičius pagal administracinę teritoriją (Lentelė 3140312), 2010 metai. Kintamasis Z_j, $j = 1, 2, \ldots, K$, buvo sukonstruotas padalijus ligoninėse gydytų ligonių skaičių iš gyventojų skaičiaus kiekvienoje administracinėje teritorijoje.

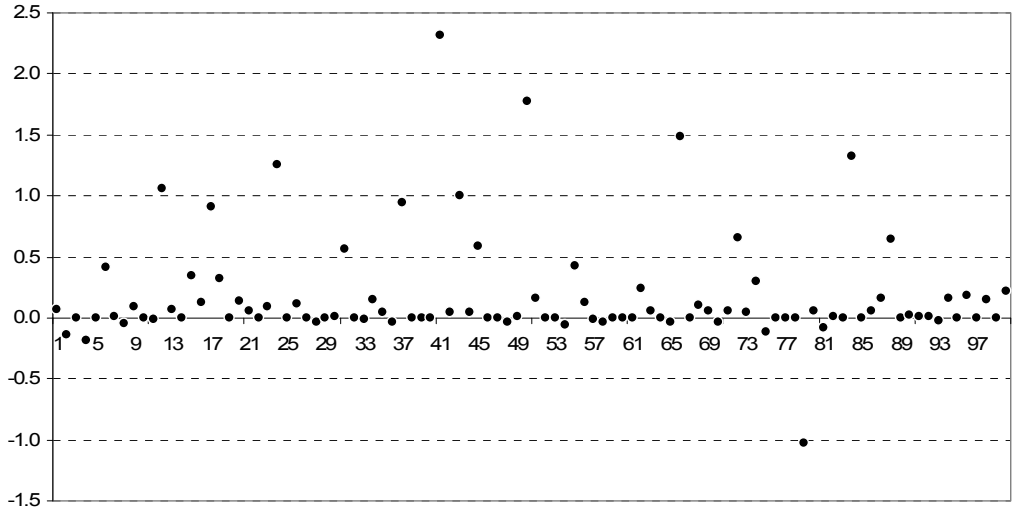


Pav. 4.3.1. Skirtumas L_{BR}–L_B (duomenų rink. 6, generavimas su modeliu (A).

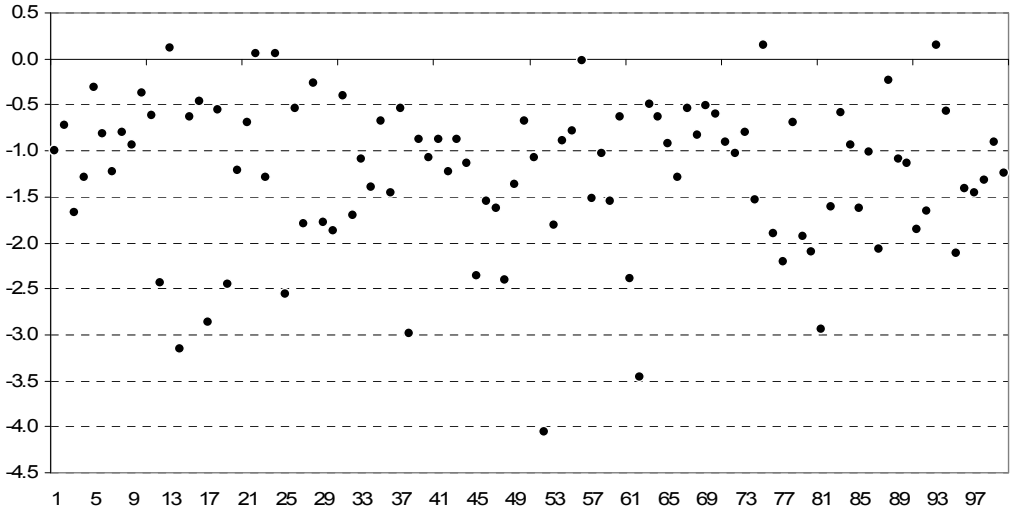


Pav. 4.3.2. Skirtumas L_{BR}–L_B (duomenų rink. 6, generavimas su modeliu (B).

Pradiniame etape realūs imties duomenys buvo įvertinti naudojant modelį (A) ir modelį (B), t. y. buvo gauti startiniai įverčiai $\{\hat{\lambda}_j\}_0$ ir $\{\tilde{\lambda}_j\}_0$. Modelio (BR) atveju startinis parametras $\mu_{.1} = 0$. Detaliau apie startinių parametrų gavimą žr. Jakimauskas (2012).



Pav. 4.3.3. Skirtumas L_A–L_B (duomenų rinkinys 6, generavimas su modeliu (B).



Pav. 4.3.4. Skirtumas L_A–L_{BR} (duomenų rinkinys 6, generavimas su modeliu (B).

Turint startinius įverčius, buvo generuojamos atsitiktinės realizacijos naudojant Monte-Carlo metodą (paprastai 100 nepriklausomų realizacijų), naudojant modelius (A), (B) ir (BR). Kiekvieno modelio atveju buvo apskaičiuoti įverčiai $\{\hat{\lambda}_j\}$, $\{\widetilde{\lambda}_j\}$, ir atitinkamos ML funkcijos reikšmės $L_A(\hat{\nu},\hat{\alpha})$, $L_B(\widetilde{\mu},\widetilde{\sigma}^2)$ ir $L_{BR}(\widetilde{\mu},\widetilde{\sigma}^2)$. Detaliau apie panaudotą optimizavimo algoritmą žr. Jakimauskas (2012).

Pagrindinis tikslas buvo palyginti modelio (BR) efektyvumą su modelio (B) efektyvumu lyginant ML funkcijų $L_B(\widetilde{\mu},\widetilde{\sigma}^2)$ ir $L_{BR}(\widetilde{\mu},\widetilde{\sigma}^2)$ skirtumus. Rezultatai rodo, kad kai kuriems duomenų rinkiniams šis skirtumas yra pakankamai didelis, kad būtų pasirinktas modelis (BR) vietoje modelio (A), nors modelis (A) yra labiau tinkamas už modelį (B). Žemiau pateikiami kai kurie modeliavimo rezultatai.

4.4. IŠVADOS

Empirinio Bajeso metodo Puasono-Gauso modelio atveju yra žinoma (Sakalauskas (1995)) šio modelio singuliarumo sąlyga. Darbe nustatyta singuliarumo sąlyga Puasono-gama modelio atveju.

Darbe atliktas kompiuterinis eksperimentas nustatant, kokią įtaką parametrų vertinimui turi modelio parametrų artėjimas prie singuliarumo, generuojant duomenis pagal Puasono-gama modelį, o vertinimui naudojant tiek Puasono-Gauso modelį, tiek Puasono-gama modelį. Taip pat ištirta, kokią įtaką generuotų duomenų modelio parametrai turi parenkant tinkamiausią modelį. Nustatyta, kad, mažėjant parametrui, kuris nusako vidutinį stebimų įvykių skaičių, tinkamesnis yra Puasono-gama modelis, o didėjant šiam parametrui, nuo tam tikros jo reikšmės, tinkamesnis yra Puasono-Gauso modelis.

Sudaryti skaitiniai algoritmai kad būtų galima apskaičiuoti empirinio Bajeso metodo parametrus Puasono-Gauso, tiek Puasono-gama modelio atveju. Monte-Carlo metodu atliktas didžiausio tikėtinumo funkcijos reikšmių palyginimas Puasono-Gauso ir Puasono-gama modelių atveju, naudojant tiek generuotus duomenis, tiek realius duomenis iš Statistikos departamento duomenų bazės.

Pritaikius empirinį Bajeso metodą sudarytas Puasono-Gauso regresinis modelis, kuris buvo pritaikytas realiems duomenims iš Statistikos departamento duomenų bazės tirti, ir parodytas jo efektyvumas, kuomet tinkamai parinktas regresijos kintamasis leidžia žymiai padidinti didžiausio tikėtinumo funkcijos reikšmę.

Naudojant tiek realius duomenis iš Statistikos departamento duomenų bazės, tiek ir modeliuotus duomenis, buvo parodyta, kad panaudojus Monte-Carlo modeliavimą, keičiant pasirinktus parametrus ir skaičiuojant maksimalaus tikėtinumo įverčius minėtiems modeliams galima parinkti tinkamiausią modelį (Puasono-Gauso, Puasono-gama, arba Puasono-Gauso regresinį) ir pateikti rekomendacijas naudoti konkretų modelį vertinant empiriniu Bajeso metodu.

5. REZULTATAI IR IŠVADOS

Sprendžiant darbe suformuluotus uždavinius, gauti šie nauji rezultatai:

1. Sudarytas didelio matavimo duomenų binarinio skaidymo metodas, paremtas erdvės skaidymu, naudojamu duomenų klasifikavime, įgalinantis efektyviai atlikti didelio matavimo duomenų skaidymą.

2. Sukurtas naujas metodas tikrinti didelio matavimo nekoreliuotų duomenų pasirinktų komponenčių nepriklausomumą, naudojant didelio matavimo duomenų binarinio skaidymo metodą.

3. Naudojant skirtingus matematinius modelius sudarytas naujas metodas parenkant didelių populiacijų retų įvykių optimalų empirinio Bajeso modelį ir atitinkamą empirinį Bajeso įvertį.

Darbe gauti šie praktiniai rezultatai:

1. Sudarytas ir ištirtas didelio matavimo duomenų skaidymo algoritmas, leidžiantis pagreitinti klasifikavimo algoritmus, nesumažinus klasifikavimo tikslumo.

2. Sudarytas ir ištirtas didelio matavimo duomenų komponenčių nepriklausomumo hipotezės tikrinimo algoritmas.

3. Sudarytas ir ištirtas retų dažnių tikimybių vertinimo empirinio Bajeso algoritmas, parenkant optimalų didelių populiacijų retų įvykių modelį, pritaikytas Lietuvos gyventojų socialinių-medicininių rodiklių analizei.

Gauti rezultatai ir atlikti tyrimai leidžia padaryti šias išvadas:

1. Darbe sudaryta binarinė skaidymo procedūra labiausiai tinka duomenims su išreikšta klasterine struktūra. Atlikus palyginti nedidelį žingsnių skaičių galima žymiai sumažinti pradinę normuotos vidutinės paklaidos reikšmę iki tokio lygio, kad, viena vertus, gauname gerokai trumpesnę grupuotą duomenų seką, antra vertus, ši normuotos vidutinės paklaidos reikšmė leidžia pakankamai tiksliai atlikti skaičiavimus su grupuota seka (vietoje pradinės sekos), taip sumažinant skaičiavimų laiką.

2. Nagrinėjant modelio adekvatumo testavimo algoritmas didelio matavimo duomenims, rezultatai parodė, kad yra labai silpna priklausomybė nuo parinkto mišinio modelio ir erdvės matavimo. Testinės statistikos, atmetus 5 proc.

didžiausių ir 5 proc. mažiausių reikšmių, maksimumas yra tinkamas kriterijus priimti ar atmesti nagrinėjamą hipotezę.

3. Nagrinėjant didelio matavimo duomenų komponenčių nepriklausomumo testavimo algoritmą, Monte-Carlo modeliavimų rezultatai rodo, kad pasiūlyta procedūra yra tinkama didelio matavimo duomenims. Ši procedūra pradeda lenkti klasikinį Blum-Kiefer-Roselblatt testą palyginti nedidelio matavimo duomenims. Kritinė reikšmė c_α mažai priklauso nuo matavimo d ir padalijimo procedūros ir gali būti dar sumažinta pridėjus papildomus reikalavimus padalijimo procedūrai.

