

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Jurgita ŽIDANAVIČIŪTĖ

KATEGORINIŲ POŽYMIŲ
PRIKLAUSOMYBIŲ STRUKTŪROS
STATISTINĖ ANALIZĖ
IR JOS TAIKYMAS GENETIKOJE

DAKTARO DISERTACIJA

FIZINIAI MOKSLAI,
MATEMATIKA (01P)



Vilnius LEIDYKLA
TECHNIKA 2010

Disertacija rengta 2003–2009 metais Vilniaus Gedimino technikos universitete.

Disertacija ginama eksternu.

Mokslinis konsultantas

doc. dr. Marijus RADAVIČIUS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

VG TU leidyklos TECHNIKA 1711-M mokslo literatūros knyga
<http://leidykla.vgtu.lt>

ISBN 978-9955-28-530-4
© VG TU leidykla TECHNIKA, 2010
© Židanavičiūtė, J. 2010
jurgita.zidanaviciute@fm.vgtu.lt

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Jurgita ŽIDANAVIČIŪTĖ

DEPENDENCE STRUCTURE ANALYSIS
OF CATEGORICAL VARIABLES
WITH APPLICATIONS IN GENETICS

DOCTORAL DISSERTATION

PHYSICAL SCIENCES,
MATHEMATICS (01P)



Vilnius LEIDYKLA TECHNICA 2010

Doctoral dissertation was prepared at Vilnius Gediminas Technical University in 2003–2009.

The dissertation is defended as an external work.

Scientific consultant

Assoc Prof Dr Marijus RADA VIČIUS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P).

Santrauka

Disertacijoje nagrinėjami kai kurie kategorinių požymių statistinės analizės uždaviniai – ryšių struktūros bei modelio parinkimo ir jo parametrų įvertinimo uždaviniai. Disertacijos tikslas – pasiūlyti ryšių struktūros tarp kategorinių požymių įvertinimo metodą tuo atveju, kai turime didelės dimensijos duomenų rinkinius ir išsklaidytas dažnių lenteles, t. y. kai daugumoje dažnių lentelės ląstelių yra mažas stebinių skaičius arba jos yra tuščios. Ši situacija iliustruojama DNR genetinių sekų statistine analize.

Disertaciją sudaro įvadas, trys pagrindiniai skyriai, išvados, naudotos literatūros ir autoriaus publikacijų sąrašai.

Įvadiniame skyriuje pristatoma tiriamoji problema ir jos aktualumas, aprašomas tyrimų objektas, formuluojamas darbo tikslas ir uždaviniai, aptariamas darbo mokslinis naujumas bei ginamieji teiginiai. Įvado pabaigoje pateikiami autoriaus pranešimai konferencijose disertacijos tema.

Pirmame skyriuje pateikta kategorinių požymių statistinėje analizėje taikomų matematinių modelių apžvalga, šių modelių ryšys su Markovo laukų teorija ir Gibso skirstiniu.

Antrame skyriuje atlikta statistinė analizė kai kurioms realioms DNR sekoms Markovo eilės jose įvertinimui bei pirminių ir antrinių DNR grandinių vijų palyginimui: pasiūlyta kategorinių duomenų statistinės analizės metodika, pagrįsta specialia stebimų duomenų forma, apibendrintu logit modeliu bei savirankos testais.

Trečiame skyriuje išsklaidytų dažnių lentelės problemai spręsti pasiūlytas semiparametrinis duomenų glodinimo metodas, kuriuo remiantis sukonstruotas semiparametrinis (suglodintas) savirankos kriterijus hipotezėms apie išsklaidytų kategorinių požymių tikimybinį skirstinį tikrinti. Modeliavimo būdu ištirtos pasiūlyto metodo galimybės DNR genetinių sekų statistinėje analizėje.

Disertacijos tema paskelbti 6 straipsniai. Du iš jų yra referuojami ISI Web of Science duomenų bazėje. Disertacijos tema perskaityti 6 pranešimai tarptautinėse ir 5 respublikinėse konferencijose.

Abstract

The dissertation considers some problems in the statistical analysis of categorical variables – the dependence structure between categorical variables and the problems of selecting and assessing the models for this structure. The aim of this dissertation is to propose the method to estimate dependence structure between multivariate categorical variables in case of sparse frequency tables and to apply the proposed method in the statistical analysis of genetic sequences.

The thesis layout consists of introduction chapter, three main chapters, conclusions, list of authors’s publications and bibliography chapter.

The introduction reveals the investigated problem, importance of the thesis and the object of research and describes the purpose and tasks of the dissertation, research methodology, scientific novelty, the practical significance and defended statements. The introduction end in presenting the author’s publications on the subject of the defended dissertation.

In the first chapter of the dissertation various models available for describing the nature of the association between categorical variables are introduced and their link with Markov field theory and Gibbs distribution.

In the second chapter the basic notions of DNA sequences and a special structure of genetic data is introduced. The logit models and Markov field theory are applied to assess the dependence structure (interactions) between DNA nucleotides and to test hypothesis about Markov order of these dependencies and hypothesis about reverse-complement symmetry between the leading and the lagging strand of DNA.

In the third chapter of dissertation a simple resampling technique based on semi-parametric smoothing (*semi-parametric or smoothed bootstrap*) is introduced and its application for hypothesis testing in sparse contingency tables is discussed.

The main results of the thesis are published in six scientific papers. Two of them are published in Thomson ISI Web of Science data base. The results were presented at five national and five international conferences.

Turinys

IVADAS	1
Problemos formulavimas ir darbo aktualumas	2
Tyrimų objektas	3
Darbo tikslas ir uždaviniai	3
Tyrimų metodai	4
Mokslinis darbo naujumas	4
Darbo rezultatų praktinė reikšmė	5
Ginamieji disertacijos teiginiai	5
Darbo rezultatų aprobavimas	6
Disertacijos struktūra	7
Padėka	8
 1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE	
NAUDOJAMI MATEMATINIAI MODELIAI	9
1.1. Bendrosios sąvokos	10
1.2. Apibendrintas tiesinis modelis	12
1.2.1. Eksponentinė skirstinių šeima	13
1.2.2. Apibendrinto tiesinio modelio apibrėžimas	17
1.2.3. Binarinė logistinė regresija	18
1.2.4. Puasono regresija	19
1.2.5. Logtiesiniai modeliai	20

1.2.6. Apibendrintas (daugialypis) logit modelis	24
1.2.7. Logtiesiniuose modeliuose naudojamos statistikos	26
1.2.8. Apibendrintas logit modelis ir Markovo laukai	27
2. LOGIT ANALIZĖ GENETINIAMS DUOMENIMS	33
2.1. Nukleotidų sekos	34
2.2. Nukleotidų sekų evoliucija	35
2.3. Literatūros apžvalga apie Markovo modelio eilės įvertinimą	35
2.4. Markovo savybė ir apibendrintas logit modelis	37
2.5. Statistinė analizė	41
2.5.1. Markovo savybė ir jos ryšys su sekų nehomogeniškumu ..	41
2.5.2. Homogeniškumo testas DNR sekoms	43
2.6. DNR grandinės pirminės ir antrinės vijų palyginimas	45
2.7. Logit statistinės analizės išvados	54
3. IŠSKLADYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS GLODINIMAS	55
3.1. Išsklaidytų dažnių lentelių problema	55
3.2. Literatūros apie išsklaidytas dažnių lenteles apžvalga	56
3.3. Semiparametrinis glodinimas ir saviranka	58
3.3.1. Semiparametrinis glodinimas	58
3.3.2. Semiparametrinis savirankos metodas	60
3.3.3. Suderinamumo testai	61
3.4. Logit modelis ir semiparametrinis savirankos metodas	63
3.4.1. Savirankos metodas	65
3.4.2. Kompiuterinis eksperimentas	66
3.5. Išvados pasiūlytam semiparametriniam glodinimo metodui	68
BENDROSIOS IŠVADOS	71
LITERATŪROS SĄRAŠAS	73
AUTORIAUS PUBLIKACIJOS DISERTACIJOS TEMA	79

Contents

INTRODUCTION	1
Scientific problem and topicality of the work	2
Research object	3
The aim and tasks of the work	3
Applied methods	4
Scientific novelty	4
Practical value of the work results	5
Statements presented for defence	5
Approval of the work results	6
The scope of the scientific work	7
Acknowledgment	8
 1. THE MATHEMATICAL MODELS USED IN STATISTICAL ANALYSIS OF CATEGORICAL VARIABLES	 9
1.1. Definitions and notations	10
1.2. Generalized linear model	12
1.2.1. Exponential family of distribution	13
1.2.2. Definition of generalized linear model	17
1.2.3. Dichotomous logistic regression	18
1.2.4. Poisson regression	19
1.2.5. Loglinear models	20

1.2.6. Generalized logit	24
1.2.7. Statistics used in the loglinear models	26
1.2.8. Generalized logit model and Markov fields	27
2. LOGIT ANALYSIS OF GENETIC DATA	33
2.1. Basics of DNA sequences	34
2.2. Evolution of DNA sequences	35
2.3. Review of Markov model estimating	35
2.4. Markov property and generalized logit	37
2.5. Statistical analysis	41
2.5.1. Markov property and inhomogeneity of DNA	41
2.5.2. Homogeneity test for DNA sequences	43
2.6. Reverse-complement symmetry of DNA	45
2.7. Conclusions for logit analysis of genetic data	54
3. SEMI-PARAMETRIC SMOOTHING OF SPARSE CONTINGENCY TABLES	55
3.1. The problem of sparse contingency tables	55
3.2. Review of literature	56
3.3. Semi-parametric smoothing and resampling	58
3.3.1. Semi-parametric smoothing	58
3.3.2. Semi-parametric resampling	60
3.3.3. Test statistics	61
3.4. Logit model and semi-parametric bootstrap	63
3.4.1. Semi-parametric bootstrap	65
3.4.2. Computer experiment	66
3.5. Conclusions for semi-parametric smoothing method	68
GENERAL CONCLUSIONS	71
REFERENCES	73
LIST OF THE AUTHOR’S SCIENTIFIC PUBLICATIONS ON THE TOPIC OF DISSERTATION	79

Įvadas

Bet kuri realiai egzistuojanti reiškinį apibūdinantys požymiai gali būti susieti tarpusavyje tam tikrais ryšiais, lemiančiais vienokias ar kitokias to reiškinio būsenas. Statistinė analizė, kurios metu pagal stebėtus duomenis įvertinami minėti ryšiai, dažnai vadinama terminu *Struktūrų analizė* (angl. *Structure learning*). Priklausomai nuo to, tarp kokių požymių siekiama įvertinti priklausomybę, atitinkamai pasirenkami ir matematiniai modeliai bei metodai jų parametrams įvertinti.

Pastaruoju metu sparti kompiuterinių technologijų pažanga bei vis didėjantys kaupiamų duomenų kiekiai ir galimybės juos apdoroti padidino mokslininkų susidomėjimą diskrečiais modeliais. Tokio tipo duomenys aptinkami praktiškai visose srityse: medicinoje, genetikoje ir kt.

Disertacijoje sprendžiamas požymių priklausomybės struktūrų įvertinimo uždavinys naudojant kai kurių realių DNR sekų duomenis. DNR sekų duomenų bazėse (*GenBank*) informacijos kiekiai dėl vis tobulėjančios genomų nuskaitymo ir dešifravimo technikos auga dideliu tempu, todėl juos apdorojant atsiranda vis didesnis poreikis ne tik intensyvaus kompiuterinių resursų panaudojimo, bet ir naujų algoritmų, o tuo pačiu ir naujų matematinių metodų bei modelių išvystymo.

Problemos formulavimas ir darbo aktualumas

Statistikos taikymo metodologijoje gerai žinomi ir praktikoje gana dažnai pasitaiko atvejai, kai papildomo kintamojo įtraukimas į statistinę analizę iš esmės pakeičia ankstesnio tyrimo rezultatus ir jų interpretaciją, o kartais net gaunamos priešingos išvados. Kategoriniams duomenims toks atvejis vadinamas *Simpsono paradoksu*. Vadinasi, atliekant statistinę analizę į tyrimą reikia įtraukti visus kintamuosius, kurie tik prieinami tyrėjui. O šiuo metu tyrėjui prieinami informacijos kiekiai yra pakankamai dideli, ir dėl to vis dažniau statistikoje susiduriama su didelio matavimo arba išsklaidytų duomenų analizės problemomis.

Kategoriniams duomenims didelės duomenų dimensijos problema – ypač aktuali. Kiekybinių duomenų modeliai paprastai aprašo vidurkių priklausomybės struktūrą, ir todėl šiuo atveju nežinomų parametrų skaičius nuo duomenų dimensijos n priklauso tiesiškai. Jei modeliuojama ir sudėtinga kovariacinė struktūra, nežinomų parametrų skaičius auga kaip $O(n^2)$. Tuo tarpu bendro modelio kategoriniams duomenims nežinomų parametrų skaičius dimensijos n atžvilgiu auga eksponentiškai. Todėl, netgi santykinai nedidelio skaičiaus kategorinių kintamųjų duomenys gali būti *išsklaidyti, išretinti, (angl. sparse)* – jų stebėtų dažnių lentelėje atsiranda daug tuščių ląstelių arba ląstelių su mažu stebinių skaičiumi jose, o nežinomų parametrų skaičius net didesnis už imties dydį.

Gerai žinoma, kad įprastiniai statistiniai metodai išsklaidytoms dažnių lentelėms nėra adekvatūs, nes taikomų kriterijų (statistikų) skirstiniams χ^2 aproksimacija nėra pakankamai tiksli. Be to, pačios kriterijų statistikos tampa mažai informatyvios, todėl sunku tikėtis parinkti adekvatų modelį. Ši problema formuluojama kaip išsklaidytų dažnių lentelių modelio parinkimo problema. Kita kategorinių duomenų analizės problema yra stebinių priklausomumas, kuris būdingas genetinių sekų, tekstų ar vaizdų analizėje. Šiuo atveju standartiniai statistiniai metodai taip pat nėra tinkami.

Išsklaidytų dažnių lentelių problema yra seniai žinoma. Yra didelis kiekis straipsnių, kuriuose nagrinėjami su šia problema susiję klausimai, tačiau pasiūlyti metodai turi tam tikrų trūkumų, tiriant išsklaidytas dažnių lenteles ir priklausomus kategorinius stebinius. Genetiniai duomenys yra tipinis tokio atvejo pavyzdys. Dėl to disertacijoje yra sprendžiami du genetikos taikomieji uždaviniai. Vienas iš jų – tinkamas Markovo grandinės eilės parinkimas. Jis svarbus dėl to, kad Markovo eilė yra susijusi su kaimyninių nukleotidų, įtakojančių evoliuciją kontekstiniuose evoliucijos modeliuose, skaičiumi. Yra parodyta (Jensen 2002), kad jeigu DNR sekų atsitiktinės mutacijos priklauso tik

nuo m artimiausių kaimyninių nukleotidų, tai stacionarusis nukleotidų evoliucijos skirstinys sudaro m -os eilės Markovo grandinę. Taigi, Markovo eilės įvertinimas yra taip pat Markovo eilės įvertinimas DNR sekų evoliucijos kontekstiniuose modeliuose.

Kitas DNR sekų analizės uždavinys, sprendžiamas tiek genetikų, tiek matematikų – DNR grandinės vijų simetriškumo savybė. Bendru atveju ši savybė palygina, ar nukleotidų sekos generavimo taisyklė, kuri galioja skaitant pirminę viją iš kairės į dešinę sutampa su antrinės vijos atitinkama generavimo taisykle, kai pastaroji skaitoma iš dešinės į kairę. Tačiau kol kas nėra pripažinto formalaus vijų simetriškumo savybės apibrėžimo, o juo labiau – jos matematinio/tikimybinio formulavimo (Baisnée *et al* 2002; Zhang ir Huang 2008).

Tyrimų objektas

Pagrindinis disertacijos objektas yra išsklaidytos dažnių lentelės, gautos stebint priklausomus kategorinius kintamuosius. Ryšių struktūrai tarp kategorinių požymių įvertinti naudojamas apibendrintas logit modelis. Taikymuose yra naudojami kai kurių bakterijų genomų nekoduojantys fragmentai, gauti iš *Genbank* duomenų bazės. Yra laikoma, kad nekoduojantys genomų fragmentai yra biologiškai mažiau svarbūs, dėl to juos galima laikyti genetiniu triukšmu.

Darbo tikslas ir uždaviniai

Disertacijos tikslas – pasiūlyti ryšių struktūros tarp kategorinių kintamųjų įvertinimo metodą tuo atveju, kai turime išsklaidytas dažnių lenteles, t. y. kai daugumoje lentelės ląstelių yra mažas stebinių skaičius arba jos yra tuščios. Pasiūlytą metodą pritaikyti DNR genetinių sekų analizėje.

Siekiant numatyto tikslo buvo sprendžiami šie uždaviniai:

1. Sukonstruoti adaptyvią išsklaidytų kategorinių duomenų glodinimo procedūrą bei pritaikyti ją hipotezių tikrinime.
2. Ištirti siūlomos glodinimo procedūros ir testų konstravimo metodo savybes (privalumus ir trūkumus), jų pritaikymo genetinių sekų analizėje galimybes.
3. Pasiūlyti DNR sekų statistinės analizės modelį, suderintą su jų evoliu-

cijos modeliu.

4. Naudojant realių bakterijų DNR duomenis, patikrinti hipotezę apie jų nekoduojančios dalies homogeniškų fragmentų markoviškumą bei hipotezę apie pirminės ir antrinės grandinės simetriškumo savybę.

Tyrimų metodai

Tyrime naudojami logtiesiniai kokybinių požymių modeliai, jų savybės, diskrečių Markovo grandinių ir laukų teorija, pakartotinių imčių metodai. Siūlomų metodų savybės tiriamos kompiuteriu atliekant imitacinius eksperimentus.

Darbo mokslinis naujumas

Disertacijoje pasiūlyta duomenų statistinės analizės metodika priklausomų kategorinių kintamųjų ir išsklaidytų dažnių lentelių statistinei analizei. Metodas remiasi apibendrintu logit modeliu, specialia duomenų struktūra, semiparametriniu glodinimu ir perrinkimo algoritmu. Šis glodinimo metodas ypač tinkamas kategoriniams duomenims glodinti, nes glodinimo procedūros, paremtos stebinių *artumu*, turi ribotas galimybes, yra nenatūralios ir sunkiai interpretuojamos.

Pasiūlytas semiparametrinis savirankos metodas, skirtingai negu neparametrinis, pritaikomas ir labai išsklaidytoms stebėtų dažnių lentelėms bei skirtingai negu parametrinis, jis yra mažiau jautrus neteisingai parametrinio modelio specifikacijai. Darbe pasiūlyto metodo galimybės tiriamos modeliavimo būdu.

Kaip jau buvo minėta anksčiau, disertacijoje DNR sekų Markovo savybės įvertinimui pristatyta speciali stebėtų duomenų struktūra, kuri užtikrina pagrindinių logit modelio prielaidų galiojimą ir dėl to statistinei analizei atlikti gali būti panaudota standartinė programinė įranga. Disertacijoje Markovo eilės nustatymo problema susiejama su sekų homogeniškumu. Atlikus specialų homogeniškumo testą buvo parodyta, kad bakterijų genomo nekoduojančiuose fragmentuose (aukštesnės eilės) Markovo priklausomybė tarp nukleotidų gali būti identifikuota net ir nepriklausomose sekose, jeigu jose pažeista homogeniškumo prielaida. Iš kitos pusės, nustatyta, kad statistiškai reikšmingoje tų fragmentų, kuriuos galima laikyti homogeniškais (jiems neatmestas homogeniškumo testas), dalyje nukleotidai nėra tarpusavyje nepriklausomi.

Daugelis darbų, susijusių su DNR sekų simetriškumo tematika, yra publikuoti biologų, genetikų ir informatikų. Išimtis yra Simons ir bendraautorių

(Simons *et al.* 2005) darbas, kuriame jie pasiūlė *globalų* komplementarių vijų simetriškumo tikimybinį modelį. Vis dėl to jų pasiūlytas modelis yra labai bendras ir negali būti įvertintas atskiram genomui. Disertacijoje pateiktas tikimybinis DNR pirminės ir antrinės vijų *lokalaus* simetriškumo apibrėžimas ir jo formulavimas naudojant apibendrintą logit modelį. Lokalus simetriškumas yra susijęs su kontekstiniais evoliucijos modeliais ir gali būti interpretuojamas kaip evoliucijos sąlygų tapatumas abiejose DNR komplementariose grandinėse. Preliminari bakterijų statistinė analizė parodė, kad DNR sekų vijos nėra *lokaliai* simetriškos, tačiau ta asimetrija įvairiuose sekos fragmentuose pasireiškia skirtingai ir *globaliai* neturi jokios bendros tendencijos. Tai leidžia daryti prielaidą, kad sekų vijos gali turėti skirtingą biologinę prasmę.

Darbo rezultatų praktinė reikšmė

Pasiūlytas semiparametrinio glodinimo metodas gali būti taikomas atliekant statistinę analizę bet kokioje srityje, kurioje reikia įvertinti ryšių tarp kategorinių požymių struktūrą tuo atveju, kai duomenys yra išsklaidyti. Taikomųjų uždavinių rezultatai yra naudingi genetikoje, tiriant DNR sekų evoliucijos modelius.

Ginamieji disertacijos teiginiai

1. Pasiūlytas adaptyvus semiparametrinis glodinimo metodas yra universalus ir atitinkamai parenkant optimalaus glodinimo kriterijų gali būti taikomas reguliarizacijos uždaviniams spręsti. Nors šį metodą galima taikyti tolydiems dydžiams, tačiau jis ypatingai naudingas išsklaidytiems kategoriniams duomenims.
2. Pasiūlyta glodinimo procedūra leidžia pritaikyti neparametrinį (semiparametrinį) metodą įvairioms hipotezėms tikrinti ir tuo atveju kai dažnių lentelės yra išsklaidytos.
3. Apibendrintas logit modelis bei įvesta speciali duomenų struktūra leidžia atlikti DNR sekų statistinę analizę, panaudojant standartinę programinę įrangą, o taip pat tiesiogiai pritaikyti semiparametrinį glodinimą.
4. Dalis nekoduojančių DNR fragmentų yra nehomogeniški ir tai gali sąlygoti jų ne-Markoviškumą.

5. Reikšminga dalis homogeninių fragmentų nėra pirmos eilės Markovo. Vadinasi, įprasti DNR sekų tikimybiniai modeliai, tiek paslėptos Markovo grandinės, tiek ir k -tosios eilės homogeninės Markovo grandinės, nėra tinkami modeliai *visos* DNR sekos analizei.
6. Disertacijoje tirtų bakterijų DNR sekų vijos lokaliai nėra simetriškos, tačiau ta asimetrija įvairiose sekos fragmentuose pasireiškia skirtingai. Praktiškai tai reiškia, kad sekų vijos gali turėti skirtingą biologinę prasmę.

Darbo rezultatų aprobavimas

Disertacijos tema paskelbti 6 straipsniai. Du iš jų yra referuojami ISI Web of Science duomenų bazėje. Disertacijos tema perskaityti 6 pranešimai tarptautinėse ir 5 respublikinėse konferencijose:

1. Radavičius M., Židanavičiūtė J. Local reverse-complement symmetry of DNA sequences // *2nd Nordic-Baltic Biometric Conference, NBBC09*, 10–12 June 2009, Tartu, Estonia.
2. Židanavičiūtė J., Radavičius M. DNR komplementarių grandinių palyginimas // *Lietuvos matematikų draugijos L konferencija*. Vilnius, Birželio 18–19 d., 2009.
3. Radavičius M., Židanavičiūtė J. Hypothesis testing for sparse categorical data // *22nd Nordic Conference on Mathematical Statistics*, 16–19 June 2008, Vilnius, Lithuania.
4. Židanavičiūtė J., Rekašius T. Apibendrinto logit modelio taikymas, tiriant DNR sekų savybes // *Lietuvos matematikų draugijos XLIX konferencija*. Kaunas, Birželio 25–26 d., 2008.
5. Židanavičiūtė J., Rekašius T. Logit Analysis of Genetic Data. // *12th International Conference Mathematical Modelling and Analysis*, Trakai, May 30–June 2, 2007.
6. Radavičius M., Židanavičiūtė J. Analysis of Sparse Contingency Tables with Applications. // *The 8th Tartu Conference on Multivariate Statistics*, 26–29 June 2007, Tartu, Estonia.
7. Radavičius R., Židanavičiūtė J. Model Testing for High-Dimensional Contingency Tables with Application in Genetics. // *The 8th International conference Computer Data Analysis and Modeling: Complex*

- Stochastic Data and Systems*, Minsk, Belarus, September 11–15, 2007.
8. Rekašius T., Židanavičiūtė J. Dependence structure of noncoding DNA regions // *The 9th international Vilnius conference on probability theory and mathematical statistics*, June 25–30, 2006, Vilnius.
 9. Rekašius T., Židanavičiūtė J. Genetinių sekų markoviškumo ir apverčiamumo tyrimas // *Lietuvos matematikų draugijos XLVII konferencija*. Kaunas, Birželio 20–21 d., 2006.
 10. Radavičius M., Židanavičiūtė J. Statistinė struktūrų analizė; kai kurios jos taikymo problemos // *Lietuvos matematikų draugijos XLVI konferencija*. Vilnius, Birželio 15–16 d., 2005.
 11. Židanavičiūtė J. Kokybinių rodiklių tarpusavio ryšių struktūrinė analizė ir jos taikymai // *Lietuvos matematikų draugijos XLV konferencija*. Kaunas, Birželio 17–18 d., 2004.

Disertacijos struktūra

Disertaciją sudaro įvadas, trys pagrindiniai skyriai, išvados, naudotos literatūros sąrašas ir autoriaus publikacijų disertacijos tema sąrašas. Darbo apimtis yra 80 puslapių, tekste panaudotos 139 formulės, 1 paveikslas, 8 lentelės. Rašant disertaciją buvo remtasi 70 literatūros šaltiniais.

Pirmame skyriuje pateikta kategorinių požymių statistinėje analizėje taikomų matematinių modelių apžvalga, šių modelių ryšys su Markovo laukų teorija ir Gibso skirstiniu. Antrame skyriuje atlikta statistinė analizė kai kurioms realioms DNR sekoms Markovo eilės jose įvertinimui bei pirminių ir antrinių DNR grandinių vijų palyginimui: pasiūlyta kategorinių duomenų statistinės analizės metodika, pagrįsta specialia stebimų duomenų forma, apibendrintu logit modeliu bei savirankos testais. Trečiame skyriuje išsklaidytų dažnių lentelės problemai spręsti pasiūlytas semiparametrinis duomenų glodinimo metodas.

Padėka

Labai džiaugiuosi, kad, rašant šį darbą, teko galimybė pažinti daug įdomių, puikių, geranoriškų mokslo aplinkos žmonių. Labai dėkoju darbo vadovui doc. dr. Marijui Radavičiui už galimybę mokytis iš jo, už skirtą laiką, įdomias idėjas, nuoširdų bendravimą ir buvimą patarėju bet kokių klausimų. Taip pat dėkoju Vilniaus Gedimino technikos universiteto Matematinės statistikos katedros darbuotojams bei Matematikos ir informatikos instituto kolektyvui už pagalbą, moralinį palaikymą bei konsultacijas. Lietuvos banko Statistikos departamento kolegoms dėkoju už supratimą ir geranoriškumą. Ačiū artimiesiems ir visiems draugams už visakeriopą pagalbą, paramą ir laukimą.

1

Kategorinių požymių statistinėje analizėje naudojami matematiniai modeliai

Priklausomai nuo keliamų tikslų, struktūrų analizėje naudojami įvairūs modeliai: logtiesiniai (Avery, Henderson 1999; Avery 2002), paslėptojo kintamojo (Loehlin 1997) ir paslėptieji Markovo modeliai (Murphy 2001), grafiniai modeliai (Murphy 2001; Koehly *et al* 2003) ir kt. Paslėptojo kintamojo modelio idėja – pagal stebimo dydžio Y , kuris priklauso ir nuo X , reikšmes atstatyti nestebimas X reikšmes ir įvertinti jų tikimybinės charakteristikas. Analogiškas uždavinys yra ir paslėptuose Markovo modeliuose, tik čia Y yra atsitiktinis procesas, o X yra nestebima Markovo grandinė. Šio tipo modeliams priskiriami klasterinė, faktorinė, logistinė analizės (Stokes *et al* 2001; Agresti 1990; Santer, Duffy 1989), modeliai su cenzūruotais duomenimis. Šis modelių skirstymas yra gana sąlyginis, nes išvardintos modelių klasės yra persipynusios tarpusavyje.

Toliau šiame skyriuje aptariama kategorinio kintamojo ir nekorektiškų uždavinių sąvokos, pristatyta dažnių lentelė, kaip vienintelė prasminga statistika kokybiniams požymiams, aprašoma apibendrinta tiesinių modelių klasė, kuriai priklauso ir disertacijoje naudojami logtiesiniai modeliai bei logtiesinių modelių ryšys su Markovo atsitiktinių laukų teorija.

