VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Jurgita ŽIDANAVIČIŪTĖ

DEPENDENCE STRUCTURE ANALYSIS OF CATEGORICAL VARIABLES WITH APPLICATIONS IN GENETICS

SUMMARY OF DOCTORAL DISSERTATION
PHYSICAL SCIENCES,
MATHEMATICS (01P)

Vilnius LEIDYKLA TECHNIKA 2010

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

The dissertation is defended as an external work.

Scientific Consultant

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

The dissertation is being defended at the Council of Scientific Field of Mathematics at Vilnius Gediminas Technical University:

Chairman

Prof Dr Habil Leonas SAULIS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P).

Members:

Prof Dr Habil Mindaugas BLOZNELIS (Vilnius University, Physical Sciences, Mathematics – 01P),

Assoc Prof Dr Rimantas EIDUKEVIČIUS (Vilnius University, Physical Sciences, Mathematics – 01P),

Prof Dr Habil Kęstutis KUBILIUS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P),

Prof Dr Habil Rimantas RUDZKIS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P).

Opponents:

Prof Dr Habil Kęstutis DUČINSKAS (Klaipėda University, Physical Sciences, Mathematics – 01P),

Assoc Prof Dr Mečislavas MEILŪNAS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P).

The dissertation will be defended at the public meeting of the Council of Scientific Field of Mathematics in the Senate Hall of Vilnius Gediminas Technical University at 10 a. m. on 12 February 2010.

Adress: Saulėtekio al. 11, LT-10223 Vilnius, Lietuva.
Tel. +370 5 274 4952, +370 5 274 4956; fax +370 5 270 0112;
e-mail: doktor@vgtu.lt

The summary of the doctoral dissertation was distributed on 11 January 2010.

A copy of the doctoral dissertation is available for review at the Libraries of Vilnius Gediminas Technical University (Saulėtekio al. 14, LT-10223 Vilnius, Lithuania) and the Institute of Mathematics and Informatics (Akademijos 4, LT-08663 Vilnius, Lithuania).

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Jurgita ŽIDANAVIČIŪTĖ

KATEGORINIŲ POŽYMIŲ PRIKLAUSOMYBIŲ STRUKTŪROS STATISTINĖ ANALIZĖ IR JOS TAIKYMAS GENETIKOJE

DAKTARO DISERTACIJOS SANTRAUKA
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).
Disertacija ginama Vilniaus Gedimino technikos universiteto Matematikos mokslo krypties taryboje:
Pirmininkas
prof. habil. dr. Leonas SAULIS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).
Nariai:
prof. habil. dr. Mindaugas BLOZNELIS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),
doc. dr. Rimantas EIDUKEVIČIUS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),
prof. habil. dr. Kęstutis KUBILIUS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P),
prof. habil. dr. Rimantas RUDZKIS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P).
Oponentai:
prof. habil. dr. Kęstutis DUČINSKAS (Klaipėdos universitetas, fiziniai mokslai, matematika – 01P),
doc. dr. Mečislavas MEILŪNAS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

Disertacija bus ginama viešame Matematikos mokslo krypties tarybos posėdyje 2010 m. vasario 12 d. 10 val. Vilniaus Gedimino technikos universiteto senato posėdžių salėje.
Adresas: Saulėtekio al. 11, LT-10223 Vilnius, Lietuva.
Tel.: (8 5) 274 4952, (8 5) 274 4956; faksas (8 5) 270 0112;
el. paštas doktor@vgtu.lt
Disertacijos santrauka išsiuntinėta 2010 m. sausio 11 d.
Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto (Saulėtekio al. 14, LT-10223 Vilnius, Lietuva) ir Matematikos ir informatikos instituto (Akademijos g. 4, LT-08663 Vilnius, Lietuva) bibliotekose.
VGTU leidyklos „Technika“ 1709-M mokslo literatūros knyga.

Introduction

Scientific problem and topicality of the work

Sometimes the inclusion of additional variables in a statistical analysis changes previous conclusions completely. The Simpson paradox demonstrates this situation in case of the statistical analysis of categorical variables. Consequently, all variables known to an investigator should be included into this kind of analysis. Currently the amount of information accessible to a statistician is very extensive, therefore problems related to a large dimension and/or sparsity of data arise rather frequently. The problem of sparsity is especially topical for categorical data. Models for quantitative (continuous) variables describe the relationship structure between the means of these variables. In this case the number of model parameters linearly depends on n, the dimension of the data. When a sophisticated covariance structure of data is modelled, the number of unknown parameters increases as $O(n^2)$. For the models of qualitative (categorical) variables, the number of unknown parameters increases exponentially with respect to the data dimension n. Consequently, even for a moderate number of categorical variables, a corresponding contingency table can be sparse, i. e. many cells in the table are empty or have small counts. Sometimes the number of unknown parameters (the number of cells) is even greater than the sample size (very sparse categorical data).

Traditionally, expected (under the null hypothesis) frequencies are required to exceed 5 in (almost) all cells of the contingency table. If this condition is violated, the χ^2 approximations of goodness-of-fit statistics may be inaccurate. For very sparse data, classical goodness-of-fit statistics becomes simply uninformative. Another problem in the categorical data analysis is caused by a statistical dependence of observations. The dependence of observations is typical when dealing with genetic sequences, in the text and image processing.