4. Taikant empirinį Bajeso metodą testuojant didelio matavimo duomenų komponenčių nepriklausomumą, pradinė didelio matavimo duomenų testavimo problema suvedama į papildomą testavimo problemą. Nulinė hipotezė H_0^η gali būti performuluota kaip $G = \delta_0$, kur G yra apriorinis nežinomų parametrų θ_i, $i = 1, \ldots, n$, skirstinys, o δ_0 yra išsigimęs skirstinys taške 0. Todėl bet kuris nukrypimo matas tarp δ_0 ir skirstinio G didžiausio tikėtinumo įverčio $\hat{G}_{ML}$ gali būti naudojamas testavimui, pvz., χ^2 testas arba neparametrinis tikėtinumo santykio kriterijus. Modeliavimo būdu buvo nagrinėjamos baigtinių imčių savybės testui, paremtam neparametrine empirine Bajeso statistika $\hat{\mu}_1^2$.

Modeliavimo rezultatai rodo neparametrinio empirinio Bajeso testo privalumus, palyginus su χ^2 testu. Kadangi didžiausio tikėtinumo įverčio $\hat{G}_{ML}$ skaičiavimas yra iteracinis ir užima daug laiko, rezultatai gali priklausyti nuo skaičiavimo metodo iteracijų skaičiaus.

5. Suformuluota Puasono-Gauso modelio, naudojant empirinį Bajeso metodą, nesinguliarumo sąlyga ir pateiktas „paprastų iteracijų“ iteracinis metodas įverčių skaičiavimui. Pateiktas metodas buvo pritaikytas socialinių ir medicininių duomenų analizei, parodant jo paprastumą ir pritaikomumą.

6. Nustatyta Puasono-gama modelio singuliarumo sąlyga ir sudarytas skaitinis algoritmas empirinio Bajeso įverčiams gauti.

7. Nagrinėjant Puasono-gama ir Puasono-Gauso modelius empiriniame Bajeso mažų tikimybių vertinime, panaudojus Monte-Carlo modeliavimą,

keičiant pasirinktus parametrus ir skaičiuojant maksimalaus tikėtinumo įverčius (atskirai imant šiuos modelius, taip pat pasiūlytą Puasono-Gauso modelį su papildomu regresijos kintamuoju) galima parinkti tinkamiausią modelį ir tokiu būdu pateikti rekomendacijas parinkti tinkamiausią modelį Bajeso vertinimui.

LITERATŪRA

1. Abramovich M., Stegun I. A. (1968). *Handbook of Mathematical Functions*, Dover, New York.
2. Aivazyan S. A. (1996) Mixture approach to clustering via maximum likelihood, criteria of model complexity and projection pursuit. In *Data Science, Classification and Related Methods*, Abstracts of 5th IFCS Conference. IFCS, Cobe, **1**, 36.
3. Aivazyan S. A., Enyukov I. S., Meshalkin L. D. (1983) Applied Statistics: Principles of Modeling and Primary Data Processing, A Handbook (in Russian) (rusų k. – Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных, М., Финансы и статистика, 1983, 471 с.).
4. Aivazyan S. A., Enyukov I. S., Meshalkin L. D. (1989) Applied Statistics: Classification and Reduction of Dimensionality (in Russian).
5. Andrieu C., de Freitas N., Doucet A., Jordan M. I. (2003) An Introduction to MCMC for Machine Learning, Machine Learning, Vol. 50, No. 1, pp. 5–43.
6. Asmussen S., Glynn, P. W. (2007) Stochastic Simulation: Algorithms and Analysis, Springer. Series: Stochastic Modelling and Applied Probability, Vol. 57.
7. Baringhaus L., Henze N. (1988) A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, **35**:339–348.
8. Behboodian J. (1970). On a mixture of normal distributions. *Biometrika*, **57**, 215–217.
9. Berg B. A. (2004) Markov Chain Monte Carlo Simulations And Their Statistical Analysis, ISBN: 978-981-238-935-0, 380p, World Scientific Publishing Co Pte Ltd.

10. Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution free tests for independence based on the sample distribution function. Annals of Mathematical Statistics, 35, 138-149.
11. Bousquet O., Boucheron S., Lugosi G. (2004) Introduction to statistical learning theory. In O. Bousquet, U. von Luxburg and G. Ratsch(Eds.), *Advanced Lectures on Machine Learning, 2004*, Lecture Notes in Artificial Intelligence, vol. 3176, pp. 169–207. Springer.
12. Bowman A. W., Foster P. J. (1993) Adaptive smoothing and density based tests of multivariate normality. *J. Amer. Statist. Assoc.*, **88**:529–537.
13. Bradley P. C., Thomas A. L., Bradley C. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC.
14. Carlin B. P., Louis T. A. (1996) Bayes and Empirical Bayes Methods for Data Analysis, New York, Chapman and Hall (CRC Press).
15. Clayton D., Kaldor J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, 43, 671–681.
16. van de Geer S. (2003) Asymptotic theory for maximum likelihood in nonparametric mixture models. *Computational Statistics and Data Analysis*, **41**:453–464.
17. DeGroot M. H. (1970). Optimal statistical Decisions, ISBN 0-471-68029-X, McGraw-Hill Company.
18. Dennis J. E., Schnabel R. B. (1996) *Numerical Methods for unconstrained optimization and nonlinear Equations*. In: Classics for Applied Mathematics, 16, SIAM.
19. Diaconis P. (2009) The Markov chain Monte Carlo revolution, 2009, Bull. Amer. Math. Soc. Vol. 46, No. 2, April 2009, pp 179–205. S 0273-0979(08)01238-X.
20. Diuk V.A., Samoilenko A.P. (2002). *Data Mining. Learning Course*, ISBN 5-318-00227-7, Piter, St. Petersburg, 368p. (in Russian).

21. DuMouchel, W.; Volinsky, C.; Johnson, T.; Cortes, C.; Pregibon, D. (1999). Squashing flat files flatter. In KDD '99 Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ISBN:1-58113-143-7. p. 6–15.
22. DuMouchel, W. (2002). Data Squashing: Constructing summary data sets. In Handbook of Massive Data Sets. ISBN 1-4020-0489-3. Kluwer Academic Publishers, Printed in the Netherlands. p. 579–592.
23. Everitt B.S., Hand D.J. (1981). *Finite Mixture Distributions*. Chapman and Hall, London.
24. Genest, C., and Remillard, B. (2004). Tests of independence and randomness based on the empirical copula process. Test, 13, 335-370.
25. Gurevičius R., Jakimauskas G, Sakalauskas L. (2009) Empirical Bayesian estimation of small mortality rates, 5th international Vilnius conference [and] EURO-mini conference „Knowledge-based technologies and OR methodologies for decisions of sustainable development (KORSD-2009), Vilnius, Technika, pp. 290-295
26. Friedman J. H. (1987) Exploratory projection pursuit. *J. Amer. Statist. Assoc.*, **82**, 249-266.
27. Friedman J. H., Tukey J. W. (1974) A projection pursuit algorithm for exploratory data analysis. IEEE Trans. Comput., **C-21**, 881-889
28. Hasselblad V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics*, **8**, 431–444.
29. Hastie, T., Tibshirani, R., Friedman, J. H. (2001). The elements of Statistical Learning. New York, Berlin: Springer (paskutinė versija (2013), žr. http://www- stat.stanford.edu/~tibs/ElemStatLearn/ printings/ESLII_print10.pdf).
30. Hirukawa J. (2012) Large-Deviation Results for Discriminant Statistics of Gaussian Locally Stationary Processes, Advances in Decision Sciences, Vol. 2012, Faculty of science, Niigata University.
31. Huang L.-S. (1997) Testing goodness-of-fit based on a roughness measure. *J. Amer. Statist. Assoc.*, **92**:1399–1402.

32. Hyvarinen, A., Karhunen, J., and Oja, E. (2001). Independent Component Analysis. New York: John Wiley and Sons.
33. Jakimauskas G. (1997) Efficiency analysis of one estimation and clusterization procedure of one-dimensional Gaussian mixture. *Informatica*, **8**(3), p. 331-343.
34. Jakimauskas G. (2002) Procedure of the removal of the outliers from the sample satisfying the multidimensional Gaussian mixture model, *Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai*, ISBN 9986-680-16-6, **41**, p. 523-528
35. Jakimauskas, G. (2009) Efficient algorithm for testing goodness-of-fit for classification of high dimensional data, *Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai*, ISSN 0132-2818. **50**, p. 293-297.
36. Jakimauskas G. (2012) Gamma and logit models in empirical Bayesian estimation of probabilities or rare events. Proceedings of the Workshop STOPROG-2102 (Stochastic Programming for Implementation and Advanced Applications), July 3-6, 2012, Neringa, Lithuania, pp. 43–48.
37. Jakimauskas G., Krikštolaitis R. (2000a) Influence of projection pursuit on classification errors: computer simulation results. *Informatica*, ISSN 0868-4952, **11**(2), p. 115-124
38. Jakimauskas G., Krikštolaitis R. (2000b) Bootstrap methods in selection of the discriminant subspace. *Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai*, **40**, p. 281-286
39. Jakimauskas G., Radavičius M., Sušinskas J. (2008) A simple method for testing independence of high-dimensional random vectors. *Austrian J. Statist.*, **44**:101–108.
40. Jakimauskas G., Sakalauskas L. (2010) Empirical Bayesian estimation for Poisson-gamma model. Proceedings of the 24th Mini EURO conference on continuous optimization and information-based technologies in the financial sector (MEC EurOPT 2010). Vilnius: Technika, p. 254-257.

41. Jakimauskas G., Sakalauskas L. (2012) Empirical Bayesian regression model for estimation of small rates. *Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai*, ser. A, **53**, p. 42-47
42. Jakimauskas G., Sušinskas J. (2010) Application of the empirical Bayes approach to nonparametric testing for high-dimensional data. *Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai*, **51**, p. 402-407
43. Jiang W.; Zhang C.-H. (2009) General maximum likelihood empirical bayes estimation of normal means. *Ann. Statist.*, **37**:1647–1684, 2009.
44. Kantorovich, L. V., Akilov, G. P. (1982) *Functional Analysis*, Pergamon Press, Oxford.
45. Knorr-Held L., Rasser G. (1999). *Bayesian detection of clusters and discontinuities in disease maping*. Sonderforschungsbereich 386, Discussion Paper 107.
46. Leite J. G., Rodrigues J., Milan L. A. (2000) *A Bayesian analysis for estimating the number of species in a population using nonhomogeneous Poisson process*. Statistics & Probability Letters, vol. 48, pp. 153-161.
47. Madigan, D., Raghavan, N., DuMouchel, W., Nason, M., Posse, C., Ridgeway, G. (2002). Likelihood-based data squashing: A modeling approach to instance construction. Journal of Data Mining and Knowledge Discovery, **6**, p. 173-190.
48. Marcoulides G. A., Hershberger S. L. (1997) Multivariate Statistical Methods: A First Course. Lawrence Erlbaum Associates, ISBN-080582572X, Mahwah, New Jersey.
49. Meza J. L. (2003) Empirical Bayes estimation smoothing of relative risks in disease mapping, *Journal of Statistical Planning and Inference*, 112:43–62.
50. Nisbet, R., Elder, J., Miner, G. (2009); *Handbook of Statistical Analysis & Data Mining Applications*, Academic Press/Elsevier, ISBN 978-0-12-374765-5.

