Applied problems of the identification and regression analysis in stochastic structures with multimodal properties

Vladimir Kulikov¹-⁶, Alexander Kulikov¹, vb.kulikov@yandex.ru

¹Nizhny Novgorod State Technical University n. a. R. E. Alekseev, Nizhny Novgorod, Russia

Abstract. We have been elaborating an approach founded on the identification of multimodal laws of the complex structure distribution in medicine, biology, chemistry of ultrapure materials and membrane technology as well as in technical applications. The method is based on the formulation and solution of inverse problems in mathematical physics for the respective probability density functions. The verification of the used algorithmic tools is carried out on model limited-scope samples. For stochastic structures and systems under study the method is supplemented with an original option of a regression analysis taking into account the identified stochastic laws displaying numerical parameters into the binary space. The proposed approach has been tested on clinical material in practical medicine.

1 Introduction

Applied problems for the identification of distribution laws with multimodal specificity may be based on the solution of inverse problems in mathematical physics. Nevertheless, there are some approaches based on the classical methods of constructing histograms, the Parzen-Rosenblatt algorithms, the application of the goodness-of-fit criteria for testing of hypotheses regarding stochastic phenomena under study. In publications [1, 2] the issues are discussed, which relate to the selection of the (kernel) weight function in constructing the probability density estimate by the Parzen-Rosenblatt type and the conditions for using sign-variable higher-order kernels to improve the identification algorithm quality. D.B.H. Klein [1] introduced the term of "the admissible weight function" and established the necessary ratios to minimize an integral squared error within the $L^2$ approach and in the assumption of sufficient smoothness of the $p (x)$ function. In metric $C [-\infty, \infty]$ the admissibility condition may be omitted [2].

The kernel function selection is herewith always limited by a priori suggestion of a researcher on the smoothness of the estimated density and the sample scope.

To this block of problems quite naturally the regression analysis method and the least-squares method are linked up. However, without a proper analysis of the stability of solutions obtained their development, for example, with the approximation of stochastic data by linear varieties with the dimensionality less than the dimension of initial data spaces (method of principal components) is not quite correct.

In publications of a number of the USA scientists, to study the considerable data arrays on the level of gene expression in the mRNA (messenger ribonucleic acid) they initially applied the principal components analysis (PCA), which generated linear representations of uncorrelated components [3].

Later works have already shown in the context of the gene expression the significance of the ICA (Independent Components Analysis), which is especially effective with the assumption of dependence within the data generation process – Liebermeister [4].

The approach using the solution of inverse problems [5] enables to suggest in addition a method combining the said principles for the identification of the distribution densities and regression algorithms. This combination forms a method for the interpretation of experimental data and includes a new version of regression.

The method is universal and its testing has been carried out with the factual material in clinical immunology. It is expected that the results of works such as the solution principles of inverse stochastic problems, correct algorithms, programs will become a tooling in the field of biotechnology and modern "personalized" medicine. As it has been previously shown [5], a number of indicators in immunology have such peculiarities as complex distribution laws. To great disturbances in the immune system (in response to infection or therapy) shall correspond in this case the similar features of processes in non-equilibrium dissipative structures leaving the point of dynamic equilibrium.

The structure changes from the normal state to the chaotic one with loosing its conservative properties, its parameters outreach normal ranges. With further dynamics external fluctuations become significant in bifurcation points.

¹Corresponding author: vb.kulikov@yandex.ru

This is an Open Access article distributed under the terms of the Creative Commons Attribution License 4.0 which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
In mathematical immunology the most known are the studies of the viral infections dynamics.

The review www.cytokines.uspb.ru of S.R. Kuznetsov, V.I. Shishin (the St. Petersburg State University) includes the latest achievements of mathematical immunology.

The review “describes examples of the effective use of major mathematical methods in order to study the immune system dynamics, to estimate values inaccessible for the direct measurement, to carry out computer experiments”.

It is noted that “in the recent decade a number of new measurement techniques has come into existence, which enable to make high-quality analyzes in vivo, it is worth noting among them the multiphoton microscopy and the resonance transfer of fluorescence energy; the X-ray investigation, MRI and ultrasonic diagnostics have been also extended to a micro-level” [6, 7].