Statistical inference problems caused by sparsity of contingency tables are widely discussed in the literature. They all are not applicable or have some limitations in case of dependent categorical observations and very sparse contingency tables containing a large portion of zeros. Counted frequencies of long words (tuples) in DNA (Deoxyribonucleic acid) sequences are typical examples of such data. Therefore two applied problems in genetics are considered in the dissertation. The first one is the estimation of the Markov order of DNA sequences. The problem is closely related to the context-dependent evolutionary model of DNA, where the mutation probability of each nucleotide depends on m its nearest-neighbouring nucleotides from each side (context). It is shown (Jensen 2002) that under the assumption that the DNA evolution is reversible, this implies that the stationary distribution of the evolution is a homogeneous Markov field of order m. Thus, the estimation of the Markov order is, in fact, the estimation of the order of a context dependence in the DNA evolution.

The second problem is testing of a higher order symmetry of two complementary strands of the DNA helix. The first order DNA strand symmetry is sometimes referred to as the Chargraff's second parity rule (Rudner *et al* 1968). Since the second strand is read in the reverse order, the extension of this first-order similarity to higher-orders is called a reverse-complement symmetry or simply a strand symmetry. Scientific literature falls short of providing a comprehensive study of the strand symmetry of a higher order. It also lacks a formal definition and characterization of this symmetry, as well as a convincing explanation or clarification of its origin (Baisnée *et al* 2002; Zhang and Huang 2008).

Research object

The main research object of the dissertation are sparse contingency tables obtained from possibly dependent multivariate categorical observations. Generalized logit model is used for estimating dependence structure between categorical variables. In applications non-coding sequences of bacterial genomes from GenBank database are investigated. It is supposed that non-coding regions of primitive organisms are less important biologically and thus can be thought of as genetic noise.

The aim and tasks of the work

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.

We consider the following problems:

1. To propose a method for adaptive smoothing of sparse categorical data and to apply it to hypothesis testing.
2. To investigate properties of the proposed smoothing and testing procedures and their applicability to genetic sequence analysis.
3. To propose a model suitable for the statistical analysis of DNA sequences and consistent with the context-dependent model of DNA evolution.
4. To assess the Markov property and to test hypothesis about symmetry of complementary DNA strands in non-coding regions of bacterial genomes.

Research methods

The following research methods are used in the dissertation: loglinear models for categorical variables, theory of discrete Markov field, resampling methods and computer simulations.

Scientific novelty

In the dissertation a method of the statistical inference for dependent multivariate categorical observations and sparse contingency tables is proposed. The method is based on the generalized logit model linked up with a specially introduced regression-like data structure, semi-parametric smoothing and corresponding re-sampling technique (semi-parametric or smoothed bootstrap).

Although the semi-parametric smoothing and re-sampling can be applied to any data, it is especially suitable for sparse categorical data, since it provides a natural way to define the kernel for data smoothing. Alternative smoothing procedures for categorical data by making use of the nearest neighbours with respect to some similarity measure are artificial, depend on appropriateness of the similarity measure, and are difficult to interpret.

The semi-parametric re-sampling unlike a non-parametric bootstrap is applicable for very sparse contingency tables containing a large portion of zeros. When comparing it to the parametric bootstrap, it is less sensitive to a parametric model (mis)specification.

As has already been mentioned, the investigation of the Markov property of DNA sequences is based on a new framework for the categorical data analysis. The introduced special structure of genetic data ensures that classical assumptions of the generalized logit model do hold and standard techniques can be applied for the statistical inference. It is worthwhile noticing that the generalized logit model is consistent with the context-dependent model of the genome evolution. In the study it is suggested that non-Markovity of DNA sequences can be caused by their in-homogeneity. A special homogeneity test shows that a significant part of the longest non-coding regions of bacterial genoms which are established to be non-Markovian are non-homogeneous as well. Thus, the results of the Markov order estimation for DNA sequences can be misleading. On the other hand, it is found that even homogeneous DNA regions cannot be treated as the first order Markov chains.

Most papers devoted to the strand symmetry are published by researchers in biology, genetics and bioinformatics. An important exception is the paper by Simons *et al* (2005). The paper proposes a *global* probabilistic model of the strand symmetry and investigates its goodness-of-fit empirically. The model, however, is very general and cannot be identified for a single genome sequence. In the dissertation a probabilistic model of *local* the strand symmetry is introduced and its characterization via a generalized logit is presented. The local strand symmetry is related to a context-dependent evolutionary model and can be interpreted as the equality of evolutionary conditions for both strands in the DNA helix. Preliminary results of testing of the local symmetry of non-coding regions in bacterial genoms shows that long composite sequences tend to be more symmetric than their separate parts.

Practical value of the work results

The proposed smoothing method can be used in any type of the statistical analysis, where it is necessary to assess dependence structure between multivariate categorical variables in case when data are sparse. The results of the genetic applications can be of interest for researches in genetics and biology and in investigations of DNA evolutionary models.

