VILNIAUS UNIVERSITETAS

Lina
DREIŽIENĖ

Erdvinių Gauso duomenų klasifikavimo rizika naudojant tiesines diskriminantines funkcijas

DAKTARO DISERTACIJOS SANTRAUKA

Fiziniai mokslai,
matematika 01P

VILNIUS 2019

Disertacija rengta 2014-2018 metais Vilniaus universitete

Moksliniai vadovai:
prof. dr. Marijus Radavičius (Vilniaus universitetas, fiziniai mokslai, matematika, 01P). Nuo 2014-10-01 iki 2017-09-27;
prof. dr. Kęstutis Dučinskas (Vilniaus universitetas, fiziniai mokslai, matematika, 01P). Nuo 2017-09-28 iki 2018-09-30.
Mokslinis konsultantas: prof. dr. Marijus Radavičius (Vilniaus universitetas, fiziniai mokslai, matematika, 01P). Nuo 2017-09-28 iki 2018-09-30.

Gynimo taryba:
Pirmininkas – prof. habil. dr. Kęstutis Kubilius (Vilniaus universitetas, fiziniai mokslai, matematika, 01P).
Nariai:
prof. dr. Ričardas Krikštolaitis (Vytauto Didžiojo universitetas, fiziniai mokslai, matematika, 01P);
prof. habil. dr. Remigijus Leipus (Vilniaus universitetas, fiziniai mokslai, matematika, 01P);
prof. habil. dr. Rimantas Rudzkis (Vilniaus universitetas, fiziniai mokslai, matematika, 01P);
prof. dr. Jūratė Šaltytė-Benth (Oslo universitetas, Norvegija, fiziniai mokslai, matematika, 01P).

Disertacija ginama viešame gynimo tarybos posėdyje 2019 m. kovo mėn. 1 d. 12 val. Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų instituto 203 auditorijoje.
Adresas: Akademijos g. 4, LT-08412 Vilnius, Lietuva.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu: *https://www.vu.lt/naujienos/ivykiu-kalendorius*

VILNIUS UNIVERSITY

Lina
DREIŽIENĖ

Classification risk of Gaussian spatial data using linear discriminant fuctions

SUMMARY OF DOCTORAL DISSERTATION

Physical Sciences,
Mathematics 01P

VILNIUS 2019

This dissertation was written between 2014 and 2018 at Vilnius University.

Academic supervisors:
Prof. Dr. Marijus Radavičius (Vilnius University, Physical sciences, Mathematics, 01P). From 01/10/2014 to 27/09/2017;
Prof. Dr. Kęstutis Dučinskas (Vilnius University, Physical sciences, Mathematics, 01P). From 28/09/2017 to 30/09/2018.
Academic consultant: Prof. Dr. Marijus Radavičius (Vilnius University, Physical sciences, Mathematics, 01P). From 28/09/2017 to 30/09/2018.

This doctoral dissertation will be defended at the public meeting of the Dissertation Defence Panel:
Chairman – Prof. Habil. Dr. Kęstutis Kubilius (Vilnius University, Physical sciences, Mathematics, 01P).
Members:
Prof. Dr. Ričardas Krikštolaitis (Vytautas Magnus University, Physical sciences, Mathematics, 01P).
Prof. Habil. Dr. Remigijus Leipus (Vilnius University, Physical sciences, Mathematics, 01P).
Prof. Habil. Dr. Rimantas Rudzkis (Vilnius University, Physical sciences, Mathematics, 01P).
Prof. Dr. Jūratė Šaltytė-Benth (University of Oslo, Physical sciences, Mathematics, 01P).

The dissertation will be defended at the public meeting of the Dissertation Defence Panel at 12:00 a.m. on 1 March, 2019 at Vilnius University, Institute of Data Science and Digital Technologies, Room 203.
Address: Akademijos st. 4, LT-08412 Vilnius, Lithuania.
The text of this dissertation can be accessed at the library of Vilnius University and the website www.vu.lt/lt/naujienos/ivykiu-kalendorius

Tiriamoji problema ir jos aktualumas

Pagrindinis disertacijos tyrimų objektas yra erdvėje koreliuotų duomenų diskriminantinė analizė (DA). Klasikinė diskriminantinė analizė yra daugiamatės statistikos metodas tiriantis objektų klasifikavimą remiantis mokymo imtimi, ir dažnai vadinamas klasifikavimo su mokymu metodu (angl. *supervised classification*). Viena iš prielaidų, kuria grindžiamas šis metodas, yra stebinių nepriklausomumas. Erdvinių duomenų atveju ši prielaida dažnai yra nepagrįsta, nes erdvėje arčiau vieni kitų esantys stebiniai yra labiau susiję nei tie stebiniai, kuriuos skiria didesnis atstumas. Šis reiškinys vadinamas erdvine koreliacija arba autokoreliacija (angl. *spatial correlation, autocorrelation*). Erdvėje koreliuotų duomenų pavyzdžių gausu įvairiuose geomoksluose, pvz. astronomijoje, augalininkystėje, tiriant dirvožemio struktūrą, įvairias vandens telkinių charakteristikas ar net vertinant nekilnojamojo turto pardavimo kainas. Akivaizdu, kad atliekant erdvinių duomenų diskriminantinę analizę ir siekiant minimizuoti klasifikavimo riziką (arba klaidingo klasifikavimo tikimybę) ar maksimizuoti teisingo klasifikavimo tikimybę, būtina atsižvelgti į erdvinę koreliaciją - jos ignoravimas gali reikšmingai paveikti klasifikavimo procedūrų tikslumą. Klasifikavimo procedūrų tikslumui įtakos taip pat gali turėti ir kita erdviniams duomenims būdinga savybė – anizotropija (angl. *anisotropy*). Įtraukiant šią duomenų savybę į modelį pasikeičia kovariacijų funkcijos struktūra – ji papildoma anizotropiškumo parametrais, o tai yra reikšmingas žingsnis siekiant sumažinti klasifikavimo riziką.

Tradiciškai statistiniai erdvinių duomenų modeliai (Cressie [6], Cressie, Wikle [7]) yra skirstomi į dvi klases: geostatistiniai modeliai su tolydžiu erdviniu indeksu ir gardelės (angl. *lattice*) tipo modeliai. Remiantis šia klasifikacija disertacijoje tiriamos dvi Gauso atsitiktinių laukų klasės: geostatistiniai Gauso atsitiktiniai laukai (GGRF) ir Gauso Markovo atsitiktiniai laukai (GMRF).

Priminsime, kad klasifikavimo su mokymu metode naudojamas planas, kai mokymo imties stebiniai imami ne iš populiacijų mišinio (angl. *mixture sampling design*), o iš kiekvienos populiacijos atskirai (angl. *separate sampling design*) (McLachlan [17]).

Pagrindinis šio darbo tikslas yra ištirti Bajeso klasifikavimo taisykle pagrįstas procedūras, kurios atsižvelgtų į erdvinę stebinių koreliaciją ir, formuojant klasifikavimo taisykles, naudotų ne marginalinius, o sąlyginius populiacijų skirstinius klasifikuojamame taške. Remiantis pasiūlytomis klasifikavimo procedūromis, priskirti atsitiktinio Gauso lauko (angl. *Gaussian random field* (GRF)) stebinį vienai iš keleto populiacijų ir įvertinti klasifikavimo riziką.

Erdvinė statistika, kaip mokslas, pradėjo vystytis sąlyginai neseniai, apie 1980 m., todėl natūralu, kad darbų, susijusių su erdvėje koreliuotų stebinių diskriminantine analize nėra labai daug. Pradininku šioje srityje laikomas Switzer [24], vėliau jo darbą pratęsė Mardia [15] įtraukdamas erdvinius diskriminavimo metodus formuojant klasifikavimo žemėlapius. Taip pat šioje srityje dirbo Klein ir Press [13], Shekhar ir kiti [23], Okamoto [18], McLachlan [17], Batsidis ir Zografos [1], tačiau nė vienas iš paminėtų autorių neanalizavo klasifikavimo rizikos. Klasifikavimo riziką, susijusią su nekoreliuotais stebėjimais ir įvairiais mokymo imčių planais tyrė Dučinskas [9]. Šaltytė [20], Šaltytė ir Dučinskas [20] pasiūlė aktualiosios klaidingo klasifikavimo tikimybės vidurkio aproksimacijos formulę skalariniam GGRF stebiniui dviejų klasių atveju, vėliau rezultatus apibendrino daugiamačiam erdvės-laiko modeliui (Šaltytė-Benth, Dučinskas [22]). Išsamų empirinį skirtingų klasifikavimo procedūrų palyginimą galime rasti Atkinson ir Lewis [1] bei Berret ir Calder [4]) darbuose. Visose paminėtose publikacijose buvo *daroma prielaida, kad stebinys, kurį siekiama klasifikuoti (vadinsime jį fokaliniu stebiniu, angl. focal observation) ir mokymo imtis yra nepriklausomi*, t.y. naudojami marginaliniai fokalinio stebinio tankiai formuojant diskriminantines funkcijas. Šios nepriklausomumo prielaidos erdvinių duomenų klasifikavimo uždaviniuose pirmą kartą atsisakė K. Dučinskas [10], [11]

pateikdamas klasifikavimo klaidos aproksimacijos formulę, kai nežinomos tikrosios vidurkio parametrų reikšmės bei naudojama kovariacijų funkcija su nežinomu vieninteliu mastelio parametru, kitus parametrus laikant žinomais, t.y. tiriama ne pilno populiacijų neapibrėžumo situacija.

Disertacijoje pateikiamos formulės klasifikavimo rizikai, kai nežinomi visi populiacijų parametrai, įtraukiant ir anizotropijos parametrus, kurie aukščiau minėtuose darbuose buvo ignoruojami. Taip pat disertacijoje pateiktas išplėtimas į daugelio klasių ir daugiamatį atvejus bei sprendžiamas klasifikavimo į vieną iš dviejų klasių uždavinys GMRF stebiniui.

Darbo tikslas ir uždaviniai

Sudaryti erdvinių Gauso stebinių tiesines diskriminantines funkcijas ir ištirti su jomis susijusią klasifikavimo riziką.
Siekiant numatyto tikslo buvo suformuluoti šie uždaviniai:

- Išvesti klasifikavimo rizikos formulę ir jos įvertinių analitines išraiškas GGRF stebiniui ir ištirti jų savybes.
- Išvesti aktualiosios klasifikavimo rizikos vidurkio aproksimacijos formules skaliariniam ir vektoriniam GGRF dviejų klasių atveju.
- Išvesti aktualiosios klaidingo klasifikavimo tikimybės vidurkio aproksimacijos formulę skaliariniam GGRF daugelio klasių atveju.
- Išvesti klasifikavimo rizikos ir jos aproksimacijos formules GMRF stebinio klasifikavimo į dvi klases atveju.
- Implementuoti pasiūlytas erdvinių duomenų diskriminantinės analizės procedūras ir modeliavimo būdu ištirti jų veikimo priklausomybę nuo Gauso lauko modelio parametrų bei palyginti modeliavimo būdu įvertintą riziką su pasiūlytomis jos aproksimacijomis.
- Pritaikyti pasiūlytą klasifikavimo procedūrą realiems duomenims.

