Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 3, August 2020, pp.1517-1524

Symptom analysis of multidimensional categorical data with applications

N. P. Alexeyeva, F. S. Al-Juboori, E. P. Skurat
Department of Mathematics, St. Petersburg State University, 28, University Avenue, 198504, St. Petersburg, Russia

ABSTRACT

The linear combinations of dichotomous variables over the field $\mathbb{F}_2$, which are called symptoms, form the projective space from which it is possible select the more informative subspaces for reducing the dimensionality of binary data. In this article, the symptom space expands to the super-symptom space. The super-symptom means a linear combination of various multiplications of k dichotomous variables over a field of characteristic 2 without repeating. In algebra, such functions are called Zhegalkin polynomials or algebraic normal forms. It is known that each logical function can be represented in the form of a Zhegalkin polynomial in a unique way, therefore using them to iterate one can find a logical function to best describe a risk group. The search algorithm of a more informative super-symptom for classification is based on the superposition of impulse sequences with different types of operations: first multiplication and then addition over the $\mathbb{F}_2$ field. Also the super-symptom analysis is a convenient method for a study of the correlation between two sets of categorical variables. This method was applied to identify the most severe forms of the disease by combining hormonal, immunological and genetic tests in patients with breast cancer (data from Cancer Oncology Hospital in Medicine City in Baghdad) and to identify genetic risk factors by patients with alcohol dependence syndrome, receiving alcohol dependence therapy (St.Petersburg V.M. Bekhterev Psychoneurological Research Institute).

Keywords: Multidimensional categorical data, dimension reduction, Zhegalkin polynomials, symptom analysis, impulse sequence.

Corresponding Author:

Skurat Evgeniia,
Departement of Mathematics,
St. Petersburg State University,
28, Universitetskiy pr., St. Petersburg, 198504, Russia

E-mail: skuratevgenia@gmail.com

1. Introduction

The statistical task of comparing a single dependent variable with a complex of several independent dichotomous variables remains actual, especially, when the influence of various factors on the dependent variable is being studied separately, and all interrelations are insignificant. Separate factors are sometimes not enough to describe the risk group. If many factors are considered then there is a problem of the dimension reduction which means search for a few functions of factors with the least information loss. Models of such functions may be different.

In this article, we can use the symptom-syndrome models [1] in which the predicate is expressed in terms of independent factors as linear combinations over field $\mathbb{F}_2$, which form the finite projective space [2].

If we construct a finite projective space not for k dichotomous variables, but for their $2^k - 1$ various non-degenerate multiplications without repetition, then we receive Zhegalkin polynomials which describe all sorts of logical functions – all possible combinations of logic operations: addition, negation, multiplication of these

k variables. The main problem lies in the complexity of calculations. Modern computing capabilities make it easy to operate with up to three-four of dependent factors. But if you first find the three or four most significant variables, then they will be enough to determine the risk group using one variable, which is expressed through a logical combination of these factors.
To check the independence of a pair of categorical variables, one can use the uncertainty coefficient or the p-value of Fisher exact test [3], application example in symptom analysis [4]. In the case of the metric dependent variable, the homogeneity tests: t-test, the Kruskal Wallis test or ANOVA are used [5]. The Gehan Wilcoxon test [6] is used in survival symptom analysis [7].

2. Material and methods

2.1. Symptom and super-symptom

Consider random vector $X = (X_1, \dots, X_m)^{\mathrm{T}}$ with components taking values 0 and 1. Usually 0 and 1 mean lack and presence of factors respectively. The new variable $X_{ij} = X_i + X_j (\mathrm{mod} 2)$ [1] also takes only values 0 and 1 and means presence of any one in the absence of another factor. This latent variable is called the symptom. For example, if X_i and X_j direct to the big height and the big weight respectively then $X_{ij} = 1$ means inadequacy of height to weight: tall and thin or short and fat. Contrariwise $X_{ij} = 0$ means that weight is adequate for height: short and lightweight or tall and heavyweight. We can consider the such sum of k variables, the meaning of which at $k > 2$ is more complicated and will be discussed further.
Definition 1 Let be $\tau = (\mathrm{t}_1, \dots, \mathrm{t}_k) \subseteq (1,2, \dots, \mathrm{m})$. Then $\mathrm{X}_\tau = \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{X}_{\mathrm{t_i}} (\mathrm{mod} 2)$ is called the symptom X_τ of the rank k.

