Statistical methods of SNP data analysis with applications

Alexander Bulinski, Oleg Butkovsky, Alexey Shashkin, Pavel Yaskov

Statistical methods of SNP data analysis with applications

A.V.Bulinski[1,2], **O.A.Butkovsky**[1], **A.P.Shashkin**[1], **P.A.Yaskov**[1,3]

Abstract

Various statistical methods important for genetic analysis are considered and developed. Namely, we concentrate on the multifactor dimensionality reduction, logic regression, random forests and stochastic gradient boosting. These methods and their new modifications, e.g., the MDR method with "independent rule", are used to study the risk of complex diseases such as cardiovascular ones. The roles of certain combinations of single nucleotide polymorphisms and external risk factors are examined. To perform the data analysis concerning the ischemic heart disease and myocardial infarction the supercomputer SKIF "Chebyshev" of the Lomonosov Moscow State University was employed.

Keywords and phrases: Genetic data statistical analysis; multifactor dimensionality reduction; logic regression; random forests; stochastic gradient boosting; independent rule; single nucleotide polymorphisms; external factors; ischemic heart disease; myocardial infarction; supercomputer.

AMS 2010 classification: 92B15, 92D10, 65C20.

1 Introduction

The detection of genetic susceptibility to complex diseases (such as cardiovascular, oncological ones etc.) has recently drawn much attention in many leading research centers, see, e.g., [6] and [43]. According to the forecast of the World Health Organization (www.who.int), in 2030 the deaths related to cardiovascular diseases will exceed 23 millions (this year about 17 millions), the oncological diseases will take the lives of more than 11 millions of our planet inhabitants and at least 2 millions of people will be the victims of the diabetes. Thus this research domain is important since one expects to provide for each person the prophylactic measures and medical treatment taking into account his/her genetic peculiarities which increase the risk of some diseases and protect from the others, see, e.g., [28]. Individual's DNA variations are typically described in terms of *single nucleotide polymorphisms* (SNP), i.e. the fragments of genetic code where a nucleotide change is possible. For more details see, e.g., [46]. The first examples of genetically based diseases (e.g., sicklemia) were related with a single mutation. Contrariwise many hard diseases such as diabetes, Alzheimer's disease and others have a complex character as they

[1]Dept. of Mathematics and Mechanics, Lomonosov Moscow State University, Moscow, 119991, Russia.
[2]LPMA, Université Pierre et Marie Curie (Paris-6), 4, place Jussieu 75252, Paris Cedex 05, France.
[3]Steklov Mathematical Institute, Gubkina str. 8, Moscow, 119991, Russia.

can be provoked by mutations in different parts of the DNA code which are responsible for the formation of certain types of proteins. Quite a number of recent studies (see, e.g., [14], [41] and [40]) support the paradigm that the increasing risks of complex diseases can be explained by combinations of certain SNP whereas separate mutations have no dangerous effects.

Thereupon it should be mentioned that there existed a longstanding demand for statistical analysis of biological and medical data. However, only in the first part of the 20th century, due to the classical contributions by K.Pirson, R.Fisher, H.Cramér, A.Kolmogorov, N.Smirnov, A.Wald and other prominent statisticians, the essential progress was achieved both in theory and applications. The methods developed were sufficient, e.g., for investigation of the efficiency of new medicaments. The situation has changed radically at the beginning of the 2000's when the laboratory methods of DNA analysis provided the data related to the personal human code structure. The achievements in decoding of the human genome have led to formation of vast data bases in the frameworks of the International Research projects, see, e.g., GAW16 [17] and HapMap [20]. Note also that software engineering plays an important role in such studies, see, e.g., [23] and [49]. The cost of genomic analysis has fallen considerably in the last 10 years, allowing to collect large volumes of genetic data for genetic mapping of complex diseases. However statistical problems arising here require new methods of inference rather than classical ones. Indeed, the modern statistical models involve huge number of variables, parameters, hypotheses etc., while the sample size is usually moderate (several hundred or sometimes several thousand of observations, see, e.g., [24]). The sample design is limited both by costs of analysis which are still high and by difficulties due to the sample selection. In particular, the ethnic homogeneity should be taken into account, as well as the influence of external risk factors such as obesity, smoking etc.

To perform reliable statistical inference, it is necessary to apply new powerful tools developed in high-dimensional statistics, artificial intelligence, information retrieval, econometrics etc. Some of them have been adapted and further generalized in numerous papers by biostatisticians. Among the most important SNP analysis methods are the multifactor dimensionality reduction (MDR), logic regression (LR), random forests (RF) and stochastic gradient boosting (SGB). All approaches based on these methods do not impose any strong restrictions on the dependence structure of variables under consideration (apart from independence and identical distribution of observations within certain groups). Thus a broad class of statistical models is defined and the model providing the best out-of-sample fit is selected.

If one deals with too many parameters, overfitting is likely to happen, i.e. the estimated parameters depend too much on the given sample. As a result the constructed estimators give poor prediction on new data. On the other hand, application of a very sophisticated model may not capture the studied dependence structure of various factors efficiently. However the trade-off between the model's complexity and its predictive power allows to perform reliable statistical inference via new model validation techniques (see, e.g., [2] and [30]). The main tool of model selection is the *cross-validation*, see, e.g., [48]. Its idea is to estimate parameters by involving only a part of the sample (*training sample*) and afterwards use the remaining observations (*test sample*) to test the predictive power of the obtained estimates. Then an average over several realizations of randomly chosen

training and test samples is taken, see [21].

There are two closely connected research directions in genomic statistics. The first one is aimed at the disease risk estimation when the genetic portrait of a person is known (in turn this problem involves estimation of disease probability and classification of genetic data into high and low risk domains, see, e.g., [30]). The second trend is to identify relevant combinations of SNPs having the most significant pathogenic (or, in other way, protective) influence. Both directions are presented in this paper. Moreover, the authors propose further development of various statistical methods and apply them to study of the risks of cardiovascular diseases. For this purpose the new software concerning the employment of the mentioned statistical methods was designed and used.

Due to high-dimensionality of data many numerical procedures based on the above mentioned statistical methods are very time consuming. The authors are grateful to the Chancellor of the Lomonosov Moscow State University (MSU) Professor V.A.Sadovnichy and to the Deputy Director of the MSU Research Computing Center Professor V.V.Voevodin for the opportunity to use the supercomputer SKIF MSU "Chebyshev".

This investigation was started in the framework of the project headed by Professor V.A.Tkachuk, the Dean of the Faculty of the Fundamental Medicine of the MSU. An overview of preliminary results of the work was presented at the International conference "Postgenomic methods of analysis in biology, and laboratory and clinical medicine" in the talk by Professor A.V.Bulinski (see [9] and [10]).

2 Methods

We start with some definitions. Let N be the the number of patients in the sample and the vector $X^j = \left(X_1^j, \ldots, X_n^j\right)$ consist of genetic (SNP) and external risk factors of j-th individual, $j = 1, ..., N$. Here n is the total number of factors, and X_i^j is the value of i-th variable (characterizing SNP or external factor) of j-th individual. These variables are also called *explanatory variables* or *predictors*. If X_i stands for an SNP, we set

$$X_i = \begin{cases} 0, & \text{no mutation in } i\text{-th SNP}, \\ 1, & \text{heterozygous mutation}, \\ 2, & \text{homozygous mutation}. \end{cases} \tag{1}$$

We assume that the external risk factors also take no more than three values, denoted by 0, 1 and 2. For example, we can specify a presence or an absence of obesity (or hypercholesterolemia etc.) by values 1 and 0 respectively. If the external factor takes more values (e.g., blood pressure), we can divide individuals into three groups according to its values.

Further on $X_1^j, \ldots, X_m^j$ stand for genetic data and $X_{m+1}^j, \ldots, X_n^j$ for external risk factors. Let a binary variable Y^j (*response variable*) be equal to 1 for a *case*, i.e. whenever j-th individual is diseased, and to -1 otherwise (that is for a *control*). Set

$$\xi = (\xi^1, \ldots, \xi^N) \text{ where } \xi^j = (X^j, Y^j), \quad j = 1, \ldots, N. \tag{2}$$

Suppose $\xi^1, \ldots, \xi^N$ are i.i.d. discrete random vectors having the same law as a vector (X, Y) and independent of this vector. Assume that $X = (X_1, \ldots, X_n)$. All random

vectors (and random variables) are considered on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$, E denotes the integration w.r.t. P.

The main problem is to find a function in genetic and external risk factors describing the phenotype (that is the individual healthy or sick) in the best way.

2.1 Prediction algorithms

Let $\mathcal{X} := \{0, 1, 2\}^n$ denote the space of all possible values of explanatory variables. Any function $f : \mathcal{X} \to \{-1, 1\}$ will be called a *theoretical prediction function*. Define the *balanced* or *normalized prediction error* for the theoretical prediction function f as

$$Err(f) := \mathsf{E}|Y - f(X)|\psi(Y)$$

where the *penalty function* $\psi : \{-1, 1\} \to \mathbb{R}_+$. Obviously,

$$Err(f) = 2\psi(-1)\mathsf{P}(f(X) = 1, Y = -1) + 2\psi(1)\mathsf{P}(f(X) = -1, Y = 1). \tag{3}$$

Clearly $Err(f)$ depends also on the law of (X, Y). Following [45] and [48] we put

$$\psi(y) = \frac{1}{4\mathsf{P}(Y = y)},\ y \in \{-1, 1\}, \tag{4}$$

the trivial cases $\mathsf{P}(Y = -1) = 0$ and $\mathsf{P}(Y = 1) = 0$ are excluded. Then

$$Err(f) = \frac{1}{2}\mathsf{P}(f(X) = 1|Y = -1) + \frac{1}{2}\mathsf{P}(f(X) = -1|Y = 1). \tag{5}$$

For a *balanced* sample considered in [36], $\mathsf{P}(Y = -1) = \mathsf{P}(Y = 1) = 1/2$ and $Err(f) = \mathsf{E}|Y - f(X)|/2$ is equal to the *classification error* $\mathsf{P}(Y \neq f(X))$.