Tyrimų metodika

Pagrindinis disertacijoje naudojamas metodas – diskriminantinės analizės teorija pritaikyta daugiamačiams Gauso skirstiniams. Daugelis įrodymų pagrįsti daugiamačio Gauso skirstinio savybėmis. Aktualiosios klasifikavimo rizikos vidurkio aproksimacijos formulėms išvesti naudojamas Teiloro skleidinys. Nežinomi parametrai vertinami maksimalaus tikėtinumo (MT) metodu. Remtasi Mardia ir Marshal [16] teorema apie asimptotinį MT parametrų normalumą. Skaitiniai eksperimentai atlikti su R pagalba.

Darbo mokslinis naujumas ir jo reikšmė

Šiame darbe pateikti nauji rezultatai:

- Užrašyta išreikštinė asimptotinės kovariacijų matricos išraiška geometriškai anizotropiniam eksponentiniam kovariacijos modeliui.
- Pasiūlytas neparametrinis testas erdvinių duomenų geometrinei anizotropijai nustatyti.
- Pilno populiacijų neapibrėžtumo atveju (angl. *case of complete uncertainty*) išvestos aktualiosios klasifikavimo į dvi klases rizikos (angl. *actual risk*) ir jos aproksimacijos formulės skaliariniu ir vektoriniu atvejais.
- Išvesta aktualiosios klasifikavimo į dvi klases rizikos formulė ir jos aproksimacija vektoriniu atveju GMRF stebiniui.
- Išvesta aktualiosios klasifikavimo rizikos vidurkio aproksimacijos formulė skaliariniam GGRF kelių klasių atveju.
- Išvesta išreikštinė aktualiosios klasifikavimo rizikos vidurkio aproksimacijos išraiška geometriškai anizotropiniam eksponentiniam kovariacijos modeliui.

Darbo struktūra ir apimtis

Disertaciją sudaro įvadas, trys skyriai, išvados bei literatūros sąrašas. Pirmajame skyriuje pristatomas disertacijoje naudojamas matematinis erdvinių duomenų modelis, apibūdinamos pagrindinės erdvinių procesų charakteristikos. Antrasis skyrius skirtas diskriminantinei analizei – čia pateikiami pagrindiniai disertacijos rezultatai. Trečiajame skyriuje pateikti skaitiniai eksperimentai ir siūlomų klasifikavimo procedūrų taikymai. Disertacija parašyta anglų kalba, bendra darbo apimtis 101 puslapis.

Disertacijos turinys

Čia pateiksime pagrindinius disertacijos rezultatus.

Erdvinių duomenų modeliavimas

Gauso atsitiktiniai laukai (angl. *Gaussian random fields* (GRF)) užima svarbią vietą erdvinėje statistikoje ir ypač geostatistikoje (Cressie [6]; Cressie, Wikle [7]; Diggle, Ribeiro [8]; Chiles, Delfiner [5]). Tradiciškai erdvinių duomenų modeliai yra skirstomi į dvi klases: geostatistiniai modeliai, susieti su tolydžiu erdviniu indeksu ir gardelės tipo modeliai, t.y. diskretaus erdvinio indekso modeliai. Remiantis šia klasifikacija, disertacijoje tiriami du GRF tipai: geostatistiniai Gauso atsitiktiniai laukai (GGRF) ir Gauso-Markovo atsitiktiniai laukai (GMRF).

Atsitiktinis laukas $Z(\mathbf{s})$ vadinamas Gauso, jei visiems $n \in \mathbb{N}$ ir bet kokiam erdvės taškų rinkiniui $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n \in \mathbb{R}^d$, jungtinis $\mathrm{Z}(\mathbf{s}_1), \dots, \mathrm{Z}(\mathbf{s}_n)$ skirstinys yra daugiamatis Gauso. Gauso-Markovo atsitiktinis laukas yra GRF su diskrečiu erdviniu indeksu ir galiojančia Markovo savybe apibrėžtoje kaimynystės sistemoje (schemoje).

Pirmajame disertacijos skyriuje pateiktos erdvinių procesų stacionarumo sąvokos, taip pat trumpai aprašyta erdvinių duomenų anizotropiškumo savybė, jos nustatymo metodai. Skyriaus pabaigoje

pateikiamas neparametrinis testas geometrinei anizotropijai nustatyti (30-32 p.). Taip pat šiame skyriuje aptariami parametrų vertinimo metodai, didesnį dėmesį skiriant maksimalaus tikėtinumo (MT) metodui – pateikti vidurkio ir dispersijos įvertiniai skirtingo parametrinio neapibrėžtumo situacijoms.

Dėl patogumo Mardia bei Marshal [16] teoremoje suformuluotas reguliarumo sąlygas GRF pažymėsime (MM).

Lema (Lemma 1.1, 29 p.). Tegul stebėjimų vektorius $\mathbf{Z}_n \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ ir galioja (MM) sąlygos. Tuomet MT įvertiniai $\widehat{\boldsymbol{\beta}}$ ir $\widehat{\boldsymbol{\theta}}$ tenkina šias savybes (kai $n \to \infty$):

$$\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta} \text{ ir } \widehat{\boldsymbol{\theta}} \sim AN_p\left(\boldsymbol{\theta}, \mathbf{J}_\theta^{-1}\right),$$

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta} \text{ ir } \widehat{\boldsymbol{\beta}} \sim AN_q\left(\boldsymbol{\beta}, \mathbf{J}_\beta^{-1}\right),$$

čia $\mathbf{J}_\theta$ ir $\mathbf{J}_\beta$ yra asimptotinės kovariacijų matricos, AN žymi asimptotinį normalumą, o $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ žymi konvergavimą pagal tikimybę.

Diskriminantinės analizės elementai

Tarkime stebimas atsitiktinis Gauso laukas $\{Z(\mathbf{s}): \mathbf{s} \in D \subset \mathbb{R}^d\}$ apibrėžtas tam tikroje tikimybinėje erdvėje $(\Omega, \boldsymbol{\mathcal{F}}, \mathbb{P})$ ir įgyjantis reikšmes iš $\mathcal{Z} = \mathbb{R}^p$, kur $\mathcal{Z}$ yra požymių erdvė. Pagrindinis tikslas yra priskirti GRF stebinį $Z_0 = Z(\mathbf{s}_0)$, $\mathbf{s}_0 \in D$ vienai iš m populiacijų (klasių, grupių) $\Omega_l, l = 1, \dots, m$ ir įvertinti klasifikavimo riziką.

Stebinio Z_0 modelis populiacijoje Ω_l yra $Z(\mathbf{s}) = \mu_l(\mathbf{s}) + \varepsilon(\mathbf{s})$, čia $\mu_l(\mathbf{s})$ yra vidurkio funkcija arba erdvinis trendas.

- Laikysime, kad vidurkio modeliai populiacijose Ω_l, $l = 1, \dots, m$, yra skirtingi parametriniai modeliai $\boldsymbol{\mu}_l(\mathbf{s}) = \mathbf{x}'(\mathbf{s})\boldsymbol{\beta}_l$, t.y. modeliai su vienodais regresoriais (kovariatėmis), bet skirtingais regresijos parametrais.
- Atsitiktinės klaidos $\varepsilon(\mathbf{s})$ generuojamos tuo pačiu Gauso erdviniu procesu $\{\varepsilon(\mathbf{s}): \mathbf{s} \in D \subset \mathbb{R}^d\}$ su nuliniu vidurkiu ir kovariacija, kuri yra arba Mattern tipo ar kita parametrinė

funkcija, susijusi su tolydžiu erdviniu indeksu, arba funkcija, susieta su gardele, kurioje įvestas priklausomybės grafas su apibrėžta kaimynystės schema.

$p_l(Z_0|\mathbf{t}), l = 1, \dots, m$, žymėsime stebinio Z_0, kai $\mathbf{T} = \mathbf{t}$, sąlyginį tankį populiacijoje Ω_l. Čia $\mathbf{T}$ žymi atsitiktinę mokymo imtį, o $\mathbf{t}$ – mokymo imties realizaciją. Nuostolius (angl. *loss function*), patiriamus stebinį iš $l - osios$ populiacijos priskiriant $k - ajai$ populiacijai, žymėsime $L(l, k), l, k = 1, \dots, m$.

Remsimės šiomis prielaidomis:

(A1) Apriorinės klasių tikimybės $\pi_l, l = 1, \dots, m$, $\sum_{l=1}^{m} \pi_l = 1$, yra žinomos, jos gali būti pastovios arba determinuotos erdvinio indekso ir mokymo imčių dydžių funkcijos.

(A2) Nuostolių funkcijos $L(l, k)$ reikšmės yra neneigiamos ir baigtinės bei nepriklauso nei nuo $\mathbf{s}_0$, nei nuo mokymo imties erdvinės konfigūracijos.