We note again, since all operations are performed over the field $\mathbb{F}_2$ all symptoms take only values of 0 or 1. Let $\tau = (1,2,3)$ and $X_{123} = X_1 + X_2 + X_3 (\mathrm{mod} 2)$ takes a value of 0 when all factors are equal to 0 or the value of 0 occurs only once. The symptom $X_{123} = 1$ when all factors are equal to 1 or the value of 1 occurs only once. The variable X_{12} means that if $X_1 = 0$ then $X_2 = 1$ or if $X_1 = 1$ then $X_2 = 0$. The elements of vector $X = (X_1, \dots, X_m)^{\mathrm{T}}$ are trivial symptoms of rank 1. We define symptom of rank zero degenerate, it takes a value of 0 with a probability of 1.

Definition 2 Let $\tau = (\mathrm{t}_1, \dots, \mathrm{t}_k) \subseteq (1,2, \dots, \mathrm{m})$ be one of $2^{\mathrm{m}} - 1$ subsets with the exception of the empty set $\emptyset$, where $\mathrm{k} \leq \mathrm{m}$. We denote the result of multiplying several dichotomous variables $\mathrm{X}_{\mathrm{t}_1}, \dots, \mathrm{X}_{\mathrm{t}_k}$ by $\mathrm{X}^\tau = \mathrm{X}_{\mathrm{t}_1} \cdot \dots \cdot \mathrm{X}_{\mathrm{t}_k}$. For different $\tau_1, \dots, \tau_{\mathrm{L}}$ polynomials over the field $\mathbb{F}_2$ of the form $\sum_{\mathrm{i}=1}^{\mathrm{L}} \mathrm{X}^{\tau_\mathrm{i}} (\mathrm{mod} 2)$ are called super-symptoms.
In particular, the super-symptom $s_1 = X_1 + X_2 + X_1 X_2 (\mathrm{mod} 2)$ means presence X_1 or X_2 or both together when $s_1 = 1$, which corresponds to the logical sum. When $s_1 = 0$ then $X_1 = 0$ and $X_2 = 0$ at the same time. The super-symptom $s_2 = X_1 X_2 + X_1 X_3 + X_2 X_3 (\mathrm{mod} 2)$ means presence not less two out three factors X_1, X_2, X_3 at $s_2 = 1$. When $s_2 = 0$ then all X_1, X_2, X_3 are equal 0 or $X_i = 1$ separately, i.e the vector $X = (X_1, X_2, X_3)$ takes values $(0,0,0), (1,0,0), (0,1,0), (0,0,1)$.
Remark that the super-symptom $\bar{s} = s + 1 (\mathrm{mod} 2)$ differs from the super-symptom s only in order of graduations. For example $s = 1$ and $s = 0$ mean the big and small height respectively. On the contrary $\bar{s} = 1$ and $\bar{s} = 0$ mean the small and big height respectively. Thus we can treat s and $\bar{s}$ as one and the same super-symptom when it meets in isolation without other super-symptoms. So we consider as super-symptoms all polynomials over the field $\mathbb{F}_2$ with zero free member without loss of generality.
In [5], there is the definition of projective geometry $PG(n, q)$ that the vectors $X_0, \dots, X_n$ form over a finite field $\mathbb{F}_q$. Similarly introduce the following definition.
Definition 3 Let be there are linearly independent symptoms $\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{k}$, coefficients $\gamma_\mathrm{j} \in \mathbb{F}_2, \mathrm{j} = 1, \dots, \mathrm{k}$, γ_j are not zero at the same time. Then a collection of $2^{\mathrm{k}} - 1$ symptoms in the form $\gamma_1 \mathrm{X}_1 + \dots + \gamma_\mathrm{k} \mathrm{X}_\mathrm{k} (\mathrm{mod} 2)$ is called the syndrome S_k.

[1] The expression $a(\mathrm{mod} 2)$ means the remainder of a division of the number a by 2. This corresponds to that all operations are performed over the field $\mathbb{F}_2$.

