UDC 519.2

Asymptotic Methods and Limit Theorems

A. V. Bulinski*

* *Department of Probability Theory,*
Faculty of Mathematics and Mechanics,
Lomonosov Moscow State University,
Leninskie Gory 1, Moscow, 119234, Russia

Abstract. The talk is devoted to problems related to asymptotic analysis of dependent functions constructed by means of arrays of independent observations. Such functions are employed in statistics, e.g., in the framework of regression analysis, and have a number of applications in medical and biological studies. A new version of the conditional central limit theorem is established and applied to data analysis. The feature selection problems are considered as well.

Keywords: arrays of random variables, conditional central limit theorem, law of large numbers, feature selection.

1. Introduction

We study the models described by systems of dependent random variables and discuss their applications. In the first part of the talk we prove the conditional CLT for arrays of random variables. Conditional probabilities and conditional expectations play an important role in the modern probability theory. It suffices to indicate the classes of Markov stochastic processes and random fields, martingales, conditionally associated processes and others. The concept of conditionally independent variables goes back to A.A.Markov. The relationship between exchangeability and conditional independence was studied by B. de Finetti. Later, in the works by B. K. L. Prakasa Rao, G. G. Roussas (see, e.g., [11] and [13]) and other researchers, the classes of random variables possessing different forms of conditional independence were considered. During the last decade, for such classes the analogues of some classical limit theorems of probability theory were established. In particular, we employ our conditional CLT ([5]) for an extension of the recent result by L.Györfi and H.Walk [9] concerning the regression function estimation. In the second part of the talk we discuss the feature selection problems (see, e.g., [1]) and asymptotic behavior of the corresponding statistics. We concentrate on the multifactor dimensionality reduction (MDR) method proposed by M.Ritchie and coauthors in 2001. The review [8] shows that more than 800 papers published between 2001 and 2014 were devoted to extensions, modifications and applications of the initial idea. The development of our MDR-EFE (MDR - error function estimation) method ([2], [4]) to stratified samples in the case of a binary response variable, describing, e.g., the sick and healthy state of a patient, is obtained. The study of stratified samples ([6]) is essential when

the disease probability is small. We establish a criterion of strong consistency of estimates, involving K-cross-validation procedure and penalty, for a specified prediction error function. The generalization of the XOR-model important for genetic data analysis is introduced. The cost approach is proposed to compare experiments with random and non-random number of observations. Analytic results are accompanied by simulations.

2. Main results

Assume as usual that there exists a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and all random variables under consideration are defined on it. Let a σ-algebra $\mathcal{A}$ be such that $\mathcal{A} \subset \mathcal{F}$. The events $A_1, \ldots, A_n$ are called conditionally independent w.r.t. $\mathcal{A}$ or $\mathcal{A}$-independent if

$$\mathrm{P}^{\mathcal{A}}\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} \mathrm{P}^{\mathcal{A}}(A_k), \tag{1}$$

here $\mathrm{P}^{\mathcal{A}}(A) := \mathrm{E}^{\mathcal{A}}(I\{A\})$ for $A \in \mathcal{F}$, $I\{A\}$ being an indicator of a set A, and $\mathrm{E}^{\mathcal{A}}X$ stands for conditional expectation (when it exists) of a random variable X w.r.t. $\mathcal{A}$. The $\mathcal{A}$-independence of σ-algebras $\mathcal{A}_1, \ldots, \mathcal{A}_n$ means that (1) holds for any $A_i \in \mathcal{A}_i$, $i = 1, \ldots, n$. The random vectors $X_1, \ldots, X_n$ are $\mathcal{A}$-independent whenever σ-algebras generated by these vectors are conditionally independent w.r.t. $\mathcal{A}$. An infinite system of random variables is called $\mathcal{A}$-independent if any finite collection of these random variables possesses such property.

Clearly, (1) is valid for $\mathcal{A} = \mathcal{F}$. We obtain the classical definition of independent events (σ-algebras, random variables) taking $\mathcal{A} = \{\varnothing, \Omega\}$. Independence of events (or random variables) can disappear if we take nontrivial σ-algebra $\mathcal{A}$. At the same time some dependent events (or random variables) could be considered as conditionally independent for an appropriate choice of σ-algebra $\mathcal{A}$. The corresponding examples one can find in [7], [11] and [13]. The relationship between conditional independence and exchangeability is considered in Section 7.3 of the monograph [7].