The reason to consider this weighted scheme is that a misclassification in a more rare class should be taken into account with a greater weight. Otherwise, if the probability of disease $\mathsf{P}(Y = 1)$ is small, then the trivial function $f(x) \equiv -1$ may have the least prediction error. The approach to calculation of the prediction error based on penalty functions is not the only one possible. Nevertheless Velez et al. [45] showed that for models with high computational costs it outperforms substantially other methods such as over- and undersampling.

It is easy to prove that the *optimal* theoretical prediction function minimizing the balanced prediction error is given by

$$f^*(x) = \begin{cases} 1, & p(x) > \mathsf{P}(Y = 1), \\ -1, & \text{otherwise.} \end{cases} \tag{6}$$

where

$$p(x) = \mathsf{P}(Y = 1|X = x),\ \ x \in \mathcal{X}. \tag{7}$$

Then each multilocus genotype (with added external risk factors) $x \in \mathcal{X}$ is classified as high-risk if $f^*(x) = 1$ or low-risk if $f^*(x) = -1$.

Since $p(x)$ and $\mathsf{P}(Y = 1)$ are unknown, the immediate application of (6) is not possible. Thus we try to find an approximation of unknown function f^* using a *prediction algorithm*

that is a function $f_{PA} = f_{PA}(x, \xi(S))$ with values in $\{-1, 1\}$ (recall that $Y \in \{-1, 1\}$ a.s.) which depends on $x \in \mathcal{X}$ and the sample

$$\xi(S) = \{\xi^j,\ j \in S\} \ \text{ where } \ S \subset \{1, \dots, N\}. \tag{8}$$

The simplest way is to employ formula (6) with $p(x)$ and $\mathsf{P}(Y = 1)$ replaced by their statistical estimates. For example introduce

$$\widehat{p}(x, \xi(S)) = \frac{\sum_{j \in S} I\{Y^j = 1, X^j = x\}}{\sum_{j \in S} I\{X^j = x\}}, \ \ x \in \mathcal{X}, \tag{9}$$

and take

$$\widehat{\mathsf{P}}_S(Y = 1) = \frac{1}{\sharp S} \sum_{j \in S} I\{Y^j = 1\} \tag{10}$$

where $I\{A\}$ stands for the indicator of an event A and $\sharp D$ denotes the cardinality of a finite set D.

Along with (10) we will consider

$$\widehat{\mathsf{P}}_S(Y = 1 | X \in C) = \frac{\sum_{j \in S} I\{Y^j = 1, X^j \in C\}}{\sum_{j \in S} I\{X^j \in C\}}, \ \ C \subset \mathcal{X}. \tag{11}$$

Thus (9) is a special case of (11) for $C = \{x\}$ with $x \in \mathcal{X}$. Note that more difficult way is to search for the estimators of f^* using several subsamples of ξ.

Assume that we constructed a prediction algorithm f_{PA}. Then taking in mind (5) set

$$Err(f_{PA}(\cdot, \xi(S))) = \frac{1}{2} \sum_{y \in \{-1,1\}} \mathsf{P}\left(f_{PA}(X, \xi(S)) \neq y | Y = y\right). \tag{12}$$

As a law of (X, Y) is unknown one can only construct an estimate $\widehat{Err}(f_{PA}(\cdot, \xi(S)))$ of $Err(f_{PA}(\cdot, \xi(S)))$. In Section 3 we use the *estimated prediction error* of a prediction algorithm f_{PA} which is based on K-fold cross-validation and has the form

$$\widehat{Err}_K(f_{PA}(\cdot, \xi), \xi) = \frac{1}{2} \sum_{y \in \{-1,1\}} \frac{1}{K} \sum_{k=1}^{K} \frac{\sum\limits_{j \in S_k} I\left\{f_{PA}(X^j, \xi(\overline{S_k})) \neq y, Y^j = y\right\}}{\sum\limits_{j \in S_k} I\{Y^j = y\}} \tag{13}$$

where

$$S_k = \left\{(k-1)\left[\frac{N}{K}\right] + 1,\ \dots,\ k\left[\frac{N}{K}\right] I\{k < K\} + N I\{k = K\}\right\}, \tag{14}$$

$\overline{S_k} = \{1, \dots, N\} \setminus S_k$ and $[a]$ is the integer part of $a \in \mathbb{R}$.

A very important problem is to make sure that the prediction algorithm f_{PA} gives statistically reliable results. The quality of an algorithm is determined by its prediction error (12) which is unknown and therefore the inference is based on consistent estimates of this error. Clearly the high quality of an algorithm means that it captures the dependence between predictors and response variables, so the error is made more rarely than it would be if these variables were independent. Consider a null hypothesis H_0 that X and Y are

independent. If they are in fact dependent, then for any significant prediction algorithm f_{PA} an appropriate test procedure involving f_{PA} should reject H_0 at the approximate significance level α, e.g., 5%. Intuitively, this shows that results of the algorithm could not be obtained by chance. For such a procedure, we take a *permutation test* (see [18]). Its idea is as follows.

Permutation test for a given statistic $\widehat{L}(\xi)$ (we consider $\widehat{L}(\xi) = \widehat{Err}(f_{PA}(\cdot, \xi))$ is done by the following steps.

1. Generate B independent random vectors $(\pi_1^b, \ldots, \pi_N^b)$, $1 \le b \le B$, with the uniform distribution over all permutations Π_N of $1, \ldots, N$.
2. Compute $\widehat{Err}_{K,b} = \widehat{Err}_K(f_{PA}, \bar{\xi}_b)$, $1 \le b \le B$, with
$$\bar{\xi}_b = \big(\big(X^1, Y^{\pi_1^b}\big), \ldots, \big(X^N, Y^{\pi_N^b}\big)\big).$$
3. Find the *Monte Carlo* p*-value* (see, e.g., [27, p. 63]):
$$\widehat{\mathsf{p}} = \widehat{F}\big(\widehat{Err}_K(f_{PA}, \xi)\big) \tag{15}$$
where $\widehat{F} = \widehat{F}(z)$ is the empirical cumulative distribution function (c.d.f.) defined by the relation
$$\widehat{F}(z) = \frac{1}{B}\sum_{b=1}^{B} I\big\{\widehat{Err}_{K,b} \le z\big\}, \quad z \in \mathbb{R}.$$
4. If $\widehat{\mathsf{p}} < \alpha$, reject H_0, otherwise not.

According to [18], one ideally has to use all permutations belonging to Π_N but this is impractical in view of computational costs. Thus the Monte Carlo approximations for the *true* p*-value* $\mathsf{p} = F\big(\widehat{Err}_K(f_{PA}, \xi)\big)$ are employed, here F is the c.d.f. of $\widehat{Err}_{K,b}$. The upper bound for $|\mathsf{p} - \widehat{\mathsf{p}}|$ is $1/2\sqrt{B}$ (see [18]). This could be used to determine the number B of simulations for a desired accuracy.

Note also that if the estimate of the error function for the algorithm $f_{PA}(\cdot, \xi)$ is *asymptotically optimal*, i.e. converges in probability to the error of the optimal prediction function f^* as $N \to \infty$ (ξ depends on N), then the rule of thumb is to suspect overfitting if $\widehat{Err}_K(f_{PA}, \xi)$ is close to $1/2$, which is a probability limit of this error under H_0 as $N \to \infty$.

We use complementary approaches to analyze dataset related to complex diseases. Each approach (MDR, LR and machine learning) is characterized by its own way of constructing prediction algorithms. For each method one or several prediction algorithms admitting the least estimated prediction error are found (a typical situation is that there are several ones with almost the same estimated prediction error). These prediction algorithms provide a way to determine the domains where the disease risk is high or low (depending on the value of the corresponding prediction function). It is also possible to select combinations of SNPs and external risk factors whose presence influences the liability to disease to a great extent. Some methods allow to present such combinations immediately. Others, which employ more complicated forms of dependence between explanatory and response variables, need further analysis based on modifications of permutation tests.

Now we pass to the description of various statistical methods and their applications to the cardiovascular risk detection.

2.2 Multifactor dimensionality reduction

Ritchie et al. [36] introduced *multifactor dimensionality reduction* (MDR) as a new method of analyzing gene-gene and gene-environment interactions. Rather soon the method became very popular. Since the first publication more than 200 papers applying MDR in genetic studies were written (see, e.g., references in [45]).

MDR is a flexible non-parametric method not depending on a particular inheritance model. We give a rigorous description of the method following ideas of [36] and [45]. As mentioned earlier, the probability $p(x)$ introduced in (7) is unknown. To find its estimate one can apply maximum likelihood approach assuming that the random variable $I\{Y=1\}$ conditionally on $X=x$ has a Bernoulli distribution with unknown parameter $p(x)$. Then we come to (9).

A direct calculation of estimate in (9) with exhaustive search over all possible values of x is highly inefficient, since the number of different values of x grows exponentially with number of risk factors. Moreover, such a search leads to overfitting. Instead, it is usually supposed that $p(x)$ non-trivially depends not on all, but certain variables x_i. That is, there exist $l \in \mathbb{N}$, $l < n$, and $(k_1^*, \dots, k_l^*)$, where $1 \le k_1^* < \dots < k_l^* \le n$, such that for each $x = (x_1, \dots, x_n) \in \mathcal{X}$, the following relation holds:

$$p(x) = \mathsf{P}(Y = 1 | X_{k_1^*} = x_{k_1^*}, ..., X_{k_l^*} = x_{k_l^*}). \tag{16}$$

In other words only few factors influence the disease and the others can be neglected. A minimal combination of factors $(X_{k_1^*}, \dots, X_{k_l^*})$ in formula (16) is called *the most significant*. Clearly it is the most significant combination which has the least prediction error. Indeed, if we consider any other combination of pairwise different indices $k_1, \dots, k_r$ and set

$$f_{k_1,\dots,k_r}(x) = \begin{cases} 1, & \mathsf{P}(Y = 1 | X_{k_1} = x_{k_1}, ..., X_{k_r} = x_{k_r}) > \mathsf{P}(Y = 1), \\ -1, & \text{otherwise}, \end{cases}$$

then we obviously have

$$Err\left(f_{k_1^*,\dots,k_l^*}\right) \le Err(f_{k_1,\dots,k_r}) \tag{17}$$

where $Err(f)$ is calculated according to (5).