The single symptom X_τ can be considered as a syndrome S_1 which is analogous to the point $PG(0,2)$. Symptoms X_τ, X_μ and $X_{\tau\mu} = X_\tau + X_\mu(\mathrm{mod}2)$ form the syndrome S_2 which is analogous to the projective line $PG(1,2)$. Just as a projective line can only be defined by two points, the syndrome S_2 can be built on any pair of variables because $X_{\tau\mu} = X_\tau + X_\mu(\mathrm{mod}2)$, $X_\mu = X_{\tau\mu} + X_\tau(\mathrm{mod}2)$, $X_\tau = X_{\tau\mu} + X_\mu(\mathrm{mod}2)$. The syndrome S_3 or projective plane $PG(2,2)$ can be built on any three of linear independent variables and so on.
Similar to inductive construction of projective geometries the introduction of a new point, not belonging to finite geometry, and through the construction of all lines containing this point, it is possible to determine the syndrome $k-1$ - th order on the base of linearly independent symptoms $X_{t_1} \dots, X_{t_k}$ recurrently:

$$\begin{aligned} &S_1 = S_1(X_{t_1}) = X_{t_1}\,, \\ &S_2 = S_2(X_{t_1}, X_{t_2}) = (X_{t_1}, X_{t_2}, X_{t_1} + X_{t_2}(\mathrm{mod}2)), \\ &\vdots \\ &S_i = S_i(X_{t_1}, \dots, X_{t_i}) = (\,S_{i-1}, X_{t_i}, S_{i-1} + X_{t_i}(\mathrm{mod}2)), \\ &\quad where\ \ X_{t_i} \notin S_{i-1}, i > 2\,. \end{aligned} \tag{1}$$

The symptoms of $S_k(X_1, \dots, X_k)$ form an impulse sequence with $2^k - 1 = K$ elements. For example, $S_3(X_1, X_2, X_3)$ looks like $(X_1, X_2, X_{12}, X_3, X_{12}, X_{13}, X_{123})$. The elements on the 2^i-nd place are called basic. If we take as basic the symptoms X_{12}, X_{13}, X_2 then $S_3(X_{12}, X_{13}, X_2) = (X_{12}, X_{13}, X_{23}, X_2, X_1, X_{123}, X_3)$. Remark that syndromes $S_3(X_1, X_2, X_3)$ and $S_3(X_{12}, X_{13}, X_2$ different only by order of elements. We can generalize an impulse sequence (1) when we use any operation on the field $\mathbb{F}_2$, not just addition. We define a special impulse sequence from $(X_{t_1}, \dots, X_{t_k}), X_{t_i} \in S_k$ with operation (*).

$$\begin{aligned} &M_1 = M_1(X_{t_1}|*) = X_{t_1}\,, \\ &M_2 = M_2(X_{t_1}, X_{t_2}|*) = (X_{t_1}, X_{t_2}, X_{t_1} * X_{t_2}), \\ &\vdots \\ &M_i = M_i(X_{t_1}, \dots, X_{t_i}|*) = (\,M_{i-1}, X_{t_i}, M_{i-1} * X_{t_i}), \\ &\quad where\ \ X_{t_i} \notin M_{i-1}\,, i > 1\,. \end{aligned} \tag{2}$$

One can see a match $M_k(X|+) = S_k(X)$ for $X = (X_1, \dots, X_k)$. In case of multiplication ($\odot$) over the field $\mathbb{F}_2$ and $k = 2$ we have $M_2(X_1, X_2|\odot) = (X_1, X_2, X_1X_2)$ which consists of $K = 2^2 - 1 = 3$ elements. In case $k = 3$, $M_3(X_1, X_2, X_3|\odot) = (X_1, X_2, X_1X_2, X_3, X_1X_3, X_2X_3, X_1X_2X_3)$ consists of $K = 2^3 - 1 = 7$ elements and so on. We can use these elements as basic for the syndrome. Thus, we get a simple way to construct *a super-syndrome* as $S_K(M_k(X_1, X_2, \dots, X_k|\odot))$, where $K = 2^k - 1$, which includes all $2^K - 1$ polynomials over the field $\mathbb{F}_2$ with free member equal 0.

For example, the impulse ordered syndrome over the field $\mathbb{F}_2$ at $k = 2$ includes $2^{2^2-1} - 1 = 7$ polynomials: $S_3(M_1(X_1, X_2|\odot)) = S_3(X_1, X_2, X_1X_2) = (X_1, X_2, X_1 + X_2, X_1X_2, X_2 + X_1X_2, X_2 + X_1X_2, X_1 + X_2 + X_1X_2)$ All these super-symptoms have a simple logical interpretation: the new factor X_1X_2 direct to that factors X_1 and X_2 are present at the same time, $X_2 + X_1X_2(\mathrm{mod}2) = \bar{X}_1X_2$ means that the factor X_2 is present and the other X_1 is absent[2], $X_1 + X_1X_2 = X_1\bar{X}_2$ contrariwise. The super-symptom $X_1 + X_2 + X_1X_2(\mathrm{mod}2) = 1 + \bar{X}_1\bar{X}_2(\mathrm{mod}2)$ corresponds to the logic sum. As mentioned earlier, the symptom $X_{12} = X_1 + X_2(\mathrm{mod}2)$ means the presence of one factor in the absence of another which corresponds to the logic sum of factors $X_1\bar{X}_2$ and $\bar{X}_1X_2$,

