Limit Theorems in Probability Theory and Their Applications
15 - 21 August
Novosibirsk - 2011

Statistical analysis of complex diseases models in genetics

A.V.Bulinski
Lomonosov Moscow State University

The susceptibility to complex diseases (such as cardiovascular, oncological ones etc.) determined by the genetic mechanisms has drawn much attention in the leading biomedical centers worldwide. The goal is to provide the prophylactic measures and medical treatment taking into account personal genetic peculiarities which increase the risks of some diseases and protect from the others. In particular one investigates the relations between the deviations in the genetic code and certain maladies. Individual's DNA variations are typically described in terms of single nucleotid polymorphisms (SNP), i.e. the fragments of genetic code where a nucleotide change is possible. The first examples of genetically based diseases (e.g., sicklemia) were related with a single mutation. However, many diseases, e.g., diabetes, have a complex character as they can be provoked by mutations in different parts of the DNA which are responsible for the formation of certain types of proteins. Quite a number of recent studies demonstrate that the increasing risks of complex diseases can be explained by combinations of certain SNP whereas separate mutations have no dangerous effects.

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 of 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.

There are two closely related 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). The second trend is to identify relevant combinations of SNP having the most significant pathogenic influence. Both directions are discussed in the talk. Moreover, further development of the mentioned statistical methods is proposed as well as their applications to study of the risks of cardiovascular diseases (see [1]).

Due to high-dimensionality of data, the numerical imlementation of statistical methods is very time consuming. So the supercomputer SKIF MSU "Chebyshev" was employed. This statistical data analysis started within 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 in [2].

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

[1] A.Bulinski, O.Butkovsky, A.Shashkin, P.Yaskov. Statistical methods of SNP data analysis with applications. Université de Paris 6 et Paris 7 - CNRS (UMR 7599). Prepublication n° 1450 (juin 2011) du Laboratoire de Probabilités et Modèles Aléatoires, p. 1-30; arXiv:1106.4989v1 [math.PR].

[2] 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.