Statements presented for defence

1. The method of semi-parametric smoothing is flexible and universal. Supplemented with an appropriate data-driven smoothing criteria, it can be applied to any statistical regularization problem. Although semi-parametric smoothing can be useful for continuous data, it is especially suitable for the smoothing of sparse categorical data.
2. For the hypothesis testing, the non-parametric (semi-parametric) bootstrap can be applied even to (very) sparse categorical data by making use of semiparametric smoothing.
3. The generalized logit model linked up with a specially introduced regression-like data structure allows the usage of standard statistical methods (software) and provides a framework for a direct implementation of semiparametric smoothing and bootstrap.
4. For non-coding regions of some bacterial genoms DNA sequences are rather inhomogeneous and this can lead to their non-Markovity. On the other hand, even homogeneous DNA regions cannot be treated as the first order Markov chain.
5. For some bacterial DNA sequences, the first order local strand symmetry generally does not hold, but the asymmetry has a different manifestation in different DNA regions. It means that DNA strands can have different evolutionary conditions and a biological sense.

Approval of the work results

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 six international conferences.

The scope of the scientific work

The dissertation consists of introduction, three chapters and conclusions. The total scope of the dissertation – 80 pages, 1 picture, 8 tables. The work cites 70 references. The dissertation is written in Lithuanian.

1. The mathematical models used in statistical analysis of categorical variables

Categorical data arise from different sampling frameworks. The goal of statistical analysis is to find a dependence structure between a set of categorical variables. In the first chapter of the dissertation various models available for describing the nature of the association between categorical variables are introduced: Generalized linear models, Loglinear models, Binary logistic regression, Generalized logit and their link with Markov field theory and Gibbs distribution.

2. Logit analysis of genetic data

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 DNA.

DNA sequences are long sequences of nucleotides. At each position it has one of the nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). The binding between the bases is through hydrogen bonds: two between A and T and three between C and G. Thus, a DNA sequence can be viewed as a sequence of categorical random variables taking their values from a finite alphabet $\mathcal{A}$ with four letters $\{A, T, C, G\}$. Typical models for DNA sequences are homogeneous m-order Markov chains (models Mm) on a finite alphabet (state space) $\mathcal{A}$.

Let $\mathcal{A}$ be a finite alphabet with $|\mathcal{A}| = \mathrm{card}\mathcal{A}$ elements and $N := \{1, \ldots, n\}$. In the case of DNA sequences $|\mathcal{A}| = 4$ and $\mathcal{A} = \{A, C, G, T\}$. Fix some positive integer $m < n/2$ and define the "interior" N° and "boundary" ∂N of N: $N^\circ := (m+1, \ldots, n-m), \quad \partial N := N \setminus N^\circ$ and a collection of neighbourhoods $U(\ell) = U_m(\ell) := [\ell - m, \ell + m] \setminus \{\ell\}, \quad \ell \in N^\circ$. Here $[i, j] := (i, i+1, \ldots, j),\ i < j,\ i, j \in N$, is an interval of integers. Given $x \in \mathcal{A}^n$ and a set of indices $I \subset N$, let $x_I := (x_i, i \in I)$ denote the corresponding subsequence of x. Suppose that values of the random sequence x are fixed on the boundary ∂N, i. e. $x_{\partial N} = c_{\partial N}$ for some $c \in \mathcal{A}^n$, and set $\mathcal{X}_+ := \{a \in \mathcal{A}^n \colon\ a_{\partial N} = c_{\partial N}\}$.

Definition 2.1. The random sequence $x \in \mathcal{X}_+$ is a homogeneous Markov random field of order m (denoted by Mm) if for each $\ell \in N^\circ$ and each $a \in \mathcal{A}^n$

$$\mathbf{P}\{x_\ell = a_\ell | x_j = a_j, j \neq l\} = \mathbf{P}\{x_\ell = a_\ell | x_{U_m(\ell)} = a_{U_m(\ell)}\}, \qquad (1)$$

where the conditional probabilities in (1) are independent of $\ell \in N^\circ$. Denote them by $p(a_\ell | a_{U_m(\ell)})$.

Definition 2.2. *Conditional odds* $O_{y|b}(z)$ of y versus b given the values $z = x_{U_m(\ell)}$ of the m nearest neighbours $U_m(\ell)$ of $\ell \in N^\circ$, for some reference value $b \in \mathcal{A}$ with $p(b|z) > 0$, is the ratio

$$O_{y|b} = Q_{y|b}(z) := \frac{p(y|z)}{p(b|z)}, \quad y \in \mathcal{A},\ z \in \mathcal{A}^{2m}, \tag{2}$$

where the probabilities $p(y|z)$ are introduced in (1).

Brook's lemma (Brook 1964) and *Hamersley-Clifford theorem* (Besag 1974) imply the following proposition.

Proposition 2.1. If $\mathbf{P}\{x = a\} > 0$ for all $a \in \mathcal{X}_+$, the (conditional) distribution of the homogeneous Markov random field Mm is uniquely determined by the conditional odds $O_{y|b}(z),\ y \in \mathcal{A},\ z \in \mathcal{A}^{2m}$, for some reference value $b \in \mathcal{A}$, and there exists a function $\lambda_m \colon \mathcal{A}^{m+1} \to \mathbf{R}$ such that for each $a = (a_1, \dots, a_{2m+1}) \in \mathcal{A}^{2m+1}$

$$\log\Big(Q_{a_{m+1}|b}(a_{U_m(m+1)})\Big) = \sum_{j=1}^{m+1} \Big[\lambda_m\Big(a_{[j,m+j]}\Big) - \lambda_m\Big(a^{(b)}_{[j,m+j]}\Big)\Big], \tag{3}$$

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

This means that a way to identify the model Mm is to determine its conditional odds (2).