Klasifikavimo taisyklę, kai $\mathbf{T} = \mathbf{t}$, pažymėsime $D_t(\bullet): \mathcal{Z} \to \{1, \dots, m\}$. Tuomet tikėtini nuostoliai (angl. *expected loss*) arba sąlyginė rizika (angl. *conditional risk*), priėmus sprendimą stebinį Z_0 priskirti Ω_l klasei, yra

$$R_0\big(l, D_t(\bullet)\big) = \int_{\mathcal{Z}} L(l, D_t(Z_0)) p_l(Z_0|\mathbf{t}) dZ_0 = E_{Z_0|t,l}\big\{L\big(l, D_t(Z_0)\big)\big\}.$$

Tuo tarpu *bendra rizika* (angl. *total risk*) arba *bendri tikėtini nuostoliai* (angl. *total excepted losses*) reikš vidutinius nuostolius, patiriamus Z_0 klasifikavus naudojant sprendimo taisyklę $D_t(\bullet)$:

$$R_0\big(D_t(\bullet)\big) = \sum_{l=1}^{m} \pi_l R_0\big(l, D_t(\bullet)\big).$$

Taisyklė, minimizuojanti šią riziką, vadinama Bajeso klasifikavimo taisykle, ją žymėsime $D_t^B(\bullet)$. Klasifikuojant stebinį Z_0, ji užrašoma tokiu būdu:

$$D_t^B(Z_0) = arg \min_{\{k=1,\dots,m\}} \{\sum_{l=1}^{m} \pi_l p_l(Z_0|\mathbf{t}, \mathbf{\Psi}) L(l, k)\}. \qquad (1)$$

Čia $\mathbf{\Psi}$ žymi nežinomų populiacijos parametrų vektorių. Tuomet *Bajeso rizika*, susieta su Bajeso klasifikavimo taisykle (1), yra

$$R_0^B = R_0\left(D_t^B(\bullet)\right) = \sum_{l=1}^{m} \pi_l E_{Z_0|t,l} L\left(l, D_t^B(Z_0)\right). \quad (2)$$

Tegul $G_{lk}^B(Z_0)$ žymi porines Bajeso diskriminantines funkcijas (angl. *pairwise Bayes discriminant functions*)

$$G_{lk}^B(Z_0, \boldsymbol{\Psi}) = \sum_{j=1}^{m} \pi_j \, p_j(Z_0|t, \boldsymbol{\Psi}) d(j,l,k), \quad (3)$$

$d(j,l,k) = L(j,l) - L(j,k),\ l,k = 1,\dots,m$. Bajeso rizika, susijusi su šiomis diskriminantinėmis funkcijomis, užrašoma (žr. Dučinskas [9])

$$R_0^B(\boldsymbol{\Psi}) = \sum_{l,k=1}^{m} \pi_l E_{Z_0|t,l} L(l,k) \prod_{j=1, j\neq k} e(G_{kj}^B(Z_0, \boldsymbol{\Psi})).$$

Čia $e(x) = \begin{cases} 0, if\ x < 0 \\ 1, if\ x \geq 0 \end{cases}$ žymi Hevisaido funkciją.

Bajeso taisyklė yra optimali taisyklė minimizuojanti klasifikavimo riziką tais atvejais, kai populiacijos yra pilnai apibrėžtos ir yra žinoma nuostolių funkcija, tačiau praktiniuose uždaviniuose retai pasitaiko pilno populiacijų apibrėžtumo atvejų, tuomet nežinomos parametrų reikšmės yra pakeičiamos jų įvertiniais (parametrų įvertinių vektorių žymėsime $\widehat{\boldsymbol{\Psi}}$), kurie randami iš mokymo imties. Tokiu būdu gauta diskriminantinė funkcija vadinama *įterpta* (angl. *plug-in*) *Bajeso diskriminantine funkcija* (PBDF). Taigi nežinomas parametrų reikšmes pakeitus jų įvertiniais, gauname

$$\hat{G}_{lk}^B(Z_0) = G_{lk}^B\left(Z_0, \widehat{\boldsymbol{\Psi}}\right). \quad (4)$$

Apibrėžimas. *Aktualioji (įvertinta) klasifikavimo rizika* (angl. *actual risk*), kai $\mathbf{T} = \mathbf{t}$, susieta su PBDF (4), yra apibrėžiama

$$R_0^B(\widehat{\boldsymbol{\Psi}}) = \sum_{l,k=1}^{m} \pi_l E_{Z_0|\mathbf{t},l} L(l,k) \prod_{j=1, j\neq k} e(\hat{G}_{kj}^B(Z_0)).$$

Apibrėžimas. Aktualiosios klasifikavimo rizikos vidurkis pagal mokymo imties $\mathbf{T}$ skirstinį vadinamas *aktualiosios klasifikavimo rizikos vidurkiu* (angl. *expected risk* (*ER*)) $ER = E_T\left(R_0^B(\widehat{\boldsymbol{\Psi}})\right)$.

Dauguma disertacijoje pateiktų rezultatų yra išvesti dviejų klasių atveju, todėl žemiau yra pateiktos diskriminantinės funkcijos bei klasifikavimo rizikos formulių išraiškos šiam atvejui.

Bajeso klasifikavimo taisyklė dviejų klasių atveju įgauna tokį pavidalą

$$D_t^B(Z_0, \mathbf{\Psi}) = arg \max_{\{l=1,2\}} \{g_l p_l(Z_0|\mathbf{t}, \mathbf{\Psi})\},$$

čia $g_l = \pi_l\big(L(l, 3-l) - L(l,l)\big),\ l = 1{,}2.$

Disertacijoje dviejų klasių atveju naudojama Bajeso diskriminantinė funkcija, apibrėžiama kaip sąlyginių skirstinių santykio logaritmas. Kadangi turint dvi populiacijas yra reikalinga tik viena diskriminantinė funkcija, todėl žymėsime ją be indeksų, nusakančių klases, t.y.:

$$W^B(Z_0, \mathbf{\Psi}) = ln\left(\frac{p_1(Z_0|\mathbf{t},\mathbf{\Psi})}{p_2(Z_0|\mathbf{t},\mathbf{\Psi})}\right) + \gamma^*,$$

čia $\gamma^* = \ln\left(\frac{g_1}{g_2}\right),\ g_l = \pi_l\big(L(l, 3-l) - L(l,l)\big), l = 1{,}2.$ Remiantis šia taisykle, stebinys Z_0, kai $\mathbf{T} = \mathbf{t}$, yra priskiriamas populiacijai Ω_1, jeigu $W^B(Z_0, \mathbf{\Psi}) \geq 0$ arba populiacijai Ω_2 priešingu atveju.

Bajeso rizika dviejų klasių atveju gali būti įvertinta pagal

$$R_0^B(\mathbf{\Psi}) = \sum_{l=1}^{2}(\pi_l L(l,l) + g_l PM_l),$$

kur $PM_l = P_l((-1)^l W^B(Z_0, \mathbf{\Psi}) > 0)$ yra klaidingo klasifikavimo tikimybė.

Visos aukščiau aprašytos formulės galioja esant bendrai nuostolių funkcijai, tačiau dažnai susiduriama su atskiru atveju, t.y. *0-1 nuostolių funkcija* (angl. *zero-one loss function*), kuri apibrėžiama taip: $L(l,k) = 1 - \delta_{lk}$, kur δ_{lk} yra Kronekerio delta.

Tuomet klasifikavimo rizika tampa *klaidingo klasifikavimo tikimybe* (angl. *probability of misclassification*) (Dučinskas [9]) arba klasifikavimo klaida (angl. *error rate*). Tokiu atveju porinės Bajeso diskriminantinės funkcijos, apibrėžtos (3), įgauna paprastesnį pavidalą

$$G_{lk}^{B}(Z_0, \boldsymbol{\Psi}) = \pi_l p_l(Z_0|\mathbf{t}, \boldsymbol{\Psi}) - \pi_k p_k(Z_0|\mathbf{t}, \boldsymbol{\Psi})$$

arba, naudojant logaritmuotą jos išraišką, gauname

$$W_{lk}^{B}(Z_0, \boldsymbol{\Psi}) = \ln\left(\frac{p_l(Z_0|\mathbf{t}, \boldsymbol{\Psi})}{p_k(Z_0|\mathbf{t}, \boldsymbol{\Psi})}\right) + \gamma_{lk},$$

$\gamma_{lk} = \ln\left(\frac{\pi_l}{\pi_k}\right), l, k = 1, \dots, m, k \neq l.$

Naudojant pastarąją diskriminantinę funkciją, stebinys Z_0, kai $\mathbf{T} = \mathbf{t}$, priskiriamas populiacijai Ω_l, jeigu $W_{lk}^{B}(Z_0, \boldsymbol{\Psi}) \geq 0$, visiems $l, k = 1, \dots, m, k \neq l$. Tuomet klaidingo klasifikavimo tikimybė apskaičiuojama pagal formulę:

$$P_0^{B}(\boldsymbol{\Psi}) = 1 - \sum_{l=1}^{m} \pi_l PC_l,$$

kur $PC_l = P_l(W_{lk}^{B}(Z_0, \boldsymbol{\Psi}) \geq 0), l = 1, \dots, m, l \neq k$ yra teisingo klasifikavimo tikimybė.

Dviejų klasių su 0-1 nuostoliais atveju Bajeso klasifikavimo taisyklė yra ekvivalenti maksimalios aposteriorinės klasės tikimybės taisyklei.

Pagrindiniai rezultatai

Skaliarinio GGRF stebinio klasifikavimo uždavinys į vieną iš dviejų populiacijų

Pagrindinis šio skyrelio tikslas yra priskirti GGRF stebinį vienai iš dviejų populiacijų, kai nežinomi visi populiacijų parametrai, ir įvertinti klasifikavimo riziką. Tai yra [10], [11] publikacijose pateiktų rezultatų išplėtimas. Minėtose publikacijose analizuojama klasifikavimo klaida, o ne rizika, be to, išvestos formulės yra nepilno parametrinio neapibrėžtumo atveju.

Šiame skyriuje laikoma, kad visi populiacijų parametrai $\boldsymbol{\Psi} = (\boldsymbol{\beta}', \boldsymbol{\theta}')'$ yra nežinomi. Čia $\boldsymbol{\beta}$ yra vidurkio parametrų vektorius, o $\boldsymbol{\theta}$ žymi kovariacijų funkcijos parametrų vektorių. Analizuojamu atveju pastarąjį sudaro 5 parametrai: grynuolis (angl. *nugget*), slenkstis (angl. *sill*) arba mastelio parametras, plotis (angl. *range*) bei

geometrinės anizotropijos parametrai – anizotropijos santykis (angl. *anisotropy ratio*) ir anizotropijos kampas (angl. *anisotropy angle*), t.y.:

$$\boldsymbol{\theta} = (\tau^2, \sigma^2, \alpha, \lambda, \varphi)'.$$

$\mathbf{S}_n = \{\mathbf{s}_i \in D, i = 1, \dots, n\}$ žymėsime rinkinį erdvės taškų, kuriuose stebime mokymo imtį $\mathbf{T} = (\mathbf{T}_1', \mathbf{T}_2')' = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$. $\mathbf{S}_n = \mathbf{S}^{(1)} \cup \mathbf{S}^{(2)}$ sudarytas iš dviejų poaibių, kur $\mathbf{S}^{(l)}$ sudaro n_l stebėjimų iš populiacijos Ω_l, $l = 1,2$. Mokymo imties modelis užrašomas

$$\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E},$$

kur $\mathbf{X} = \bigoplus_{l=1}^{2} \mathbf{X}_l$ yra plano matrica, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ vidurkio parametrų vektorius ir $\mathbf{E}$ yra atsitiktinių klaidų vektorius, kurio skirstinys yra $N_n(0, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ žymėsime kovariacijų matricą tarp mokymo imties komponenčių, o $\mathbf{c}_0$ - kovariacijų vektorių tarp $\mathbf{T}$ ir Z_0.