1.1. Bendrosios sąvokos

Kintamieji ir jų matavimo skalės. Atliekant statistinę analizę, paprastai naudojami duomenys, gauti matuojant populiacijos elementų tam tikrus požymius (rodiklius). Šie požymiai yra vadinami kintamaisiais. Jie skirstomi į kokybinius (kategorinius) ir kiekybinius. Kokybiniais kintamiesiems matuoti naudojama pavadinimų arba rangų skalė, kiekybiniais kintamiesiems – intervalų arba santykių skalė. Šias 4-ių tipų matavimo skales, pavadinimų (vardų, nominalioji), rangų, intervalų ir santykių, po ilgų mokslininkų debatų įvedė Stevens (1946).

Matavimo skalę charakterizuoja prasmingos (leistinos) duotoje skalėje operacijos ir operacijos, kurios nekeičia turimos informacijos, t. y. transformacijos, kurių atžvilgiu matavimų duomenyse esanti informacija yra invariantiška. Taip pat matavimo skalės apsprendžia statistinių metodų pasirinkimą, atliekant tolimesnį gautų duomenų apdorojimą.

Naudojant vardų skalę, objektus pagal išmatuotas kintamųjų reikšmes galima priskirti tik vienai ar kitai grupei (klasei), t. y. matuojami požymiai aprašomi tik lygybės ar nelygybės savybe. Kintamieji matuojami vardų skalėje vadinami nominaliaisiais.

Rangų skalė naudojama tada, kai galima nustatyti objektų tiriamo požymio skirtumus ir pagal tai objektus išrikiuoti į eilę. Vadinasi, prasmingas tik objektų palyginimas pagal duoto požymio išreikštumą: daugiau, lygu ar mažiau.

Intervalinėje skalėje matuojamų pavyzdžių yra nedaug: temperatūra, laikas. Šioje skalėje objektus galima ne tik klasifikuoti, tvarkyti, bet ir kiekybiškai įvertinti skirtumus. Šioje skalėje nenusakyta koordinačių pradžia, ją galima pasirinkti laisvai ir tai neįtakoja gautų rezultatų interpretacijos ir išvadų.

Santykinė skalė – tai pati turtingiausia skalė. Joje prasmingos visos aritmetinės operacijos ir palyginimai. Nuo intervalinės skalės ji skiriasi tuo, kad nusakytą reali prasminga (interpretuojama) koordinačių pradžia.

Parametriniai ir neparametriniai statistiniai modeliai. Atsitiktinio n -mačio vektoriaus Y skirstinių šeima $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ vadinama parametriniu modeliu jei parametras $\theta = (\theta_1, \dots, \theta_m)' \in \Theta \subset \mathbb{R}^m$ yra baigtiniamatis. Statistinis modelis $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ vadinamas neparametriniu, jei parametras θ yra begalinio matavimo ir neturi baigtinio matavimo komponentų (Bagdonavičius, Kruopis 2007).

Jeigu diskretūs kintamieji gali įgyti tik baigtinį skaičių skirtingų reikšmių, tai jų (bendras) tikimybinis skirstinys pilnai nusakomas baigtiniu skaičiumi skaliarinių parametru. Taigi, tokiu atveju visi tikimybiniai modeliai formaliai yra parametriniai. Tačiau tradiciškai visų galimų skirstinių aibė laikoma

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI MATEMATINIAI MODELIAI

11

neparametrine, o tuo pačiu ir visi metodai, kurie nesiremia jokia papildoma specialia parametrine struktūra ir taikomi visų tikimybinių skirstinių klasėje, yra taip pat neparametriniai.

Atrodytų, kad diskrečių kintamųjų, įgijančių baigtinių kategorijų skaičių, atveju pavyksta išvengti vienos iš opiausių neparametrinio vertinimo problemų – nekorektiškojo (statistinio) uždavinio (*angl. ill-posed problem*). Su šia problema neretai susiduriama taikant neparametrinius metodus tolydiesiems kintamiesiems. Tipinis uždavinys: turint baigtinių skaičių stebinių, įvertinti begalinį skaičių regresijos (pasiskirstymo tankio ir pan.) funkcijos Furje koeficientų Furje skleidinyje. Aišku, kad be papildomų prielaidų šis uždavinys (vienareikšmiškai) neišsprendžiamas. Tokie uždaviniai ir vadinami *nekorektiškaisiais*. Tuo tarpu diskrečių kintamųjų su baigtiniu būsenų (kategorijų) skaičiumi atveju vertinamų (skaliarinių) parametrų skaičius visada yra tik baigtinis, ir atrodytų turint pakankamą imties dydį juos būtų galima įvertinti. Deja, praktikoje dažnai pasitaiko atvejis, kai dažnių lentelės ląstelėse pasitaiko labai mažai stebinių arba jų iš viso nėra. Kaip tada įvertinti tikimybę, kad atsitiktinis stebinis pateks į tokia ląstelę? Taigi, ir šiuo (teoriškai *parametriniu*, o praktiškai *neparametriniu*) atveju susiduriama su nekorektiškiesiems statistiniams uždaviniams būdingais keblumais. Teoriškai *parametrinis* uždavinys praktiškai virsta *neparametriniu*.

Kokybinių kintamųjų dažnių lentelės. Iš nominaliosios (vardų) skalės apibrėžimo išplaukia, kad vienintelis prasmingas statistikos pavidalas kokybiniais požymiams yra jų įvairių reikšmių kombinacijų pasitaikymo duomenyse stebėjimo dažniai, kurie paprastai pateikiami kryžminių dažnių lentelių pavidalu.

Tarkime, yra stebimi 2 požymiai – X_1 ir X_2 . Požymio X_1 įgyjamas reikšmes pažymėkime $i = 1, \dots, I$. Analogiškai požymis X_2 gali įgyti reikšmes $j = 1, \dots, J$. Tada minėtų požymių stebimus dažnius Y_{ij} galima pavaizduoti dvimatėje dažnių lentelėje (1.1).

Dažnių lentelė gali būti laikoma pirmąja statistine priemone apdoroti kokybinio tipo duomenims ir įvertinti jų ryšių struktūrą. Dažnių lentelių analizės pradininku laikomas Pearson (1900), kuris pirmasis pasiūlė klasikinį χ^2 testą. Pearson teigė, kad visų tipų duomenys, tame tarpe ir kiekybiniai, gali būti diskretizuojami ir jiems taikoma kokybinių duomenų analizė. Panašiu metu Yule (1900) pasiūlė naują požiūrį dažnių lentelių analizėje. Jis tvirtino, kad kokybiniai duomenys yra fiksuoti ir nėra diskretizuoto proceso rezultatas. Taip pat jis apibrėžė šansų santykį, tiesiogiai naudodamas ląstelių dažnius be kintamojo tolydumo prielaidų – jis teigė, kad dauguma kintamųjų yra iš prigimties diskretūs. Tai buvo priešingas požiūris Pearsonui. Ilgą laiką vyko debatai šia

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI MATEMATINIAI MODELIAI

tema, ir iki tol buvo nagrinėjamos tik 2x2 lentelės. Tuo metu Fisher (1934) pristatė tikslųjį testą 2x2 lentelėms. Barlett (1935) apibrėžė trimatę 2x2x2 dažnių lentelę, o Lancaster (1951) išvystė metodą χ^2 statistikai, pritaikytą $r \times s \times t$ lentelėms. Barlett darbus apibendrino Roy ir Kastenbaum (1956).

Turint didelį požymių skaičių, atlikti gilesnę statistinę analizę šiais metodais yra sudėtinga, o kartais iš vis neįmanoma. Bet visi šie metodai buvo pagrindu, iš kurių 1960 m. atsirado ir šiuo metu plačiai naudojami logtiesiniai modeliai (plačiau apžvalgą galima rasti Fienberg (2000) straipsnyje). Ir nors daugelis tyrinėtojų idėjo tam tikrą reikšmingą indėlį į logtiesinių modelių kūrimą (Agresti 1990), bet labiausiai išpopuliarinusių šiuos modelius socialiniuose moksluose buvo laikomas Leo Goodman (1968, 1970).

Logtiesinių modelių klasei (Stokes *et al* 2001; Agresti 1990; Christensen 1990; Santer, Duffy 1989) priklauso ne vien tik kokybinių kintamųjų modeliai, o tam tikra prasme ir multiplikatyvūs ekonometriniai modeliai yra atskiri jos atvejai.

1.1 lentelė. Dviejų požymių X_1 ir X_2 kryžminė dažnių Y_{ij} lentelė

Table 1.1. Two-Way contingency table

		X_2				
		Y_{ij}	1	2	...	J
X_1	1		Y_{11}	Y_{12}	...	Y_{1J}
	2		Y_{21}	Y_{22}	...	Y_{2J}
	.		.	.	...	.
	.		.	.	...	.
	I		Y_{I1}	Y_{I2}	...	Y_{IJ}
			Y_{+1}	Y_{+2}	...	Y_{+J}
						Y_{++}

1.2. Apibendrintas tiesinis modelis

Paprasčiausiu atveju įvertinti tiesinį ryšį tarp priklausomo (atsako) kintamojo Y ir nepriklausomų kintamųjų (aiškinančiųjų kintamųjų, dar vadinamų kovariantėmis) $X_1, X_2, \dots, X_m$ gali būti naudojamas tiesinis regresijos modelis (Čekanavičius, Murauskas 2001; Sengupta, Jammalamadaka 2003; Agresti 2007), kuriame laikoma, kad Y turi normalųjį skirstinį su vidurkiu μ ir dispersija σ^2 : $Y \sim N(\mu, \sigma^2)$.

Tačiau yra ryšių, kurių negalima adekvačiai įvertinti, naudojant paprastą

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI MATEMATINIAI MODELIAI

13

tiesinį modelį dėl dviejų pagrindinių priežasčių: kai priklausomo kintamojo Y skirstinys nėra Normalusis ar priklausomybė tarp Y ir $X_1, X_2, \dots, X_m$ nėra tiesinė. Abi šios savybės būdingos kokybinio tipo duomenimis. Dėl šių priežasčių Nelder ir Wedderburn (1972) praplėtė ir apibendrinė tiesinių modelių klasę bei pavadino ją apibendrintais tiesiniais modeliais (*angl. generalized linear model*) (Sengupta, Jammalamadaka 2003; Agresti 2007). Šiuose modeliuose, priklausomai nuo esančių prielaidų priklausomam kintamajam Y , jo vidurkis gali priklausyti nuo tiesinių prediktorių per atitinkamą sąryšio funkciją f (*angl. link function*), kuri gali būti log, logit ir kt., o Y skirstinys priklauso (apibendrintai) eksponentinei skirstinių šeimai.

1.2.1. Eksponentinė skirstinių šeima

1 apibrėžimas. Sakoma, kad n -matis atsitiktinis vektorius Y arba jo skirstinys priklauso (apibendrintai) eksponentinei skirstinių šeimai su parametru $\theta \in \Theta \subset \mathbf{R}^d$, jeigu jo pasiskirstymo tankis σ -baigtinio mato ν atžvilgiu turi tokį pavidalą:

$$p(y | \theta) = h(y) \exp \{ \eta(\theta)' T(y) - b(\theta) \}, \quad y \in \mathbf{R}^n, \quad (1.1)$$

čia $\eta: \mathbf{R}^d \rightarrow \mathbf{R}^m$, $h: \mathbf{R}^n \rightarrow [0, \infty)$, $T: \mathbf{R}^n \rightarrow \mathbf{R}^m$ yra žinomos funkcijos, o $b: \mathbf{R}^d \rightarrow \mathbf{R}$ yra normuojantis narys. Funkcijos η, h ir T kartu su matu ν yra tarsi eksponentinės šeimos hiperparametrai. Jie nusako konkrečios eksponentinės šeimos pavidalą, jos skirstinių formą.

Funkciją h galima eliminuoti pakeitus pradinį matą ν matu $\nu_h(dy) = h(y)\nu(dy)$. Taigi, funkcija h eksponentinės šeimos apibrėžime yra perteklinė, tačiau ji leidžia eksponentinės šeimos apibrėžime kaip bazinį matą ν naudoti standartinius matus. Paprastai tai būna Lebego arba skaičiuojantysis matas.

Matas ν vadinamas skaičiuojančiuoju suskaičiuojamoje aibėje S , jeigu bet kurio aibės S elemento matas (svoris) yra lygus vienetui. Kitais žodžiais tariant, mato ν atrama (*angl. support*) yra aibė S , ir $\nu(A) = |A|$ visiems poaibiams $A \subset S$.

Funkcija T yra eksponentinės šeimos (minimali) pakankama statistika. Funkcija $b(\theta)$ nusako normavimo daugiklį ir yra vadinama kumuliantine (*angl. cumulant*) arba logaritmine dalinimo (*angl. logpartition*) funkcija:

$$\exp \{ b(\theta) \} = \int_{\mathbf{R}^n} h(y) \exp \{ \eta(\theta)' T(y) \} \nu(dy).$$

Kai $\eta(\theta) \equiv \theta$, parametras θ vadinamas natūraliuoju (kanoniniu). Kai θ

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI MATEMATINIAI MODELIAI

yra natūralusis parametras ir $T(y) \equiv y$, eksponentinė skirstinių šeima vadinama natūraliąja. Natūraliosios eksponentinės skirstinių šeimos tankis turi tokį pavidalą:

$$p(y | \theta) = h(y) \exp \{ \theta' y - b(\theta) \}, \quad y \in \mathbf{R}^n, \quad (1.2)$$

kuris vadinamas kanoniniu.

Pasiskirstymo tankis (1.1) suvedamas į šį pavidalą (žinoma, jau su kitais hiperparametrais h ir ν ir, gal būt, kita y reikšmių sritimi) atlikus atitinkamus kintamųjų pakeitimus.

Diskrečių (kategorinių) duomenų atveju galima laikyti, kad $Y \in \mathbf{Z}_+^n$ (čia $\mathbf{Z}_+$ žymi sveikųjų neneigiamų skaičių aibę, $\mathbf{Z}_+ := \{0, 1, \dots\}$). Todėl paprasčiausia skaityti, kad matas ν yra skaičiuojantysis matas, apibrėžtas aibėje $\mathbf{Z}_+^n$. Šiuo atveju lygybė (1.1) tiesiog reiškia, kad

$$\mathbf{P}(Y = y) = p(y | \theta) = h(y) \exp \{ \eta(\theta)' T(y) - b(\theta) \}, \quad y \in S \quad (S = \mathbf{Z}_+^n). \quad (1.3)$$

Paminėsime kai kurias pagrindines natūraliosios eksponentinės šeimos savybes:

1. Tarp θ ir Y vidurkio bei kovariacijų matricos yra paprastas ryšys:

$$\mathbf{E}_\theta Y := \mu(\theta) = \frac{\partial b(\theta)}{\partial \theta}, \quad (1.4)$$

$$\text{Var}(Y) := \mathbf{E}_\theta Y Y' - \mu(\theta) \mu(\theta)' = \frac{\partial^2 b(\theta)}{\partial \theta \partial \theta'}. \quad (1.5)$$

2. Tegu turime N nepriklausomų ir vienodai pasiskirsčiusių atsitiktinių vektorių $Y_1, \dots, Y_N$ su bendru skirstiniu iš natūraliosios eksponentinės šeimos. Tada kanoninio parametro θ *minimali, pakankama* statistika yra aritmetinis vidurkis $\bar{Y} := \frac{1}{N} \sum_{j=1}^N Y_j$.

Iš pirmo žvilgsnio eksponentinė skirstinių šeima neatrodo labai „turinga“, tačiau jai priklauso dauguma gerai žinomų ir plačiai taikomų skirstinių. Be to, eksponentinei skirstinių šeimai priklauso visi diskretieji skirstiniai su ta pačia baigtine galimų reikšmių aibe, t. y. su ta pačia skirstinio atrama. Toliau mums svarbūs bus tik diskretieji skirstiniai: binominis, Puasono ir polinominis.

Polinominis skirstinys. Atsitiktinis dydis X su reikšmėmis aibėje $I_n := \{0, 1, \dots, n\}$ turi polinominį skirstinį su parametrais $N \in \mathbf{N}$ ir $p \in \mathcal{P}_n$, jeigu:

$$p(x) = \mathbf{P}\{X = x\} = \frac{N!}{\prod_{j=0}^n x_j!} \prod_{j=0}^n p_j^{x_j}, \quad x \in I_n^n, \quad \sum_{j=0}^n x_j = N, \quad (1.6)$$

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI
MATEMATINIAI MODELIAI

15

$$\mathcal{P}_n := \left\{ p = (p_0, p_1, \dots, p_n)' \in \mathbf{R}^{n+1} : p_i > 0, i = 0, \dots, n, \sum_{j=0}^n p_j = 1 \right\}.$$

Trumai tai žymėsime $X \sim Multinomial_n(N, p)$.

Polinominiai skirstiniai sudaro tokią eksponentinės šeimos skirstinių klasę:

$$\begin{aligned} p(y) = p_N(y \mid \theta) &:= h_N(y) \exp\{(\theta, \theta')y - Nb(\theta)\}, y \in S_N, \theta \in \mathbf{R}^n, \\ h_N(y) &:= \frac{N!}{\prod_{j=0}^n y_j!}, \\ b(\theta) &:= \log \left(1 + \sum_{j=1}^n \exp\{\theta_j\} \right), \\ S_N &:= \left\{ x = (x_0, x_1, \dots, x_n)' \in I_N^{n+1} : \sum_{j=0}^n x_j = N \right\}, \end{aligned} \quad (1.7)$$

o mato ν atrama yra S_N .

Kai $N = 1$, polinominės eksponentinės šeimos skirstinio pavidalas ženkliai supaprastėja:

$$p(y) = p(y \mid \theta) = \exp \left\{ \sum_{j=1}^n \theta_j \delta_j(y) - b(\theta) \right\}, \quad y = 0, \dots, n, \quad (1.8)$$

čia $\delta_j(\cdot)$ yra Dirako matas, sukoncentruotas taške j , o $b(\theta)$ jau apibrėžėme anksčiau (žr. 1.7).

Fiksuokime baigtinę aibę $S_n = \{s_0, s_1, \dots, s_n\}$. Tada atvaizdavimas $T: S_n \rightarrow I_n$ su $T(s_j) = j, j = 0, 1, \dots, n$, apibrėžia abipus vienareikšmę atitinkamybę tarp skirstinių su atrama S ir eksponentinės polinominių skirstinių šeimos. Taigi, eksponentinei skirstinių šeimai priklauso visi diskretieji skirstiniai su ta pačia baigtine galimų reikšmių aibe S_n .

Skirstinys (1.8) faktiškai yra Gibso skirstinys (žr. žemiau) su potencialų funkcijomis

$$\lambda_j(x) := -\theta_j \delta_j(x), \quad x = 0, \dots, n \quad (1.9)$$

ir padalijimo (*angl. partition*) funkcija $Z = \exp\{b(\theta)\}$.

Šiuo atveju potencialų struktūra labai paprasta, o tokiu būdu apibrėžti kanoniniai parametrai nėra labai informatyvūs. Informatyvesnę struktūrą turi toliau pateiktas sekų su baigtiniu būsenų skaičiumi Gibso skirstinys.

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI
16 MATEMATINIAI MODELIAI

Sekų su baigtiniu būsenų skaičiumi Gibso skirstinys. Tegu $\mathcal{A}$ yra baigtinė būsenų aibė (alfabetas), o $\mathcal{X} = \mathcal{X}_n$ yra aibė visų n ilgio sekų x su reikšmėmis iš $\mathcal{A}$.

2 apibrėžimas. Gibso skirstiniu (matu) aibėje $\mathcal{X}$ vadinamas tikimybinis skirstinys p_g , nusakomas potencialų funkcijų rinkiniu $E := \{\lambda_s: \mathcal{X} \rightarrow \mathbf{R}, s \in S\}$ ir lygybėmis:

$$p_g(x) := \frac{1}{Z} \exp \left\{ - \sum_{s \in S} \lambda_s(x) \right\}, \quad Z := \sum_{x \in \mathcal{X}} \exp \left\{ - \sum_{s \in S} \lambda_s(x) \right\}. \quad (1.10)$$

Normuojantis daugiklis Z vadinamas padalijimo (*angl. partition*) funkcija.

Trumpai aptarsime potencialų rinkinio E parinkimo uždavinį, pasinaudodami Mobiuso formule.

Mobiuso formulė. Duotai aibei $J \subset I_n := \{1, \dots, n\}$ apibrėžkime atvaizdavimą $y = H_J(x)$, $H_J: \mathcal{X} \rightarrow \mathcal{X}$, tokiu būdu: $y_j = x_j$, $\forall j \in J$ ir $y_j = b$, $\forall j \in I_n \setminus J$, čia $b \in \mathcal{A}$ – pasirinkta bazinė arba atskaitos (*angl. reference*) būseną (bazinė alfabeto $\mathcal{A}$ raidė).

Tegu $\psi: \mathcal{X} \rightarrow \mathbf{R}$ yra bet kokia funkcija. Tada Mobiuso formulė (tapatybė) jai užrašoma taip:

$$\psi(x) = \sum_{S: S \subset I_n} \sum_{J: J \subset S} (-1)^{|S|-|J|} \psi(H_S(x)) =: \sum_{S: S \subset I_n} \lambda_S(x_S), \quad (1.11)$$

$$\lambda_S(x_S) := \sum_{J: J \subset S} (-1)^{|S|-|J|} \psi(H_S(x)), \quad x_J = (x_j, j \in J), \quad x \in \mathcal{X}, \quad (1.12)$$

čia $|J|$ žymi aibės J elementų skaičių (galia). Formulė atrodo gana įmantri, bet iš tikrųjų ji yra tik žemiau pateiktų paprastų tapatybių (1.13) – (1.14) apibendrinimas:

$$\begin{aligned} \psi(x_1) &= \psi(b) + (\psi(x_1) - \psi(b)) =: \lambda_\emptyset + \lambda_1(x_1); \\ \psi(x_1, x_2) &= \psi(b, b) + (\psi(x_1, b) - \psi(b, b)) + (\psi(b, x_2) - \psi(b, b)) \\ &\quad + (\psi(x_1, x_2) - \psi(x_1, b) - \psi(b, x_2) + \psi(b, b)) \\ &=: \lambda_\emptyset + \lambda_1(x_1) + \lambda_2(x_2) + \lambda_{12}(x_1, x_2); \end{aligned} \quad (1.13)$$

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI MATEMATINIAI MODELIAI

17

$$\begin{aligned}
 \psi(x_1, x_2, x_3) &= \psi(b, b, b) + (\psi(x_1, b, b) - \psi(b, b, b)) \\
 &+ (\psi(b, x_2, b) - \psi(b, b, b)) \\
 &+ (\psi(b, b, x_3) - \psi(b, b, b)) + (\psi(x_1, x_2, b) \\
 &- \psi(x_1, b, b) - \psi(b, x_2, b) + \psi(b, b, b)) \\
 &+ (\psi(x_1, b, x_3) - \psi(x_1, b, b) - \psi(b, b, x_3) + \psi(b, b, b)) \\
 &+ (\psi(b, x_2, x_3) - \psi(b, x_2, b) - \psi(b, b, x_3) + \psi(b, b, b)) \\
 &+ (\psi(x_1, x_2, x_3) - \psi(x_1, x_2, b) - \psi(x_1, b, x_3) - \psi(b, x_2, x_3) \\
 &+ \psi(x_1, b, b) + \psi(b, x_2, b) + \psi(b, b, x_3) - \psi(b, b, b)) \\
 &=: \lambda_\emptyset + \lambda_1(x_1) + \lambda_2(x_2) + \lambda_3(x_3) + \lambda_{12}(x_1, x_2) \\
 &+ \lambda_{13}(x_1, x_3) + \lambda_{23}(x_2, x_3) + \lambda_{123}(x_1, x_2, x_3), \quad (1.14)
 \end{aligned}$$

čia trumpumo dėlei vietoje $\lambda_{\{i,j,\dots\}}$ rašoma $\lambda_{ij\dots}$.

Funkcijos λ_J naudingos tuo, kad jos virsta 0, jeigu kuriam nors $j \in J$ funkcija ψ nepriklauso nuo x_j .

1.2.2. Apibendrinto tiesinio modelio apibrėžimas

Turime stebinius $(Y(t), X(t)), t = 1, \dots, N$, čia $Y(t) \in \mathbf{R}^n$ yra pagrindinis tyrimo kintamasis, o $X(t) \in \mathbf{R}^k$ yra papildoma informacija ($t = 1, \dots, N$). Pažymėkime $X = (X(1), \dots, X(N))$, $Y = (Y(1), \dots, Y(N))$.

3 apibrėžimas. Sakoma, kad stebiniai Y su papildomais duomenimis X tenkina apibendrintąjį tiesinį modelį (ATm), jeigu:

1) $(Y(t), t = 1, \dots, N)$ yra sąlyginai nepriklausomi, kai žinomos atitinkamos aiškinančiųjų kintamųjų $(X(t), t = 1, \dots, N)$ reikšmės, ir $Y(t)$ sąlyginis skirstinys priklauso eksponentinei skirstinių šeimai su kanoniniu parametru $\theta(t) = \theta(t | X(t)) \in \Theta$ ($t = 1, \dots, N$).

2) Apibrėžtas tiesinis prediktorius $\eta(t) := BX(t)$, čia B yra $n \times k$ nežinomų ATm parametrų matrica.

3) Nusakytos sąryšio funkcijos (*angl. link function*) $g_t: \mathbf{R}^n \rightarrow \mathbf{R}^n$, susiejančios sąlyginį $Y(t)$ vidurkį (kai žinomas $X(t)$) $\mu(t) = \mu(t | \theta(t)) = \mu(t | X(t)) := \mathbf{E}[Y(t) | X(t)]$ su tiesiniu prediktoriumi $\eta(t)$ lygybe:

$$g_t(\mu(t)) = \eta(t), \quad t = 1, \dots, N. \quad (1.15)$$

1.2.3. Binarinė logistinė regresija

Tegu Y žymi atsitiktinį dydį, įgyjantį dvi reikšmes: 0 ir 1, t. y. laikoma, kad ATm klasėje priklausomo kintamojo Y skirstinys yra Binominis:

$$Y(t) \Big|_X \sim \text{Binomial}(1, p_t), \quad p_t = p(X(t)), \quad t = 1, \dots, N,$$

$$l_1(p) := \ln(f(Y | p)) = Y \ln p + (1 - Y) \ln(1 - p).$$

Tada vienintelė tikimybinė charakteristika, nusakanti įvykį, yra jo tikimybė $p = P(Y = 1)$, nes $EY = p$.