To choose the most significant combination, exhaustive search over all possible combinations of factors is applied. For each $\{k_1, \dots, k_r\} \subset \{1, \dots, n\}$ and any $x \in \mathcal{X}$ consider

$$C_{k_1,\dots,k_r}(x) = \{u = (u_1, \dots, u_n) \in \mathcal{X} : u_{k_i} = x_{k_i},\, i = 1, \dots, r\}$$

and for S appearing in (8) define a prediction algorithm (cf. (6)) by

$$\widehat{f}_{k_1,\dots,k_r}(x, \xi(S)) := \begin{cases} 1, & \widehat{\mathsf{P}}_S(Y = 1 | X \in C_{k_1,\dots,k_r}(x)) > \widehat{\mathsf{P}}_S(Y = 1), \\ -1, & \text{otherwise}, \end{cases} \tag{18}$$

here we use formulas (10) and (11). It is easy to show that $\widehat{Err}_K(\widehat{f}_{k_1,\dots,k_r}, \xi) \to Err(f_{k_1,\dots,k_r})$ in probability as $N \to \infty$ (ξ depends on N). Consequently, (17) implies that, for any $\varepsilon > 0$ and all N large enough, with probability close to 1 one has

$$\widehat{Err}_K(\widehat{f}_{k_1^*,\dots,k_l^*}, \xi) < \widehat{Err}_K(\widehat{f}_{k_1,\dots,k_r}, \xi) + \varepsilon.$$

Hence, it is natural to pick one or a few combinations of factors with the smallest empirical prediction errors as an approximation for the most significant combination.

The last step in MDR is to determine statistical significance of the results. Here we test a null hypothesis of independence between X and Y i.e. between risk factors X and a disease Y. This can be done via the permutation test described in Section 2.1.

MDR method with "independent rule". We propose *multifactor dimensionality reduction* with *"independent rule"* (MDRIR) method to improve the estimate of probability $p(x)$. This approach is motivated by Park [35], who deals with classification of large array of binary data. The principal difficulty with employment of formula (9) is that the number of observations in numerator and denominator of the formula might be small even for large N (see, e.g., [26]). This can lead to inaccurate estimates and finally to a wrong prediction algorithm. Moreover, for some samples the denominator of (9) can equal zero.

The Bayes formula implies that

$$p(x) = \frac{\mathsf{P}(X = x|Y = 1)\mathsf{P}(Y = 1)}{\mathsf{P}(X = x|Y = 1)\mathsf{P}(Y = 1) + \mathsf{P}(X = x|Y = -1)\mathsf{P}(Y = -1)}, \tag{19}$$

where the trivial cases $\mathsf{P}(Y = -1) = 0$ and $\mathsf{P}(Y = 1) = 0$ are excluded. Substituting (19) into (6) we obtain the following expression for prediction function:

$$f^*(x) = \begin{cases} 1, & \mathsf{P}(X = x|Y = 1) > \mathsf{P}(X = x|Y = -1), \\ -1, & \text{otherwise.} \end{cases} \tag{20}$$

As in standard MDR method described above, we will assume that formula (16) holds. It was proved in [35] that for a broad class of models (e.g., *Bahadur model* [3], *logit model* [13]) the conditional probability $\mathsf{P}\left(X_{k_1} = x_1, \ldots, X_{k_r} = x_r \,|Y = y\right)$, where $y = \pm 1$, can be estimated in the following way:

$$\widehat{\mathsf{P}}_S\left(X_{k_1} = x_1, \ldots, X_{k_r} = x_r \,|Y = y\right) = \prod_{i=1}^{r} \widehat{\mathsf{P}}_S\left(X_{k_i} = x_i \,|Y = y\right), \tag{21}$$

here (cf. (11))

$$\widehat{\mathsf{P}}_S\left(X_{k_i} = x \,|Y = y\right) = \frac{\sum\limits_{j \in S} I\{X_{k_i}^j = x, Y^j = y\}}{\sum\limits_{j \in S} I\{Y^j = y\}}. \tag{22}$$

Combining (16), (20), (21) and (22) we find the desired estimate of $f^*(x)$.

A number of observations in numerator and denominator of (22) increases considerably comparing with (18). It allows to estimate the conditional probability more precisely whenever the estimate introduced in (21) is reasonable.

Thus, as opposed to standard MDR method, MDRIR uses alternative estimates of conditional probabilities. All other steps (prediction algorithm construction, prediction error calculation) remain the same. Let us mention that as far as we know this modification of MDR has not been applied before. It is based on a combination of the original MDR method [36] and the ideas of [35].

2.3 Logic regression

The *logic regression* (LR) was proposed in [38]. Further generalizations are given in [16], [25], [40], [39] and other works. LR is based on the classical binary logistic regression [22] and exhaustive search for relevant predictor combinations. The main difficulty is to organize the search efficiently. The LR method was applied to identification of the most significant SNP combinations in [1], [33] and [40]. Note that for genetic analysis it is convenient to use explanatory variables taking 3 values. Thus we employ *ternary variables*, whereas the authors of the above-mentioned papers employ binary ones.

Let $p(x)$ be the conditional probability of a disease defined in (7). We suppose that trivial situations when $p(x) \in \{0, 1\}$ do not occur and omit them from the consideration. To estimate $p(x)$ we pass now to the *logistic transform*

$$q(x) = \lambda\left(p(x)\right) \tag{23}$$

where $\lambda(z) = \ln(z/(1-z))$, $z \in (0,1)$, is the *inverse logistic function*. The *logistic function* itself equals to $\Lambda(t) = (1+e^{-t})^{-1}$, $t \in \mathbb{R}$. Note that we are going to estimate the unknown disease probability with the help on linear statistics with appropriately selected coefficients. Therefore it is natural to avoid restrictions on possible values of the function estimated. Thus the logistic transform is convenient, because $p(x) \in (0,1)$ for $x \in \mathcal{X}$ while $q(x)$ can take all real values.

Consider a class $\mathcal{G}$ of all real-valued functions in ternary variables $x_1, \ldots, x_n$. We call a *model* of the dependence between the disease and explanatory variables any subclass $\mathcal{M} \subset \mathcal{G}$. Set

$$\widehat{\psi}(y, \xi(S)) = \frac{1}{4\widehat{\mathsf{P}}_S(Y=y)}, \;\; y \in \{-1, 1\},$$

here $\widehat{\mathsf{P}}_S(Y=y)$ was introduced in (10). Define the *normalized smoothed score function*

$$L(h, \xi(S)) = \frac{1}{\sharp S} \sum_{j \in S} \phi(-Y^j h(X^j)) \widehat{\psi}(Y^j, \xi(S)) \tag{24}$$

where S is introduced in (8), $\phi(t) = \log_2(1+e^t)$ for $t \in \mathbb{R}$, and $h \in \mathcal{M}$. In contrast to previous works our version of LR scheme involves normalization (cf. (3)), i.e. taking the observations with weights dependent on the proportion of cases and controls in subsample $\xi(S)$.

An easy computation yields that $\arg\min_{h \in \mathcal{M}} L(h, \xi(S))$ equals to

$$\arg\max_{h \in \mathcal{M}} \frac{1}{\sharp S} \sum_{j \in S} \left(\ln \Lambda(h(X^j)) \frac{I\{Y^j = 1\}}{2\widehat{P}_S(Y=1)} + \ln(1 - \Lambda(h(X^j))) \frac{I\{Y^j = -1\}}{2\widehat{P}_S(Y=-1)} \right).$$

That is, minimizing the score function is equivalent to the search of normalized maximal likelihood estimate of q. Note that estimating the disease probability in this setup is closely connected with the problem of data classification, i.e. predicting the disease by the value of $x \in \mathcal{X}$. Recall that in standard classification problem instead of the score function (24) one uses the following normalized estimate of the error probability

$$\widetilde{L}(h, \xi(S)) = \frac{1}{\sharp S} \sum_{j \in S} I\{Y^j h(X^j) < 0\} \widehat{\psi}(Y^j, \xi(S)).$$

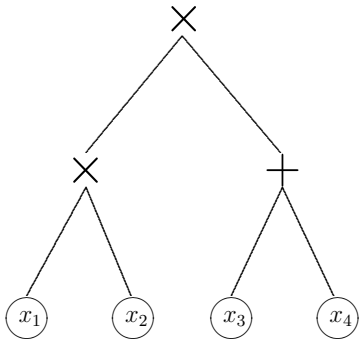

Figure 1: A tree T representing a function $T(x) = (x_1 \times x_2) \times (x_3 + x_4)$.

In fact the optimal choice of h for these problems coincide if the underlying model $\mathcal{M}$ is correctly specified (i.e. $q \in \mathcal{M}$), see [5]. However the usage of score function L has an important advantage over $\widetilde{L}$ because one has to evaluate the minimum of a smooth function.