Pilno populiacijų apibrėžtumo atveju Bajeso diskriminantinė funkcija (BDF) yra

$$W^B(Z_0, \boldsymbol{\Psi}) = (Z_0 - (\mu_{1t} + \mu_{2t})/2)'\,(\mu_{1t} - \mu_{2t})/\sigma_t^2 + \gamma^*,$$

čia μ_{lt} ir σ_t^2 yra sąlyginis vidurkis ir sąlyginė dispersija:

$$\mu_{lt} = E(Z_0|\mathbf{T} = \mathbf{t}; \Omega_l) = \mathbf{x}_0'\boldsymbol{\beta}_l + \boldsymbol{\alpha}_0'(\mathbf{t} - \mathbf{X}\boldsymbol{\beta}),\ l = 1,2,$$

$$\sigma_t^2 = var(Z_0|\mathbf{T} = \mathbf{t}; \Omega_l) = C(\mathbf{0}) - \mathbf{c}_0'\boldsymbol{\Sigma}^{-1}\mathbf{c}_0,$$

$$\mathbf{x}_0' = \big(x_1(\mathbf{s}_0), \dots,\ x_q(\mathbf{s}_0)\big) \text{ ir } \boldsymbol{\alpha}_0' = \mathbf{c}_0'\boldsymbol{\Sigma}^{-1}.$$

Įstačius sąlyginio vidurkio ir dispersijos išraiškas į BDF gauname

$$W^B(Z_0, \boldsymbol{\Psi}) = (Z_0 - \boldsymbol{\alpha}_0'(\mathbf{t} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{x}_0'\mathbf{I}_+\boldsymbol{\beta}/2)'(\,\mathbf{x}_0'\mathbf{I}_-\boldsymbol{\beta})/\sigma_t^2 + \gamma^*, \quad (5)$$

čia $\mathbf{I}_+ = (\mathbf{I}_q, \mathbf{I}_q)$, $\mathbf{I}_- = (\mathbf{I}_q, -\mathbf{I}_q)$.

Lema (Lema 2.1, 42 p.). Galiojant prielaidoms (A1) ir (A2) *Bajeso rizika* susieta su BDF (5) yra

$$R_0^B(\boldsymbol{\Psi}) = \sum_{l=1}^{2}\{\pi_l L(l,l) + g_l \Phi(-d/2 + (-1)^l\gamma^*/d)\}.$$

Čia $\Phi(\cdot)$ žymi standartinį normalųjį skirstinį, $d^2 = (\mu_{1t} - \mu_{2t})^2/\sigma_t^2$ yra Mahalanobiso atstumo tarp sąlyginių skirstinių kvadratas. Šios lemos įrodymas remiasi Gauso skirstinio savybėmis.

Pakeitus vidurkį ir dispersiją jų MT įvertiniais, gauname įterptą Bajeso diskriminantinę funkciją (PBDF)

$$W^B\left(Z_0, \widehat{\mathbf{\Psi}}\right) = \left(Z_0 - \widehat{\boldsymbol{\alpha}}_0'\left(\mathbf{t} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) - \mathbf{x}_0'\mathbf{I}_+\widehat{\boldsymbol{\beta}}/2\right)' \times$$

$$\times \left(\mathbf{x}_0'\mathbf{I}_-\widehat{\boldsymbol{\beta}}\right)/\widehat{K}\hat{\sigma}^2 + \gamma^*. \qquad (6)$$

Ši funkcija taikoma skaliarinio GGRF stebinio priskyrimui į vieną iš dviejų klasių.

Lema (Lema 2.2, 43 p.). Aktualioji funkcijos $W^B\left(Z_0, \widehat{\mathbf{\Psi}}\right)$ klasifikavimo rizika yra

$$R_0^B(\widehat{\mathbf{\Psi}}) = \sum_{l=1}^{2}\left\{\pi_l L(l,l) + g_l \Phi(\hat{Q}_l)\right\}. \qquad (7)$$

Čia

$$\hat{Q}_l = (-1)^l\left((a_l - \hat{b})\mathrm{sgn}(\mathbf{x}_0'\mathbf{I}_-\widehat{\boldsymbol{\beta}})/\sigma_t + \gamma^*\hat{\sigma}_t^2/\left|\mathbf{x}_0'\mathbf{I}_-\widehat{\boldsymbol{\beta}}\right|\sigma_t\right),$$

$$a_l = \mathbf{x}_0'\boldsymbol{\beta}_l + \boldsymbol{\alpha}_0'(\mathbf{t} - \mathbf{X}\boldsymbol{\beta}), l = 1{,}2,$$

$$\hat{b} = \widehat{\boldsymbol{\alpha}}_0'\left(\mathbf{t} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) + \mathbf{x}_0'\mathbf{I}_+\widehat{\boldsymbol{\beta}}/2.$$

Aktualiosios klasifikavimo rizikos vidurkio formulė gaunama skaičiuojant $R_0^B\left(\widehat{\mathbf{\Psi}}\right)$ vidurkį pagal mokymo imties **T** skirstinį

$$ER = E_T\left(R_0^B\left(\widehat{\mathbf{\Psi}}\right)\right) = E_T\left\{\sum_{l=1}^{2}\left(\pi_l L(l,l) + g_l\Phi(\hat{Q}_l)\right)\right\}.$$

Ši charakteristika įvertina diskriminantinės funkcijos efektyvumą ir kokybę, tačiau dažniausiai jos analitinės išraiškos nepavyksta gauti. Dėl šios priežasties tenka ieškoti būdų, kaip ją aproksimuoti. Šiame darbe siūloma aktualiosios klasifikavimo rizikos vidurkio aproksimacijos formulė, pagrįsta asimptotiniu skleidiniu. Parametrų vertinimui naudojamas maksimalaus tikėtinumo (MT) metodas. Gausiniu atveju MT parametrų įvertinių savybės įrodytos Mardia ir Marshall [16] teoremoje.

Apibrėžimas. *Aktualiosios klasifikavimo rizikos vidurkio aproksimacija* (AER) vadinsime $R_0^B(\widehat{\boldsymbol{\Psi}})$ Teiloro eilutės asimptotinį skleidinį iki antrųjų išvestinių taškų $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ aplinkoje, iš jo pašalinus liekamąjį narį.

Tolimesniems rezultatams išvesti remsimės šia prielaida:

(A3) Mokymo imtis $\mathbf{T}$ ir įvertinys $\widehat{\boldsymbol{\theta}}$ yra statistiškai nepriklausomi.

Šia prielaida vadovavosi daugelis autorių (pvz., Zhu ir Stein [25]), kadangi Abt [1] įrodė, kad MSPE aproksimacijos, laikant, kad $\mathbf{T}$ ir $\widehat{\boldsymbol{\theta}}$ tarpusavyje koreliuoja, tikslesnių rezultatų neduoda.

Teorema (Teorema 2.1, 45 p.). Tarkime, kad fokalinis stebinys Z_0 yra klasifikuojamas pagal PBDF (6) ir laikykime, kad galioja (MM) sąlygos bei (A3) prielaida. Tuomet aktualiosios klasifikavimo rizikos vidurkio aproksimacija yra

$$AER = \sum_{l=1}^{2} g_l \Phi(Q_l) + g_1 \varphi(Q_1) d(K_\beta + K_\theta)/2\sigma_t^2, \qquad (8)$$

$$K_\beta = \boldsymbol{\Lambda}' \mathbf{J}_\beta^{-1} \boldsymbol{\Lambda}, \boldsymbol{\Lambda}' = \boldsymbol{\alpha}_0' \mathbf{X} - \mathbf{x}_0' (\mathbf{I}_+/2 + \gamma^* \mathbf{I}_-/d^2), \mathbf{J}_\beta = \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X},$$

$$K_\theta = tr\left(\boldsymbol{\Sigma} \mathbf{A}_\theta \mathbf{J}_\theta^{-1} \mathbf{A}_\theta'\right) + (\gamma^*)^2 \mathbf{s}_\theta' \mathbf{J}_\theta^{-1} \mathbf{s}_\theta / d^2 \sigma_t^2,$$

$$\mathbf{A}_\theta = \partial \widehat{\boldsymbol{\alpha}}_0 / \partial \widehat{\boldsymbol{\theta}}', \mathbf{s}_\theta = (\widehat{\sigma}_t^2)_\theta^{(1)}.$$

Čia $\varphi(\cdot)$ – standartinio normaliojo skirstinio tankio funkcija.
Teoremos įrodymas pateiktas disertacijos 45-48 p., taip pat [A11].

Skaliarinio GGRF aktualiosios klasifikavimo rizikos vidurkio aproksimacijos išreikštinė forma eksponentiniam kovariacijos modeliui

Aktualiosios klasifikavimo rizikos vidurkio aproksimacijos formulėje (8) $\mathbf{A}_\theta$, $\mathbf{J}_\theta^{-1}$ ir $\mathbf{s}_\theta$ yra dalinės matricų ir vektorių išvestinės nežinomų kovariacijos parametrų $\boldsymbol{\theta} = (\tau^2, \sigma^2, \alpha, \lambda, \varphi)'$ atžvilgiu. Norint AER realizuoti praktiškai, reikia turėti išreikštines šių išvestinių formas, kurios priklauso nuo pasirinkto kovariacijų funkcijos modelio. Disertacijoje (49-51 p.) yra užrašytos $\mathbf{A}_\theta$, $\mathbf{J}_\theta^{-1}$ ir

$\mathbf{s}_\theta$ išreikštinės formos geometriškai anizotropinės eksponentinės kovariacijų funkcijos modeliui.

Skaliarinio GGRF stebinio klasifikavimo uždavinys į vieną iš keleto populiacijų

Šiuo atveju sprendžiamas stebinio Z_0 klasifikavimo į vieną iš keleto populiacijų ($m > 2$) uždavinys bei tiriama klaidingo klasifikavimo tikimybė, t.y. taikoma *0-1 nuostolių funkcija*. Pasirinkta faktorizuota, be grynuolio kovariacijų funkcija $C(h) = \sigma^2 r(h)$, kur σ^2 yra nežinomas mastelio parametras arba slenkstis, o $r(h)$ yra žinoma erdvinių koreliacijų funkcija. Taigi tiriama dalinio populiacijų neapibrėžtumo situacija.

Kai klasių skaičius $m > 2$ sudaroma $m - 1$ Bajeso diskriminantinė funkcija

$$W_{lk}^B(Z_0, \mathbf{\Psi}) = (Z_0 - \boldsymbol{\alpha}_0'(\mathbf{t} - \mathbf{X}\boldsymbol{\beta}) - \mu_{lk})' d_{lk}/\sigma\sqrt{K} + \gamma_{lk},$$

kur $\mu_{lk} = \mathbf{x}_0'(\boldsymbol{\beta}_l + \boldsymbol{\beta}_k)/2$, $\boldsymbol{\alpha}_0' = \mathbf{c}_0'\boldsymbol{\Sigma}^{-1}$ ir d_{lk} yra sąlyginis Mahalanobiso atstumas $d_{lk} = (\mu_{lt} - \mu_{kt})/\sigma_t$. $K = 1 - \mathbf{r}_0'\mathbf{R}^{-1}\mathbf{r}_0$, $\mathbf{r}_0$ yra erdvinių koreliacijų tarp mokymo imties $\mathbf{T}$ ir stebinio Z_0 vektorius, o $\mathbf{R}$ žymi erdvinių koreliacijų tarp $\mathbf{T}$ komponenčių matricą.