Tegu $X \in R^m$ yra prediktorių vektorius. Natūralu laikyti, kad prediktoriai įtakoja įvykio tikimybę, t. y. $p = p(X) = P(Y = 1|X)$. Logistinės regresijos modelyje ši priklausomybė aprašoma logistine funkcija, priklausančia nuo nežinomo parametro $\beta \in R^m$:

$$p(X) = p(X|\beta) = \frac{e^{\beta' X}}{1 + e^{\beta' X}}.$$

Atvirkštinė logistinei funkcijai yra vadinama logit funkcija, kuri susieja tikimybę $p(X)$ su prediktoriais tokiu būdu:

$$\text{logit}(p(X)) = \ln \left(\frac{p(X)}{1 - p(X)} \right) = \beta' X. \quad (1.16)$$

Santykis po logaritmo ženklų (1.16) vadinamas *šansu*, *galimybe* arba *šansais*. Ši sąvoka yra svarbi, interpretuojant logistinės regresijos rezultatus. Ji parodo, kiek kartų įvykio $Y = 1$ tikimybė yra didesnė (mažesnė) už įvykio $Y = 0$ tikimybę prie sąlygų X :

$$\Lambda_{1|0}(X) = \frac{p}{1 - p} = \frac{P(Y = 1|x)}{P(Y = 0|x)} = e^{\beta' X}, \quad p = \frac{\Lambda_{1|0}(X)}{1 + \Lambda_{1|0}(X)}.$$

Taip pat svarbi sąvoka rezultatų interpretavime yra šansų santykis (*angl. odds ratio*). Tarkime, kad aiškinamasis kintamasis x įgyja tik dvi reikšmes $x = 1$ ir $x = 0$. Tada šansų santykis apibrėžiamas kaip įvykio $Y = 1$ dviejų šansų $\Lambda_{1|0}(1)$ ir $\Lambda_{1|0}(0)$ santykis, kai $\Lambda_{1|0}(1)$ skaičiuojamas prie sąlygų, kad egzistuoja požymio x įtaka, o $\Lambda_{1|0}(0)$ – kai požymio x poveikis neįtraukiamas

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI MATEMATINIAI MODELIAI

19

į modelį:

$$\begin{aligned} OR = \frac{\Lambda_{1|0}(1)}{\Lambda_{1|0}(0)} &= \frac{P(Y=1|x=1)}{P(Y=0|x=1)} \cdot \frac{P(Y=1|x=0)}{P(Y=0|x=0)} \\ &= \frac{P(Y=1|x=1)}{P(Y=0|x=1)} \cdot \frac{P(Y=0|x=0)}{P(Y=1|x=0)}. \end{aligned} \quad (1.17)$$

Jeigu įvykio tikimybė maža, tai šansų santykis apytiksliai lygus atitinkamų įvykių tikimybių santykiui. Šansų santykio įvertis gaunamas tiesiogiai iš parametro β didžiausio tikėtimumo įverčio $\hat{\beta}$. Šiuo atveju, kai turime tik vieną aiškinamąjį kintamąjį x , įgyjantį tik dvi reikšmes $x=1$ ir $x=0$, logit įvertinti modeliai šansui $\hat{\Lambda}_{1|0}(1)$ ir $\hat{\Lambda}_{1|0}(0)$ bus atitinkami:

$$\hat{\Lambda}_{1|0}(1) = \frac{P(Y=1|x=1)}{P(Y=0|x=1)} = e^{\hat{\beta}_0 + \hat{\beta}_1 x}, \quad \hat{\Lambda}_{1|0}(0) = \frac{P(Y=1|x=0)}{P(Y=0|x=0)} = e^{\hat{\beta}_0},$$

$$\hat{OR} = \frac{\hat{\Lambda}_{1|0}(1)}{\hat{\Lambda}_{1|0}(0)} = e^{\hat{\beta}_1 x}.$$

Logtikėtimumo funkcija:

$$\begin{aligned} l(p) &= l(p | Y^N, X) = \sum_{t=1}^N [Y_t \ln p_t - (1 - Y_t) \ln(1 - p_t)] \\ &= \sum_{t=1}^N Y_t \ln \frac{p_t}{1 - p_t} + \sum_{t=1}^N \ln(1 - p_t) \\ &= \sum_{t=1}^N \beta' X(t) Y(t) - \sum_{t=1}^N \ln(1 + e^{\beta' X(t)}). \end{aligned}$$

1.2.4. Puasono regresija

Puasono regresiją ATm klasėje turime tada, kai atsako kintamojo Y skirstinys yra Puasono, t. y:

$$Y \sim \text{Poisson}(\lambda) \implies f(y | \lambda) = \frac{\lambda^y}{y!} e^{-\lambda},$$

$$\ln f(y | \lambda) = y \ln \lambda - \lambda - \ln y! \quad y = 0, 1, \dots$$

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI MATEMATINIAI MODELIAI

Vadinasi, kanoninis parametras yra $\theta = \ln(\lambda)$, o logtikėtinumo funkcija:

$$l(\lambda) = l(\lambda \mid Y^N, X) = \sum_{t=1}^N [Y(t) \beta' X(t) - e^{\beta' X(t)}].$$

1.2.5. Logtiesiniai modeliai

Logtiesiniai modeliai, priklausantys ATm klasei, aprašo sudėtingus kategorinių požymių tarpusavio sąryšius ir skirti dažnių lentelių daugiamatei statistinei analizei (Bergsma 1997). Šiuose modeliuose nėra kintamųjų klasifikavimo į aiškinamuosius ir aiškinančiuosius. Juose visi kintamieji traktuojami kaip aiškinantieji, o aiškinamuoju kintamuoju laikomas konkrečios kategorinių kintamųjų reikšmių (būsenų) kombinacijos stebėtas dažnis. Ir nors logtiesinių modelių analizės rezultatai primena standartines dispersinės ir regresinės analizės lenteles, tačiau bendru atveju jie žymiai sudėtingesni, nes didesnio nei 3 matavimo duomenims pilnai nusakyti neužtenka apibrėžti visų kintamųjų porų tarpusavio priklausomybių, o reikia aprašyti ir (visas) aukštesnės negu antrosios eilės kintamųjų tarpusavio sąveikas (*angl. interactions*).

Logtiesiniu modeliu vadinamas toks stebinių modelis, kai tų stebinių vidurkio logaritmas yra tiesinė funkcija nuo nežinomų modelio parametrų. Pateiksime formalų apibrėžimą:

Tegu $\{Y_j, j \in J\}$ yra stebiniai su reikšmėmis erdvėje $\mathbf{R}$. Čia J yra baigtinė indeksų aibė, turinti N elementų, $N := |J|$. Indeksų aibėje J įveskime kokią nors numeravimo tvarką, kurią žymėsime simboliu $\tau, \tau: \{1, \dots, N\} \rightarrow J$, ir ją fiksuokime. Tada $Y := (Y_j, j \in J)$ žymės N -matį atsitiktinį vektorių, kurio komponentės yra τ tvarka išdėstyti pradiniai stebiniai $\{Y_j, j \in J\}$. Tiksliau $Y := (Y_{\tau(i)}, i = 1, \dots, N)$.

Pažymėkime $\mu_j := \mathbf{E}Y_j, j \in J$ ir analogiškai kaip ir vektorių Y apibrėžkime N -matį vektorių μ , t. y. $\mu = \mathbf{E}Y \in \mathbf{R}^N$. Tarę, kad $\mu > 0$, galime apibrėžti ir vektorių $\log(\mu) \in \mathbf{R}^N$ (čia laikoma, kad visos operacijos, tiek palyginimo, tiek ir logaritmovimo, atliekamos pakomponenčiai). Kai Y yra stebėti dažniai, prielaida $\mu > 0$ yra natūrali, bet nevisada būna išpildyta.

4 apibrėžimas. Sakoma, kad stebiniai $\{Y_j, j \in J\}$ tenkina logtiesinį modelį, jeigu egzistuoja tiesinis poerdvis $\mathcal{M} \subset \mathbf{R}^N$, kurio matavimas $k := \dim(\mathcal{M}) < N$ toks, kad $\log(\mu) \in \mathcal{M}$.

1 pavyzdys. Tegu tiriami požymiai A ir B , $A \in \{1, 2\}$, $B \in \{1, 2, 3\}$. Laikysime, kad numeravimo tvarka τ požymių A ir B dažnių lentelėje $\mathcal{L} = \mathcal{L}(A, B)$ yra numeracija iš kairės į dešinę ir iš viršaus į apačią, o Y_j yra stebėtas

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI
MATEMATINIAI MODELIAI

21

dažnis ląstelėje $j \in J$, indeksų aibė $J := \{1,2\} \times \{1,2,3\}$ su $N = 6$:

$A \backslash B$	1	2	3
1	Y_{11}	Y_{12}	Y_{13}
2	Y_{21}	Y_{22}	Y_{23}

Trumpumo dėlei pažymėsime $Y_{(i)} := Y_{\tau(i)}$ ($i = 1, \dots, 6$), $Y_{(i)}$ yra stebėta reikšmė i -ojoje pagal tvarką τ lentelės $\mathcal{L}$ ląstelėje. Tada gauname, kad:

$$\begin{array}{cccccc} Y_{(1)} & Y_{(2)} & Y_{(3)} & Y_{(4)} & Y_{(5)} & Y_{(6)} \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ Y_{1,1} & Y_{1,2} & Y_{1,3} & Y_{2,1} & Y_{2,2} & Y_{2,3} \end{array}$$

ir

$$Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \\ Y_{(3)} \\ Y_{(4)} \\ Y_{(5)} \\ Y_{(6)} \end{pmatrix} = \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ Y_{1,3} \\ Y_{2,1} \\ Y_{2,2} \\ Y_{2,3} \end{pmatrix}.$$

1 pastaba. Nagrinėjant logtiesinius modelius kartais patogų logtiesinių modelių šeimą praplėsti, papildant ją *fiktyviu* (neparametriniu) modeliu, kuriam $\mathcal{M} = \mathbf{R}^N$. Tokie modeliai vadinami pilnaisiais, prisotintais (*angl. saturated*) logtiesiniais modeliais.

2 pastaba. Palyginimui paimkime daugialypės tiesinės regresijos modelį su $N \times k$ plano matrica $\mathbf{X}$:

$$Y = \mathbf{X}\beta + \varepsilon, \quad \beta \in \mathbf{R}^k, Y \in \mathbf{R}^N, \quad \mu := \mathbf{E}Y = \mathbf{E}[Y \mid \mathbf{X}] = \mathbf{X}\beta,$$

Taigi, $\mathcal{M} := \{\mathbf{X}\beta, \beta \in \mathbf{R}^k\}$ yra tiesinis poerdvis, generuotas plano matricos $\mathbf{X}$. Vadinasi, $\mu \in \mathcal{M} \subset \mathbf{R}^N$ ir $\dim(\mathcal{M}) = k$.

Logtiesinės modelio specifikacijos privalumas lyginant ją su įprasta daugialypės tiesinės regresijos modelio specifikacija yra tas, kad pirmuoju atveju aiškinančiųjų kintamųjų vektoriai nėra tiksliai nusakyti, juos galima pasirinkti laisvai: jais gali būti bet kurie tiesiškai nepriklausomi vektoriai $Z_1, \dots, Z_k$, priklausantys poerdviui $\mathcal{M}$. Tai ypač patogus, kai reikia įvesti fiktyvius kintamuosius.

3 pastaba. Logtiesinio modelio apibrėžime stebinių vektorių Y skirstinys nėra nusakytas, tam tikri apribojimai įvedami tik jo vidurkiui. Kokybinių požymių statistinėje analizėje Y skirstinys paprastai būna (nepriklaus-

22 1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI MATEMATINIAI MODELIAI

somų) Puasono skirstinių sandauga (trumpai jį vadinsime Puasono ir žymėsime $Y \sim Poisson$), polinominis skirstinys (žymėsime $Y \sim Multinomial$) arba polinominių skirstinių sandauga (žymėsime $Y \sim MultiProduct$). Visi 3 atvejai priklauso ATm klasei. Jeigu galioja Puasono logtiesinis modelis, tai logtikėtinumo funkcija yra:

$$l(\mu) = l(\mu | Y) := \sum_{j \in J} (Y_j \log \mu_j - \mu_j) + C(Y), \quad (1.18)$$

čia $C(Y)$ žymi statistiką, nesusijusią su modelio parametrais.

Rasime logtikėtinumo funkciją polinominiame logtiesiniame modelyje. Tegul $Y \sim Multinomial_N(n, p)$,

$$p \in \mathcal{P}_N := \left\{ p \in \mathbf{R}^N : p_j > 0, \forall j \in J, \sum_{j \in J} p_j = 1 \right\}.$$

Tada tikėtinumo funkcija atžvilgiu parametro $p \in \mathcal{P}_N$ užrašoma taip:

$$L(p) = L(p | Y) := \frac{N!}{\prod_{j \in J} Y_j!} \prod_{j \in J} p_j^{Y_j}.$$

Kadangi šiuo atveju $\mu_j = n \cdot p_j$, čia $n = Y_+ := \sum_{j \in J} Y_j$, tai pakeitę modelio parametrizaciją su p į parametrizaciją su μ , gauname:

$$\begin{aligned} l(\mu) = l(\mu | Y) &:= \log(n!) - n \log(n) + \sum_{j \in J} \ln(Y_j!) + \sum_{j \in J} Y_j \ln \mu_j = \\ &= \sum_{j \in J} Y_j \ln \mu_j + \tilde{C}(Y). \end{aligned} \quad (1.19)$$

Tokiu būdu, Puasono ir polinominio skirstinio atveju logtikėtinumo funkcija turi tą patį pavidalą $l(\mu) = \sum_{j \in J} Y_j \ln \mu_j - \mu_+ + C(Y)$, kuriame nuo skirstinio priklauso tik funkcija $C(Y)$ ir polinominio skirstinio atveju galioja papildomas apribojimas:

$$\mu_+ = Y_+, \quad \mu_+ := \sum_{j \in J} \mu_j. \quad (1.20)$$

Kadangi funkcija $\log(\cdot)$ yra įgaubta, tai logtikėtinumo funkcija $l(\mu)$ taip pat yra įgaubta atžvilgiu μ . Todėl didžiausio tikėtinumo įvertinių aibė yra iškila

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI
MATEMATINIAI MODELIAI

23

ir susideda iš vienintelio taško prie gana bendrų sąlygų. Kai $\mathcal{M} = \{Z = \mathbf{X}\beta, \beta \in \mathbf{R}^k\}$, tai tikėtinumo lygtis parametrai $l(\mu)$ atrodo taip:

$$\mathbf{X}'Y = \mathbf{X}'\hat{\mu}_{DT}.$$

Padauginę abi šios lygties puses iš matricos $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ (laikoma, kad $\mathbf{X}'\mathbf{X}$ yra neišsigimusi, t. y. $\text{rank}\mathbf{X} = k$), gauname $\Pi_{\mathbf{X}}Y = \Pi_{\mathbf{X}}\hat{\mu}_{DT}$, ir, vadinasi,

$$\Pi_{\mathcal{M}}Y = \Pi_{\mathcal{M}}\hat{\mu}_{DT}, \quad (1.21)$$

čia $\Pi_{\mathcal{M}} := \Pi_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ yra projektorius į tiesinį poerdvį $\mathcal{M}$.

Iš (1.21) matosi, kad sąlyga (1.20) su $\mu = \hat{\mu}_{DT}$ yra išpildyta automatiškai, jeigu tik vienetų vektorius $(1, 1, \dots, 1)'$ priklauso tiesiniam poerdviui $\mathcal{M}$. Vadinasi, tokiu atveju Puasono ir polinominis logtiesiniai modeliai yra ekvivalentūs, jų pagrindų gauti didžiausio tikėtinumo įvertiniai sutaps.

Tokiu būdu, kai išpildyta sąlyga

$$\mathbf{1}_N := (1, 1, \dots, 1)' \in \mathcal{M}, \quad (1.22)$$

$$l(\mu) \longrightarrow \max_{\log \mu \in \mathcal{M}} \iff G(\mu) := 2 \sum_{j \in J} y_j \log \frac{y_j}{\mu_j} \longrightarrow \min_{\log \mu \in \mathcal{M}}.$$

Matome, kad didžiausio tikėtinumo įvertinys $\hat{\mu}_{DT}$ minimizuoja logtikėtinumo santykio statistikos $G = G(\mu)$ reikšmę. Kadangi funkcija $\log(1/z) = -\log z$ yra iškila, tai ir statistika $G(\mu)$ yra iškila atžvilgiu μ .

Logtiesinių modelių interpretacija. Nagrinėkime 3-jų kategorinių požymių A, B ir C dažnių lentelę $\mathcal{L} = \mathcal{L}(A, B, C)$. Šiuo atveju pilnasis (prisotintas) logtiesinis modelis $\mathcal{M} = \mathcal{M}(A, B, C)$ bus:

$$\log \mu_{ijs} = u_0 + u_i^A + u_j^B + u_s^C + u_{ij}^{AB} + u_{is}^{AC} + u_{js}^{BC} + u_{ijs}^{ABC}.$$

1.2 lentelėje pateikti visi galimi logtiesinių modelių atvejai trims kokybiniais požymiais A, B ir C . Pateiksime jų interpretaciją:

0 modelis. A, B ir C yra tarpusavyje (statistiškai) nepriklausomi.

1 modelis. požymis A nepriklauso nuo poros B ir C , bet B ir C yra tarpusavyje priklausomi.

2 ir 3 modeliai turi analogišką interpretaciją kaip ir 1 modelis.

4 modelis. Šiuo atveju A ir B yra sąlyginai nepriklausomi, kai žinomas C . Kitaip tariant, A ir B tarpusvio priklausomumas pasireiškia tik per C .

5 ir 6 modeliai turi analogišką interpretaciją kaip ir 4 modelis.

7 modelis neturi paprastos interpretacijos: šansų (galimybių) santykiai dvi-

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI MATEMATINIAI MODELIAI

24

matėse lentelėse $\mathcal{L}(AB)$, kai duota požymio C reikšmė nepriklauso nuo tos reikšmės.

8 modelis yra pilnasis modelis.

1.2 lentelė. Galimi logtiesiniai modeliai trijų požymių dažnių lentelėje

Table 1.2. Loglinear models for three-way contingency table

Nr.	Specifikacija
0	$\log \mu_{ijs} = u_0 + u_i^A + u_j^B + u_s^C$
1	$\log \mu_{ijs} = u_0 + u_i^A + u_j^B + u_s^C + u_{js}^{BC}$
2	$\log \mu_{ijs} = u_0 + u_i^A + u_j^B + u_s^C + u_{is}^{AC}$
3	$\log \mu_{ijs} = u_0 + u_i^A + u_j^B + u_s^C + u_{ij}^{AB}$
4	$\log \mu_{ijs} = u_0 + u_i^A + u_j^B + u_s^C + u_{is}^{AC} + u_{js}^{BC}$
5	$\log \mu_{ijs} = u_0 + u_i^A + u_j^B + u_s^C + u_{ij}^{AB} + u_{js}^{BC}$
6	$\log \mu_{ijs} = u_0 + u_i^A + u_j^B + u_s^C + u_{ij}^{AB} + u_{is}^{AC}$
7	$\log \mu_{ijs} = u_0 + u_i^A + u_j^B + u_s^C + u_{ij}^{AB} + u_{is}^{AC} + u_{js}^{BC}$
8	$\log \mu_{ijs} = u_0 + u_i^A + u_j^B + u_s^C + u_{ij}^{AB} + u_{is}^{AC} + u_{js}^{BC} + u_{ijs}^{ABC}$

1.2.6. Apibendrintas (daugialypis) logit modelis

Kai nominalus atsako kintamasis įgyja daugiau nei dvi kategorijas, turime taip vadinamą apibendrintą logit arba daugialypį logit modelį (*angl. generalized logits*). Logistinė regresija yra atskiras jo atvejis – kai stebimos tik dvi atsako kintamojo kategorijos.

Tegu stebime kokybinį kintamąjį A , kuris įgyja reikšmes iš $J := \{1, \dots, m\}$ ir turime stebinius $(A(t), X(t))$, $t \in \mathcal{T} := \{1, \dots, N\}$, iš kurių suskaičiuojame dažnius $Y_j(X^*) = |\{t : A(t) = j, X(t) = x^*\}|$, $x^* \in \mathcal{X}$, čia aibė $\mathcal{X}$ yra kintamųjų X reikšmių visų stebėtų skirtingų kombinacijų aibė.

Tegu: $p_j(x) := \mathbf{P}\{A(t) = j \mid X(t) = x\} = \mathbf{P}\{Y_j(t) = 1 \mid X(t) = x\}$, $\forall t \in \mathcal{T}, x \in \mathcal{X}$. Tada binominis logistinės regresijos modelis gaunamas kai $m = 2$:

$$\log \left(\frac{p(x)}{1 - p(x)} \right) = \beta' X, \quad p(x) := \mathbf{P}\{Y = 1 \mid x\}, \quad x \in \mathcal{X}. \quad (1.23)$$

O apibendrintas logit arba daugialypis (multinominis) logit modelis (ALm)

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI MATEMATINIAI MODELIAI

25

turi tokį pavidalą:

$$\log \frac{p_j(x)}{p_m(x)} = \beta'_j X, \quad x \in \mathcal{X}, j \in J, j \neq m, \quad (1.24)$$

čia m yra bazinė reikšmė (*angl. reference value*), pasirinkta iš aibės J . Bazinė reikšmė gali būti bet kuris (stebėtas) aibės J elementas, bet patogumo dėlei (statistiniuose paketuose) paprastai pasirenkama paskutinė kategorija pagal ko-
kią nors aibės J elementų išdėstymo tvarką (pavyzdžiui, abėcėlinę).

ALm aprašo šansų, lyginant su bazine kategorija, priklausomybę nuo aiš-
kinamųjų kintamųjų X . Parodysime, kad ALm yra logtiesinio modelio atskiras
atvejis. Imsime atskirą atvejį kai $m = 2$, t. y. kai ALm yra logistinės regresijos
modelis. Tegu

$$\log \frac{p_{1X}}{1 - p_{1X}} = \beta' X, \quad \beta \in \mathbf{R}^k, X \in \mathcal{X}.$$

Tada atitinkamas logtiesinis modelis užrašomas taip:

$$\begin{cases} \log p_{1X} = \beta' X + \eta_X, \\ \log p_{2X} := \log(1 - p_{1X}) =: \eta_X, \end{cases} \quad X \in \mathcal{X}.$$

Kadangi $Y \in \{1, 2\}$ ir turime $M := |\mathcal{X}|$ skirtingų stebinių X reikšmių, tai
tikimybių p_{iX} iš viso yra $2M$. Tokiu būdu, $\dim(\mathcal{M}) = k + M \ll 2M$,
kai M yra (santykinai) didelis lyginant su k (k yra parametrų vektoriaus β
dimensija).

ALm modeliuoja sąlyginį aiškinamojo kintamojo Y skirstinį, kai yra ži-
nomos aiškinamųjų kintamųjų X reikšmės. Taigi, ALm yra klasikinių regresi-
nės ir dispersinės analizės modelių analogas. Šiuo atveju atskirai apie patį X
skirstinį jokių prielaidų nedaroma, jis gali būti bet koks.

Jeigu kintamieji X taip pat yra kategoriniai, tai kintamojo Y ALm galima
interpretuoti, kaip bendrą kintamųjų Y ir X logtiesinį modelį. Kadangi Y AL
modelyje X skirstinys gali būti bet koks, tai jis aprašomas pilnuoju (priso-
tintu) modeliu, kuriuo ir papildomas pradinis sąlyginis kintamojo Y modelis.
Taip gaunamas bendras kintamųjų Y ir X logtiesinis modelis, ekvivalentus
pradiniam Y AL modeliui. Tuo lengva įsitikinti iš gauto logtiesinio modelio
išskaičiavus sąlyginį kintamojo Y modelį kai žinomi X .

1 teiginys. Apibendrintas aiškinamojo kategorinio kintamojo Y su aiški-
nančiaisiais kategoriniais kintamaisiais X logit modelis yra ekvivalentus ben-
dram kintamųjų Y ir X logtiesiniam modeliui ir duoda tiksliai tuos pačius
statistinės analizės rezultatus, tiek kokybiškai, tiek kiekybiškai, jeigu į minėtą

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI 26 MATEMATINIAI MODELIAI

logtiesinį modelį kaip jo atskira dalis įeina pilnasis aiškinančiųjų požymių X modelis.

Šis teiginys yra gerai žinomas, nors surasti tinkamą nuorodą būtų, matyt, sudėtinga. Jis tiesiogiai išplaukia iš logtiesinio modelio, apibendrinto logit modelio ir sąlyginės tikimybės apibrėžimų.

2 pavyzdys. Tegu X yra fiktyvusis kintamasis, atitinkantis kokybinius požymius B ir C , o tiriamo požymius A, B ir C . Tada ALm galima užrašyti tokioje formoje:

$$\log \frac{p_{aX}}{p_{mX}} = \alpha_a + \alpha_{ai}^B + \alpha_{aj}^C + \alpha_{aij}^{BC},$$

čia $X = (i, j)'$, i yra B galima reikšmė, o j yra C galima reikšmė. Jeigu A gali įgyti tik dvi reikšmes, tai indeksą a galima praleisti (eliminuoti), ir gauname:

$$\lambda_X := \log \left(\frac{p_{1X}}{1 - p_{1X}} \right) = \alpha_0 + \alpha_i^B + \alpha_j^C + \alpha_{ij}^{BC}. \quad (1.25)$$

Galimos modelio 1.25 specifikacijos ir apibendrinto logit modelio atitinkamybė logtiesiniam modeliui parodyta 1.3 lentelėje. Pavyzdžiui, 1.3 lentelėje 3-io modelio simbolinis užrašas yra: $\lambda_t = \alpha_0 + \alpha_i^B + \alpha_j^C$

1.3 lentelė. Atitinkamybė tarp apibendrinto logit ir logtiesinio modelio

Table 1.3. Relationship between generalized logit and loglinear model

ALm	Logtiesinis modelis
(0) $A = \emptyset$	$[A][BC]$
(1) $A = B$	$[BC][BA]$
(2) $A = C$	$[BC][CA]$
(3) $A = B + C$	$[BC][CA][BA]$
(4) $A = B + C + BC$	$[ABC]$

Modelis $[A][BC]$ reiškia, kad požymis A ir požymiai B, C yra tarpusavyje (statistiškai) nepriklausomi. Savo ruožtu, $[BC][BA]$ tvirtina, kad požymiai A ir C yra tarpusavyje nepriklausomi, kai yra žinomos požymio B reikšmės. Sąveika tarp A ir C pasireiškia tik per B .

1.2.7. Logtiesiniuose modeliuose naudojamos statistikos

Hipotezėms apie logtiesinio modelio parametrų ir jų sąveikų statistinį reikšmingumą, o taip pat parinkto modelio suderinamumą (*angl. goodness of fit*)

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI MATEMATINIAI MODELIAI

27

tikrinti naudojamos statistikos, kurios lygina stebimus dažnius Y_{ij} dažnių lentelėje su tikėtiniais dažniais $\hat{m}_{ij}$ (indeksai (ij) reiškia, kad šiuo atveju dažniai užrašomi dviejų požymių dažnių lentelėje (1.1). Įprasta tam naudoti Pearson χ^2 arba G^2 statistiką:

$$\chi^2 = \sum_{ij} \frac{(Y_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}, \quad G^2 = 2 \sum Y_{ij} \log(Y_{ij}/\hat{m}_{ij}). \quad (1.26)$$

Kai pasirinktas modelis yra teisingas, abi statistikos (1.26) turi (prie tam tikrų sąlygų) asimptotinę χ^2 skirstinį su laisvės laipsnių skaičiumi, lygiu ląstelių kiekio dažnių lentelėje ir į modelį įtrauktų parametru skaičiaus skirtumui. Statistika G^2 , lyginant su Pearsono χ^2 statistika, turi svarbų privalumą, nes yra adityvi ir gali būti taikoma lygiai taip pat, kaip kvadratų sumų statistikos dispersinėje analizėje.

Read ir Cressie (1988) įvedė naują statistikų šeimą vadinamą divergencijos (nukrypimo) statistika (*angl. power divergence statistics*):

$$\frac{2}{\lambda(\lambda + 1)} \sum_{ij} Y_{ij} \left[\left(\frac{Y_{ij}}{\hat{m}_{ij}} \right)^\lambda - 1 \right]. \quad (1.27)$$

Šis bendras statistikos atvejis (1.27) yra lygus χ^2 statistikai kai $\lambda = 1$ ir G^2 statistikai – kai $\lambda \rightarrow 0$. Read ir Cressie (1998) rekomendavo naudoti statistiką (1.27) su $\lambda = 2/3$, parodydami, kad ji mažiau jautri išsklaidytų duomenų efektui.

1.2.8. Apibendrintas logit modelis ir Markovo laukai

Ryšį tarp Markovo atsitiktinių laukų ir logtiesinių modelių pirmieji pastebėjo ir aprašė Darroch, Lauritzen ir Speed (1980). Pastaruoju metu labai plačiai taikoma diskrečių atsitiktinių Markovo laukų teorija, kuri remiasi Brooko lema apie diskrečių atsitiktinių laukų su teigiamomis visų būsenų tikimybėmis (t. y. pilna skirstinio atrama) lokalia specifikacija ir Hammersley-Clifford teorema apie ryšį tarp Markovo ir Gibso atsitiktinių laukų.

Toliau įvesime reikalingus apibrėžimus ir trumpai aprašysime šios teorijos sąryšį su apibendrintais (daugialypiais) logit (ir logtiesiniais) modeliais.

Terminas *atsitiktinis laukas* yra stochastinio proceso $\{Y_t : t \in T \subset R\}$ apibendrinimas. Atsitiktinį lauką taip pat sudaro atsitiktinių dydžių rinkinys $\{Y_v : v \in V\}$, tik čia indeksų aibė V priklauso nebūtinai realių skaičių tiesei,

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI
28 MATEMATINIAI MODELIAI

pavyzdžiui, V gali būti plokštumos R^2 poaibis, grafo viršūnių aibė ar DNR sekų atveju – nukleotidų aibė.

Tegu $Y := \{Y(v), v \in V\}$ (V reiškia „viršūnės“) yra duotas atsitiktinių dydžių $Y(v)$ su reikšmėmis baigtinėje būsenų aibėje $\mathcal{A}_v$ rinkinys ($v \in V$). Čia V žymi abstrakčią (baigtinę) indeksų aibę, $n := |V|$ yra aibės V elementų skaičius. Atsitiktinių dydžių rinkinį Y traktuosime kaip aibės $\mathcal{A}^V := \prod_{v \in V} \mathcal{A}_v$ elementą.

Indeksų aibėje V įvesime diskrečiąją topologiją $\mathcal{B}$, kuri yra nusakoma neorientuotu, be kilpų grafu $(V, \mathcal{B})$ ir yra susieta su aibės V kaimyninių elementų (aplinkų) rinkiniu $\mathcal{U} := \{U_\ell, \ell \in V\}$. Aibėje $U_\ell \subset V$ yra išvardinti visi pozicijos ℓ kaimynai (grafo viršūnės). Viršūnė v priklauso U_ℓ tada ir tik tada, kai egzistuoja briauna $b \in \mathcal{B}$, jungianti viršūnes v ir ℓ . Taigi, aplinkų (kaimynų) sistema $\mathcal{U}$ tenkina tokias sąlygas: $\ell \notin U_\ell$, ir $v \in U_\ell$ tada ir tik tada, kai $\ell \in U_v$.