51. Polonik, W. (1999). Concentration and goodness-of-fit in higher dimensions: (asymptotically) distribrution free methods. Annals of Statistics, 27, 1210-1229.
52. Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P. (2007) Numerical Recipes 3rd Edition: The Art of Scientific Computing, New York: Cambridge University Press, ISBN 978-0-521-88068-8.
53. Quigley J., Bedford T., Walls L. (2007) *Estimating rate of occurrence of rare events with Empirical Bayes: A railway application. Reliability Engineering and System Safety.* 92, pp. 619–627
54. Richey M. (2010) The Evolution of Markov Chain Monte Carlo Methods, The American Mathematical Monthly, May 2010, 383-413.
55. Rossi P. E., Allenby G. M. (2003) Bayesian Statistics and Marketing, Marketing Science, 22 (3), 304–328.
56. Rudzkis R., Radavičius M. (1995). Statistical Estimation of a Mixture of Gaussian Distributions. *Acta Applicandae Mathematicae*, **38**, 37–54.
57. Rudzkis R., Radavičius M. (1997). Projection pursuit in Gaussian mixture models preserving information about cluster structure. *Lithuanian Math. J.*, **37**, No. 4, 416–425.
58. Rudzkis R., Radavičius M. (1999). Characterization and Statistical Estimation of a Discriminant Space for Gaussian Mixtures. *Acta Applicandae Mathematicae*, **58**, 279–290.
59. Sakalauskas L. (1995) On Bayes analysis of small rates in medicine, Proc. of the Internat. Conf. "Computer Data Analysis and Modeling", 1995, September 14-19, Minsk, vol. 1, pp. 127-130.
60. Sakalauskas L., Vaičiulytė I. (2012) Daugiamatis mažų dažnių vertinimo algoritmas, Lietuvos matematikos rinkinys. LMD darbai. Vilnius: Matematikos ir informatikos institutas. t. 53, ser. B, p. 260-263.
61. Szekely G. J., Rizzo M. L. (2005) A new test for multivariate normality. *J. Multiv. Anal.*, **93**:58–80.
62. Szekely, G. J., and Rizzo, M. L. (2006). Testing for equal distributions in high dimension.

63. Tan, P.-N., Steinbach, M., Kumar, V. (2005); *Introduction to Data Mining*, ISBN 0-321-32136-7
64. Tsutakava R. K., Shoop G. L., Marienfield C. J. (1985) *Empirical Bayes estimation of cancer mortality rates*. Statistics in medicine, No 4, p.p. 201-212.
65. Vaičiulytė I., Sakalauskas L. (2011) Daugiamačio antisimetrinio t-skirstinio parametrinis įvertinimas, Jaunųjų mokslininkų darbai, Šiaulių universitetas. Šiauliai: Šiaulių universiteto leidykla. 2011, nr. 4, p. 157-163.
66. Vapnik V. N. (1998) *Statistical Learning Theory*. Wiley, New York.
67. Vapnik, V. N., and Chervonenkis, A. (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. Theory Probabability and its Applications, 26, 821-832.
68. Vaurioa L. K., Jankala K. E. (2006). *Evaluation and comparison of estimation methods for failure rates and probabilities*. Reliability Engineering and System Safety, 91, pp. 209–221
69. Verdinelli I., Wasserman L. (1995) Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.*, **26**(4):1215–1241.
70. Yasui Y., Liu H., Benach J., Winget M. (2000) *An empirical evaluation of various priors in the empirical Bayes estimation of small area disease risks*. Statistics in Medicine, vol. 19, pp. 2409-2420
71. Ye, N. (2003); *The Handbook of Data Mining*, Mahwah, NJ: Lawrence Erlbaum.
72. Zhou, J.; Sander, J. (2003). Data Bubbles for Non-Vector Data: Speeding-up Hierarchical Clustering in Arbitrary Metric Spaces. Proceedings of the 29th international conference on Very large data bases - Volume 29. p. 452–463.
73. Zhu L.-X., Fang K. T. and Bhatti M. I. (1997) On estimated projection pursuit-type Crameer–von Mises statistics. *J. Multiv. Anal.*, **63**:1–14.

74. Zhu L.-X., Neuhaus G. (2000) Nonparametric Monte Carlo tests for multivariate distributions *Biometrika*, **87**:919–928.

PRIEDAI

PRIEDAS 1. NAUDOJAMŲ STATISTINIŲ DUOMENŲ SĄRAŠAS

Skyrelyje 4.1 buvo naudojami 2003–2004 metų nužudymų ir savižudybių mirtingumo Lietuvoje duomenys (žr. Lent. 1 - Lent. 4, visi įvykiai populiacijoje, vyrai ir moterys atskirai, pagal administracinę teritoriją. Duomenis pateikė Lietuvos Higienos institutas.

Skyreliuose 4.2 ir 4.3 modeliavimui buvo naudojami realūs duomenys iš Lietuvos Statistikos departamento duomenų bazės – darbingo amžiaus asmenys, pirmą kartą pripažinti neįgaliaisiais pagal administracinę teritoriją (Lentelė M3140706), 2010 metai (teritorijų skaičius $K = 60$), iš viso 9 duomenų rinkiniai (žr. Lent. 5, čia Lietuvos Respublikos ir 10 apskričių duomenys pateikti tik aiškumo dėlei, jie skaičiavimuose nenaudojami):

1 – Iš viso (15432 atvejai) (Iš viso)

2 – Tuberkuliozė (399) (A15A19)

3 – Piktybiniai navikai (2706) (C00C97)

4 – Psichikos ir elgesio sutrikimai (1303) (F00F99)

5 – Nervų sistemos ligos (1501) (G00G99)

6 – Kraujotakos sistemos ligos (3525) (I00I99)

7 – Jungiamojo audinio ir skeleto-raumenų sistemos ligos (2566) (M00M99)

8 – Traumos, apsinuodijimai (1116) (S00T98)

9 – Kitos priežastys (2316) (Kita)

Taip pat buvo naudojamas vidutinis metinis gyventojų skaičius pagal administracinę teritoriją (Lentelė M3010211), iš viso gyventojų skaičius visose administracinėse teritorijose lygus 3286820. Skyrelyje 4.3 buvo naudojamas 2010 metais ligoninėse gydytų ligonių skaičius pagal administracinę teritoriją (Lentelė M3140312), (žr. Lent. 6.).

	Gyv.sk.	Suic	Hom.
Lietuvos Respublika	**3442501**	**1434**	**2010**
Alytaus apskritis	**185574**	**89**	**98**
Alytaus m. sav.	32132	13	5
Alytaus r. sav.	71100	22	6
Druskininkų sav.	25116	10	10
Lazdijų r. sav.	26647	16	26
Varėnos r. sav.	30579	28	51
Kauno apskritis	**693794**	**254**	**200**
Birštono sav.	5383	5	8
Jonavos r. sav.	52282	30	11
Kaišiadorių r. sav.	37338	16	60
Kauno m. sav.	371292	102	16
Kauno r. sav.	83420	25	15
Kėdainių r. sav.	65273	33	18
Prienų r. sav.	35190	17	35
Raseinių r. sav.	43616	26	37
Klaipėdos apskritis	**383597**	**135**	**196**
Klaipėdos m. sav.	190906	45	20
Klaipėdos r. sav.	46708	17	22
Kretingos r. sav.	45953	26	24
Neringos sav.	2518	1	21
Palangos m. sav.	17606	8	23
Skuodo r. sav.	25245	12	44
Šilutės r. sav.	54661	26	42
Marijampolės apskritis	**187172**	**93**	**160**
Kalvarijos sav.	13710	10	14
Kazlų Rūdos sav.	14856	3	17
Marijampolės sav.	70414	33	27
Šakių r. sav.	38289	22	39
Vilkaviškio r. sav.	49903	25	63
Panevėžio apskritis	**296341**	**141**	**199**
Biržų r. sav.	34949	18	9
Kupiškio r. sav.	24302	11	25
Panevėžio m. sav.	118208	38	62
Panevėžio r. sav.	42987	28	32
Pasvalio r. sav.	34442	18	33
Rokiškio r. sav.	41453	28	38

(tęsinys kitame puslapyje)

Lentelė 1. Gyventojų skaičius, bendras savižudybių (Suic.) ir nužudymų (Hom.) skaičius pagal administracinę teritoriją, 2003 metai.

(tęsinys)

	Gyv.sk.	Suic	Hom.
Šiaulių apskritis	**365621**	**178**	**211**
Akmenės r. sav.	29698	12	4
Joniškio r. sav.	31561	16	12
Kelmės r. sav.	40353	22	19
Pakruojo r. sav.	29069	14	31
Radviliškio r. sav.	51482	33	36
Šiaulių m. sav.	131948	50	54
Šiaulių r. sav.	51510	31	55
Tauragės apskritis	**133101**	**73**	**130**
Jurbarko r. sav.	37199	24	13
Pagėgių sav.	12134	11	30
Šilalės r. sav.	31340	18	41
Tauragės r. sav.	52428	20	46
Telšių apskritis	**178639**	**106**	**165**
Mažeikių r. sav.	66935	41	28
Plungės r. sav.	44130	22	34
Rietavo sav.	10569	14	56
Telšių r. sav.	57005	29	47
Utenos apskritis	**182104**	**98**	**260**
Anykščių r. sav.	34268	20	7
Ignalinos r. sav.	22357	15	59
Molėtų r. sav.	24774	19	29
Utenos r. sav.	49604	32	50
Visagino sav.	28710	6	58
Zarasų r. sav.	22391	6	57
Vilniaus apskritis	**836558**	**267**	**391**
Elektrėnų sav.	28623	18	61
Šalčininkų r. sav.	38752	17	40
Širvintų r. sav.	19898	15	43
Švenčionių r. sav.	32528	9	45
Trakų r. sav.	37127	18	48
Ukmergės r. sav.	47854	27	49
Vilniaus m. sav.	541330	137	53
Vilniaus r. sav.	90446	26	52

Lentelė 1. Gyventojų skaičius, savižudybių (Suic.) ir nužudymų (Hom.) skaičius pagal administracinę teritoriją, 2003 metai (lentelės pabaiga).

	Vyrai	Suic.	Hom.	Moterys	Suic.	Hom.
Lietuvos Respublika	**1607616**	**1199**	**232**	**1834903**	**256**	**100**
Alytaus apskritis	**88768**	**76**	**10**	**96806**	**13**	**7**
Alytaus m. sav.	15741	12	1	16391	1	1
Alytaus r. sav.	34085	17	3	37015	5	3
Druskininkų sav.	11543	8	1	13573	2	0
Lazdijų r. sav.	12788	14	4	13859	2	0
Varėnos r. sav.	14611	25	1	15968	3	3
Kauno apskritis	**320119**	**203**	**50**	**373675**	**51**	**20**
Birštono sav.	2467	4	0	2916	1	0
Jonavos r. sav.	24587	16	6	27695	14	3
Kaišiadorių r. sav.	18366	13	1	18972	3	0
Kauno m. sav.	167308	85	27	203984	17	10
Kauno r. sav.	39525	18	7	43895	7	4
Kėdainių r. sav.	30590	31	5	34683	2	1
Prienų r. sav.	16714	15	0	18476	2	2
Raseinių r. sav.	20562	21	4	23054	5	0
Klaipėdos apskritis	**180312**	**130**	**22**	**203285**	**22**	**7**
Klaipėdos m. sav.	88308	45	16	102598	8	4
Klaipėdos r. sav.	22598	17	3	24110	5	1
Kretingos r. sav.	21776	26	1	24177	3	0
Neringos sav.	1212	1	0	1306	0	0
Palangos m. sav.	8062	8	0	9544	1	0
Skuodo r. sav.	12099	11	1	13146	1	1
Šilutės r. sav.	26257	22	1	28404	4	1
Marijampolės apskritis	**89186**	**79**	**13**	**97986**	**14**	**3**
Kalvarijos sav.	6556	9	2	7154	1	0
Kazlų Rūdos sav.	7037	2	1	7819	1	0
Marijampolės sav.	33536	30	6	36878	3	0
Šakių r. sav.	18271	17	1	20018	5	2
Vilkaviškio r. sav.	23786	21	3	26117	4	1
Panevėžio apskritis	**138135**	**107**	**27**	**158206**	**38**	**9**
Biržų r. sav.	16462	12	5	18487	6	0
Kupiškio r. sav.	11366	11	1	12936	4	2
Panevėžio m. sav.	53913	26	9	64295	12	3
Panevėžio r. sav.	20587	24	1	22400	4	0
Pasvalio r. sav.	16365	13	4	18077	5	0
Rokiškio r. sav.	19442	21	7	22011	7	4

(tęsinys kitame puslapyje)

Lentelė 2. Gyventojų skaičius, savižudybių (Suic.) ir nužudymų (Hom.) skaičius pagal lytį ir administracinę teritoriją, 2003 metai.