Sprendimo priėmimo taisyklė: Z_0 priskiriamas populiacijai Ω_l, jei

$$W_{lk}^B(Z_0, \mathbf{\Psi}) \geq 0, \text{ visiems } l, k = 1, \dots, m, \ k \neq l.$$

Pakeitus sąlyginio vidurkio ir dispersijos išraiškas jų įvertiniais, gauname PBDF išraišką

$$W_{lk}^B(Z_0, \widehat{\mathbf{\Psi}}) = \left(Z_0 - \boldsymbol{\alpha}_0'\left(\mathbf{t} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) - \hat{\mu}_{lk}\right)' \hat{d}_{lk}/\hat{\sigma}\sqrt{K} + \gamma_{lk}. \quad (9)$$

Lema (Lema 2.3, 52 p.). Gauso populiacijų klaidingo klasifikavimo tikimybė (klasifikavimo klaida), pagrįsta Bajeso klasifikavimo taisykle, kai $m > 2$, yra

$$P_0^B(\mathbf{\Psi}) = 1 - \sum_{l=1}^m \pi_l \int_{B_l} \varphi(u) du,$$

$$B_l = \{u: u \in R^1, \ d_{lk}u + d_{lk}^2/2 + \gamma_{lk} \geq 0; l = 1, \dots, m, \ k \neq l\}.$$

Lema (Lema 2.4, 53 p.). Klaidingo klasifikavimo tikimybė susieta su PBDF, kai $m > 2$, yra

$$P_0^B(\widehat{\mathbf{\Psi}}) = 1 - \sum_{l=1}^{m} \pi_l \int_{A_l} \varphi(u)du,$$

$$A_l = \{u: u \in R^1,\ \hat{d}_{lk}u + \left(\mu_k + \mathbf{r}_0'\mathbf{R}^{-1}\mathbf{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \hat{\mu}_{lk}\right)\hat{d}_{lk}/\sigma\sqrt{K} +$$

$$+ \gamma_{lk}\hat{\sigma}/\sigma \geq 0; l = 1, \dots, m,\ k \neq l\}.$$

Apibrėžimas. Aktualiosios klaidingo klasifikavimo tikimybės vidurkis mokymo imties **T** skirstinio atžvilgiu vadinamas *aktualiosios klaidingo klasifikavimo tikimybės vidurkiu* (angl. *expected error* (*EER*))

$$EER = E_T\left(P_0^B(\widehat{\mathbf{\Psi}})\right).$$

Apibrėžimas. Aktualiosios klaidingo klasifikavimo tikimybės vidurkio aproksimacija (AEER) vadinsime $P_0^B(\widehat{\mathbf{\Psi}})$ Teiloro eilutės asimptotinį skleidinį iki antrųjų išvestinių taške $\widehat{\mathbf{\Psi}} = \mathbf{\Psi}$, iš kurio pašalintas liekamasis narys.

Teorema (Teorema 2.2, 58 p.). Tarkime, kad stebinys Z_0 klasifikuojamas pagal PBDF (9) ir laikykime galiojančiomis sąlygas (B1)-(B3) (žr. 56 p.). Tuomet aktualiosios klaidingo klasifikavimo tikimybės vidurkio aproksimacijos formulė yra

$$AEER = P_0^B(\mathbf{\Psi}) + C/2 + D,$$

čia $C = \sum_{l=1}^{m}\sum_{k>l}\pi_l\varphi\left(\frac{\gamma_{lk}}{d_{lk}} + \frac{d_{lk}}{2}\right)d_{lk}\mathbf{\Lambda}_{lk}'\mathbf{R}_\beta\mathbf{\Lambda}_{lk}\prod_{j\neq l,k}e(w_{lkj})/K,$

$$D = \sum_{l=1}^{m}\sum_{k>l}\frac{\gamma_{lk}^2}{n-2q}\pi_l\varphi\left(\frac{\gamma_{lk}}{d_{lk}} + \frac{d_{lk}}{2}\right)\prod_{j\neq l,k}e(w_{lkj})/d_{lk}.$$

Įrodymas yra pagrįstas $P_0^B(\widehat{\mathbf{\Psi}})$ Teiloro eilutės skleidiniu taške $\widehat{\mathbf{\Psi}} = \mathbf{\Psi}$ iki antrųjų išvestinių.

Vektorinio GGRF stebinio klasifikavimo uždavinys į vieną iš dviejų populiacijų

Vektorinio GGRF $\{\mathbf{Z}(\mathbf{s})\colon \mathbf{s} \in D \subset R^2\}$ stebinio klasifikavimo uždavinys buvo sprendžiamas [12], tačiau šioje publikacijoje analizuojama klasifikavimo klaida, o ne klasifikavimo rizika. Be to, tiriama dalinio parametrinio neapibrėžtumo situacija. Disertacijoje šie rezultatai išplėsti iki pilno parametrinio neapibrėžtumo bei taikoma bendra nuostolių funkcija, t.y. tiriama klasifikavimo rizika.

Stebinio $\mathbf{Z}(\mathbf{s})$ populiacijoje Ω_l modelis yra

$$\mathbf{Z}(\mathbf{s}) = \mathbf{B}_l'\mathbf{x}(\mathbf{s}) + \boldsymbol{\varepsilon}(\mathbf{s}),$$

$\mathbf{x}(\mathbf{s})$ yra $q \times 1$ regresorių vektorius, $\mathbf{B}_l$ yra $q \times p$ parametrų matrica, o atsitiktinė klaida generuojama $p - mačiu$ nulinio vidurkio GGRF $\{\boldsymbol{\varepsilon}(\mathbf{s})\colon \mathbf{s} \in D\}$ su faktorizuota kovariacijų funkcija be grynuolio, kuri visiems $\mathbf{s}, \mathbf{u} \in D$ apibrėžiama $\mathrm{cov}\{\boldsymbol{\varepsilon}(\mathbf{s}), \boldsymbol{\varepsilon}(\mathbf{u})\} = r(\mathbf{s} - \mathbf{u})\mathbf{S}$. Čia $r(\mathbf{s} - \mathbf{u})$ erdvinė koreliacijų funkcija, o $\mathbf{S}$ yra $p \times p$ dydžio kovariacijų matrica tarp stebimų požymių. Reikėtų akcentuoti, kad pasirinkta faktorizuota kovariacijų funkcija sąlygoja šiek tiek siauresnį parametrinį neapibrėžtumą nei buvo analizuojamas skaliariniu atveju.

Stebima mokymo imtis $\mathbf{T}$, kurios modelis yra $\mathbf{T} = \mathbf{XB} + \mathbf{E}$. Čia $\mathbf{X}$ yra $n \times 2q$ plano matrica, $\mathbf{B}' = (\mathbf{B}_1', \mathbf{B}_2')$ – $p \times 2q$ vidurkio prametrų matrica ir $\boldsymbol{E}$ yra $n \times p$ atsitiktinių paklaidų matrica, kurios skirstinys $\boldsymbol{E} \sim N_{n\times p}(\mathbf{0}, \mathbf{R} \otimes \mathbf{S})$. Siekiama fokalinį stebinį $\mathbf{Z}_0$ priskirti vienai iš dviejų klasių. Sąlyginis $\mathbf{Z}_0$ skirstinys, kai $\mathbf{T} = \mathbf{t}$, yra Gauso su sąlyginiu vidurkiu ir sąlygine dispersija

$$\boldsymbol{\mu}_{lt} = \mathbf{B}_l'\mathbf{x}_0 + \boldsymbol{\alpha}_0'(\mathbf{t} - \mathbf{XB}),$$

$$\mathbf{S}_t = K\mathbf{S},\ K = 1 - \mathbf{r}_0'\mathbf{R}^{-1}\mathbf{r}_0.$$

MT įvertiniais pakeitus nežinomas parametrų reikšmes, gaunama PBDF išraiška

$$W^B\left(Z_0, \widehat{\boldsymbol{\Psi}}\right) = \left(\mathbf{Z}_0 - \widehat{\boldsymbol{\alpha}}_0'\left(\mathbf{t} - \mathbf{X}\widehat{\mathbf{B}}\right) - \mathbf{x}_0'\mathbf{I}_+\widehat{\mathbf{B}}/2\right)' \times$$

$$\times \widehat{\mathbf{S}}^{-1}(\mathbf{x}_0'\mathbf{I}_-\widehat{\mathbf{B}})/\widehat{K} + \gamma^*. \qquad (10)$$

Tuomet aktualioji klasifikavimo rizika, atitinkanti šią PBDF, yra

$$R_0^B(\widehat{\mathbf{\Psi}}) = \sum_{l=1}^{2}\{\pi_l L(l,l) + g_l\Phi(\hat{Q}_l)\},$$

$$\hat{Q}_l = (-1)^l\left((\mathbf{a}_l - \hat{\mathbf{b}})\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}}'\mathbf{I}_-'\mathbf{x}_0 + \gamma^*\widehat{K}\right)/\sqrt{\mathbf{x}_0'\mathbf{I}_-\widehat{\mathbf{B}}\widehat{\mathbf{S}}^{-1}\mathbf{S}\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}}'\mathbf{I}_-'\mathbf{x}_0 K},$$

$$\mathbf{a}_l = \mathbf{x}_0'\mathbf{B}_l + \boldsymbol{\alpha}_0'(\mathbf{t} - \mathbf{X}\mathbf{B}),\ l = 1{,}2,\ \mathbf{b} = \widehat{\boldsymbol{\alpha}}_0'\left(\mathbf{t} - \mathbf{X}\widehat{\mathbf{B}}\right) + \mathbf{x}_0'\mathbf{I}_+\widehat{\mathbf{B}}/2.$$

Teorema (Teorema 2.3, 62 p.). Tarkime, kad siekiama klasifikuoti fokalinį stebinį $\mathbf{Z}_0$, naudojant PBDF (10), bei laikykime, kad galioja (MM) sąlygos ir (A3) prielaida. Tuomet *aktualiosios klasifikavimo rizikos vidurkio aproksimacijos* išraiška yra

$$AER = R_0^B(\mathbf{\Psi}) + g_1\varphi_1\{\mathbf{\Lambda}'\mathbf{R}_0\mathbf{\Lambda}d/K + (p-1)\mathbf{x}_0'\mathbf{I}_-\mathbf{R}_0\mathbf{I}_-'\mathbf{x}_0/Kd + \\ +\mathrm{tr}(\mathbf{F}_1\mathbf{V}_\eta) + \mathrm{tr}(\mathbf{F}_2\mathbf{V}_\vartheta) + 2\mathrm{tr}(\mathbf{F}_3\mathbf{V}_{\eta\vartheta})\}/2,$$