Toliau mums svarbi bus aibės V poaibių, sudarytų iš grafo $(V, \mathcal{B})$ klikų viršūnių, sistema $\mathcal{K}$. (Grafo $\mathcal{G}$ klikomis vadinami jo „maksimalūs“ jungūs poagrafių. Pažymėkime:

$$Y(U) := \{Y(u), u \in U\}, \quad W_{-v} := W \setminus \{v\}, \quad v \in W \subset V.$$

5 apibrėžimas. Diskretus atsitiktinis laukas $Y = Y(V)$ vadinamas diskrečiu Markovo atsitiktiniu lauku (MAL) atžvilgiu grafo $(V, \mathcal{B})$ arba topologijos $\mathcal{B}$ (arba Markovo atsitiktiniu lauku su kaimynų sistema $\mathcal{U}$), jeigu su visais $v \in V$ atsitiktinis dydis $Y(v)$ ir atsitiktiniai dydžiai $Y(V_{-v} \setminus U_v)$ yra sąlyginai nepriklausomi, kai yra žinomos dydžių $Y(U_v)$ reikšmės. Trumpai tai užrašoma taip:

$$Y(v) \perp Y(V_{-v} \setminus U_v) \mid Y(U_v). \quad (1.28)$$

Vadinasi, Markovo atsitiktiniam laukui Y su visais $a \in \mathcal{A}^V$ galioja lygybės

$$\begin{aligned} \mathbf{P}\{Y(v) = a(v) \mid Y(V_{-v}) = a(V_{-v})\} = \\ \mathbf{P}\{Y(v) = a(v) \mid Y(U_v) = a(U_v)\}. \end{aligned} \quad (1.29)$$

Kai $V \subset \{1, \dots, n\}$ ir Y yra k -osios eilės Markovo laukas (grandinė), tai aibių rinkinys $\mathcal{K}$ yra sudarytas iš aibių $[i, i+k]$, $i = 1, \dots, n-k$. Čia patogumo dėlei pažymėta $[i, i+k] := \{i, \dots, i+k\}$.

Iš (1.29) matome, kad MAL lokali charakteristika

$$g_v(a(v) \mid a(U_v)) := \mathbf{P}\{Y(v) = a(v) \mid Y(U_v) = a(U_v)\}, \quad a \in \mathcal{A}^V \quad (1.30)$$

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI
MATEMATINIAI MODELIAI

29

nusako pilnąsias sąlygines tikimybes (*angl. full conditional distributions*)

$$p_v(a) = p_v(a(v) \mid a(V_{-v})) := \mathbf{P}\{Y(v) = a(v) \mid Y(V_{-v}) = a(V_{-v})\}. \quad (1.31)$$

Bet ar pakanka nusakyti vien tik pilnąsias sąlygines tikimybes, kad būtų korektiškai nusakytas visas (pilnasis) tikimybinis skirstinys P_Y ? Į šį klausimą atsako Brooko lema, kurią atskiru diskrečiųjų atsitiktinių laukų atveju galima suformuluoti taip:

1 lema. Tegu $V = [1, n] := \{1, 2, \dots, n\}$. Jeigu $a, r \in \mathcal{A}^V$ ir $\mathbf{P}\{Y = a\} > 0$ bei $\mathbf{P}\{Y = r\} > 0$, tai:

$$\frac{\mathbf{P}\{Y = a\}}{\mathbf{P}\{Y = r\}} = \prod_{i=1}^n \frac{\mathbf{P}\{Y(i) = a(i) \mid Y([1, i-1]) = a([1, i-1]), Y([i+1, n]) = r([i+1, n])\}}{\mathbf{P}\{Y(i) = r(i) \mid Y([1, i-1]) = a([1, i-1]), Y([i+1, n]) = r([i+1, n])\}} \quad (1.32)$$

Prielaida $V = [1, n]$ supaprastina teiginio formulavimą, bet nemažina jo bendrumo, nes bet kuri aibė V su n elementų yra izomorfiška aibei $[1, n]$.

Iš 1 lemos išplaukia, kad pilnieji sąlyginiai (lokalieji) skirstiniai pilnai nusako bendrą (globalųjį) skirstinį, jeigu tik išpildyta teigiamų tikimybių sąlyga:

$$\mathbf{P}\{Y = a\} > 0 \quad \forall a \in \mathcal{A}^V. \quad (1.33)$$

Gibso skirstinys ir lokalios sąveikos. Apie Gibso skirstinį jau buvo rašyta aptariant apibendrintus tiesinius modelius. Dabar susiesime jį su aplinkų sistema $\mathcal{U}$ (diskrečiąja topologija $\mathcal{B}$, nusakyta grafu $(V, \mathcal{B})$).

Gibso skirstinys (fizikoje) yra apibrėžiamas per energijos funkciją $E(a)$:

$$P_G(a) := Z^{-1} \exp\{-E(a)\}, \quad a \in \mathcal{A}^V, \quad (1.34)$$

čia Z yra normuojantis daugiklis, vadinamas *dalinimo funkcija*, o energijos funkcija yra išreiškiama per (*Gibso*) *potencialus* $\lambda_K(a)$, $K \in \mathcal{K}$:

$$E(a) = \sum_{K \in \mathcal{K}} \lambda_K(a). \quad (1.35)$$

Susiesime šį potencialų rinkinį su anksčiau (aptariant MAL) įvesta klikų viršūnių aibių sistema $\mathcal{K}$, o tuo pačiu ir su grafu $(V, \mathcal{U})$ bei aplinkų sistema $\mathcal{U}$.

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI
30 MATEMATINIAI MODELIAI

Laikysime, kad funkcijos $\lambda_K: \mathcal{A}^V \rightarrow \mathbf{R}$, $K \in \mathcal{K}$, priklauso tik nuo a koordinačių iš aibės K , t. y. λ_K priklauso tik nuo $a(K)$. Formaliai ši sąlyga užrašoma taip:

$$\lambda_K(a) = \lambda_K(z) \quad \forall a, z \in \mathcal{A}^V, a(K) = z(K). \quad (1.36)$$

6 apibrėžimas. Skirstinį, nusakytą lygybėmis (1.34) ir (1.35), kuriose $\mathcal{K}$ yra aibių, sudarytų iš grafo $(V, \mathcal{B})$ klikų viršūnių, sistema, ir potencialų funkcijos λ_K , $K \in \mathcal{K}$, tenkina sąlygą (1.36), vadinsime Gibso skirstiniu su lokalsios sąveikos aplinkomis $\mathcal{U}$ arba grafu $(V, \mathcal{B})$.

1 teorema. (Hammersley-Clifford (Besag 1971)). Tarkime, kad atsitiktinio lauko Y skirstinys P_Y tenkina teigiamų tikimybių sąlygą (1.33). Tuomet P_Y yra Markovo atsitiktinis laukas su aplinkų sistema $\mathcal{U}$ tada ir tik tada, kai jis yra Gibso skirstinys su lokalsios sąveikos aplinkomis $\mathcal{U}$ ir koku nors potencialo funkcijų $u_K(a)$, $K \in \mathcal{K}$, tenkinančių sąlygą (1.36), rinkiniu.

3 pavyzdys. Logtiesinis modelis.

Tegu aibė V yra sudaryta iš kokybinių (nominaliųjų) požymių identifikatorių, aibė $\mathcal{A}_v$ yra požymio $v \in V$ galimų reikšmių (būsenų) aibė. Tegu $Y = \{Y(v), v \in V\}$ yra stebėtos tų požymių reikšmės. Tarkime, kad Y skirstinys aprašomas logtiesiniu modeliu su požymių sąveikomis, nusakytomis grafu $(V, \mathcal{B})$. Vadinas, Y yra Markovo atsitiktinis laukas su lokalių sąveikų grafu $(V, \mathcal{B})$, ir remiantis 1 teorema ir (1.34) bei (1.35), Y skirstinys P_Y aibėje $\mathcal{A}^V$ tenkina sąryšį:

$$\log(P_Y(a)) = \lambda_0 - \sum_{K \in \mathcal{K}} \lambda_K(a), \quad a \in \mathcal{A}^V, \quad (1.37)$$

čia $\lambda_0 := -\log(Z)$ yra normuojanti konstanta, $\mathcal{K}$ yra aibių, sudarytų iš grafo $(V, \mathcal{B})$ klikų viršūnių (trumpai tariant, klikų) rinkinys. Kiekviena (potencialo) funkcija λ_K priklauso nuo $|K|$ kintamųjų ir todėl bendru atveju $n_K - 1$ skaliarinių parametrų nusako ją konstantos tikslumu. Čia $n_K := \prod_{v \in K} n_v$, $n_v := |\mathcal{A}_v|$. Vadinas, logtiesinis modelis (1.37) bendru atveju priklauso nuo ne daugiau kaip $n_V := \prod_{K \in \mathcal{K}} (n_K - 1)$ skaliarinių parametrų, bet jų gali būti gerokai mažiau dėl to, kad aibės $K \in \mathcal{K}$ gali kirstis, ir dėl papildomų apribojimų potencialo funkcijoms u_K , $K \in \mathcal{K}$.

Kadangi pagal Brooko leimą skirstinį P_Y nusako ir pilnieji sąlyginiai skirstiniai, tai jį nusako ir apibendrintų (daugialypių) logit modelių sistema:

$$\log \left(\frac{\mathbf{P}\{Y(v) = a(v) \mid Y(V_{-v}) = a(V_{-v})\}}{\mathbf{P}\{Y(v) = r \mid Y(V_{-v}) = a(V_{-v})\}} \right) =$$

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI
MATEMATINIAI MODELIAI

31

$$= \sum_{K \in \mathcal{K}: v \in K} (\lambda_K(a_v^r) - \lambda_K(a)), \quad a \in \mathcal{A}^V, r \in \mathcal{A}_v, v \in V, \quad (1.38)$$

čia trumpumo dėlei naudojamas toks pažymėjimas: duotiems $a \in \mathcal{A}^V$, $v \in V$ ir $r \in \mathcal{A}_v$, $a_v^r := \{z \in \mathcal{A}^V: z(v) = r, z(h) = a(h) \forall h \in V_{-v}\}$.

4 pavyzdys. m -tos eilės homogeninė Markovo grandinė (HMG).

Tegu $Y = \{Y(\ell), \ell \in [1, n]\}$ yra m -tos eilės homogeninė Markovo grandinė su būsenų aibe $\mathcal{A}$. Tada, kaip jau minėjome, $\mathcal{K} = \{[k, k+m], k \in [1, n-m]\}$. Kadangi šiuo atveju dėl homogeniškumo daugialypių logit modelių sistema nepriklauso nuo pozicijos $\ell \in [1, n]$, tai Y skirstinį vienareikšmiškai nusako vienintelis daugialypis logit modelis pozicijai $\ell = m+1$ ir aplinkai $U := [1, m] \cup [m+2, 2m+1]$:

$$\log \left(\frac{\mathbf{P}\{Y(m+1) = a \mid Y(U) = z\}}{\mathbf{P}\{Y(v) = r \mid Y(U) = z\}} \right) = \sum_{\ell=1}^{m+1} (\lambda_{[\ell, \ell+m]}(a_v^r) - \lambda_{[\ell, \ell+m]}(a)),$$

$$a, r \in \mathcal{A}, \quad z \in \mathcal{A}^{2m}. \quad (1.39)$$

Gibso ėmimas. Tarkime, kad aibėje $\mathcal{A}^V$ apibrėžtas tikimybinis skirstinys P , tenkinantis tikimybių teigiamumo sąlygą $P(a) > 0, \forall a \in \mathcal{A}^V$. Uždavinys yra sukonstruoti homogeninę Markovo grandinę (HMG) X , kurios stacionarusis skirstinys yra P . Remiantis tokia HMG X galima generuoti atsitiktines reikšmes $Y \in \mathcal{A}^V$, kurių skirstinys yra (apytiksliai) lygus P . Tai sudaro Markovo grandinių Monte Karlo (*angl. MCMC = Markov Chain Monte-Carlo*) metodo pagrindą ir pastaruoju metu labai plačiai taikoma Bajeso metode aposterorinių skirstinių atsitiktinių reikšmių generavimui.

Tegu duota homogeninė Markovo grandinė $\nu = \{\nu_t, t = 0, 1, 2, \dots\}$ su reikšmėmis aibėje V ir asimptotiškai teigiamomis visų būsenų vidutinėmis tikimybėmis (teigiamų tikimybių sąlygos analogas). Dažniausiai taikomi tokie du HMG ν variantai:

Pirmame variante $\nu = \{\nu_t, t = 0, 1, 2, \dots\}$ yra nepriklausomi ir vienodai pasiskirstę atsitiktiniai dydžiai su bendru skirstiniu Q , $Q(v) > 0 \forall v \in V$. Paprastai $Q(v) \equiv 1/n$.

Antrame variante ν yra cikliška seka $\nu_t = t \pmod{n} + 1, t = 0, 1, 2, \dots$.

Pateiktame žemiau HMG $X = \{X_t, t = 0, 1, 2, \dots\}$ generavimo algoritme yra realizuota *Glauberio dinamika*: kiekviename žingsnyje t yra keičiama HMG X_t reikšmė tik vienoje „viršūnėje“ $v \in V$ ($t = 0, 1, 2, \dots$).

1. KATEGORINIŲ POŽYMIŲ STATISTINĖJE ANALIZĖJE NAUDOJAMI
32 MATEMATINIAI MODELIAI

Apskaičiuokime skirstinio P pilnąsias sąlygines tikimybes (1.31):

$$p_v(a(v) | a(V_{-v})) := p_v(a) = \frac{P(a)}{\sum_{z \in \mathcal{A}_v} P(a_v^z)}, \quad a \in \mathcal{A}^V, v \in V, z \in \mathcal{A}_v. \quad (1.40)$$

Gibso ėmimo algoritmas. Tarkime, kad pradinė sekos X reikšmė $X_0 \in \mathcal{A}^V$ yra pasirinkta bet kaip, ir jau yra sugeneruotos atsitiktinės poros (ν_j, X_j) , $j = 0, \dots, t-1$, $t > 0$. Tada nauja atsitiktinė pora (ν_t, X_t) yra gaunama tokiu būdu:

(G1) Remiantis ankstesne reikšme ν_{t-1} , generuojame ν_t , HMG ν reikšmę momentu t .

(G2) Viršūnėje (pozicijoje) $v := \nu_t$ atsitiktinai pagal formulę (1.40) apibrėžtą sąlyginę tikimybę $p_v(a | X_{t-1}(V_{-v}))$, $a \in \mathcal{A}_v$, generuojama nauja reikšmė $X_t(v)$, o likusiose X_t reikšmės viršūnėse (pozicijose) $\ell \in V_{-v}$ lieka tokios pat kaip ir X_{t-1} , t. y.: $X_t(\ell) = X_{t-1}(\ell)$ $\ell \in V, \ell \neq v$.

Iš aprašyto algoritmo išplaukia, kad homogeninės Markovo grandinės X iš z į a perėjimo tikimybė:

$$\Pi(a | z) = p(a(v) | a(V_{-v})) \mathbf{1}\{a(V_{-v}) = z(V_{-v})\} \quad a, z \in \mathcal{A}^V, v \in V. \quad (1.41)$$

Taigi, HMG X yra necikliška, neredukuojama ir jos perėjimo tikimybės (1.41) tenkina lokalaus balanso lygtį:

$$P(z) \Pi(a | z) = P(a) \Pi(z | a) \quad \forall a, z \in \mathcal{A}^V. \quad (1.42)$$

Vadinasi, Markovo grandinė X yra apverčiama, ir jos stacionarusis skirstinys yra P .

2

Logit analizė genetiniams duomenims

Dirbant su genetiniais duomenimis, iškyla įvairaus pobūdžio uždavinių (Rekašius 2007; Usotskaya, Ryabko 2008). Disertacijoje nagrinėjami ir šiame skyriuje pateikiami: Markovo eilės įvertinimas DNR grandinėse, DNR sekų homogeniškumas bei pirminių ir antrinių DNR sekos vijų palyginimas – DNR sekų simetriškumo savybė. Visus šiuos uždavinius galima laikyti priklausomybių struktūros tarp DNR sekos nukleotidų vertinimo uždaviniais, kai turimi didelės dimensijos kategorinių požymių duomenų masyvai ir tam tikrais atvejais išretintos dažnių lentelės.

2.1 ir 2.2 skyreliuose trumpai aprašoma, kas yra DNR seka ir jos evoliucija. Toliau kiekvienam iš minėtų uždavinių apžvelgiami literatūroje jau žinomi rezultatai (2.3) bei išdėstomi autorės straipsnyje (2008) publikuoti rezultatai, kuriuose pasiūlyta speciali duomenų struktūra bei logit modelio taikymas Markovo eilės DNR sekose įvertinimui (skyr. 2.4, 2.6 ir 2.5). Taip pat pateikiama procedūra, ieškanti nekoduojančių DNR grandinės posekių, kurie yra homogeniški, bet hipotezė apie pirmos eilės Markoviškumą jiems yra atmetama. 2.6 skyrelyje pasiūlytas DNR grandinės pirminės ir antrinės vijų simetriškumo hipotezės teorinis formulavimas naudojant logit modelius. 2.5 skyrelyje pateikiami gauti Markovo eilės, sekų homogeniškumo bei simetriškumo tikrinimo rezultatai kai kurioms realioms bakterijų DNR sekoms.

2.1. Nukleotidų sekos

DNR molekulė yra ilga, dviguba, spiralės formos nukleotidų seka. Ją sudaro keturių tipų nukleotidai: du pirimidinai – timinas (T) ir citozinas (C), ir du purinai – guaninas (G) ir adeninas (A). Jų aibę žymėsime $\mathcal{A} = \{A, C, G, T\}$.

Nukleotidų sekos fragmentas:

$$\begin{aligned} & \dots GCAATACGCCTA \dots \\ & \dots CGTTATGCGGAT \dots \end{aligned} \quad (2.1)$$

Taigi, DNR nukleotidų grandinė $X =: \{x_1, \dots, x_n\}$ gali būti laikoma kokybinių kintamųjų, įgyjančių reikšmes iš baigtinės raidžių aibės $\mathcal{A} = \{A, T, C, G\}$, seka.

Taisyklingai DNR grandinei susisukti į dvigubą spiralę reikalinga, jog purinas vienoje grandinės pusėje būtų pakeičiamas pirimidinu kitoje. Todėl nukleotidas A stovi priešais T , o nukleotidas C – priešais G . Be to, A ir T jungiasi per dvi, o C ir G – per tris vandenilines jungtis. Tokios bazių poros ir iš jų sudarytos DNR vijos vadinamos komplementariomis. Pagal susitarimą, dvigubos DNR spiralės viršutinė (pirminė) vija skaitoma iš kairės į dešinę, o apatinė (antrinė) – iš dešinės į kairę.

Pirminės ir antrinės vijos yra suskirstytos į koduojančias sritis (genus) ir nekoduojančias (tarpus). Genas yra baltymą koduojanti nukleotidų seka, o DNR sekos tarp genų vadinamos nekoduojančiomis genomo sekomis. Genai abiejose DNR grandinės vijose gali būti išsidėstę pavieniai ar grupėse, taip pat gali persikloti vienas su kitu. Šiame darbe yra naudojamos nekoduojančios sekos, manant, kad jos yra labiau homogeniškos.

Kadangi kiekvieną nukleotidą apibūdina du požymiai: jungčių skaičius (dvi arba trys) ir molekulės tipas (purinas ar pirimidinas), todėl yra patogiu DNR seką pagal minėtas kiekvieno nukleotido dvi savybes perkoduoti į dvi-matę binarinę seką (2.1 lentelė).

2.1 lentelė. DNR nukleotidų savybių perkodavimo taisyklė

Table 2.1. The rule for code two properties of DNA

	Purinas (Jungtis)		Pirimidinas	$\beta(v)$
Dvi jungtys	A	($=$)	T	-1
Trys jungtys	G	($\equiv$)	C	$+1$
$\pi(v)$	-1		$+1$	

2.2. Nukleotidų sekų evoliucija

Visa paveldima organizmo genetinė informacija, esanti DNR sekose, nėra pastovi – vyksta atsitiktinės tų sekų mutacijos, vienus nukleotidus keičia kiti, atskiri sekų fragmentai prapuola, įsiterpia kiti. Kalbant apie DNR sekų evoliuciją, išskiriamos dvi jos kryptys:

DNR sekos evoliucija laike: $X(t) \rightarrow X(t+1)$.

DNR sekos evoliucija erdvėje: $x_l \rightarrow x_{l+1}$.

Ilgą laiką buvo daroma prielaida, kad nukleotidai DNR sekoje mutuoja nepriklausomai vienas nuo kito. Modeliai, kuriuose x_l pasikeitimo tikimybė priklauso nuo dviejų kaimyninių nukleotidų (ar kodonų) x_{l-1} ir x_{l+1} reikšmių, vadinami kontekstiniais mutavimo modeliais. Pastaraisiais metais pasiūlytuose kontekstiniuose evoliucijos modeliuose (Jensen 2005) kiekvieno nukleotido mutacija priklauso nuo m artimiausių jo kaimynų (iš abiejų pusių) ir nepriklauso nuo vietos DNR grandinėje. Paprastai $m = 1, 2, 3$.

Yra žinoma (Jensen 2005), kad laike apverčiamai evoliucijai, stacionaraus nukleotidų skirstinio DNR sekose Markovo eilė įtakojama lokalių kaimyninių nukleotidų sąveikų. Jensen (2005) parodė, kad jeigu DNR sekų atsitiktinės mutacijos priklauso tik nuo artimiausių kaimyninių nukleotidų (tai yra visiškai natūrali prielaida nekoduojančioms sekoms), tai stacionarusis nukleotidų skirstinys jose sudaro pirmos eilės Markovo grandinę. Taigi, iš aukštesnės eilės nukleotidų priklausomybės nekoduojančiose sekose galime manyti, kad evoliucija gali būti laike neapverčiama. Todėl iškyla aktualus Markovo eilės DNR sekose įvertinimo uždavinys.

DNR sekų analizėje nemažas dėmesys skiriamas taip vadinamam ilgos priklausomybės fenomenui (Yu *et al* 2001). Šis terminas atėjęs iš stacionarių laiko eilučių teorijos. Yra žinoma, kad tai gali būti priežastimi atsirandančio nehomogeniškumo DNR sekose. Taigi, kitas disertacijoje keliamas klausimas – ar Markovo savybė DNR sekose yra susijusi su sekų nehomogeniškumu.

2.3. Literatūros apžvalga apie Markovo modelio eilės įvertinimą

Žodžiu ar motyvu yra vadinama seka, sudaryta iš aibės $\mathcal{A} = \{A, T, C, G\}$ elementų. DNR sekoje žodžių tikimybės ir statistinės savybės pakankamai nemažai nagrinėtos daugelyje darbų (apžvalgą žiūr. Reinert *et al* 2000). Minėtuose darbuose akcentuojami du pagrindiniai žodžių pasirodymo DNR sekose

aspektai: kur jie pasirodo ir kaip dažnai jie pasirodo? Tačiau žodžių statistinės savybės turi prasmę tik tuo atveju, kai remiamasi tam tikru tikimybinio modeliu, kuris padeda prognozuoti, tikrinti įvairias hipotezes ar atrasti tam tikras taisykles ir dėsningumus sekose. Taigi, turint konkrečią sekos realizaciją, genotipinių sekų modeliavimas susideda iš: tikimybinio modelio parinkimo ir parinkto modelio parametrų įvertinimo.

Tipiniai modeliai DNR sekoms yra homogeninės m -eilės Markovo grandinės (Markovo laukai), įgyjančios būsenas iš baigtinės aibės $\mathcal{A}$. Šie modeliai įvertina lokalią nukleotidų priklausomybę DNR sekose kai tam tikros būsenos (raidės) pasirodymas konkrečioje sekos pozicijoje priklauso nuo m prieš tai buvusių nukleotidų (raidžių) ir nepriklauso nuo pasirinktos vietos sekoje. Taigi, Markovo modelio įvertinimas reiškia Markovo eilės m parinkimą ir perėjimo tikimybių įvertinimą:

$$\Pi := \{\pi(a, a_{m+1}), a = (a_1, \dots, a_m) \in \mathcal{A}^m, a_{m+1} \in \mathcal{A}\}.$$

Paprastai m eilės Markovo modelis turi $(c - 1)c^m$ parametrų kai $c = |\mathcal{A}|$ yra elementų skaičius aibėje $\mathcal{A}$. Naudojant šiuos modelius statistinėje analizėje, iškyla didžiausia problema parametrų (perėjimo tikimybių) vertinime, nes augant Markovo eilei, parametrų skaičius auga eksponentiškai. Tai dažniausiai atbaido tyrėjus nuo šių modelių naudojimo.

Markovo eilės įvertinimui literatūroje yra siūloma įvairių metodų. Raftery (1985) pasiūlė aukštesnės eilės Markovo modelio įvertinimui modelį analogišką standartiniam AR(m) modeliui, kuriame aukštesnė vienetu Markovo eilė įvertinama pridėdant vieną papildomą parametą. Taip pat šiuo atveju parodoma, kad autokoreliacijos atitinka tiesinių lygčių sistemą, panašią į *Yule-Walker* lygtis. Šis Raftery modelis vėliau buvo išplėstas ir patobulintas (Ching *et al* 2004; Ching *et al* 2005).

Sudėtingesnė taisyklė, sumažinanti modelio parametrų skaičių, yra pasiūlyta Weinberger *et al* (1995) ir Buhlmann, Wyner (1999). Čia pasiūlyto *Variable length Markov chain* modelio idėja yra ta, kad sekos atminties ilgis (*angl. depth*) bet kuriuo laiko momentu gali priklausyti nuo prieš tai buvusios istorijos, vaizduojamos *kontekstiniais medžiais*. Šiuo atveju modelio eilės parinkimas yra nagrinėjamas kaip kontekstinio medžio (ar priklausomybių struktūros) parinkimo problema.

Reinert *et al* (2000) publikacijoje aprašomas įprastas χ^2 kriterijus, kuris DNR seką vaizduoja kaip dvimatę dažnių lentelę ir tikrina, kuri Markovo modelio eilė yra tinkama. Eggar (2002) pasiūlė grafų teorija paremtą procedūrą, pagal kurią galima patikrinti, ar turimiems duomenims yra tinkamas pirmos

eilės Markovo modelis.

Markovo eilės vertinimo problema ar bendru atveju kontekstinių medžių įvertinimas nagrinėjams daugelyje publikacijų: Csizsár, Paul (2000); Zhao *et al* (2001); Morvai, Weiss (2005); Peres, Shields (2005); Csizsár, Talata (2006); Csizsár, Talata (2006); Streliaff (2007) ir jų nuorodose.

Markovo eilės DNR sekose įvertinimo uždavinį panašiai kaip disertacijoje siūlomą variantą, nagrinėjo taip pat Avery ir Henderson (1999, 2002). Jie naudojo konkretaus geno (*preproglucagon*) nekoduojančią dalį (*introną*) iš žmogaus DNR. Laikydami, kad DNR seka sudaro baigtinės eilės Markovo grandinę, jie taikė logtiesinius modelius nukleotidų porų ir jų tripletų dažnių bendram tikimybiniam skirstiniui įvertinti bei Markovo grandinės eilei nustatyti. Pirmos eilės Markovo modelis nepakankamai gerai aprašė turimus duomenis, o antros eilės nebuvo atmestas.

Jų sudaryta imtis dėl stebėjimų tarpusavio priklausomybės netenkina standartinių logtiesinio modelio prielaidų, ir todėl jie modeliavimo būdu ieškojo sprendimų tinkamai χ^2 statistikos aproksimacijai.

Paprastai yra daroma prielaida (ir tai taikoma visuose anksčiau minėtuose darbuose), kad Markovo grandinė yra homogeninė. Tačiau atrodo, jog bendru atveju ši prielaida DNR sekoms negalioja.

2.4. Markovo savybė ir apibendrintas logit modelis

Šio skyrelio pradžioje apibrėžiamos pagrindinės diskrečių baigtinių būsenų Markovo laukų teorijos sąvokos ir jų taikymas DNR sekų atveju. Taip pat pasiūlytas Markovo grandinės aprašymas, pasinaudojant sąlyginių šansų bei pristatoma speciali genetinių duomenų struktūra, kuri yra patogi logit analizei, šiuo atveju skirtai Markovo eilės DNR grandinėje įvertinimui ir tikimybės lygybės tarp dviejų komplementarių DNR grandinių testavimui.