The arrays with rows consisting of conditionally independent random variables w.r.t. certain σ-algebras are studied. An analogue of the Lindeberg - Feller theorem known for systems of independent random variables is established. This result is based on the theorem proved by D-M. Yuan, L-R. Wei, L. Lei in [14] where the authors considered a sequence of random variables conditionally independent w.r.t. a given σ-algebra. They were interested in a.s. convergence whereas our version of the Lindeberg condition in a weak form (involving convergence in probability) is less restrictive. An application of the mentioned new result for arrays provides an extension of conditions for asymptotic normality of the estimates of the regression function second moment obtained in a recent paper by L. Györfi and H. Walk [9].

The research direction combining probability, statistics and machine learning for analysis of mathematical problems of feature selection is vastly represented in literature along with various applications of this theory. Let us consider a response variable Y depending on factors (features) $X_1, \dots, X_n$. The challenging problem is to identify a collection of relevant factors $X_{k_1}, \dots, X_{k_r}$ such that Y depends on them essentially in a sense. Quite a number of powerful methods were developed for different models in the course of such investigations. Several new variable selection procedures have emerged during the last 20 years. Note that many exhaustive, stochastic and heuristic methods to detect epistasis (in genetics) are considered in [10]. In the paper by M.Ritchie et al. [12] the MDR method was proposed to identify the relevant factors having influence on a binary response variable. The review [8] demonstrates great popularity of the method. One can mention the following versions of this method: MDR method with independent rule, the generalized MDR (GMDR) method employing the framework of generalized linear models for scoring in conjunction with MDR, the model-based MDR (MB-MDR) method which allows a more flexible definition of risk cells than the application of MDR techniques, the MDR pedigree disequilibrium test (MDR-PDT). Gene-MDR method and a robust MDR method (RMDR) have been introduced also among others.

We are interested in identification of a collection of relevant factors which determine in a sense the behavior of a random response. For instance a binary response 1 or -1 can characterize the state of a patient health (1 means that a person is sick and -1 corresponds to healthy person). In the study of limit behavior of the proposed statistics special attention will be paid to using the versions for arrays of random variables the law of large numbers and the central limit theorem, see, e.g., [3], [4] and [6]. We compare two approaches concerning the application of the MDR-EFE method for different sample plans. The first one was described in [2] and consists in the employment of nonrandom number of i.i.d. observations. The second one is considered in [6] and involves the stratified sample. More exactly, stratification means the separation of observations taking into account the values of a response variable under consideration. We will denote these methods as iMDR-EFE and sMDR-EFE, respectively.

To compare two approaches in the sense of the total cost of experiment assume that there is a fixed amount of money C ($C \in \mathbb{N}$) for research. Let each observation (X_i, Y_i) cost 1 and let the ratio of the price of measuring Y to that of X be $w \in R^+$. Since comparison for equal sample sizes is not interesting, we consider the maximal sizes $s_{ind}(C; w)$ and $s_{str}(C; w)$ of the samples which are available in experiments organized to apply iMDR-EFE and sMDR-EFE, respectively.

We turn to the popular XOR-model to compare iMDR-EFE and sMDR-EFE. This model is used in genetics to describe epistasis without main effects. Namely, let $\mathbb{X} = \{0, 1, 2\}^n$ ($0, 1, 2$ correspond to the number of minor alleles of a specified gene). Assume now that the components

of a random vector $X = (X_1, \ldots, X_n)$ are independent and, for each $i \in \{1, \ldots, n\}$, there exists such $p_i \in (0, 0.5]$ that $\mathrm{P}(X_i = 0) = (1 - p_i)^2$, $\mathrm{P}(X_i = 1) = 2p_i(1 - p_i)$, $\mathrm{P}(X_i = 2) = p_i^2$. This situation is typical for genome-wide association studies (GWAS) where each X_i corresponds to a single nucleotide polymorphism (SNP) and p_i is a minor allele frequency (MAF). We propose a generalization of the XOR-model described in [15] to the case of more than 2 relevant factors. Fix a collection $\{k_1, \ldots, k_r\} \subset \{1, \ldots, n\}$. Set, for each $x = (x_1, \ldots, x_n) \in \mathbb{X}$,