$$\begin{aligned} &X_1\bar{X}_2 + \bar{X}_1X_2 + \bar{X}_1X_2\bar{X}_2X_1(\mathrm{mod}2) = \\ &= X_1 + X_1X_2 + X_2 + X_1X_2(\mathrm{mod}2) = X_1 + X_2(\mathrm{mod}2) = X_{12}\,. \end{aligned}$$

2.2. Selection algorithm

Let the dichotomous variables $X_1, \dots, X_m$ be involved separately in some statistical test. We consider all possible combinations $X_{t_1}, X_{t_2}, X_{t_3}$ and then calculate all possible elements $Y_j \in S_7(M_3(X_{t_1}, X_{t_2}, X_{t_3}))$, where $t_1, t_2, t_3 \in \{1,2, \dots, m\}$, $j = 1, \dots, 2^7 - 1$. In each case entropy H_j of the super-symptom Y_j and appropriate p-value p_j are calculated. Factor Y_j is considered significant when $p_j < 0.005$ adjusted for multiple comparisons and $H_j > 0.05$. Thus we choose the most significant super-symptom.
Each of $127 = 2^7 - 1$ possible super-symptoms constructed from three variables $a, b, c \in \{X_1, \dots, X_m\}$ can be expressed both as a polynomial modulo 2 and as a combination of logical operations. Almost half of the

[2] The factor $\bar{X} = X + 1(\mathrm{mod}2)$ is opposite to the factor X.

super-symptoms (63 out of 127) have a fairly simple interpretation. First of all, expressions $\alpha^{k_1}\beta^{k_2}\gamma^{k_3}$ are easily interpreted, where

$$\alpha \in \{a, \bar{a}\}, \beta \in \{b, \bar{b}\}, \gamma \in \{c, \bar{c}\}. \quad (3)$$

Multiplication of several symptoms means that all of there take a value 1 at the same time. Degrees k_1, k_2, k_3 take values 0 or 1 and these $23 = C_3^1 + \sum_{j=2}^{3} C_3^j 2^j$ expressions look like Table 1.

Table 1. Types of expressions

C_3^1
a
b
c

$2^2C_3^2$		
ab	ac	bc
$\bar{a}b$	$\bar{a}c$	$\bar{b}c$
$a\bar{b}$	$a\bar{c}$	$b\bar{c}$
$\bar{a}\bar{b}$	$\bar{a}\bar{c}$	$\bar{b}\bar{c}$

$2^3C_3^3$	
abc	$\bar{a}bc$
$ab\bar{c}$	$\bar{a}b\bar{c}$
$a\bar{b}c$	$\bar{a}\bar{b}c$
$a\bar{b}\bar{c}$	$\bar{a}\bar{b}\bar{c}$

The expression $\alpha\beta + \alpha\gamma + \beta\gamma(\mathrm{mod}2)$ means the presence of two or more factors from α, β, γ. There are eight variants for such expressions depending on (3). One can see that if we replace all the variables with opposite then we get the same polynomial but with the free member equal 1, i.e. $\bar{\alpha}\bar{\beta} + \bar{\alpha}\bar{\gamma} + \bar{\beta}\bar{\gamma}(\mathrm{mod}2) = \alpha\beta + \alpha\gamma + \beta\gamma + 1(\mathrm{mod}2)$. But since the super-symptoms are determined with precision about the permutation of gradations, we have four polynomials of the form: $ab + ac + bc, \bar{a}b + \bar{a}c + bc, a\bar{b} + ac + \bar{b}c, ab + a\bar{c} + b\bar{c}(\mathrm{mod}2)$.

We can construct $24 = C_3^1 \cdot 2^3$ super-symptoms of the form $\alpha(\beta + \gamma + \beta\gamma)$, corresponding to the presence of factors α simultaneously with β or γ, where C_3^1 indicates that three kind of variables can be aside and 2^3 is the number of variants depending on (3).