2.4.1. Atsitiktinis Markovo laukas ir sąlyginiai šansai DNR sekų atveju

DNR seką galime apibrėžti kaip atsitiktinį Markovo lauką (1.28), kuriame kiekvienas nukleotidas nepriklauso nuo visų kitų sekos nukleotidų kai žinomi jo dešinės ir kairės pusės kaimynai, ir taikyti MRF teoriją (*angl. Markov random fields theory*).

Tarkime $\mathcal{A}$ yra baigtinė aibė su $|\mathcal{A}| = \text{card}\mathcal{A}$ elementų ir $N := \{1, \dots, n\}$. Fiksuokime tam tikrą teigiamą sveiką skaičių $m < n/2$ ir apibrėžkime N „vidų“ N° ir „kraštus“ $\partial: N^\circ := (m+1, \dots, n-m)$, $\partial N := N \setminus N^\circ$. Indeksų rinkinys, nusakantis kaimyninius nukleotidus:

$$U(\ell) = U_m(\ell) := [\ell - m, \ell + m] \setminus \{\ell\}, \quad \ell \in N^\circ,$$

čia $[i, j] := (i, i+1, \dots, j)$, $i < j$, $i, j \in N$ yra sveikų skaičių intervalas. Turint $x \in \mathcal{A}^n$ ir indeksų aibę $I \subset N$, $x_I := (x_i, i \in I)$ apibrėš atitinkamą posekį x .

1 apibrėžimas. Atsitiktinė seka $X \in \mathcal{A}^n$ yra m eilės homogeniškas Markovo laukas (grandinė) (toliau – Mm) jei $\forall \ell \in N^\circ$ ir $a \in \mathcal{A}^n$:

$$\mathbf{P}\{x_\ell = a_\ell | x_j = a_j, j \neq \ell\} = \mathbf{P}\{x_\ell = a_\ell | x_{U_m(\ell)} = a_{U_m(\ell)}\} =: p(a_\ell | a_{U_m(\ell)}). \quad (2.2)$$

2 apibrėžimas. Sąlyginis šansas $\Lambda_{v|b}(u)$ nusako, kiek kartų skiriasi ℓ -tojo nukleotido tikimybė įgyti reikšmę v nuo tikimybės įgyti bazinę reikšmę (*reference value*) $b \in \mathcal{A}$ su $p(b|u) > 0$, kai žinomos m jo artimiausių kaimynų $U_m(\ell)$ reikšmės $u = x_{U_m(\ell)}$, $\ell \in N^\circ$:

$$\Lambda_{v|b} = \Lambda_{v|b}(u) := \frac{p(v|u)}{p(b|u)}, \quad v \in \mathcal{A}, u \in \mathcal{A}^{2m}, \quad (2.3)$$

čia tikimybės $p(v|u)$ yra apibrėžtos (2.2).

Paimkime atskirą atvejį, t.y. pirmos eilės Markovo grandinę ($m = 1$):

$$\rightarrow x_{n-2} \rightarrow x_{n-1} \rightarrow x_n \rightarrow x_{n+1} \rightarrow x_{n+2}$$

Šiuo atveju klikos turės pavidalą $c = [n-1, n]$. Tada sekos skirstinys:

$$\mathbf{P}(x) = p(x_0) \prod_{n=1}^N p(x_n | x_{n-1}) = p(x_0) \exp\left\{\sum_{n=1}^N \log p(x_n | x_{n-1})\right\}. \quad (2.4)$$

Vadinasi, pirmos eilės Markovo grandinės atveju Gibso skirstinyje klikų potencialų funkcija turės tokį pavidalą:

$$\lambda(x_n, x_{n-1}) = \log(p(x_n | x_{n-1})) := \log(p_{(n-1)n}). \quad (2.5)$$

Užrašysime šanso 2.3 išraišką homogeniniam pirmos eilės Markovo laukui,

pasinaudodami Gibso skirstiniu ir potencialų funkcija (2.5):

$$\begin{aligned} N^\circ &:= (2, 4, \dots, n-1), \quad U(\ell) = [\ell-1, \ell+1] \setminus \{\ell\}, \quad \ell \in N^\circ, \\ \mathbf{P}\{x_\ell = a_\ell | x_{U(\ell)} = a_{U(\ell)}\} &= \frac{\mathbf{P}\{x_\ell = a_\ell, x_{U(\ell)} = a_{U(\ell)}\}}{\mathbf{P}\{x_{U(\ell)} = a_{U(\ell)}\}} \\ &= \frac{\mathbf{P}\{x_\ell = a_\ell, x_{\ell-1} = a_{\ell-1}, x_{\ell+1} = a_{\ell+1}\}}{\mathbf{P}\{x_{\ell-1} = a_{\ell-1}, x_{\ell+1} = a_{\ell+1}\}} \\ &:= \frac{p_{\ell-1}p_{(\ell-1)\ell}p_{\ell(\ell+1)}}{p_{\ell-1}p_{(\ell-1)0}p_{0\ell} + p_{\ell-1}p_{(\ell-1)1}p_{1\ell}} = \frac{p_{(\ell-1)\ell}p_{\ell(\ell+1)}}{p_{(\ell-1)0}p_{0\ell} + p_{(\ell-1)1}p_{1\ell}}. \end{aligned}$$

Tada šansas:

$$\begin{aligned} \log \left(\Lambda_{a_l|b}(a_{U(l)}) \right) &= \log \left(\frac{\mathbf{P}\{x_\ell = a_\ell | x_{U(\ell)} = a_{U(\ell)}\}}{\mathbf{P}\{x_\ell = b | x_{U(\ell)} = a_{U(\ell)}\}} \right) = \\ &= \log(p_{(\ell-1)\ell}) + \log(p_{\ell(\ell+1)}) - \log(p_{(\ell-1)b}) - \log(p_{b(\ell+1)}) \\ &= \lambda(x_{\ell-1}, x_\ell) + \lambda(x_\ell, x_{\ell+1}) - \lambda(x_{\ell-1}, x_\ell^b) - \lambda(x_\ell^b, x_{\ell+1}). \end{aligned} \quad (2.6)$$

Kiekvienam sekos tripletui $a = (a_1, a_2, a_3) \in \mathcal{A}^3$:

$$\log \left(\Lambda_{a_2|b}(a_{U(2)}) \right) = \sum_{j=1}^2 \left[\lambda_1 \left(a_{[j,1+j]} \right) - \lambda_1 \left(a_{[j,1+j]}^{(b)} \right) \right], \quad a^{(b)} = (a_1, b, a_3). \quad (2.7)$$

Tarkime, kad Markovo grandinės X kraštai ∂N yra fiksuoti $X_{\partial N} = c_{\partial N}$ tam tikram $c \in \mathcal{A}^n$, ir aibė $\mathcal{X}_+ := \{a \in \mathcal{A}^n: a_{\partial N} = c_{\partial N}\}$. Tada iš Hamersley-Clifford teoremos (Besag 1974) bet kokios eilės Markovo laukui seka tvirtinimas:

1 tvirtinimas. Jei $\mathbf{P}\{x = a\} > 0$ visiems $a \in \mathcal{X}_+$, sąlyginis homogeninio Markovo atsitiktinio lauko skirstinys yra vienareikšmiškai nusakomas sąlyginio šansu: $\Lambda_{v|b}(u)$, $v \in \mathcal{A}$, $u \in \mathcal{A}^{2m}$ su tam tikra bazine reikšme $b \in \mathcal{A}$, ir egzistuoja tokia funkcija $\lambda_m: \mathcal{A}^{m+1} \rightarrow \mathbf{R}$, kad kiekvienam $a = (a_1, \dots, a_{2m+1}) \in \mathcal{A}^{2m+1}$:

$$\log \left(\Lambda_{a_{m+1}|b}(a_{U_m(m+1)}) \right) = \sum_{j=1}^{m+1} \left[\lambda_m \left(a_{[j,m+j]} \right) - \lambda_m \left(a_{[j,m+j]}^{(b)} \right) \right], \quad (2.8)$$

čia $a^{(b)} = (a_1, \dots, a_m, b, a_{m+2}, \dots, a_{2m+1})$.

Tai reiškia, kad norint identifikuoti Markovo modelį, reikia apibrėžti jo

sąlyginį šansą (2.3).

Apibrėžkime stebimų sekų $x \in \mathcal{A}^n$ struktūrą, kai jų ilgis $n = n_m \cdot (m + 1) + m$, dydis n_m yra sveikas skaičius. Pažymėkime:

$$X := \{(v_\ell, u_\ell), \ell \in S\}, \quad S = S_{n,m} = \{m+1, 2(m+1), \dots, n-m\}, \quad (2.9)$$

čia $v_\ell := x_{(m+1)\ell}$ yra atsako kintamasis, o $u_\ell = x_{U_m(\ell)}$ yra vektorius aiškinamųjų kintamųjų, $\ell \in S$.

Prielaidos:

- (A1) $\{v_\ell, \ell \in S\}$ yra sąlyginai nepriklausomi, kai žinomi $\{u_j, j \in S\}$;
 (A2) Sąlyginės tikimybės $\mathbf{P}\{v_\ell = a | u_j, j \in S\}$, $a \in \mathcal{A}$ nepriklauso nuo pozicijos sekoje. $\ell \in S$.

1 pastaba. Pastebėkime, kad šios prielaidos yra išpildytos, jei x yra generuota pagal homogeninę m eilės Markovo grandinę. Markovo savybė šansui (2.3) numato papildomas sąlygas. Iš lygybės (2.8) matome, kad šansas priklauso nuo $v = x_\ell$ ir $u = x_{U_m(\ell)}$ tik per sąveikas $x_{I_j}, j = 0, \dots, m$, čia $I_j = U_m(\ell) \cap [\ell - m + j, \ell + j]$. Tai suteikia pagrindą markoviškumo tikrinimui ir Markovo eilės įvertinimui. (žiūrėti sekantį skyrelį).

2 pastaba. Prielaidos (A1) ir (A2) užtikrina, kad bendros apibendrinto logit modelio sąlygos yra išpildytos (Agresti 1990). Apibendrintas logit modelis yra regresinio tipo modelis, t. y. sąlyginis modelis su duotosiomis aiškinamųjų kintamųjų u reikšmėmis. Tai reiškia, kad aiškinamųjų kintamųjų $\{u_\ell, \ell \in S\}$ tikimybinius modelis yra neapibrėžtas ir gali būti traktuojamas kaip nepatogi neparametrinė modelio komponentė. Tai taip pat reiškia, kad 1 pastaboje minimos sąlygos Markovo savybei yra būtinos, bet nepakankamos.

2.4.2. Apibendrintas logit modelis

Iš tikrųjų apibendrintas logit modelis – tai logtiesinis modelis sąlyginiam šansui (2.3):

$$\log(\Lambda_{y|b}(v, w)) = \lambda(v, y, w), \quad y \in \mathcal{A}, \quad v, w \in \mathcal{A}^m. \quad (2.10)$$

Funkcija $\lambda: \mathcal{A}^{2m+1} \rightarrow \mathbf{R}$ atitinka $\lambda(\cdot, b, \cdot) = 0$ ir bendru atveju (tiesiškai) priklauso nuo $K = (|\mathcal{A}| - 1)|\mathcal{A}|^{2m}$ parametrų. Modelis (2.10) yra laikomas *saturated* apibendrintu logit modeliu.

Atsižvelgiant į (2.3), (2.10) ir prielaidas (A1), (A2), $\{y_\ell, \ell \in S\}$ sąlyginis skirstinys, kai žinomos $\{z_\ell, \ell \in S\}$ reikšmės, yra sandauga multinominių skirstinių. Žodžių $a \in \mathcal{A}^{2m+1}$ pasirodymo sekoje dažniai $N(a) := |\{j \in$

$S: x_{[j-m, j+m]} = a\}$, $a \in \mathcal{A}^{2m+1}$ sudaro pakankamas statistikas pradiniam modeliui. Mm modelio atveju funkcija λ duota (2.8) ir įgyja tokią formą:

$$\lambda(a) := \lambda\left(a_{[1,m]}, a_{m+1}, a_{[m+1, 2m+1]}\right) = \lambda_{\text{Mm}}(a). \quad (2.11)$$

$$\lambda_{\text{Mm}}(a) := \sum_{j=1}^{m+1} \left[\lambda_m\left(a_{[j, m+j]}\right) - \lambda_m\left(a_{[j, m+j]}^{(b)}\right) \right] \quad a \in \mathcal{A}^{2m+1}. \quad (2.12)$$

ir priklauso nuo $k = (|\mathcal{A}| - 1)|\mathcal{A}|^m$ parametrų.

Nagrinėkime atvejį, kai $m = 1$. Apibendrintas logit modelis (2.10) bendru atveju (saturated model) ir Mm eilės Markovo modelio atveju (2.11), (2.12) įgyja žemiau pateiktas paprastas formas:

$$\log(\Lambda_{y|b}(v, w)) = \lambda_0(y) + \lambda_L(v, y) + \lambda_R(y, w) + \lambda_{LR}(v, y, w) \quad (2.13)$$

$$\log(\Lambda_{y|b}(v, w)) = \lambda_0(y) + \lambda_1(v, y) + \lambda_1(y, w), \quad v, y, w \in \mathcal{A}, \quad (2.14)$$

čia šios funkcijos λ praranda vieną iš savo argumentų, įgyjančių bazinės reikšmės b .

Taigi, pirmos eilės homogeninei Markovo grandinei M1 nulinė hipotezė:

$$H_0 : \lambda_{LR}(v, y, w) = 0, \quad \lambda_L(v, w) = \lambda_R(v, w), \quad \forall v, y, w \in \mathcal{A}. \quad (2.15)$$

2.5. Statistinė analizė

Šiame skyrelyje yra pateikiama logit statistinė analizė kai kurioms realioms DNR sekoms, t. y. tikrinama hipotezė apie Markovo eilę, sekų homogeniškumą bei pirminių ir antrinių vijų simetriškumą.

Trijų bakterijų genomų *Escherichia coli*, *Bordetella bronchiseptica* ir *Coxiella burnetti* duomenys yra gauti iš duomenų bazės *GenBank*. Statistinei analizei altikti taip pat panaudojama SAS programinė įranga bei jos procedūros CATMOD, LOGISTIC ir NLMIXED.

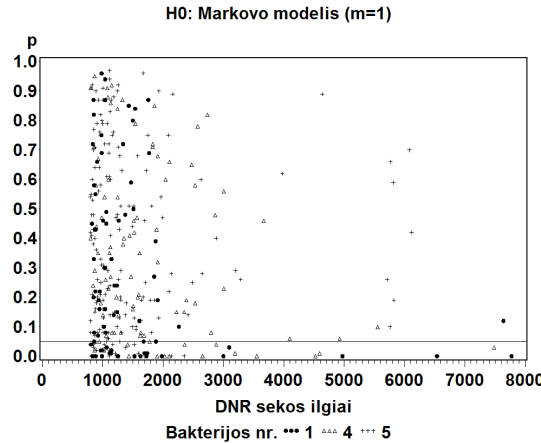
2.5.1. Markovo savybė ir jos ryšys su sekų nehomogeniškumu

Manoma, kad DNR sekos gali būti nehomogeniškos, dėl to darbe yra naudojamos tik jų nekoduojančios sritys. Laikoma, kad primityviems organizmams šios sritys neturi įtakos paveldimumui ar išgyvenimui. Be to, homo-

geniškumo reikalavimas Markovo grandinėms yra susilpninamas, pakeičiant logtiesinius modelius, naudotus Avery ir Henderson (1999, 2002), lankstesniu apibendrintu logit modeliu ir Markoviškumo hipotezėje (2.15) paliekant tik pirmąją lygybę:

$$H_0 : \lambda_{LR}(v, y, w) = 0 \quad \forall v, y, w \in \mathcal{A}. \quad (2.16)$$

Sekoms, sudarytoms iš pasirinktos bakterijos visų nekoduojančių sričių, nulinė hipotezė (2.16) yra atmetama ($p < 0,0001$). Tikrinant šią hipotezę (2.16) kiekvienam bakterijos nekoduojančiam fragmentui (ilgesniam nei 800) atskirai, kai kuriems iš jų minėta savybė taip pat atmetama (2.1 pav.) Be to, gauti rezultatai rodo, kad sekos yra gerokai nehomogeniškos. Hipotezė apie Markovo eilę $m =$



2.1 pav. Hipotezė apie 1-os eilės Markovo savybę įvairių ilgių DNR

Fig. 2.1. Hypothesis about first order Markov property in different length of DNA

2 ir $m = 3$ visų nekoduojančių fragmentų sekoms taip pat yra atmetama su $p < 0,0001$. Pastebėsime, kad Markovo modelis su eile $m = 3$ turi 624 parametrus.

Žemiau pateiktas paprastas teorinis pavyzdys rodo, kad sudėjus dvi skirtingas homogenines pirmos eilės Markovo grandines, gaunama nehomogeninė seka atrodys kaip aukštesnės (negu 1-os) eilės Markovo seka, jeigu laikyti ją homogenine.

1 pavyzdys. Imkime dvi to paties ilgio n homogenines (pirmos eilės) Markovo grandines $x^{(1)}$ ir $x^{(2)}$ su reikšmėmis iš aibės $\mathcal{A}_1 := \{0, 1\}$ ir perėjimo tikimybėmis $\mathbf{P}\{x_{j+1}^{(i)} = w | x_j^{(i)} = v\} =: p_{vw}^{(i)}$, apibrėžiamomis atitinkamai: $p_{01}^{(1)} = p_{10}^{(1)} = 0,8$ ir $p_{01}^{(2)} = p_{10}^{(2)} = 0,2$. Atkreipkime dėmesį, kad stacionarus

skirstinys π_v yra tas pats abejoms sekoms: $\pi_0 = \pi_1 = 1/2$.

Tarkime, kad Markovo seka $x^{(1)}$ prasideda nuo stacionaraus skirstinio π , o seka $x^{(2)}$ prasideda nuo būsenos $x_n^{(1)}$ ir tada nauja ilgio $N = 2n - 1$ seka $x = (x_1^{(1)}, \dots, x_n^{(1)}, x_2^{(2)}, \dots, x_n^{(2)})$ yra nehomogeninė pirmos eilės Markovo grandinė. Tikėtini dažniai visiems sekos x skirtingiems dvejetų rinkiniams

$$T_2(v, w) := |\{j = 1, \dots, 2n - 2 : x_j = v, x_{j+1} = w\}|, \quad v, w \in \mathcal{A}_1$$

yra lygūs $ET_2(v, w) \equiv (n - 1)/2$ ir todėl x laikant homogeniška Markovo seka jos visi elementai yra statistiškai nepriklausomi. Tuo remiantis gauname, kad visų trejetų tikimybės yra lygios $1/8$. Tačiau pradinei sekai, laikant kad n yra pakankamai didelis, gauname, pavyzdžiui, kad $p_{111} \approx 0,17$, $p_{110} \approx 0,08$

Iškyla klausimas, ar iš tikrųjų Markovo eilė DNR sekose yra didesnė už, pavyzdžiui, 3 ar tiesiog net ir tos pačios bakterijos sekos yra ženkliai nehomogeniškos.

Standartinis būdas nehomogeniškumo problemai spręsti yra mišriųjų (*angl. mixed*) modelių, t. y. modelių su atsitiktiniais koeficientais taikymas. Buvo bandyta sudaryti mišrųjų logistinių modelių y -kui su nekoduojančių posekių identifikatoriumi (eilės numeriu), kuris buvo laikomas atsitiktiniu efektu modelyje. Skaičiavimams buvo naudojama SAS procedūra NLMIXED. Deja, turiningų rezultatų gauti nepavyko. To priežastimi gali būti didelis parametų skaičius, didelis problemos mastas ir duomenų nehomogeniškumas, kuris neatitinka mišraus modelio prielaidų. Todėl nehomogeniškumo uždaviniui spręsti yra taikomas kitame skyrelyje pasiūlytas metodas.

2.5.2. Homogeniškumo testas DNR sekoms

Šiame skyrelyje pateiktos statistinės analizės tikslas – išsiaiškinti, ar Markovo savybės atmetimas yra įtakotas sekų nehomogeniškumo. Pasiūlyta homogeniškumo procedūra identifikuoja, kuriems posekiams nehomogeniškumas yra mažiau išreikštas nei ne-Markoviškumo savybė, t. y. homogeniškumo hipotezė yra neatmesta, tuo tarpu Markovo savybė (M1) atmetama. Procedūra susideda iš keletos žingsnių:

1. Kiekvienai iš trijų bakterijų išrenkamos kelios ilgiausių nekoduojančių sekų, kurioms pirmos eilės Markovo savybė (M1) atmetama. Sekos yra padalinamos į maždaug vienodo ilgio posekius (posekių ilgis yra nuo 712 iki 848).
2. Posekiai, kuriems M1 modelis atmetamas, yra naudojami tolimesnei analizei.
3. Gautiems posekiams taikomas homogeniškumo testas, kuris remiasi są-

lygomis (A1) ir (A2), apibendrintu logit modeliu (2.13) bei saviranka. Kiekvienai sekai suskaičiuojamos tam tikros LR statistikos p reikšmė ir ji lyginama su tos pačios statistikos p reikšmių skirstiniu, gautu generuojant homogeninę Markovo grandinę su iš minėtų posekių įvertintomis tikimybėmis bei pakartojant tą patį savirankos algoritmą 100 kartų. Detalizuosime minėtą savirankos algoritmą:

a) Pasirinktas sekos posekis suskaidomas į 10 daugmaž vienodo ilgio fragmentų, sudarytų iš eilės einančių nukleotidų.

b) Sudaromos visos galimos skirtingos 5 fragmentų iš 10 kombinacijos ir naujai įvestas požymis *grupė* įvardija išrinktą penketukų kombinaciją, priskirdamas jiems reikšmę 1. Likę penketukai patenka į grupę, pažymėtą 2-tu. Tokių kombinacijų skaičius bus lygus skaičiui derinių $C_{10}^5 = \frac{10!}{5!5!} = 252$.

c) Kiekvienai penketukų kombinacijai ir ją atitinkančiam požymiui *grupė* yra vertinamas pilnas (*saturated*) logit modelis su požymiu *grupė*, t. y. tikrinama nulinė hipotezė, kad požymis *grupė* ir visos kitų modelyje esančių kintamųjų sąveikos su šiuo požymiu, yra statistiškai nereikšmingos. Išsaugojamos atitinkamo testo LR statistikos ir jos p reikšmės. Iš 252 gautų p reikšmių išsirenkama mažiausia.

Homogeninės Markovo sekos Y^* generavimas:

1. Iš posekio Y suskaičiuojami tripletų santykiniai dažniai:

$$p(k|c,d) = \frac{f(k,c,d)}{f(k,+,+)}, \quad p(k,c,d) = \frac{f(k,c,d)}{f(+,+,+)},$$

čia $f(k,c,d) = f(kcd)$ yra tripleto (kcd) stebėtas dažnis sekoje Y (su pridėta diskretumo pataisa (apsauga 0,5)). Be pridėtos pataisos būtų $f(+,+,+) = n$. Čia n yra tiriamo posekio ilgis.

2. Generuojama n ilgio Markovo grandinė, kurioje pirmasis tripletas $(k_0c_0d_0)$ generuojamas su tikimybe $p(k_0,c_0,d_0)$, kitas tripletas prasidedantis raide d_0 – su sąlygine tikimybe $p(c_1,d_1|k_1 = d_0)$. Trečias tripletas prasidedantis raide d_1 , generuojamas su sąlygine tikimybe $p(c_2,d_2|k_2 = d_1)$ ir t. t. Generuojama tiek tripletų, kol gaunama n ilgio homogeninė seka Y^* .

3. Ankstesniame žingsnyje gauta seka Y^* suskaidoma į 10 fragmentų, kaip aprašyta pirmame žingsnyje. Toliau 100 kartų vykdomas aukščiau aprašytas penketukų išrinkimo algoritmas ir išsisaugoma 100 atitinkamų p reikšmių.

Posekiai, kuriuose nulinė homogeniškumo hipotezė atsitiktinėms grupėms nebuvo atmesta, laikomi homogeniškais. Kiekvienai bakterijai buvo paimtos 6-ios ilgiausios sekos. 2.2 lentelėje pateikti rezultatai rodo, kad beveik 41,37 % šių sekų posekių yra nehomogeniški.

2.2 lentelė. Posekių homogeniškumas

Table 2.2. The results of homogeneity of subsequences

Bakterija	Nr. 1	Nr. 2	Nr. 3
<i>Escherichia coli</i>	14	6 (42,8 %)	3 / 6 (50 %)
<i>Coxiella burnetti</i>	35	9 (25,7 %)	3 / 9 (33 %)
<i>Bordetella bronchiseptica</i>	37	14 (37,8 %)	6 / 14 (42,85 %)
Viso	86	29 (33,7 %)	12 / 29 (41,37 %)

Nr. 1 - posekių skaičius.

Nr. 2 - posekių skaičius, kuriems pirmos eilės Markovo hipotezė buvo atmesta.

Nr. 3 - posekių skaičius, kuriems homogeniškumo hipotezė buvo atmesta.

2.6. DNR grandinės pirminės ir antrinės vijų palyginimas

DNR grandinės komplementarių vijų sudėties palyginimas dažnai literatūroje įvardijamas kaip apverstoji komplementarioji simetrija (*angl. reverse-complement symmetry*) (Baisnee *et al* 2002) ar apverčiamumo savybė (*angl. reversibility*) (Gordon *et al* 2005) pasidarė aktualiu uždaviniu kai 1953 m. Watson ir Cricks atrado dvigubą DNR spiralės struktūrą, t. y. buvo parodyta, jog DNR grandinė yra sudaryta iš dviejų, pirminės ir antrinės, vijų (*angl. leading and lagging strand*), kurios jungiasi per hidrogenines jungtis tarp purinų ir pirimidinų (A purinas jungiasi su pirimidinu T, G su C). Šiam atradimui pasitarnavo ir Chargaff(1952) rezultatai, gauti tyrinėjant DNR sudėtį ir atradus pirmąją tapatumo (atitikimo) taisyklę (*angl. first parity rule*), kuri teigė, jog visoje DNR sekoje A nukleotidų kiekis labai artimas T, o G kiekis yra artimas C. Antroji tapatumo (atitikimo) taisyklė (*angl. second parity rule*) tvirtina, kad apytikslės lygybės $A \approx T$ ir $G \approx C$ galioja ir atskirose komplementariose vijose. Šios (apytikslės) lygybės yra vienas iš būdų, kaip empiriškai apibrėžti pirminės ir antrinės vijų simetriškumo savybę, t. y. laikoma, kad jei minėtos apytikslės lygybės $A \approx T$ ir $G \approx C$ galioja kiekvienoje atskiroje nagrinėjamos DNR sekos vijoje, tai ta DNR seka yra apverčiama (*angl. reversible*) arba simetriška. Toliau naudosime pastarąjį terminą, nes terminas *apverčiama seka* naudojamas ir Markovo grandinių teorijoje ir turi kitą prasmę (Jensen 2005).

Remiantis tokiu pačiu principu galima apibrėžti DNR simetriškumo savybę imant ir aukštesnės eilės polinukleotidus, t. y. imant nukleotidų poras, triple-

tus ir t. t. Tokiu atveju minėtą DNR sekos simetriškumo savybę turėsime tada, kai (poli)nukleotidų proporcijos jos atskiroje pirminėje ir atskiroje antrinėje vijoje, atitinkamai perkoduotoje pagal komplementarumo savybę bei *perskaitytoje* iš priešingos pusės, bus lygios. Pavyzdžiui, jei DNR seka yra simetriška tai jos pirminėje (arba antrinėje) vijoje dinukleotidų AA ir AC bei trinukleotidų ACT ir CAT proporcijos bus lygios tos pačios vijos atitinkamų dinukleotidų TT ir GT bei trinukleotidų AGT ir ATG proporcijoms. Literatūroje yra parodyta, kad eukariotams ši empirinė savybė galioja (Gordon *et al* 2005).

Simetriškumo savybę galima apibrėžti nebūtinai per atitinkamų polinukleotidų proporcijas. Bendru atveju ši savybė palygina, ar nukleotidų sekos generavimo taisyklė, kuri galioja skaitant pirminę viją iš kairės į dešinę sutampa su antrinės vijos atitinkama generavimo taisykle, kai pastaroji skaitoma iš dešinės į kairę. Toliau jau viskas priklauso nuo to, kaip tos generavimo taisyklės bus apibrėžiamos.

DNR sekų simetriškumo savybės uždavinys sprendžiamas tiek genetikų, biostatistikų, tiek matematikų. Tačiau straipsnyje (Baisnee *et al* 2002) pateiktoje gana plačioje šios tematikos darbų apžvalgoje akcentuojama, kad yra trūkumas mokslinės literatūros, visapusiškai, išsamiai aprašančios DNR sekų apverstojo komplementaraus simetriškumo savybę. Ypatingai yra trūkumas darbų nagrinėjančių aukštesnės eilės simetriją.