Let us introduce the following structure of the observed random sequence $x \in \mathcal{A}^n$ of the length $n = n_m \cdot (m+1) + m$, the quantity n_m being an integer. Set

$$X := \{(y_\ell, z_\ell), l \in S\}, \quad S := \{m+1, 2(m+1), \dots, n-m\}, \tag{4}$$

where $y_\ell := x_{(m+1)\ell}$ is a target variable and $z_\ell = x_{U_m(\ell)}$ is a vector of explanatory variables, $\ell \in S$.

Assume: (A1) $\{y_\ell,\ \ell \in S\}$ are conditionally independent given $\{z_j,\ j \in S\}$. (A2) The conditional probabilities $\mathbf{P}\{y_\ell = a | z_j,\ j \in S\},\ a \in \mathcal{A}$, do not depend on the position $\ell \in S$.

Remark 2.1. Note that these assumptions are fulfilled if x is generated by a homogeneous Markov field (chain) of the order m. The Markov property implies the additional conditions on odds (2). Namely, as (3) shows the odds depend on $y = x_\ell$ and $z = x_{U_m(\ell)}$ only through interactions $x_{I_j},\ j = 0, \dots, m$, where $I_j = U_m(\ell) \cap [\ell - m + j, \ell + j]$. This gives a basis for testing Markovity and selection of the Markov order.

Remark 2.2. Assumptions (A1) and (A2) ensure that common conditions of the generalized logit model are satisfied. The generalized logit is a regression-type model, i. e. a *conditional* model with given values of the explanatory variables z. This means that the probabilistic model of the explanatory variables $\{z_\ell,\ l \in S\}$ is not specified and can be treated as a nuisance nonparametric component of the model. This also means that the conditions mentioned in Remark 2.1 are necessary but not sufficient for the Markov property.

Let us consider the case $m = 1$ and let v, c, w denote, respectively, the left, center, right element of a triplet (v, c, w) and let $\{(v_l, c_l, w_l), l \in S\} := \{(x_{2l-1}, x_{2l}, x_{2l+1}),\ l \in S\}$ be the set of every second triplet in the sequence x. Note that $v_{l+1} = w_l$. Then generalized logit models in the general case (the saturated model) and in the case of Markov model M1 take the following simple forms:

$$\begin{aligned} \log\big(Q_{c|b}(v,w)\big) &= \lambda_0(c) + \lambda_L(v,c) + \lambda_R(c,w) + \lambda_{LR}(v,c,w), && (5)\\ \log\big(Q_{c|b}(v,w)\big) &= \lambda_0(c) + \lambda_1(v,c) + \lambda_1(c,w), \quad v,c,w \in \mathcal{A} && (6) \end{aligned}$$

respectively, where all these functions λ vanish provided any of their arguments take the reference value b.

Thus, the null hypothesis for the first order homogeneous Markov chain M1 is given by

$$H_0:\ \lambda_{LR}(v,c,w) = 0,\ \lambda_L(v,w) = \lambda_R(v,w) \quad \forall\, v,c,w \in \mathcal{A}. \tag{7}$$

In the statistical analysis only the first condition in (7) has been tested: $H_0 : \lambda_{LR}(v,c,w) \equiv 0$. If it holds for each position of the sequence (not only in every second position like in our data structure), the Hammersley-Clifford theorem implies $\lambda_L(v,w) \equiv \lambda_R(v,w)$.

Logit statistical analysis. We apply the logit analysis to assess the Markov property of bacterial genomes treated as random sequences. Non-coding regions of three bacteria, *Esherichia coli*, *Bordetella bronchiseptica* and *Coxiella burnetti*, from the GenBank database have been used in the analysis. The null hypothesis (7) is rejected for whole non-coding DNA sequence ($p < 0.0001$) and for some of non-coding subsequences of the each bacteria. For the whole non-coding sequences the null hypothesis of the Markov order $m = 2$ and $m = 3$ is rejected as well with p-value $p < 0.0001$. Notice that Markov model of the order m=3 has 624 parameters whereas the mean length of non-coding regions for bacteria is about 200.

Inhomogenity of data. The rejection of the null hypothesis of the first order Markov property can be caused merely by the data inhomogeneity. A rather sophisticated procedure based on resampling is performed to check this suggestion. The results obtained show that, within the generalized logit framework, DNA sequences are indeed rather inhomogeneous and this can lead to their non-Markovity. On the other hand, it is found that a significant part of non-Markov sequences (approximately 60%) can be viewed as homogeneous. Thus, even homogeneous DNA regions can not be treated as the first order Markov sequences.

Reverse-complement symmetry of DNA. DNA is double stranded helix. There is complementary between both DNA strands, as an A on one strand, always binds with T on the other, and C always binds with a G. The binding between the bases is through hydrogen bonds: two between A and T and three between C and G. The comparing of two complementary DNA strands often called reverse-complement symmetry (Baisne *et al* 2002) or reversibility property (Simons *et al* 2005). This property can be described as rule which is same in leading strand reading it in direction from left to right side and lagging strand – from right to left. Next step is to describe this rule.