Finally, we can add $12 = C_3^1 \cdot 2^2$ expressions of the form $\alpha\beta + (\alpha + 1)\gamma$, corresponding to the presence of factor α together with β or the opposite $\bar{\alpha}$ with factor γ. There are four expression instead of eight variants depending on (3), because for all $\alpha \in \{a, \bar{a}\}$ we have $\alpha b + \bar{\alpha}c = \alpha\bar{b} + \bar{\alpha}\bar{c} + 1$ and $\alpha\bar{b} + \bar{\alpha}c = \alpha b + \bar{\alpha}\bar{c} + 1$.

Other super-symptoms can also be expressed through logical functions, but in a more complex way. Sometimes several super-symptoms have comparable p-values. In this case, it becomes possible to choose as a nominative representative the factor which is more accessible for interpretation.

3. Calculation and results

In statistical analysis the data or variable selection is an active research area. Before starting to apply the classification method, the variable selection methods could be used to minimize the number of features in the research dataset. Therefore, the purpose standing behind variable selection is to select a subset of variables by ignoring features with less important information. This is especially importantly for categorical data.

3.1. Symptom survival analysis and application in genetics

The study was conducted on the basis of the Department of narcology of the National medical Research center of Psychiatry and Neurology (SMRC PN) to them. V. M. Bekhtereva in the period 2013-2017. Within the double-blind placebo-controlled study, 100 patients with alcohol dependence syndrome (ICD-10) were randomly distributed (randomized) into 2 groups: patients of the main group (50 people) received pregabalin at a dosage of 150 mg / day (at night), patients of the comparison group (control group) (50 people) received an identical-looking placebo. The study drug was prescribed for 3 months (12 weeks), during which the subjects had to visit the research center on a weekly to control remission, drug compliance, as well as for psychometric assessments. In addition, blood samples were taken from patients. For technical reasons, 86 patients were available for analysis, blood samples of the remaining patients were lost or DNA isolation and genotyping was impossible. Differences in the duration of remission in the treatment program in carriers of different polymorphic variants of genes and their combinations were carried out using Kaplan-Meier survival analysis. The significance of differences in survival curves was assessed using the Gehan's Wilcoxon test. Interrelation of outcomes and separate polymorphic variants of genes were evaluated independently from the group of therapy. As a time scale were considered the time before retiring from the program (time before retiring from the program for any reason relapse, violation of the conditions of participation).

The dichotomous variable was considered as the dropout scale based on the following:

• 1 (dropped out of the treatment program),

• 0 (completed the treatment program).

The recessive model was used for the analysis - three genotypes for each polymorphic locus were aggregated into two groups:

• 0 (carriers only of major allele in homozygous state),

• 1 (all other genotypes).

Consider composition of the genetic panel selected in this study which includes code, recessive model and decryption.

• Single nucleotide polymorphisms (SNP) in the DRD2 gene of the dopamine receptor D2 gene rs1799732 (type 2 dopamine receptor gene) is encoded by a variable G_1 where $G_1 = 0$ when the genotype is CC and there is a carrier only of major allele in homozygous; $G_1 = 1$ when there are polymorphisms CT, TT. This gene is detected in significant amounts in the limbic system of the brain and plays an important role in the functioning of the central nervous system. It is considered an autoreceptor for dopamine (DA). Polymorphisms CT, TT determines the reduced concentration or high severity of alcoholism.

• SNP in the DRD4 gene of the dopamine D4 receptor gene rs1800955 is encoded by a variable G_2, where $G_2 = 0$ when the genotype is TT; $G_2 = 1$ when there are polymorphisms CT, CC. This gene is the main acceptor of the neuronal impulse in the dopamine neurotransmitter system, it is located at the terminal of the neuron that receives the nerve impulse, and mediates the effects of dopamine as a neurotransmitter. It is expressed at high levels in the prefrontal cortex and is the dominant DA receptor localized in this region of the brain. Carriers of the minor allele CT, CC are less effectively treated with serotonin and have an increased susceptibility to novelty seeking.

• SNP in the gene of the opioid μ-receptots OPRM1 rs1799971 is encoded by a variable G_3 where $G_3 = 0$ when the genotype is AA; $G_3 = 1$ when there are polymorphisms AG, GG. Opioid receptors of μ types (OPRM1) are the most important effector of the opioid reinforcing effect. The minor allele in exon 1 of the μ opioid receptor OPRM1 gene causes the normal amino acid at residue 40, asparagine, to be replaced by aspartic acid. Carriers of at least one minor allele AG, GG appear to have stronger cravings for alcohol than carriers of two major alleles.