Šioje srityje vyrauja darbai, kuriuose empiriškai skaičiuojami komplementarių vijų įvairių (paprastai iki 3-ios eilės) polinukleotidų dažniai ir atitinkamos jų reikšmės palyginamos tarpusavyje grafiškai (vizualiai) arba naudodant įvairius simetriškumo matus. Gauti rezultatai yra gana prieštaringi. Priklausomai nuo to, kokių organizmų DNR naudota, koku būdu lyginta, gaunamos išvados patvirtinančios tiek DNR sekų simetriją (Baisnee *et al* 2002), tiek ir jų asimetriją (Mrazek, Karlin 1998). Be to Baisnee *et al* (2002) atliktas empirinis tyrimas parodė, kad žemesnės eilės DNR sekų simetrija nebūtinai sąlygoja jų aukštesnės eilės simetriją.

Kaip jau buvo minėta anksčiau, DNR sekų simetriškumo problema nagrinėjama ir matematinėje literatūroje Gordon *et al* (2005) pasiūlė *globalų* komplementarių vijų simetriškumo tikimybinį modelį ir tyrė jo suderinamumą, naudodamas įvairių biologinių rūšių duomenis. Disertacijoje pateiktas DNR grandinės vijų *lokalaus* simetriškumo savybės formulavimas ir sudarytas atitinkamas testas nulinei simetriškumo hipotezei tikrinti.

Apibendrintas logit modelis (2.13) yra patogus dviejų komplementarių DNR grandinės vijų *lokalaus* simetriškumo savybei apibrėžti. Šiuo atveju patogiau pradinį alfabetą $\mathcal{A} = \{A, C, G, T\}$ pakeisti nauju alfabetu $\mathcal{A}_2 := \mathcal{A}_1 \times$

$\mathcal{A}_1, \mathcal{A}_1 := \{-1, +1\}$ ir perkodavimui naudoti taisyklę c_1 :

$$(c_1(A), c_1(C), c_1(G), c_1(T)) = ((-1, -1), (+1, +1), (+1, -1), (-1, +1)).$$

Vaizdžiau perkodavimo taisyklę pateikta 2.1 lentelėje. Toliau laikysime, kad DNR sekos elementai priklauso alfabetui $\mathcal{A}_2$. Be to, prisiminę, kad nukleotidai charakterizuojami dviem savybėmis, jungčių skaičiumi ir savybe pirimidinas-purinas, elemento $v \in \mathcal{A}_2$ pirmąją koordinatę simboliškai žymėsime $\beta(v)$, o antrąją žymėsime $\pi(v)$. Pastebėkime, kad jeigu v ir w yra DNR grandinės komplementarūs elementai (nukleotidai), tai $\beta(v) = \beta(w)$ o $\pi(v) = -\pi(w)$.

Tegu $v = (v_1, \dots, v_n) \in \mathcal{A}_2^n$ yra DNR grandinės pirminės vijos n ilgio nukleotidų seka. Kadangi antrinė vija yra sudaryta iš komplementarių nukleotidų ir yra skaitoma iš kito galo (priešinga kryptimi), tai nukleotidų seką v atitiks antrinės vijos nukleotidų seka:

$$\bar{v} := (\bar{v}_n, \dots, \bar{v}_1), \quad \bar{v}_i := (\beta(v_i), -\pi(v_i)), \quad i = 1, \dots, n.$$

Nagrinėkime DNR grandinės pirminės vijos n ilgio nukleotidų seką $x = (x_1, \dots, x_n) \in \mathcal{A}_2^n$. DNR grandinės vijų k -tos eilės simetriškumo savybės empirinė išraiška reiškia, kad bet kuris ilgio k nukleotidų fragmentas $v \in \mathcal{A}_2^k$ pirminėje vijoje x ir antrinėje vijoje $\bar{x}$ pasitaikys apytiksliai vienodą skaičių kartų.

Tuomet bendras tikimybinis empirinės k -tos eilės simetriškumo savybės formulavimas yra išreiškiamas lygybėmis:

$$\sum_{s=0}^{n-k} P_1\{x_{[s+1, s+k]} = v\} = \sum_{t=0}^{n-k} P_2\{\bar{x}_{[t+1, t+k]} = v\}, \quad \forall v \in \mathcal{A}_2^k. \quad (2.17)$$

Kadangi įvykiai $\{x_{[s+1, s+k]} = v\}$ ir $\{x_{[s+1, s+k]} = \bar{v}\}$ sutampa, tai su $\mathbf{P} = P_1$ sąlyga (2.17) ekvivalenti lygybėms:

$$\sum_{s=0}^{n-k} \mathbf{P}\{x_{[s+1, s+k]} = v\} = \sum_{t=0}^{n-k} \mathbf{P}\{x_{[t+1, t+k]} = \bar{v}\}, \quad \forall v \in \mathcal{A}_2^k. \quad (2.18)$$

Remiantis šiuo pastebėjimu ir apibrėžiama DNR vijų simetriškumo savybė.

3 apibrėžimas. (globali simetrija, žiūr. Gordon *et al* 2005) Sakysime, kad DNR sekai $x = (x_1, \dots, x_n) \in \mathcal{A}_2^n$ galioja m -osios eilės *globalaus* simetriškumo savybė, jeigu (2.18) galioja visiems $k \leq m$.

Globalios simetrijos savybė yra labai bendra ir be papildomų prielaidų

apie sekos x skirstinį ją analizuoti sudėtinga. Šios savybės statistinei analizei atlikti neužtenka vienos sekos, reikia turėti tokių sekų imtį. Todėl tarkime, kad x , o tuo pačiu ir $\bar{x}$, yra m -os eilės stacionari (homogeninė) seka. Tai reiškia, kad:

$$\mathbf{P}\{x_{[s+1,s+k]} = v\} = \mathbf{P}\{x_{[t+1,t+k]} = v\}, \quad \forall v \in \mathcal{A}_2^k, s, t \in [0, n-k], k \leq m. \quad (2.19)$$

Kai $m = \infty$, seka x yra stacionari siaurąja prasme, o kai $m = 2$ turime seką stacionarią plačiąja prasme. Jeigu galioja (2.19), tai (2.18) yra ekvivalenti sąlygai:

$$\mathbf{P}\{x_{[s+1,s+k]} = v\} = \mathbf{P}\{x_{[t+1,t+k]} = \bar{v}\}, \quad \forall v \in \mathcal{A}_2^k, s, t \in [0, n-k], k \leq m. \quad (2.20)$$

Vadinasi, jeigu x yra homogeninė m -os eilės Markovo grandinė ir turi $(m+1)$ -os eilės globalaus simetriškumo savybę, tai:

$$\mathbf{P}\{x_\ell = v | x_{U_m(\ell)} = u\} = \mathbf{P}\{x_{\ell'} = \bar{v} | x_{U_m(\ell')} = \bar{u}\} \quad (2.21)$$

su visais $\ell, \ell' \in N^\circ$, $v \in \mathcal{A}_2$, $u \in \mathcal{A}_2^{2m}$.

Lygybės (2.21) sudaro *lokalios* simetrijos apibrėžimo pagrindą. Aišku, jog tam, kad ją apibrėžtume, nebūtina reikalauti, kad seka x būtų m -os eilės homogeninė Markovo grandinė ir turi $(m+1)$ -os eilės globalaus simetriškumo savybę. Svarbu tik, kad sąlyginės tikimybės lygybėse (2.21) būtų apibrėžtos ir nepriklausytų nuo pozicijos $\ell, \ell' \in N^\circ$.

4 apibrėžimas. (lokali simetrija) Sakysime, kad atsitiktinė seka $x = (x_1, \dots, x_n) \in \mathcal{A}_2^n$, kuriai su duotu $m < n/2$ yra apibrėžtos sąlyginės tikimybės

$$\mathbf{P}\{x_\ell = v | x_{U_m(\ell)} = u\}, \quad \ell \in N^\circ, v \in \mathcal{A}_2, u \in \mathcal{A}_2^{2m},$$

ir jos nepriklauso nuo vietos $\ell \in N^\circ$, turi m -osios eilės *lokalus* simetriškumo savybę, jeigu su visais $v \in \mathcal{A}_2$ ir $u \in \mathcal{A}_2^{2m}$:

$$\mathbf{P}\{x_{m+1} = v | x_{U_m(m+1)} = u\} = \mathbf{P}\{x_{m+1} = \bar{v} | x_{U_m(m+1)} = \bar{u}\}. \quad (2.22)$$

Įvestą sekų lokalaus simetriškumo savybę (2.22) galima aprašyti jau anksčiau įvestų (2.3) sąlyginių tikimybių $p(v|u)$ arba *sąlyginių šansų* su bazine reikšme $b = (-1, -1)$ logaritmo (*apibendrintų logit funkcijų*) $\Lambda_{v/b}(u)$ terminais ($v \in \mathcal{A}_2, u \in \mathcal{A}_2^{2m}$). Toliau pateikiamuose teiginiuose bus naudojamas *sąlyginių šansų* žymėjimas, nekartoiant indekse bazinės reikšmės b , t.y. $\Lambda_v(u)$.

5 apibrėžimas. Funkciją $\psi : \mathcal{A}_2^{2m} \rightarrow \mathbf{R}$ vadinsime *simetrine (antisimetrine)* jeigu $\psi(u) = \psi(\bar{u})$ (atitinkamai, $\psi(u) = -\psi(\bar{u})$) su visais $u \in \mathcal{A}_2^{2m}$.

1 teiginys. Jeigu x yra m -os eilės simetrinė seka, tai egzistuoja simetrinė funkcija ψ ir dvi antisimetrinės funkcijos ψ_- ir ψ_+ tokios, kad:

$$\Lambda_{-1,1}(u) = \psi_-(u), \quad (2.23)$$

$$\Lambda_{1,1}(u) - \Lambda_{1,-1}(u) = \psi_+(u), \quad (2.24)$$

$$\Lambda_{1,1}(u) + \Lambda_{1,-1}(u) - \Lambda_{-1,1}(u) = \psi(u). \quad (2.25)$$

Kita lygčių (2.23)–(2.25) užrašymo forma:

$$\log \left(\frac{p((-1,1) | u)}{p((-1,-1) | u)} \right) = \psi_-(u), \quad (2.26)$$

$$\log \left(\frac{p((1,1) | u)}{p((1,-1) | u)} \right) = \psi_+(u), \quad (2.27)$$

$$\log \left(\frac{p((1,1) | u) \cdot p((1,-1) | u)}{p((-1,1) | u) \cdot p((-1,-1) | u)} \right) = \psi(u), \quad \forall u \in \mathcal{A}_2^{2m} \quad (2.28)$$

Irodymas. Naudojantis apibendrintos logit funkcijos (sąlyginių šansų) apibrėžimu (2.3) ir atsižvelgiant į koduotę (2.1) galima parašyti:

$$\Lambda_{-1,1}(u) = \log \left(\frac{p((-1,1) | u)}{p((-1,-1) | u)} \right),$$

$$\Lambda_{1,1}(u) - \Lambda_{1,-1}(u) = \log \left(\frac{p((1,1) | u)}{p((1,-1) | u)} \right),$$

$$\Lambda_{1,1}(u) + \Lambda_{1,-1}(u) - \Lambda_{-1,1}(u) = \log \left(\frac{p((1,1) | u) \cdot p((1,-1) | u)}{p((-1,1) | u) \cdot p((-1,-1) | u)} \right)$$

Lieka patikrinti, kad funkcijos (2.26)–(2.28) turi reikiamas savybes.

Pradėsime nuo (2.26). Remiantis simetriškumo apibrėžimu:

$$p((\beta, \pi) | u) = p((\beta, -\pi) | \bar{u}), \quad \forall u \in \mathcal{A}_2^{2m}. \quad (2.29)$$

Todėl su visais $u \in \mathcal{A}_2^{2m}$

$$\psi_-(u) := \log \left(\frac{p((-1,1) | u)}{p((-1,-1) | u)} \right) = \log \left(\frac{p((-1,-1) | \bar{u})}{p((-1,1) | \bar{u})} \right) \quad (2.30)$$

$$= -\log \left(\frac{p((-1,1) | \bar{u})}{p((-1,-1) | \bar{u})} \right) = -\psi_-(\bar{u}). \quad (2.31)$$

Taigi, funkcija $\psi_-(u)$ yra antisimetrinė.

Analogiškai:

$$\psi_+(u) := \log \left(\frac{p((1,1) | u)}{p((1,-1) | u)} \right) = \log \left(\frac{p((1,-1) | \bar{u})}{p((1,1) | \bar{u})} \right) \quad (2.32)$$

$$= -\log \left(\frac{p((1,1) | \bar{u})}{p((1,-1) | \bar{u})} \right) = -\psi_+(\bar{u}) \quad (2.33)$$

ir

$$\psi(u) := \log \left(\frac{p((1,1) | u)p((-1,1) | u)}{p((-1,1) | u)p((-1,-1) | u)} \right) \quad (2.34)$$

$$= \log \left(\frac{p((1,-1) | \bar{u})p((1,1) | \bar{u})}{p((-1,-1) | \bar{u})p((-1,1) | \bar{u})} \right) = \psi(\bar{u}). \quad (2.35)$$

Pastarosios dvi lygybės užbaigia teiginio įrodymą.

Sprendžiant nežinomų funkcijų $\psi_+(u)$, $\psi_-(u)$ ir $\psi(u)$ statistinio vertinimo uždavinį, šias funkcijas reikia parametrizuoti. Bendru atveju parametrų skaičius priklauso nuo m ir labai greitai auga didėjant m . Kai $m = 1$ antisimetrinė funkcija priklauso nuo 6, o simetrinė funkcija nuo 10 (laisvų) parametrų.

Seką $v = (v_1, \dots, v_k) \in \mathcal{A}_2^k$, ($k \in N$) išskaidykime į dvi dalis w ir z :

$$w := \beta(v), \quad \beta(v) = (\beta(v_1), \dots, \beta(v_k)) \in \mathcal{A}_1^k, \quad (2.36)$$

$$z := \pi(v), \quad \pi(v) = (\pi(v_1), \dots, \pi(v_k)) \in \mathcal{A}_1^k. \quad (2.37)$$

Toliau laikysime, kad tiriama funkcija $\psi = \psi(w, z)$, $\psi : \mathcal{A}_1^k \times \mathcal{A}_1^k \rightarrow \mathbf{R}$.

Tegu aibė $J \subset K := \{1, \dots, k\}$. Įveskime jai atitinkamą *apverstą* aibę J_* :

$$J_* := k + 1 - J = \{k + 1 - j : j \in J\}. \quad (2.38)$$

Duotam $x = (x_1, \dots, x_k) \in \mathbf{R}^k$ žymėsime:

$$x^J := \prod_{i \in J} x_i, \quad x^\emptyset := 1. \quad (2.39)$$

Tuomet bet kurią funkciją $\psi : \mathcal{A}_1^k \times \mathcal{A}_1^k \rightarrow \mathbf{R}$ galima vienareikšmiškai užrašyti tokiu pavidalu:

$$\psi(w, z) = \sum_{J, J' \subset K} a_{J, J'} w^J z^{J'}, \quad w, z \in \mathcal{A}_1^k, \quad (2.40)$$

čia sumavimas vyksta pagal visus aibės K poaibius (įskaitant ir tuščią poaibį $\emptyset$), $a_{J,J'} = a_{J,J'}(\psi)$, $J, J' \subset K$, yra laisvi parametrai, nusakantys funkciją ψ . Iš viso turime 4^k laisvų parametrų. Formulė (2.40) yra gerai žinoma ir (lengvai) išplaukia iš Mobiuso išskaidymo (1.11).

Jeigu funkcija ψ yra simetrinė (antisimetrinė), tai:

$$\psi(w, z) = \psi(-w, z), \quad \text{atitinkamai } \psi(-w, z) = -\psi(w, z). \quad (2.41)$$

Vadinasi, simetrinės funkcijos ψ atveju:

$$\sum_{J', J \subset K} a_{J', J} w^{J'} z^J = \sum_{J, J' \subset K} (-1)^{|J|} a_{J', J} w^{J'} z^{J*}, \quad \forall w, z \in \mathcal{A}_1^k, \quad (2.42)$$

ir vadinasi, jos koeficientai a turi tenkinti lygybes

$$a_{J', J} = (-1)^{|J|} a_{J', J*}, \quad J', J \subset K. \quad (2.43)$$

Taigi, tuo atveju kai $J_* = J$ ir $J'_* = J'$ (poaibių pora (J', J) yra „apverstai simetriška“), poaibio J dydis $|J|$ yra lyginis skaičius, lygybės (2.43) virsta tapatybėmis. Todėl koeficientai $a_{J', J}$ gali įgyti bet kokias reikšmes. „Apverstai simetriškų“ aibės K poaibių skaičių žymėsime $k_* = k_*(k)$.

Jeigu visų aibės K poaibių porų (J', J) klasėje įvesti kokią nors tvarką τ (t. y. tas poaibių poras koku nors būdu sunumeruoti), tai iš (2.43) išplaukia, kad koeficientai $a_{J', J}$, $\tau(J', J) < \tau(J'_*, J_*)$, vienareikšmiškai nusakys likusių (ne „apverstai simetriškų“ poaibių porų) koeficientų $a_{J', J}$, $\tau(J', J) > \tau(J'_*, J_*)$, reikšmes.

Apibrėžkime:

$$\begin{aligned} \mathcal{K}_2 &:= \{(J', J) : \tau(J', J) < \tau(J'_*, J_*)\}, \\ \mathcal{K}_{20} &:= \{(J', J) : \tau(J', J) = \tau(J'_*, J_*)\}. \end{aligned} \quad (2.44)$$

Tokiu būdu iš (2.40), (2.41), (2.43) atsižvelgiant į pažymėjimus (2.44) išvedama, kad simetrinės funkcijos ψ parametrinis pavidalas yra toks:

$$\psi(w, z) = \sum_{(J, J') \in \mathcal{K}_{20}} a_{J, J'} w^J z^{J'} + \sum_{(J, J') \in \mathcal{K}_2} a_{J, J'} \left(w^J z^{J'} + \bar{w}^J \bar{z}^{J'} \right). \quad (2.45)$$

Vadinasi, simetrinė funkcija ψ turi

$$k_s = k_s(k) := k_*^2 + (4^k - k_*^2)/2 \quad (2.46)$$

laisvų parametų.

Kai $m = 1$, $k_* = 1 + 1 = 2$ ir $k_s = 2^2 + (4^2 - 2^2)/2 = 4 + (16 - 4)/2 = 10$.

Pakartojus tuos pačius argumentus, kaip ir simetriniu atveju, gaunasi, kad antisimetrinė funkcija ψ turi tokį pavidalą:

$$\psi(w, z) = \sum_{(J, J') \in \mathcal{K}_2} a_{J, J'} (w^J z^{J'} - \bar{w}^J \bar{z}^{J'}) \quad (2.47)$$

ir priklauso nuo

$$k_a = k_a(k) := (4^k - k_*^2)/2 \quad (2.48)$$

laisvų parametų.

Antisimetrinio atvejo ypatumas lyginant jį su simetriniu atveju tik tame, kad lygybių (2.43) vienoje pusėje atsiranda papildomas minuso ženklas, ir todėl aibės K „apverstai simetriškoms“ poroms (J', J) šios lygybės galioja tik tada, kai koeficientai $a_{J', J} = 0$. Dėl to formulėse (2.47) ir (2.48) prapuola formulių (2.45) ir (2.46) dešinės pusės pirmasis dėmuo.

Kai $m = 1$, $k_a = (4^2 - 2^2)/2 = 6$. Užrašysime detalią simetrinės funkcijos ψ išraišką, kai $m = 1$, t.y. kai turime pirmos eilės Markovo grandinę. Šiuo atveju aibė $K = \{\emptyset, 1, 3\}$ ir

$$J = J' = \{\emptyset, 1, 3, 13\}, J_* = J'_* = \{\emptyset, 3, 1, 31\}$$

Poaibių J ir J' poros ir atitinkamai šių porų „apverstos“ poros J_* ir J'_* :

$$\begin{aligned} JJ' = & \{(\emptyset \emptyset), (\emptyset 1), (\emptyset 3), (\emptyset 13), (1 \emptyset), (1 1), (1 3), (1 13), \\ & (3 \emptyset), (3 1), (3 3), (3 13), (13 \emptyset), (13 1), (13 3), (13 13)\} \end{aligned} \quad (2.49)$$

$$\begin{aligned} J_* J'_* = & \{(\emptyset \emptyset), (\emptyset 3), (\emptyset 1), (\emptyset 31), (3 \emptyset), (3 3), (3 1), (3 31), \\ & (1 \emptyset), (1 3), (1 1), (1 31), (31 \emptyset), (31 3), (31 1), (31 31)\} \end{aligned} \quad (2.50)$$

Jeigu sunumeruosime visas galimas poaibių J ir J' poras eilės tvarka, taip kaip surašyta 2.49, tai aibę $\mathcal{K}_{20}$ sudarys keturios „simetriškai apverčiamos“ poros:

$$\mathcal{K}_{20} = \{(\emptyset \emptyset), (\emptyset 13), (13 \emptyset), (13 13)\}$$

Pagal (2.44) aibė $\mathcal{K}_2$ sudalija poras JJ' į dvi dalis:

$$\mathcal{K}_2^I = \{(\emptyset 1), (1 \emptyset), (1 1), (1 13), (13 1)\}$$

$$\mathcal{K}_2^{II} = \{(\emptyset 3), (3 \emptyset), (3 3), (3 13), (13 3)\}$$

Tada simetrinė funkcija (2.45) $m = 1$ atveju atrodys taip:

$$\begin{aligned} \psi(w, z) = & a_{\emptyset\emptyset} + a_{\emptyset 13} z^{13} + a_{13 \emptyset} w^{13} + a_{13 13} w^{13} z^{13} \\ & + a_{\emptyset 1} (z^1 + z^3) + a_{1\emptyset} (w^1 - w^3) + a_{13} (w^1 z^3 - w^3 z^1) \\ & + a_{11} (w^1 z^1 - w^3 z^3) + a_{1 13} (w^1 z^{13} - w^3 z^{13}) + a_{13 1} (w^{13} z^1 + w^{13} z^3) \end{aligned} \quad (2.51)$$

O asimetrinė funkcija (2.47) $m = 1$ atveju atrodys taip:

$$\begin{aligned} \psi(w, z) = & a_{\emptyset 1} (z^1 - z^3) + a_{1\emptyset} (w^1 + w^3) + a_{11} (w^1 z^1 + w^3 z^3) \\ & + a_{13} (w^1 z^3 + w^3 z^1) + a_{1 13} (w^1 z^{13} + w^3 z^{13}) + a_{13 1} (w^{13} z^1 - w^{13} z^3) \end{aligned} \quad (2.52)$$

2.3 lentelė. DNR vijų simetriškumo hipotezės rezultatai

Table 2.3. The results of reverse-complement symmetry of DNA

Bakterija	Sekos nr.	Sekos ilgis	p*	p**	p***
<i>Esherichia coli</i>	320	1696	0,003	0,0006	0,122
	950	1528	0,097		
	1087	2138	0,045		
	1092	1594	0,022		
	1598	2352	0,279		
	1599	1532	0,008		
<i>Coxiella burnetti</i>	70	7484	0,005	<0,0001	0,91
	161	3194	0,0004		
	355	4526	<0,0001		
	515	3552	0,0011		
	600	4594	0,0005		
	758	2886	0,282		
<i>Bordetella bronchiseptica</i>	854	7774	<0,0001	0,966	0,064
	1471	6540	<0,0001		
	441	4974	0,0283		
	1411	3100	0,1185		
	1445	3002	0,0003		

* – kiekvienai 2-ame stulpelyje nurodytai sekai.

** – 2-ame stulpelyje nurodytoms sekoms bendrai.

*** – visos bakterijos nekoduojančioms sekoms bendrai.

2.7. Logit statistinės analizės išvados

Įvesta speciali stebėtų duomenų struktūra, kuri remiasi prielaidomis (A1), (A2) ir apibendrintu logit modeliu. Tai tiesiogiai susiję su baigtinių būsenų Markovo laukų specifikacija ir dėl to yra patogų pasiūlytą metodiką taikyti Markovo priklausomybių statistinei analizei. Be to, apibendrintas logit modelis yra lankstesnis nei kaip alternatyva naudojami nesąlyginiai logtiesiniai modeliai ir leidžia praplėsti sekų nehomogeniškumo tyrimus.

Yra natūralu tikėtis, kad DNR sekų nehomogeniškumas įtakoja, pavyzdžiui, jų pirmos eilės Markovo modelio atmetimą. Atlikta procedūra, kuri remiasi perrinkimu, yra skirta nustatyti, ar nemarkoviškumas sekose yra įtakojamas posekių nehomogeniškumo.

Gauti rezultatai rodo, kad remiantis apibendrintu logit modeliu, DNR sekos yra gerokai nehomogeniškos ir tai gali įtakoti jų nemarkoviškumą. Iš kitos pusės, buvo nustatyta, kad reikšminga dalis nemarkovo sekų (apie 58,63%) yra homogeniškos. Taigi, net ir homogeniški DNR regionai negali būti traktuojami pirmos eilės Markovo sekomis.

Kadangi augant Markovo eilei, vertinamų parametrų skaičius Markovo modelyje auga eksponentiniu greičiu, o daugelis nekoduojančių sekų yra trumpos ir/arba nehomogeniškos, vertinti aukštesnės eilės Markovo modelį pasidaro sudėtinga.

Pasinaudojant logit modeliu buvo suformuluota ir DNR komplementarių vijų simetriškumo hipotezė. Statistinė analizė parodė (2.3 lentelė), kad lokaliai DNR sekų vijos nėra simetriškos.

3

Išsklaidytų kategorinių duomenų semiparametrinis glodinimas

Šiame skyriuje nagrinėjama išsklaidytų dažnių lentelių problema. Pirmiausia trumpai apžvelgiami literatūroje žinomi jos sprendimo būdai, toliau pateikiami disertacijos rezultatai, publikuoti straipsnyje (2009). Jame pasiūlytas (išsklaidytų) kategorinių duomenų glodinimo metodas bei juo paremtas savirankos algoritmas, skirtas hipotezei apie kategorinių požymių tarpusavio priklausomybę tikrinti. Šis metodas glaudžiai siejasi su prieš tai skyriuje (2) aprašytu lokalsios priklausomybių struktūros DNR sekose vertinimo uždaviniu bei tam tikslui naudotu apibendrintu logit modeliu – tai priklausomų kategorinių duomenų atvejis, kai galima taikyti pasiūlytą semiparametrinio glodinimo metodą.

3.1. Išsklaidytų dažnių lentelių problema

Išsklaidyti duomenys dažniausiai gaunami tada kai į tyrimą įtraukiama didelis skaičius kategorinių kintamųjų arba kintamųjų su daug kategorijų, t.y. kai turime dažnių lenteles su (santykinai) dideliu, lyginant su imties dydžiu, ląstelių skaičiumi. Tokiu atveju dažnių lentelėje yra daug tuščių ląstelių arba ląstelių su mažu stebinių skaičiumi jose. Gera žinoma, kad tokiu atveju įprastiniai statistiniai metodai lentelėms nėra tinkami, nes taikomų kriterijų skirstiniams χ^2 aproksimacija nėra pakankamai tiksli.

3. IŠSKLAIDYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS GLODINIMAS

Sąvoka „mažas stebėjimų skaičius ląstelėse“ nėra tiksliai apibrėžta, yra pateikta įvairių rekomendacijų, kada dažnių lentelę jau reiktų laikyti išsklaidyta: Kendall (1952) siūlė, kad visi tikėtini dažniai ląstelėse būtų didesni už 20, Cramer teigė, kad 10, Fisher (1938) siūlė mažiausiai 5, Jeffreys (1961), Slakter (1965) ir Kempthorne (1966) laikė pakankamu 1 stebėjimą. Iki šiol yra laikomasi rekomendacijos, kad (galiojant nulinei hipotezei), dažniai daugumoje ląstelių būtų didesni už 5 (Hu 1999; Agresti, Hitchcock 2005). Jei ši sąlyga negalioja, χ^2 aproksimacija taikomai statistikai gali būti netinkama ir tada paprastai sakoma, kad dažnių lentelė yra išsklaidyta.

3.2. Literatūros apie išsklaidytas dažnių lenteles apžvalga

Išsklaidytų dažnių lentelių problema yra seniai žinoma, ir literatūroje jai skirta daug dėmesio (žr. apžvalgas Marhoul 1985; Hu 1999). Viena iš pirmųjų ir paprasčiausių tuščioms ląstelėms siūlomų taisyklių – sujungti gretimas lentelės eilutes (Craig 1953). Šį būdą logtiesiniams modeliams taikė Bishop (1971). Vis dėlto didelėms lentelėms su daug tuščių ląstelių šis metodas nėra veiksmingas, nes jis pradinę lentelę gerokai sumažina, prarandant dalį informacijos.

Kita siūloma taisyklė – tuščias ląsteles pakeisti tam tikru teigiamu skaičiumi ir atlikti analizę su pakeista lentele. Paprasčiausias šios taisyklės metodas yra pasiūlytas Berkson (1955), jis siūlė 0 keisti skaičiumi 1/2.