(tęsinys)

	Vyrai	Suic.	Hom.	Moterys	Suic.	Hom.
Šiaulių apskritis	**171177**	**144**	**29**	**194444**	**34**	**12**
Akmenės r. sav.	13920	10	3	15778	2	1
Joniškio r. sav.	14815	15	2	16746	1	1
Kelmės r. sav.	19334	20	2	21019	2	3
Pakruojo r. sav.	13855	11	2	15214	3	0
Radviliškio r. sav.	24469	25	3	27013	8	2
Šiaulių m. sav.	60221	43	14	71727	7	4
Šiaulių r. sav.	24563	20	3	26947	11	1
Tauragės apskritis	**63184**	**64**	**8**	**69917**	**9**	**4**
Jurbarko r. sav.	17627	20	2	19572	4	0
Pagėgių sav.	5770	10	1	6364	1	2
Šilalės r. sav.	15183	15	2	16157	3	0
Tauragės r. sav.	24604	19	3	27824	1	2
Telšių apskritis	**84460**	**90**	**17**	**94179**	**16**	**5**
Mažeikių r. sav.	31594	37	6	35341	4	1
Plungės r. sav.	20967	18	3	23163	4	0
Rietavo sav.	5044	13	1	5525	1	0
Telšių r. sav.	26855	22	7	30150	7	4
Utenos apskritis	**85932**	**87**	**9**	**96190**	**11**	**7**
Anykščių r. sav.	16099	19	4	18187	1	1
Ignalinos r. sav.	10578	14	0	11779	1	1
Molėtų r. sav.	11899	15	2	12875	4	1
Utenos r. sav.	23205	29	0	26399	3	2
Visagino sav.	13653	4	2	15057	2	1
Zarasų r. sav.	10498	6	1	11893	0	1
Vilniaus apskritis	**386343**	**219**	**47**	**450215**	**48**	**26**
Elektrėnų sav.	13644	17	4	14979	1	1
Šalčininkų r. sav.	18534	15	3	20218	2	3
Širvintų r. sav.	9392	10	3	10506	5	0
Švenčionių r. sav.	15174	7	1	17354	2	4
Trakų r. sav.	17400	14	3	19727	4	0
Ukmergės r. sav.	22335	23	3	25519	4	0
Vilniaus m. sav.	246412	110	22	294918	27	15
Vilniaus r. sav.	43452	23	8	46994	3	3

Lentelė. 2. Gyventojų skaičius, savižudybių (Suic.) ir nužudymų (Hom.) skaičius pagal lytį ir administracinę teritoriją, 2003 metai (lentelės pabaiga).

	Gyv.sk.	Suic	Hom.
Lietuvos Respublika	**3423841**	**1381**	**294**
Alytaus apskritis	183829	101	12
Alytaus m. sav.	32127	22	3
Alytaus r. sav.	70288	25	4
Druskininkų sav.	24952	17	1
Lazdijų r. sav.	26295	18	1
Varėnos r. sav.	30167	19	3
Kauno apskritis	**688584**	**250**	**60**
Birštono sav.	5343	4	1
Jonavos r. sav.	52346	17	6
Kaišiadorių r. sav.	37048	22	0
Kauno m. sav.	366486	93	28
Kauno r. sav.	84336	40	7
Kėdainių r. sav.	64853	34	9
Prienų r. sav.	34901	20	2
Raseinių r. sav.	43271	20	7
Klaipėdos apskritis	**382714**	**128**	**33**
Klaipėdos m. sav.	189477	41	17
Klaipėdos r. sav.	47420	19	1
Kretingos r. sav.	46064	23	5
Neringos sav.	2731	0	0
Palangos m. sav.	17607	3	3
Skuodo r. sav.	25025	9	1
Šilutės r. sav.	54390	33	6
Marijampolės apskritis	**186078**	**87**	**12**
Kalvarijos sav.	13644	12	0
Kazlų Rūdos sav.	14875	6	1
Marijampolės sav.	70120	28	6
Šakių r. sav.	37907	15	3
Vilkaviškio r. sav.	49532	26	2
Panevėžio apskritis	**293769**	**124**	**25**
Biržų r. sav.	34627	17	5
Kupiškio r. sav.	24038	10	1
Panevėžio m. sav.	116920	43	9
Panevėžio r. sav.	43147	18	5
Pasvalio r. sav.	34087	14	1
Rokiškio r. sav.	40950	22	4

(tęsinys kitame puslapyje)

Lentelė. 3. Gyventojų skaičius, savižudybių (Suic.) ir nužudymų (Hom.) skaičius pagal administracinę teritoriją, 2004 metai.

(tęsinys)

	Gyv.sk.	Suic	Hom.
Šiaulių apskritis	**362415**	**184**	**32**
Akmenės r. sav.	29288	25	4
Joniškio r. sav.	31334	20	6
Kelmės r. sav.	39934	29	1
Pakruojo r. sav.	28819	23	5
Radviliškio r. sav.	51041	23	4
Šiaulių m. sav.	130600	41	9
Šiaulių r. sav.	51399	23	3
Tauragės apskritis	**132105**	**69**	**11**
Jurbarko r. sav.	36828	18	8
Pagėgių sav.	12008	10	0
Šilalės r. sav.	31165	17	2
Tauragės r. sav.	52104	24	1
Telšių apskritis	**177575**	**76**	**26**
Mažeikių r. sav.	66571	26	10
Plungės r. sav.	44030	24	6
Rietavo sav.	10498	5	0
Telšių r. sav.	56476	21	10
Utenos apskritis	**180045**	**92**	**9**
Anykščių r. sav.	33873	28	3
Ignalinos r. sav.	21803	8	1
Molėtų r. sav.	24405	15	0
Utenos r. sav.	49208	22	2
Visagino sav.	28767	6	2
Zarasų r. sav.	21989	13	1
Vilniaus apskritis	**836727**	**270**	**74**
Elektrėnų sav.	28468	12	1
Šalčininkų r. sav.	38546	22	5
Širvintų r. sav.	19782	4	3
Švenčionių r. sav.	32176	15	6
Trakų r. sav.	37063	25	5
Ukmergės r. sav.	47443	29	6
Vilniaus m. sav.	541180	123	40
Vilniaus r. sav.	92069	40	8

Lentelė. 3. Gyventojų skaičius, savižudybių (Suic.) ir nužudymų (Hom.) skaičius pagal administracinę teritoriją, 2004 metai (lentelės pabaiga).

	Vyrai	Suic.	Hom.	Moterys	Suic.	Hom.
Lietuvos Respublika	**1598009**	**1124**	**200**	**1825912**	**257**	**93**
Alytaus apskritis	87816	84	8	96013	17	4
Alytaus m. sav.	15734	18	2	16393	4	1
Alytaus r. sav.	33650	19	3	36638	6	1
Druskininkų sav.	11474	14	0	13478	3	1
Lazdijų r. sav.	12595	16	0	13700	2	1
Varėnos r. sav.	14363	17	3	15804	2	0
Kauno apskritis	**317496**	**211**	**38**	**371088**	**39**	**22**
Birštono sav.	2442	4	0	2901	0	1
Jonavos r. sav.	24635	15	4	27711	2	2
Kaišiadorių r. sav.	18188	20	0	18860	2	0
Kauno m. sav.	164889	72	17	201597	21	11
Kauno r. sav.	40001	37	4	44335	3	3
Kėdainių r. sav.	30375	26	8	34478	8	1
Prienų r. sav.	16574	18	0	18327	2	2
Raseinių r. sav.	20392	19	5	22879	1	2
Klaipėdos apskritis	**179716**	**105**	**27**	**202998**	**23**	**6**
Klaipėdos m. sav.	87483	33	15	101994	8	2
Klaipėdos r. sav.	22963	15	1	24457	4	0
Kretingos r. sav.	21789	19	3	24275	4	2
Neringos sav.	1350	0	0	1381	0	0
Palangos m. sav.	8043	2	3	9564	1	0
Skuodo r. sav.	11989	8	0	13036	1	1
Šilutės r. sav.	26099	28	5	28291	5	1
Marijampolės apskritis	**88556**	**77**	**10**	**97522**	**10**	**2**
Kalvarijos sav.	6524	12	0	7120	0	0
Kazlų Rūdos sav.	7050	5	1	7825	1	0
Marijampolės sav.	33308	24	6	36812	4	0
Šakių r. sav.	18083	13	2	19824	2	1
Vilkaviškio r. sav.	23591	23	1	25941	3	1
Panevėžio apskritis	**137017**	**98**	**20**	**156752**	**26**	**5**
Biržų r. sav.	16340	13	5	18287	4	0
Kupiškio r. sav.	11253	9	1	12785	1	0
Panevėžio m. sav.	53329	28	7	63591	15	2
Panevėžio r. sav.	20685	13	4	22462	5	1
Pasvalio r. sav.	16199	14	1	17888	0	0
Rokiškio r. sav.	19211	21	2	21739	1	2

(tęsinys kitame puslapyje)

Lentelė. 4. Gyventojų skaičius, savižudybių (Suic.) ir nužudymų (Hom.) skaičius pagal lytį ir administracinę teritoriją, 2004 metai.

(tęsinys)

	Vyrai	Suic.	Hom.	Moterys	Suic.	Hom.
Šiaulių apskritis	**169587**	**141**	**20**	**192828**	**43**	**11**
Akmenės r. sav.	13702	18	4	15586	7	0
Joniškio r. sav.	14700	18	3	16634	2	3
Kelmės r. sav.	19127	24	1	20807	5	0
Pakruojo r. sav.	13733	19	2	15086	4	3
Radviliškio r. sav.	24262	15	1	26779	8	3
Šiaulių m. sav.	59551	29	7	71049	12	2
Šiaulių r. sav.	24512	18	2	26887	5	0
Tauragės apskritis	**62716**	**57**	**6**	**69389**	**12**	**5**
Jurbarko r. sav.	17447	16	5	19381	2	3
Pagėgių sav.	5712	7	0	6296	3	0
Šilalės r. sav.	15090	13	0	16075	4	2
Tauragės r. sav.	24467	21	1	27637	3	0
Telšių apskritis	**83878**	**59**	**19**	**93697**	**17**	**7**
Mažeikių r. sav.	31362	16	8	35209	10	2
Plungės r. sav.	20930	22	4	23100	2	2
Rietavo sav.	5009	4	0	5489	1	0
Telšių r. sav.	26577	17	7	29899	4	3
Utenos apskritis	**84978**	**76**	**5**	**95067**	**16**	**4**
Anykščių r. sav.	15888	25	2	17985	3	1
Ignalinos r. sav.	10324	5	0	11479	3	1
Molėtų r. sav.	11715	12	0	12690	3	0
Utenos r. sav.	22996	20	1	26212	2	1
Visagino sav.	13721	5	1	15046	1	1
Zarasų r. sav.	10334	9	1	11655	4	0
Vilniaus apskritis	**386249**	**216**	**47**	**450558**	**54**	**27**
Elektrėnų sav.	13557	10	1	14991	2	0
Šalčininkų r. sav.	18421	20	3	20125	2	2
Širvintų r. sav.	9325	4	3	10457	0	0
Švenčionių r. sav.	15025	12	5	17151	3	1
Trakų r. sav.	17350	21	4	19713	4	1
Ukmergės r. sav.	22139	25	3	25304	4	3
Vilniaus m. sav.	246191	96	21	294989	27	19
Vilniaus r. sav.	44241	28	7	47828	12	1

Lentelė. 4. Gyventojų skaičius, savižudybių (Suic.) ir nužudymų (Hom.) skaičius pagal lytį ir administracinę teritoriją, 2004 metai (lentelės pabaiga).