Let us introduce some notation. Taking into account the two basic properties of nucleotides, it is convenient to replace the initial alphabet $\mathcal{A} = \{A, C, G, T\}$ with $\mathcal{A}_2 := \mathcal{A}_1 \times \mathcal{A}_1$, $\mathcal{A}_1 := \{-1, +1\}$ and to define the one-to-one mapping $c_1 : \mathcal{A} \to \mathcal{A}_2$ by the following rule:

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

Let β and π denote the corresponding components of the mapping c_1. Thus the components $\beta(v) \in \mathcal{A}_1$ and $\pi(v) \in \mathcal{A}_1$ of $v \in \mathcal{A}_2$ represent the property "number of bonds" and the property "purine versus pyrimidine", respectively. Let $v = (v_1, \ldots, v_n) \in \mathcal{A}_2^n$ be the nucleotide sequence in the first strand of DNA. Since the second strand is complementary to the first one and is read in the reversed order the nucleotide sequence in the former is the *reversed complement* of v: $\bar{v} := (\bar{v}_n, \ldots, \bar{v}_1)$, $\quad \bar{v}_i := (\beta(v_i), -\pi(v_i))$, $i = 1, \ldots, n$.

Definition 2.3. It is said that a random sequence $x = (x_1, \ldots, x_n) \in \mathcal{A}_2^n$ with independent of $\ell \in N^\circ$ conditional probabilities

$$\mathbf{P}\{x_\ell = v | x_{U_m(\ell)} = u\}, \quad \ell \in N^\circ,\ v \in \mathcal{A}_2, u \in \mathcal{A}_2^{2m}, \tag{8}$$

possesses the *local* strand symmetry of the mth order if for each $v \in \mathcal{A}_2$ and each $u \in \mathcal{A}_2^{2m}$

$$\mathbf{P}\{x_{m+1} = v \mid x_{U_m(m+1)} = u\} = \mathbf{P}\{x_{m+1} = \bar{v} \mid x_{U_m(m+1)} = \bar{u}\}. \tag{9}$$

Definition 2.4. Function $\psi : \mathcal{A}_2^{2m} \to \mathbf{R}$ is called *symmetric* (*antisymmetric*) if $\psi(u) = \psi(\bar{u})$ (respectively, $\psi(u) = -\psi(\bar{u})$) for each $u \in \mathcal{A}_2^{2m}$.

Proposition 2.2. If x is a locally-symmetric sequence of order m then there exist a symmetric function ψ and two antisymmetric functions ψ_- and ψ_+ such that:

$$\begin{aligned}\Lambda_{-1,1}(u) &= \psi_-(u),\\ \Lambda_{1,1}(u)-\Lambda_{1,-1}(u) &= \psi_+(u),\\ \Lambda_{1,1}(u)+\Lambda_{1,-1}(u)-\Lambda_{-1,1}(u) &= \psi(u),\end{aligned}$$

where $\Lambda_{\beta,\pi}(u) := \log\left(p((\beta,\pi)\mid u)/p((-1,-1)\mid u)\right),\ \beta,\pi\in\mathcal{A}_1$.

When estimating the functions $\psi_+(u), \psi_+(u)$ and $\psi(u), u\in\mathcal{A}_2^{2m}$, one needs some parametrization of them. In the dissertation an explicit parametric representation of the symmetric and antisymmetric functions is given. The symmetric (antisymmetric) function depends on k^+ (respectively k^-) free parameters where $k^\pm = 2^{2m-1}\pm k_*^2/2$, k_* is the number of subsets J of $\{1,\ldots,m\}$ such that $J = m+1-J$. Note that a general function defined on $\mathcal{A}_2^{2m}$ has 2^{2m} parameters. When $m=1$, for example, it has 16 parameters, whereas $k^+=10, k^-=6$.

Concluding remarks. A new framework of genetic sequence statistical analysis based on assumptions (A1), (A2), and generalized logit model is introduced. It is directly related to finite-state Markov field specification and thus convenient for statistical analysis of Markov dependence structure.

The results obtained show that, within the generalized logit framework, DNA sequences are rather inhomogeneous and this can lead to their non-Markovity. On the other hand, it is found that a significant part of non-Markov sequences (approximately 60%) are homogeneous. Thus, even homogeneous DNA regions can not be treated as the first order Markov sequences. For some bacterial DNA sequences, the second order local strand symmetry in general does not hold, but the asymmetry has different manifestation in different DNA regions. It means that DNA strands can have different evolutionary conditions and a biological sense.

3. Semi-parametric smoothing of sparse contingency tables

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.

Semi-parametric smoothing and re-sampling. In this section a method for hypothesis testing based on the semi-parametric smoothing and re-sampling is introduced. Although the method is rather general and can be applied to any kind of data, here we deal with finite-state random sequences.

Let $Y = \big(Y_l,\ l=1,\ldots,n\big)\in\mathcal{A}^n,\quad \mathcal{A}^n = \bigotimes_{i=1}^{n}\mathcal{A}$, be a random sequence with an unknown distribution μ. Here $\mathcal{A}$ is a finite state space (alphabet). Further, let $X(h) = \big(X_l(h),\ l=1,\ldots,n\big)\in\mathcal{A}^n,\quad h\in\mathbf{Z}_+ := \{0,1,2,\ldots\}$ be a discrete time homogeneous Markov chain with a transition probability matrix (Markov kernel) Π and initial distribution π, $\pi(a) = \mathbf{P}\{X(0) = a\},\ a \in \mathcal{A}^n$.