Kad galima būtų naudoti visus išsklaidytos dažnių lentelės duomenis, vėliau buvo pasiūlyti tikslieji testai (*angl. exact tests*), pavyzdžiui, gerai žinomas tikslusis Fišerio testas bei jo apibendrinimai ir išvystymai, ar χ^2 ir kitų kriterijų statistikų skirstinių asimptotinės aproksimacijos (Agresti 1990; Hu 1999; Muller, Osius 2003).

Logtiesinių modelių analizėje naudojamai G^2 statistikai Koehler (1986) gavo įdomius rezultatus, kurie teigė, kad prie tam tikrų sąlygų normalioji aproksimacija gali būti daug tikslesnė nei χ^2 aproksimacija. Kai lentelėje ląstelių yra labai daug, natūralu tikėtis, kad G^2 skirstinys bus artimas normaliajam, bet bendru atveju literatūroje parodyta, kad gali būti tinkamesnės aproksimacijos ir kitais skirstiniais: lognormaliuoju (Lawal, Upton 1980), gama (Hu 1999) ar skirstinių mišiniais (Yarnold 1970; Cochran 1952).

Iš tikrųjų, nė vienas bandymas išvystyti bendrą asimptotinę teoriją nebuvo sėkmingas, nepavyko pasiūlyti bendros taisyklės, kuri be papildomų sąlygų

3. IŠSKLAIDYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS GLODINIMAS

57

nusakytų tinkamą bazinį (*angl. reference*) testo statistikos skirstinį.

Pasitelkiant šiuolaikinius kompiuterinius resursus, atsiranda vis daugiau galimybių vystyti ir tam tikrus statistinius metodus, tokius kaip glodinimas, savirankos testai (von Davier 1997), Bajeso metodai (Agresti, Hitchcock 2005; Congdon 2005) ir kt.

Turint tolydžius kintamuosius, glodinimo metu iš duomenų galima įvertinti neparametrinį tankio, pasiskirstymo ar regresijos funkcijas. Neparametrinis tankio įvertis yra tam tikra prasme geras visiems tinkamai suglodintiems tankiams, bet parametrinio tankio įverčio gerumas priklausys nuo to, kaip parinktas modelis: jei gerai, tai glodinimo rezultatas bus puikus, bet jei nėra tinkamai parinktas modelis – glodinimo rezultatas bus labai prastas. Dėl to apibrėžiamos glodinimo taisyklės, kuriose atskiras parametras kontroliuoja santykį tarp naudojamo parametrinio ir neparametrinio glodinimo. Statistikos teorijoje tokie modeliai vadinami semiparametriniais. Šiuo terminu norima pabrėžti, kad tyrėjo pats skirstinys nedomina, jam svarbūs tik keli jo parametrai, kurie pačio skirstinio dar pilnai nenusako.

Pirmieji semiparametrinio glodinimo darbai yra atlikti tolydiems duomenims (Olkin, Spiegelman 1987; Copas 1995; Hjort, Glad 1995 ir kt.) Pirmieji semiparametrinį glodinimą kokybiniais duomenims su parametrine komponente iš standartinių diskrečių skirstinių klasės (Binominis, Puasono) pasiūlė Faddy ir Jones (1998). Jų pasiūlytame metode parametrinė komponentė yra stacionarus DNR grandinės skirstinys ir kaip glodinimo mechanizmas naudojamos atitinkamos perėjimo tikimybių matricos.

Simonoff (1995, 1998) kokybinių duomenų tikimybinio skirstinio įvertinimą nurodo kaip mažiausiai išvystytą glodinimo metodų kryptį ir ją įvardija kaip vieną iš perspektyviausių būdų, taikytinų išsklaidytų dažnių lentelių analizėje. Glodinimas suteikia galimybę užpildyti tuščias ląsteles, pasinaudojant kaimyninių dažnių lentelės ląstelių informacija. Tačiau didžioji dalis darbų yra skirta ranginių kokybinių kintamųjų glodinimui (Simonoff 1996, 1998), nes nominaliems požymiams yra gana sudėtinga apibrėžti natūralų „artumo“ matą tarp jo įgijamų kategorijų.

Visi šiame skyrelyje paminėti metodai nėra pritaikomi arba turi tam tikrų apribojimų kai stebimi kokybiniai duomenys yra priklausomi ir jie sudaro stipriai išsklaidytas dažnių lenteles, kuriose daug nulinių stebėjimų. Genetinės sekos yra tipinis tokių duomenų pavyzdys.

3.3. Semiparametrinis glodinimas ir saviranka

Šiame skyrelyje yra aprašytas hipotezių tikrinimo metodas, paremtas semiparametriniu glodinimu ir savirankos algoritmu. Metodas yra gana bendras ir gali būti taikomas bet kokio tipo duomenims (2.2 pavyzdys), bet disertacijoje šis metodas naudojamas atsitiktinėms sekoms su baigtiniu būsenų skaičiumi.

Toliau bus naudojami tokie pažymėjimai: $\mathbf{Z}_+ := \{0, 1, 2, \dots\}$, $\|\cdot\|_p$ žymi L_p – normą, δ_y yra Dirako matas sukoncentruotas taške y , I yra vienetinė matrica (operatorius), $|A|$ – aibės A elementų skaičius.

3.3.1. Semiparametrinis glodinimas

Tarkime $Y = (Y_l, l = 1, \dots, n) \in \mathcal{A}^n$, $\mathcal{A}^n = \bigotimes_{i=1}^n \mathcal{A}$ yra atsitiktinė seka su nežinomu skirstiniu μ . Čia $\mathcal{A}$ yra baigtinė būsenų aibė (baigtinis alfabetas). Toliau, tegu $X(h) = (X_l(h), l = 1, \dots, n) \in \mathcal{A}^n$, $h \in \mathbf{Z}_+$ yra diskretaus laiko homogeninė Markovo grandinė su perėjimo tikimybių matrica (Markovo branduoliu) Π ir pradiniu skirstiniu π , $\pi(a) = \mathbf{P}\{X(0) = a\}$, $a \in \mathcal{A}$.

Tarkime, kad Markovo grandinė X yra neredukuojama ir acikliška. Tada ji yra ergodiška. Pažymėkime q jos vienintelį stacionarų (invariantišką) skirstinį. Jei perėjimo tikimybių matricas Π imsime iš duotosios (parametrinės) Markovo branduolių šeimos $\mathcal{P}$, tai atitinkami invariantiški „tikslai“ (*target*) skirstiniai q sudarys (parametrinę) skirstinių šeimą $\mathcal{Q}$. Taigi, $q \in \mathcal{Q}$ tada ir tik tada, kai q yra invariantiškas skirstinys tam tikram $\Pi \in \mathcal{P}$.

1 uždavinys. Turint Y stebinį y , patikrinti suderinamumo (*angl. goodness-of-fit*) hipotezę:

$$H_0 : \mu \in \mathcal{Q} \quad \text{su alternatyva} \quad H_1 : \mu \notin \mathcal{Q},$$

kai žinoma tik perėjimo tikimybių matrica $\Pi \in \mathcal{P}$ arba dar bendriau – kai yra žinomas tik išrinkimo (*angl. sampling*) iš $q \in \mathcal{Q}$ metodas.

Tegu $X(h) = X(h|y) = \{X_l(h), l = 1, \dots, n\}$, $h \in \mathbf{Z}_+$, yra homogeninė Markovo grandinė su perėjimo tikimybių matrica Π , prasidedanti taške y , t. y. Markovo grandinė su pradiniu skirstiniu $\pi(a) = \delta_y(a)$, $a \in \mathcal{A}$. Pažymėkime $q_h(a) = q_h(a|y) := \mathbf{P}\{X(h) = a | X(0) = y\}$, $a \in \mathcal{A}$, $h \in \mathbf{Z}_+$. Tada $q_0 = \delta_y$ ir $q_h \rightarrow q$, kai $h \rightarrow \infty$, eksponentiniu konvergavimo greičiu su tam tikru $\kappa > 0$:

$$\|q_h - q\| \leq e^{-\kappa h} \|q_0 - q\|, \quad \forall h \in \mathbf{Z}_+, \quad (3.1)$$

3. IŠSKLAIDYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS GLODINIMAS

59

čia $\|\nu\| := \sum_{a \in \mathcal{A}} \nu(a)$ žymi (apibendrinto) mato pilnąją variaciją aibėje $\mathcal{A}$.

Norint sumažinti (pakeisti) skirstinių q_h glodinimo efektą, galima imti jų svertinį vidurkį:

$$\bar{q}_h(a) = \sum_{t=0}^h \omega_h(t) q_t(a|y), \quad h \in \mathbf{Z}_+, \quad (3.2)$$

čia $\omega_h(t) \geq 0$ yra svorių rinkinys, $\sum_{t=0}^h \omega_h(t) = 1 \quad \forall h \in \mathbf{Z}_+$. Natūralu reikalauti, kad $\omega_h(t) \rightarrow 0$, kai $h \rightarrow \infty$, su visais $t \in \mathbf{Z}_+$. Toliau bus laikoma, kad šis rinkinys yra fiksuotas, ir mes jo žymėjimuose nenurodysime.

Iš (3.1) nelygybės išplaukia toks teiginys, kuriuo remiasi semiparametrinio glodinimo taikymai hipotezių tikrinimo uždaviniuose:

1 teiginys. Tegu homogeninė Markovo grandinė X yra ergodiška. Jeigu su kuriuo nors $h \in \mathbf{Z}_+$ ir $\delta > 0$ galioja nelygybė $\|q_h - q\| > \delta$, tai ir $\|\mu - q\| > \delta$.

Taigi, fiksuotam $h > 0$, $\mu_h(\cdot) := \mathbf{E}\bar{q}_h(\cdot|Y)$ gali būti traktuojamas kaip Y skirstinys „suglodintas link skirstinio q “ su glodinimo parametru h . Parametro h parinkimo uždavinys aptariamas tolesniuose skyreliuose. Skirstinys q yra traktuojamas kaip bazinis, tikslinis (*angl. target*) (parametrinis) tikimybinis modelis.

Faddy ir Jones (1998) šį metodą yra pavadinę semiparametriniu glodiniu. Aplamai, bet kuri glodinimo procedūra yra tam tikra prasme semiparametrinė, nes ji iš vienos pusės nesiremia jokių parametriniu modeliu, o iš kitos pusės priklauso nuo tam tikrų skaliarinių glodinimo parametrų. Šiuo atveju šis terminas naudojamas pabrėžti tai, kad jeigu „glodinimo“ branduolys Π priklauso nuo kokių nors skaliarinių parametrų, tai nuo tų parametrų gali priklausyti taip pat ir branduolį Π atitinkantis invariantiškas skirstinys q .

2 uždavinys. Pasiūlyta taisyklė suteikia galimybę glodinti kategorinius duomenis. Glodinimas remiasi branduoliais, kurie susieja nestebimo objekto tikėtinumą su stebimų jo kaimynų tikėtinumu, t. y. su tikėtinumu objektų, kurie tam tikra prasme yra panašūs į jį patį.

Semiparametriniame glodiniame branduoliai atsiranda natūraliai kaip ergodinė transformacija, kuri konverguoja į duotą tikslo skirstinį. Praktikoje tai atitinka stabilizuojančią evoliuciją arba paklaidų korekcijos (*angl. error correction*) dinamiką, kurios yra būdingos natūralioms gyvosioms ar socialinėms sistemoms. Bet kuriuo atveju perėjimo branduolio pasirinkimas semiparametrinio glodinimo metode, konstruojant semiparametrinio glodinimo procedūrą, bent jau yra pagrįstas tikrinama nuline hipoteze.

3.3.2. Semiparametrinis savirankos metodas

Pagrindinis neparametrinio savirankos metodo pranašumas, lyginant jį su parametriniu, yra jo paprastumas ir tai, kad jis nepriklauso nuo to, ar tinkamai parinkta parametrinio modelio specifikacija bei procedūrų, vertinančių nežinomus modelio parametrus.

1 pavyzdys. Nepriklausomumo tikrinimas naudojant neparametrinį savirankos metodą.

Tarkime, kad $Y = (Y_l, l = 1, \dots, n)$ yra seka nepriklausomų ir vienodai pasiskirsčiusių m -mačių atsitiktinių dydžių:

$$Y_l = (Y_l^{(k)}, k = 1, \dots, m), \quad Y_l^{(k)} \in \mathcal{A}_k, \quad k = 1, \dots, m, l = 1, \dots, n.$$

Uždavinys – patikrinti komponentų $Y^{(k)}, k = 1, \dots, m$ nepriklausomumą, turint išsklaidytus duomenis, t. y. tuo atveju, kai $\prod_{k=1}^m |\mathcal{A}_k| > n$. Šiuo atveju dažnių lentelėje vidutinis stebėtas dažnis yra mažesnis už 1, ir paprastai joje yra daug nulinių ląstelių, todėl neparametrinis savirankos metodas yra nepritaikomas. Tačiau savirankos metodą galima taikyti sugludintai dažnių lentelei, gautai pagal žemiau aprašytą glodinimo algoritmą:

1 algoritmas. Tarkime y yra gauta imtis iš Y ir $\mathcal{L}(y)$ apibrėžia dažnių lentelę, gautą iš duomenų y . Taip pat pažymėkime $J_n := \{1, \dots, n\}$.

a) Pažymėkime $X(0) = y$ ir $h = 1$;

b) Pasirinkime atsitiktinai su lygiomis tikimybėmis porą $(i, j) \in J_n \times J_n, i \neq j$, ir $k \in J_m$;

c) Pažymėkime $X_i^{(k)}(h) = X_j^{(k)}(h - 1)$ ir $X_j^{(k)}(h) = X_i^{(k)}(h - 1)$. Visiems kitiems atvejams taip pat: $X_l(h) = X_l(h - 1) \forall l \in J_n \setminus \{i, j\}$ ir $X_l^{(r)}(h) = X_l^{(r)}(h - 1)$ visiems $l \in \{i, j\}, r \in J_m \setminus \{k\}$;

d) Jei visos tuščios ląstelės dažnių lentelėje $\mathcal{L}(y)$ gautoje iš y yra išnaudotos (perrinktos), suskaičiuojamas dažnių lentelių $\mathcal{L}(X(t)), t = 0, 1, \dots, h$ vidurkis ir algoritmas sustabdomas. Priešingu atveju – pasirenkamas $h+1 \rightarrow h$ ir grįžtama į žingsnį b.

Akivaizdu, kad stacionarus dydžio X skirstinys turi tas pačias marginalines tikimybes kaip ir y , bet jau nepriklausomas komponentes.

Atsitiktinė seka $X(h|Y)$ yra seka, gauta iš skirstinio μ_h , kuris, esant mažoms h reikšmėms, yra artimas pradiniam (neparametriniam) skirstiniui μ , tuo tarpu didelėms h reikšmėms – jis artimas parametriniam skirstiniui q .

Taigi, pakartotinių imčių iš $\bar{q}_h$ arba tiesiog iš q_h taikymas statistinėje analizėje yra tarsi tarpinis atvejis tarp parametrinio ir neparametrinio savirankos

3. IŠSKLAIDYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS GLODINIMAS

61

metodo. Perrinkimo procedūrą panaudojant $\bar{q}_h$ (arba q_h) vadinsime *semiparametriniu savirankos metodu*.

Kitame skyrelyje aptariamas klausimas, kaip parinkti glodinimo parametro h reikšmę taikant semiparametrinį savirankos metodą hipotezių tikrinimo uždavinyje.

3.3.3. Suderinamumo testai

Pradžioje metodo idėjos iliustracijai pateiksime klasikinį pavyzdį apie kiekybinių (tolydžių) duomenų Gauso glodinimo metodą.

2 pavyzdys. Neparametrinis pasiskirstymo tankio įvertis su Gauso branduoliu. Tegu φ yra skirstinio tankis, sudarytas iš $Z \sim N(0,1)$ ir $Y^n := (Y_1, \dots, Y_n)$ yra nepriklausomi vienodai pasiskirstę standartizuoti atsitiktiniai dydžiai ($\mathbf{E}Y_1 = 0, \mathbf{E}Y_1^2 = 1$) su nežinomu glodžiu pasiskirstymo tankiu f_Y . Uždavinys yra įvertinti f_Y ir patikrinti nulinę suderinamumo hipotezę:

$$H_0: f_Y = \varphi. \quad (3.3)$$

Tegu $Z^n(t) := (Z_1(t), \dots, Z_n(t))$, $t \in \mathbf{Z}_+$, yra seka nepriklausomų standartinių Gauso atsitiktinių dydžių taip pat nepriklausančių ir nuo Y^n .

Fiksuokime $h > 0$ ir nagrinėkime Markovo grandinę $\{Z_h^n(t), t \in \mathbf{Z}_+\}$, prasidedančią taške $X_h^n(0) = Y^n$, o toliau apibrėžiamą lygybėmis:

$$X_h^n(t+1) = \frac{1}{\sqrt{1+h^2}} \left(X_h^n(t) + h \cdot Z^n(t) \right), \quad t \in \mathbf{Z}_+.$$

Sekos $\{X_h^n(t), t \in \mathbf{Z}_+\}$ stacionarus skirstinys yra Gauso. Kadangi $X_h^n(j) = X_H^n(1)$ pagal pasiskirstymą su $H = 1 - (1+h^2)^{-(j+1)}$, tai užtenka nagrinėti tik vieną Markovo grandinės žingsnį. Trumpumo dėlei tegu $X_h^n = X_h^n(1)$.

Tegu $\hat{F}_n(\cdot | Y^n)$ žymi X_h^n sąlyginę empirinę pasiskirstymo funkciją, kai duotas Y^n . Tada

$$\hat{f}_{Y,h}(y) := \frac{d\hat{F}_n(y | Y^n)}{dy} = \frac{\sqrt{1+h^2}}{hn} \sum_{j=1}^n \varphi\left(\frac{y\sqrt{1+h^2} - Y_j}{h}\right)$$

yra klasikinis (standartinis) neparametrinis pasiskirstymo tankio įvertis su Gauso branduoliu ir glodinimo pločiu h .

Jei neatitikimas tarp $\hat{f}_{Y,h}$ ir φ yra matuojamas naudojant integruotą kvadratinę paklaidą (*angl. integrated square error*) $\|\varphi - \hat{f}_{Y,h}\|_2^2$, tai optimalus glodinimo plotis (3.3) hipotezei yra apibrėžiamas kaip santykinės vidutinės inte-

3. IŠSKLAIDYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS 62 GLODINIMAS

gruotos kvadratinės paklaidos maksimizavimo uždavinio sprendinys:

$$\frac{\mathbf{E}\|\varphi - \hat{f}_{Y,h}\|_2^2}{\|\mathbf{Var} \hat{f}_{Z,h}\|_1} \longrightarrow \max_{h>0}. \quad (3.4)$$

Galima tikėtis, kad tokiu būdu parinktas h^* maksimizuoja taip pat ir suderinamumo kriterijaus, kuris remiasi atstumu $\|\varphi - \hat{f}_{Y,h}\|_2^2$, galią. Pavyzdžiui, jei

$$f_Y(y) = \frac{1}{2\sigma_a} \varphi\left(\frac{y-a}{\sigma_a}\right) + \frac{1}{2\sigma_a} \varphi\left(\frac{y+a}{\sigma_a}\right)$$

su $\sigma_a^2 := 1 - a^2$ ir $a = a_n \rightarrow 0$ (kai $n \rightarrow \infty$), tiesioginis suskaičiavimas rodo, kad $h^* = h_0 + o(1)$ ir $\mathbf{E}\|\varphi - \hat{f}_{Y,h^*}\|_2^2 = \kappa n a_n^8 (1 + o(1))$ su tam tikromis konstantomis $h_0 > 0$ ir $\kappa > 0$. Taigi, tam, kad hipotezės $H_0: a = 0$ prieš kontigualią alternatyvą $H_1: a = a_n, a_n \rightarrow 0 (n \rightarrow \infty)$ kriterijus būtų suderintas, būtina, kad $a_n^8 n \rightarrow \infty$.

Pastebėjime, kad skirtingai nei (3.4) taisyklėje, optimalus glodinimo plotis h_n pagal klasikinę taisyklę (Silvermann 1986) yra parenkamas taip, kad minimizuotų vidutinę integruotą kvadratinę paklaidą (*angl. mean integrated square error – MISE*) $\mathbf{E}\|f_Y - \hat{f}_{Y,h}\|_2^2$, ir $h_n \rightarrow 0$ nuo f_Y glodumo priklausančiu greičiu.

Ir nors išvados apie didelės dimensijos kategorinius duomenis yra labiau komplikotos, šis paprastas pavyzdys suteikia minčių, kaip apibrėžti testo statistiką ir optimalų glodinimo parametą. Kitame skyrelyje yra pateikiami kai kurie šio metodo taikymai hipotezių tikrinime turint atsitiktines sekas su baigtiniu būsenų skaičiumi.

Testo statistikos. Nagrinėkime vektorinį funkcionalą $\lambda = \lambda(q)$, apibrėžtą tikimybių matų aibėje $\mathcal{Q}$, ir apibrėžkime χ^2 tipo statistikas:

$$\eta(h) = \eta(h|\ell) := (\hat{\lambda}_h - \hat{\lambda}_\ell)^\top \widehat{W}^- (\hat{\lambda}_h - \hat{\lambda}_\ell). \quad (3.5)$$

Čia $\hat{\lambda}_h = \lambda(\hat{q}_h)$, $\hat{q}_h$ yra empirinis $\bar{q}_h$ skirstinys, gautas savirankos būdu iš $\{X(t), t = 1, \dots, h\}$ su $X(0) = y$, $\widehat{W}$ yra įvertinta dydžių $\hat{\lambda}_h - \hat{\lambda}_\ell$ kovariacinė matrica, o $\widehat{W}^-$ yra apibendrinta atvirkštinė matrica matricai $\widehat{W}$. Kad gautume išsprendžiamą uždavinį tuo atveju kai turime didelio matavimo funkcionalus λ , reikia įvesti tam tikrus apribojimus matricos $\widehat{W}$ struktūrai. Įprastai ši matrica yra diagonalinė (χ^2 statistikoms) arba sferinė ($W = \sigma^2 I, \sigma^2 > 0$, kai nagrinėjamas L_2 -atstumas).

1 pastaba. Aišku, kad vietoj (3.5) galima naudoti kitus normuotus ma-

3. IŠSKLAIDYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS GLODINIMAS

63

tus, matuojančius skirtumus tarp $\hat{q}_h$ ir $\hat{q}_\ell$. Pavyzdžiui, galima naudoti laipsninės nukrypimų šeimos *angl. power-divergence family* statistikas (Cressie, Read 1984).

Nulinės hipotezės tikrinimas remiasi statistika:

$$\eta^* = \eta^*(\ell) := \max_{h \in \mathcal{H}} \eta(h|\ell), \quad (3.6)$$

čia $\mathcal{H} = \{h_1, \dots, h_k\}$, $h_1 < h_2 < \dots < h_k < \ell$ yra duota glodinimo parametro reikšmių aibė. Parametrą ℓ reiktų imti kiek galima didesnį, bet tai didina kompiuterinio laiko sąnaudas.

Statistikos (3.6) skirstinio įvertinimui ir jos kritinių bei p reikšmių nustatymui gali būti naudojami savirankos ir kiti pakartotinių imčių metodai.

Optimali glodinimo parametro h reikšmė yra apibrėžiama lygtimi:

$$h^* = h^*(\ell) := \arg \max_{h \in \mathcal{H}} \eta(h|\ell).$$

2 pastaba. Pasiūlytas semiparametrinis savirankos metodas gali būti naudojamas ir kaip atskiras adaptyvaus glodinimo būdas, kai yra duotas (žinomas) tikslo (parametrinis) skirstinys q .

Statistikos η^* apibrėžime paprastumo dėlei buvo laikoma, kad statistikos $\eta(h)$, $h \in \mathcal{H}$ turi apytiksliai tą patį skirstinį. Tai yra natūrali prielaida, jei $\hat{\lambda}_h$ skirstinys yra artimas Gauso. Bendru atveju, vis gi statistikos $\eta(h)$ skirstinys gali kisti priklausomai nuo h . Todėl šių statistikų *standartizuota* maksimali reikšmė

$$\eta_0^*(\ell) := \max_{h \in \mathcal{H}} (\eta(h, \ell) - a(h))/b(h),$$

(ar dar sudėtingesni dydžių $\eta(\cdot|\ell)$) funkcionalai) gali būti labiau tinkama negu (3.6). Čia $a(h)$ ir $b(h)$ yra centravimo ir normavimo parametrai, kuriuos reiktų tinkamai parinkti. Tai yra atskira tema ir šiame darbe mes jos nenagrinėsime.

3.4. Apibendrintas logit modelis ir semiparametrinis savirankos metodas

Efektyvų semiparametrinio glodinimo ir savirankos algoritmo realizacija reikalauja apibrėžti atitinkamą Markovo branduolį Π (iš šeimos $\mathcal{P}$), funkcionalą λ statistikoje (3.5) ir svorius $\omega_h(\cdot)$ (žiūr. (3.2)). Bendru atveju tai gana sudėtinga problema. Šis skyrelis yra skirtas lokalioms priklausomybių struktūros baigtinio alfabeto atsitiktinėse sekose įvertinimui orientuojantis į taikymus DNR

64 3. IŠSKLAIDYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS GLODINIMAS

genetinių sekų analizėje. Tam naudojama ta pati, duomenų struktūra, kaip ir 2 skyriuje. Paprastumo dėlei čia apsiribojama tik artimiausių kaimynų sąveikomis.

Tegu $y = \{y_i, i = 1, \dots, 2m + 1\}$, $y_i \in \mathcal{A}$, yra stebima seka. DNR sekoms $\mathcal{A} = \{A, C, G, T\}$, $|\mathcal{A}| := \text{card}(\mathcal{A}) = 4$. Tegu l, c, r apibrėžia atitinkamai kairįjį, centrinį ir dešinįjį tripleto (l, c, r) elementus, ir tegu

$$\{(l_i, c_i, r_i), i \in J_m\} := \{(y_{2i-1}, y_{2i}, y_{2i+1}), i \in J_m\} \quad (3.7)$$

yra tripletų rinkinys, sudarytas iš kas antro sekos y tripleto. Čia pažymėta $J_m := \{1, \dots, m\}$. Pastebėkime, kad $l_{i+1} = r_i$, $i = 1, \dots, m - 1$.

Darome prielaidą, kad tripletų rinkiniui (3.7) galioja 2-ame skyriuje suformuluotos sąlygos A1 ir A2, kurios šiuo konkrečiu atveju reiškia, kad $p(c|l, r) := \mathbf{P}\{Y_{2i} = c | Y_{2i-1} = l, Y_{2i+1} = r\} \forall i \in J_m$ ir $|\mathcal{A}| \times |\mathcal{A}| \times |\mathcal{A}|$ kokybinių kintamųjų L, C, R dažnių lentelė $\mathcal{L} = \mathcal{L}(y) := \{f_{lcr}, (l, c, r) \in \mathcal{A}^3\}$ su stebėtais tripleto (l, c, r) dažniais:

$$f_{lcr} = |\{i \in J_m : (L_i, C_i, R_i) = (l, c, r)\}| \quad (3.8)$$

atitinkamoje lentelės $\mathcal{L}$ ląstelėje yra pakankama statistika įvertinti sąlygines tikimybes $p = p(c|l, r)$, $(l, c, r) \in \mathcal{A}^3$. Sąlyginė tikėtinumo funkcija yra sandauga multinominių skirstinių:

$$L(p) = L(p|Y) = \prod_{i=1}^m p(C_i | L_i, R_i) = \prod_{(l, c, r) \in \mathcal{A}^3} p(c|l, r)^{f_{lcr}}.$$

Pilnas (*angl. saturated*) apibendrintas logit modelis šiuo atveju:

$$\log \left(\frac{p(c|l, r)}{p(a|l, r)} \right) = \beta_c^C + \beta_{lc}^{LC} + \beta_{cr}^{CR} + \beta_{lcr}^{LCR}, \quad l, c, r \in \mathcal{A} \setminus \{a\}, \quad (3.9)$$

čia β žymi nežinomus modelio parametrus.

Homogeninės pirmos eilės Markovo grandinės atveju L_i ir R_i yra sąlyginai nepriklausomi, kai yra žinomos C_i reikšmės, ir $\mathbf{P}\{C_i = u | L_i = v\} = \mathbf{P}\{R_i = u | C_i = v\} \forall u, v \in \mathcal{A}$, $i \in J_m$, todėl turėtų galioti nulinė hipotezė $H_0 = H_{M1}$:

$$H_{M1}: \quad \beta_{uvz}^{LCR} = 0, \beta_{uv}^{LC} = \beta_{uv}^{CR} \quad \forall u, v, z \in \mathcal{A} \setminus \{a\}. \quad (3.10)$$

Tačiau H_{M1} ir sąlygos A1, A2 nereiškia, kad stebima seka Y yra homogeninė pirmos eilės Markovo grandinė.

3. IŠSKLAIDYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS GLODINIMAS

65

Pasiūlyta struktūra, kuri remiasi sąlygomis A1, A2 ir (3.9) apima ir kai kuriuos specialius antros eilės Markovo modelius (M2). Jie yra aprašomi perėjimo tikimybėmis:

$$p_2(c, r | l) = \mathbf{P}\{C_i = c, R_i = r | L_i = l\}, \quad (l, c, r) \in \mathcal{A}^3, i \in J_m. \quad (3.11)$$

Jei sekos Y elementai yra laikomi nepriklausomais (tai – klasikinis tikimybinių modelis DNR sekoms), nulinė hipotezė $H_0 = H_I$ su

$$H_I: \quad \beta_{uvz}^{LCR} = 0, \beta_{uv}^{LC} = \beta_{uv}^{CR} = 0 \quad \forall u, v, z \in \mathcal{A} \setminus \{a\}. \quad (3.12)$$

Paprasčiausias modelis DNR sekų koduojančioms sritims remiasi prielaida, kad seka Y sudaryta iš nepriklausomų tripletų. Tačiau šis modelis negali būti aprašomas, naudojant pasiūlytą struktūrą. Jį žymėmsime H_T .

3.4.1. Savirankos algoritmas

Dėl DNR sekų heterogeniškumo (vidutinis nekoduojančių bakterijos sekų ilgis yra apie 200), net ir apsiribojant tik antros eilės Markovo modeliu (M2 modeliu), dažnių lentelės, kurios suformuojamos naudojant atitinkamus logitinesius ar apibendrintus logit modelius, yra išsklaidytos. Tačiau atlikus saviranką iš $\bar{q}_h$ su pakankamai dideliu h , kai generuotų replikacijų skaičius taip pat yra pakankamai didelis, gautos dažnių lentelės jau nėra išsklaidytos.

Išnaudojant apibendrinto logit modelio pavidalą (3.9) su nagrinėjama nulinė hipoteze H_0 , atitinkančia (nežinomų) parametrų specifikaciją, yra nesudėtinga suskaičiuoti sąlygines tikimybes $p(c|l, r)$, $(l, c, r) \in \mathcal{A}^3$, ir pagal jas generuoti homogeninę Markovo grandinę X , prasidedančią nuo y ir turinčią H_0 hipotezėje nurodytą stacionarų skirstinį. Markovo sekos X generavimui galima naudoti, pavyzdžiui, Gibso išrinkimą su randomizacija (*angl. Gibbs sampling with randomization*), t. y. su kiekviename žingsnyje atsitiktinai pasirenkamu modifikuojamos koordinatės numeriu. Tokia generavimo schema (*Glauberio dinamika*) aprašyta disertacijos 2-ame skyriuje.

Kadangi tripletų dažniai (3.8) mūsų pasirinktoje struktūroje sudaro pakankamą statistiką, yra natūralu imti juos kaip pagrindą suderinamumo hipotezės (3.5) statistikai apibrėžti. Kitos pasirinktos statistikos, naudojamos atliktoje baigtinių būsenų statistinėje analizėje yra m -rinkiniai (m -tuples)

$$t_w = \left| \{i \in J_{n-m+1} : (Y_i, \dots, Y_{i+m-1}) = w\} \right|, \quad w \in \mathcal{A}^m. \quad (3.13)$$

Praktiniuose skaičiavimuose "tikrosios" apibendrintos logit funkcijos (3.9) turi

būti pakeistos jų statistiniais įverčiais.

Teorinis pasiūlyto metodo tyrimas yra gana sudėtingas, kadangi statistikos $\eta(h)$ komponentės, būtent, seka y ir jos ilgis n , spektrinės perėjimo tikimybių matricos Π savybės, funkcionalo λ matavimas ir struktūra sudėtingai sąveikauja vienas su kitu. Taigi, standartiniai asimptotiniai ir aproksimavimo metodai tiesiogiai negali būti pritaikyti. Iš kitos pusės, aprašyta procedūra reikalauja daug kompiuterinių resursų ir tai apsunkina jos tyrimą kompiuterinio modelavimo būdu (Monte Karlo metodu).

Su tikslu ištirti procedūros veikimą buvo atliktas supaprastintas kompiuterinis eksperimentas, kurio rezultatai aptariami kitame skyriuje.

3.4.2. Kompiuterinis eksperimentas

Paprastumo dėlei imamas binarinis alfabetas $\mathcal{A} = \{0, 1\}$. Nulinė hipotezė H_0 tvirtina, kad Y yra stacionari pirmos eilės Markovo grandinė su perėjimo tikimybėmis $p_{01} = p_{10} = \rho$. Parametras $\rho \in (0, 1)$ pilnai nusako Y skirstinį ir yra laikomas žinomu (žemiau pateikiami rezultatai su $\rho = 0,6$ ir $\rho = 0,8$). Marginalinis Y -ko skirstinys (invariantis sekos Y_l , $l = 1, \dots, n$ skirstinys) yra $\text{Binomial}(1, 1/2)$, jis nepriklauso nuo ρ .

Uždavinys yra patikrinti paprastąją nulinę hipotezę H_0 su sudėtinga alternatyva H_1 , kuri yra priešinga teiginiui H_0 . Yra tiriama pradinės (nesuglodintos) statistikos $\eta(0)$ (3.5) ir suglodintos statistikos η^* (3.6) galia kai kurioms alternatyvoms: nepriklausomų nukleotidų alternatyva H_I (3.12), nepriklausomų tripletų seka H_T (3.10) ir antros eilės Markovo seka su perėjimo tikimybėmis (3.11) (toliau šis modelis bus vadinamas H_{M2}) ir kt.

Alternatyvos H_I atveju, $Y_i, i = 1, \dots, n$ yra nepriklausomi ir vienodai pasiskirstę atsitiktiniai dydžiai su $Y_i \sim \text{Binomial}(1, 1/2)$.

Alternatyva H_T tvirtina, kad Y yra rinkinys $n/3$ nepriklausomų tripletų, kai kiekvieno nepriklausomo tripleto $(l, c, r) \in \mathcal{A}^3$ tikimybės p_{lcr} sekoje Y abiejų hipotezių H_0 ir H_T atveju sutampa ir yra lygios $p_{lcr} = 0,5 \cdot \rho^{|l-c|+|c-r|} \cdot (1-\rho)^{2-|l-c|-|c-r|}$. Hipotezių H_0 ir H_T atskyrimo uždavinys yra pakankamai sudėtingas.

Alternatyvai H_{M2} yra naudojamos tokios tikimybių (3.11) reikšmės:

$$\begin{aligned} p_2(0, 0 | 0) &= 0,4 & p_2(1, 0 | 0) &= 0,5 \\ p_2(0, 1 | 0) &= 0,1 & p_2(1, 1 | 0) &= 0,1 \\ p_2(c, r | l) &= p_2(1-c, 1-r | 1-l), & (l, c, r) &\in \mathcal{A}^3. \end{aligned} \quad (3.14)$$

Dalis gautų rezultatų H_T ir H_{M2} alternatyvoms pateikti 3.1 ir 3.2 lentelėse.

3. IŠSKLAIDYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS GLODINIMAS

67

Markovo evoliucija $X(h)$, $h \in \mathbf{Z}_+$, prasidedanti būsenoje $X(0) = y$ ir turinti stacionarųjį skirstinį, nusakyta nuline hipoteze H_0 , apibrėžiama per sąlygines tikimybes (3.1 algoritmas):

$$p(c|l, r) = \frac{p_{lcr}}{p_{l0r} + p_{l1r}}, \quad (l, c, r) \in \mathcal{A}^3. \quad (3.15)$$

Svoriai 3.2 formulėje yra parinkti tokiu būdu: $\omega_h(t) = 1/(h+1)$, $t = 0, 1, \dots, h$ ir yra lygūs nuliui likusiais atvejais. Kol kas svorių parinkimo poveikis statistikų galiai buvo nežymi. Kovariacijų matrica $\widehat{W}$ (3.5) yra laikoma diagonaline ir yra įvertinama iš atskirų nepriklausomų imčių, generuotų pagal galiojančią nulinę hipotezę H_0 . Paprastai su šiuo pasirinkimu gaunami šiek tiek galingesni testai nei imant empirinę kovariacinę matricą $\widehat{W}$.

Trumpoms sekoms (tarkim su $n < 20$) tikslią testo galią (*angl. power of test*) galima suskaičiuoti tiesiogiai. Čia žodis „tiksliai“ reiškia, kad sąlyginiai savirankos (pakartotinių imčių) vidurkiai su pradine sąlyga $Y = y$, o taip pat (nesąlyginiai) savirankos (pakartotinių imčių) iš Y vidurkiai yra pakeičiami atitinkamais (sąlyginiais ir nesąlyginiais) teoriniais vidurkiais. Remiantis nukleotidų trejetų ir penketų rinkiniais (*angl. m-tuples*), rezultatai testo statistikoms $\eta(0)$ ir η^* su $n = 9$ yra apibrėži 3.13 formulėje ir apibendrinti 3.1 lentelėje.

Reiktų pastebėti, kad imant trejetų ir penketų rinkinius sekose, vidutinis ląstelių dažnis stebinių lentelėse atitinkamai lygus $7/8$ ir $5/32$. Vadinasi, dažnių lentelė šiuo atveju yra labai išsklaidyta.

Tiesioginis skaičiavimas ilgesnėms sekoms yra neįmanomas, ir testo galiai įvertinti naudojamas kompiuterinis modeliavimas. Kai kurie rezultatai $n = 21$ ilgio sekoms yra pateikiami 3.2 lentelėje. Kompiuteriniame modeliavime buvo generuojama 1000 nepriklausomų Y sekos replikacijų ir kiekvienai replikacijai atskirai generuojama 100 nuo jos prasidedančių evoliucijų. Formulėje (3.6) glodinimo parametro reikšmių rinkinys tiksliam skaičiavimui atlikti yra $\mathcal{H} = \{1, \dots, 50\}$ ir $\mathcal{H} = \{1, \dots, 10, 20, 40, 160\}$ – modeliavimo atveju.

Gauti rezultatai rodo, kad glodinis reikšmingai padidina galią tuo atveju, kai turime alternatyvą H_{M2} su $\rho = 0,6$, ypatingai tada, kai imami penketukai sekoje. Kai $\rho = 0,8$ glodinimo efektas yra nežymus visais atvejais. Tai reiškia, kad arba net vienas glodinimo žingsnis su perėjimo tikimybėmis (3.15) per daug suglodina duomenis arba gal seka $\eta(h)$, $h \in \mathbf{Z}_+$, nėra stacionari. Didesniam reikšmingo lygmeniui ($\alpha \geq 0,05$) paprastai abi statistikos $\eta(0)$ ir η^* , naudojančios penketukų rinkinių statistiką $\hat{\lambda}$, yra galingesnės negu naudojančios trejetukų rinkinių statistiką (3.2 lentelė).

3. IŠSKLAIDYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS GLODINIMAS

68

3.1 lentelė. Tiksli testo galia, kai $n = 9$

Table 3.1. The exact power of the tests, $n = 9$

Reikšming. lygmuo α			0,01	0,05	0,1	0,2	0,01	0,05	0,1	0,2
ρ	H_1	Testas	3-jetai				5-etai			
0,6	H_{NT}	$\eta(0)$	0,016	0,057	0,117	0,224	0,021	0,070	0,121	0,260
		η^*	0,016	0,051	0,113	0,217	0,016	0,070	0,136	0,252
0,8	H_{NT}	$\eta(0)$	0,032	0,114	0,268	0,395	0,033	0,141	0,209	0,443
		η^*	0,033	0,105	0,268	0,396	0,033	0,145	0,215	0,443
0,6	H_{M2}	$\eta(0)$	0,052	0,136	0,197	0,369	0,054	0,110	0,158	0,232
		η^*	0,060	0,149	0,263	0,434	0,048	0,146	0,218	0,375
0,8	H_{M2}	$\eta(0)$	0,241	0,419	0,480	0,541	0,254	0,357	0,494	0,621
		η^*	0,252	0,426	0,480	0,574	0,254	0,379	0,496	0,644

3.2 lentelė. Testo galia, gauta kompiuteriniu modeliavimu, kai $n = 21$

Table 3.2. The power of the tests, simulation results for $n = 21$

Reikšmingumo lygmuo α			0,01	0,05	0,1	0,01	0,05	0,1
ρ	Testas	m-motyvai	$H_1 = H_{NT}$			$H_1 = H_{M2}$		
0,8	$\eta(0)$	3-jetai	0,079	0,227	0,325	0,437	0,612	0,681
		5-etai	0,063	0,184	0,348	0,453	0,648	0,728
0,8	η^*	3-jetai	0,082	0,175	0,309	0,431	0,565	0,654
		5-etai	0,058	0,184	0,343	0,448	0,654	0,730

Akivaizdi išimtis yra jau minėtas atvejis su alternatyva H_{M2} ir $\rho = 0,6$. Alternatyvai H_I glodinimo efektas ir m -rinkinių eilės m įtaka testo galiai yra panaši kaip ir alternatyvos H_T , bet galia yra ženkliai didesnė (kai $\alpha = 0,05$, galia yra didesnė už 0,5).

3.5. Išvados pasiūlytam semiparametriniam glodinimo metodui

Pasiūlytas semiparametrinio glodinimo ir savirankos metodas gali būti taikomas tolydiems duomenims kaip parodyta 2 pavyzdyje, bet jis ypatingai tinkamas išsklaidytiems kategoriniams duomenims, kadangi tai suteikia natūralų būdą glodinimo duomenims apibrėžti branduolį.

Metodas yra pakankamai paprastas, bet jo efektyvus įgyvendinimas – gana

3. IŠSKLAIDYTŲ KATEGORINIŲ DUOMENŲ SEMIPARAMETRINIS GLODINIMAS

69

sudėtinga problema. Tikslus tiesioginis skaičiavimas yra įmanomas tik mažo matavimo duomenims. Tuo tarpu modeliavimas reikalauja ne mažų kompiuterinių resursų ir laiko net vidutinėms duomenų dimensijoms.

Išskyrus tą atvejį kai turime branduolio glodinimą tolydiems dydžiams, glodinimo parametras šiuo atveju yra diskretus. Glodinimo efektas kiekvienam glodinimo žingsniui artimai priklauso nuo spektrinės Markovo branduolio Π struktūros, funkcionalo λ ir jų dimensijos. Todėl yra sudėtinga taikyti standartinius asimptotinius metodus ir aproksimacijas.

Kompiuterinis modeliavimas rodo, kad tam tikrais atvejais semiparametrinio glodinimo efektas testo galiai yra nežymus, nors paprastai didinant m -rinkinių eilę, jis padidėja. Manoma, kad tai įtakoja eksperimente naudojami gerokai trumpi binarinių sekų ilgiai, diskretus glodinimo parametro pobūdis ir gal būt ne stacionarus statistikų $\eta(h)$, $h \in \mathbf{Z}_+$ elgesys.

Yra tikimasi, kad semiparametrinis glodininimas ir saviranka yra statistiškai labiau efektyvus didelės dimensijos kategoriniams duomenims su artima alternatyva. Kitaip tariant – kai diskretūs duomenys tam tikra prasme gali būti traktuojami kaip tolydūs.

Ši problema yra tolimesnių tyrimų objektas.

Bendrosios išvados

1. Disertacijoje pasiūlytas semiparametrinio glodinimo metodas, kuris gali būti taikomas kategorinių požymių glodinimui ir leidžia pritaikyti neparametrinį (semiparametrinį) savirankos metodą išsklaidytų duomenų dažnių lentelėms. Semiparametriniam glodinimui realizuoti disertacijoje pasiūlyta apibendrinto logit modelio struktūra.
2. Apibendrintas logit modelis, pritaikytas DNR duomenims su darbe įvesta specialia duomenų struktūra, yra suderintas su DNR sekų kontekstiniu evoliucijos modeliu.
3. Atliktoje DNR sekų statistinėje analizėje parodyta, kad DNR sekos yra nehomogeniškos ir tai gali įtakoti jų ne-Markoviškumą. Iš kitos pusės, buvo nustatyta, kad reikšminga dalis (apie 60%) tirtų ne-Markovo sekų yra homogeniškos. Taigi, net homogeniški DNR fragmentai negali būti laikomi pirmos eilės Markovo sekomis. Vadinasi, įprasti DNR sekų tikimybiniai modeliai, paslėptosios Markovo grandinės ir k -tosios eilės homogeninės Markovo grandinės, nėra tinkami *visos* DNR sekos analizei.
4. Atlikta DNR sekų statistinė analizė parodė, kad įvertinti aukštesnės eilės Markovo modelį dėl DNR nekoduojančių sekų trumpumo (vidutinis sekų ilgis apie 200 *bp*) ir nehomogeniškumo yra sudėtinga, nes

modelio parametrų skaičius, didėjant Markovo eilei m , auga eksponentiniu greičiu.

5. Panaudojant DNR vijų lokalaus simetriškumo charakterizaciją logit funkcijos terminais nustatyta, kad tirtuose bakterijų genomuose yra ženkliai nekoduojančių fragmentų, kurie nėra 1-os eilės lokaliai simetriški, dalis (13 fragmentų iš 17). Praktiškai tai reiškia, kad sekų vijos gali turėti skirtingą biologinę prasmę.

Literatūros sąrašas

- Agresti, A. 1990. *Categorical data analysis*. Wiley & Sons, New York.
- Agresti, A. 2007. *An introduction to categorical data analysis*. Wiley & Sons, New York, second edition.
- Agresti, A.; Hitchcock, B. 2005. Bayes inference for categorical data analysis, *Statistical methods and applications* 14: 297–330.
- Avery, P. 2002. Fitting interconnected Markov chain models-DNA sequences and test cricket matches, *The Statistician* 51(2): 267–278.
- Avery, P.; Henderson, D. 1999. Fitting Markov chain models to discrete state series such as DNA sequences, *Journal of the royal statistical society: Series C (Applied Statistics)* 48(1): 53–61.
- Bagdonavičius, V.; Kruopis, J. 2007. *Matematinė statistika*. Vilnius.
- Baisnee, P.; Hampson, S.; Baldi, P. 2002. Why are complementary DNA strands symmetric?, *Bioinformatics* 18(8): 1021–1033.
- Bansal, M. 2003. DNA structure: Revisiting the Watson-Crick double helix, *Current science* 85(11): 1556–1563.
- Bergsma, W. 1997. *Marginal models for categorical data*. Tilburg University Press, Tilburg.
- Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the royal statistical society: series B (Statistical methodology)*

36: 192–236.

Buhlmann, P.; Wyner, A. 1999. Variable length markov chains, *The annals of statistics* 27: 480–513.

Burman, P. 2004. On some testing problems for sparse contingency tables, *Journal of multivariate analysis* 88(1): 1–18.

Charles, S. D. 1993. The computer generation of multinomial random varieties, *Computational statistics & Data analysis* 16: 205–217.

Marhoul, J. C. 1985. *A model for large sparse contingency tables*. PhD thesis, Stanford University, California.

Ching, W.; Fung, S. E.; Michael, K. 2004. Higher-Order Markov chain models for categorical data sequences, *Naval Research Logistics (NRL)* 51(4): 557–574.

Ching, W.; Michael, K.; Zhang, S. 2005. On computation with higher-order Markov chains, *Current trends in high performance computing and its applications* 1: 15–24.

Christensen, R. 1990. *Log-linear models*. New York.

Cochran, W. 1952. The χ^2 Test of goodness-of-fit, *Annals of mathematical statistics* 23: 315–345.

Congdon, P. 2005. *Bayesian models for categorical data*. Wiley & Sons, New York.

Coull, B.; Agresti, A. 2003. Generalized log-linear models with random effects, with application to smoothing contingency tables, *Statistical modelling* 3: 251–271.

Cressie, N.; Read, T. 1984. Multinomial goodness of fit tests, *Journal of the royal statistical society, series B* 46: 440–464.

Csiszár, I.; Paul, C. S. 2000. The consistency of the BIC Markov order estimator, *The annals of statistics* 28(6): 1601–1619.

Csiszár, I.; Talata, Z. 2006. Consistent estimation of the basic neighbourhood of Markov random fields, *The annals of statistics* 34(1): 123–145.

Csiszár, I.; Talata, Z. 2006. Context tree estimation for not necessarily finite memory processes, via BIC and MDL, *IEEE Transactions on information theory* 52: 107–1016.

Čekanavičius, V.; Murauskas, G. 2001. *Statistika ir jos taikymai*. I ir II dalys. Vilnius.

Dalevi, D.; Dubhashi, D.; Hermansson, M. 2006. A new order estimator for fixed and variable length Markov models with applications to DNA sequence similarity, *Statistical applications in genetics and molecular biology* 5, Article 8.

Darroch, J.; Lauritzen, S.; Speed, T. P. 1980. Markov fields and log-linear interaction models for contingency tables, *The annals of statistics* 8(3): 522–539.

Dalevi D.; Dubhashi, D.; Hermansson, M. 2006. Bayesian classifiers for detecting

- HGT using fixed and variable order Markov models of genomic signatures, *Bioinformatics* 22(5): 517–522.
- Eggar, M. 2002. Validity of fitting a first order Markov chain model, *The Statistician* 51(2): 259–265.
- Faddy, M.; Jones, M. 1998. Semiparametric smoothing for discrete data, *Biometrika* 85: 131–138.
- Fienberg, S. 2000. Contingency tables and log-linear models: basic results and new developments, *Statistics in the social sciences* 95(450): 1371–1383.
- Gordon, S.; Yi-Ching, Y.; Glenn, M. 2005. Global Markov models for eukaryote nucleotide data, *Journal of statistical planning and inference* 130: 251–275.
- Gusfield, D. 1997. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press.
- Hu, M. 1999. *Model checking for incomplete high dimensional categorical data*. PhD thesis, University of California, Los Angeles.
- Jensen, J. 2005. Context dependent DNA evolutionary models, *Paper research reports* 458, *Department of Mathematical Sciences*. University of Aarhus.
- Kim, S.; Choi, H.; Lee, S. 2009. Estimate-based goodness-of-fit test for large sparse multinomial distributions, *Computational statistics & data analysis* 53(4): 1122–1131.
- Koehler, K. 1986. Goodness-of-fit tests for log-linear models in sparse contingency tables, *Journal of the American statistical association* 81: 483–493.
- Koehly, L.; Goodreau, S.; Morris, M. 2003. The link between exponential random graph models and loglinear models for networks, *CSDE working paper 03-05, Center for studies in demography and Ecology*.
- Kuss, O. 2002. Global goodness-of-fit tests in logistic regression with sparse Data, *Statistics in medicine* 21: 3789–3801.
- Lawal, B.; Upton, J. 1980. An Approximation of the distribution of χ^2 test of goodness-of-fit statistics for use with small expectations, *Biometrika* 67(2):447–453.
- Loehlin, J. 1997. *Latent variable models: an introduction to factor, path, and structural analysis*. New York: Erlbaum.
- Morvai, G.; Weiss, B. 2005. Order estimation of Markov chains, *IEEE Transactions on information theory* 51: 1496–1497.
- Mrazek, J.; Karlin, S. 1998. Strand compositional asymmetry in bacterial and large viral genomes, in *Proceedings of the national academy of sciences* 95(7): 3720–3725.
- Mueller, R. 1996. *Basic principles of structural equation modeling*. Springer-Verlag, New York.

- Muller, U.; Osius, G. 2003. Asymptotic normality of goodness-of-fit statistics for sparse Poisson data, *Statistics* 37(2): 119–143.
- Murphy, K. 2001. *An introduction to graphical models*.
- Niu, D.; Lin, K.; Zhang, D. 2003. Strand compositional asymmetries of nuclear DNA in eukaryotes, *Journal of molecular biology* 57: 325–334.
- Peres, Y.; Shields, P. 2005. Two new Markov order estimators, *arXiv:math/0506080v1*.
- Qi, D.; Cuticchia, A. 2001. Compositional symmetries in complete genomes, *Bioinformatics* 17: 557–559.
- Raftery, A. 1985. A model for high-order Markov chains, *Journal of the royal statistical society: series B* 47: 528–539.
- Reinert, G.; Schbath, S.; Waterman M. S. 2000. Probabilistic and statistical properties of words, *Journal of computational biology* 7(1/2): 1–46.
- Kindermann R.; Snell J. L. 2000. *Markov random fields and their applications*. American mathematical society.
- Santer, T.; Duffy, D. 1989. *The Statistical analysis of discrete data*. New York.
- Sengupta, D.; Jammalamadaka, S. R. 2003. *Linear models an integrated approach*. World scientific.
- Silverman, B. 1986. *Density estimation for statistics and data Analysis*. Chapman & Hall, London.
- Simonoff, J. 1995. Smoothing categorical data, *Journal of statistical planning and inference* 47: 41–69.
- Simonoff, J. 1996. *Smoothing methods in statistics*. SAS Institute and Wiley.
- Simonoff, J. 1998. Three sides of smoothing: categorical data smoothing, nonparametric regression, and density estimation, *International statistical review* 66(2): 137–156.
- Stevens, S. 1946. On the theory of scales of measurement. *Science* 103(2684): 677–680.
- Stokes, M.; Davis, C.; Koch, G. 2001. *Categorical data analysis using the SAS(R) System*. SAS Institute and Wiley, North Carolina, 2nd edition.
- Streliaoff, C.; Crutchfield, J.; Hubler, A. 2007. Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling, *arXiv:math.ST/0703715*.
- Timm, N. 2002. *Applied multivariate analysis*. Springer, New York.

Rekašius, T. 2007. *Evoliucinis neinformatyvių genetinių sekų modelis*. PhD thesis, Vilniaus Gedimino technikos universitetas, Vilnius.

Usatenko, O.; Yampolskii, V. 2003. Binary N-step Markov chains and long-range correlated systems, *Physical review letters* 90(11): 110601.

Usotskaya, N.; Ryabko, B. 2008. Application of information-theoretic tests for the analysis of DNA sequences based on Markov chain models, *Computational statistics & data Analysis* 53(5): 1861–1872.

von Davier, M. 1997. Bootstrapping goodness-of-fit statistics for sparse categorical data – results of a Monte Carlo study, *Methods of psychological research online* 2.

Weinberger, M.; Rissanen, J.; Feder, M. 1995. A universal finite memory source, *IEEE Transactions on information theory* 41: 643–652.

Yarnold, J. 1970. *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions*. Springer-Verlag, New York, 2nd edition.

Yu, Z.; Anh, V.; Lau, K. 2001. Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome, *Physica* 301(1-4): 351–361.

Zhao, L.; Dorea, C.; Gonçaves, C. 2001. On determination of the order of a Markov chain, *Statistical inference for stochastic processes* 4: 273–282.

Autoriaus publikacijos disertacijos tema

Recenzuojamuose mokslo leidiniuose

Radavičius, M.; Židanavičiūtė, J. 2009. Semiparametric smoothing of sparse contingency tables, *Journal of statistical planning and inference* 139(11): 3900–3907 (Thomson ISI Web of Science).

Židanavičiūtė, J. 2008. Logit analysis of genetic Data, *Mathematical modelling and analysis* 13(1): 135–144 (Thomson ISI Web of Science).

Radavičius, M.; Židanavičiūtė, J. 2007. Model testing for high-dimensional contingency tables with application in genetics, in *Computer data analysis and modelling: complex stochastic data and systems: proc. of the eighth intern. conf., Minsk, September 11–15, 2007* 1: 269–272.

Radavičius, M.; Židanavičiūtė, J.; Rekašius, T. 2007. Kai kurie matematikos uždaviniai genetikoje, *Lietuvos matematikos rinkinys* 47(spec. nr.): 21–28.

Židanavičiūtė, J.; Rekašius, T. 2006. Genetinių sekų markoviškumo tyrimas, *Lietuvos matematikos rinkinys* 46(spec. nr.): 280–285.

Radavičius, M.; Židanavičiūtė, J. 2005. Statistinė struktūrų analizė: kai kurios jos taikymo problemos, *Lietuvos matematikos rinkinys* 45(spec. nr.): 354–362.

Jurgita ŽIDANAVIČIŪTĖ

KATEGORINIŲ POŽYMIŲ
PRIKLAUSOMYBIŲ STRUKTŪROS
STATISTINĖ ANALIZĖ
IR JOS TAIKYMAS GENETIKOJE

Daktaro disertacija
Fiziniai mokslai, matematika (01P)

DEPENDENCE STRUCTURE ANALYSIS
OF CATEGORICAL VARIABLES
WITH APPLICATIONS IN GENETICS

Doctoral Dissertation
Physical Sciences, Mathematics (01P)

2010 01 05. 7,5 sp. l. Tiražas 20 egz.
Vilniaus Gedimino technikos universiteto leidykla „Technika“,
Saulėtekio al. 11, LT-10223 Vilnius, <http://leidykla.vgtu.lt>
Spausdino UAB „Baltijos kopija“,
Kareivių g. 13B, LT-09109 Vilnius, <http://www.kopija.lt>