A wide and easy to handle class of models is obtained by taking functions linear in variables $x_1, \ldots, x_n$ or in their products. In turn these functions admit a convenient representation by elementary polynomials. Recall that an *elementary polynomial* (EP) is a function T in ternary variables $x_1, \ldots, x_n$ belonging to $\{0, 1, 2\}$ which can be represented as a finite sum of products $x_1^{u_1} \ldots x_n^{u_n}$ where $u_1, \ldots, u_n \in \mathbb{Z}_+$. The addition and multiplication of ternary variables is considered by modulo 3. Any EP can be represented as a *binary tree*[4] in which *knots* (vertices which are not *leaves*) contain either addition or multiplication sign, and each leaf corresponds to a variable. Figure 1 provides an example of a binary tree. Different trees may correspond to the same EP, thus this relation is not one-to-one. However, it does not influence our problem, so we regain the notation T for a tree. A finite set of trees $F = (T_1, \ldots, T_s)$ is called a *forest*. For a tree T, its *complexity* $C(T)$ is the number of leaves. The complexity $C(F)$ of a forest F is the maximal complexity of trees constituting F.

It is clear that if $g \in \mathcal{G}$ then there exists $s \geq 1$ such that g has the following form:

$$g(x_1, \ldots, x_n) = \beta_0 + \sum_{i=1}^{s} \beta_i T_i(x_1, \ldots, x_n), \tag{25}$$

here $\beta_0, \beta_1, \ldots, \beta_s \in \mathbb{R}$ and $T_1, \ldots, T_s$ are EP.

Let us say that function g belongs to a class $\mathcal{G}_r(s)$, where $s, r \in \mathbb{N}$, if there exist a decomposition (25) of g such that all trees T_i $(i = 1, \ldots, s)$ have complexity less or equal r. We identify a function $g \in \mathcal{G}_r(s)$ with pair (F, β) where F is the corresponding forest and $\beta = (\beta_0, \ldots, \beta_s)$ is the vector of coefficients in (25).

Minimization of $L(h, \xi(S))$ defined by (24) over all functions $h \in \mathcal{M} \subset \mathcal{G}_r(s)$ is done in two alternating steps. First we find the optimal value of β while F is fixed (which is the minimization of a smooth function in several variables) and then we search for the best F. Here one uses stochastic algorithms, since the number of such forests increase rapidly

[4]For the basic concepts of the graph theory see, e.g., [7].

when the complexity r grows. For $s \in \mathbb{N}$, a forest $F = (T_1, \ldots, T_s)$ and a subsample $\xi(S)$ (see (8)) consider a prediction algorithm f^F_{LR} setting

$$f^F_{\mathrm{LR}}(x, \xi(S)) = \begin{cases} 1, & \widehat{h}(x) > 0, \\ -1, & \text{otherwise}, \end{cases}$$

where $\widehat{h} = (F, \widehat{\beta})$ and

$$\widehat{\beta} = \arg\min_{\beta} L\left(\beta_0 + \sum_{j=1}^{s} \beta_j T_j(\cdot), \xi(S)\right). \tag{26}$$

Define also the *normalized prediction error of a forest* $F = (T_1, \ldots, T_s)$ as

$$\widetilde{\varphi}(F) = \widehat{Err}_K(f^F_{\mathrm{LR}}(\cdot, \xi), \xi).$$

A subgraph B of a tree T is called a *branch* if it is itself a binary tree (i.e. it can be obtained by selecting one vertex of T together with its offspring). Sum and product signs standing in a knot of a tree are called *operations*, thus $*$ stands for sum or product. Following [38], call the tree $\widetilde{T}$ *a neighbor* of T if it is obtained from T via one and only one of the following transformations.

1. Changing one variable to another in a leaf of the tree T (*variable change*).
2. Replacing an operation in a knot of a tree T with another one, i.e. sum to product or vice versa (*operator change*).
3. Changing a branch of two leaves to one of these leaves (*deleting a leaf*).
4. Changing a leaf to a branch of two leaves, one of which contains the same variable as in initial leaf (*splitting a leaf*).
5. Replacing a branch $B_1 * B_2$ with the branch B_1 (*branch pruning*).
6. Changing a branch B to a branch $x_j * B$ (*branch growing*), here x_j is a variable.

Figure 2 depicts results of these operations applied to the tree T of Figure 1. We say that forests F and $\widetilde{F}$ are *neighbors* if they can be written as $F = \{T_1, T_2, \ldots, T_s\}$ and $\widetilde{F} = \{\widetilde{T}_1, T_2, \ldots, T_s\}$ where T_1 and $\widetilde{T}_1$ are neighbors. The neighborhood relation defines a finite connected graph on all forests of equal size s with complexity not exceeding r. To each vertex F of this graph we assign a number $\widetilde{\varphi}(F)$. To find the global minimum of a function defined on a finite graph we employ the *simulated annealing method* (see, e.g., [19], [32] and [37]). This method constructs some specified Markov process which takes values in the graph vertices and converges with high probability to the global minimum of the function. To avoid stalling at a local minimal point the process is allowed to pass with some small probability to a point F having greater value of $\widetilde{\varphi}(F)$ than current one. We propose a new modification of this method in which the output is the forest corresponding to the minimal value of a function $\widetilde{\varphi}(F)$ over all (randomly) visited points.

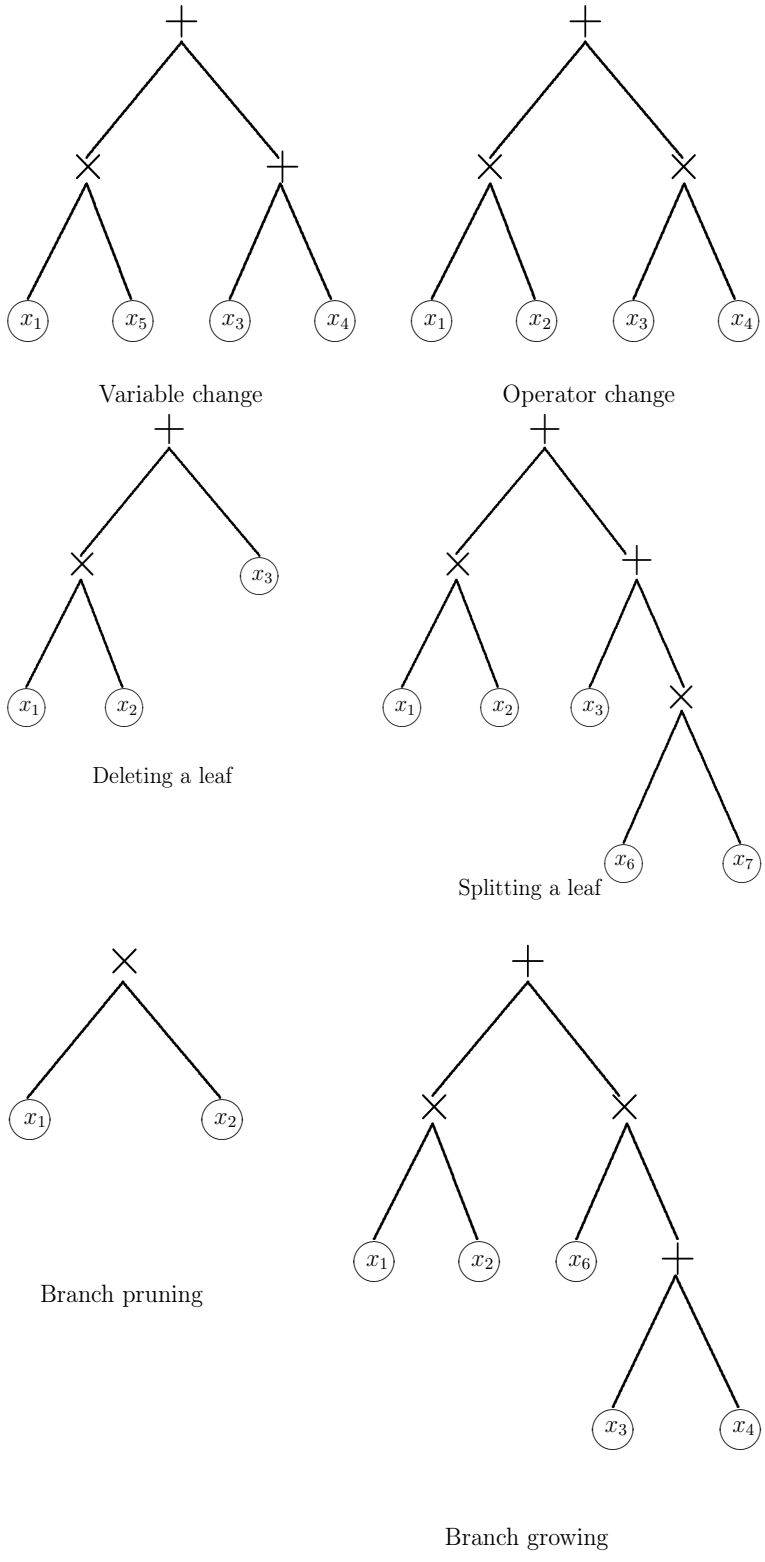

Figure 2: Neighbors of the tree T.

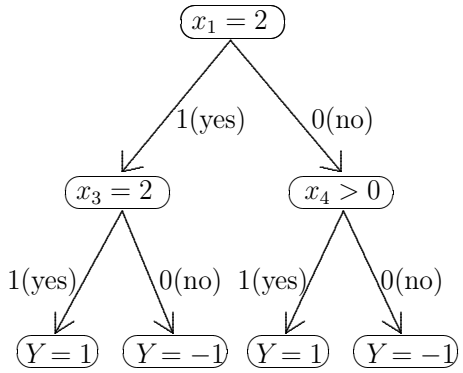

Figure 3: CART representing the prediction $((x_1 = 2)$ and $(x_3 = 2))$ or $(x_4 > 0)$.

2.4 Machine learning methods

Let us describe (see, e.g., [43]) two among the most popular machine learning methods – *random forests* (RF) and *stochastic gradient boosting* (SGB). They belong to *ensemble methods* which combine multiple predictions from a certain base algorithm to obtain better predictive power (i.e. less estimated prediction error). We use *classification and regression trees* (CART) for a base learning algorithm because it showed good performance in a number of studies (see [21]).