Assume that the Markov chain X is irreducible and acyclic. Then it is ergodic, and let q denote its unique stationary (invariant) distribution. If we take transition matrices Π from a given (parametric) family $\mathcal{P}$, the corresponding (parametric) family $\mathcal{Q}$ of target distributions q will be obtained. Thus, $q \in \mathcal{Q}$ if q is the invariant distribution for some $\Pi \in \mathcal{P}$.

Problem 3.1. Given an observation y of Y, to check the goodness-of-fit hypothesis $H_0 : \mu \in \mathcal{Q}$ versus $H_1 : \mu \notin \mathcal{Q}$, when only the transition probability matrices $\Pi \in \mathcal{P}$ are available, or more generally, when only a method of sampling from $q \in \mathcal{Q}$ is available.

Let $X(h) = X(h|y) = \{X_l(h),\ l = 1, \dots, n\},\ h \in \mathbf{Z}_+$ be a homogeneous Markov chain with the transition probability matrix Π starting at y, i.e. with the initial distribution $\pi(a) = \delta_y(a),\ a \in \mathcal{A}^n$. Set $q_h(a) = q_h(a|y) := \mathbf{P}\{X(h) = a|X(0) = y\},\quad a \in \mathcal{A}^n,\ h \in \mathbf{Z}_+$. Then $q_0 = \delta_y$ (δ_y denotes the Dirac measure supported at a point y) and $q_h \to q$ as $h \to \infty$ with the exponential rate of convergence for some $\kappa > 0$:

$$\|q_h - q\| \leq \mathrm{e}^{-\kappa h}\, \|q_0 - q\|, \quad \forall h \in \mathbf{Z}_+. \tag{10}$$

Here $\|\nu\|$ denotes the total variation of the (generalized) measure $\|\nu\|$ on the set $\mathcal{A}^n$.

Application of the semi-parametric smoothing to problems of hypothesis testing is based on the following proposition, which follows from inequality (10):

Proposition 3.1. Let the homogeneous Markov chain X be ergodic and for a given $\delta > 0$ and some $h \in \mathbf{Z}_+$ the inequality $\|q_h - q\| > \delta$ holds. Then $\|\mu - q\| > \delta$.

To decrease (change) the smoothing effect of q_h one can take their weighted averages

$$\bar{q}_h(a) = \sum_{t=0}^{h} \omega_h(t) q_t(a|y), \quad h \in \mathbf{Z}_+, \tag{11}$$

here $\omega_h(t) \geq 0$ is a collection of weights, $\sum_{t=0}^{h} \omega_h(t) = 1\ \forall h \in \mathbf{Z}_+$.

Thus, for fixed $h > 0$, $\mu_h(\cdot) := \mathbf{E}\bar{q}_h(\cdot|Y)$ can be treated as a "smoothed (toward q)" distribution of Y with the smoothing parameter h. The distribution q is thought of as a *target* (parametric) probability model. Faddy and Jones (1998) have called this method *semi-parametric smoothing*. The term "semi-parametric" is used to stress that the "smoothing" kernel Π depends on parameters of the expected parametric model of the data (i. e. the parameters of the distribution q), which are not directly related to the smoothing level.

Problem 3.2. The approach also provides a way to solve a problem of smoothing of categorical data. Smoothing is based on kernels which relate the likelihood of an (unobserved) object with that of its (observed) neighbours, i. e. objects similar to it in some sense. Usually distance measures of edit type (Hamming distance, Levenshtein metric) are applied in order to evaluate similarity between two sets of categorical features. In general, however, this forthright method is unnatural and too simplistic.

In semi-parametric smoothing, kernels arise naturally as an ergodic transformation which converges to a given target distribution. In practice, this corresponds to a stabilizing evolution or to "error correction" dynamics which are typical for natural living and social systems. Anyway, the choice of transition kernel in the framework of semi-parametric smoothing is justified at least by the null hypothesis to be tested.

Semi-parametric bootstrap. The main advantages of nonparametric bootstrap in comparison with parametric one is its simplicity and independence of parametric model (mis)specification and estimation procedures of unknown parameters of the model. In the case of sparse categorical data, the corresponding contingency table contains (a lot of) zeros, and therefore the nonparametric bootstrap is inapplicable. However, (nonparametric) re-sampling can be readily applied to the *smoothed contingency table* and it is in a sense an intermediate between nonparametric and parametric bootstrap.

Test statistics for the measure of the discrepancy between $\hat{q}_h$ and $\hat{q}_\ell$. Consider a vector functional $\lambda = \lambda(q)$ defined on a set $\mathcal{Q}$ of probability distributions on $\mathcal{A}^n$ and define χ^2-type statistics

$$\eta(h) = \eta(h|\,\ell) := (\hat{\lambda}_h - \hat{\lambda}_\ell)^\top \widehat{W}^-(\hat{\lambda}_h - \hat{\lambda}_\ell), \tag{12}$$

here $\hat{\lambda}_h = \lambda(\hat{q}_h)$, $\hat{q}_h$ is the empirical distribution of $\bar{q}_h$ based on resampling from $\{X(t),\ t = 1,\dots,h\}$ given $X(0) = y$, $\widehat{W}$ is an estimated covariance matrix of $\hat{\lambda}_h - \hat{\lambda}_\ell$, and $\widehat{W}^-$ is the generalized inverse of $\widehat{W}$. For the high-dimensional functionals λ, some restrictions on the structure of $\widehat{W}$ are necessary in order to get a feasible task. The common structures are diagonal (for χ^2 statistics) or spherical matrices ($W = \sigma^2 I$, $\sigma^2 > 0$, for L_2-distances).