	Iš viso	A15 A19	C00 C97	F00 F99	G00 G99	I00 I99	M00 M99	S00 T98	Kita
Lietuvos Respublika	**15432**	**399**	**2706**	**1303**	**1501**	**3525**	**2566**	**1116**	**2316**
Alytaus apskritis	**769**	**30**	**130**	**73**	**92**	**164**	**136**	**63**	**81**
Alytaus m. sav.	306	10	57	28	22	66	65	27	31
Alytaus r. sav.	139	4	17	18	13	27	32	10	18
Druskininkų sav.	88	1	17	7	21	18	11	8	5
Lazdijų r. sav.	110	7	21	11	21	12	21	11	6
Varėnos r. sav.	126	8	18	9	15	41	7	7	21
Kauno apskritis	**3295**	**96**	**523**	**261**	**328**	**869**	**422**	**265**	**531**
Birštono sav.	32	1	4	2	6	9	5	0	5
Jonavos r. sav.	210	14	37	12	12	66	26	11	32
Kaišiadorių r. sav.	210	8	31	11	15	47	42	28	28
Kauno m. sav.	1698	39	257	150	157	493	198	135	269
Kauno r. sav.	475	10	70	30	68	123	45	36	93
Kėdainių r. sav.	302	6	59	29	22	54	59	28	45
Prienų r. sav.	181	11	32	15	28	49	15	8	23
Raseinių r. sav.	187	7	33	12	20	28	32	19	36
Klaipėdos apskritis	**1698**	**41**	**281**	**155**	**183**	**309**	**444**	**83**	**202**
Klaipėdos m. sav.	723	14	137	72	69	148	159	33	91
Klaipėdos r. sav.	256	6	42	13	12	61	77	12	33
Kretingos r. sav.	182	3	33	16	13	40	46	11	220
Neringos sav.	10	0	1	0	1	2	4	2	0
Palangos m. sav.	86	0	19	6	4	16	26	0	15
Skuodo r. sav.	83	4	15	15	7	16	8	5	13
Šilutės r. sav.	358	14	34	33	77	26	124	20	30
Marijampolės apskritis	**778**	**31**	**92**	**99**	**65**	**179**	**112**	**74**	**126**
Kalvarijos sav.	54	4	4	5	0	13	12	4	12
Kazlų Rūdos sav.	66	5	6	9	4	13	9	8	12
Marijampolės sav.	300	4	45	40	28	74	39	21	49
Šakių r. sav.	139	9	14	23	14	24	18	19	18
Vilkaviškio r. sav.	219	9	23	22	19	55	34	22	35
Panevėžio apskritis	**1361**	**18**	**252**	**101**	**105**	**295**	**270**	**103**	**217**
Biržų r. sav.	169	4	24	8	10	29	54	12	28
Kupiškio r. sav.	116	2	24	9	9	21	20	12	19
Panevėžio m. sav.	540	3	113	39	44	114	95	39	93
Panevėžio r. sav.	214	5	40	14	18	54	38	18	27
Pasvalio r. sav.	134	2	23	9	11	26	34	7	22
Rokiškio r. sav.	188	2	28	22	13	51	29	15	28

(tęsinys kitame puslapyje)

Lentelė. 5. Darbingo amžiaus asmenys, pirmą kartą pripažinti neįgaliaisiais pagal administracinę teritoriją, 2010 metai.

(tęsinys)

	Iš viso	A15A19	C00C97	F00F99	G00G99	I00I99	M00M99	S00T98	Kita
Šiaulių apskritis	**1742**	**30**	**319**	**197**	**135**	**304**	**346**	**129**	**282**
Akmenės r. sav.	143	2	22	36	10	17	20	9	27
Joniškio r. sav.	115	3	25	15	16	21	13	6	16
Kelmės r. sav.	166	3	32	15	9	31	29	14	33
Pakruojo r. sav.	184	5	32	22	17	23	49	18	18
Radviliškio r. sav.	259	8	47	23	25	31	56	17	52
Šiaulių m. sav.	595	5	111	63	35	114	117	49	101
Šiaulių r. sav.	280	4	50	23	23	67	62	16	35
Tauragės apskritis	**532**	**17**	**99**	**46**	**46**	**108**	**77**	**41**	**98**
Jurbarko r. sav.	152	5	35	16	9	29	20	13	25
Pagėgių sav.	55	1	8	7	5	9	9	1	15
Šilalės r. sav.	113	5	20	9	6	24	14	12	23
Tauragės r. sav.	212	6	36	14	26	46	34	15	35
Telšių apskritis	**707**	**35**	**136**	**77**	**52**	**137**	**95**	**59**	**116**
Mažeikių r. sav.	276	12	61	10	24	58	40	22	49
Plungės r. sav.	156	7	30	26	7	35	16	17	18
Rietavo sav.	40	4	6	4	4	7	5	4	6
Telšių r. sav.	235	12	39	37	17	37	34	16	43
Utenos apskritis	**817**	**14**	**146**	**47**	**89**	**186**	**147**	**45**	**143**
Anykščių r. sav.	149	2	21	10	27	28	22	12	27
Ignalinos r. sav.	100	1	17	12	8	19	25	4	14
Molėtų r. sav.	112	7	11	8	15	24	24	6	17
Utenos r. sav.	222	1	40	8	17	53	44	14	45
Visagino sav.	132	2	42	2	11	32	18	4	21
Zarasų r. sav.	102	1	15	7	11	30	14	5	19
Vilniaus apskritis	**3733**	**87**	**728**	**247**	**406**	**974**	**517**	**254**	**520**
Elektrėnų sav.	123	2	28	8	14	24	18	14	15
Šalčininkų r. sav.	280	14	34	6	28	78	82	12	26
Širvintų r. sav.	80	3	6	9	12	11	11	9	19
Švenčionių r. sav.	133	5	20	10	14	36	18	13	17
Trakų r. sav.	170	4	35	12	19	40	28	10	22
Ukmergės r. sav.	218	9	38	18	17	49	29	18	40
Vilniaus m. sav.	2257	36	483	162	262	595	265	129	325
Vilniaus r. sav.	472	14	84	22	40	141	66	49	56

Lentelė. 5. Darbingo amžiaus asmenys, pirmą kartą pripažinti neįgaliaisiais pagal administracinę teritoriją, 2010 metai (lentelės pabaiga).

	Vidutinis metinis gyventojų skaičius	Ligoninėse gydytų ligonių skaičius
Lietuvos Respublika	**3286820**	**805994**
Alytaus apskritis	**170344**	**40077**
Alytaus m. sav.	65242	16061
Alytaus r. sav.	30290	5509
Druskininkų sav.	23676	5569
Lazdijų r. sav.	23973	6076
Varėnos r. sav.	27163	6862
Kauno apskritis	**656959**	**164204**
Birštono sav.	5136	861
Jonavos r. sav.	50759	12686
Kaišiadorių r. sav.	34918	9745
Kauno m. sav.	342768	88760
Kauno r. sav.	89729	19514
Kėdainių r. sav.	61034	14678
Prienų r. sav.	32548	8361
Raseinių r. sav.	40067	9599
Klaipėdos apskritis	**371725**	**96814**
Klaipėdos m. sav.	180282	47132
Klaipėdos r. sav.	51884	13720
Kretingos r. sav.	44563	11707
Neringos sav.	3766	589
Palangos m. sav.	17439	4642
Skuodo r. sav.	22793	5662
Šilutės r. sav.	50998	13362
Marijampolės apskritis	**176030**	**40082**
Kalvarijos sav.	13035	2882
Kazlų Rūdos sav.	14005	3034
Marijampolės sav.	67378	16081
Šakių r. sav.	35078	8211
Vilkaviškio r. sav.	46534	9874
Panevėžio apskritis	**274605**	**72271**
Biržų r. sav.	31582	8395
Kupiškio r. sav.	22028	5977
Panevėžio m. sav.	110493	29328
Panevėžio r. sav.	41869	9644
Pasvalio r. sav.	31287	8034
Rokiškio r. sav.	37346	10893

(tęsinys kitame puslapyje)

Lentelė. 6. Vidutinis metinis gyventojų skaičius pagal administracinę teritoriją ir ligoninėse gydytų ligonių skaičius pagal administracinę teritoriją 2010 metais.

(tęsinys)

	Vidutinis metinis gyventojų skaičius	Ligoninėse gydytų ligonių skaičius
Šiaulių apskritis	**335403**	**87946**
Akmenės r. sav.	26234	8025
Joniškio r. sav.	28785	7847
Kelmės r. sav.	36191	9011
Pakruojo r. sav.	26151	6268
Radviliškio r. sav.	46703	11414
Šiaulių m. sav.	123211	34766
Šiaulių r. sav.	48128	10615
Tauragės apskritis	**122729**	**29359**
Jurbarko r. sav.	33646	8930
Pagėgių sav.	10937	2187
Šilalės r. sav.	29097	6988
Tauragės r. sav.	49049	11254
Telšių apskritis	**168728**	**37685**
Mažeikių r. sav.	63471	13235
Plungės r. sav.	42421	10115
Rietavo sav.	9710	1996
Telšių r. sav.	53126	12339
Utenos apskritis	**165709**	**42368**
Anykščių r. sav.	30681	7823
Ignalinos r. sav.	19133	5453
Molėtų r. sav.	22237	5572
Utenos r. sav.	46343	11643
Visagino sav.	27724	6385
Zarasų r. sav.	19591	5492
Vilniaus apskritis	**844588**	**195188**
Elektrėnų sav.	27273	6198
Šalčininkų r. sav.	36661	8257
Širvintų r. sav.	18385	4517
Švenčionių r. sav.	29570	6973
Trakų r. sav.	35438	9366
Ukmergės r. sav.	43926	11357
Vilniaus m. sav.	557126	129645
Vilniaus r. sav.	96209	18875

Lentelė. 6. Vidutinis metinis gyventojų skaičius pagal administracinę teritoriją ir ligoninėse gydytų ligonių skaičius pagal administracinę teritoriją 2010 metais (lentelės pabaiga).