Classification tree T is a binary tree having the following structure. Any leaf of T contains either 1 or -1 and for any vertex P in T (including leaves) there exists a subset A_P of the explanatory variable space $\mathcal{X}$ such that the following properties hold:

1. $A_P = \mathcal{X}$ if P is the root of T;
2. if vertices P' and P'' are children for P, then $A_{P'} \cup A_{P''} = A_P$ and $A_{P'} \cap A_{P''} = \varnothing$.

In particular, subsets corresponding to the leaves form the partition of $\mathcal{X}$. To obtain a prediction of Y given a certain value $x \in \mathcal{X}$ of the random vector X, one should go along the path which starts from the root and ends in some leaf turning at each parent vertex P to that child P' for which $A_{P'}$ contains x. At the end of the x-specific path, one gets either 1 or -1 which serves as a prediction of Y. Figure 3 provides an example of a classification tree. Namely, the partition of $\mathcal{X}$ is formed by values of boolean functions standing in parent vertices. For each x starting from the root of the tree we calculate the value of a boolean function and move along the edge marked with the value obtained (1 or 0). The left child of the root corresponds to the subset $\{x \in \mathcal{X} : x_1 = 2\}$, while the right one to its complement in $\mathcal{X}$. Next, the leftmost leaf stands for a subset $\{x \in \mathcal{X} : x_1 = 2, x_3 = 2\}$, and if X falls in this subset, we predict that $Y = 1$; the rightmost leaf stands for a subset $\{x \in \mathcal{X} : x_1 < 2, x_4 = 0\}$, and if X takes values in this subset, we predict $Y = -1$.

Classification tree could be constructed via CART *algorithm* (if it is the case, we will call it CART). The algorithm proceeds iteratively. That is, on the l-th step of the algorithm ($l = 1, 2, \ldots$), each element A of the current partition $\mathcal{A}_l$ ($\mathcal{A}_1 = \mathcal{X}$) of the set $\mathcal{X}$ is divided into two disjoint parts

$$A^+(i,t) = \{(x_1, \ldots, x_n) \in A : x_i \le t\} \text{ and } A^-(i,t) = \{(x_1, \ldots, x_n) \in A : x_i > t\}$$

minimizing the sum $\widehat{G}(A^+(i,t)) + \widehat{G}(A^-(i,t))$ over $i = 1, \ldots, n$ and $t \in \{0, 1\}$. Here the *empirical Gini index*

$$\widehat{G}(C) = 2\widehat{\mathsf{P}}_S(Y = 1|X \in C)\big(1 - \widehat{\mathsf{P}}_S(Y = 1|X \in C)\big)$$

with $C \subset \mathcal{X}$ and $\widehat{\mathsf{P}}_S(Y = 1|X \in C)$ (see (11)) measures the heterogeneity of the subsample $\{j \in S : X^j \in C\}$ w.r.t. response variable Y. Any uninformative partition with

$$\min_{(i,t)}\big(\widehat{G}(A^+(i,t)) + \widehat{G}(A^-(i,t))\big) > \widehat{G}(A),$$

is not made.

The algorithm stops whenever a number of leaves D reaches some critical level which is chosen via some data-dependent criteria (see [21], page 308). For a subsample $\xi(S)$ of ξ, each CART defines a prediction algorithm

$$f(x, \xi(S)) = \sum_{d=1}^{D} a_d(\xi(S)) I\{x \in A_d(\xi(S))\} \tag{27}$$

where $\{A_1(\xi(S)), \ldots, A_D(\xi(S))\}$ is the partition of $\mathcal{X}$ corresponding to the leaves,

$$a_d(\xi(S)) = \begin{cases} 1, & \sharp\{j \in S\colon Y^j = 1, X^j \in A_d(\xi(S))\} > \sharp\{j \in S\colon Y^j = -1, X^j \in A_d(\xi(S))\}; \\ -1, & \text{otherwise.} \end{cases}$$

RF is a non-parametric method of estimating conditional probability $p(x)$. It was successfully applied to genetics data in a number of papers (see references in [43]). It could be briefly described as follows (see chapter 15 in [21] for details). Generate B bootstrap samples from the initial sample where one could choose $B = \max\{[N \log N], 1000\}$ according to [34]. For b-th bootstrap sample ($1 \le b \le B$) construct a CART prediction algorithm $f_b : \mathcal{X} \times (\mathcal{X} \times \{-1, 1\})^N \to \{-1, 1\}$ defined according to (27) and take

$$\widehat{p}_{\mathrm{RF}}(x, \xi(S)) = \big(B^{-1} \sum_{b=1}^{B} f_b(x, \xi(S)) + 1\big)/2$$

as an estimate of $p(x)$.

It is shown in [5] that generally RF method gives consistent estimates of $p(x)$ only if the number of partitions used in CART grows slower than the sample size. A final prediction algorithm $f_{\mathrm{RF}}(x, \xi(S))$ is constructed from the estimate $\widehat{p}_{\mathrm{RF}}(x, \xi(S))$ similarly to (6), i.e.

$$f_{\mathrm{RF}}(x, \xi(S)) = \begin{cases} 1, & \widehat{p}_{\mathrm{RF}}(x, \xi(S)) > \widehat{\mathsf{P}}_S(Y = 1), \\ -1, & \text{otherwise.} \end{cases}$$

The distinctive features of this method are low computational costs and the ability to extract relevant predictors when the number of irrelevant ones is large (see [4]).

SGB is another non-parametric method of estimating conditional probability $p(x)$. This method is used in a number of procedures for studying genetics data (see, e.g., [47]). SGB method can be described as follows ([15]).

1. Pass on the input of the algorithm[5] initial parameters $D, M \in \mathbb{N}$ and $\rho, \eta \in (0,1)$.

2. Put $m = 0$, $\xi_0(S) = \xi(S)$ and

$$f_0(x, \xi_0(S)) \equiv \frac{1}{2} \ln \frac{\widehat{\mathsf{P}}_S(Y=1)}{\widehat{\mathsf{P}}_S(Y=-1)}.$$

3. Increase m by 1 and define

$$\bar{Y}_m^j := \frac{2Y^j}{1 + \exp\{2Y^j f_{m-1}(X^j, \{\xi_l(S)\}_{l=0}^{m-1})\}}.$$

Choose a random subset in $\xi_m(S) = \{(X^j, \bar{Y}_m^j)\}_{j \in S}$ with $[\eta \sharp S]$ elements. Construct CART prediction algorithm (with D leaves) $\sum_{d=1}^{D} a_d^m(\xi_m(S)) I\{x \in A_d^m(\xi_m(S))\}$ on the chosen subset. Compute weight coefficients

$$w_d^m(\xi_m(S)) = \frac{\sum_{j \in J} \bar{Y}_m^j}{\sum_{j \in J} |\bar{Y}_m^j|(2 - |\bar{Y}_m^j|)}, \quad d = 1, \ldots, D,$$

where the random set $J = \{j : X^j \in A_d^m(\xi_m(S))\}$, and put

$$f_m = f_{m-1} + \rho \sum_{d=1}^{D} w_d^m(\xi_m(S)) I_{A_d^m(\xi_m(S))},$$

here ρ is the *memory relaxation* parameter.

4. If $m < M$, go to Step 3, otherwise determine a final estimate

$$\widehat{p}_{\mathrm{SGB}}(x, \xi(S)) = \frac{1}{1 + \exp\{-2 f_M(x, \xi(S), \{\xi_m(S)\}_{m=1}^{M})\}}.$$

This algorithm is to be run for several times with different parameters D, M, ρ and η. Then their optimal values could be chosen via cross-validation (see section 16.3.1 in [21]). Small values of η (= 0.1, 0.05, 0.0225 etc.) help to get accurate estimates for relatively noisy data.

Standard RF and SGB work poorly for unbalanced samples. One needs either to balance given datasets (as in [11]) before these methods are applied or use special modifications of RF ([8]) and SGB ([29]). To avoid overfitting, permutation test needs to be done.

A common problem of all machine learning methods is a complicated functional form of the final probability estimate $\widehat{p}(x, \xi)$ (w.r.t. x). In genetic studies, one wants to pick up all relevant combinations of SNP and risk factors, based on a biological pathway causing the disease. Therefore, the final estimate $\widehat{p}(x, \xi)$ is to be analyzed.

[5] This algorithm is not to be confused with prediction algorithms.

We describe one of the possible methods of such analysis within RF framework and called *conditional variable importance measure* (CVIM). One could determine CVIM for each predictor X_i in X and range all X_i in terms of this measure. Following [42], CVIM of predictor X_i given certain subvector Z_i of X is calculated as follows (supposing Z_i takes values $z_{i1}, \dots, z_{im(i)}$).

1. Construct a vector $(l_1, \dots, l_N)$, randomly permuting $1, \dots, N$ in each subset

$$A_{ik} = \{j : Z_i^j = z_{ik}\}, \quad k = 1, \dots, m(i).$$

2. Generate B bootstrap samples $\xi_b = \big((X^{jb}, Y^{jb}), j = 1, \dots, N\big)$, $b = 1, \dots, B$. For each of these samples, construct a classifier $f_b(x, \xi_b)$ and calculate

$$\mathrm{CVIM}_b = \frac{1}{|C_b|} \sum_{j \in C_b} I\{Y^j = f_b(X^j, \xi_b)\} - \frac{1}{|C_b|} \sum_{j \in C_b} I\{Y^j = f_b(X^{l_j}, \xi_b)\}$$

   where $C_b = \{j \in \{1, \dots, N\} : (X^j, Y^j) \notin \xi_b\}$.