The results of monitoring for such processes (experimental data) require correct interpretation models and stable processing algorithms.

2 Statement of identification and verification problems

The method used by the author for the reconstruction of distribution densities with many modes [5] is supplemented by stochastic verification: the analysis of "simulated" distributions with known properties.

For this purpose based on a number of fundamental work results there were carried out tests of samples generated under programs developed by the corporate authors under the direction of Professor B.Y. Lemeshko (NSTU, Novosibirsk) - following different distribution laws as well as their identification as per the suggested approach. The sample volume L had herewith a range of 100-600 units.

The random variable samples were studied, which related to theoretically quite continuous distributions: gamma-, Cauchy, exponential, normal, Weibull distribution.

Samples of discrete readings of “reference” distribution densities were studied but in fact there were excluded from the consideration the mechanisms for the generation of singular distributions typical to theoretically fractal realizations (Cantor "dust" etc.).

For example, such factual data may be obtained by using the EEG (Electrogastroenterography). They possess the persistence characteristic of singular (fractal-like) processes. It is noted in the author’s publications that the use of regularizing methods for the identification of distribution laws combined with prediction algorithms enables also to expand significantly the scope of the analysis for fractal-like structures of any natural, technological, biological, social and other nature.

The research results show that the conclusion persists on the available peculiarities of composite (Cauchy, exponential) distributions, which result in the need for an increasing number of smooth functions to represent the approximate solution: about 20 harmonics for Cauchy, 10-23 for exponential distributions. In the latter case the shape of the distribution curve is structurally improved by increasing the sample scope from 100 to 300 units.

![Figure 1. The Reconstructed Exponential Distribution (L=300 readings)](image1)

A similar conclusion may be also made in normal law at increasing the sample scope. Though, even for $L = 100$ the goodness-of-fit criteria give a close fit of the sample to the Gaussian distribution.

The proposed method fundamentally distinguishes the main difference for samples generated as per the normal type (scopes were studied with $L = 100, 200, 300, 600$), that is the number of expansion terms $N$ is low and is within the range of $N = 2-4$. The solution expansion coefficients in the system of trigonometric functions for the sample of 200 units have, for example, the following value: $\alpha = 0.01393; \beta = -0.83330; \gamma = -0.50181$.

![Figure 2. The Reconstructed Normal Distribution (L=300 readings)](image2)

The authors suppose that this is a fundamental feature for this type of distributions. The identified property of random variables with the Gaussian probability density distribution may be used as an additional argument in the identification of the distribution law in applied mathematical statistics. The "bell" form of the Gaussian distribution and the three sigma rule are complemented by new identification criterion.

3 Regression approach

To complement identification algorithms for multimodal stochastic structures, the following regression approach is suggested, which takes into account the identified distribution laws.
The fact is that a researcher prior to constructing any mathematical model has a priori information on the interval of hypothetical or practically accepted norms for desired variable $x_i$ (usually assumed with a significant margin). Such norm may be a permissible instrumental error, limits of the physiological norm for clinical evidence, the stability zone in adjustment, the degree of the technological component quality (the said concept is not to be confused with the norm of elements for normalized spaces of the functional analysis). This a priori information on the parameter norm is taken as the basis for the proposed method. Let us consider its essence on the example of immunological clinical tests within the framework of the suggested regression model.

We assume normal ranges for each indicator according to recognized intervals. We convert an initial data table to a binary-form table, i.e., we compress the initial information in a special way. In this case, in each binary matrix cell the figure of one is recorded, if the resultant value of the initial cell is within the normal range, and zero is recorded, if otherwise. Zero values for binary matrix elements are resulting from those readings, which are above or below the norm. The actual testing of the method was carried out with a binary matrix for several dozens of patients in 12 parameters of blood, lymph, hormones. There were obtained unique vectors formallyizing the status of patients with respect to the “parameter norm” characteristics.

The coded data differ from the original laboratory data in the most comprehensive provision of information about an object just by two gradations. As shown below, it gives the considerable obviousness and simplicity for the analysis in the "patient-index" plane, forms in a naturally way algorithms and methods for the immunological data processing. Such binary matrices are used in the graph theory.