• SNP in the gene of the GABA-α receptors rs567926 is encoded by a variable G_4, where $G_4 = 0$ when the genotype is TT; $G_4 = 1$ when there are polymorphisms CT, CC. The gamma-amino butyric acid alpha receptors GABA-α is an ionotopic receptor and ligand-activated ion channel. The endogenous GABA ligand, upon binding of which hyperpolarization of the neuron membrane occurs, which is the basis of the inhibitory effect of GABA. There is evidence of the dependence of a number of polymorphisms CT, CC of the genes of these subunits with alcohol addictive.

• SNP in the gene of glutamate receptor rs2832407 is encoded by a variable G_5, where $G_5 = 0$ when the genotype is CC; $G_5 = 1$ when there are polymorphisms AC, AA. The study of the structure of genes coding for glutamate and GABA systems indicates a link between their polymorphism and the presence of motivation for alcohol consumption. In participants with the two major alleles genotype, treatment enhances self-efficacy and reduces heavy drinking.

The super-symptom method allows us to define that patients differ best in duration of remission by means the composite genetic factors. Groups patients with the genotypes $[G_1(CC), G_3(AG, AA), G_4(TT)]$ or $[G_1(CC), G_3(GG), G_4(CC, CT)]$ were combined into one favorable group with the least number of relapses, which can be described by a super-symptom of the form

$$S = (1 + G_1)(G_3 + G_4)(\mathrm{mod}2)\,.$$

This composite factor S means the presence one of two risk factors G_3, G_4 at absence of G_1. Average time spent in the program is equal to 9.92 at $S = 1$ and 6.56 at $S = 0$, significance [3] of the Gehan's Wilcoxon Test is equal to $p = 1.8 \cdot 10^{-5}$. The probability of outcome is equal to $p_1 = 0.35(37)$ at $S = 1$ compared to $p_0 = 0.82(45)$ in another group at $S = 0$, significance of the Fisher Exact Test is equal to $p = 2.8 \cdot 10^{-5}$.

In combination with genes of factor S, other factors $T = G_1G_3 + G_1G_4 + G_3G_4(\mathrm{mod}2)$ can be identified. This composite factor T means presence of at least two from three risk factors. Average time of

[3] Further we denote the significance p-value of statistical tests by p.

remission is equal to 5.68 at $T = 1$ and 8.8 at $T = 0$, significance of the Gehan's Wilcoxon Test is equal to $p = 0.0003$.

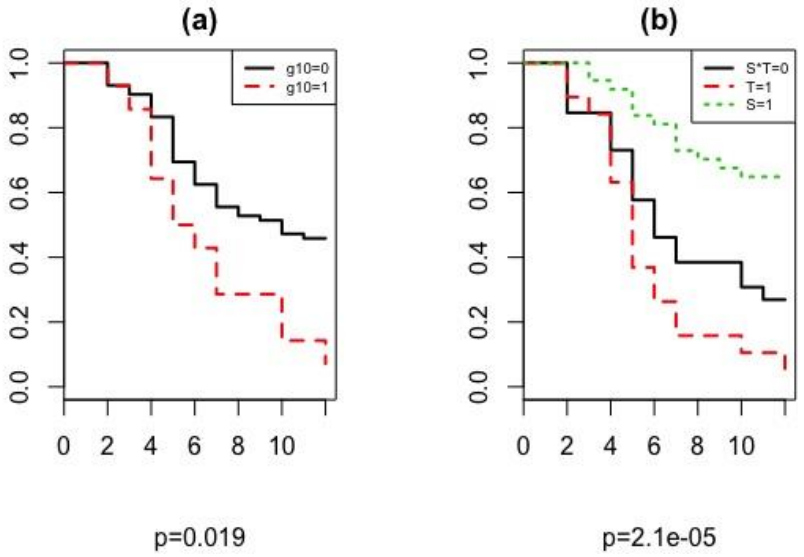

Figure 1. The Kaplan-Meier survival curves for the duration of the relapse-free period with factors: (a) G_1, (b) combination of $S = (1 + G_1)(G_3 + G_4)(\mathrm{mod}2)$ and $T = G_1G_3 + G_1G_4 + G_3G_4(\mathrm{mod}2)$

Thus, the outcomes from the treatment program are significantly associated with the combination of genetic data. Patients with two of the three genetic risk factors have a shorter recurrence-free period. If add G_2 in selection algorithm then the most significant factor looks like

$$U = (1 + G_1)(G_3 + G_4)G_2(\mathrm{mod}2).$$

Average time spent in the program is equal to 10.59 at $U = 1$ and 7.08 at $U = 0$, significance of the Gehan's Wilcoxon Test is equal to $p = 2.9 \cdot 10^{-5}$. The probability of outcome at $U = 1$ is equal to $p_1 = 0.22(27)$ compared to $p_0 = 0.78(59)$ in another group at $U = 0$, significance of the Fisher Exact Test is equal to $p = 2 \cdot 10^{-6}$.