3. Compute the final CVIM using the formula

$$CVIM = B^{-1} \sum_{b=1}^{B} \mathrm{CVIM}_b. \tag{28}$$

Any permutation $(l_1, \dots, l_N)$ in the CVIM algorithm destroys dependence between X_i and (Y, Z_{-i}) where Z_{-i} consists of all components of X which are not in Z_i. At the same time it preserves initial empirical distribution of (X_i, Z_i) calculated for the sample ξ. After that the average loss of correctly classified Y is calculated. If it is relatively large w.r.t. CVIM of other predictors, then X_i plays important role in classification and vice versa.

For Z_i, one could take all components X_k $(k \neq i)$ such that the hypothesis of the independence between X_k and X_i is not rejected at some significance level (e.g., 5%). Note also that CVIM-like algorithm could be used to range pairs of SNP and risk factors w.r.t. the level of association to the disease. This will be done elsewhere.

3 Applications: risks of IHD and MI

We employ here the various statistical methods described above to analyze the influence of genetic and external factors on risks of ischemic heart disease (IHD) and myocardial infarction (MI) using the data for 454 individuals (333 cases, 121 controls) and 333 individuals (165 cases, 168 controls) respectively. These data contain values of seven SNPs (PAI-1, GpIa, GpIIIa, FXIII, FVII, IL-6, Cx37), as well as four external risk factors, namely, obesity (Ob), arterial hypertension (AH), smoking (Sm) and hypercholesterolemia (HC). The age of all individuals in case and control groups ranges from 35 to 55 years, which

reduces its influence on the risk analysis. For each of considered methods, K-fold cross-validation is used with $K = 6$. As shown in [31] and [48], the standard choice of partition number of cross-validation from 6 to 10 does not change the prediction error significantly. We take $K = 6$ as the sample sizes do not exceed 500. The supercomputer SKIF MSU "Chebyshev" was involved to perform computations. All applied methods have prediction error less than 0.25, so predictions constructed have significant predictive power. Indeed, in [12] and [44] the interplay between genotype characteristics and MI development was also studied, with estimated prediction errors 0.30–0.40. Further on we write prediction error instead of estimated prediction error.

3.1 MDR and MDRIR methods

Ischemic heart disease

Table 1 contains (estimated) prediction errors of the most significant combinations obtained by MDR analysis of ischemic heart disease data. At Figure 4 a plot of empirical distribution function of prediction error is given when the disease is not linked with explanatory variables. We use here the simulated samples $\overline{\xi}_b$ introduced in Section 2.1, with $b = 1, \ldots, B$ where $B = 100$. One can see that out of these 100 simulations, the corresponding prediction error was not less than 0.42. Note that Monte Carlo p-value (15) of all three combinations is less than 0.01 (since their prediction errors are much less than 0.42), which is usually considered as a good performance.

Factors	Prediction error
GpIa, FXIII, AH, HC	0.231
Cx37, AH, HC	0.238
GpIa,Cx37, AH, HC	0.241

Table 1: The most significant combinations obtained by MDR analysis for IHD data.

Table 2 contains the results of MDRIR method, which are similar to results of MDR method. However, it is worth mentioning that MDRIR method allows to identify additional combinations with prediction error around 0.24.

Factors	Prediction error
FXIII, FVII, AH, HC	0.240
FXIII, AH, HC	0.242
GpIa, Cx37, AH, HC	0.247

Table 2: The most significant combinations obtained by MDRIR analysis of IHD data.

It follows from Tables 1 and 2 that hypertension and hypercholesterolemia are the most important external risk factors. Indeed, these two factors appear in every of 6 combinations.

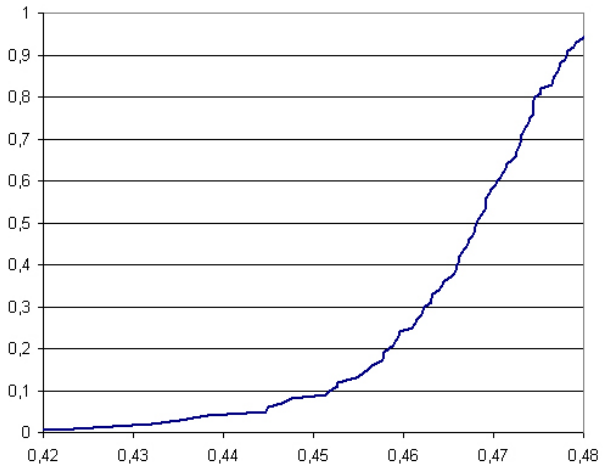

Figure 4: The empirical distribution function of the estimated error for the case when disease risk and predictors are independent (permutation test), IHD dataset.

To perform a more precise analysis of influence of SNPs on IHD provoking we analyze gene-gene interactions. We used two different strategies. Namely, we applied MDR method to a subgroup of individuals who are not subject to any of the external risk factors (i.e. to non-smokers without obesity and without hypercholesterolemia, 51 cases and 97 controls). Another strategy is to apply MDR method to the whole sample, but to take into account only genetic factors rather than all factors. Table 3 contains the most significant combinations of SNPs and their prediction errors.

Method	Genetic factors	Prediction Error
MDR on a subgroup of individuals who do not have any risk factors	GpIa, Cx37	0.281
MDR method on the whole group taking into account only genetic factors	GpIa, Cx37	0.343

Table 3: Comparison of the most significant SNP combinations obtained by two different ways of MDR analysis of IHD data.

It turned out that both methods yield similar results. Combination of SNPs GpIa and Cx37 has the biggest influence on IHD. Prediction error is about 0.28-0.34, and smaller error corresponds to a *risk-free sample*. Moreover it follows from Tables 2 and 3 that prediction error significantly dropped after additional exogenous factors were taken into account (the error is 0.247 if additional external factors are taken into account and 0.343 if not).

Thus based on ischemic heart disease data with the help of Tables 1–3 we can make the following conclusions. Combination of two SNPs (GpIa and Cx37) and two external

factors (hypertension and hypercholesterolemia) has the biggest influence on IHD. Also FXIII gives additional predictive power if AH and HC are taken into account.

Myocardial infarction

Prediction errors of the most significant combinations obtained by MDR analysis of MI data are presented in Table 4. Figure 5 contains the plot of empirical c.d.f. of prediction error if disease is not linked with risk factors. This curve shows that for all 100 simulations of $\overline{\xi}_b$ the estimated prediction error was not less that 0.38. Note that Monte Carlo p-value of all combinations is less than 0.01.

Factors	Prediction error
GpIIIa, FXIII, Cx37, AH	0.343
GpIIIa, FXIII, FVII, Cx37	0.347
Cx37, Sm	0.356

Table 4: The most significant combinations obtained by MDR analysis of MI dataset.

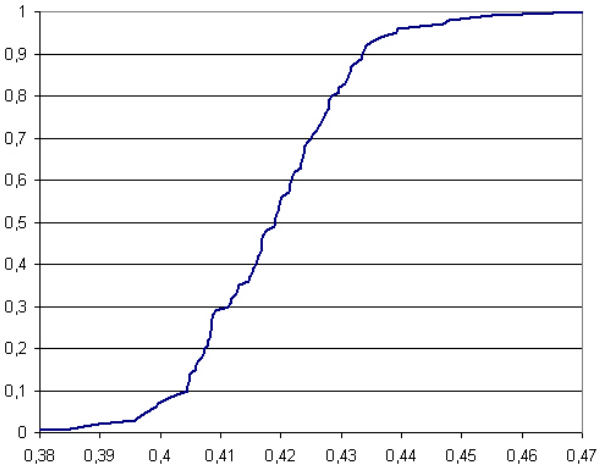

Figure 5: The empirical c.d.f. of the estimated error for the case when disease risk and predictors are independent (permutation test), MI dataset.

MDRIR analysis of the same dataset gives a clearer picture (see Table 5).

Apparently, combination of smoking and SNP Cx37 is the most significant. These two factors appear in all combinations in Table 5. Involving any additional factors only increases prediction error.

The explicit form of the prediction algorithm based on Cx37 and Sm shows that these factors interact nonlinearly. Smoking as well as Cx37 homozygote leads to the disease. However wild-type allele can protect from consequences of smoking, because combination

Factors	Prediction error
Cx37, Sm	0.351
GpIIIa, Cx37, Sm	0.353
GpIIIa, Cx37, Sm, HC	0.355

Table 5: The most significant combinations obtained by MDRIR analysis of MI dataset.

of smoking and Cx37 wild-type is a protective one (i.e. value of prediction algorithm of this combination is -1).

3.2 Logic regression

We performed several research procedures both for IHD and MI data, with different restrictions imposed on the statistical model. To describe these models set

$$(X_1, \ldots, X_n) = (Z_1, \ldots, Z_m, R_1, \ldots, R_k)$$

where variables $Z = (Z_1, \ldots, Z_m)$ stand for SNP values (PAI-1, GpIa, GpIIIa, FXIII, FVII, IL-6, Cx37 respectively) and $R = (R_1, \ldots, R_k)$ denote external risk factors (Ob, AH, Sm, HC), $m = 7$, $k = 4$. We consider four different models in order to analyze both total influence of genetic and external factors and losses in predictive force appearing when some factors are excluded. In our applications we will take $s = 3$, as search over larger forests for samples with modest sizes can give very complicated and unreliable results.

Model 1. We consider the class $\mathcal{M}$ (see Section 2.3) consisting of the functions h having a form

$$h(Z, R) = \beta_0 + \sum_{v=1}^{s} \beta_v T_v(Z_1, \ldots, Z_m) + \sum_{v=1}^{k} \beta_{s+v} R_v$$

where the coefficients $\beta_i \in \mathbb{R}$ and T_i are polynomials identified with trees. In other words we require that external factors are present only in trees consisting of one variable.

Model 2. Now we assume that any function $h \in \mathcal{M}$ has the representation

$$h(Z, R) = \beta_0 + \sum_{v=1}^{s} \beta_v T_v(Z_1, \ldots, Z_m, R_1, \ldots, R_k) \tag{29}$$

where $\beta_i \in \mathbb{R}$ and T_i are polynomials identified with trees. Thus we allow the interaction of genes and external factors in order to find significant gene-environment interactions. However we impose additional restrictions to avoid too complex combinations of external risk factors. We do not tackle here effects of interactions where several external factors are involved. Namely, we consider only the trees satisfying the following two conditions.