Clearly, any other (normalized) measure of the discrepancy between $\hat{q}_h$ and $\hat{q}_\ell$, e.g. the power-divergence (Cressie and Read (1984)), can be used instead of (12).

Testing the null hypothesis is based on the statistic

$$\eta^* = \eta^*(\ell) := \max_{h\in\mathcal{H}} \eta(h|\,\ell), \tag{13}$$

here $\mathcal{H} = \{h_1,\dots,h_k\},\ h_1 < h_2 < \dots < h_k < \ell$, is a given set of smoothing parameter values. The parameter ℓ of the method should be taken as large as possible but is limited by computer resources.

The resampling methods (bootstrap) can be applied to estimate the distribution of statistic (13) and to evaluate its critical values and p-values.

The "optimal" value of the smoothing parameter h is

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

Generalized logit and semi-parametric re-sampling. An efficient implementation of semi-parametric smoothing and re-sampling requires to specify the appropriate transition matrix Π (the family $\mathcal{P}$), the functional λ in (12), and the weights $\omega_h(\cdot)$ (see (11)). In general setting, it is a difficult problem. In the dissertation a specific problem of fitting local dependence structure of finite-state random sequences is addressed with potential applications to genetic data DNA in mind. The data structure and generalized logit model, described in chapter 2 are used in the study. For simplicity, here we restrict ourselves to the nearest-neighbour interactions (6) and the binary alphabet. By taking advantage of "true" generalized logits (6) with a corresponding specification of parameters imposed by the null hypothesis H_0 under consideration, it is easy to calculate the conditional probabilities $p(c|v,w),\ (c,v,w)\in\mathcal{A}^3$, and hence to generate via Gibbs sampling with randomization a homogeneous Markov chain X which starts at y and has the stationary distribution determined by H_0.

Algorithm 3.1. Assume, that conditional probabilities $p(c\,|\,v,w),\ (v,c,w)\in\mathcal{A}^3$, and the initial state $X(0)=y$ are given and the random sequences $X(t),\ t=1,\dots,h,$ are generated. Then a new random sequence $X(h+1)$ is obtained as follows:

1. A location $j\in\{2,\dots,n-1\}$ is chosen at random with equal probabilities;

2. Given the location j, a value $c\in\mathcal{A}$ of $X_j(h+1)$ is chosen at random according to the conditional probabilities $p\big(c\,|\,X_{j-1}(h),X_{j+1}(h)\big)$, while the rest of the sequence remains the same: $X_i(h+1)=X_i(h),\ \forall\, i\in J_m,\ i\neq j.$

Computer experiment. A computer experiment is carried out to get a better insight on the practical performance of the procedure. The null hypothesis H_0 asserts that Y is a stationary first-order Markov chain with transition probabilities $p_{01}=p_{10}=\rho$, where ρ is assumed to be known. Then the marginal distribution of Y (the invariant distribution of the sequence $Y_l,\ l=1,\dots,n$) is $\mathrm{Binomial}(1,1/2)$.

The problem is to test simple null hypothesis H_0 versus the composite alternative H_1, the opposite hypothesis to H_0. The power of initial (non-smoothed) test statistic $\eta(0)$ (12) and smoothed statistic η^* (13) for various cases of the alternative H_1 have been considered: complete independence H_I, a sequence of independent triplets H_T, a special second-order Markov chain H_{M2} and others. In the case of

H_I, $Y_i, i = 1, \ldots, n$, are iid with $Y_1 \sim \text{Binomial}(1, 1/2)$. The alternative H_T states that Y is a collection of $n/3$ independent triplets with the same as in H_0 probabilities of triplets. To separate the hypotheses H_0 and H_T is a difficult task.

The weights in (11) are assumed to be $\omega_h(t) = 1/(h+1), \ t = 0, 1, \ldots, h$. Thus far their impact on the power of the statistics has been negligible. The covariance matrix $\widehat{W}$ in (12) is assumed to be diagonal and is estimated from separate independent samples generated under the valid hypothesis H_0.

For short sequences (with $n < 20$, say), the exact powers of the tests can be obtained by direct calculations. Here, "exact" means that conditional re-sampling averages, given $Y = y$, as well as (unconditional) averages of re-sampling from Y are replaced by the corresponding (conditional and unconditional) expectations. The results for the test statistics $\eta(0)$ and η^* with $n = 9$ based on 3-tuples and 5-tuples are summarized in chapter 3. Note that the corresponding contingency tables are very sparse. The average cell frequencies for 5-tuples are $5/32$.

For longer sequences, the direct calculations are infeasible and simulations are used to assess the power of the tests.

Concluding remarks. The proposed method of semi-parametric smoothing and re-sampling is very flexible and can be used for any ill-posed statistical problem. Although the method can be applied to continuous data, it is especially suitable for sparse categorical data, since it provides a natural way to define a kernel for data smoothing. The computer experiment shows that the effect of semi-parametric smoothing sometimes is insignificant, however it usually increases with the order of m-tuples. This can be caused by the rather short lengthes of the binary sequences considered in the experiment and the discrete character of the smoothing parameter. It is expected that the semi-parametric smoothing and re-sampling is statistically more efficient for high-dimensional categorical data in the case of contiguous alternatives, or in other words, when discrete data, in a sense, can be treated as continuous.