If we assume that polymorphisms of gene G_2 are responsible for risk-taking, polymorphisms of gene G_1 contribute to cognitive disorders and severe alcohol dependence, polymorphisms of gene G_3 indicate high doses of alcohol, and gene G_4 helps to quickly become addicted, then we will get a description of a group of patients with longer remission. These patients have increased sensitivity to the search for new sensations ($G_2 = 1$), not prone to cognitive disorders ($G_1 = 0$), or have separately a tendency to large doses of alcohol ($G_3 = 1$), or a fast addiction ($G_4 = 1$). If we construct the vector (G_1, G_2, G_3, G_4) then this group corresponds to two realizations (0,1,0,1) or (0,1,1,0) with probability of outcome 0.35. At combination (0,1,1,1) with two risk factors simultaneously, the probability of outcome from the treatment is a lot higher being equal to 0.86(7).

In the case of combination (0,1,0,0), it would seem that the situation should be better, however we have the probability of outcome is equal to 0.78(18). Perhaps this is due to the fact that 10 out of 18 patients have $G_5 = 1$, which means $G_5(AS, AA)$associated with heavy drinking and treatment difficulties, and the probability of outcome is 0.9(10). The rest 8 patients which have only one risk factor $G_2 = 1$, while $G_i = 0, i = 1,3,4,5$, have nearly the same chances to complete the program or leave it since the probability of the outcome is 0.625(8).

It follows from this study that outcomes from the treatment program can be predicted on the base of composition of the genetic panel selected in this study.

3.2. Canonical symptom analysis and biometrical example

Symptom analysis can be used to identify the structure of the relationship between two sets of categorical variables by analogy with the canonical correlation analysis. The research data set was proposed to

be collected from Cancer Oncology Hospital in Medicine City in Baghdad for a set of patients who were scheduled for biopsy, mammograms interpreted by radiologists. Laboratory and Clinical Investigations, Ultrasound of mammary glands and elastography provided data on mammographic results as a part of the standard mammographic workup.

There are some features important for a physician to make decision whethere dangerous operation is necessary for treatment or not. According to the results of initial statistical processing, surgery confirms malignancy in 85% of cases. Only a few patients with malignancy not undergoing surgery. Encoding of 11 selected variables is presented in Table 2.

Table 2. Encoding signs

code	name	indicat
A	Age	1- age less than 59 (73%)
		0 - age greater than 59 (27%)
D	The Oncotype DX test	1 - ILC nodal severe type tumore (13%)
		0 - IDC ducts type tumor
		in Pipe lactiferous (87%)
G	Grades of best cancer	1 - poorly differentiated tumor (67%)
		0 - moderately differentiated tumor (33%)
E	Estrogen receptor positive	1 - yes(73%), 0 - no(37%)
P	Progesterone receptor positive	1 - yes(75%), 0 - no(25%)
H	The human epidermal growth factor receptor	1 - HER2-positive breast cancer(73%)
		0 - there is no antigen in tissue(27%)
K	Proliferative activity of cells	1 - greater than 15(60%)
		0 - ki67 less than 15 (60%)
S	Operation removal of the tumor or breast	1 - mastectomy or Lumpectomy (56%)
		0 - excisional biobsy (44%)
T	Advanced type of tumor	1 - the size of the main tumor
		more than 3 cm(68%), 0 - otherwise (32%)
L	Lymph nodes	1 - tumor spreading
		to the lymph node(82%) , 0 - no (18%)
M	Metastasis	1 - distant metastasis(92%)
		0 - No distant metastases (8%)

We are interested in the dependence between left set of variables T, L, M, which indicate the severity of the disease, and the test results which belong to the right set.

Table 3. The most significant canonical super-symptoms: $\mathrm{R}_1 = \mathrm{P}(1 + \mathrm{ES})$, $\mathrm{L}_1 = \mathrm{ML}$; $\mathrm{R}_2 = \mathrm{G}(1 + \mathrm{A}(1 + \mathrm{K}))$, $\mathrm{L}_2 = \mathrm{M}(1 + \mathrm{TL})$ $\mathrm{R}_3 = \mathrm{HS}(1 + \mathrm{D})$, $\mathrm{L}_3 = \mathrm{L}(1 + \mathrm{M})$, over $\mathbb{F}_2$.