čia $\varphi_1 = \varphi(-d/2 - \gamma^*/d)$, kur $d^2 = (\boldsymbol{\mu}_{1t} - \boldsymbol{\mu}_{2t})'\mathbf{S}_t^{-1}(\boldsymbol{\mu}_{1t} + \boldsymbol{\mu}_{2t})$ – sąlyginis Mahalanobiso atstumo kvadratas.

$$\mathbf{\Lambda} = \mathbf{X}'\boldsymbol{\alpha}_0 - (\mathbf{I}_+'/2 + \gamma^*\mathbf{I}_-'/d^2)\mathbf{x}_0,\ \mathbf{R}_0 = \left(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\right)^{-1}.$$

$$\mathbf{F}_1 = \mathbf{D}_p'\Big((\mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1} \otimes \mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1})(\gamma^*)^2K/\Delta^4 + \\ +\mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1} \otimes \left(\mathbf{S}^{-1} - \mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1}/d^2\right)\Big)\mathbf{D}_p/(d\sqrt{K}),$$

$$\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2.$$

$$\mathbf{F}_2 = (tr(\mathbf{A}_\vartheta'\mathbf{R}\mathbf{A}_\vartheta\mathbf{V}_\vartheta)\Delta^2 + (\gamma^*)^2\mathbf{K}_\vartheta'\mathbf{V}_\vartheta\mathbf{K}_\vartheta)/\Delta^3\sqrt{K},$$

$$\mathbf{F}_3 = \mathbf{D}_p'(\mathbf{S}^{-1}\Delta\boldsymbol{\mu} \otimes \mathbf{S}^{-1}\Delta\boldsymbol{\mu})(\gamma^*)^2\mathbf{K}_\vartheta/\Delta^4\sqrt{K}.$$

$\mathbf{D}_p$ žymi $p^2 \times (p(p+1)/2)$ dydžio dublikacijos matricą (angl. *duplication matrix*), $\mathbf{A}_\vartheta = \partial\widehat{\boldsymbol{\alpha}}_0/\partial\boldsymbol{\vartheta}'$ $(\boldsymbol{\vartheta} = (\alpha, \lambda, \varphi)')$ yra $n \times k$ dalinių išvestinių taške $\widehat{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}$ matrica, o $\mathbf{K}_\vartheta = \partial K/\partial\boldsymbol{\vartheta}'$ – $k \times 1$ dydžio vektorius sudarytas iš dalinių išvestinių taške $\widehat{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}$. $\mathbf{V}_\vartheta$ – informacijos matricos atvirkštinė.

Vektorinio GMRF stebinio klasifikavimo uždavinys į vieną iš dviejų populiacijų

Šiame skyriuje sprendžiamas vektorinio Gauso-Markovo atsitiktinio lauko (GMRF) $\{\mathbf{Z}(\mathbf{s})\colon \mathbf{s} \in D \subset R^2\}$ stebinio klasifikavimo uždavinys į vieną iš dviejų populiacijų su *0-1 nuostolių funkcija.*

Kaip jau buvo minėta, geostatistiniai modeliai taikomi tolydiems erdviniams procesams su tiesiogiai aprašoma parametrine kovariacijų funkcija. Tokie modeliai reikalauja daug kompiuterinio laiko operacijoms su kovariacijų matricomis atlikti (Lindgren ir kt. [14]). Tuo tarpu GMRF modeliai yra tiesiogiai aprašomi retomis (angl. *sparse*) tikslumo matricomis (angl. *precision matrix*). Erdviniai ryšiai yra modeliuojami pagal tam tikrą kaimynystės schemą, tokiu būdu ženkliai sumažinamas skaičiavimo uždavinių sudėtingumas (operacijų skaičius). Taigi pagrindinis skirtumas, lyginant GGRF ir GMRF, yra kovariacijų matricos struktūra. Naudosime de Oliveiros ir Ferreiros [18] pasiūlytą kovariacijų funkcijos parametrizavimo formą, kuri pasirinkta dėl patogios klasifikavimui išraiškos bei pilnai ištirtų MT įvertinių savybių baigtiniam n.

Stebinio $\mathbf{Z}(\mathbf{s})$ modelis apibrėžiamas taip pat kaip ir vektorinio GGRF atveju, t.y. $\mathbf{Z}(\mathbf{s}) = \mathbf{B}_l'\mathbf{x}(\mathbf{s}) + \boldsymbol{\varepsilon}(\mathbf{s})$. Pirmoji komponentė sutampa su GGRF modelio, o $\boldsymbol{\varepsilon}(\mathbf{s})$ yra generuojama ant gardelės $p - mačiu$ nulinio vidurkio daugiamačiu GMRF $\{\boldsymbol{\varepsilon}(\mathbf{s})\colon \mathbf{s} \in D\}$ ir yra susijusi su tam tikra kaimynystės schema.

$\mathbf{S}_n = \{\mathbf{s}_i \in D, i = 1, \dots, n\}$ žymėsime erdvės taškų rinkinį, kuris formuoja taisyklingą arba netaisyklingą gardelę, ir jame stebime mokymo imtį $\mathbf{T}$. $\mathbf{S}_n = \mathbf{S}^{(1)} \cup \mathbf{S}^{(2)}$ sudarytas iš dviejų poaibių, kur $\mathbf{S}^{(l)}$ sudaro n_l stebėjimų iš populiacijos Ω_l, $l = 1{,}2$. $\mathbf{S}_n$ elementus sieja kaimynystės schema $N = \{N_l\colon l = 1, \dots, n\}$, o $\mathbf{S}_0^n = \mathbf{S}_n \cup \{\mathbf{s}_0\}$ – $N^0 = \{N_l^0\colon l = 0{,}1, \dots, n\}$, čia N_l – taško $\mathbf{s}_l$ kaimynų aibė.

Remiantis de Oliveira ir Ferreira [18], sudaromos svorių matricos, su kurių pagalba erdviniai svoriai susieti kaimynystės schema bus įtraukti į kovariacijų matricą.

$\mathbf{H}^0 = (h_{ij}^0 : i,j = 0,1,\dots,n)$ ir $\mathbf{H} = (h_{ij} : i,j = 1,\dots,n)$, kurių elementai yra

$$h_{ij}^0 = \begin{cases} h_i^0 \quad if \quad i = j \\ -w_{ij} \;\; if \; i \in N_j^0 \\ 0 \quad kitu\; atveju \end{cases}, \text{ kur } h_i^0 = \sum_{j \in N_i^0} w_{ij},\; i,j = 0,1,\dots,n,$$

$$h_{ij} = \begin{cases} h_i \quad if \quad i = j \\ -w_{ij} \;\; if \; i \in N_j \\ 0 \quad kitu\; atveju \end{cases}, \text{ kur } h_i = \sum_{j \in N_i} w_{ij},\; i,j = 1,\dots,n.$$

Čia $w_{lk} > 0$ $(w_{kl} = w_{lk})$ yra erdviniai svoriai, kurie nusako panašumą tarp $l - ojo$ ir $k - ojo$ erdvės taškų. Tolimesniuose skaičiavimuose naudosime vektorius, sudarytus iš svorių: $\mathbf{w}_0' = (w_{01},\dots,w_{0n})$ ir $\mathbf{w}_i' = (w_{i0},\dots,w_{ii-1},w_{ii+1},\dots,w_{in})$, $i = 1,\dots,n$.

Mokymo imties skirstinys yra Gauso, jį galima užrašyti dviem būdais:

- vektorizuotai mokymo imčiai $\mathbf{T} \sim N_{np}(vec(\mathbf{B}'\mathbf{X}'), \sigma^2\mathbf{V}(\alpha) \otimes \mathbf{\Lambda}^{-1})$,
- matriciniu pavidalu $\mathbf{T}^* \sim N_{p\times n}(\mathbf{XB}, \sigma^2\mathbf{V}(\alpha) \otimes \mathbf{\Lambda}^{-1})$.

Čia $\mathbf{X}$ yra plano matrica, $\mathbf{B}' = (\mathbf{B}_1', \mathbf{B}_2')$ – vidurkio parametrų matrica, $\mathbf{V}(\alpha) = (\mathbf{I}_n + \alpha\mathbf{H})^{-1}$ žymi mokymo imties $\mathbf{T}$ erdvinių koreliacijų matricą, o $\mathbf{\Lambda}$ yra $p \times p$ dydžio koreliacijos tipo matrica, kurios pagrindinės įstrižainės elementai yra lygūs 1, o nediagonaliniai elementai yra $\{-\lambda_{ij}\}$.

Duotai mokymo imties realizacijai $\mathbf{T} = \mathbf{t}$ (arba $\mathbf{T}^* = \mathbf{t}^*$) stebinio $\mathbf{Z}_0$ populiacijoje Ω_l sąlyginis skirstinys yra $p - matis$ Gauso su sąlyginiu vidurkiu $\boldsymbol{\mu}_{lt}$ ir sąlygine dispersija $\mathbf{S}_t$:

$$\boldsymbol{\mu}_{lt} = \mathbf{B}_l'\mathbf{x}_0 + \left(\boldsymbol{\alpha}_0' \otimes \mathbf{I_p}\right)\left(\mathbf{t} - \mathrm{vec}(\mathbf{XB})\right) =$$

$$= \mathbf{B}_l'\mathbf{x}_0 + \left(\boldsymbol{\alpha}_0' \otimes \mathbf{I_p}\right)(\mathbf{t}^* - \mathbf{XB})',$$

$$\mathbf{S}_t = \rho_0\mathbf{\Lambda}^{-1}, \text{ kur } \rho_0 = \frac{\sigma^2}{1+\alpha h_0}.$$

Tuomet Bajeso diskriminantinė funkcija ir ją atitinkanti klaidingo klasifikavimo tikimybė yra

$$W^B(\mathbf{Z}_0, \mathbf{\Psi}) = (1 + \alpha h_0)(\mathbf{Z}_0 - \boldsymbol{\alpha}_0'(\mathbf{T}^* - \mathbf{XB}) - \mathbf{x}_0'\mathbf{I}_+\mathbf{B}/2)' \times$$

$$\times \mathbf{\Lambda}\mathbf{x}_0'\mathbf{I}_-\mathbf{B}/\sigma^2 + \gamma,$$

$$P_0^B(\mathbf{\Psi}) = \sum_{l=1}^2 \pi_l \Phi(-d/2 + (-1)^l \gamma/d),$$

kur $\gamma = \ln(\pi_1/\pi_2)$, $d^2 = (\boldsymbol{\mu}_1^0 - \boldsymbol{\mu}_2^0)'\mathbf{\Lambda}(\boldsymbol{\mu}_1^0 - \boldsymbol{\mu}_2^0)/\rho_0$ – sąlyginio Mahalanobiso atstumo kvadratas, o $\boldsymbol{\mu}_l^0 = \mathbf{B}_l'\mathbf{x}_0, l = 1{,}2$.