General conclusions

1. The proposed method of semi-parametric smoothing can be used for smoothing of categorical variables and it enables one to apply re-sampling technique (*semi-parametric or smoothed bootstrap*) for sparse contingency tables. A convenient framework of logit model for implementation of semi-parametric smoothing is proposed. The method is quite simple but its efficient implementation is a difficult problem. Exact straightforward calculations are feasible only for low-dimensional data whereas simulations are very computer-intensive even for the medium dimensionality.

2. Generalized logit model, when applied to DNA data with the special data

structure introduced in the thesis, is consistent with the context-dependend model of DNA evolution.

3. DNA sequences are inhomogeneous and this can lead to their non-Markovity. On the other hand, it is found that a significant part of non-Markov sequences (approximately 60 %) are homogeneous. Thus, even homogeneous DNA regions can not be treated as the first order Markov sequences.

4. The typical length of DNA sequences is 200 *bp*. The longer sequences are known to be rather inhomogeneous and it is difficult to fit the model for higher Markov order because the number of model parameters increases exponentially fast with respect to the order m of the model.

5. It is shown, that for some bacterial DNA sequences, the first order local strand symmetry generally does not hold, but the asymmetry has a different manifestation in different DNA regions (for 13 regions of 17 first order local strand symmetry does not hold). It means that DNA strands can have different evolutionary conditions and a biological sense.

List of scientific publications on the topic of dissertation

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.

About the author

Jurgita Židanavičiūtė was born in Trakai, on 22 of November, 1979. She entered Vilnius Gediminas Technical University, Faculty of Fundamental Sciences in 1997. First degree in Informatics, 2001. Master of Science in Statistics in 2003. In 2003–2008 – PhD student of Vilnius Gediminas Technical University, Department of Mathematical Statistics. At present works as a junior lecturer at Vilnius Gediminas Technical University. In 2003–2008 was working at the Institute of Mathematics and Informatics. Since 2008 are working at the Bank of Lithuania.

KATEGORINIŲ POŽYMIŲ PRIKLAUSOMYBIŲ STRUKTŪROS STATISTINĖ ANALIZĖ IR JOS TAIKYMAS GENETIKOJE

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 dideli, ir dėl to vis dažniau statistikoje susiduriama su didelio matavimo arba išsklaidytų duomenų analizės problemomis.

Kategoriniams (diskretiems) duomenims didelės duomenų dimensijos problema yra ypač aktuali. Kiekybinių duomenų modeliai paprastai aprašo vidurkių priklausomybės struktūrą ir 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 atžvilgiu dimensijos n 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 – 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. Daugelyje straipsnių 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 DNR sekose. Jis svarbus dėl to, kad Markovo eilė yra susijusi su kaimyninių nukleotidų, įtakojančių evoliuciją konteksiniuose evoliucijos modeliuose, skaičiumi. Jeigu DNR sekų atsitiktinės mutacijos priklauso tik nuo m artimiausiųjų kaimyninių nukleotidų, tai stacionarusis nukleotidų evoliucijos skirstinys sudaro m-os eilės Markovo grandinę (Jensen 2005).

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 iš vis 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 genetinių sekų statistinės analizės modelį, suderintą su jų evoliucijos 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, pakartotinų 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. Svarbi išimtis yra Simons *et al* (2005) darbas, kuriame pasiūlytas *globalus* komplementarių vijų simetriškumo tikimybinis modelis, kuris vis tik 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 statitinė 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 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 įvairių hipotezių tikrinimui leidžia pritaikyti neparametrinį (semiparametrinį) savirankos metodą 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 nemarkoviškumą.

5. Reikšminga dalis homogeninių fragmentų nėra pirmos eilės Markovo.

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.

Disertacijos struktūra

Darbo apimtis yra 80 puslapių, pateiktas 1 paveikslas, 8-ios lentelės ir remtasi 70 literatūros šaltinių. Disertaciją sudaro įvadas, trys skyriai ir išvados.

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.

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 ženkli 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.

Apie autorių

Jurgita Židanavičiūtė gimė 1979 m. lapkričio 22 d. Trakuose. 1997 m. baigė Lentvario M. Šimelionio vidurinę mokyklą. 2001 m. įgijo informatikos bakalauro laipsnį Vilniaus Gedimino technikos universiteto Fundamentinių mokslų fakultete. 2003 m. tame pačiame universitete baigė magistrantūros studijas ir įgijo statistikos magistro laipsnį. 2003–2008 m. – VGTU Matematinės statistikos katedros doktorantė. Nuo 2003 m. – tos pačios katedros asistentė, lektorė. Nuo 2008 m. dirba Lietuvos banko Statistikos departamente.

Jurgita ŽIDANAVIČIŪTĖ

DEPENDENCE STRUCTURE ANALYSIS
OF CATEGORICAL VARIABLES
WITH APPLICATIONS IN GENETICS

Summary of Doctoral Dissertation
Physical Sciences, Mathematics (01P)

KATEGORINIŲ POŽYMIŲ PRIKLAUSOMYBIŲ STRUKTŪROS
STATISTINĖ ANALIZĖ IR JOS TAIKYMAS GENETIKOJE

Daktaro disertacijos santrauka
Fiziniai mokslai, matematika (01P)

2010 01 05. 1,5 sp. l. Tiražas 70 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*