	$(\mathrm{R}_1, \mathrm{L}_1)$	$(\mathrm{R}_2, \mathrm{L}_2)$	$(\mathrm{R}_3, \mathrm{L}_3)$
Fisher Exact Test, p	0.000003	0.027	0.003
Sensitivity, %	51	25	100
Accuracy, %	63	66	66
Specificity, %	100	95	64

The significant super-symptoms are presented in the Table 3. Sensitivity is the True Positive Rate (TPR), Specificity is the True Negative Rate (TNR), Accuracy is equal to the proportion of correctly classified observations. This table presents the results and analysis done for this study, the symptom analysis is used to identify the most significant combination of factors for predicting breast cancer severity.

The first canonical left factor $L_1 = ML$ means the presence of distant metastases and the tumor spreading to the lymph nodes and the first canonical right factor $R_1 = P(1 + ES)(\mathrm{mod}2)$ means practically that progesterone receptor was positive and as a surgical procedure the excisional biopsy was performed. When

$R_1 = P(1 + ES)(\mathrm{mod}2) = 1$ then $L_1 = ML = 1$ with probability equal to 1, but when $R_1 = 0$ then $L_1 = 1$ with probability equal to 0.6. In this case we have a smaller $p = 0.000003$ of Fishers Exact Test.
The second canonical left factor $L_2 = M(1 + TL)$ means presence of distant metastases and the small tumor non spreading to the lymph nodes, which can match the lateral breast cancer. This can be explained by the fact that the second canonical right factor $R_2 = G(1 + A(1 + K))$ that highlights a small group of 9 patients with poorly differentiated tumor and which are of older age or have high proliferative activity of cells. In this group probability of the lateral breast cancer is equal to 0.78, while in the rest this probability is equal to 0.36, $p = 0.027$.
The third canonical left factor $L_3 = L(1 + M)$ means that the tumor was spreading to the lymph nodes but without do distant metastases. This probably corresponds to an early stage aggressive cancer. There are only six such patients and all of them have the third canonical right factor $R_3 = HS(1 + D) = 1$ that means surgery and combination of IDC type with tests positive for a protein called human epidermal growth factor receptor 2 (HER2). Among the rest, this factor R_3 occurs with probability to 0.36. In addition, the third canonical pair has maximal Sensitivity 100% and comparable Accuracy 66% but lower Specificity.
The study shows that the precision (accuracy) for the prediction analysis of breast cancer data is acceptable and can help physicians in decision making for early diagnosis.

4. Conclusions

In the case of multidimensional data analysis, when the individual factors are insignificant, it is possible to detect a risk group with a special combination of factors. That the search method described in the article is based on the parameterization of all possible combinations of factors by polynomials over $\mathbb{F}_2$. Examples of symptom survival analysis and canonical symptom analysis show that if is not limited to the analysis of one-dimensional samples, interesting patterns can be obtained.

Acknowledgement
The work was carried out with the support of the RFBR, project in (Grant No. 20-01-00096-a).

References

[1] N. Alexeyeva, Analysis of biomedical systems. Reciprocity. Ergodicity. Synonymy, Saint-Petersbeurg: Publishing jf the Saint-Petersburg, 2013.

[2] N. Alexeyeva, P. Gracheva, E. Podkhalyuzina and K. Usevich, "Proceedings, 3rd International Conference on BioMedical Engineering and Informatics BMEI," in *Symptom and syndrome analysis of categorial series, logical principles and forms of logic*, China, 2010.

[3] C. R. Mehta and N. R. Patel, *IBM SPSS Exact tests*, IBM Corporation, 2011.

[4] N. Alexeyeva, A. Alexeyev, E. Verbitskaya and E. Krupitsky, "Final geometry and a logic principle of projectivity in the statistical analysis of the medical data," *Bulletin of the international statistical institute LXII-2007*, pp. 3021-3024, 2009.

[5] A. Afifi and S. Azen, The Analysis of Variance, New York: Academic Press, 1972.

[6] E. A. Gehan, "Statistical methods for survival time studies," in *Cancer Therapy: Prognostic Factors and Criteria. M. J. Staquet*, New York, Raven Press, 1975, pp. 7-35.

[7] N. Alexeyeva, P. Gracheva, B. Martynov and I. Smirnov, "The 2nd International Conference on BioMedical Engineering and Informatics (BMEI'09)," in *The finitely geometric symptom analysis in the glioma survival*, China, 2009.