Nežinomus populiacijų parametrus pakeitus MT įvertiniais $\widehat{\mathbf{\Psi}} = \left(\widehat{\mathbf{B}}, \hat{\sigma}^2\right)$, gauname PBDF

$$W^B\left(\mathbf{Z}_0, \widehat{\mathbf{\Psi}}\right) = (1 + \alpha h_0)\left(\mathbf{Z}_0 - \left(\mathbf{T}^* - \mathbf{X}\widehat{\mathbf{B}}\right)'\boldsymbol{\alpha}_0 - \widehat{\mathbf{B}}'\mathbf{I}_+'\mathbf{x}_0/2\right)' \times$$

$$\times \mathbf{\Lambda}\widehat{\mathbf{B}}'\mathbf{I}_-'\,\mathbf{x}_0/\hat{\sigma}^2 + \gamma.$$

Ją atitinkanti aktualioji klaidingo klasifikavimo tikimybė yra

$$P_0^B\left(\widehat{\mathbf{\Psi}}\right) = \sum_{l=1}^2 \pi_l \Phi(\hat{Q}_l),$$

kur

$$\hat{Q}_l = (-1)^l \frac{(1+\alpha h_0)\left(\mathbf{x}_0'\left(\mathbf{B}_l - \mathbf{I}_+\widehat{\mathbf{B}}/2\right) + \boldsymbol{\alpha}_0'\mathbf{X}\left(\Delta\widehat{\mathbf{B}}\right)\right)\mathbf{\Lambda}\widehat{\mathbf{B}}'\mathbf{I}_-'\mathbf{x}_0 + \gamma\hat{\sigma}^2}{\hat{\sigma}\sqrt{\mathbf{x}_0'\mathbf{I}_-\widehat{\mathbf{B}}\mathbf{\Lambda}\widehat{\mathbf{B}}'\mathbf{I}_-'\mathbf{x}_0(1+\alpha h_0)}}.$$

Čia $\Delta\widehat{\mathbf{B}} = \widehat{\mathbf{B}} - \mathbf{B}$.

Kaip ir GGRF atveju, pateikiame aktualiosios klaidingo klasifikavimo tikimybės vidurkio aproksimacijos formulę (Teorema 2.4, 70 p.)

$$AEER = \sum_{l=1}^2 \pi_l \Phi(-d/2 + (-1)^l \gamma/d) + \pi_1 \varphi(-d/2 - \gamma/d) \times$$

$$\times \{\mathbf{F}_0'\mathbf{R}_B\mathbf{F}_0 d/k_0 + (p-1)\mathbf{x}_0'\mathbf{I}_-\mathbf{R}_B\mathbf{I}_-'\mathbf{x}_0/(k_0 d) + 2\gamma^2/d(np - 2q)\}/2,$$

$$\mathbf{F}_0 = \mathbf{X}'\boldsymbol{\alpha}_0 - (\mathbf{I}_+'/2 + \gamma\mathbf{I}_-/d^2)\mathbf{x}_0,$$

$$\mathbf{R}_B = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}, k_0 = 1/(1 + \alpha h_0).$$

Pastaba. Aktualiosios klasifikavimo rizikos vidurkio aproksimacijos formulė skaliariniam GMRF publikuota [A1].

Skaitiniai eksperimentai ir taikymai

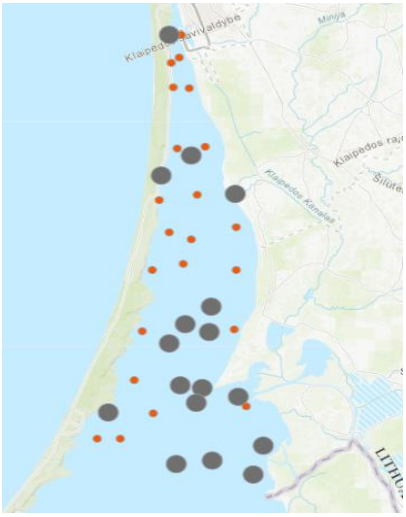
Disertacijos trečiajame skyriuje pateikti tokie skaitiniai eksperimentai ir taikymai:

- Neparametrinio testo galios tyrimas. Testo galia buvo tiriama statistinio modeliavimo būdu, vertinamas nulinės hipotezės atmetimų skaičius. Nustatyta, kad šio kriterijaus galia didėja augant slenksčio atstumo parametro reikšmei bei didinant atstumų tarp erdvės taškų (angl. *lags*) skaičių.
- Kovariacijos parametrų įtaka AER. Tiriama, kaip skirtingi statistiniai parametrai veikia klasifikavimo rizikos aproksimaciją. Nustatyta, kad AER labiausiai turi įtakos Mahalanobiso tarpklasinis atstumas ir slenksčio atstumas (erdvinės koreliacijos plotis).
- PBDF taikymas vertinant invazinių moliuskų (lot. *Dreissena polymorpha*) pasiskirstymą Kuršių mariose.

Pastarojo uždavinio tikslas – sudaryti matematinį modelį, leidžiantį įvertinti vienos iš invazinių moliuskų rūšių (*dreisenų*) paplitimą Kuršių mariose, taikant tiesinę diskriminantinę analizę, pagrįstą disertacijoje pasiūlytomis įterptomis Bajeso diskriminantinėmis funcijomis, bei įvertinti teisingo klasifikavimo tikimybę.

Pristatysime Gauso-Markovo modelį. Mokymo imtį sudaro 39 stebėjimo taškai ($n = 39$) – stacionarios stotys Kuršių mariose, kuriose stebimi trys požymiai: druskingumas, vandens atsinaujinimo greitis bei gylis. Taip pat stotyse fiksuojamas dreisenų buvimo/nebuvimo faktas. Pagal pastarąjį požymį mokymo imtį sudarančios stotys priskiriamos vienai iš dviejų klasių. 1 pav. pavaizduotas stočių išsidėstymas. Raudonais taškeliais pažymėtos stotys, kuriose dreisenų neaptikta ($n_1 = 22$), o pilkais – stotys, kuriose jų aptikta ($n_2 = 17$). Duomenų fragmentas pateiktas 1 lentelėje.

Priklausomu požymiu $Z(\mathbf{s})$, pagal kurį atliekamas klasifikavimas, gali būti bet kuris iš trijų stebimų požymių. Likę požymiai gali būti įtraukiami į vidurkio modelį kaip regresoriai (kovariatės). Erdvinė informacija į modelį gali būti įtraukiama trimis skirtingais būdais: į vidurkio modelį, į kovariacijų funkciją bei vertinant apriorines tikimybes.



1 pav. Stebėjimo taškų išsidėstymas Kuršių mariose

1 lentelė. Duomenų fragmentas

x	y	vandens atsinaujinimo greitis	druskingumas	gylis	Klasė (dreisenų aptikta/neaptikta)
21.1302	55.6553	24.71	2.89747	6.14803	1
21.1448	55.6352	26.89	2.53391	3.15062	0
21.1334	55.6309	28.58	2.49225	3.16278	0
...	...	...	...	...	...

Šio tyrimo tikslas yra, remiantis turima mokymo imtimi, sudaryti tiksliausią modelį, kurį būtų galima taikyti nestebimo taško Kuršių mariose (fokalinio taško) priskyrimui vienai iš dviejų klasių: 1) aptinkama dreisenų, 2) dreisenų neaptinkama.

Ieškant geriausio modelio disertacijoje buvo varijuojama tiek priklausomo požymio $Z(\mathbf{s})$, tiek kovariačių parinkimu, tiek sukauptos geografinės informacijos panaudojimu. Nustatyta, kad geriausiai dreisenų pasiskirstymą (77% tikslumu) galima klasifikuoti pagal *vandens atsinaujinimo greitį*, naudojant pastovaus vidurkio modelį ir tik artimiausius kaimynus, erdvinę informaciją įtraukiant į kovariacijų funkciją ir apriorinių tikimybių vertinimą.

Taip pat pastebėta, kad atsižvelgiant į erdvinę koreliaciją, klasifikavimo tikslumas gana ženkliai išauga, o vertinant apriorines tikimybes tikslinga įtraukti tik artimiausių kaimynų duomenis.

Išvados

Pasiūlytas neparametrinis testas geometrinei anizotropijai nustatyti yra nesunkiai realizuojamas ir gali būti taikomas kaip alternatyva kitų autorių siūlomiems kriterijams. Šio kriterijaus galia didėja augant slenksčio atstumo parametro reikšmei bei didinant atstumų tarp erdvės taškų skaičių.

Išvesta aktualiosios klasifikavimo rizikos vidurkio aproksimacijos formulė bei gautos parametrų asimptotinių kovariacijų matricų išraiškos gali būti panaudotos erdvinių imčių planų optimizavime, t.y. AER gali būti panaudota, kaip tikslo funkcija sprendžiant globalaus optimizavimo uždavinį.

Remiantis atliktų eksperimentų rezultatais, nustatyta, kad:

- erdvinės koreliacijos įtraukimas į modelį gerina siūlomų klasifikatorių efektyvumą;
- aktualiosios klasifikavimo rizikos vidurkio aproksimacijai labiausiai daro įtaką Mahalanobiso tarpklasinis atstumas ir slenksčio atstumas (erdvinės koreliacijos plotis);
- didesnis anizotropijos santykis didina AER reikšmes esant žemoms slenksčio atstumo reikšmėms, tačiau jo įtaka nėra labai žymi.

Pritaikius disertacijoje siūlomas diskriminantines funkcijas invazinių moliuskų pasiskirstymo Kuršių mariose vertinimui, nustatyta, kad:

- erdvinės informacijos įtraukimas į modelį ženkliai padidina klasifikavimo tikslumą;
- didesnis klasifikavimo tikslumas pasiekiamas apriorinių tikimybių vertinimui naudojant tik artimiausių kaimynų duomenis;
- tiriamam atvejui GMRF modelis yra pranašesnis nei GGRF.

Publikacijų sąrašas

Disertacijos tema yra atspausdinta 19 mokslinių straipsnių (3 iš jų žurnaluose įtrauktuose į CA Web of Science bazę).

[A1] Dučinskas K., Dreižienė L. (2018). Risks of classification of the Gaussian Markov random field observations. *Journal of Classification*, 35:422–436.