$$\mathsf{P}(Y = 1 | X = x) = \begin{cases} \gamma, & (x_{k_1} + \ldots + x_{k_r}) \bmod 2 = 1, \\ 0, & \text{otherwise} \end{cases}$$

where $\gamma \in (0, 1)$ and $p_{k_1} = \cdots = p_{k_r} = 0.5$. According to Lemma 3 of [6] a response variable Y depends on $(X_{m_1}, \ldots, X_{m_l})$ where $\{m_1, \ldots, m_l\} \subset \{1, \ldots, n\}$ if and only if $\{k_1, \ldots, k_r\} \subset \{m_1, \ldots, m_l\}$. Thus $(X_{k_1}, \ldots, X_{k_r})$ is a collection of relevant factors. In our simulations we employ the XOR-model of dependence between predictors and response variable to compare iMDR-EFE and sMDR-EFE. In order to measure the method performance power we use TMR (true model rate). It is shown in [6] that, for the fixed total cost C a stratified sampling gives better results than independent one.

We tackle also the problem of stability of feature selection methods. We consider various stability measures (indexes) and discuss their properties.

3. Conclusions

We mentioned some problems concerning the asymptotic analysis of the specified stochastic models. Besides a survey the new results with the sketches of their proofs will be provided in the talk. Special attention will be paid to applications.

Acknowledgments

This work is supported by Russian Science Foundation under grant No. 14-21-00162.

References

1. *Bolon-Canedo V., Sanchez-Marono N., Alonso-Betanzos A.* Feature Selection for High-Dimensional Data. — Springer, 2015.
2. *Bulinski A.* On foundation of the dimensionality reduction method for explanatory variables // Journal of Mathematical Sciences. — 2014. — Vol. 199, no. 2. — P. 113–122.

3. *Bulinski A.* Central limit theorem related to MDR-method // Asymptotic Laws and Methods in Stochastics. A volume in Honour of Miklos Csorgo. Fields Institute Communications. — Springer, 2015. — Vol. 76. — P. 113–128.
4. *Bulinski A., Rakitko A.* MDR method for nonbinary response variable // J. of Multivariate Analysis. — 2015. — Vol. 135 — P. 25–42.
5. *Bulinski A.* Conditional central limit theorem // Theory of Probability and Applications. — 2016. — Vol. 61, no. 4. — P. 1–23.
6. *Bulinski A., Kozhevin A.* New version of the MDR method for stratified samples // Statistics, Optimization and Information Computing. — 2017. — Vol. 5. — P. 1–18.
7. *Chow Y. S., Teicher H.* Probability Theory. Independence, Interchangeability, Martingales. 3rd ed. Springer, 1997.
8. *Gola D., John J.M.M., van Steen K., Konig R.* A roadmap to multifactor dimensionality reduction methods // Briefings in Bioinformatics. — 2015. — P. 1–16.
9. *Györfi L., Walk H.* On the asymptotic normality of an estimate of a regression functional // J. Mach. Learn. Res. — 2015. — Vol. 16. — P. 1863–1877.
10. *Moore J.H., Williams S.M.* (Eds.). Epistasis: Methods and Protocols. Methods in Molecular Biology. Vol. 1253. — Springer Science + Business Media, New York, 2015.
11. *Prakasa Rao B.L.S.* Conditional independence, conditional mixing and conditional association // Ann. Inst. Stat. Math. — 2009. — Vol. 61. — P. 441–460.
12. *Ritchie M.D., Hahn L.W., Roodi N., Bailey R., Dupont W.D., Parl F.F., Moore J.H.* Multifactor dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer // Amer. J. Human Genetics. — 2001. — Vol. 69. — P. 139–147.
13. *Roussas G. G.* On conditional independence, mixing, and association // Stochastic Anal. Appl. — 2008. — Vol. 26, no. 6. — P. 1274–1309.
14. *Yuan D-M., Wei L-R., Lei L.* Conditional central limit theorems for a sequence of conditional independent random variables // J. Korean Math. Soc. — 2014. — Vol. 51, no. 1. — P. 1–15.
15. *Winham S.J., Slater A.J., Motsinger-Reif A.A.* A comparison of internal validation techniques for multifactor dimensionality reduction // BMC Bioinformatics — 2010. — Vol. 11, no. 1. — Article 394.