The regression coefficients shall be determined from the matrix equation $A\mathbf{z} = \mathbf{u}$ (the system coefficients - are binary digits) with regularization algorithm. The identified parameters belong to the set $\mathbf{R}^1$ and assign the mathematical model of a process or structure.

Further, if the norm-limits of the index are very wide, they may be "compressed" by using the form of the previously identified distribution density for each immunological parameter. In this case, information is used on the availability or lack of higher modes, their offset to one or another direction of the parameter norm, calculated point estimates in the form of the mathematical expectation and variance. Such compression may be performed based on the Jaynes’ maximum entropy principle or otherwise. Then, the algorithm for determining the regression coefficients is repeated.

Let us compare the proposed interpretation method to the classical regression model in the standard Euclidean space of the initial data. In the both cases, the linear regression recovery occurs on a sample of short length ($L = 50 \ldots 100$).

3.1 Classical Regression Model Based on the Initial Data Matrix

1. The linearization of the model of a phenomenon under study based on a volitional decision.
2. The method for the recovery of sample regression coefficients by the linear regularization methods in case of ill-conditioning system.
3. The ranking of the solution components in a natural way – by small values of the eigenvalues corresponding to these components; the assessment of the each individual factor actual contribution to the output value is generally hindered.
4. There is no possibility for the direct functional relationship of this regression model with hypothetical or practically accepted norms for each variable $x_i$ (permissible error range, the limits of physiological norm for clinical evidence, the stability zone in adjustment, etc.).
5. The introduction of the quantitative characteristics for the solution components, the variable regression and factors is hindered from the position of a universal and formalized approach if the information theory.
6. There is only one degree of freedom in the study of the behavior characteristics of the subject under study – an increase in the sample length.

3.2 Regression Model Based on the Binary Coding of Input Data Matrix - Capabilities and Peculiarities of the Approach

1. The linearization of the model of the phenomenon under study in a natural way is the display of input data into binary space $D[0,1]$ according to the sign of belonging to the field of norm $x_i$.
2. The mode for the recovery of sample regression coefficients is by the linear regularization methods under ill conditio sine qua non (ill conditioning).
3. The regression recovery by binary matrices occurs with the certain decorrelation of sample data (both on lines and columns of the matrix operator), thus, increasing the degree of the problem correctness.
4. The availability of the natural bond of this regression model with hypothetical or practically accepted norms $x_i$.
5. The second degree of freedom is added in the study of the behaviour characteristics of the subject of research-the change in the length of the field of the norm for each variable model (usually the contraction) on the input principle of the error regularization - the principle of Pareto (80%-20%), Jaynes or any other.
6. The said approach implements a new operator on $D$ (binary matrix), the relevant regression model and it checks the properties of the object or phenomenon under study for its stability with respect to the norm of the subject parameter.
7. Such functions derivative from the "matrix of norms" are introduced in a natural way as the monitoring frequency of the number of norms on the input parameter vector, the model arguments in the sample series, their cumulative characteristics as well as the “measure” of the output value amount in binary units of the norm.
8. The ranking of the solution components apparently occurs in the following combination: a) a natural
component of the rank – by small values of the eigenvalues corresponding to these components – as per the method of solving the problem based on conditionality, b) the assessment of the partial contribution of each factor to the output value pursuant to the homogeneity of the binary description of vectors of input parameters and the output value.

9. The introduction of the quantitative characteristics of the variable regression and factors becomes possible from the viewpoint of the information theory – each model factor for estimates of probability for occurrence of ones and zeros is considered as the source of messages with respective entropy $H$.

The proposed method is in general enhances greatly possibilities for the correct, formalized and systematic study of phenomena in interpretation problems in engineering, economics, medicine and other fields.

Acknowledgement

The work was performed with the financial support by the RF Ministry of Education and Science (Agreement №14.577.21.0098 on Granting Aid Grant of 08.26.2014, Unique Project IdentifierRFMEFI57714X0098

References

5. V. Kulikov, Applied Mathematical Sciences, 9, 6827 (2015)