[A2] Dreižienė L., Dučinskas K., Šaltytė-Vaisiauskė, L. (2018). Statistical classification of multivariate conditionally autoregressive Gaussian random field observations. *Spatial Statistics*, 28:216–225.

[A3] Dučinskas K., Dreižienė L. (2016). Expected error rates in classification of Gaussian CAR observations. *Computer data analysis and modelling: theoretical and applied stochastics: proceedings of the 11th international conference*. Publishing center of BSU, Minsk, 127–130.

[A4] Dreižienė L., Dučinskas K. (2015). Error rates in multi-category classification of the spatial multivariate Gaussian data. *Procedia Environmental Sciences. Spatial Statistics conference 2015: Emerging Patterns*. Elsevier, Science Direct, 26:78–81.

[A5] Dreižienė L., Dučinskas K., Paulionienė L. (2015). Correct classification rates in multi-category discriminant analysis of spatial Gaussian data. *Open Journal of Statistics*, 5(1):21–26.

[A6] Dučinskas K., Dreižienė L., Zikarienė E. (2015). Multiclass classification of the scalar Gaussian random field observation with known spatial correlation function. *Statistics and Probability Letters*, 98:107–114.

[A7] Dučinskas K., Zikarienė E., Dreižienė L. (2014). Comparison of performances of plug-in spatial classification rules based on Bayesian and ML estimators. *In proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods*. SciTePress, 161-166.

[A8] Dučinskas K., Dreižienė L. (2013). Optimal classification of multivariate GRF observations. *Multivariate Statistics: Theory and Applications*, World Scientific, 61-72. (knygos skyrius).

[A9] Dreižienė L., Karaliutė M. (2012). The influence of training sample size on the expected error rate in spatial classification. *Lietuvos matematikos rinkinys. Proc. of the Lithuanian Mathematical Society*, Ser. A, 53:24–29.

[A10] Dreižienė L. (2011). Linear discriminant analysis of spatial Gaussian data with estimated anisotropy ratio. *Lietuvos matematikos rinkinys. LMD darbai*, 52:315–320.

[A11] Dučinskas K., Dreižienė L. (2011). Application of Bayes discriminant functions to classification of the spatial multivariate Gaussian data. *Procedia Environmental Sciences. Spatial Statistics 2011- Mapping Global Change*. Elsevier; ScienceDirect, 7:212-217.

[A12] Dučinskas K., Dreižienė L. (2011). Supervised classification of the scalar Gaussian random field observations under a deterministic spatial sampling design. *Austrian Journal of Statistics*, 40(1,2):25-36.

[A13] Dučinskas K., Dreižienė L. (2010). Supervised classification of the scalar Gaussian random field observation. *Computer data analysis and modeling: complex stochastic data and systems.*

Proceedings of the 9th International Conference. 2010 September 7-11, Minsk, Publishing center of BSU, 1:33-36.

[A14] Dučinskas K., Dreižienė L. (2010). Nonparametric test for spatial geometric anisotropy. *Lietuvos matematikos rinkinys. LMD darbai*. Vilnius: MII, 51:397–401.

[A15] Dreižienė L., Dučinskas K. (2009). The influence of the anisotropy ratio on the expected error rates in classification of stationary GRF observations. *Applied Stochastic Models and Data Analysis*, Vilnius: Technika, 101–105.

[A16] Dučinskas K., Dreižienė L. (2007). Effect of anisotropy coeficient on error rates of linear discriminant functions. *Lietuvos matematikos rinkinys*. Vilnius: MII. 47(spec. nr.):359–363.

[A17] Budrikaitė L. (2005). Modeling of zonal anisotropic semivariograms. *Lietuvos Matematikos rinkinys*. Vilnius: MII. 45(spec.nr.):339–342.

[A18] Budrikaitė L., Dučinskas K. (2005). Modelling of geometric anisotropic spatial variation. *Mathematical Modelling and Analysis*. Vilnius: Technika, 361–366.

[A19] Budrikaitė L., Dučinskas K. (2004). Forms of anisotropy for spatial variograms. *Lietuvos matematikos rinkinys*, Vilnius: MII. 44 (spec.nr.):542–546.

Rezultatų sklaida

Dalyvauta 7 tarptautinėse konferencijoje, kurių metu pristatyti 2 stendiniai ir skaityti 5 žodiniai pranešimai:

1. NORDSTAT 2018. Tartu, Estija, 2018 m. birželio 26-29 d.
2. Spatial Statistics 2017: One World, One Health. Lankasteris, Jungtinė Karalystė, 2017 m. liepos 4-7 d.
3. NORDSTAT 2016. Copenhaga, Danija, 2016 m. birželio 27-30 d.

4. Spatial statistics 2015: Emerging Patterns. Avinjonas, Prancūzija, 2015 m. birželio 9-12 d.
5. The 9-th Tartu Conference on Multivariate Statistics & the 20-th International Workshop on Matrices and Statistics. Tartu, Estija, 2011 m. birželio 26 – liepos 1 d.
6. Applied stochastic models and data analysis. Vilnius, Lietuva, 2009 m. birželio 30 – liepos 3 d.
7. Mathematical modelling and analysis. Trakai, Lietuva, 2005 m. birželio 1 – 5 d.

Taip pat skaityti 7 pranešimai LMD konferencijose, vykusiose 2018, 2016, 2012, 2011, 2010, 2005 ir 2004 m.

Literatūra

1. Abt, M. (1999). Estimating the prediction mean squared error in Gaussian stochastic processes with correlation structure. *Scandinavian Journal of Statistics*, 26:563-578.
2. Atkinson, P. M., Lewis, P. (2000). Geostatistical classification for remote sensing: an introduction. *Computers&Geosciences*, 26(4):361-371.
3. Batsidis, A., Zografos, K. (2006). Discrimination of observations into one of two elliptic populations based on monotone training samples. *Metrika*, 64, 221–241.
4. Berret, C., Calder, C. A. (2016). Bayesian spatial binary classification. *Spatial Statistics*. 16, 72–102.
5. Chiles, J. P., Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. Second Edition. John Wiley & Sons, New York.
6. Cressie, N. (1993). *Statistics for Spatial Data*. Wiley & Sons, New York.
7. Cressie, N. Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ.
8. Diggle, P. J., Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer.

9. Dučinskas, K. (1997). An asymptotic analysis of the regret risk in discriminant analysis under various training schemes. *Lithuanian Mathematical Journal*, 37(4):337–351.
10. Dučinskas, K. (2009). Approximation of the expected error rate in classification of the Gaussian random field observations. *Statistics and Probability Letters*, 79:138–144.
11. Dučinskas, K. (2009). Statistical classification of the observation of nuggetless spatial Gaussian process with unknown sill parameter. *Nonlinear Analysis: Modelling and Control*, 14(2):155–163.
12. Dučinskas, K. (2011). Error rates in classification of multivariate Gaussian random field observation. *Lithuanian Mathematical Journal*, 51:477–485.
13. Klein, R., Press, S. J., (1992). Adaptive Bayesian classification of spatial data. *Journal of the American Statistical Association*, 87:844–851.
14. Lindgren, F., Rue, H., Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Societ*y. *Series B: Statistical Methodology*, 73(4):423–498
15. Mardia, K. V., (1984). Spatial discrimination and classification maps. *Communications in Statistics – Theory and Methods*, 13:2181-2197.
16. Mardia, K. V., Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–146.
17. McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition*. Wiley, New York.
18. Okamoto, M. (1963). An asymptotic expansion for the distribution of linear discriminant function. *Annals of Mathematical Statistics*, 34:1286–1301.

19. de Oliveira, V., Ferreira M. A. R. (2011). Maximum likelihood and restricted maximum likelihood estimation for class of Gaussian Markov random fields. *Metrika*, 74(2):167–183.
20. Šaltytė, J., (2001). *The asymptotic expansion of the expected risk for the LDA of spatially correlated Gaussian observations* (Daktaro disertacija).
21. Šaltytė, J., Dučinskas, K. (2002). Comparison of ML and OLS estimators in discriminant analysis of spatially correlated observations. *Informatica*, 13(2):297–238.
22. Šaltytė-Benth, J., Dučinskas, K. (2005). Linear discriminant analysis of multivariate spatial-temporal regressions. *Scandinavian Journal of Statistics*, 32:281–294.
23. Shekhar, S., Schrater, P. R., Vatsavai, R. R., Wu, W., and Chawla, S. (2002). Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transactions on Multimedia*, 4(2):174–188.
24. Switzer, P. (1980). Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery. *Mathematical Geology*, 12(4):367–376.
25. Zhu, Z., Stein, M. L. (2006). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 11:24–44.

Summary

The thesis is devoted to the linear discriminant analysis of spatial data. We propose original discriminant functions based on plug-in Bayes classification rule and examine classification risk of the proposed classifiers.

This doctoral thesis consists of the introduction, three chapters, conclusions and bibliography.

Chapter 1 is designated for Gaussian models and their characteristics. It includes the issues of modelling spatial data, discusses the estimators for spatial models and presents a non-parametric test for detecting the geometric anisotropy.

Chapter 2 presents the main results of the thesis concerned with discriminant analysis of spatial data. Firstly the general definitions related with discriminant analysis are introduced. Later the formulas for Bayes risk and actual risk as well as formulas for error rates and actual error rates for the different number of populations are presented. Then the problem of classification of GGRF observation is analyzed. The Bayes risk associated with Bayes discriminant function and the asymptotic approximation formula of expected risk are derived. The above mentioned results are obtained for the univariate and multivariate cases and for different number of populations. Also the closed-form expression of asymptotic covariance matrix for exponential covariance model is presented. Finally, the classification problem of GMRF observation into one of two populations is solved.

The last chapter introduces the numerical experiments and applications.

Trumpos žinios apie disertantą

Išsilavinimas

2003 – Klaipėdos universiteto matematikos magistras

2001 – Vytauto Didžiojo universiteto matematikos bakalauras

1997 – Žemaičių Naumiesčio vidurinė mokykla

Pedagoginio darbo patirtis

Nuo 2001 m. Klaipėdos valstybinės kolegijos (anksčiau Klaipėdos verslo ir technologijų kolegija, Klaipėdos aukštesnioji žemės ūkio mokykla) lektorė;

Nuo 2003 m. Klaipėdos universiteto lektorė;

Nuo 2017 m. LCC tarptautinio universiteto lektorė.

Mokslinio darbo patirtis

Nuo 2018 m. Jūrinių tyrimų instituto jaunesnioji mokslo darbuotoja.

Vilniaus universiteto leidykla
Universiteto g. 1, LT-01513 Vilnius
El. p. info@leidykla.vu.lt,
www.leidykla.vu.lt
Tiražas 20 egz.