PRIEDAS 2. NAUDOJAMŲ ALGORITMŲ SĄRAŠAS

P.2.1. Skaitinio integravimo algoritmas naudojant Ermito polinomus

Tam, kad apskaičiuotume integralus, turinčius pavidalą

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{(z-a)^2}{2\sigma^2}} f(z)dz,$$

kur *f* yra tam tikra duota funkcija, naudojama kvadratūrinė formulė, naudojanti Ermito polinomus (žr. Abramovich, Stegun (1968), lent. 25.10):

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{(z-a)^2}{2\sigma^2}} f(z)dz = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} \sigma f(a+\sigma y)dy =$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} f(a+\sigma y)dy \cong \frac{1}{\sqrt{\pi}}\sum_{i=1}^{n} w_i f(a+\sigma\sqrt{2}x_i),$$

kur w_i yra svoriai, x_i yra mazgai (Ermito polinomo nuliai), o *n* yra parenkamas mazgų skaičius (žr. Pav. 5, kaip jie pateikti Abramovich, Stegun (1968).). Buvo pasirinktas mazgų skaičius *n* = 20. Taigi,

w_{11} = 4.622436696006E-01;

w_{12} = 2.866755053628E-01;

w_{13} = 1.090172060200E-01;

w_{14} = 2.481052088746E-02;

w_{15} = 3.243773342238E-03;

w_{16} = 2.283386360163E-04;

w_{17} = 7.802556478532E-06;

w_{18} = 1.086069370769E-07;

w_{19} = 4.399340992273E-10;

w_{20} = 2.229393645534E-13;

o $w_i = w_{20-i+1}$, i = 1, 2,...,10. Analogiškai,

x_{11} = 0.2453407083009;

x_{12} = 0.7374737285454;

x_{13} = 1.2340762153953;

x_{14} = 1.7385377121166;

$x_{15} = 2.2549740020893;$

$x_{16} = 2.7888060584281;$

$x_{17} = 3.3478545673832;$

$x_{18} = 3.9447640401156;$

$x_{19} = 4.6036824495507;$

$x_{20} = 5.3874808900112;$

o $x_i = -x_{20-i+1}$, $i = 1, 2,...,10.$

Table 25.10 ABSCISSAS AND WEIGHT FACTORS FOR HERMITE INTEGRATION

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx \approx \sum_{i=1}^{n} w_i f(x_i) \qquad \int_{-\infty}^{\infty} g(x)dx \approx \sum_{i=1}^{n} w_i e^{x_i^2} g(x_i)$$

Abscissas = $\pm x_i$ (Zeros of Hermite Polynomials) Weight Factors = w_i

$\pm x_i$	w_i	$w_i e^{x_i^2}$
$n=2$		
0.70710 67811 86548	(-1) 8.86226 92545 28	1.46114 11826 611
$n=3$		
0.00000 00000 00000	(0) 1.18163 59006 04	1.18163 59006 037
1.22474 48713 91589	(-1) 2.95408 97515 09	1.32393 11752 136
$n=4$		
0.52464 76232 75290	(-1) 8.04914 09000 55	1.05996 44828 950
1.65068 01238 85785	(-2) 8.13128 35447 25	1.24022 58176 958
$n=5$		
0.00000 00000 00000	(-1) 9.45308 72048 29	0.94530 87204 829
0.95857 24646 13819	(-1) 3.93619 32315 22	0.98658 09967 514
2.02018 28704 56086	(-2) 1.99532 42059 05	1.18148 86255 360
$n=6$		
0.43607 74119 27617	(-1) 7.24629 59522 44	0.87640 13344 362
1.33584 90740 13697	(-1) 1.57067 32032 29	0.93558 05576 312
2.35060 49736 74492	(-3) 4.53000 99055 09	1.13690 83326 745
$n=7$		
0.00000 00000 00000	(-1) 8.10264 61755 68	0.81026 46175 568
0.81628 78828 58965	(-1) 4.25607 25261 01	0.82868 73032 836
1.67355 16287 67471	(-2) 5.45155 82819 13	0.89718 46002 252
2.65196 13568 35233	(-4) 9.71781 24509 95	1.10133 07296 103
$n=8$		
0.38118 69902 07322	(-1) 6.61147 01255 82	0.76454 41286 517
1.15719 37124 46780	(-1) 2.07802 32581 49	0.79289 00483 864
1.98165 67566 95843	(-2) 1.70779 83007 41	0.86675 26065 634
2.93063 74202 57244	(-4) 1.99604 07221 14	1.07193 01442 480
$n=9$		
0.00000 00000 00000	(-1) 7.20235 21560 61	0.72023 52156 061
0.72355 10187 52838	(-1) 4.32651 55900 26	0.73030 24527 451
1.46855 32892 16668	(-2) 8.84745 27394 38	0.76460 81250 946
2.26658 05845 31843	(-3) 4.94362 42755 37	0.84175 27014 787
3.19099 32017 81528	(-5) 3.96069 77263 26	1.04700 35809 767

$\pm x_i$	w_i	$w_i e^{x_i^2}$
$n=10$		
0.34290 13272 23705	(-1) 6.10862 63373 53	0.68708 18539 513
1.03661 08297 89514	(-1) 2.40138 61108 23	0.70329 63231 049
1.75668 36492 99882	(-2) 3.38743 94455 48	0.74144 19319 436
2.53273 16742 32790	(-3) 1.34364 57467 81	0.82066 61264 048
3.43615 91188 37738	(-6) 7.64043 28552 33	1.02545 16913 657
$n=12$		
0.31424 03762 54359	(-1) 5.70135 23626 25	0.62930 78743 695
0.94778 83912 40164	(-1) 2.60492 31026 42	0.63962 12320 203
1.59768 26351 52605	(-2) 5.16079 85615 88	0.66266 27732 669
2.27950 70805 01060	(-3) 3.90539 05846 29	0.70522 03661 122
3.02063 70251 20890	(-5) 8.57368 70435 88	0.78664 39394 633
3.88972 48978 69782	(-7) 2.65855 16843 56	0.98969 90470 923
$n=16$		
0.27348 10461 3815	(-1) 5.07929 47901 66	0.54737 52050 378
0.82295 14491 4466	(-1) 2.80647 45852 85	0.55244 19573 675
1.38025 85391 9888	(-2) 8.38100 41398 99	0.56321 78290 882
1.95178 79909 1625	(-2) 1.28803 11535 51	0.58124 72754 009
2.54620 21578 4748	(-4) 9.32284 00862 42	0.60973 69582 560
3.17699 91619 7996	(-5) 2.71186 00925 38	0.65575 56728 761
3.86944 79048 6012	(-7) 2.32098 08448 65	0.73824 56222 777
4.68873 89393 0582	(-10) 2.65480 74740 11	0.93687 44928 841
$n=20$		
0.24534 07083 009	(-1) 4.62243 66960 06	0.49092 15006 667
0.73747 37285 454	(-1) 2.86675 50536 28	0.49384 33852 721
1.23407 62153 953	(-1) 1.09017 20602 00	0.49992 08713 363
1.73853 77121 166	(-2) 2.48105 20887 46	0.50967 90271 175
2.25497 40020 893	(-3) 3.24377 33422 38	0.52408 03509 486
2.78880 60584 281	(-4) 2.28338 63601 63	0.54485 17423 644
3.34785 45673 832	(-6) 7.80255 64785 32	0.57526 24428 525
3.94476 40401 156	(-7) 1.08606 93707 69	0.62227 86961 914
4.60368 24495 507	(-10) 4.39934 09922 73	0.70433 29611 769
5.38748 08900 112	(-13) 2.22939 36455 34	0.89859 19614 532

Compiled from H. E. Salzer, R. Zucker, and R. Capuano, Table of the zeros and weight factors of the first twenty Hermite polynomials, J. Research NBS **48,** 111–116, 1952, RP2294 (with permission).

Pav. 5. Skaitinio integravimo, naudojant Ermito polinomus, koeficientai (iš Abramovich, Stegun (1968) – Table 25.10)

P.2.2. Gama funkcijos skaičiavimas naudojant Hornerio schemą

Tam, kad apskaičiuotume gama funkcijos reikšmę tam tikrame taške x, panaudota formulė (žr. Abramovich, Stegun (1968), sk. 6.1.36)

$$\Gamma(x+1) = 1 + b_1 \cdot x + b_2 \cdot x \cdot x + ... + b_8 \cdot x \cdot x \cdot x \cdot x \cdot x \cdot x \cdot x \cdot x + \varepsilon(x), 0 \le x \le 1,$$

$b_1 = -0.577191652;$

$b_2 = 0.988205891;$

$b_3 = -0.897056937;$

$b_4 = 0.918206857;$

$b_5 = -0.756704078;$

$b_6 = 0.482199394;$

$b_7 = -0.193527818;$

$b_8 = 0.035868343;$

kur $|\varepsilon(x)| \leq 3 \cdot 10^{-7}$, o taip pat panaudota lygybė $\Gamma(z+1) = z \cdot \Gamma(z), -\infty < z < \infty$. Polinomo reikšmė skaičiuojama panaudojant Hornerio schemą

$$\Gamma(x+1) \cong (...((b_8 \cdot x + b_7) \cdot x + b_6) \cdot x + ... + b_1) \cdot x + 1.$$

P.2.3. Atsitiktinių skaičių, tolygiai pasiskirsčiusių intervale [0,1] generavimo algoritmas

Atsitiktinių skaičių, tolygiai pasiskirsčiusių intervale [0,1], realizacijų $u_1, u_2, ..., u_n$, generavimui panaudotas algoritmas iš Aivazyan *et al*. (1983).

Pažymėkime $M = 2718281821$. Tuomet

$$u_j = K_j \cdot 2^{-35}, j = 1, 2, ..., n,$$

kur

$$K_0 = (r \cdot M) \bmod 2^{35},$$
$$K_j = (K_{j-1} \cdot M) \bmod 2^{35}, j = 1, 2, ..., n,$$

o r – realizacijos numeris (naudojami nelyginiai numeriai $r = 1, 3, 5, 7, ...$, tam, kad numerį įeinantis daugiklis, kuris yra dvejeto laipsnis, nesusiprastintų su 2^{35}).

Kadangi šis algoritmas naudojamas kelis dešimtmečius nuo pirmųjų didelių kompiuterių laikų, jame pagal tuose kompiuteriuose naudojamų slankaus kablelio skaičių užimamą atmintį naudojamas modulis mod 2^{35}. Tam, kad šis algoritmas būtų pritaikytas dabartiniams kompiuteriams (kuriuose galima atlikti modulio operacijas iki mod 2^{32}, taip kad visi tarpiniai veiksmai su sveikais skaičiais duotų reikšmes iš sveikųjų skaičių intervalo $[-2^{32} ... 2^{32} - 1]$), buvo panaudoti papildomi dydžiai ir kintamieji. Taigi, tarkime, kad turime realizacijos numerį r (programiškai jis turi būti sveikasis skaičius iš intervalo $[-2^{15} ... 2^{15} - 1]$, pvz. *Pascal* kalboje – *integer* tipo, taigi nagrinėjamu neneigiamų nelyginių skaičių atveju – iš aibės (1, 3, 5, ... , 32767)), ir apibrėžkime kintamuosius

$x_1 = 0,$

$x_2 = 0,$

$$x_3 = r,$$

(programiškai jie turi būti sveikieji skaičiai iš intervalo $[-2^{32} \ldots 2^{32} - 1]$, pvz. *Pascal* kalboje – *longint* tipo). Taip pat apibrėžkime tokio pat tipo kintamuosius q_1, q_2, q_3, ir konstantas

$$\begin{aligned}
y_1 &= 2,\\
y_2 &= 17419,\\
y_3 &= 12381,\\
L_5 &= 32 = 2^5,\\
L_{15} &= 32768 = 2^{15},\\
L_{20} &= 1048576 = 2^{20}.
\end{aligned}$$