1. If there is a leaf containing external factor variable then the root of that leaf contains product operator.

2. Moreover, another branch growing from the same root is also a leaf and contains a genetic (SNP) variable.

Models 3 and 4 have additional restrictions that polynomials $T_v\,(v = 1, \ldots, s)$ in (29) depend only on external factors and only on SNPs respectively. These models are considered to compare their results with ones obtained with all information taken into account, in order to demonstrate the importance of genetic (resp. external) data for risk analysis.

Ischemic heart disease

We have the following results.

Model	1	2	3	4
Prediction error	0.19040	0.20364	0.22812	0.33990

Table 6: Results of LR for IHD dataset.

Note that prediction error in Model 1 is only about 0.19. For the same model we performed also *fast simulated annealing* search of the optimal forest which is much more time-efficient, and a reasonable error of 0.23 was obtained. Model 3 application shows that external factors play an important role in IHD genesis, as classification based on external factors only gives the error less than 0.23, while usage of SNPs only (Model 4) lets the error grow to 0.34.

Model 1 gave the minimal prediction error. For the optimal forest $(T_1, \ldots, R_4)$ the function $\widehat{h}(Z, R)$ given before formula (26) with $S = \{1, \ldots, N\}$ is provided by the expression

$$-0.597T_1 - 0.354T_2 + 0.521T_3 - 0.444R_1 + 1.311R_2 - 0.146R_3 + 2.331R_4 - 0.226 \quad (30)$$

where[6]

$$T_1 = (Z_4Z_3 + Z_6Z_7 + Z_2Z_2 + Z_3Z_7)(Z_1)^2Z_3Z_7,$$
$$T_2 = Z_1(Z_3)^2(Z_6Z_7 + Z_7(Z_4)^2Z_2), \;\; T_3 = Z_2 + 2Z_2(Z_6)^2Z_7.$$

The external factors 2 and 4 (i.e. AH and HC) are the most influential since the coefficients at them are the greatest ones (1.311 and 2.331). As was shown above, MDR yields the same conclusion. If the gene-environment interactions are allowed (Model 2), no considerable increase in predictive force has been detected. However we list the pairs of SNPs and external factors present in the best forest: Z_7 and R_2, Z_7 and R_1, Z_7 and R_4, Z_5 and R_1. It is seen that Cx37 SNP is of substantial importance as it appears in combination with all risk factors except for smoking.

As formula (30) is hard to interpret, we select the most significant SNPs via a variant of permutation test. Consider a random rearrangement of the column with first SNP in IHD dataset. Calculate the prediction error using these new simulated data and the same function $\widehat{h}$ as before. The analogous procedure is done for other columns (containing the

[6]The sums and products are modulo 3.

values of other SNPs) and the errors found are given in Table 7. It is seen that the error increases considerably when the values of GpIa and Cx37 are permutated. The statement that they are the main sources of risk agrees with what was obtained above by MDR method.

Prediciton error for Model 1	GpIa	Cx37	IL-6	PAI-1	GpIIIa	FXIII	FVII
0.19040	0.26283	0.25987	0.22590	0.21212	0.20798	0.20173	0.19040

Table 7: The SNP significance test for IHD in Model 1.

Myocardial infarction

For the MI dataset, under the same notations that above, the following results for our four models were obtained.

Model	1	2	3	4
Prediction error	0.30526	0.33058	0.39057	0.36455

Table 8: Results of LR for IHD dataset.

To comment the Table 8 we should first underline that external risk factors play less important role compared with IHD risk: if they are used without genetic information, the error increases by 0.09, see Models 1 and 3 (while the same increase for IHD was 0.03). The function $\widehat{h}(Z,R)$ defined before formula (26) with $S=\{1,\dots,N\}$ is is equal to

$$-1.144T_1+0.914T_2-0.45T_3-0.285R_1-0.675R_2+0.828R_3-0.350R_4-0.055$$

where

$$T_1=Z_1Z_3(Z_5)^2,\ \ T_2=Z_7,\ \ T_3=Z_4+Z_3+Z_7+Z_6.$$

Thus the first tree has the greatest weight (coefficient equals -1.144), the second tree (i.e. Cx37 SNP) is on the second place, and external factors are less important.

As for IHD we performed a permutation test to compare the significance of different SNPs. Its results are presented in Table 9.

Prediciton error for Model 1	Cx37	GpIIIa	IL-6	FXIII	FVII	PAI-1	GpIa
0.30526	0.44420	0.35345	0.33998	0.32761	0.32427	0.31918	0.30526

Table 9: The SNP significance test for MI in Model 1.

As seen from this table, the elimination of Cx37 SNP leads to a noticeable increase in the prediction error. This fact agrees with results obtained by MDR analysis of the same dataset.

3.3 Results obtained by RF and SGB methods

The given datasets were unbalanced w.r.t. response variable and we first applied the resampling technique to them. That is, enlargement of the smaller of two groups case-control in the sample by additional bootstrap observations till the final proportion case:control would be 1:1. We also employed modifications of RF by [8] and SGB by [29] for unbalanced samples, but those worked poorly for permutation tests and we do not give their results here. Note that due to the resampling techniques the following effect arise. Some observations in small groups (case or control) appear in the new sample more frequently than other ones. Therefore, we took the average over 1000 iterations.

Ischemic heart disease

Data	RF	SGB
with SNP	0.20/0.454	0.134/0.473
without SNP	0.23/0.51	0.261/0.503

Table 10: Prediction error/prediction error in permutation test calculated via cross-validation for IHD dataset with employment of RF and SGB methods.

Results of RF and SGB methods are given in Table 10. It shows that RF and SGB methods give statistically reliable results (prediction error in the permutation test is close to 50%). Moreover, additional SNP information improves predicting ability on 11% and 13% (SGB). It seems that SGB method is better fitted to IHD data than RF.

Computing CVIM for each X_i, we constructed Z_i as follows. We included in Z_i all predictors X_j, $j \neq i$, for which χ^2-criteria rejected hypothesis of independence between X_j and X_i at 5% significance level. Since the genetic information has second order effect on prediction of Y comparing to the risk factors, we ran the program 1000 times and then took the average CVIM to get a reliable estimate. An error over different runs of the program was around 0.01. The results are given in Table 11.

AH	HC	Cx37	Ob	FXIII	Sm	GpIa	FVII	PAI-1	GpIIIa	IL-6
8.9	5.3	5.1	0.56	0.53	0.11	0.1	0.07	0.03	0.02	0.01

Table 11: Predictors are ranged in terms of their CVIM for IHD dataset.

Thus, the most relevant predictors for IHD are AH, HC and Cx37.

Myocardial infarction

Results of RF and SGB methods are given in the following table.

Table 12 shows that RF and SGB methods give statistically reliable estimates (prediction error in the permutation test is close to 50%). Moreover, additional SNP information improves predicting ability on 10%.

Data	RF	SGB
with SNP data	0.36/0.497	0.399/0.53
without SNP data	0.473/0.527	0.482/0.562

Table 12: Prediction error/prediction error in permutation test calculated via cross-validation for MI dataset with employment of RF and SGB methods.

Cx37	Sm	AH	GpIIIa	FVII	FXIII	HC	GpIa	Ob	IL-6	PAI-1
7.5	2	1.86	0.03	0.02	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0

Table 13: Predictors are ranged in terms of their CVIM for MI dataset.

CVIM was calculated according to (28) and is given below.
Thus, the most relevant predictors for MI are Cx37, Sm and AH.

4 Conclusions and final remarks

Let us briefly summarize the main results obtained. The analysis of IHD dataset showed that two external risk factors out of four considered (AH and HC) have a strong connection with the disease risk (the error of classification based on external factors only is 0.25–0.26 with p-value less than 0.01). Also, the classification based on SNPs only gives a relatively low error of 0.28. Moreover, the most influential SNPs are Cx37 and GpIa (FXIII also enters the analysis only when AH and HC are present). Prediction error decreases to 0.13 if both SNP information and external risk factors are taken into account. Note that excluding any of the 5 remaining SNPs (all except for two most influential) from data increases the error by 0.01–0.02 approximately. So, while the most influential data are responsible for the situation within a large part of population, there are smaller parts where other SNPs come to effect and provide a more efficient prognosis ("small subgroups effect").

The MI dataset gave the following results. The most significant factors of MI risk are the Cx37 SNP (more precisely, homozygous mutation) and smoking with a considerable gene-environment interaction present. The smallest prediction error of methods applied was 0.33–0.35 (with p-value less than 0.01). The classification based on external factors only yields a much greater error of 0.42. Thus genetic data improves the prognosis quality essentially. While two factors are of great importance, other SNPs considered actually do not improve the prognosis essentially, i.e. no small groups effect is observed.

The conclusions given above are based on several complementary methods of modern statistical analysis. These new data mining methods allow to analyze other datasets as well. The study can be continued with larger datasets, in particular, involving new SNP data.

Acknowledgments

The authors are grateful to Professor V.A.Tkachuk, Associate Professor L.M.Samokhodskaya and MD. A.V.Balatsky for providing the data concerning the complex diseases. On the basis of this preprint the joint paper will be prepared with medical interpretation of the obtained results.

A.V.Bulinski also would like to thank Professors I.Kourkova and G.Pagès for invitation to LPMA of the University Pierre and Marie Curie, he is grateful to all the members of LPMA for hospitality.

The work is partially supported by RFBR grant 10-01-00397.

References

[1] A. Albrechtsen, S. Castella, G. Andersen, T. Hansen, O. Pedersen and R. Nielsen. *A Bayesian multilocus association method: allowing for higher-order interaction in association studies.* Genetics, vol. 176 (2007), pp. 1197-1208.