Tuomet nuosekliai atlikę veiksmus

$$\begin{aligned}
q_1 &= x_1 \cdot y_3 + x_2 \cdot y_2 + x_3 \cdot y_1,\\
q_2 &= x_2 \cdot y_3 + x_3 \cdot y_2,\\
q_3 &= x_3 \cdot y_3,\\
x_3 &= q_3 \bmod L_{15},\\
q_2 &= q_2 \bmod L_{20} + q_3 \operatorname{div} L_{15},\\
x_2 &= q_2 \bmod L_{15},\\
q_1 &= q_1 \bmod L_5 + q_2 \operatorname{div} L_{15},\\
x_1 &= q_1 \bmod L_5,
\end{aligned} \qquad (1)$$

gausime, kad

$$K_0 = x_1 \cdot 2^{30} + x_2 \cdot 2^{15} + x_3,$$

o toliau, nuosekliai kartodami veiksmus (1) ir naudodami vis besikeičiančius kintamuosius x_1, x_2, x_3, pagal tą pačią formulę, kaip ir K_0, gausime

$$K_j = x_1 \cdot 2^{30} + x_2 \cdot 2^{15} + x_3,\ j = 1, 2, \ldots, n.$$

Programiškai,

$$u_j = K_j \cdot 2^{-35} = (1073741824.0 \cdot x_1 + 32768.0 \cdot x_2 + x_3) / 34359738368.0,\ j = 1, 2, \ldots, n.$$

PRIEDAS 3. ALGORITMINIS BINARINĖS DIDELIO MATAVIMO DUOMENŲ SKAIDYMO PROCEDŪROS APRAŠYMAS

P.3.1. Įvadas

Nagrinėjamo algoritmo (pažymėkime jį A) aprašyme tam tikru pseudokodu naudosime pažymėjimus:

`X` – pradinė seka

`q` – pradinės sekos elementų pasikartojimų skaičius

`dim` – pradinės sekos matavimas

`N` – pradinės sekos ilgis

`XGrp` – grupuota seka

`qGrp` – grupuotos sekos elementų pasikartojimų skaičius

`kGrp` – grupuotos sekos elementų skaičius (be pasikartojimų)

`NGrp` – grupuotos sekos ilgis

`maindone` – valdymo parametras

P.3.2. Bendras algoritmo aprašymas

P.3.2.1. Algoritmo aprašymas bendriausiu lygiu

Bendriausiu lygiu algoritmas A išskaidomas į tris dalis:

A1. Pradinis žingsnis. Įvedami duomenys ir parametrai (tarp jų pradinė seka `X`, su elementų pasikartojimų skaičiumi `q` ir matavimu `dim`, kurios ilgis yra `N`), pagal juos atliekamas pirmasis grupavimo žingsnis ir gaunama pirmoji grupuota seka `XGrp` su elementų pasikartojimų skaičiumi `qGrp`, kurios ilgis yra `NGrp`. Po to apskaičiuojamas valdymo parametras `maindone` (jei jis lygus `true`, tai procedūros pagrindinė dalis baigta ir pereinama prie baigiamojo žingsnio A3, o jei jis lygus `false`, tai algoritmo pagrindinė dalis nėra baigta, ir pereinama prie žingsnio A2).

A2. Pagrindinė algoritmo dalis. Tol, kol valdymo parametras `maindone` lygus `false`, atliekami pagrindiniai algoritmo žingsniai, po kiekvieno iš jų grupuotos sekos ilgis padidėja vienetu:

```
While not maindone do
  begin
  A2.1. Pagrindinis algoritmo žingsnis. Po kiekvieno žingsnio grupuotos sekos ilgis NGrp
   padidėja vienetu, ir iš naujo apskaičiuojamas valdymo parametras maindone.
  end;
```

A3. Baigiamasis žingsnis. Jei reikia (tuo atveju, kai nurodoma, kad kiekvieno kvadratėlio didžiausia kraštinė turi skirtis nuo mažiausios ne daugiau kaip nurodytą skaičių kartų), atliekami papildomi skaidymai, ir juos pabaigus apskaičiuojama vidutinė grupavimo paklaida ir maksimali grupavimo paklaida.

P.3.2.2. Algoritmo pagrindinių žingsnių aprašymas

Pateiksime algoritmo A pagrindinių žingsnių bendrą aprašymą, kaip pavyzdį duodami nuorodas į Pascal programavimo kalba parašytos programos tekstą.

A1. Pradinis žingsnis

A1.1. Konstantų nustatymas ir parametrų nurodymas.

A1.1.1. Nustatomos konstantos `RealNmax` ir `IntegerNmax`, priklausančios nuo naudojamo kompiuterio ir programinės įrangos, kurios nusako maksimalius atitinkamai realių skaičių ir sveikų skaičių masyvų ilgius.

A1.1.2. Nurodomas parametras `minNGrp` (mažiausias pakankamas grupuotos duomenų sekos ilgis, t. y. kada jis pasiekiamas, procedūra baigiama, išskyrus tuo atveju, kai nėra pasiekta pakankamai maža vidutinė kvadratinė grupavimo paklaida).

A1.1.3. Nurodomas parametras `maxNGrp` (grupuotos duomenų sekos ilgis, kuomet esant mažesniam pradinės sekos ilgiui, grupavimas gali būti neatliekamas, tokiu atveju perkopijuojant pradinę seką `X` su elementų pasikartojimų skaičiumi `q` į grupuotą seką `XGrp` su elementų pasikartojimų skaičiumi `qGrp`) (žr. A1.4.1).

A1.1.4. Nurodomas parametras `ForceGrp` (jei jis lygus `true`, tai grupavimas atliekamas priverstinai, o jei jis lygus `false`, tai grupavimas gali būti neatliekamas, tokiu atveju perkopijuojant pradinę seką `X` su elementų pasikartojimų skaičiumi `q` į grupuotą seką `XGrp` su elementų pasikartojimų skaičiumi `qGrp`) (žr. A1.4.1).

A1.1.6. Nurodomas mažiausias kraštinės padalijimų skaičius `minM` ir didžiausias kraštinės padalijimų skaičius `maxM` (abu jie yra dvejeto laipsniai, t. y. galimos reikšmės yra iš {1, 2, 4, 8, ...}).

A1.1.7. Nurodomas koks pradedant procedūrą turi būti mažiausias kraštinės padalijimų skaičius `minresM` ir koks procedūros veikimo metu turi būti didžiausias kraštinės padalijimų skaičius `maxresM` (abu jie yra dvejeto laipsniai, t.y. galimos reikšmės yra iš {1, 2, 4, 8, ...}).

A1.1.8. Nurodomas koks turi būti pabaigus procedūrą didžiausias leidžiamas santykis tarp kvadratėlio didžiausios ir mažiausios kraštinių `resdifM` ir didžiausias kraštinės padalijimų skaičius `maxresM` (abu jie yra dvejeto laipsniai, t. y. galimos reikšmės yra iš {1, 2, 4, 8, ...}).

A1.1.9. Nurodomas valdymo parametras `MSErrmode`, kuris nurodo, ar skaičiuojama kvadratinė paklaida:

`MSErrmode = 0` – kvadratinė paklaida neskaičiuojama;

`MSErrmode = 1` – kvadratinė paklaida skaičiuojama tik baigiant algoritmą;

`MSErrmode = 2` – kvadratinė paklaida naudojama parinkti kvadratėlį;

`MSErrmode = 3` – kvadratinė paklaida naudojama kvadratėlio optimaliam skaidymui.

A1.1.10. Nurodomas valdymo parametras `Exitmode`, kuris nurodo algoritmo pabaigos sąlygą priklausomai nuo tuo metu esančio grupuotos sekos ilgio `NGrp` ir atitinkamos kvadratinės paklaidos `sGrp` (jei `MSErrmode<=1` , tai `Exitmode = 4`):

`Exitmode = 0` – `sGrp <= maxMSErr`;

`Exitmode = 1` – `NGrp >= minNGrp or sGrp <= maxMSErr`;

`Exitmode = 2` – `NGrp >= minNGrp and sGrp <= maxMSErr`;

`Exitmode = 3` – `(NGrp>=minNGrp and sGrp<=maxMSErr) or sGrp<=minMSErr`;

`Exitmode = 4` – `NGrp >= minNGrp`.

A1.1.11. Nurodomas valdymo parametras `Selectmode`, kuris nurodo sąlygą kvadratėlio parinkimo skaidymui (jei `MSErrmode>1` , tai `Selectmode = 3`):

`Selectmode = 1` – parenkama pagal didžiausią kvadratėlyje esantį taškų skaičių (su pasikartojimais);

`Selectmode = 2` – parenkama pagal didžiausią kvadratėlyje esantį taškų skaičių (su pasikartojimais) padaugintą iš kvadratėlio diagonalės ilgio;

`Selectmode = 3` – parenkama pagal maksimalų kvadratinės paklaidos sumažėjimą;

A1.1.12. Nurodomas valdymo parametras `Splitmode`, nurodantis kurios kvadratėlio kraštinės parenkamos galimam skaidymui:

`Splitmode = 0` – parenkama pirmoji kraštinė su maksimaliu ilgiu;

`Splitmode = 1` – parenkamos visos kraštinės su maksimaliu ilgiu;

`Splitmode = 2` – parenkamos visos leistinos (priklausomai nuo kitų sąlygų) kraštinės;

`Splitmode = 3` – parenkamos visos galimos kraštinės;

A1.1.13. Nurodomas valdymo parametras `SplitFncNo`, nurodantis, kaip parenkama kraštinė kvadratėlio skaidymui (jei `MSErrmode=3` , tai `SplitFncNo = 4`):

`SplitFncNo = 1` – parenkama kraštinė, kuriai gaunamas mažiausias maksimalus iš dviejų suskaidytų kvadratėlių taškų skaičius (su pasikartojimais);

`SplitFncNo = 2` – parenkama kraštinė, kuriai gaunama mažiausia suma, gaunama padauginus kiekvieno iš dviejų suskaidytų kvadratėlių taškų skaičių (su pasikartojimais) iš jo maksimalaus kraštinės ilgio;

`SplitFncNo = 3` – parenkama kraštinė, kuriai gaunama mažiausia sumos, gaunamos padauginus kiekvieno iš dviejų suskaidytų kvadratėlių taškų skaičių (su pasikartojimais) iš jo diagonalės ilgio;

`SplitFncNo = 4` – parenkama pagal maksimalų kvadratinės paklaidos sumažėjimą;

A1.2. Masyvų sudarymas.

A1.3.1. Iš `dim` (`dim` – pradinės sekos matavimas) komponenčių sudaromas masyvas `X` (kurio kiekviena komponentė yra `real` tipo masyvas ilgio `N`), kuriame bus talpinami pradinės sekos taškai.

A1.3.2. Sudaromas masyvas `q` (`integer` tipo, ilgio `N`), kuriame bus talpinamas pradinės sekos taškų pasikartojimų skaičius. Dažnai visos šios reikšmės yra lygios vienetui.

A1.3.3. Sudaromas masyvas `qGrp` (`integer` tipo, ilgio `N`), kuriame bus talpinamas grupuotos sekos taškų pasikartojimų skaičius.

A1.3.4. Sudaromas masyvas `kGrp` (`integer` tipo, ilgio `N`), kuriame talpinamas grupuotos sekos taškų skaičius (be pasikartojimų).

A1.3.5. Iš `dim` komponenčių sudaromas masyvas `lx` (kurio kiekviena komponentė yra `integer` tipo masyvas ilgio `N`), kuriame bus talpinami kvadratėlių kraštinių ilgiai mažiausio galimo kvadratėlio atžvilgiu (dvejeto laipsniai, t. y. galimos reikšmės yra iš {1, 2, 4, 8, ... , `maxM`}).