[2] A. Arlot and A. Celisse. *A survey of cross-validation procedures for model selection.* Statist. Surv., vol.4 (2010), pp. 40-79.

[3] R. Bahadur. *A representation of the joint distribution of responses to n dichotomous items.* Studies in Item Analysis and Prediction, Stanford University Press, H. Solomon (ed.), 1961, pp. 158-168.

[4] G. Biau. *Analysis of a Random Forests model.* J. of Machine Learning Research, LSTA, LPMA, Université Paris-6, 2010.

[5] G. Biau, L. Devroye and G. Lugosi. *Consistency of Random Forests and Other Averaging Classifiers.* J. of Machine Learning Research, vol.9 (2008), pp. 2015-2033.

[6] C. Bock and T. Lengauer. *Computational epigenetics.* Bioinformatics, vol. 24 (2008), pp. 1-10.

[7] A. Bondy and U.S.R. Murty. *Graph Theory.* Springer, 2008.

[8] L. Breiman, C. Chen and A. Liaw. *Using random forest to learn imbalanced data.* J. of Machine Learning Research, no. 666, Department of Statistics, University of California, Berkeley, CA, 2004.

[9] A. Bulinski. *Stochastic methods of identification of SNP interactions.* The 1st Int. Research and Practice Conference on Postgenomic Methods of Analysis in Biology, and Laboratory and Clinical Medicine, MSU, 2010, p. 146.

[10] A. Bulinski, O. Butkovsky, A. Shashkin, P. Yaskov, M. Atroshchenko and A. Khaplanov. *Investigations in the framework of the MSU Research project Postgenomic Medical Studies and Technologies.* MSU, 2010.

[11] N.V. Chawla. *Data mining for imbalanced datasets: An overview.* Data mining and knowledge discovery handbook, Part 6, O. Maimon and L. Rokach (eds.), Springer, 2010, pp. 875-886.

[12] C.S. Coffey, P.R. Hebert, M.D. Ritchie, H.M. Krumholz et al. *An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation.* Atherosclerosis, vol. 30 (2004), pp. 5-49.

[13] D.R. Cox. *The analysis of multivariate binary data.* Applied Statistics, vol. 21 (1972), pp. 113-120.

[14] J.Y.Dai et al. *SHARE: an adaptive algorithm to select the most informative set of SNPs for candidate genetic association.* Biostatistics, vol. 10 (2009), pp. 680-693.

[15] J.H. Freedman. *Greedy function approximation: A gradient boosting machine.* Ann. Statist., vol. 29 (2001), pp. 1189-1232.

[16] A. Fritsch and K. Ickstadt. *Comparing logic regression based methods for identifying SNP interactions.* Lecture Notes in Computer Science 4414, Springer, 2007, pp. 90-103.

[17] GAW16, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000128.v3.p3.

[18] P. Golland, F. Liang, S. Mukherjee and D. Panchenko. *Permutation tests for classification*, Lecture notes in Computer Science, Springer, vol. 3559 (2005), pp. 501-515.

[19] B. Hajek. *Cooling schedules for optimal annealing.* Math. Oper. Res., vol. 13 (1988), pp. 311-329.

[20] HapMap, http://www.hapmap.org.

[21] T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009.

[22] D. Hosmer and S. Lemeshow. *Applied Logistic Regression.* Wiley, 2000.

[23] R. Karchin. *Next generation tools for the annotation of human SNPs.* Briefings in Bioinformatics, vol. 10 (2009), pp. 35-52.

[24] K. Kohara, Y. Tabara, J. Nakura, Y.Imai et al. *Identification of Hypertension-Susceptibility Genes and Pathways by a Systemic Multiple Candidate Gene Approach: The Millennium Genome Project for Hypertension.* Hypertension Research, vol. 31 (2008), pp. 203-212.

[25] C. Kooperberg, J.C. Bis, K.D. Marciante, S.R. Heckbert, T. Lumley and B.M. Psaty. *Logic regression for analysis of the association between genetic variation in the renin-angiotensin system and myocardial infarction or stroke.* Am. J. of Epidemiology, vol. 165 (2007), pp. 334-343.

[26] S.Y. Lee, Y. Chung, R.C. Elston, Y. Kim and T. Park. *Log-linear model-based multifactor dimensionality reduction method to detect gene-gene interactions.* Bioinformatics, vol. 23 (2007), pp. 2589-2595.

[27] E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses*, Springer, 2005.

[28] T. Lengauer (ed.). *Bioinformatics From Genomes to Therapies*, Wiley VCH Verlag GmbH and KGaA, Weihhein, 2007.

[29] X.-Y. Liu, J. Wu and Z.-H. Zhou. *Exploratory Undersampling for Class-Imbalance Learning.* Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions, vol. 39 (2008), pp. 539-550.

[30] P. Massart. *Concentration Inequalities and Model Selection.* Springer, 2003.

[31] A.A. Motsinger and M.D. Ritchie. *The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction.* Genetic Epidemiology, vol. 30 (2006), pp. 546-555.

[32] A.G. Nikolaev and S.H. Jacobson. *Simulated Annealing.* Handbook of Metaheuristics, Springer, 2010, pp. 1-39.

[33] R. Nunkesser, T. Bernholt, H. Schwender, K. Ickstadt and I. Wegener. *Detecting high-order interactions of single nucleotide polymorphisms using genetic programming.* Bioinformatics, vol. 23 (2007), pp. 3280-3288.

[34] D.J. Olive, *The number of samples for resampling algorithms.* http://www.math.siu.edu/olive/ppresamp.pdf, 2010.

[35] J. Park. Independent rule in classification of multivariate binary data. J. Multivar. Anal., vol. 100 (2009), pp. 2270-2286.

[36] M.D. Ritchie, L.W. Hahn, N. Roodi, R. Bailey, W.D. Dupont, F.F. Parl and J.H.Moore. *Multifactor-dimensionality reduction reveals high-order iteractions among estrogen-metabolism genes in sporadic breast cancer.* Am. J. Hum. Genet., vol. 69 (2001), pp. 138-147.

[37] S. Rubenthaler, T. Ryden and M. Wiktorsson. *Fast simulated annealing in $\mathbb{R}^d$ with an application to maximum likelihood estimation in state-space models.* Stoch. Proc. Appl., vol. 119 (2009), pp. 1912-1931.

[38] I. Ruczinski, C. Kooperberg and M. LeBlanc. it Logic regression. J. Comp. Graph. Statist., vol. 12 (2003), pp. 475-511.

[39] H. Schwender and K. Ickstadt. *Identification of SNP interactions using logic regression.* Biostatistics, vol. 9 (2008), pp. 187-198.

[40] H. Schwender and K. Ickstadt. *Empirical Bayes analysis of single nucleotide polymorphisms.* Bioinformatics, vol. 9 (2008), pp. 144-159.

[41] H. Schwender and I. Ruczinski. *Testing SNPs and sets of SNPs for importance in association studies.* Biostatistics, vol. 11 (2010), pp. 1-15.

[42] C. Strobl, A.L. Boulesteix, T. Kneib, T. Augustin and A. Zeileis. *Conditional variable importance for random forests.* BMC Bioinformatics, vol. 9 (2008), p. 307.

[43] S. Szymczak, J.M. Biernacka, H.J. Cordell, G.-R. Oscar, I.R. König, H. Zhang and Y.V. Sun. *Machine Learning in Genome-Wide Association Studies.* Genetic Epidemiology, vol. 33 (2009), pp. 51-57.

[44] C.-T. Tsai, J.-J. Hwang, M.D. Ritchie and J.H. Moore. *Reninangiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: Detection of high order gene−gene interaction.* Atherosclerosis, vol. 195 (2007), pp. 172-180.

[45] D. Velez, B.C. White, A.A. Motsinger, W.S. Bush, M.D. Ritchie, S.M. Williams and J.H. Moore. *A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction.* Genetic Epidemiology, vol. 31 (2007), pp. 306-315.

[46] J. Venter, M.D. Adams, E.W. Myers, P.W. Li et al. *The Sequence of the Human Genome.* Science, vol. 291 (2001), pp. 1304-1351.

[47] X. Wan, C. Yang, Q. Yang, H. Xue, N.L.S. Tang and W. Yu. *MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study.* BMC Bioinformatics, vol. 10 (2009), p. 13.

[48] S. Winham, A.J. Slater and A.A. Motsinger-Reif. *A comparison of internal validation techniques for multifactor dimensionality reduction.* BMC Bioinformatics, vol. 11 (2010), pp.394.

[49] Z. Zhang, E.S. Buckler, T.M. Casstevents and P.J. Bradburg. *Software engineering the mixed model for genome wide assosiation studies on large samples*, Briefings in Bioinformatics, vol. 10 (2009), pp. 664-675.

Alexander BULINSKI,

Faculty of Mathematics and Mechanics, Lomonosov Moscow State University, GSP-1, Leninskie gory, Moscow, 119991, Russia

and

LPMA UPMC University Paris-6,
4 Place Jussieu, 75252 Paris CEDEX 05, France

E-mail address: bulinski@mech.math.msu.su

Oleg BUTKOVSKY,

Faculty of Mathematics and Mechanics, Lomonosov Moscow State University, GSP-1, Leninskie gory, Moscow, 119991, Russia

E-mail address: oleg.butkovskiy@gmail.com

Alexey SHASHKIN,

Faculty of Mathematics and Mechanics, Lomonosov Moscow State University, GSP-1, Leninskie gory, Moscow, 119991, Russia

E-mail address: ashashkin@hotmail.com

Pavel YASKOV,

Steklov Mathematical Institute, Gubkina str. 8, Moscow, 119991, Russia

and

Faculty of Mathematics and Mechanics, Lomonosov Moscow State University, GSP-1, Leninskie gory, Moscow, 119991, Russia

E-mail address: pavel.yaskov@mi.ras.ru