A1.3.6. Iš `dim` komponenčių sudaromas masyvas `u` (kurio kiekviena komponentė yra `integer` tipo masyvas ilgio `N`), kuriame bus talpinamos kvadratėlių pirmosios viršūnės koordinatės (galimos reikšmės yra iš {0, 1, 2, 3, , ... , `maxM-1` }).

A1.3.7. Iš `dim` komponenčių sudaromas masyvas `uu` (kurio kiekviena komponentė yra `integer` tipo masyvas ilgio `N`), kuriame bus talpinamos kiekvieną pradinės sekos tašką atitinkančių mažiausių kvadratėlių pirmosios viršūnės koordinatės (galimos reikšmės yra iš {0, 1, 2, 3, , ... , `maxM-1` }).

A1.3.8. Sudaromas masyvas `multMs` (`integer` tipo, ilgio `dim`), kuriame bus talpinami papildomi daugikliai (dvejeto laipsniai, t. y. galimos reikšmės yra iš {1, 2, 4, 8, ...}) maksimaliam kraštinės dalijimo skaičiui mų `maxM` (pagal nutylėjimą jie visi lygūs vienetui).

A1.3.9. Sudaromas masyvas `kDF` (`integer` tipo, ilgio `N`), kuriame bus talpinami indeksai, nuo kurių masyve `Nums` yra nuosekliai talpinami kvadratėlyje esančių taškų indeksai.

A1.3.10. Sudaromas masyvas `Nums` (`integer` tipo, ilgio `N`), kuriame bus talpinami taškų esančių kvadratėliuose indeksai.

A1.3.11. Sudaromas masyvas `kNeighb` (`integer` tipo, ilgio `N`), kuriame gali būti talpinami pradinės sekos taškų kaimynų skaičiai (vartotojo suskaičiuojami atskirai).

A1.3.12. Sudaromas masyvas `sGrp` (`real` tipo, ilgio `N`), kuriame bus talpinami kvadratėlių grupavimo paklaidos sumažėjimo dydžiai.

A1.3.13. Sudaromas masyvas `kdsGrp` (`integer` tipo, ilgio `N`), kuriame gali būti talpinami tarpiniai indeksai.

A1.3.14. Sudaromas masyvas `ind` (`integer` tipo, ilgio `N`), kuriame bus talpinami indeksai, nusakantys, ar kvadratėlis gali būti skaidomas (jei reikšmė didesnė už 0), ar ne (jei reikšmė lygi 0).

A1.3.15. Sudaromi masyvai `X1` ir `X2` (`real` tipo, ilgio `dim`), kuriame bus nurodomi minimumai ir maksimumai pagal koordinačių ašis (visos duomenų sekos `X` reikšmės turės patekti į šias ribas).

A1.3.16. Sudaromas masyvas `EX` (`real` tipo, ilgio `N`), kuriame bus patalpinamos grupuotos sekos reikšmės.

A1.3.17. Sudaromas masyvas `tmpNums` (`integer` tipo, ilgio `N`), kuriame bus talpinami taškų, esančių kvadratėliuose, tarpiniai indeksai.

A1.3.18. Sudaromas masyvas `dimind` (`integer` tipo, ilgio `N`), kuriame bus talpinami indeksai, nusakantys, pagal kurį matavimą turėtų būti skaidomas kvadratėlis.

A1.3.19. Sudaromi masyvai `n1pts, n2pts` (`integer` tipo, ilgio `N`), kuriuose (jei naudojamas parametras `NFirst`, nurodantis, kad pradinėje sekoje, susidedančioje iš dviejų sekų, toks yra pirmosios sekos ilgis) bus talpinami pirmosios ir antrosios sekų taškų skaičiai (su pasikartojimais). Taip pat sudaromas masyvas `spts` (`real` tipo, ilgio `N`), kuriame bus talpinamos atitinkamos chi-kvadrat statistikos reikšmės. Be to, sudaromas masyvas `kdifGrp` (`integer` tipo, ilgio `N`), kuriame bus talpinami taškų skaičių skirtumai (absoliučiu dydžiu) tarp pirmosios ir antrosios sekų.

A1.3. Duomenų įvedimas.

A1.3.1. Patikrinus, ar sekos ilgis `N` neviršija konstantos `Nmax = RealNmax`, o sekos matavimas `dim` neviršija konstantos `dimmax = IntegerNmax`, įvedama pradinė duomenų seka `X` su elementų pasikartojimų skaičiumi `q`.

A1.3.2. Įvedami parametrų masyvai `X1` ir `X2`, kuriuose nurodomi minimumai ir maksimumai pagal koordinačių ašis (visos duomenų sekos `X` reikšmės turi patekti į šias ribas).

A1.4. Pradiniame žingsnyje atliekami skaičiavimai.

A1.4.1. Duomenų sekos perkopijavimas ir išėjimas iš procedūros, jei grupavimo atlikti nereikia:

```
if (not ForceGrp and (N <= maxNGrp))  then
  begin
  < pradinės sekos X perkopijavimas grupuotą seką XGrp>
  maindone := true;
  exit;
  end;
```

A1.4.2. Maksimalaus galimo grupuotos duomenų sekos ilgio `NGrp_max` apskaičiavimas, naudojant maksimalų kraštinės padalijimų skaičių `maxM`.

A1.4.3. Maksimalaus galimo grupuotos duomenų sekos ilgio `NGrp_resmax` apskaičiavimas, naudojant maksimalų kraštinės padalijimų skaičių `maxresM`.

A1.4.4. Minimalaus galimo grupuotos duomenų sekos ilgio `NGrp_resmin` apskaičiavimas, naudojant minimalų kraštinės padalijimų skaičių `minresM`.

A1.4.5. Minimalaus galimo grupuotos duomenų sekos ilgio `NGrp_min` apskaičiavimas, naudojant minimalų kraštinės padalijimų skaičių `minM`.

A1.4.6. Parametrų `(ind, lx, u)` apskaičiavimas, kai yra tik vienas taškas kvadratėlyje. Tai yra tarpiniai masyvai, kurių ilgis `NGrp`. `ind` yra vienmatis masyvas, kuris nurodo, kvadratėlio statusą, jei reikšmė lygi 0, tai kvadratėlis negali būti daugiau skaidomas. `lx` ir `u` yra kvadratėlių ilgių ir pirmų taškų koordinačių matavimo `dim` masyvai (abiejų visos reikšmės yra dvejeto laipsniai, t.y. galimos reikšmės yra iš {1, 2, 4, 8, ...})

A1.4.7. Kvadratėlių su koordinatėmis `(lx, u)` sumažinimas iki minimalaus dydžio, t. y. pagal kiekvieną koordinatę kvadratėliai, jei reikia, sumažinami per pusę, kol kiekvienoje pusėje yra bent vienas sekos `X` taškas.

A1.4.8. Izoliuotų taškų atskyrimas, jei nurodomas parametras `isolLevel > 0`, kuris nurodo kad tiek ar mažiau kaimynų turintis kvadratėlis (sekos `X` kaimynų skaičius, suskaičiuotas pagal norimą algoritmą, patalpinamas į masyvą `kNeighb`) laikomas izoliuotu ir atliekamas priverstinis jo skaidymas. Taip pat atliekamas kvadratėlių papildomo skaičiaus `(NGrpPlus)` apskaičiavimas.

A1.4.9. Išėjimo sąlygos (maindone) nustatymas:

```
maindone := false;
if NGrp + NGrpPlus >= NGrp_resmax  then
  maindone := true;
if NGrp + NGrpPlus >= maxNGrp  then
  maindone := true;
if NGrp_min = NGrp_max  then
  maindone := true;
```

A1.4.10. Vidutinio kvadratinio ir maksimalaus nuokrypių (MSErr_min, maxErr_min) apskaičiavimas

A1.4.11. Išėjimo sąlygos maindone nustatymas pagal nurodytą vidutinio kvadratinio nuokrypio ribą MSErr_min, pagal nurodytą pakankamą grupuotos sekos ilgį minNGrp ir valdymo parametrus MSErrmode ir Exitmode, taip pat jei suskaidyti visi kvadratėliai:

```
if ((MSErrmode > 1) and (Exitmode <> 4))  then
  begin
  sumErr2 := sqr(MSErr_min);
  case Exitmode of
    0:if sumErr2 < maxErr2  then
        maindone := true;
    1:if ((sumErr2 < maxErr2) or (NGrp + NGrpPlus >= minNGrp))  then
        maindone := true;
    2:if ((sumErr2 < maxErr2) and (NGrp + NGrpPlus >= minNGrp))  then
        maindone := true;
    3:if (((sumErr2 < maxErr2) and (NGrp + NGrpPlus >= minNGrp)) or
           (sumErr2 < minErr2))  then
        maindone := true;
    end;
  end
else
  begin
  if NGrp + NGrpPlus >= minNGrp  then
    maindone := true;
  end;

i := 0;
for j:=0  to NGrp-1  do
  if ind^[j] <> 0  then
    Inc(i);
if i < 1  then
  maindone := true;
```

A1.4.12. Apskaičiavimas pirmos ir antros sekų taškų skaičiaus (n1pts, n2pts) kai pirmos sekos ilgis yra Nfirst.

A2. Pagrindinė algoritmo dalis.

A2.1. Radimas indekso `i`, kuriam grupavimo paklaidos sumažėjimas didžiausias, naudojant reikšmes iš masyvo `sGrp`, arba, jei `MSErrmode = 1`, indekso parinkimas pagal valdymo parametrą `Selectmode`.

A2.2. Nustatymas parametro `IsPlus`, ar kvadratėlis turi būti būtinai skaidomas

A2.3. Nustatymas parametro `IsSquare`, ar kvadratėlis yra tiksliai kvadratinės formos, t. y ar jo kraštinės lygios.

A2.4. Maksimalaus kraštinės ilgio `maxlen` ir atitinkamos dimensijos indekso `indmaxlen` nustatymas.

A2.5. Nustatymas indekso `indsplit`, pagal kurį bus skaidomas kvadratėlis.

A2.6. Apskaičiavimas suskaidyto kvadratėlio abiejų dalių kvadratinių nuokrypių `MSErr1` ir `MSErr2`.

A2.7. Kvadratėlio paruošimas skaidymui, kad taškai pagal nurodytą ašį iš pradžių būtų kairėje pusėje, o toliau būtų dešinėje pusėje.

A2.8. Kvadratėlio išskaidymas į du (`NGrp` reikšmė padidėja vienetu) ir apkarpymas.

A2.9. Apskaičiavimas dviems naujiems kvadratėliams vidutinės kvadratinės paklaidos sumažėjimo.

A2.10. Išėjimo sąlygos `(maindone)` nustatymas

A2.11. Apskaičiavimas pirmos ir antros sekų taškų skaičiaus `(n1pts, n2pts)` kai pirmos sekos ilgis yra `Nfirst.`

A3. Baigiamasis žingsnis.

A3.1. Gautų kvadratėlių ilgiausios kraštinės dalijamos tol, kol patenkinama sąlyga, kad santykis tarp kvadratėlio didžiausios ir mažiausios kraštinių neviršija `resdifM`. Atskiru atveju, jei `resdifM` lygus 1, galų gale gauname visus „kvadratėlius“ ar „kubelius“.

A3.2. Vidutinio kvadratinio ir maksimalaus nuokrypių `(MSErr_min, maxErr_min)` apskaičiavimas.

A3.3. Apskaičiavimas pirmos ir antros sekų taškų skaičiaus `(n1pts, n2pts)`, kai pirmos sekos ilgis yra `Nfirst.`